A Method to Facilitate Uncertainty Analysis in LCAs of Buildings

Francesco Pomponi 1,*, Bernardino D’Amico 1 and Alice M. Moncaster 2

1 Institute for Sustainable Construction, Edinburgh Napier University, Colinton Road, Edinburgh EH10 5DT, UK; B.D’Amico@napier.ac.uk
2 Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK; amm24@cam.ac.uk
* Correspondence: F.Pomponi@napier.ac.uk; Tel.: +44-(0)131-455-3590

Abstract: Life cycle assessment (LCA) is increasingly becoming a common technique to assess the embodied energy and carbon of buildings and their components over their life cycle. However, the vast majority of existing LCAs result in very definite, deterministic values which carry a false sense of certainty and can mislead decisions and judgments. This article tackles the lack of uncertainty analysis in LCAs of buildings by addressing the main causes for not undertaking this important activity. The research uses primary data for embodied energy collected from European manufacturers as a starting point. Such robust datasets are used as inputs for the stochastic modelling of uncertainty through Monte Carlo algorithms. Several groups of random samplings between $10^1$ and $10^7$ are tested under two scenarios: data are normally distributed (empirically verified) and data are uniformly distributed. Results show that the hypothesis on the data no longer influences the results after a high enough number of random samplings ($10^4$). This finding holds true both in terms of mean values and standard deviations and is also independent of the size of the life cycle inventory (LCI): it occurs in both large and small datasets. Findings from this research facilitate uncertainty analysis in LCA. By reducing significantly the amount of data necessary to infer information about uncertainty, a more widespread inclusion of uncertainty analysis in LCA can be encouraged in assessments from practitioners and academics alike.

Keywords: LCA buildings; embodied energy; embodied carbon; uncertainty analysis; Monte Carlo simulation; normal distribution; uniform distribution; data distribution; data collection; environmental impact assessment

1. Introduction and Theoretical Background

Life cycle assessment (LCA) is a common method whereby the whole life environmental impacts of products can be assessed. Applied within the built environment, the focus is on assessment of products [1], assemblies [2] and whole buildings [3,4], and is predominantly concerned with energy consumption and greenhouse gas emissions [2,4,5]. However, buildings and assemblies are complex entities with long and unpredictable lifespans; they are designed and built by a fragmented industry through temporary and shifting alliances and dispersed supply chains. LCA, therefore, while providing an indication of environmental impacts, includes inherent uncertainties.

It was more than 25 years ago that the US Environmental Protection Agency [6] brought to researchers’ and practitioners’ attention the role and impact of uncertainty and variability in LCA modelling. Nearly ten years has passed since the importance of the topic resurfaced with Lloyd and Ries [7] who revealed a general lack of appreciation of uncertainty modelling and lack of application in life cycle analysis. Even so, almost all life cycle assessment studies in the built environment continue to
fail to evaluate the inherent uncertainty [8]. A similar trend has also been observed in civil engineering research [9].

This issue, of course, extends beyond the building sector and affects a significant part of the LCA research and community of practice, but the lack of uncertainty analysis in the LCAs of buildings and building products is particularly severe. Not accounting for uncertainty and variability is important, given that it can severely limit the usefulness of LCA [10], and neglecting uncertainty can have a major impact on the understanding of the problem being analysed, skewing outcomes or misleading decisions based on the analyses [11].

Uncertainty mostly occurs because of the lack of knowledge about the true value (quantity) of a datum [12] but it can also be due to subjective choices [13]. At data level, Huijbregts et al. [14] identified two main issues; data inaccuracy and lack of specific data (further distinguished between data gaps and unrepresentative data). Björklund [12] developed a broad classification of types and sources of uncertainty, these being (1) data inaccuracy; (2) data gaps; (3) unrepresentative data; (4) model uncertainty; (5) uncertainty due to choices; (6) spatial and (7) temporal variability; (8) variability between sources and objects; (9) epistemological uncertainty; (10) mistakes; and (11) estimation of uncertainty. Lloyd and Ries [7] proposed a more concise systemic approach to uncertainty, which can be related to three main elements:

Parameter uncertainty (e.g., the uncertainty refers to the values of a parameter such as the embodied energy or carbon of processes or assemblies); Scenario uncertainty (e.g., the uncertainty refers to the likelihood of different scenarios, such as the UK energy mix in 20 years’ time); Model uncertainty (e.g., the uncertainty refers to the specific model being used, such as the model developed by the IPCC to calculate the global warming potential related to GHGs over 20, 50, and 100 years’ horizons).

