Beyond Normal: Guidelines on How to Identify Suitable Model Input Distributions for Building Performance Analysis

Giorgos Petrou[1], Anna Mavrogianni[1], Phil Symonds[1], Mike Davies[1]
[1]UCL Institute of Environmental Design and Engineering, London, United Kingdom

Abstract

Building modelling is a valuable tool in the widespread efforts to decarbonise the built environment. To ensure modelling robustness, uncertainty and sensitivity analysis techniques are often used. Such techniques commonly require model input distributions to be defined. This paper describes a novel approach, within the built environment, for identifying empirically-derived probability distributions of model inputs. Following data cleaning, candidate distributions selected based on measures of skewness and kurtosis are fitted using maximum likelihood estimation. The distribution that best describes the dataset is identified using Akaike Information Criterion and its derivatives along with goodness-of-fit plots. The method was demonstrated for a dataset of wall U-value measurements in English homes.

Key innovations

- This paper presents a novel approach of identifying probability distributions of building modelling inputs that adequately describe empirical data.
- Code for implementation has been made publicly available (`https://github.com/giorgospetrou/distrmultifit`).

Practical implications

The use of inappropriate distributions in building modelling can have a substantial impact on the output. By implementing the approach detailed in this paper, a modeller can more appropriately incorporate relevant information in their modelling.

Introduction

With a share of 39 % of global greenhouse gas (GHG) emissions in 2018, building construction and operation play an important part in the efforts to reduce GHG emissions and limit global warming to 1.5 °C above pre-industrial levels (IEA, 2019; IPCC, 2018). Given the complex nature of the built environment, striking the right balance between reducing emissions while ensuring a healthy and comfortable indoor environment in a cost-effective manner is a challenging task. Building simulations can be an important part in this transition. They enable modellers to assess the different routes to decarbonisation and their unintended consequences by easily running multiple simulations and testing the efficacy of different technologies or design approaches at the individual building or urban level.

An important part of the modelling process is sensitivity and uncertainty analysis, a topic that has been widely researched and whose value is well-recognised within the field of building modelling (Tian et al., 2018). The lack of detailed uncertainty propagation can lead to overly confident and unreliable models. Sensitivity analysis can quickly identify model inputs that can greatly influence outputs. Monte Carlo methods, commonly used for forward uncertainty propagation and Bayesian methods often used for inverse uncertainty quantification require the modeller to choose appropriate distributions for model inputs (Tian et al., 2018). These distributions may represent the uncertainty around a parameter for a single building or the spread in values within a group of buildings. The normal distribution is commonly assumed in building performance simulation due its convenience and familiarity. Similarly, the uniform distribution is often used to express lack of knowledge about the possible value or distributional form of a model input. However, with data availability on the rise, distributions used for uncertainty quantification could in some cases be based on empirical evidence. If the modeller identifies the distribution that best describes the observed data relating to a model input, they can capture its expected value and shape more accurately. Alternatively, the use of inappropriate distributions could contribute to the 'performance gap' – the difference between the predicted and measured (or real) energy performance of a building (or group of buildings) (de Wilde, 2014). Examples of using non-normal and non-uniform distributions exist within the field of building modelling (Booth et al., 2012), however, no clear guidance on how to identify the most suitable distributions for a given dataset could be found.

The aim of this work is to present a step-by-step guide on how to identify the probability distribution function that best describes a given dataset of building-related parameters, potentially improving building performance modelling workflows and the accuracy of their predictions. This process is demonstrated for a set of wall U-value measurements that are available from Hulme and Doran (2014). The R code used for this analysis and a set of relevant functions have been made publicly available[1] to enable modellers to reproduce this work and apply this method for their own analysis.

