Deepening the uncertainty dimension of environmental Life Cycle Assessment: addressing choice, future and interpretation uncertainties.

Mendoza Beltran, M.A.

5.

Quantified Uncertainties in Comparative Life Cycle Assessment: What Can Be Concluded?

Angelica Mendoza Beltran
Valentina Prado
David Font Vivanco
Patrik J.G. Henriksson
Jeroen B. Guinée
Reinout Heijungs

Re-print with minor changes from:

Environmental Science & Technology (2018). 52(4): 2152–2161.
doi/10.1021/acs.est.7b06365.

Abstract

Interpretation of comparative Life Cycle Assessment results (LCA) can be challenging in the presence of uncertainty. To aid in interpreting such results under the goal of any comparative LCA, we aim to provide guidance to practitioners by gaining insights into uncertainty-statistics methods (USMs). We review five USMs - discernibility analysis, impact category relevance, overlap area of probability distributions, null hypothesis significance testing (NHST), and modified NHST-, and provide a common notation, terminology, and calculation platform. We further cross-compare all USMs applying them to a case study on electric cars. USMs belong to a confirmatory or an exploratory statistics' branch, each serving different purposes to practitioners. Results highlight that common uncertainties and the magnitude of differences per impact are key in offering reliable insights. Common uncertainties are particularly important as disregarding them can lead to incorrect recommendations. Based on these considerations, we recommend the modified NHST as a confirmatory USM. Also, we recommend discernibility analysis as an exploratory USM along with recommendations for its improvement, as it disregards the magnitude of the differences. While further research is necessary to support our conclusions, results and supporting material provided can help LCA practitioners in delivering a more robust basis for decision-making.

Keywords: Comparative LCA, Uncertainty, Interpretation, Decision-making

5.1 Introduction

One of the main applications of life cycle assessment (LCA) is to support a comparative assertion regarding the relative environmental performance of one product with respect to other functionally equivalent alternatives (ISO 2006). In such a comparative LCA, claims can be tested by comparing the inventory and/or impact assessment results for any given set of alternative products (JRC-IES 2010). To date, practitioners usually calculate and compare point-value results, an approach described as deterministic LCA (Wei et al. 2016). This practice allows one to draw conclusions such as 'alternative B causes 45% larger impacts than alternative A' or 'alternatives B and C have strengths and weaknesses, but both outperform alternative D'. Typically, deterministic comparative LCAs find trade-offs between alternatives and across environmental impacts (from here on referred to as impacts). While uncertainty estimations can be useful in understanding trade-offs between alternatives, deterministic LCAs lack an assessment of uncertainties (Ross et al. 2002).

Uncertainty appears in all phases of an LCA (Björklund 2002; Huijbregts et al. 2003; Wiloso et al. 2014) and originates from multiple sources. Some of the more prevalent are: variability, imperfect measurements (inherent uncertainty (Henriksson et al. 2014)), gaps, unrepresentativeness of inventory data (also known as parameter uncertainty) (Björklund 2002), methodological choices made by practitioners throughout the LCA (also known as scenario uncertainty or uncertainty due to normative choices) (Björklund 2002) and mathematical relationships (also known as model uncertainty) (Björklund 2002). Using analytical and stochastic approaches, e.g. Monte Carlo (MC) simulations and first order Taylor series expansion (Groen et al. 2014), LCA practitioners have propagated these sources of uncertainty to LCA results (Lloyd and Ries 2008; Groen et al. 2014). Unlike deterministic LCA, the quantification of uncertainties related to LCA results allows for associating a level of likelihood to and confidence in the conclusions drawn. However, interpreting overlapping ranges of results is complex and therefore requires sophisticated interpretation methods (Lloyd and Ries 2008). To this end, various statistical methods have been applied within the field of LCA, including: discernibility analysis (Heijungs and Kleijn 2001; Gregory et al. 2016), impact category relevance (Prado-Lopez et al. 2014), overlap area of probability distributions (Prado-Lopez et al. 2016), null hypothesis significance testing (NHST) (Henriksson et al. 2015a, b), and modified NHST (Heijungs et al. 2016).

