UPPSALA
UNIVERSITET

MASTER THESIS, 30 ECTS

Calibration in Urban Building Energy Modeling

Naeim Rashidfarokhi

Supervisors: Dr. Xavier Faure, Dr. Oleksii Pasichnyi
Examiner: Assistant Professor Salman Toor
Reviewer: Professor Alexander Medvedev

Thesis report
Department of Information Technology
Uppsala university

July 2021

Abstract

This thesis work tries to study the annual calibration process for building performance simulations. The bottom-up modelling is a tool to predict energy consumption in Urban Building Energy Modeling (UBEM) which is a common approach to find the most appropriate retrofitting strategies or to monitor energy use for a group of buildings (an archetype), a neighbourhood or a city. To do this, one needs to have local climate data for a site, building geometries, construction assemblies, usage schedules and HVAC system information for each building in that site. But many parameters can be uncertain or completely unknown and not all of them are equally influential. Therefore, the aim is to find representative values for affecting but unknown parameters. To do this, first a modeller needs to identify which parameters affect the thermal behaviour (energy usage) of buildings. Later instead of considering one value for each parameter, the modeller should define distributions for those parameters based on experience. The idea is to perform enough number of random simulations from parameter ranges to find the most representative areas of each influential parameter. In this study, two methods are identified to be useful for sensitivity analysis, Morris and RDB-FAST. Later a brute-force search calibration process is developed to find a good representative average set of parameters for a group of buildings, based on their annual energy consumption. In the end, to reduce the computational complexity of the calibration process an intuitive approach based on normal transformation of calibrated parameters is used to recalibrate the parameter ranges. The study shows that by starting from calibration with a low number of simulations, the sampling from the result of a recalibration can reduce the percentage of deviation between measurement and simulation and increase the number of matches for parameter combinations in comparison with the sampling from only one calibration process.

Acknowledgement

I would like to thank Dr. Xavier Faure and Dr. Oleksii Pasichnyi, the project supervisors, for their constant guidance and constructive regular feedback throughout every step of the thesis. The completion of this work would not have been possible without their support.

And finally, I sincerely thank Kimia, my wife, which without her encouragement and belief in me, the study of a second master was not possible.

Naeim / naeim@posteo.net

Contents

Nomenclature

APE Absolute Percentage Error

BEM Building Energy Modeling

BPS Building Performance Simulations

CLT Central Limit Theorem

DHN District Heating Network

EPC Energy Performance Certificate

LHS Latin HyperCube Sampling

LLN Law of Large Numbers

MCMC Monte Carlo Markov Chain

SA Sensitivity Analysis

UBEM Urban Building Energy Modeling

1 Introduction

Global warming phenomenon, with its consequences as climate change, has made it inevitable for human beings to take actions and reduce fossil fuel energy consumption. In addition to industry and transportation, a major target sector in human societies which consumes a considerable amount of energy is the building energy sector.

According to the International Energy Agency (IEA) to achieve net-zero emissions by 2050 [1], close to half of buildings in developed countries and one-third of buildings elsewhere in the world should be retrofitted by 2030.

This ambitious goal demands decision making tools for municipalities and local energy providers, in the building sector, with the ability of predicting spatial and/or temporal building energy consumption in a city. Such a tool can be used for:

- the comparison of different energy saving scenarios to identify more beneficial renovating strategies for existing buildings in an urban scale,
- and to provide a bigger picture for energy management by locally storing, producing and distributing it in a network with less possible amount of energy loss.

In this regard, there are two major categories of models to gain knowledge about the energy ecosystem in a city:

1. Top-down approaches which provide a total and high-level statistical overview, trend or correlation of energy consumption in a city with respect to population or other activities. These models essentially do not work with technical aspects of buildings. Hence, in absence of a large measured energy dataset with hour, month or annual resolutions for all buildings in a city, they can not depict details of the energy ecosystem or predict possible savings for different retrofitting strategies, see [2].
2. Bottom-up approaches, as engineering models, are based on simulation of each building with a dynamic thermal model. In principle, the same concept to simulate energy performance of a building but in a large scale (neighborhood, district or city) should be applied.

Figure 1 provides an overview of the two above approaches of Urban Building Energy Modeling (UBEM). The bottom-up modelling provides a tool for simulation of buildings at large scales.

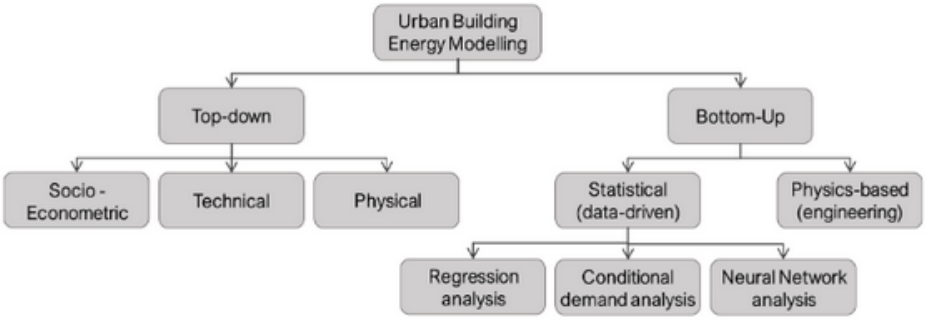


Figure 1: Schematic of the UBEM approaches, see [3]

Growing number of sensors and rapid increase of computational power has created new opportunities for bottom-up modelling. But, because technical and some specific data is rarely available for all buildings (envelope parameters, equipment characters, etc.), bottom-up models face challenges to tackle unknown simulation parameters.
Simulation of a building needs two categories of parameters, geometric and non-geometric ones. These parameters:

- geometric: correct building body dimensions and orientations,
- non-geometric: construction materials for external surfaces including roof, walls, windows, etc, electrical and mechanical parameters related to ventilation and thermal systems and internal loads simplified as maximum heat emission from occupants, lights and electrical devices with respect to some schedules such as constant or hourly varying.

To simulate a large number of buildings in urban scale (UBEM), beside the data accessibility problem, it is not practical to use individual parameters for each building. Therefore a clustering approach is usually applied to categorize buildings into different archetypes.

An archetype is defined to represent a group of buildings which are similar to each other with respect to some energy related parameters. This similarly among buildings in one archetype is enough to cover their thermal performance simulations with one set of common parameters. Any new parameter with an important impact on building thermal simulation can impose defining another archetype (for building thermal behaviour see Building physics). In other words, first, buildings need to be classified into different archetypes or groups and then characterized with a set of average representative values for all buildings in that archetype or group.

In Building Performance Simulations (BPS), lack of reliable input data is usually compensated with the judgement of a modeler in a deterministic assumption (with one value). But in UBEM, applying the same approach for all thermal models propagates the assumption error into an urban scale. Non-geometric parameters are the common source of uncertainty such as occupant related parameters but, geometric parameters due to lack of knowledge or natural deterioration can be uncertain as well, examples are infiltration rate, window's U-value, etc.

In addition, application of a deterministic assumption for an uncertain related energy parameter in UBEM means having the same behaviour for all buildings in an urban scale which leads to ignoring a considerable diversity for that parameter and consequently an unrealistic simulation result. An example can be temperature setpoints or number of people or electrical devices in buildings. To solve the problem with deterministic approaches, probabilistic approaches are suggested with possible calibration methods.

This thesis-work is concerned with Physics-based molding of UBEM under bottom-up approaches. The work is inspired by the annual calibration of probabilistic parameters introduced by Cerezo et al., see [2] but with some differences explained in Section 3.
The study aims to identify influential parameters and to highlight high probability areas of parameter ranges.
To achieve this the following objectives are addressed:

- suitable Sensitivity Analysis (SA) methods for building energy analysis,
- annual calibration of parameters based on a brute-force search comparative algorithm fed with random values from defined ranges of unknown parameters.

2 Theoretical background

As mentioned in Section 1, an archetype is a simplification to address many similar buildings in one place. As a result, a compromise or balance between accuracy and complexity of models (including computational resources) is achieved.

The quality of the results in the Physics-based approach in UBEM, depends on finding the most representative parameters for an archetype as a group of buildings. To provide a representative value for an uncertain or unknown parameter, deterministic or stochastic approaches based on probability distribution of parameters are in use, see Probability distributions. This is called characterization of an archetype. But before expanding characterisation approaches, the classification of buildings into the best representative groups or archetypes is vital.

One way of classification of buildings is based on energy related properties of buildings, available for all of them, such as building application (residential, commercial, office, etc), renovation age, type of heating systems, statistical characteristics of a population, etc, see [2]. Window to wall ratio, floor area and shape topology can also be used as classification indicators if the 3D construction of models does not have a good level of details.
But, except the typical energy properties mentioned above, measured energy data with good time resolution and the characteristics of the problem at hand can determine relevant indicators for building classification as well. This leads to customized grouping of buildings for a specific area or a city. In this respect, Pasichnyi et. al [4] studied the classification of buildings for two scenarios. First, the future effects of mass retrofitting plans to improve energy efficiency for buildings connected to District Heating Network (DHN) and second estimating power demand for electrical-based heated buildings in order to release their demand from the distribution power grid in Stockholm city, if needed.

In the first case, they used building usage, age and having district heating connection as the only source of energy. By this set of properties, they classified all buildings into three archetypes (two multi-residential groups and one office group). For the second scenario, a balance point between complexity and applicability of the model, by try and error, was derived. The complexity of such a model lies in the usage of different appliances working with electricity in addition to district heating to provide energy for buildings combined with or without other sources like fossil fuels, biofuels, natural gas, etc. At last, six different archetypes were developed to cover the diversity of all buildings for the second scenario, see [4].
At this point, after any suitable approach to classify buildings, energy simulation parameters for each archetype should be defined. Table 1 shows typical approaches for characterization of archetypes with respect to uncertain parameters.

As shown in Table 1, due to the lack of detailed documentation and measurement for existing buildings, many of the non-geometric parameters with stochastic behaviour are uncertain or unknown. Among them, the last six rows as their peak values and schedules[1] to be used in energy models are called internal loads. These parameters show human related factors which dramatically affect building energy demands, especially in residential buildings, but Cerezo et al. showed that in an annual average or total energy demand for a district, with a reasonably low resolution, the last six parameters does not typically have a large effect and can be characterized deterministically, see [2].

[1] At the annual scale, uncertainties due to schedules does not have a relevant impact on thermal performance of buildings.

Table 1: Applicability of deterministic and probabilistic approaches to different type of parameters

Parameters	Deterministic approach	Probabilistic approach
Envelope (wall, roof, slab, window, skylight, door) construction details such as size, orientation and thermal properties for materials (geometric type)	✓	✓
HVAC system properties such as temperature setpoints, heat recovery efficiency, fan efficiency, heat pump COP, etc	-	✓
Infiltration rate	-	✓
Domestic Hot Water	-	✓
Occupancy rate	-	✓
Lighting	-	✓
Electrical devices	-	✓
Schedules for occupancy	✓	✓
Schedules for Lighting	✓	✓
Schedules for devices	✓	✓

The impacts of parameters in Table 1 on energy demand are different, especially when different temporal resolutions such as annual, monthly or hourly are of interest. While a modeler can characterize all types of internal loads and their schedules with one constant value in a deterministic way for annual, monthly or hourly simulation, the realistic energy consumption and power demand need to consider them in a probabilistic way to cover the diversity of building usage by different people, especially in hourly simulations in an urban scale.

All or any of the parameters in Table 1 can be assigned in a deterministic or probabilistic way[2] but in any case, the selected approach for a parameter can be in a simple or customized/-calibrated manner, see Table 2. While the first two deterministic methods use single values for

Table 2: The nature of assigned values and their source in each method type, see [2]

Deterministic approach (single value)		Probabilistic approach (distribution/range)	
Simple type	Customized type	Simple type	Calibrated type
Literature, national building code, standards, modeler's judgement	Building surveys, individual parameters	A distribution based on literature and building surveys	Manipulation of a distribution based on a learning set of metered data

uncertain parameters[3], the two second probabilistic methods use a range of values or distribution for them. The last probabilistic method can use Bayes theorem or optimization techniques to

[2]Practically, geometric parameters can also be uncertain but, to do a useful and feasible modeling and for convenience, they are the minimum requirement in a model.