Methods

A method to identify the appropriate probability distribution functions will be demonstrated on a dataset that can be found in Appendix F of Hulme and Doran (2014). Firstly, a brief description of the dataset will be provided. The proposed method will then be outlined and its underpinning theory will be discussed. This covers the four main steps of the process:

1. Data visualization and cleaning
2. Identification of candidate distributions
3. Fit data to candidate distributions using Maximum likelihood
4. Identify the candidate distribution with the best goodness-of-fit

The dataset

During 2012-2013, fieldwork commissioned by the UK Department of Energy and Climate Change (which in 2016 became part of the Department for Business, Energy and Industrial Strategy (BEIS)), aimed to provide an assessment of the thermal performance of walls in English dwellings and compare it to the theoretical values (Hulme and Doran, 2014). The Building Research Establishment (BRE) led the data collection. For approximately 300 dwellings, in-situ measurements were taken using heat flux plates (Hukseflux HFP01) and surface temperature measurements for a period of two weeks. The homes were a sub-sample of the 2010/11 English Housing Survey (EHS); a national survey that takes place every two years and consists of household interviews and physical surveys (BRE, 2014a). Two measurements were taken for each dwelling as far away as possible from any thermal bridges. For this work, a 6 % adjustment was applied to the raw data, since following the publication of the original report it was discovered that the heat flux plates read 4-8 % lower than intended (BRE, 2016). Table 1 provides a summary of the data, where the U-value is the arithmetic mean between the two measurements taken at each dwelling and following the 6 % correction. Although interesting, the discrepancy between the theoretical U-value expected for each wall type by the Standard Assessment Procedure (SAP) (BRE, 2014b) and the measured U-values will not be the focus of this paper – Li et al. (2015) have addressed this discrepancy specifically for UK solid walls in great detail. Instead, the paper will focus only on the method of identifying the distribution that best describes the dataset. For brevity, the method will be applied only on the U-value measurements of filled cavity walls.

Step-by-step process

This section describes the four main steps undertaken:

1. Data visualisation and cleaning: A histogram is used to visualise the data. This allows for data cleaning to be performed through observation of histogram extremes and removal of outliers. This is preferred over automated procedures based on the data's interquartile range or standard deviation when the data does not appear to be normally distributed. It is easier to identify and reject outliers when there is already an established model of the measured variable and its distributional form is known. However, this is often not the case and automatic methods of outlier detection, such as the Chauvenet's Criterion that assumes a Gaussian distribution would be inappropriate (Hughes and Hase, 2014). Given that many of the model parameters within the built environment field have a physical meaning, it might be better to compare measured extreme values with their theoretical equivalents and our understanding of the physical system being studied.

2. Identify candidate distributions: Once outliers are removed, the data's empirical distribution, together with the "Cullen and Frey" graph of kurtosis against the square of skewness are used to identify candidate distributions. *Skewness*, is a measure of symmetry, with a value of zero indicating a fully symmetric distribution (Reimann et al., 2008). *Kurtosis*, indicates how heavy the tails of a distribution are (i.e is how flat or peaked the distribution is) with a value of three for a normal distribution (Reimann et al., 2008). By plotting the kurtosis and square skewness of the collected data on a graph and overlaying the values that common distributions would take, one can infer the candidate distributions that may best describe the data. Since skewness and kurtosis may easily be affected by extreme values, one can employ a bootstrap technique of random sampling (at least 1000 samples) with replacement to plot multiple possible values on the Cullen and Frey graph (Hesterberg, 2011; Delignette-Muller and Dutang, 2015). Note that it is possible that none of the distributions that appear on the Cullen and Frey Graph may be appropriate and other distributions might need to be explored.

3. Fit candidate distributions: The candidate distributions are then fitted to the data using Maximum Likelihood Estimation – this is achieved with

[1] The complete code is available here: `https://github.com/giorgospetrou/distrmultifit`

Table 1: Summary statistics, mean and percentiles, of wall U-value measurements.

	Wall Type	**Sample Size**	**Wall U-value [W/(m²K)]**					
			2.5 %	25 %	Mean	50 %	75 %	97.5 %
1	Filled cavity	109	0.3	0.6	0.7	0.7	0.8	1.2
2	Other solid	33	0.6	1.1	1.4	1.4	1.6	2.1
3	Standard solid	85	1.0	1.4	1.7	1.7	1.9	2.2
4	Unfilled cavity	50	0.8	1.3	1.5	1.5	1.7	2.0