Parameter and scenario uncertainty are those considered in LCAs and are often combined with sensitivity analysis to understand the influence of each parameter or scenario towards the final results. In some cases, there seems however to be some confusion between sensitivity and uncertainty analysis and the two terms may be confused as synonyms [15]. Rather, sensitivity analysis is aimed at understanding the influence that input parameters have on the final results whereas uncertainty analysis is aimed at understanding the variability of the final results.

Existing literature on the theoretical foundation of computational structure to undertake uncertainty analysis in LCA is slender [16]. Nonetheless, different techniques exist to analyse uncertainty such as: possibility theory, e.g., [17], fuzzy theory, e.g., [18–20], Taylor series expansions, e.g., [21], data quality indicators, e.g., [22,23], expert judgement, e.g., [24–26] and practitioners’ belief, e.g., [27] or a combination of two or more techniques, e.g., [28,29]. Despite the breadth of available approaches, uncertainty analysis in built environment LCAs still remains largely untouched.

Stochastic modelling, mainly in the form of Monte Carlo simulation, is the most widely used approach to analyse uncertainty across different products, research areas, and industrial sectors inter alia, [30–41]. The usefulness of Monte Carlo simulations relies on the assumption that a sufficient number of trials has been carried out [42] in order to claim something about data distribution and uncertainty. Yet, stochastic modelling and uncertainty analysis can be complicated tasks to undertake, and three main reasons are often found:

It requires more data than is available [36,43–45] or is, more generally, too demanding [46]; There is not a definite agreement on the number of iterations that are necessary, which adds to the length (i.e., computational time) and complexity of the calculation [7]; It is perceived as too complex by both academics and practitioners [47]. Work carried out to significantly reduce the required computational power without sacrificing accuracy, e.g., [48], or to compare two classes of error propagation methods, including sampling techniques [49], did not result in a more widespread adoption of Monte Carlo simulation. It therefore remains the case that other methods such as sensitivity and scenario analysis [50–52] are more commonly used to quantify uncertainty.

This article intends to complement the knowledge capital already published in the journal on LCAs of buildings by addressing the issues about the amount of data required, the number of iterations,
and the complexity of stochastic modelling for uncertainty analysis, offering an innovative solution for the construction sector. The next section introduces the theory and the algorithmic approach developed and used. Results follow, along with a discussion on the wider implication of the findings from this research. The final section concludes the article and presents future work that will be undertaken.

2. Theory and Methods

This research aims to investigate the impact of assumptions on data distribution on the results of uncertainty analysis and to provide an innovative simple approach to reduce the complexity of the assessment and time-consuming data collection activities. It is based on robust primary data collected from European manufacturing plants related to the glass and construction industry [53]. (The quality of the data is demonstrated by the pedigree matrix [22], for the data are characterised by the following scores (2), (2), (1), (2), (1) across the relevant five indicators which are: Reliability, Completeness, Temporal Correlation, Geographical Correlation, Further Technological Correlation—where (1) is the highest possible score and (5) is the lowest, assumed as the default value.) The algorithm has also been tested on randomly generated data and it works equally fine.

The fundamental role of uncertainty analysis in LCAs is easily understood by looking at Figure 1. The figure shows a mock example where the embodied energy (EE) of two alternatives, 1 and 2, are being compared. Most LCAs only provide deterministic, single-valued results which generally represent the most likely values resulting from the specific life cycle inventory (LCI) being considered. These values can be seen as $\mu_1$ and $\mu_2$ in the figure. If a decision were to be taken only based on such numbers one should conclude that alternative 1 is preferable over alternative 2 since $\mu_1$ is lower than $\mu_2$. This, unfortunately, is often the case in most existing literature in LCAs of buildings.