[3]Certain parameters are always and simply represented by a single value in energy modeling.

find areas with higher likelihood in parameter ranges, see Probability and likelihood.

The validation of accuracy for all approaches in Table 2 needs measured data. Based on the resolution of measured energy data, the results of UBEM simulations can be used in different time steps such as yearly, monthly, etc. In the same manner, the resolution of spatial simulations can differ from block, neighborhood, district or a whole city (with respect to the study at hand such as testing energy efficiency policies against energy supply strategies in a district). But there are other challenges.

In a Building Energy Modeling (BEM) attempt, Yousefi et al. [5] studied the influence of occupants' lifestyle on energy performance of one single multi-family building by using input from surveys (the deterministic customized approach). They showed in their study that the occupants' behaviour can affect the strategy of choosing materials for envelope renovation. In this type of study, unknown parameters should not represent more than one building therefore the modeler seeks a solution for that particular building.

In the case of UBEM, to calibrate uncertain parameters the more similar buildings in the training set the better calibration can be done, since otherwise there is a risk of biased results. In other words, the calibrated parameters should be generalized enough to represent all similar buildings (as an example in an archetype, block or neighbourhood).
On the other hand, running more random simulations from parameter ranges guarantees to not miss any influential part of unknown parameter distributions. But, this poses another problem which is the need for intensive computational resources. Many studies have tried to address this issue by surrogate modeling or metamodeling techniques, such as [6].

In general, a model is the representation of a phenomenon in the real world or nature while a metamodel or surrogate model can be a model of that model, i.e a simplification of an actual model. A metamodel can be a mathematical equation or a set of equations to relate input data to outputs to show the original model's behaviour. So a metamodel tries to highlight the properties of the main model.

In BPS and UBEM, instead of simulating all possibilities for individual buildings with different parameter choices, a statistical meta-model can be used to predict the behaviour of a bigger portion of buildings or parameter changes. This technique helps to reduce the number of simulations and hence the time and resources for computations. Østergård et. al [7] showed that among six popular meta-modeling techniques namely as: linear regression with ordinary least squares (OLS), Random Forest (RF), Support Vector Regression (SVR), Multivariate Adaptive Regression Splines (MARS), Gaussian Process Regression (GPR), and Neural Network (NN), GPR performs better with respect to accuracy, efficiency, ease-of-use, robustness, and interpretability.

In a study, Nagpal et. al [6] applied RF and NN to reduce the number of simulations for calibration of unknown parameters. Instead of running hundreds of thousands simulations, (brute-force simulations) to form a joint distribution from passed parameters and then later raw-sampling from them, they used machine learning algorithms to estimate trends from passed parameters with much lower numbers of simulations. These trends are based on the behaviour of simulations (buildings' thermal modelling results) with each combination of parameters in comparison with measured energy data for buildings in the training set. They claim that the RF model performs better than NN, especially when the number of unknown parameters are higher. This methodology yields accurate estimates 500 times faster than brute-force simulation cases but

when envelope building parameters are known.

This thesis investigates two stages, Figure 2, in UBEM that addresses uncertainty of input parameters; sensitivity analysis with Morris and RDB-FAST methods and annual calibration of selected uncertain parameters based on the interpretation of Bayes theorem proposed by Cerezo et al, see [2].

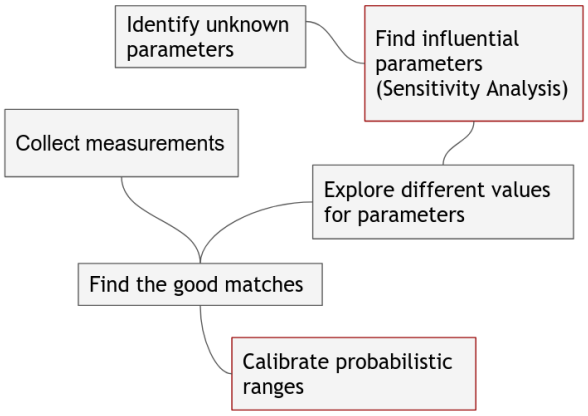


Figure 2: Typical steps in UBEM engineering approach with highlighted objectives of this study.

2.1 Sensitivity analysis or what-if analysis

Sensitivity analysis is the process of identifying how changes of independent variables as inputs may impact a dependent variable as the output or behaviour of a system under a given set of assumptions.
To do the analysis three parts should be known:

a. variables or the set of parameters to vary within a specified range,

b. how to vary the parameter(s),

c. and the objective variables in the output to observe.

Sensitivity analysis in total, and consequently in building energy analysis, can be divided into two approaches, local analysis and global analysis.
Local sensitivity analysis (or differential sensitivity analysis) is the investigation of changes around a base case (a specific building state) with respect to uncertain inputs while the global analysis investigates the impact of uncertain inputs over the whole input space, see [8]. Therefore, global analysis is a computational heavy task in comparison with local analysis because the investigation domain is much broader than local analysis and sampling techniques are needed to generate combinations of inputs.
In Global sensitivity analysis (GSA), the contribution of changes in the output with respect to changes in an input quantifies by averaging the other inputs instead of fixing them, see [9].

The global sensitivity analyses can be divided to four main categories:

a. regression methods such as: Standardised Regression Coefficients (SRC) for linear models or Standardized Rank Regression Coefficient (SRRC) for nonlinear models,

b. screening-based methods such as: Morris method,

c. variance-based methods such as: Fourier Amplitude Sensitivity Test (FAST) and Sobol methods,

d. and meta-modelling approaches such as: Multivariate Adaptive Regression Splines (MARS)

These methods need generation of samples. Usually, for regression (a) and meta-model (d) approaches, Latin Hypercube Sampling is very popular while for screening (b) and variance-based (c) methods other sampling schemes such as 'Saltelli'[4] are used.

To run a sensitivity analysis in UBEM, the same typical steps can be followed as for thermal performance analysis for one building, see [8], which are:

- define the input variables with their corresponding probability distributions as continuous uniform ranges to have equally probable samples,
- create and run building energy models with samples and collect simulation results,
- run sensitivity analysis based on inputs (samples) and outputs (results corresponding to samples from simulated energy models),
- and finally present the results by help of scatter, bar, pie chart, box or spider plots.

Usage, advantages and disadvantages of different approaches in sensitivity analysis:

A. When a current state of a building, with all parameters affecting the energy performance, is defined local sensitivity analysis can be used. In fact the study is around how and to what extent the building reacts when a parameter changes over its entire domain, while all other parameters are kept fixed at their initial mode. This type of One Factor At a Time (OFAT) study has the advantages of being easy to interpret the results and cheap in terms of simulation runs. But it only covers a limited input space around a base case and it does not provide any information about how input changes interact with each other. A major drawback of this approach is that the local analysis result can be misleading if the model is of unknown linearity [10] and variation of input parameters affect each other.

B. Global sensitivity analysis methods evaluate output changes with respect to parameter (input) variations by varying all of them over their entire ranges at the same time. Thus, they are able to provide qualitative and/or quantitative measures by ranking input parameters even for non-linear models.
The characters of four groups of global analysis methods in BEM are:

   a. regression methods are the mostly used sensitivity methods in building performance analysis. In this class among many indicators such as t-value, F-value and change of R^2 (see R^2, t-value and F-value), the sensitivity indexes SRC and SRRC after Monte Carlo sampling are widely used for sensitivity purposes in buildings, see [8]. This method is computationally cheaper than other global methods and the results are easy to interpret, but they are not very powerful with correlated inputs. This class of global methods, because of their statistical nature in contrast with other classes, can be used for not only simulation-based studies but also for observational studies.

[4]Source code to Saltelli's scheme: https://salib.readthedocs.io/en/latest/$_{m}odules/SALib/sample/saltelli.html$

b. the most important screening-based method in building performance studies is the Morris method. The main usage of this method is to identify a few important inputs among a large number of parameters but with low computational cost. The problem with this method is that it only provides qualitative measures and identifies important parameters without any measure to show to what extent each parameter affects the output.
Morris method provides two sensitivity indices:

- mu value (μ) as the effect of each parameter on the output,
- and sigma value (σ) as the interaction of each parameter with other parameters or nonlinear effect of the model,

The total effect of a parameter on the output can be calculated as the sum of the two effects above under a term as mu-star μ^*, see [8].
Morris method performs best when the number of most influential parameters is low compared with the total number of parameters, see [11].

c. In direct contrast, variance-based methods can quantify the effect of each parameter on output, i.e. they can compute each parameter's share on the uncertainty of the output.
They provide two sensitivity indices:

- the 'first order effect' measures the effect of corresponding input on the output,
- and the 'total order effect' represents the total output variance which is a combination of 'first order effect' and the interaction of the corresponding parameter with other parameters (known as 'higher-order effects').

In the process of finding critical parameters for UBEM calibration, the solo effect of each parameter on the energy performance of the building (output) is important. But the final interaction of each parameter not only on the output but also with other parameters should be considered.

This class of methods can deal with nonlinear but non-additive[5] models with the cost of much higher computational resources compared to other global methods. The two well-known methods in this class are Fourier Amplitude Sensitivity Test (FAST) and Sobol where Sobol is more powerful in terms of identifying all interaction effects and more computationally expensive,

d. and the final class of global methods are meta-models which first make a meta-model by regression methods and solve it by variance-based methods. In the case of a pure variance-based method like Sobol (c), there are needs for many detailed building energy simulations. But using a regression technique, or any other machine learning technique to estimate some input and output in the first step of meta-modeling, can reduce the number of simulations considerably. Some well-known methods in this class are Multivariate Adaptive Regression Splines (MARS), Gaussian Process (GP) and treed Gaussian process (TGP), see [8].

[5] An additive model is a nonparametric regression model, meaning that there is no predefined form to predict a relation between the data, but the form is derived from the relation of data itself. As additive in word means 'something which is added', a predictor as a form is added to the model based on sample data. On the other hand, a non-additive model is a parametric model with its own significance.

All sensitivity analysis methods mentioned above provide discrete values as sensitivity indices while it is possible to provide variations (intervals) for indices by applying bootstrapping, in order to have robust results.

Now, which method should be used?
When input parameters are correlated, it can cause collinearity which means they are aligned objects or they are in a line or row. This phenomenon prevents the parameters from independently showing their effect on the output of a model. Similarly in sensitivity analysis for building energy performance, collinearity makes large variances for some regression coefficients. Thus in presence of correlated parameters, regression methods such as SRC or t-value index cannot be used. In this case, other methods such as 'conditional variable importance from random forest' [8] can be used.

To choose appropriate SA methods, Kristensen et al. [12] studied the performance of local, Morris and Sobol analyses combined with two BEM models with hourly and monthly calculations. They found, regardless of provided input parameter resolutions such as monthly or hourly to BEM models, all methods could find similar clusters of sensitive parameters for linear models but with different rankings, knowing the fact that a complicated model such as Morris and Sobol can work with non-linear models as well.

To investigate the effectiveness of different sensitivity analyses with respect to non-linearity and level of uncertainty in energy related parameters and computational cost, Menberg et al. [13] compared Morris, linear regression and Sobol methods. They found the result of all three methods in good agreement. And compared with the computational expensive analysis Sobol (which can find higher order effects and parameters' interactions) Morris method could provide basic information about parameter interactions as well.

On the other hand, the choice of distribution type for uncertain parameters with the highest impact on an energy simulation depends completely on the purpose of the simulation, see [2].
When the purpose of the simulation is to compare different design options, a uniform distribution is more suitable for building's unknown parameters. If the purpose is to simulate the thermal performance of an existing building with respect to unknown parameters, normal distributions may suit better for most of the variables, see [8].
Since the idea in calibration is to find a value close to reality for a parameter in a given range with similar probabilities, a uniform distribution works better. This is because no more information exists to form a normal distribution and the parameter value is completely unknown or highly uncertain.

For uniform unknown parameter distributions (called non-informative parameters), Morris method can cluster and rank sensitive parameters similar to Sobol method but with much lower computational cost, see [12]. Therefore in this study, sensitivity analysis starts with Morris analysis. Note that for the Morris method, one can consider the two main approaches proposed by Morris (1991) and the extended version by Campolongo (2003).