the R package **fitdistrplus** (Delignette-Muller and Dutang, 2015; R Core Team, 2018). To fit the candidate distributions to the data, several methods exist. The cautious recommendation of this paper is the use of Maximum Likelihood Estimation (MLE) which is a commonly used method of distribution fitting and is the default option in the library **fitdistrplus** (Delignette-Muller and Dutang, 2015). Other methods may be preferred under specific circumstances, such as when placing more weight on one of the data distribution's tail is desirable. One such case might be when the interest is in the least energy efficient dwellings whose building characteristics (e.g. permeability) are described by the tails of the distributions. A probability density function, specified as $f(x_1|\theta)$, quantifies the probability of observing data point x_1, given the distribution parameters θ (i.e. assuming that θ are known) (Portet, 2020). Trying to fit a distribution to a set of known data points is the inverse problem where x_i are known and θ are unknown. Assuming $x_i = x_1, \ldots, x_n$ independent and identically distributed (i.i.d.) observations, the likelihood function is a function of parameters θ defined as (Smith, 2013):

$$L(\theta|x_i) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1}$$

The likelihood function quantifies the probability of obtaining the observed data x_i, if the parameters θ had a specific value (Portet, 2020). By employing an optimisation algorithm, for any candidate distribution ($f(\cdot|\theta)$) and observed data (x_i), parameters θ are optimised in order to maximise the log of the likelihood function (Delignette-Muller and Dutang, 2015). This process is repeated for all candidate functions separately so as to identify the parameters and density function that best describes the data.

4. Identify the candidate distribution with the best goodness-of-fit: Finally, drawing from Information Theory, the Akaike Information Criterion (AIC) and its derivates are used to identify the best fitting distribution (Burnham and Anderson, 2004). Density plots, Q-Q plots, P-P plots and Cumulative Distribution Function plots provide a supplementary measure of goodness-of-fit and inform the modeller whether the best-fitting distribution is satisfactory.

To identify the best fit amongst the candidate distributions, AIC and its derivatives may be used. AIC is defined as (Burnham and Anderson, 2004):

$$\text{AIC} = -2\log(L(\hat{\theta}|x_i)) + 2K, \tag{2}$$

where $\hat{\theta}$ is the maximum likelihood estimate of parameters θ, $\log(L(\cdot|\cdot))$ is the log likelihood and K is the number of distribution parameters (as an example, the normal distribution has two parameters: the mean and the standard deviation). For a collection of R candidate distributions (or models more generally), the best distribution given the data x_i is the one with the minimum AIC value (Portet, 2020). For a small number of observations, where $K > (N/40)$, the corrected AIC may be used instead (Portet, 2020):

$$\text{AICc} = \text{AIC} + \frac{2K(K+1)}{N-K-1}, \tag{3}$$

with AICc approaching AIC as N approaches infinity. While equations 2 - 3 enable the ranking of candidate distributions, the actual values of AIC or AICc are not themselves easily interpretable. However, some more interpretation is possible through the manipulation of the estimated AIC values. Rescaling AIC (or AICc) of each candidate distribution j, with regards to the minimum AIC (AIC_{min}) results in an estimate (Δ_j) of the information loss when distribution j is selected instead of the best candidate distribution; effectively quantifying the strength of the AIC differences (Burnham and Anderson, 2004):

$$\Delta_j = \text{AIC}_j - \text{AIC}_{min}. \tag{4}$$

Burnham and Anderson (2004) suggested that:

- Models with $\Delta_j < 2$ have substantial support (evidence)
- Models with $4 < \Delta_j < 7$ have considerably less support
- Models with $\Delta_j > 10$ have almost no support

Therefore, an alternative to the best candidate distribution (the one with the lowest AIC) with a Δ_j less than 2 may be considered a good alternative while one with Δ_j greater than 10 should not. Portet (2020) warns that these guidelines should be treated with caution if, for example, a large number of candidate distribution are assessed. Instead, one can go further and estimate the Akaike weights (or "weight of evidence") (Burnham and Anderson, 2004; Portet, 2020):

$$w_j = \frac{\exp(-\Delta_j/2)}{\sum_{r=1}^{R}\exp(-\Delta_r/2)}, \tag{5}$$

where $\exp(-\Delta_j/2)$ is the distribution likelihood. The quantity w_j is the probability that distribution j is best amongst the candidate distributions given the observations x_i. Finally, a direct comparison between two candidate distributions may be done by computing their evidence ratio w_j/w_k, quantifying the strength of evidence of model j over model k.