![Figure 1. Importance of uncertainty analysis in life cycle analyses (LCAs).](image)

However, when considering the relative uncertainty—which provides measures of the standard deviations $\sigma_1$ and $\sigma_2$ as well as the probability density function—it is evident that alternative 2 has a far less spread distribution and therefore a much narrower range within which values are likely to vary. This makes, all of sudden, alternative 2 at least as appealing as alternative 1, if not more. The information on uncertainty and data distribution does therefore clearly enable a better and more informed decision making.
In the specific case presented here the processes of the life cycle inventory (LCI) are from the manufacture of one functional unit (FU) of double skin façade, a flat glass cladding systems. In total, 200 items have been followed throughout the supply chain and therefore the dataset available for each entry of the LCI is constituted of 200 data points. For example, cutting the flat glass to the desired measure has been measured and monitored (in terms of energy inputs, outputs, and by-products) for all 200 glass panes and therefore a population of 200 data points is available for that specific process. However, the algorithm presented has been tested successfully also on randomly generated data to ensure its wide applicability. The robust primary data set served the purpose of comparing empirical vs. predicted results.

The data and related life cycle processes refer to embodied energy (MJ or kWh) and embodied carbon (kgCO₂e) (for definitions on embodied carbon see, for instance [8]) and both datasets have been used and tested in this research. Embodied energy has however been the main input to the calculation, since embodied carbon is the total carbon dioxide equivalent emissions of embodied energy given the energy mix and carbon conversion factors which are specific to the geographical context under consideration. Therefore, to avoid duplication of figures, only embodied energy results have been included in the article.

Figure 2 shows the specific embodied energy of three such processes against the frequency with which they occur.

Figure 2. Cont.
The graphs in the figure present the cumulative distribution function (CDF) of the collected data plotted against the CDF of the data if they were perfectly normally distributed. The check on the approximation of the real distribution of the dataset versus a normal distribution has been conducted through the Z-test [54]. As the plots in the figure show, there is good agreement between the collected data and normally distributed data, and therefore the hypothesis that the collected data were normally distributed was adopted.

The primary data collection represented a time-consuming and costly activity. Therefore, the research team wondered whether it was possible to infer information on the uncertainty of the result of the LCA as well as its probability distribution without embarking onto such an extensive data collection. To do so, the empirical case of collected primary data characterised by a normal distribution has been used as a reference. This was then compared with a less demanding approach to determine variability in the data. The simplest alternative is generally the use of a maximum–minimum variation range, whereby the only known piece of information is that a value is likely to vary within that range. The likelihood of different values within that range remains, however, unknown. From a probabilistic point of view, the closest case to this maximum–minimum scenario is that of a uniform distribution, which is characterised by a defined data variation range within which all values have the same probability. The comparison between these two alternatives, and their influence on the uncertainty of the result of the LCA, was the underpinning idea for this research. This has been tested through the algorithm developed and explained in detail in the next section.

The Algorithm Developed for This Research

To explain the algorithm, let us assume the life cycle inventory (LCI) of the primary data collection is arranged as a vector P:

\[ P = \{s_1 \ldots s_j \ldots s_m \} \]

with \( m \) representing the total number of sub-processes and the \( j \)th entry being, in turn, another vector \( s \), containing the entire dataset of \( n \) measures of embodied energy, \( x \), associated with the \( j \)th manufacturing sub-process:

\[ s_j = \{x_{j,1} \ldots x_{j,i} \ldots x_{j,n} \} \]

In order to perform a Monte Carlo analysis, an input domain is required from which \( x \) variables can be randomly picked according to a given probability distribution. For each \( j \) sub-process, two sets of continuous input domains are derived in here from the discrete data collection \( P \). Namely: a vector
N containing \( n \) pairs of values \( \mu_j \) and \( \sigma_j \), representing respectively the mean and standard deviation of \( x_i \) values of embodied energy associated with the \( j \)th sub-process.