In the original version, μ and σ indexes are determining factors as the former index measures the overall effect of a parameter on the final output and the latter estimates the ensemble of the second and higher-order effects in which the parameter is involved, i.e. interaction with other parameters or nonlinear effects, see [11]. In the extended version, Campolongo et al. introduced a revised version of μ namely as μ^* as the mean of the absolute values of the elementary effects (see Statistical definitions). This version is very successful in finding influential parameters when

there is a non-monotonic model to analyze, i.e. when the positive and possible negative effects can cancel each other out and we want to avoid it. This helps to have only one measure as μ^* to detect the overall influence of each parameter on the output, see [14].

To find which approach of Morris analysis suits in this study, four possible cases can happen:

- low μ and low σ shows little influence of input parameter on output and little interaction with other parameters,
- low μ and high σ shows little influence of input parameter on output but either a high nonlinear relationship with the output or high iterations with other inputs,
- high μ and low σ shows a parameter is influential and has a linear relation with output and no important interactions with other input parameters,
- and finally high μ and high σ shows an influential parameter with either a high nonlinear relationship with the output or high iterations with other inputs.

In terms of prioritization of influential parameters, when parameter ranges are large and non-linearity interactions exist among parameters the usage of μ^* is misleading. In this respect, Menberg et al. [13] observed instability of Morris method with commonly used measures for parameters and proposed to use the absolute median of elementary effects (χ^*) instead of the absolute mean value of elementary effects (μ^*).

SALib package [15] provides all three measures (σ, μ, μ^*) and since the result of Morris analysis is qualitative [11], μ^* can serve as a sign of nonlinearity and higher order of interactions among parameters. But since the first objective of this study is to detect the influential input parameters on the output, μ factor is used to find parameters with high effects.

Later in this study, sensitivity analysis continued with a Random Balance Design (RBD) extension of FAST analysis. Fourier Amplitude Sensitivity Test is a popular variance-based method. Nguyen et al. [16] found that the results of FAST and Sobol methods have consistency. By using a periodic sampling approach and a Fourier transformation to decompose the variance of an output for a model into partial variances, the method determines the contribution and effect of each parameter on the output by looking at the corresponding variance for each input parameter, see [17]. In other words, the method not only finds the ranking for the input parameters, but also quantifies the magnitude of effect for each parameter on the output.

The periodic sampling approach can be done in traditional search-curve based sampling, simple random sampling or random balance design sampling, see [9]. The latter approach which is of interest in this work, RDB, significantly reduces the computational cost of the analysis by efficiently taking samples from parameters' input space, see [18].

2.2 Calibration with Bayes theorem

Cerezo's interpretation, see [2], from Bayesian calibration starts with considering a vector of important but uncertain parameters, θ, each defined as a uniform distribution, ready for stochastic sampling. Taking a number of samples to perform building energy modeling yields simulation results named as vector $\mathbf{y}$. Later, measured energy data as vector $\mathbf{d}$ is used to identify the right area of distributed parameters in the initial ranges with respect to an Absolute Percentage Error

(APE) named as α parameter. The APE should be satisfied by a maximum threshold for measured data $\mathbf{d}$ and simulation results $\mathbf{y}$. In this way, the method modifies the initial probabilistic ranges and gives different weights of importance to different parts of distributed parameters.

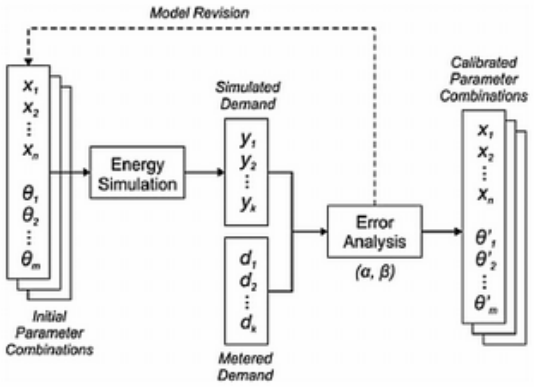


Figure 3: Bayesian calibration, Cerezo et al. [2] interpretation.

Figure 3 shows the schematic process of calibration where vector $\mathbf{x}$ represents known deterministic parameters, $\theta^{'}$ represents calibrated parameters, α is the maximum allowable APE between simulation and measurement values such that: $\mid \mathbf{y}(\theta) - \mathbf{d} \mid < \alpha$ and β is the percentage of all buildings which must pass the APE evaluation based on α.

In their test case, Cerezo et al. [2] consider four uncertain parameters and five number of samples yields 625 equally likely combinations. To find calibrated parameters for a block of buildings, they simulate each building with all 625 combinations. The definiation of APE to check the deviation of simulated energy consumption with the annual metered data for each building is:

$$APE = \frac{|EUI_{mes} - EUI_{sim}|}{EUI_{mes}} \times 100\%,$$

while EUI is the annual Energy Use Intensity for each building

At this point, any combination of parameters with an APE lower than 5% is saved as calibrated ones. They use a value called β (considered as 85%) as the ratio of passed buildings with at least one passed simulation to check the accuracy of the model, i.e. parameter selection and range definition for selected parameters.
Later by making a joint distribution of all calibrated parameters, they use it to pick 100 random samples and simulate unseen buildings in another block as a validation set. Finally the closest result for each building is presented.

Note that the calibration of probabilistic parameters in Cerezo et al. [2] is performed for all buildings disregarding of the archetypes. In other words, they characterize the buildings into four archetypes, mainly to cover the diversity of the deterministic unknown parameters, but they do not perform a separate, customized calibration for each archetype, which may average out details and be the source of losing some accuracy in the results, see [2].

The calibration of unknown parameters in UBEM needs metered energy in buildings, $\mathbf{d}$, from

the population of a particular archetype, a block of buildings, etc. for validation and calibration. This is for transforming the vector θ as samples from prior distribution to $\theta^{'}$ as accepted prior samples namely posterior distribution. In fact, the Bayes rule is used to calibrate the initial uniform distribution of parameters by highlighting areas with higher likelihood. Since Bayes theorem is in the core of this calibration procedure, reviewing it helps to understand this type of calibration better.

When the probability of an event is dependent on the probability of another event, a conditional probability exists and one way of figuring this situation is Bayes' theorem. The theorem looks at these two events with respect to their relationship. So:

- first we know that the probability of having both events A and B is commutative:
  $P(A \cap B) = P(B \cap A)$,
- while conditional probability can be written as:
  $P(A \mid B) = P(A \cap B) \,/\, P(B)$,
- we can rewrite the equation as:
  $P(A \cap B) = P(A \mid B) \,.\, P(B)$ which is called the multiplication rule,
- and similarly:
  $P(B \mid A) = P(B \cap A) \,/\, P(A)$ so: $P(B \cap A) = P(B \mid A) \,.\, P(A)$,

  therefore: $P(A \mid B) \,.\, P(B) = P(B \mid A) \,.\, P(A)$,
- which gives the theorem as:

$$P(A \mid B) = \frac{P(B \mid A) \,.\, P(A)}{P(B)} \tag{1}$$

- which means:
  Posterior-distribution = (likelihood x prior-distribution) / probability of evidence

Elements of Equation (1) for a building energy calibration can be written as:

$$P(\theta \mid d) = \frac{P(d \mid \theta) \,.\, P(\theta)}{P(d)}$$

where:

- P($\theta \mid d$) is the probability of all parameters given the fact that energy calculations with those parameters fall in the range of measured data (d) with respect to APE term called α. Finding this probability is the goal of the calibration process and is called the posterior join distribution of calibrated parameters ($\theta^{'}$),
- P($d \mid \theta$) is the likelihood as the chance of reaching a measured energy, reasonably can be explained given a set of parameters θ. This probability function is not available in the calibration process here and replaced by an aggregated error analysis, shown in Equation (2). In fact, the APE term α is used here to find the magnitude or frequencies for each sample parameter. The logical expression replaced with likelihood term is the boolean function below:

$$P(d \mid \theta) = \begin{cases} 1, & \text{if } \epsilon\,(y(\theta),\, d) < \alpha, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

which is pretty convenient because the good combinations directly fall in the APE margin to form a joint distribution in the next step.

Looking at $P(\theta \mid d) \propto P(d \mid \theta) \,.\, P(\theta)$, intuitively, the peak of posterior distribution or calibrated parameters is between the peak of prior and the peak of likelihood and since the probability of prior distribution is uniform, the maximum likelihood estimation replaced by binomial expression in Equation (2) is decisive for the shape of posterior distribution.

Also if the number of buildings with measured data (d) increases the likelihood term, $P(d \mid \theta)$, decreases and as a result the posterior distribution becomes narrower, see [19]. This shows the importance of having enough measured data, the more the better, for similar buildings in order to find the right weight of different areas in parameter distributions.

Also the other two terms can be explained as:

- $P(\theta)$ as the prior probability of parameters is a uniform distribution to provide equal chances for all values in that range,
- and finally, P(d) as the evidence is the probability of having a particular energy consumption for a building.

But, in comparison with the explained calibration process presented in Figure 3, a more complicated method can use a metadata modeling technique combined with one of the Monte Carlo variants as Chong et al. [20] introduced. Schematically, after collecting measured data and sensitivity analysis, they map the inputs of their Gaussian process (GP) model to the outputs of interest to generate a posterior distribution. Later by (Monte Carlo Markov Chain) MCMC sampling, they test the accuracy and convergence of the results. This type of calibration, including a metamodeling is out of the scope of this study, see Figure 4.

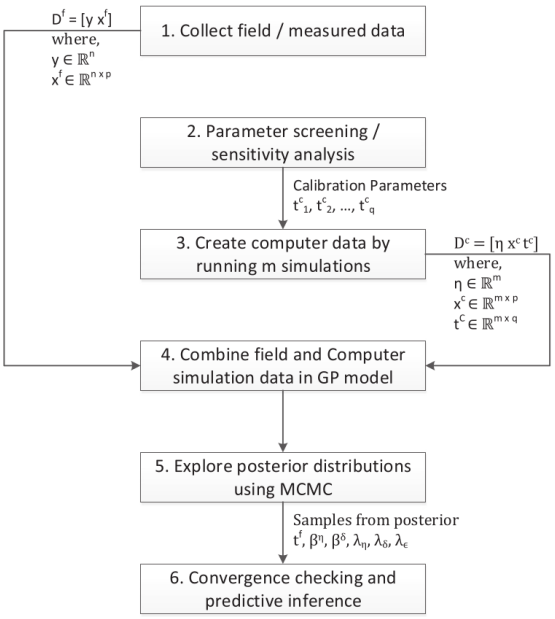


Figure 4: Bayesian calibration procedure, modified by Chong et al, see [20].

3 Methodology

This thesis work is divided into two parts, sensitivity analysis (2.1) to identify the most influential parameters on energy use in any case study and annual calibration development (2.2) based on Cerezo's procedure [2] with a few methodological differences as:

- first, the characterization of buildings is an initial step before the calibration process and is not performed in this work,
- second, the discretization of parameter ranges does not have any arbitrary name as Cerezo's work calls it "sampling size". In other words, in practice, 4 parameters are not sampled 5 times to generate 625 combinations, but each range of parameter is sampled 625 times with Latin HyperCube Sampling (LHS) technique in the first place, see Random number generators. So, LHS is in the center and this thesis uses different number of simulations for two multi-residential buildings to check the validity of the implemented code and to run some tests,
- and the third difference lies in the generation of the joint probability distribution from all acceptable parameter combinations. This work relies on covariance matrix generation to produce correlated random samples from the joint distribution, see [21], while Cerezo et al. [2] do not mention their sampling technique.

This study uses two similar multifamily buildings for both sensitivity analysis and calibrating of the parameters named as buildings 9 and 10, with 24 and 15 meters height correspondingly. Later to test the accuracy of the calibrated parameters, beside buildings 9 and 10 an additional unseen building named as building 12 with 24 meters height is used, see Figure 5.

Figure 5: Buildings 9, 10 and 12 located at Stockholm, Bromma.