While the AIC and its derivatives can help a modeller determine which model is best between the candidate models, determining whether the best model is good enough for its intended purpose has yet to be answered. This may be decided by visualising the theoretical data, originating from the best distribution, against the empirical data in four plots (see Figure 3 for an example):

1. Histogram with theoretical densities
2. Quantile-Quantile plot (Q-Q plot)
3. Empirical and theoretical Cumulative Distribution Function (CDF) plots
4. Percentile-Percentile plot (P-P plot)

A histogram of the data superimposed by the theoretical densities provides a quick and comprehensive check of the distribution fit. In a Q-Q plot, the theoretical quantiles from the assumed distribution are plotted against the empirical quantiles and a straight line would provide support for the assumed distribution. A P-P plot, will instead have the probabilities of the hypothetical distribution plotted against the probabilities of the empirical data at fixed quantiles. While a Q-Q plot is useful at exposing discrepancies in the tails of the distributions, a P-P plot focuses more on the main body (Reimann et al., 2008). The empirical CDF, is a step function, where as the number of data points increase, it should approximate the underlying distribution function (Reimann et al., 2008).

Results

A histogram with a density plot of the measured U-values for filled-cavity walls is shown in Figure 1. By simply inspecting the histogram, the data distribution seems to be positively skewed as the right tail is slightly longer than the left one. The lowest measured value is 0.2 W/(m^{2}K) while the largest value is 1.5 W/(m^{2}K). While a value of 0.2 W/(m^{2}K) is well within the expected theoretical values of well-insulated cavity walls (BRE, 2014b), a value of 1.5 W/(m^{2}K) is rather high (BRE, 2014b, 2016). This high value could be the result of surveyors incorrectly classifying the wall as filled-cavity or placing the heat flux over a thermal bridge. However, lower than nominal levels of insulation and poor workmanship could also lead to a worse than intended thermal performance. With no strong evidence to reject this data point, it was kept in the analysis.

To determine which distribution should be fitted, the skewness and kurtosis of the sample was estimated

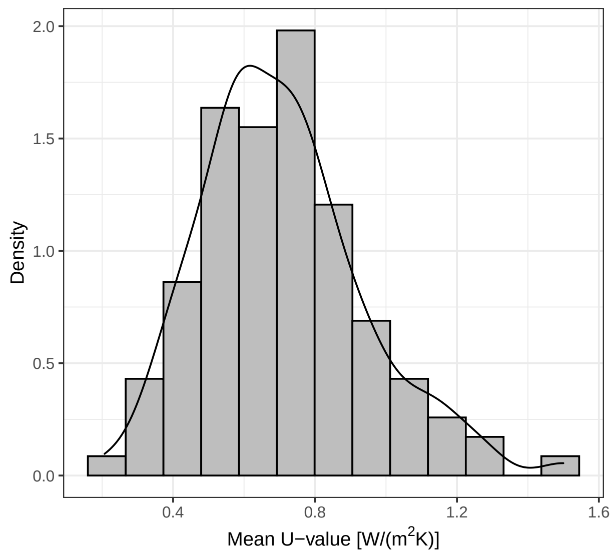

Figure 1: Histogram with a density line of the wall U-value measurements taken from filled cavity walls.

and plotted in a Cullen and Frey graph shown in Figure 2. The sample's values, indicated by the blue dot, suggest that the gamma, Weibull and lognormal distributions could all likely describe the collected data well (this will be determined by the goodness-of-fit analysis in step 4). To account for the uncertainty in the sample skewness and kurtosis, a non-parametric bootstrap analysis was run 1000 times, with the results shown as yellow rings in Figure 2. Many of the