The application of statistical methods to uncertainty analysis results, hereafter referred to as 'uncertainty-statistic methods' (USMs), can aid practitioners in various ways. First, they help to establish a level of confidence behind the trade-offs between alternatives and across environmental impacts while considering various sources of uncertainty. Second, they go beyond the practice of one at the time scenario analysis by integrating series of otherwise independent sensitivity analyses into an overall uncertainty

assessment of results (Ross et al. 2002). For instance, they enable the exploration of a broad range of possible combinations of all sorts of input data known as the scenario space (Gregory et al. 2016). Third, they allow for comparisons of alternatives in the context of common uncertainties, a crucial aspect in comparative LCAs (Henriksson et al. 2015a). Lastly, they help to identify the relative importance of different impacts for the comparison of alternatives (Hertwich and Hammitt 2001).

Choosing the most appropriate statistical method(s) to interpret the results of uncertainty analysis in the light of the goal and scope of individual LCA studies can be challenging. There is a lack of applications of these methods in real case studies, a lack of support in standard LCA software, incomprehensive and scattered documentation, and inconsistent terminology and mathematical notation. Moreover, literature is devoid of recommendations for LCA practitioners about which method(s) to use, under which LCA goal, to interpret the meaning of the uncertainty analysis results in comparative LCAs. Thus, our research question queries: "*Which statistical method(s) should LCA practitioners use to interpret the results of a comparative LCA, under the light of its goal and scope, when considering uncertainty?*" In this chapter, we answer this question by (1) critically reviewing the five above mentioned USMs, (2) comparing them for a single illustrative case study on passenger vehicles with a common calculation platform and terminology, and (3) by providing guidance to practitioners in the realm of application of these methods via a decision tree. It is the focus of this chapter to test the applicability and value of different USMs, including the visualization of results and the limitations encountered during their implementation. Testing and analyzing differences in methods to quantify and propagate uncertainties is out of the scope of this chapter, although we use some of them (e.g. Monte Carlo simulations as propagation method) for the uncertainty analysis.

5.2 Methods and case study

Statistical methods for interpretation of comparative LCA with uncertainty

In chronological order of publication, the methods we study are: discernibility analysis (Heijungs and Kleijn 2001; Gregory et al. 2016), impact category relevance (Prado-Lopez et al. 2014), overlap area of probability distributions (Prado-Lopez et al. 2016), null hypothesis significance testing (NHST) (Henriksson et al. 2015a, b), and modified NHST(Heijungs et al. 2016). The scope was narrowed to these statistical methods based on two criteria:

1) The method has been developed and published in peer reviewed journals and contains transparent and accessible algorithms. Consequently, the first-order reliability method (FORM) (Wei et al. 2016), could not be included due to incompletely documented optimization procedures.

2) The method is applied to interpret the results of uncertainty analysis of comparative LCAs with two or more alternatives and one or more emissions or impacts. This excludes studies addressing different impacts but not in a comparative way (Grant et al. 2016) and, studies focusing on methods for quantifying and/or propagating uncertainty sources through LCA. Studies developing and describing methods such as global sensitivity analysis (Groen et al. 2017) are also excluded as they are neither comparative and focus on just one emission or impact at a time. Finally, we have not revisited the enormous body of statistical literature, as the authors of the selected methods already have done this exercise.

To increase transparency in our comparison of methods and their features, we use a uniform terminology (Appendix I), and a common mathematical notation (Table 10). We interpret the state of the art for each method, and in some cases go beyond the original mathematical proposals by the authors. When this is the case, we indicate the differences.

We reviewed the methods according to the following aspects: the number of alternatives compared and approach to compare them, the inputs used by the method, the implementation, the purpose and the type of outputs. Table 11 summarizes the features of each method according to these aspects.

Some features that are consistent for all methods include: 1) they can be applied to dependently or independently sampled MC runs, meaning that the uncertainty analysis results are (dependently) or not (independently) calculated with the same technology and environmental matrices for all alternatives considered for each MC run; 2) they can be used to interpret LCA results at the inventory, characterization, and normalization level, although in our case study we only apply them at the characterization level as their use at other levels is trivial in the absence of additional uncertainties; 3) they all compare alternatives per pairs (pairwise analysis); and 4) they all originate from the idea of merging uncertainty and comparative analysis.