The target indicator in this study is total Energy Use Intensity (EUI) which is the combination of heating loads (without hot water) and electricity for building services. Cooling loads are zero for residential buildings in Energy Performance Certificate (EPC) documents and hence neglected from simulation results.

The study is conducted for five unknown parameters with ranges specified in Table 3, while to show the sensitivity results a made up code is used for easier representation of priorities. The

Table 3: Unknown parameters with probability ranges for sensitivity and calibration.

Variable names	Envelope leak	Window to wall ratio	Internal load multiplier	Area based flow rate	Temperature set-point (Winter)
Distribution range	0.4 to 2	0.2 to 0.4	0.5 to 2	0.3 to 0.6	19 to 22
Codes	1	2	3	4	5

implementation[6] in this study works upon a code developed in UrbanT research group at SEED, ABE, KTH with the calculation engine Energyplus, see Building energy calculation methods.

3.1 The accuracy of the matched cases

In this study, a transformation technique called t-SNE, see [22] is used to visualize and examine the clustering of the matched cases. When the range of parameters are selected by experience, any combination of samples can represent an actual state of a building. But when values of parameters have no physical meaning or abnormal, there should be a mechanism to detect them as anomaly cases to exclude them from the calibration process.
Considering the fact that the range of calibrated parameters and their size can vary for different archetypes of buildings, t-SNE plots produced for four perplexities (5, 30, 50, 100) with three learning rates (100, 200, 400). The method uses 'exact' approach for calibrated sample sizes less than 100 and it uses Barnes-Hut approach for bigger sample sizes and the number of iterations is 5000 for any case. For more information, Check the robustness of the results with t-SNE.

3.2 Effective number of simulations

While, the goal in the calibration process is to accurately trim the uniform distributions of unknown parameters and identify influential areas, but it is ideal to do it efficiently i.e. with the least possible number of measured data and the least number of random samples and as a result fewer simulations. In this thesis, finding the minimum number of buildings to represent calibrated parameters for an archetype is not studied. But finding an optimum number of simulations is studied. To do this, a statistical term as Effective Sample Size (ESS) is considered.

In a statistical study, an EES or the 'right size' to represent a target population [23] ensures that a scientific question gets answered correctly. In general and in Bayesian statistics, by Monte Carlo as the sampling technique more samples are needed to avoid clustering and to reach randomness. The reason is that Monte Carlo causes the drawn samples to be correlated. This means that the effective sample size could be lower in the absence of correlation in order to represent a population. For Monte Carlo, with a sufficiently large amount of samples, one can distinguish the normal distribution in the sample set known as Central Limit Theorem (CLT) and where the mean value of the sample set locates known as Law of Large Numbers (LLN), see [23]. But LHS by stratification,covers an input domain with a lower number of samples and CLT is not valid, see [24].

In UBEM, the idea with the 'right size' of sampling from uniform distributions is to reduce the number of samples and as a result the number of simulations for each building. But the sampling resolution should be fine enough to not miss possible influential values (sub-ranges

[6]Link to the developed code: https://github.com/naeimrf/Thesis-Calibration

in the defined range of each parameter). Bayesian calibration method tries to highlight these areas by tuning the frequency of discrete samples based on validating simulation outputs with measurements. A smaller or bigger number of simulations can affect either the accuracy of the results or unnecessarily increases the usage of computational resources. With LHS as sampling technique the amount of sampling points should satisfy LLN, so that the mean value of sample points gets close enough to the mean value of a parameter range, see [23].

In general and as a rule of thumb, 30 sample points is a starting point in some statistical applications. Chong et al. [20] used this idea to generate 30 samples for their Gaussian Process after sensitivity analysis. But in this study, with no surrogate model and with only a pure Bayesian approach, more than 100 samples is considered as a starting point for sampling from parameter distributions. This number may increase with respect to the APE value (α) and the parameter distribution sizes, since if a lower APE between simulations and measurements is needed, more combinations should be examined from uncertain parameter ranges to find better matches (assuming that ranges are selected carefully by experts).

4 Result

Both sensitivity analysis and calibration of uncertain parameters are performed with 4 number of simulations: 126[7], 252, 510 and 1020.

4.1 Sensitivity analysis

The purpose of running sensitivity analysis with different simulation runs is to observe when the results from different sensitivity methods become stable. Table 4 summarizes the priority of different parameters based on different sensitivity analysis.

Table 4: Parameter ranks with two methods and different number of simulations.

Runs/Sensitivity method	Morris analysis	RDB-FAST analysis
126 simulations	5, 2, 1, 4, 3	5, 2, 1, 4, 3
252 simulations	4, 1, 5, 3, 2	4, 2, 5, 1, 3
510 simulations	2, 5, 4, 1, 3	4, 2, 5, 1, 3
1020 simulations	5, 4, 2, 1, 3	5, 4, 2, 1, 3

Looking at Table 4, it is seen that RDB-FAST is able to identify similar ranks after around 250 simulations. Since the first three parameters in this method are similar, one can roughly expect to get accurate results with simulations between 250 to 500 for each building, see Figure 6 & 7.

On the other hand, the results from Morris method (Table 4) do not follow any similar pattern until 500 simulations. This is because, all five parameters have considerable effect on the energy consumption of the case study and they are highly correlated (see σ value in Figure 8). Therefore only considering an index μ with a low number of simulations could be misleading.
In addition, the high correlation among all parameters or non-linearity between parameters and the output, see Figure 9, prevents that Morris method can identify the right ranks based on μ^*.

[7]Please note that Morris analysis needs to have a multiplier of 'number of parameters + 1' for simulation runs, i.e. if you have 5 parameters you need (5+1) $\times$ natural number of simulations. Ex: $6 \times 21 = 126$

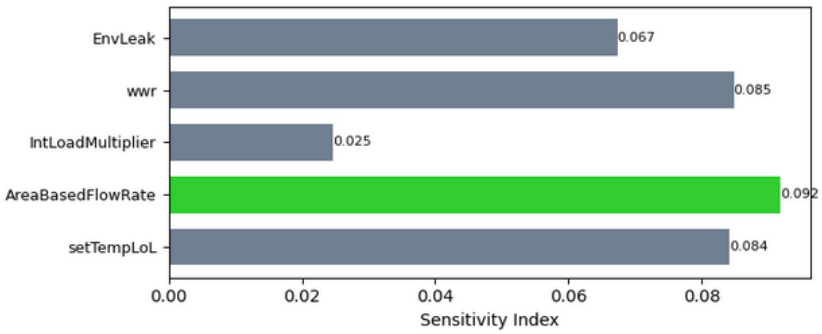


Figure 6: RDB-FAST result with 510 simulations

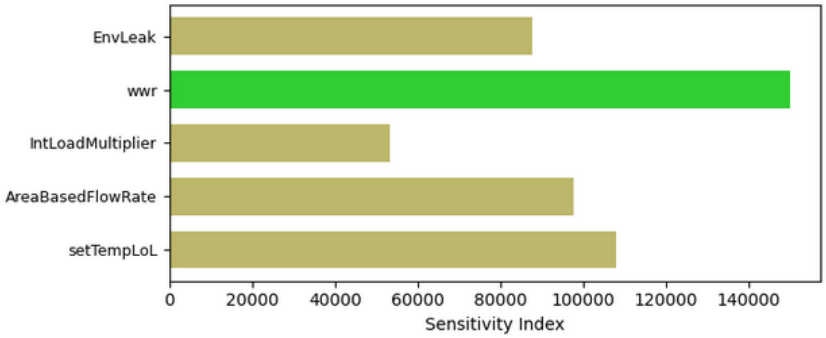


Figure 7: Morris result with 510 simulations.

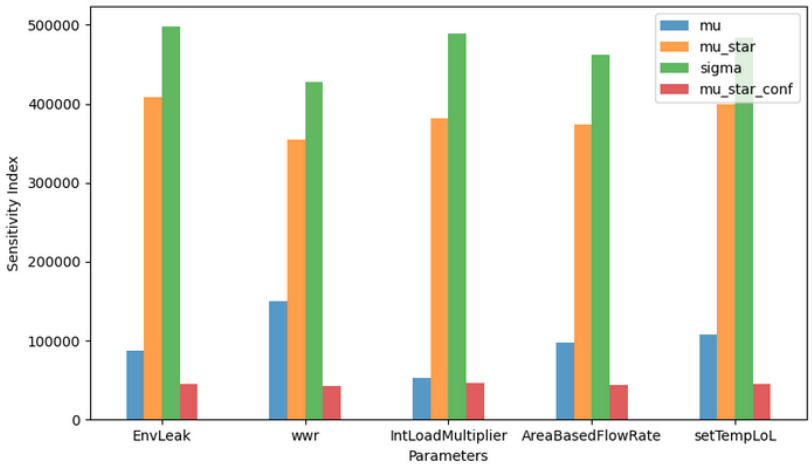


Figure 8: Sensitivity metrics for Morris method in case of each parameter on the output.

Going back to Table 4, after 1020 simulations both methods can identify similar ranks but for a lower number of runs this is usually not true. The reason is that Morris method is a screening based method which performs best when the number of influential parameters in the model is

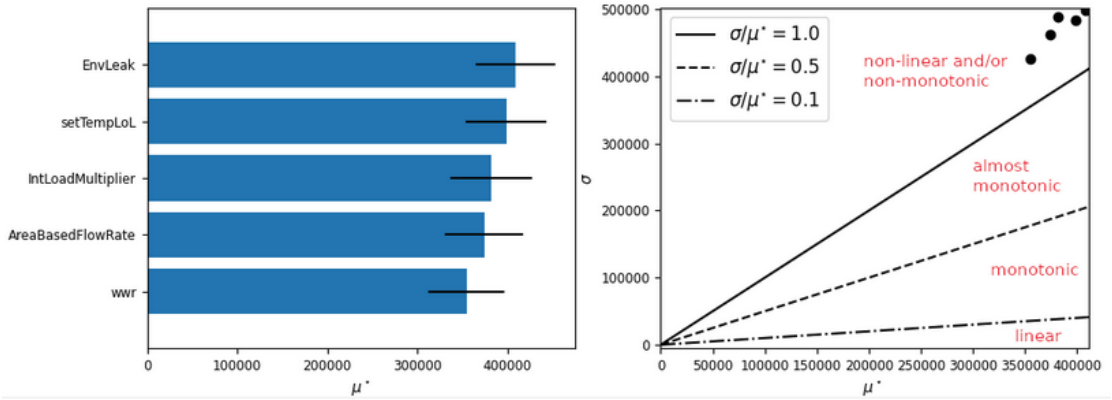


Figure 9: Parameter ranking with Morris's μ^* and 510 simulations.

small compared with the total number of parameters. The way it works is that, Morris analysis in nature is a combination of many local sensitivity runs. Saltelli et al. [11] calls it a composition of individual random experiments around input values when the impact of changing one parameter at a time is evaluated (One Factor At a Time). It means the model output for original inputs are first evaluated and then 'n' number of random points (by increasing or decreasing) around each point is selected for screening. This is defined in the SALib package by *num_resamples*. The output is re-evaluated for these new points and an 'elementary effect' is calculated by dividing the change in the output by the change in the input. Finally, the 'n' elementary effects are aggregated into an average μ and a standard deviation σ.

On the contrary, RDB-FAST is a variance based method which is able to capture the effect of parameters on the output of a model as well as any possible interactions of parameters by decomposing the output variance using Fourier transformation. The FAST method including its variant RDB highlights the importance of correlation within parameters and the index values confirms this as the sum of all indices is roughly 0.35 which means that 65% of the variance is explained by interaction between parameters, see Figure 6.

While the two methods are different in nature, for a good number of samples, their results show consistency to identify important parameters although their parameter ranking could be different, as we see for 510 runs in Table 4.

4.2 Calibration

Five uncertain parameters (Table 3) with four numbers of simulations (Table 4) are tested for buildings 9 and 10. Table 5 shows the number of passed combinations of parameters for two APE qualifications as 1% and 5%.