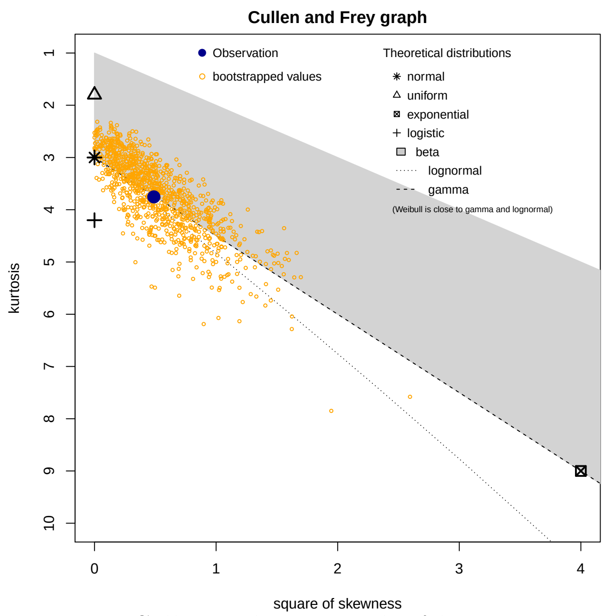

Figure 2: Cullen and Frey graph of kurtosis against square of skewness.

points lie within the shaded area that represents possible kurtosis and squared skewness values the beta distribution can take. However, as the beta distribution is bound within the [0, 1] interval it was not chosen as a candidate distribution. A few bootstrap points concentrated close to the kurtosis of three and squared skewness of zero, encouraging the inclusion of the normal in the candidate distributions.

Following the selection of the candidate distributions (normal, lognormal, Weibull and gamma), each one was fitted and their computed AIC was used to rank them in Table 2. The gamma distribution has the lowest AIC value (-16.07), indicating that it can best represent the data amongst the candidate distributions. It is followed by the lognormal with an AIC of -13.72, the normal (AIC = -8.04) and Weibull (AIC = -7.59). To enable some further interpretation of the results, equations 4-5 were used to determine the strength of the AIC differences (Δ_j), and the Akaike weights (w_j). Based on the suggestions by Burnham and Anderson (2004), with a $\Delta_j = 2.34$ there is some support for the lognormal as an alternative to the gamma, with considerably less support for the normal and Weibull distributions. This is further supported by the Akaike weights, with a 0.75 probability that the gamma distribution is the best distribution among the candidates given the observed wall U-values. A significantly lower probability of 0.23 is assigned to the lognormal while an almost negligible probability of 0.01 was assigned to the normal and Weibull distributions.

The collection of AIC metrics suggest that the gamma distribution best represents the data amongst the candidate distributions. To determine whether the gamma provides a good enough description of the dataset, a closer look of the proposed fit is required and this is provided by the plots in Figure 3. As a comparison, the best fit using a normal distribution was also included. The empirical and theoretical densities and CDFs for the gamma seem to align well, with the alignment being worse for the normal distribution. The Q-Q and P-P plots allow a closer examination of the extremes and body of the curve, respectively. The diagonal line in either plot indicates a perfect agreement between empirical and theoretical values. For both plots, the points relating to the gamma distribution align with the diagonal well. At the lower end of the Q-Q plot, a point lies below the diagonal, suggesting the theoretical prediction is not as low as the empirical evidence, while at the upper end a theoretical value is not as high as the empirical. However, the fit at either end is visibly better than for the normal distribution. The P-P plot, which allows for more attention to the body of the curve, reveals a small variation around the diagonal for the gamma, yet no sizeable deviation is observed. In comparison, deviations from the diagonal are generally greater for the normal distribution, with multiple points not located on the diagonal. Given these results, the gamma distribution does indeed describe the empirical data better than the normal distribution. In addition, the description that the gamma distribution offers is considered to be satisfactory for the dataset under investigation. Although small deviations were observed, especially in the Q-Q plot, this was at the extremes and differences were not considered large enough to reject gamma as a suitable distribution. If the fit was not deemed satisfactory, the lognormal could be assessed although it would likely not provide as good of a fit overall as the gamma did given the AIC statistics computed. However, it might provide a better fit for just a part of the data (e.g. the right tail). Alternatively, a new set of distributions would need to be fitted and assessed.