\[
N = \{(\mu_1, \sigma_1) \cdot \cdots \cdot (\mu_j, \sigma_j) \cdot \cdots \cdot (\mu_n, \sigma_n)\}
\]

where,

\[
\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{j,i}; \quad \sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{j,i} - \mu_i)^2}
\]

as well as a vector \( U \) containing \( n \) pairs of values \( x_{j,\text{min}} \) and \( x_{j,\text{max}} \), representing respectively the minimum and maximum embodied energy values, \( x_j \), associated with the \( j \)th sub-process:

\[
U = \{(x_{1,\text{min}}, x_{1,\text{max}}) \cdot \cdots \cdot (x_{j,\text{min}}, x_{j,\text{max}}) \cdot \cdots \cdot (x_{n,\text{min}}, x_{n,\text{max}})\}
\]

Two sets of Monte Carlo simulations were then run, each with a prescribed increasing number of output samples to be generated (i.e., from \( 10^1 \) to \( 10^7 \)). The inputs for the first set were randomly sampled, under the assumption that data were normally distributed (as from \( N \)), from a probability density function defined as:

\[
f(x_j \mid \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi \sigma_j^2}} e^{-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}}
\]

whereas the assumption of uniform distribution (as from \( U \)) was used for the second set, and data were randomly sampled from a probability density function defined as:

\[
f(x_j) = \begin{cases} 
\frac{1}{x_{j,\text{max}} - x_{j,\text{min}}} & \text{for } x_{j,\text{min}} \leq x_j \leq x_{j,\text{max}} \\
0 & \text{for } x_j < x_{j,\text{min}} \text{ or } x_j > x_{j,\text{max}}
\end{cases}
\]

As explained, the normal distribution was the distribution that best fitted the primary data collected, whereas the uniform distribution was chosen because it is characterised by the lowest number of required inputs where only the lower and upper bounds of the variation range are necessary. As such, it represents the least intensive option in terms of data collection to enable an uncertainty analysis.

The single output obtained at the end of each Monte Carlo iteration is a summation of the (randomly picked) embodied energy values, \( x \), associated with the \( j \)th sub-process. This represents an estimate of the embodied energy associated with the LCI, which is in this case the entire manufacturing process. This is because, mathematically, a life cycle impact assessment (LCIA) equals to:

\[
\text{LCIA} = \sum_{j=1}^{m} x_{S_j}
\]

which represents the summation of the impacts \( x \) of the \( j \)th process \( S \) across the life cycle inventory \( P \) for the relevant impact category being considered (e.g., cumulative energy demand).

The algorithmic approach is also shown visually in Figure 3.

As mentioned, the collection of data to characterise the data distribution is a significantly time-consuming and costly activity, which often severely limits or completely impedes undertaking uncertainty analysis [46]. The literature reviewed has shown that different approaches arose as alternative, more practical solutions. Out of those, one of the most often utilised in practice in the construction industry is the use of expert judgement to identify a range within which data is expected to vary [24,25]. This is then repeated for several or indeed all life cycle processes that constitute the life cycle inventory and it generally leads to calculating a minimum and maximum value for the overall impact assessment, which are then labelled as ‘worst’ and ‘best’ case scenario.
It has been shown that expert judgements can provide an accurate overview of the variability range [26], and this characteristic held true in the case of the research underpinning this article. Practitioners and professionals involved in the life cycle processes for which primary data have been collected have indeed shown a remarkably accurate sensitivity to the processes variation range. As a result, it would have been possible to identify the same (or significantly similar) data range, which resulted from the extensive data collection by asking people in the industry: “What are the best and worst case scenarios in terms of energy input for this specific process?” However, the answer to such a question tells nothing about the way data are distributed between these scenarios, and both values represent the least probable events, as they are, statistically speaking, the farthest points from the mean (i.e., most likely value). If this approach were propagated throughout the whole LCI, the result would lead to an overall range for the impact assessment which is characterised by significant confidence in terms of inclusion of the true value but the numbers it produces would likely be of very little use. As a consequence, the resulting decisions could be significantly biased, with the risk of invalidating the merit and efforts of the whole LCA [10].