The results in Table 5 imply a good selected range for all influential parameters. As the number of simulations increases (as well as simulation time[8]), the number of passed cases grows accordingly. This is more obvious in column 'All buildings' under $\alpha < 5\%$. Figure 10 shows all passed

[8]Intel Xeon E312xx, CPU(s): 4, Thread(s) per core: 1

Table 5: Number of passed simulations with respect to absolute percentage error (α)

Simulations per building	Building 9		Building 10		All buildings		Time for simulations[8]
	$\alpha < 1\%$	$\alpha < 5\%$	$\alpha < 1\%$	$\alpha < 5\%$	$\alpha < 1\%$	$\alpha < 5\%$	
126	5	36	4	33	9	69	1.5h
252	24	80	18	80	42	160	2.5h
510	31	160	33	144	64	304	5h
1020	61	321	57	272	118	593	10h

cases for 126 simulations per building.

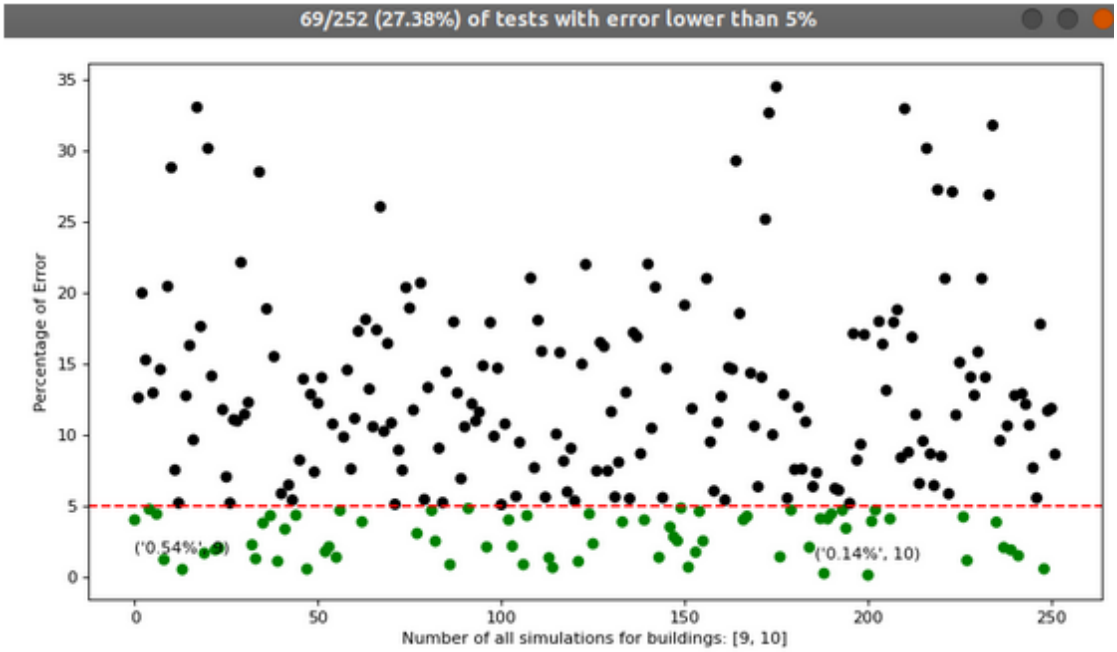


Figure 10: Visualization of Absolute Percentage Error ($\alpha < 5$) between simulations and measurements for buildings 9 and 10 where the lowest APE for building 10 is 0.14%.

Now, a covariance matrix from joint distribution of the passed cases (69 out of 252 simulations) is created. The reason behind this approach is that by sampling from uncertain parameters at the same time, we're simultaneously interested in the effect of change of uncertain variables and we are looking for a relationship among them. Figure 11 shows a grid plot of all relations between pairwise parameters with their possible linear (Pearsonr index) and nonlinear (Spearmanr index) relations.

Looking at Figure 11 the decrement of correlation between envelope leak and window to wall ratio shows that as one of them decreases, the other one can increase to meet the same energy consumption in the buildings 9 and 10. On the other hand, an increment of correlation (positive relation) between the internal load multiplier and area based flow rate might be due to a low

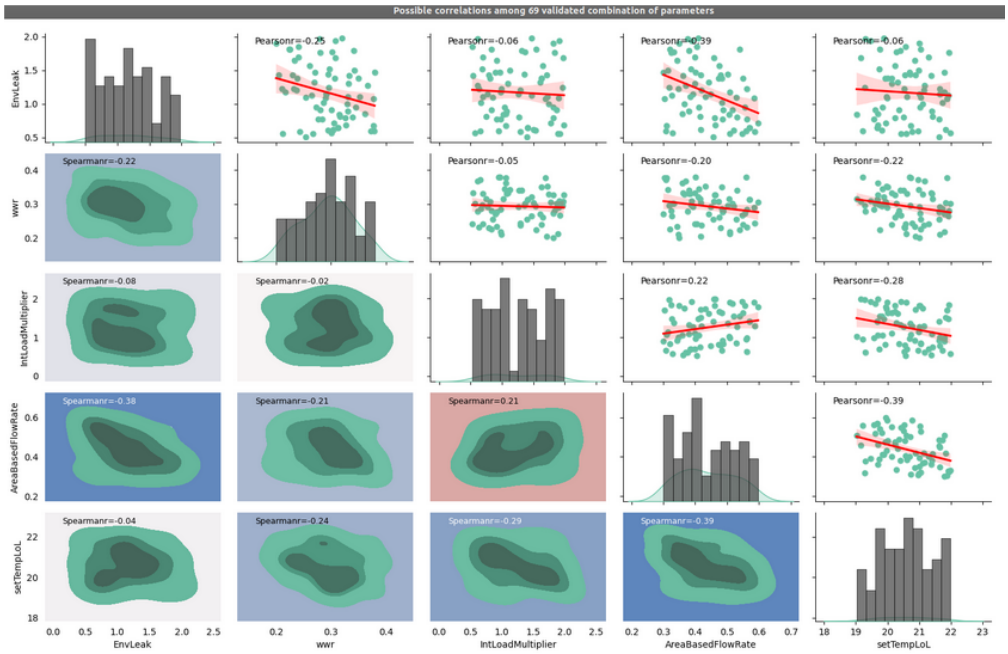


Figure 11: Pairwise illustration of relations between uncertain parameters with a first glance of calibrated parameters on the diagonal. (based on 126 simulations)

number of samples to catch the right relation. Figure 12, with more number of simulations, shows a negative relation between these two parameters.

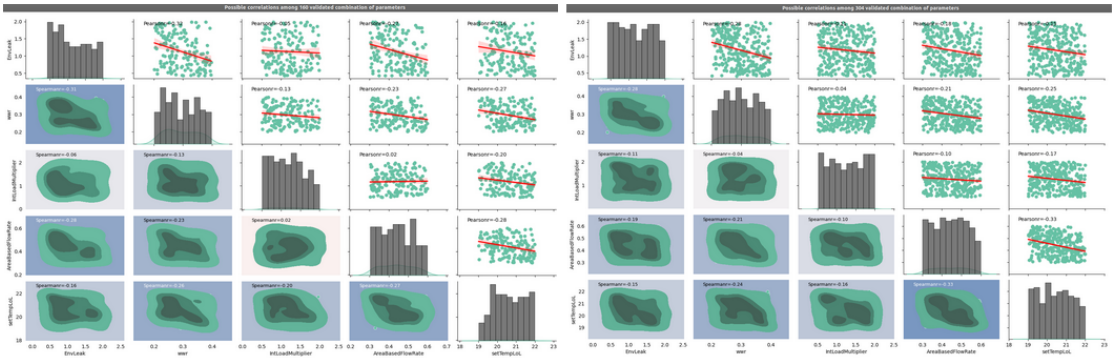


Figure 12: Pairwise illustration of relations between uncertain parameters with 252 simulations (left) and 510 simulations (right)

Now the question is how many simulations would be enough to be sure about the accuracy of a calibration process?
Looking at Figure 13, which shows the histogram and density of calibrations with different numbers of simulations, it seems the calibration process with Cerezo's interpretation is not able to catch the most influential part of uncertain parameters with the number of simulations/searches

under 500. This is because of the differences between density distribution results with 126 and 252 simulations and the similarities between 500 and 1020.

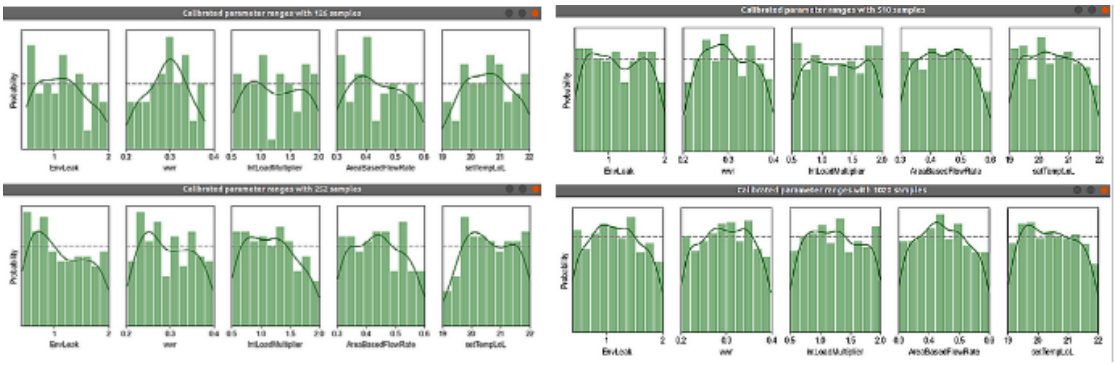


Figure 13: Calibrated parameters for buildings 9 and 10 with different numbers of simulations. First column (126, 252) and second column (510 and 1020). Prior distributions are plotted with dashed lines.

Looking at the cases with 510 and 1020 simulations, it seems when enough combinations of parameters exist, calibrated parameters start to form a more similar density distribution.
But before continuing, a doubt about the randomness of the sampling process raises a question, if any possibility of accidental matches among passed simulations exist? t-SNE technique is used to visualize the results and to identify any outlier.

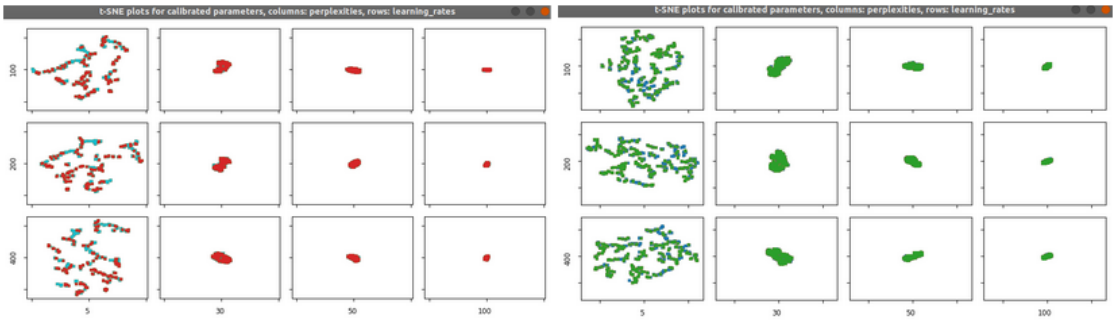


Figure 14: t-SNE technique to reduce dimensionality of 5 calibrated parameters. The left figure shows 304 passed cases for 510 simulations and the right figure shows 593 passed cases for 1020 simulations). As learning rate (X-axis) increases, all passed simulations for both buildings (two colors in each figure) converge into one point which is due to the building similarities and no outlier among passed simulations due to realistic parameter ranges.

Since the two buildings are very similar (can be classified in one archetype) their calibrated parameters gather in one cluster and more importantly with no outlier as the sign of no blind match, there is no accidental match in the result. see Figure 14 for a perplexity such as 200 and a learning rate such as 30. If there were many buildings in different archetypes distinct clusters could form and if the range of unknown parameters were very big and unrealistic, the existence

of any outlier (as matched parameter) would be possible.