Table 2: Distributions ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Distributions	**AIC**	$\boldsymbol{\Delta_j}$	$\boldsymbol{w_j}$
Gamma	-16.07	0.00	0.75
LogNormal	-13.72	2.34	0.23
Normal	-8.04	8.03	0.01
Weibull	-7.59	8.47	0.01

Discussion

This work aimed at providing step-by-step guidance on how to identify appropriate probability distributions of building modelling inputs given an empirical dataset. This novel technique was demonstrated using a dataset of wall U-value measurements in English dwellings, with the gamma distribution determined to be an appropriate theoretical model for the dataset. Assuming that this is the best dataset currently available, a modeller interested in capturing the variation of U-values for filled cavity walls in English dwellings can draw random samples from a gamma distribution with parameters Gamma(9.52, 13.5).

It is important to reflect on what this method achieves. It enables modellers to identify a *theoretical* distribution, in the form of a parametric probability model, which describes well the *empirical* distribution available (see Wild (2006) for a broader discussion around distributions). We do not claim to have identified a mathematical definition of the data generating process but merely a satisfactory description of the best indication we have of the data generating process, the empirical distribution. The data generating process is the mechanism behind the distribution of true filled cavity wall U-values within the English housing stock. This process would have multiple components, such as the effect of using different insulation materials, the change of building practices and regulations over time or workmanship. Since it is infeasible to accurately model this mathematically, we try to capture a snapshot of this process' output by taking measurements for a sample of filled cavity walls. This results in the empirical distribution of true wall U-values augmented with measurement and sampling errors shown in Figure 1. By following