Therefore, to combine the benefits of a lighter data collection with the benefit of an uncertainty analysis, the algorithmic approach developed for this research has tested whether and to what extent the knowledge of a data distribution within a known data variation range influences the outcome of an uncertainty analysis undertaken through Monte Carlo simulation. The algorithmic approach has been developed, implemented and tested in MATLAB R2015b (8.6.0). Once initial results were obtained, the robustness of the findings has been further tested for validity by one of the co-authors who independently implemented an algorithm in Python with the aim of addressing the same research problem. Findings were therefore confirmed across the two programming languages with different algorithms written by two of the authors to strengthen the reliability of our results. A comparison of the results produced in MATLAB and Python 3.5 is shown in Figures 4 and 5 for the mean values $\mu$ and the standard deviations $\sigma$ respectively.
Figure 4. Difference between results produced in MATLAB and Python for the mean values of embodied energy of the LCI.

Figure 5. Difference between results produced in MATLAB and Python for the standard deviation values of embodied energy of the life cycle inventory (LCI).

To broaden the impact and applicability of the approach developed, and to strengthen the relevance of its findings, two extreme cases have been tested:

- LCI is constituted by as few as two entries (e.g., only two life cycle processes)—an example could be a very simple construction product or material such as unfired clay;
• LCI is constituted by as many entries as those for which collected primary data were available.

In terms of number of samplings and iterations, each run of the Monte Carlo algorithm randomly and iteratively samples $10^i$ (with $i = 1, \ldots, 7$ and $10^{i+1}$ step increases) values from within each range under the pertinent assumption regarding data distribution (that is, uniform in one case and normal in the other). This process repeats across all entries of the LCI and produces the final figures for the impact assessment. The random sampling mechanism is not pre-assigned or registered, and it varies at each and new run of the algorithm. The algorithm stops after 10 different runs, a high enough number to ensure that potential biases in the variability with which the random sampling operates would emerge.

3. Results and Discussion

Figure 6 shows the results for one specific run out of the ten the algorithm runs for both hypotheses on data distribution: uniform ($S_1$) and normal ($S_2$).

The histograms refer to the LCIA (overall impact) and not to the LCI (individual inputs). All seven sampling cases are shown, from $10^1$ to $10^7$ random samplings. The figures show that for lower numbers of samplings the characteristics of the two distributions are still evident whereas from $10^4$ random samples upwards the results of the two methods converge. The algorithm calculates also the difference between the mean values and standard deviations and normalises them to a percentage. This information is presented in the last (bottom-right) graph of Figure 6. It can be seen how the difference between $\mu$ and $\sigma$ under the two hypotheses is not influential anymore after $10^3$–$10^4$ samplings. If for the mean this could be expected as a consequence of the central limit theorem, this finding is noteworthy for the standard deviation.

Figure 6. Cont.
Figure 6. Example of the results of the developed algorithm for one specific run (results refer to embodied energy (EE) (MJ) but are equal in trend in the case of embodied carbon (EC) (kgCO$_2$e)).

Figure 7 shows the $\mu$ and $\sigma$ variation (percentage) between the two hypotheses across all 10 runs in the case of a simplistic LCI composed of two entries. In all runs, regardless of the initial difference for lower number of samplings, after $10^5$ random samplings the average differences stabilise around:

- 0.01% for the $\mu$, and,
- 1% for the $\sigma$.

Figure 8 presents the same results but for the full LCI as explained in the research design section. In the case of the detailed LCI, after $10^4$ random samplings the average differences stabilise around:

- 0.01% for the $\mu$, and,
- 1% for the $\sigma$.

This demonstrates the validity of the approach developed regardless of the size of the LCI (i.e., number of entries from which the algorithm samples randomly). It should be noted that despite it takes $10^5$ samplings to achieve perfect convergence in the case of an LCI made of just two entries, the stabilisation of any variation around 0.01% for the $\mu$ and 1% for the $\sigma$ is already clearly identifiable from $10^4$ samplings.