Note that cluster sizes for parameters in the t-SNE plot do not have any meaning because t-SNE defines distances as regional variations for each parameter (variable or cluster), therefore relative sizes of calibrated parameters do not project in the cluster in t-SNE plots. In fact, t-SNE tends to expand denser regions of data (clusters) and contracts sparse ones, see [22]. This fact is valid for the distances among clusters as well.

Back to the results, the observation in Figure 13 regarding the minimum number of simulations to sufficiently strengthen the joint probability distribution can be tested by taking 100 samples from the calibrated parameters for new simulations.
The same set of buildings 9 and 10 plus an unseen but similar building named 12 are simulated with 100 samples from posterior distributions. Table 6 shows the percentage of the matched results with the prior and posterior (calibrated) distributions with respect to $\alpha < 5\%$.

Table 6: Percentage of passed simulations with respect to APE value ($\alpha < 5\%$) before and after calibration (posterior samples are taken based on covariance matrix of passed cases in prior distributions).

Building number (vertical)/number of simulations(horizontal)	Simulations with prior distributions				100 simulations with posterior distributions based on:			
	126	252	510	1020	126	252	510	1020
9	29%	32%	31%	32%	75%	79%	40%	78%
10	26%	32%	28%	27%	22%	25%	63%	8%
12	-	-	-	-	32%	36%	64%	15%
Total	27%	32%	30%	29%	43%	47%	56%	34%

The first impression from the results in Table 6 is that more simulations do not necessarily improve the result of calibration. Looking at the results for buildings 9 and 10, simulations with prior (samples from the uniform ranges) distributions keep a similar percentage of matches for both buildings. Note that this result can be expected if buildings are similar (belong to the same archetype) and the unknown parameters and their ranges are selected carefully. On the other hand, 100 simulations with posterior distributions (calibrated parameters) do not match 100 percent even with calibrated parameters based on 500 simulations or more. Therefore the reason for calibration should be seen in reducing the overall APE and reaching an acceptable result with lower number of simulations. Comparing the deviation of simulations with measurements, shown as percentage of error in Figure 10 and Figure 15 for 126 simulations, the overall APE reduces from 35% to 20%.

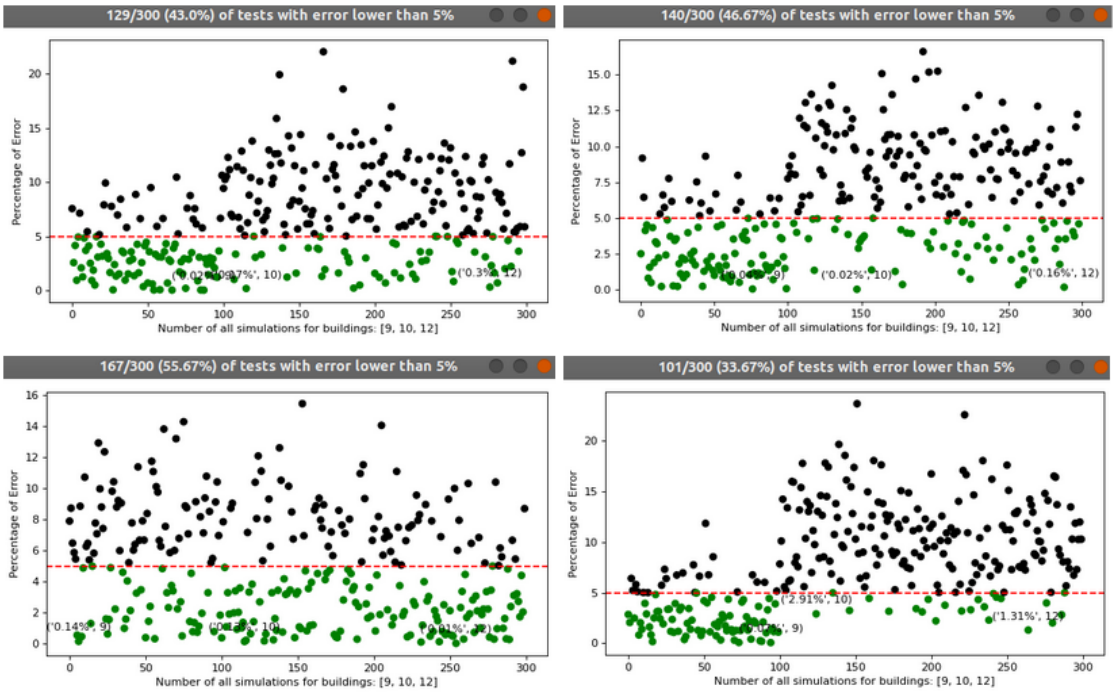


Figure 15: Distribution of matches (100 simulations per buildings 9, 10 and 12) with calibrated parameters based on 126 simulations (up-left), 252 simulations (up-right), 510 simulations (down-left) and 1020 simulations (down-right). The results are shown in posterior section of Table 6.

Looking at Figure 15, it seems that the sampling for buildings 9 and 10 in posterior distributions is in favour of building 9. This is more obvious with the case 1020 where even for the unseen building 12, the number of matches with calibrated parameters is much fewer than the case with 510 simulations. One reason could be the higher similarity between buildings 10 and 12 as their constructions are even more similar than building 9, see Figure 5, and the appearance of this result could be due to neglecting an important parameter in the simulations.

But if we compare the results with prior distributions in Table 6, it shows no biased behaviour and both buildings 9 and 10 have similar number of matched cases, in addition the number of matches for building 10 and 12 based on calibrated parameters with 510 simulations rejects the idea of biased calibrated parameters, see Figure 15.

This inconsistency shows that like any other iterative method which works with random numbers, sampling technique or starting point affects the results. Here, after the calibration the results of multiple runs are different. Either there is still a systematic problem about the iterative process or the number of iterations is not enough. A test is conducted by two samplings from calibrated results with 50 simulations instead of 100 simulations. Building 10 is put to test with these two runs each with 50 samples generated separately from posterior distributions based on 252 initial simulations. Figure 16 shows that the number of matched cases is improved from 25%, see Table 6, to 37% (combination of both runs). As expected, the results are dependent on the sampling technique from the covariance matrix based on the rvs method for the normal distribution in the stats.norm package. As already done for sampling from uniform distributions, an improvement

here is to switch from rvs sampling to LHS for the posterior distributions.

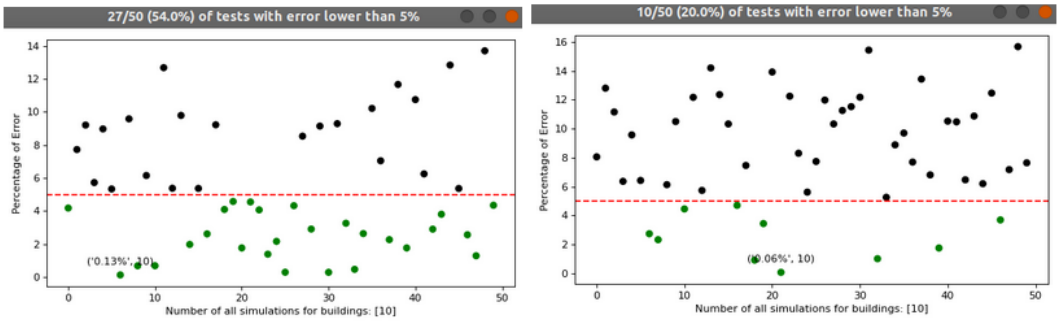


Figure 16: Two runs with 50 simulations of building 10 with posterior distributions, each 50 samples generated separately from covariance matrix based on calibrated parameters with 252 prior distributions.

By changing the sampling technique for posterior side from rvs to LHS, the result of simulation with calibrated parameters based on 1020 simulations shows a significant improvement for buildings 10 and 12, see Figure 17 and compare it with Figure 15.

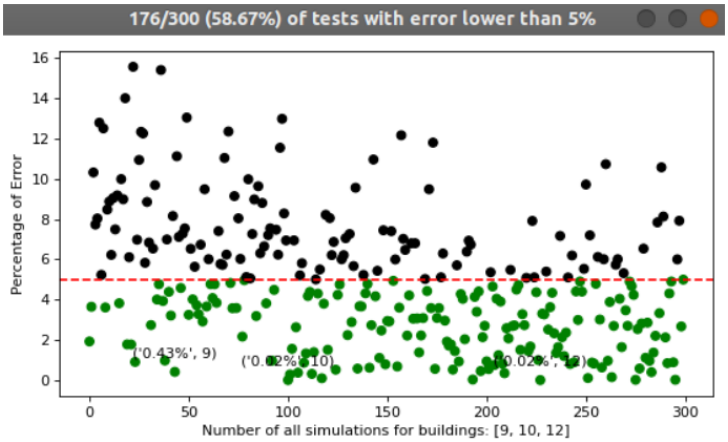


Figure 17: The result of simulations with LHS from posterior distributions based on 1020 prior sampling.

So far based on these results, to answer the first question about the number of required simulations for an accurate calibration without any metamodeling, with respect to available computational resources, between 200 to 500 simulations per building could be enough. If not enough passed simulations yield even with a low number of simulations (such as 126), then parameter ranges should be revised, assuming that all parameters in the model are based on a sensitivity analysis.
But a proposal here is to calibrate an already calibrated range for a finer resolution, i.e. instead of having 510 simulations to reach the posterior distributions, 2, 3 or 4 consecutive calibrations

with 126 simulations are tested. Note that this idea deals with the prior distributions to find the calibrated parameters at first and then uses 126 random samples from the posterior distributions to recalibrate them n times, each time from the previous round. Table 7 shows the percentage of matches and the maximum APE for 4 calibration rounds.

Table 7: Recalibration of posterior distributions with 126 sampling and $\alpha < 5\%$

	Building 9		Building 10		Total match cases
	match	max APE	match	max APE	
Calibration with prior distributions	29%	33%	26%	35%	27%
Recalibrate posterior round 1	48%	13%	55%	16%	51%
Recalibrate posterior round 2	56%	13%	44%	15%	50%
Recalibrate posterior round 3	50%	15%	48%	19%	49%

Looking at Table 7, while the number of matches increases for both buildings the maximum APE in the whole range of simulations decreases as well. Figure 18 shows the kernel distribution of ranges in each round and Figure 19 shows APE for all simulations in the consecutive calibration process.

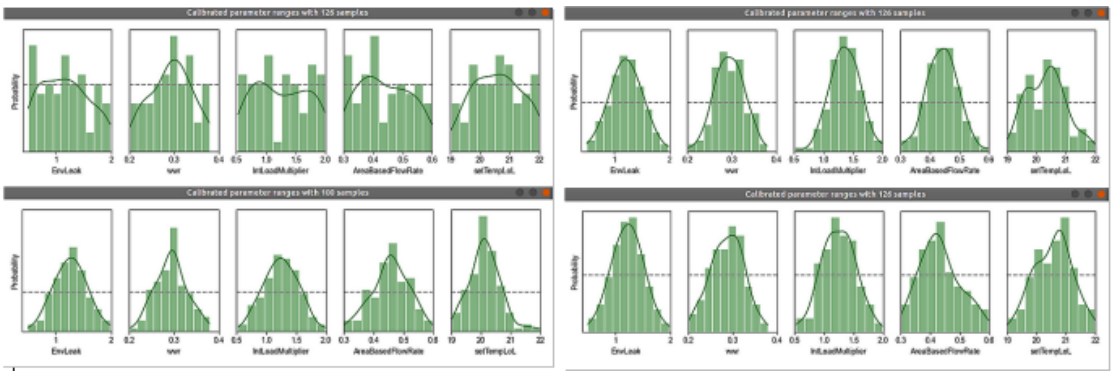


Figure 18: Calibrated distributions based on 126 sampling for buildings 9 and 10, from the prior distributions (up-left) to the last recalibration in round 3 (down-right). Prior distributions are plots with dashed lines.

By looking at Table 7, it seems after one recalibration (round 1) the results are converged and more recalibration rounds do not improve the results any more. As expected in real life, the kernel distributions of influential calibrated parameters should eventually take a normal form considering the fact that they are influential. Since if a parameter distribution after calibration does not depart far from its first prior uniform distribution, then it means a value in any area of the distribution has similar effect on the energy consumption and therefore that parameter is not influential and can be considered deterministically, see [2].