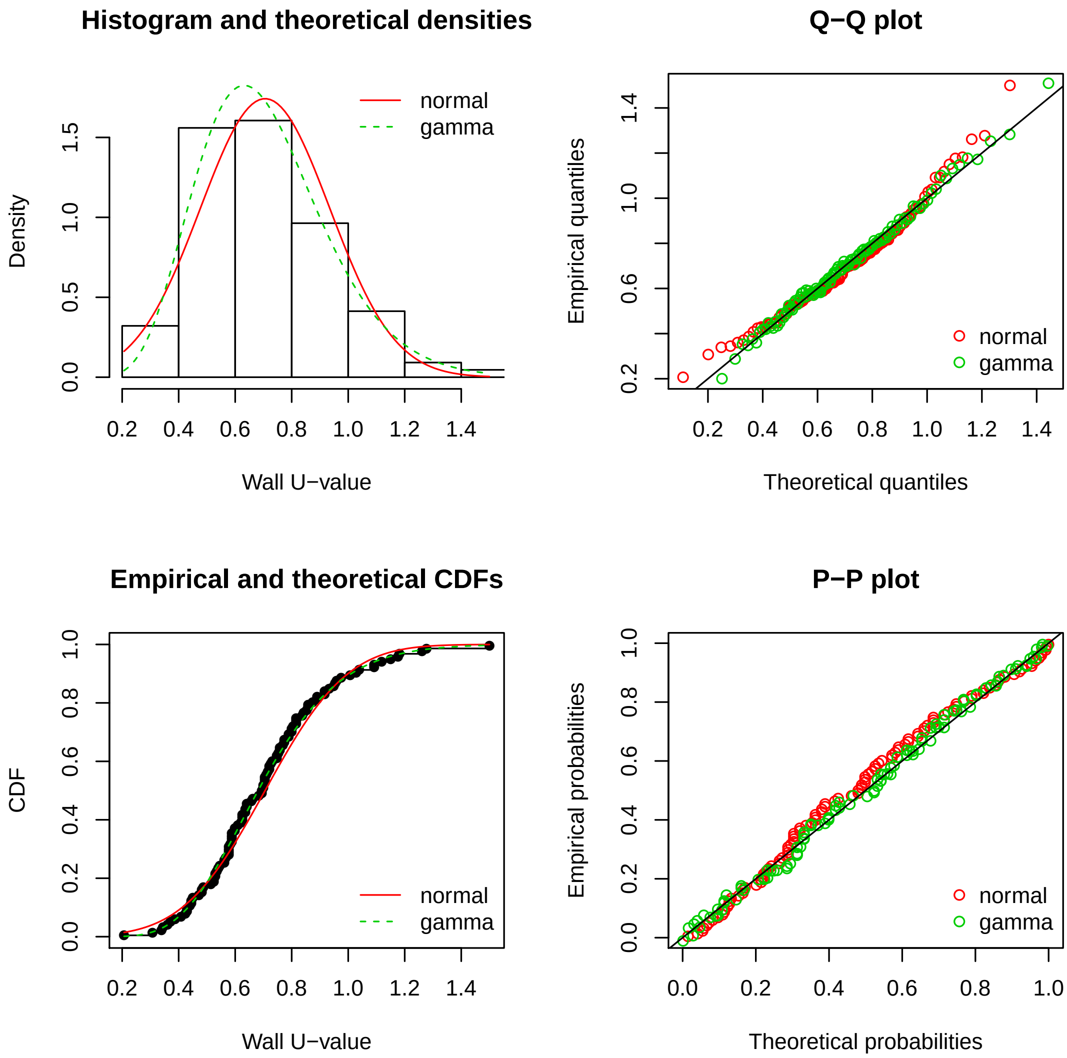

Figure 3: Goodness of fit plots with the empirical data compared to the theoretical values of a gamma and normal distribution.

the procedure described in this paper, a theoretical distribution was identified that adequately describes the dataset. The appropriateness of the assumed distribution will vary depending on the application and the potential impact that a misrepresentation of the data might have on the model output.

Through this work the use of an empirically-based distribution has been recommended over a normal or uniform distribution chosen out of convenience. However, there will be many cases where the normal or uniform distribution will indeed be the best choice of theoretical model and the method detailed in this paper should be able to identify them. For example and as a consequence of Central Limit Theorem, it is expected that a normal distribution would be appropriate when the dataset consists of mean-averaged values (Hughes and Hase, 2014, p. 31).

Although the usefulness of the proposed method in building stock modelling may be evident, its use is not only limited to this application. A similar process can be used to identify distributions of model inputs for individual buildings when repeated measurements may need to be taken, especially when the measurement method is associated with large uncertainties. The approach detailed in this paper may also be pertinent to modelling parameters that vary over the simulation year in a single building, as might be the case for occupancy-related inputs. Finally, certain model outputs may also be represented as distributions and their theoretical form can be identified using the steps summarised in the Methods section.

While the proposed approach should be effective in multiple cases, a few changes could be made if considered necessary. Firstly, it was recommended to use the MLE approach of fitting the candidate distributions. As discussed in the methods section, there might be cases where a different method is preferred that would yield slightly different results and

Delignette-Muller and Dutang (2015) provide a few such examples. In addition, while the workflow is based on the popular programming language R and the package **fitdistrplus**, other options do exist. A modeller can adapt the methodology to be used in their programming language of preference (e.g. **SciPy** in Python) and it is expected that they should have similar results. With present day computing power, the burden of fitting multiple distributions is rather small. Therefore, a modeller may choose in some cases to skip the step of identifying candidate distributions (step 2 in the Methods section) and instead fit all the distributions included in Cullen and Frey graph (see the online repository of this paper on how to easily implement this in R). The rest of the process would remain the same.

Implications

With R being a free to use programming language, the approach described within this paper could easily be implemented by modellers in academia and industry without any licensing or costs constraints. Quantifying the impact of using a distribution that describes the data well compared to a poor and simplistic approximation is rather challenging since it will depend on the specific application. Yet, it is safe to assume that in some cases, given a fairly bad initial assumption of a sensitive model input, the use of inappropriate distributions will contribute to the "performance gap" with the model not representing reality well. This could lead to the underestimation building parameters such as energy use and indoor environmental quality.

Limitations

In this paper a novel method was applied to a single dataset. Applying this across different datasets will likely result in insights relating both to the method but also to what candidate distributions should be considered for different building model parameters. We have applied this method in an ongoing modelling study focused on the English housing stock. During that application, most distributions were identified based on the candidates provided by the Cullen and Frey graph. In two instances, relating to floor area and floor-to-ceiling height, that this was not the case, based on the qualities of their empirical distribution (long right tails), the Fréchet was tested and determined suitable. It is expected that there will be other cases where the distributions provided by the Cullen and Frey graph will not provide an adequate fit. A modeller might address this limitation by reviewing the properties of different distributions and selecting candidates according to their characteristics, similarly to how we identified the Fréchet for our own application. Alternatively, one can leverage the great number of distributions already provided by R and fit several (or all) of them in a "brute-force" approach (see this paper's online repository for a demonstration). This might the preferred choice for modellers who are not familiar with several distributions, although it is the more computationally expensive approach. In either case, a careful evaluation of the goodness-of-fit must follow (step 4).