In terms of computational costs, the algorithm is extremely light and the whole lot of 10 runs, each of which has seven iterations, from $10^1$ to $10^7$ random samplings, only takes 61.594 s to run in the case of the full LCI on a MacBook Pro. Specifically, the core computations take 38 s.

To ensure that the LCI with as little as two entries would be representative of a generic case and not just of a fortunate combination of two processes that confirm the general case, we have further tested this simplified inventory through five combinations of two processes randomly picked from the database. These have been tested again on both embodied energy and embodied carbon data, and graphical results for the embodied carbon in all five cases are provided as supplementary material attached to this article (Figures S1–S5).
These findings address current challenges in uncertainty analysis in LCA as described in the introduction, and have implications for both theory and practice. Firstly, it has been shown that extensive data collection to characterise the data distribution is not necessary to undertake an uncertainty analysis. As few as two numbers (the upper and lower bounds of the data variation range) combined with a sensible use of the power of Monte Carlo simulation suffice to characterise and propagate uncertainty. Secondly, it has also been found that $10^4$ random samplings are sufficient to achieve convergence, thus establishing a reference number for random samplings within Monte Carlo simulation, at least for uncertainty analysis from LCIs similar to the one in question, i.e.,
The research presented in this article has addressed current challenges in uncertainty analysis in LCA of buildings. By means of an innovative approach based on Monte Carlo simulation, it has tested the influence of two different assumptions on data distribution towards the final results of the uncertainty analysis. The different assumptions on the data sets tested in this research were (1) data...
with a normal distribution and (2) data with a uniform distribution, both based on primary data collected as part of a robust dataset. Results have demonstrated that after $10^4$ random samplings from within the data variation range, the initial assumption on whether the data were normally or uniformly distributed loses relevance, both in terms of mean values and standard deviations.

The consequence of the findings is that an initial characterisation and propagation of uncertainty within a life cycle inventory during an LCA of buildings can be carried out without the usually expensive and time consuming primary data collection. The approach presented here therefore allows a simplified way to include uncertainty analysis within the LCA of complex construction products, assemblies and whole buildings. In turn, this should encourage increased confidence in, and therefore increased uptake of, the LCA calculation in the construction industry, leading eventually to meaningful and reliable reduction of environmental impacts.

Future work will test the influence of more data distributions used and found in LCA practice in the built environment, including triangular, trapezoidal, lognormal, and beta distributions. In its current form the approach developed requires the use of MATLAB, which could be a limit to its widespread adoption. As a consequence, we intend to develop a web-based free application hosted on an academic website where academics and practitioners alike can update their data as Excel or .csv files and get results in both graphical and datasheet formats. In the meantime, interested users are encouraged to contact the authors to find out more about the algorithm or to have their LCI data processed for uncertainty analysis.

**Supplementary Materials:** The following are available online at www.mdpi.com/1996-1073/10/4/524/s1, Figures S1–S5: $\mu$ and $\sigma$ variation (percentage) between the two hypotheses across 5 runs (for 5 random combinations of two LCI entries).

**Acknowledgments:** Funding for this research was received from EPSRC (Award Ref. EP/N509024/1) and the Isaac Newton Trust (RG74916). The authors are grateful to William Fawcett and Martin Hughes who listened to the ideas behind this research and provided useful and constructive feedback as well as positive and encouraging comments. Grateful thanks are also extended to the reviewers for their inputs and comments throughout the review process.

**Author Contributions:** Francesco Pomponi collected the primary data, conceived and designed the research, and wrote the algorithm in MATLAB; Bernardino D’Amico wrote the algorithm in Python to validate the research findings; all authors contributed to multiple drafts that gave the article its present form.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


15. Roeder, M.; Whittaker, C.; Thornley, P. How certain are greenhouse gas reductions from bioenergy? Life cycle assessment and uncertainty analysis of wood pellet-to-electricity supply chains from forest residues. *Biomass Bioenergy* 2015, 79, 50–63. [CrossRef]


© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).