The mathematical reason behind this approach is the Sequential Bayesian updating, see [25]. In presence of a good sensitivity analysis and carefully selected ranges for probabilistic parameters, by a high chance even a small number of iterations can provide enough samples to form

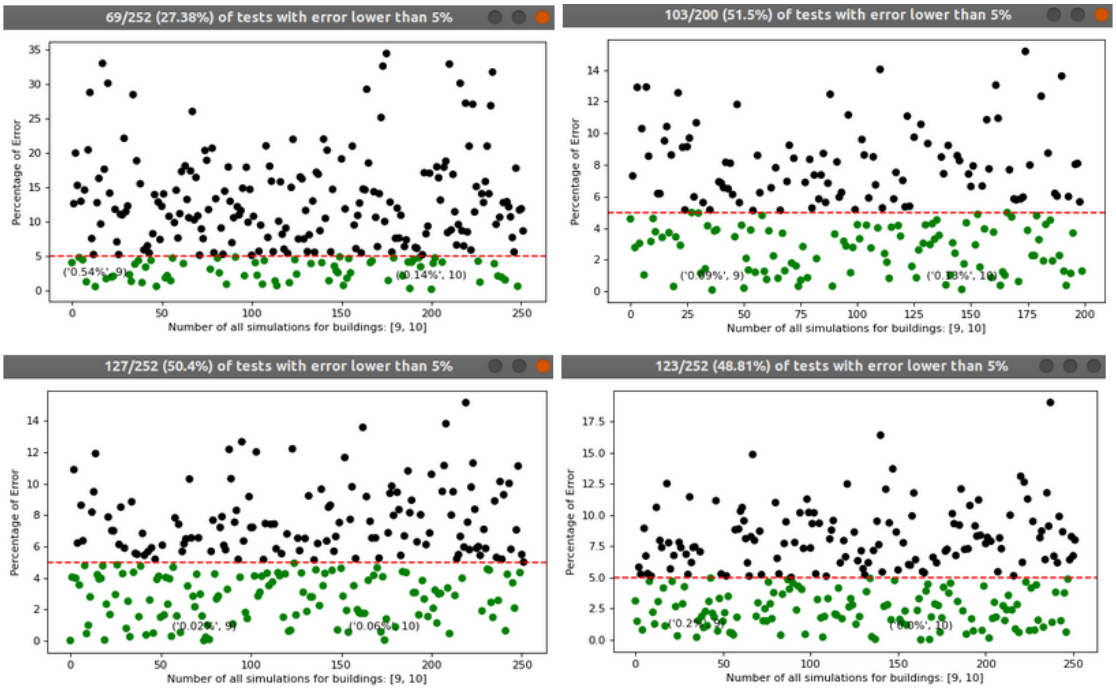


Figure 19: APE based on 126 sampling for buildings 9 and 10, from the prior distributions (up-left) to the last recalibration in round 3 (down-right).

a strong joint probability distribution to represent the most probable areas of unknown parameters. But to remedy the course search grid, another calibration can be used, i.e. the previous posterior distribution can be used as the current prior distribution. In this context, Equation (1) can be reconsidered from:

$$posterior \propto prior \times likelihood$$

to:

$$revised\,posterior \propto current\,posterior \times new\,likelihood \tag{3}$$

This recursive Baysian updating, Equation (3), only needs the current posterior information to find the next one. In other words using only the previous state to find the next one makes it a Markovian model, see [25].

After training the final calibrated parameters, they can be used to simulate unseen buildings for testing. Please note that all the procedures explained above should be applied on a good number of buildings with measured data to reach the most representative values of unknown parameters in an archetype.

5 Discussion

The results of this work are divided into two sections:

- finding reliable sensitivity analysis methods to identify the most influential unknown parameters on energy consumption of buildings,

- and developing a calibration method to find meaningful combinations of those parameters in order to explain the thermal behaviour of buildings.

For the first part, two suitable sensitivity methods (Morris and RDB-FAST) are implemented. From the reviewed studies and the results in this work, Morris, as a screening method, should be used when there are many uncertain or completely unknown parameters involved. The method is computationally cheap and can prioritize parameters while it can, to some extent, predict a level of interaction for parameters as well. But its result in the presence of few, highly correlated and influential parameters could be misleading, especially for an analysis based on a small number of simulations.

On the other hand, RDB-FAST as a variance-based method can measure the level of importance for parameters accurately, no matter how many parameters are involved in the model and how much they are correlated (order of interactions among parameters). In this sense, when having results for both methods (Morris and RDB-FAST), the results from the latter should be prioritized.

For the second part, I have developed a calibration method based on Cerezo's interpretation of Baysian calibration. The main challenges here are to use enough measured data (from buildings) and to do enough simulations for them. This work focuses on the second part and since it is a brute-force method, it is important to find an optimum number of simulations to secure the accuracy of the results while using computational resources wisely. As shown in the results, I have expected that increasing the number of simulations for each building does not necessarily increase the percentage of matched cases. What matters most is that with any selected number of simulations the number of passed cases should be enough to create a strong joint probability distribution and as a result to make a representative covariance matrix[9] for that joint distribution. I have seen that the thermal behaviour of buildings can be explained even with a low number of simulations when parameters and their uniform ranges in the first place are selected correctly. This case opens a door for recalibrating already calibrated parameter ranges.
In other words, I can suggest starting the calibration process with a low number of simulations to see the number of passed cases. If not enough passed combinations of parameters exist then the thermal model, parameters and their ranges, should be revised.

While 500 to 1000 simulations are routine in previous brute-force algorithms for calibration, for a group of similar buildings (classified as an archetype) and based on the result in this work, 100 to 200 simulations should be enough, if it provides a strong joint probability distribution. Then another Bayian inference considerably bolds the right parts of parameter ranges with the highest probability of affecting energy consumption.

I have also noticed that while the LHS method has been the technique to sample from the prior uniform distributions, using the same technique for the posterior side is very beneficial compared with a method like rvs in the scipy package[10]. To use LHS with posterior distributions, first they should be transformed to a normal distribution. The reason is that prior distributions are continuous uniform ranges and simply LHS works but without a transformation a joint probability distribution with discrete data points cannot be used as a sample source for LHS. Intuitively in the real life, due to the law of large numbers, many stochastic distributions take normal form. In calibrated ranges, this normal form is around a central value with a possible

[9] Covariance matrix used to generate random samples from calibrated parameters after the calibration process.
[10] Link to online source: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html

bias to left or right, completely depends on the lower and upper bounds selected by the modeller.

The calibration in this study is based on annual measurements, i.e. the target variable has yearly resolution and relies only on a single value for different types of energy sources as purchased heat, electricity, gas, etc. for each building. While, for the big picture, this methodology is sufficient to estimate parameters for overall retrofitting strategies but it lacks finer information about the thermal behaviour of a building, see [26]. For example, in case of predicting seasonal power demand or monthly energy usage for energy network management, annual calibration cannot be used. The reason is that (even for a building with a low APE between its annual measured data and its simulated model based on calibrated parameters) aggregation of errors based on deviation from the real situation of a building tends to average out. Therefore, this work, in the presence of measured data for validation, must be continued to monthly and even hourly calibration.

6 Conclusion

The goal of this thesis is to understand and to implement the steps of making a realibable representative thermal model for buildings. UBEM is a tool for parametric studies of retrofitting strategies or monitoring energy usage in urban scale.
But thermal models in UBEM need predefined reliable input and there are many unknown parameters which only some of them are important with respect to their effect on energy consumption of buildings and these parameters can vary in different building scenarios.

To deal with the uncertain or unknown parameters, two alternatives exist, defining them deterministically with a single value based on some assumptions, experience, national code or literature or by defining them stochastically as a distribution between two bounds with equal probabilities. The first approach is the source of discrepancy as it oversimplifies a variety of possibilities in real life into one possibility. The second approach improves the thermal models' results but it can even perform much better if we eliminate the low probable parts of parameter ranges by identifying and keeping the higher probable parts, this is called calibration.

In this work, to identify important parameters, sensitivity analysis methods Morris and RDB-FAST are found to be the most suitable with respect to their results and computational cost. Most often, electrical and mechanical system design such as setpoint temperatures and coefficient of performance, internal loads and their operational parameters, infiltration rate and even geometric values such as window to wall ratio and U-values are unknown. Among all influential parameters, the most source of inaccuracy in thermal building simulations is the uncertainty in infiltration rates, equipment loads and occupants behavior.

For the calibration, the work developed in this thesis is based on brute-force sampling of unknown parameters and validated with annual energy data. But it is computationally expensive and needs hundreds of thousands of simulations. Therefore the focus is to find a way to calibrate parameters with much less simulations but with an acceptable accuracy of the results. This work shows that in the presence of careful selection of important parameters and distribution for them, a lower number of simulations can be used for annual calibration as well.

References

[1] Achieving net-zero emissions by 2050, flagship report — october 2020.

[2] Carlos Cerezo, Julia Sokol, Saud AlKhaled, Christoph Reinhart, Adil Al-Mumin, and Ali Hajiah. Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): A residential case study in Kuwait City. *Journal of Energy and Buildings*, 154:321–334, November 2017.

[3] Martina Ferrando and Francesco Causone. An overview of urban building energy modelling (UBEM) tools. *Proceedings of the 16th IBPSA Conference, Rome, Italy*, pages 2–4, September 2019.

[4] Oleksii Pasichnyi, Jörgen Wallin, and Olga Kordas. Data-driven building archetypes for urban building energy modelling. *Journal of Energy*, 181:360–377, August 2019.

[5] Fatemeh Yousefi, Yaghob Gholipour, and Wei Yan. A study of the impact of occupant behaviors on energy performance of building envelopes using occupants' data. *Journal of Energy and Buildings*, 148:182–198, August 2017.

[6] Shreshth Nagpal, Caitlin Mueller, Arfa Aijazi, and Christoph F. Reinhart. A methodology for auto-calibrating urban building energy models using surrogate modeling techniques. *Building Performance Simulation*, 12:1–16, April 2018.

[7] Torben Østergård, Rasmus Lund Jensen, and Steffen Enersen Maagaard. A comparison of six metamodeling techniques applied to building performance simulations. *Journal of Applied Energy*, 211:89–103, February 2018.

[8] Wei Tian. A review of sensitivity analysis methods in building energy analysis. *Renewable and Sustainable Energy Reviews*, 20:411–419, April 2013.

[9] Edmund yan, Oliver Wild, Apostolos Voulgarakis, and Lindsay Lee. Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output. *Free PMC article*, 11:3131–3146, August 2018.

[10] Navid Delgarm, Behrang Sajadi, Khadijeh Azarbad, and Saeed Delgarm. Sensitivity analysis of building energy performance: A simulation-based approach using OFAT and variance-based sensitivity analysis methods. *Journal of Building Engineering*, 15:181–193, January 2018.

[11] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice : a guide to assessing scientific models.* Chichester; Hoboken, NJ : Wiley, Hoboken, NJ, USA, 2004.

[12] Martin Heine Kristensen and Steffen Petersen. Choosing the appropriate sensitivity analysis method for building energy model-based investigations. *Journal of Energy and Buildings*, 130:166–176, October 2016.

[13] Kathrin Menberg, Yeonsook Heo, and Ruchi Choudhary. Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. *Journal of Energy and Buildings*, 133:433–445, December 2016.

[14] APSIM help content. Psim: The agricultural production systems simulator.

[15] Jon Herman, Will Usher, and et al. Salib - sensitivity analysis library in python.

[16] Anh-Tuan Nguyen and Sigrid Reiter. A performance comparison of sensitivity analysis methods for building energy models. *Building Simulation*, 8:651–664, 2015.

[17] Chonggang Xu and George Gertner. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). *Journal of Computational Statistics Data Analysis*, 55:184–198, January 2011.

[18] Stefano Tarantola, Debora Gatelli, and Thierry A. Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering System Safety*, 91:717–727, June 2006.

[19] Ben Lambert. *A Student's Guide to Bayesian Statistics*. Publisher SAGE Publications Ltd, UK, 2018.