Conclusion

This paper provides a detailed and novel guidance on how to identify appropriate probability distributions of model inputs from empirical datasets for more accurate model prediction. The method requires the four following steps: (i) Data visualisation and cleaning, (ii) Identification of candidate distributions, (iii) Fit data to candidate distributions and (iv) Identify the candidate distribution with the best goodness-of-fit. To demonstrate its application, a dataset of wall U-value measurements were used and a Gamma(9.52, 13.5) was found to best represent values relating to filled cavity walls in the English housing stock. Amongst the candidate distributions, the use of Akaike weights suggested that probability of the normal being the best-fitting distribution was only 0.01 while the best fitting gamma distribution had a probability of 0.75. In future work, this technique will be applied to multiple datasets.

Acknowledgement

This research was made possible by support from the EPSRC Centre for Doctoral Training in Energy Demand (LoLo), grant number EP/L01517X/1, and the Chartered Institution of Building Services Engineers (CIBSE). In addition, this study was part-funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Environmental Change (NIHR200909). The views expressed are those of the author(s) and not necessarily those of the NIHR, Public Health England or the Department of Health and Social Care. Finally, this research was funded in part by the Wellcome Trust for the Complex Urban Systems for Sustainability and Health (CUSSH) project [Grant number 209387/Z/17/Z].

References

Booth, A. T., R. Choudhary, and D. J. Spiegelhalter (2012). Handling uncertainty in housing stock models. *Building and Environment 48*(Supplement C), 35–47.

BRE (2014a). *Energy Follow Up Survey 2011 - Report Summary of Findings.* Department of Energy and Climate Change. London (UK).

BRE (2014b). *The Government's Standard Assessment Procedure for Energy Rating of Dwellings (SAP 2012).* Department of Energy and Climate Change (DECC). Watford (England).

BRE (2016). *Consultation Paper: CONSP:16. Review of Default U-Values for Existing Buildings in*

SAP. Building Research Establishment.

Burnham, K. P. and D. R. Anderson (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research 33*(2), 261–304.

de Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction 41*, 40–49.

Delignette-Muller, M. L. and C. Dutang (2015). Fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software 64*(1), 1–34.

Hesterberg, T. (2011). Bootstrap. *WIREs Computational Statistics 3*(6), 497–526.

Hughes, I. and T. Hase (2014). *Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis.* Oxford University Press. New York (US).

Hulme, J. and S. Doran (2014). *In-Situ Measurements of Wall U-Values in English Housing.* Number 290-102. Building Research Establishment (BRE) on behalf of the Department of Energy and Climate Change (DECC).

IEA (2019). *2019 Global Status Report for Buildings and Construction: Towards a zero-emission, efficient and resilient buildings and construction sector.* Global Alliance for Buildings and Construction.

IPCC (2018). *Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty.* Intergovernmental Panel on Climate Change.

Li, F. G. N., A. Z. P. Smith, P. Biddulph, I. G. Hamilton, R. Lowe, A. Mavrogianni, E. Oikonomou, R. Raslan, S. Stamp, A. Stone, A. J. Summerfield, D. Veitch, V. Gori, and T. Oreszczyn (2015). Solid-wall U-values: Heat flux measurements compared with standard assumptions. *Building Research & Information 43*(2), 238–252.

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. *Infectious Disease Modelling 5*, 111–128.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Reimann, C., P. Filzmoser, R. G. Garrett, and R. Dutter (2008). *Statistical Data Analysis Explained: Applied Environmental Statistics with R.* John Wiley & Sons. Chichester (England).

Smith, R. C. (2013). *Uncertainty Quantification: Theory, Implementation, and Applications.* Computational Science and Engineering Series. Society for Industrial and Applied Mathematics. Philadelphia (US).

Tian, W., Y. Heo, P. de Wilde, Z. Li, D. Yan, C. S. Park, X. Feng, and G. Augenbroe (2018). A review of uncertainty analysis in building energy assessment. *Renewable and Sustainable Energy Reviews 93*, 285–301.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal 5*(2), 17.