[20] Adrian Chong and Kathrin Menberg. Guidelines for the Bayesian calibration of building energy models. *Energy Buildings*, 147:527–547, September 2018.

[21] Scipy documentation. Correlated random samples.

[22] Scikit learn. t-distributed stochastic neighbor embedding, sklearn 0.24.2.

[23] Stephanie Deviant. *The practically cheating statistics handbook*. USA, 2010.

[24] Palisade Knowledge Base. Latin hypercube versus monte carlo sampling.

[25] Steffen Lauritzen. Sequential bayesian updating, statistical inference, lectures 15 and 16, university of oxford, 2008.

[26] Julia Sokol, Carlos Cerezo Davila, and Christoph F Reinhart. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Journal of Energy and Buildings*, 134:11–24, January 2017.

[27] Faye C. McQuiston, Jerald D. Parker, and Jeffrey D. Spitler. *Heating, Ventilating, and Air Conditioning: Analysis and Design*. Hamilton Printing, United States of America, 2005.

[28] Generating quasi-random numbers, mathworks help center.

[29] StackExchange forum discussions:. What's the difference between probability and statistics?

[30] Jason Brownlee. *Statistical Methods for Machine Learning, Discover how to Transform Data into Knowledge with Python*. United States of America, 2018.

[31] Martin Wattenberg, Fernanda Viegas, and Ian Johanson. How to use t-sne effectively.

7 Appendix

7.1 R^2, t-value and F-value

R^2 is a coefficient to show how linearly a variable-change can be explained by another variable-change, i.e it provides a measure to find a possible existing correlation between two variables. While this coefficient can determine the strength of relation between two variables (correlation coefficient) it does not mean causation. R^2, as the coefficient of determination, is the square of the correlation coefficient, which among different correlation coefficients one can mention the most famous one as Peasrson's R.

The t-value and F-value are the results of t-test and F-test. The t-test is used when testing the result of restriction on one parameter in BEM is on focus, i.e. to test the hypothesis, if the parameter is sensitive in building energy performance. In contract, the F-test is used for restrictions on multiple parameters, usually to compare variances of two variables.
In general, the t-test is used to compare the mean of a sample-data to population mean. In this context, the larger the t-value the more different the sample-data is from the average of population. This is especially useful when the sample size is small or if the population standard deviation is unknown, see [23].
In UBEM, the higher the absolute value of t, F or change of R^2 is the more important the corresponding variable.

7.2 Building physics

Building Physics is the study of heat and mass transportation (air and moisture) through building materials and within buildings themselves. A building should provide a comfortable indoor environment with the least possible energy consumption (high energy efficiency) to decrease the environmental impacts on nature. Calculation of building energy consumption needs information about building location and climate, geometry and construction elements, ventilation and thermal systems and controls, infiltration rate and finally internal loads and operational profiles, see [6].

7.3 Building energy calculation methods

Occupied or unoccupied buildings may need to meet some specific thermal conditions, ventilations, lightings and noise reductions. To fulfill these requirements energy is used. It is desirable to estimate the quantity of energy necessary to heat, cool, ventilate and light up buildings. During time, some methods have been developed to calculate the energy consumption, especially for residential buildings.

In total, one can divide the methods to:

- roughly estimation, manual methods such as degree-day and Bin methods,
- and elaborated, comprehensive simulation methods such as heat balance, weighting factor and thermal network methods.

The degree-day method, heating or cooling degree-day with different modifications, is based on average conditions and does not take into account day to day weather conditions, solar effects and equipment's temperature in a building. In its original form, it assumes that when the average daily outdoor temperature equals a base value (originally $18^{\cdot C}$) the solar heat gain and internal loads compensate for heat loss through the building structure. This type of calculation,

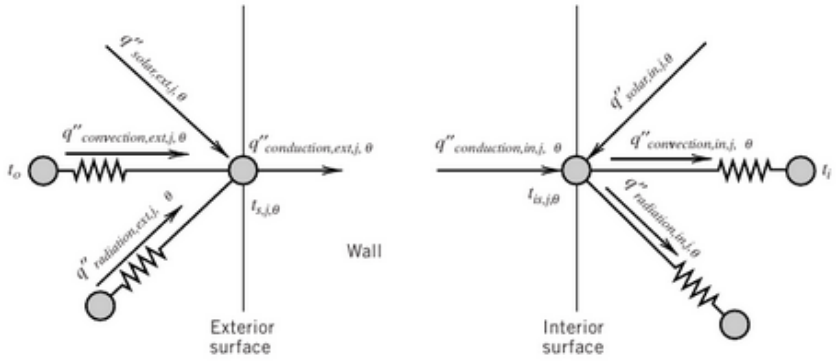
nowadays, can only be used for rough estimation of energy consumption of a building in different geographical locations, see [27].

The Bin-method achieves reduction of the error in degree-day method by sorting outdoor long-term measurement temperature for weather conditions to 'n' number of discrete bins. The procedure groups temperature-hours with respect to their occurrence. Later it considers a simplified occupant hourly schedule in one week to estimate the internal loads. In order to apply the effect of internal loads, the number of hours in each bin are then shifted to two groups of occupied and unoccupied hours by using the simplified schedule. The desired conditions inside a building for two states of presence and absence of internal loads can be calculated as a function of outdoor temperature, namely known as load profiles. Finally, the result yields by calculation of load for each bin with respect to load profiles multiplied by corresponding hours for that bin, see [27].

The simulation methods are computer-based which define the thermal behaviour of a building, its systems and a central plant as mathematical models. They use weather conditions, building body descriptions and internal loads as input. The calculation of thermal zones (sensible loads and air temperature in each zone) are the input for systems to find the amount of hot or cold water or air and the results of the systems are inputs for calculation of the central plant. This sequential approach is the base of most building software today. First the loads in each zone are calculated for every hour, followed by the amount of medium to cover the internal set point temperatures. At last the central plant summarizes power and energy needs in systems for the whole simulation period.

Three different approaches to model a building are:

- the heat balance method seeks an equilibrium among all energy flows at each point (thermal zone). A set of energy equations for each zone air, interior and exterior surfaces (wall, roof, floor, windows) are solved but with respect to algorithms for weather conditions and solar radiations. Figure below reproduced from [27] illustrates the heat balance method on an exterior wall, named j, at time θ:



  where on each surface heat balance should be satisfied (W/m^2) but the conduction heat flux on exterior and interior surfaces can be different if steady-state heat transfer is not achieved (transient heat flux through the wall). EnergyPlus program lies in this category,

- the weighting factor method also known as the transfer function method is developed to be a faster method than heat balance. It uses some coefficients called weighting factors to approximate the response of a thermal zone to a unit heat pulse,

- and the thermal network method can be considered as an extended-refined version of the heat balance method. It can decretize a building into a network of nodes with interconnecting energy flows in form of conduction, convection, radiation and/or air flow. While the heat balance method usually defines one node to represent a thermal zone, an exterior or an interior surface, a thermal network can define a set of nodes like mesh points inside one thermal zone or on different surfaces. This method can provide more accurate results with more computational cost. ESP-r program lies in this category, see [27].

7.4 Random number generators

As a single number, a random number is a positive number drawn from a set of equally uniform probable values and as a sequence of random numbers they should be also independent of each other. There are two main approaches to generate random numbers in a computer:

- Pseudo-Random Number Generators (PRNGs), which are basically algorithms to produce a set of numbers or precalculated lists that look like they are random numbers. These generators are usually very efficient and fast and if the starting point of a sequence of random numbers is known, it is possible to regenerate the exact set again in a later time (deterministic). These methods are periodic and the whole sequence of numbers will finally start over but still are very suitable for modeling and simulation software,

- true Random Number Generators (TRNGs), which are usually connected to physical phenomena such as atmospheric noise, a radioactive source decay, etc. A given set of random numbers with this approach are not reproducible (nondeterministic) and it usually takes much longer time than pseudo approaches to produce true random numbers. These methods are non periodic.

Classified under (PRNGs), Quasi-Random Number Generators (QRNGs) produce uniform samples with low discrepancy among distributed points. The low discrepancy is the measure of uniformity which makes these generators fail many statistical tests for randomness. It is important to know that their goal is not to generate true randomness but to evenly distribute samples such that no large gaps or clusters form. In fact, the goal is to spread the points with maximum possible coverage of the sample region. Among the generators in this class, one can name: Halton and Sobol. On the other hand, Latin HyperCube is not quasi in terms of minimizing the discrepancy among samples but it produces a sparse uniform distribution over the domain, see [28]. In this study, Latin HyperCube sampling is the main random number generator.

7.5 Probability and likelihood

Probability quantifies expectation of an outcome, likelihood quantifies trust in a model. In probability, we consider some underlying process which has some randomness or uncertainty modeled by random variables (as a distribution) and we want to find the chance of any sample's occurrence. In likelihood we observe something that has happened, and try to figure out what underlying process would explain that observation, see [29]. So, the problems connected to these two concepts are inverse to each other.

7.6 Probability distributions

A probability distribution describes all the possible values that a random variable can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution

depends on a number of factors. These factors include the distribution's mean (average), standard deviation, skewness, and kurtosis. (From Wikipedia)

7.7 Statistical definitions

A quick reminder to common terms and definitions in statistics:

- The median is the middle of an ordered set of numbers, for set 6, 11, 7 the median is 11,
- The mean (μ) is simply the average of a set of numbers, example: $(6+11+7)/3 = 24/3 = 8$,
- Variance is the average of the squared differences from the mean and it is the measure of how spread a data set is, example: $\sigma^2 = [(6-8)^2 + (11-8)^2 + (7-8)^2]/3 = [4+9+1]/3 = 14/3 = 4.666666667$,
- Standard deviation is the square root of variance, example: $\sigma = \sqrt{(4.666666667)} = 2.160246899$ which is the measure of how spread out numbers are, 2.16 units around 8,
- Confidence intervals are the lower and upper bounds of a distribution. If you do your experiment many times, you can expect to see the same results in the same bound and as if you examine the whole population. Confidence level is the percentage area under the distribution curve as the sampling domain between the intervals, see [23],
- Quantile is a place where a sample set is divided into equal sized subgroups or in a continuous form as probability distribution into areas with equal probability. Median is a quantile which divides a sample set into two subgroups, see [23],
- Probability density function is a mathematical expression which can be used to calculate the probability of observations in a data set, see [30],
- The law of large numbers says as the sample size increases, the mean value for the sample set tends to the average of the population.

7.8 Check the robustness of the results with t-SNE

In random sampling within probability ranges of parameters for energy simulations, there is a risk to have some combinations which, although accidentally yield results close to measured data but they may not be very feasible. To identify parameters in such combinations, t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to reduce the dimensionality of the parameters for the results.

t-SNE is a nonlinear dimensionality reduction technique which is used for visualization of data, such as mapping data points for hundreds of variables into 2 or 3 dimensional spaces but by keeping the relationship among data points. This technique is a stochastic method and its output can be slightly different for multiple runs yet it is helpful to identify outliers for each type of parameter in calibrated ranges. In other words, for one or a group of buildings in a specific archetype, it is expected to see the calibrated parameters in a distinct cluster. If a calibrated datapoint is being separated from its cluster type, it can be more considered as an accidental sample rather than a true value to represent a building parameter in the real world.

t-SNE technique needs tuning, i.e. it has hyperparameters such as:

- perplexity which is about guessing how many close neighbors each point might have, see [31]. Scikit-learn documentation suggests a value between 5 to 50 and the default is 30. This value should be selected according to the size of joint distribution of calibration data in UBEM and to get meaningful results, the perplexity should be smaller than the number of points, see [22],
- learning rate which plays an important rule in the final shape of clusters of data after dimensionality reduction. Scikit-learn suggests a value between 10 to 1000 and the default is 200 while a too low value produces dense clouds with outliers and a too high value produces equidistant points in a ball shape, see [22],
- method-type which by default uses Barnes-Hut algorithm based on distances between each point and other points accumulated in groups. This method has running time as $O(NlogN)$ with some acceptable errors. In contrast, the 'exact' method considers the distances between one point and all other points but has running time as $O(N^2)$,
- and among others, the number of iterations which is used for optimization with a default value of 1000.