Discernibility

We refer to discernibility as the method described by Heijungs and Klein (2001) as the basis of comparative evaluation of Gregory et.al (2016) is the same as proposed by Heijungs and Klein (2001). Discernibility compares two or more alternatives, using a pairwise method as the comparison takes place by pair of alternatives, comparing the results of alternative j with alternative k per MC run. It assesses the stochastic outcomes on whether the results of one alternative are higher or lower than another alternative. The purpose of discernibility is to identify whether the results of one of the alternatives are higher than (irrespective of how much higher) the results of the other. This method disregards the distance between the mean scores (or other centrality parameters). For its operationalization, practitioners count how many realizations per pair of alternatives

Table 10. Mathematical notation for comparison of uncertainty-statistics methods (USMs)

Symbol	Description
j, k	Index of alternatives e.g. products, services, systems, etc ($j = 1\ldots, n\, j = 1\ldots, n, k = 1, \ldots, n\, k = 1, \ldots, n$)
i	Impact category (Climate change, eutrophication, acidification,…)
r	Index of Monte Carlo simulations (r=1, ,…, N)
X	Random variable
x	Realization
μ	Parameter of centrality (mean)
σ	Parameter of dispersion (standard deviation)
$\bar{X}$	Statistic of centrality (estimator of mean μ)
S	Statistic of dispersion (estimator of standard deviation σ)
$\bar{x}$	Obtained value of centrality (estimate of mean μ)
s	Obtained value of dispersion (estimate of standard deviation σ)
$f_{i,j,k}$	Fraction of runs with higher results on impact category i in alternative j compared to k
$\#(x)$	Count function, counts the number of runs fulfilling condition x
$\Upsilon_{i,j,k}$	Relevance parameter for the pair of alternatives j, k on impact category i
$A_{i,j,k}$	Overlap area of two probability distributions for the pair of alternatives j, k on impact category i

Table 11. Features of the different uncertainty-statistics methods (USMs) in comparative LCA

Methods	Alternatives compared (approach)	Type of input (From uncertainty analysis)	Implementation	Purpose (Type of question)	Type of output	Reference
Deterministic LCA (Comparison of point values)	As many as required (all together)	None	Overall (i.e based on one run or point-value)	Which alternative displays the lower results? (Exploratory)	Point-value	Abundant in literature. Included as standard result in LCA software packages
Discernibility	As many as required (pairwise analysis)	Monte Carlo runs (dependently or independently sampled)	Per run	How often is the impact i higher for j than for k, or vice versa? (Exploratory)	Counts meeting "sign test" condition (equation 3)	Heijungs and Klein (Heijungs and Kleijn 2001)
Impact category relevance	As many as required (pairwise analysis)	Estimates of statistical parameters (i.e. mean and standard deviation)	Overall (i.e. based on statistical parameters)	Which are the impacts playing a relatively more important role in the comparison of j and k? (Exploratory)	Measure of influence of impacts in the comparison (equation 4)	Prado-Lopez et al.(Prado-Lopez et al. 2014)
Overlap area of probability distributions	As many as required (pairwise analysis)	Moments of the fitted distribution (e.g. maximum likelihood estimates)	Overall (i.e. based on moments of the fitted distribution)	Which are the impacts playing a relatively more important role in the comparison of *j* and k? (Exploratory)	Overlap of probability distributions of j and k (equation 5)	Prado-Lopez et al.(Prado-Lopez et al. 2016)
Null hypothesis significance testing (NHST)	As many as required (pairwise analysis)	Monte Carlo runs (dependently or independently sampled)	Per run	Is the mean impact of j significantly different from the mean impact of k? (Confirmatory)	p-values Fail to reject (no) or reject (yes) the null hypothesis	Henriksson et al.(Henriksson et al. 2015a)
Modified NHST	As many as required (pairwise analysis)	Monte Carlo runs (dependently or independently sampled)	Per run	Is the difference between the mean impact of j and k at least as different as a threshold? (Confirmatory)	p-values Fail to reject (no) or reject (yes) the null hypothesis	Heijungs et al.(Heijungs et al. 2016)

per impact i.e. $x_{i,j,r}$ and $x_{i,k,r}$ for $r = 1,\ldots,N$ meet the "sign test" condition. The counting function is indicated by the symbol $\#(\cdot)$, where the argument of the function specifies the "sign test" condition. We interpret these condition as the evaluation of whether the difference between the results per run for a pair of alternatives is bigger than zero. Equation 3 shows the calculations of the discernibility approach for each impact.

$$f_{i,j,k} = \frac{\#_{r=1}^{N}\left(x_{i,j,r} - x_{i,k,r} > 0\right)}{N}$$

Eq.3

The results of Equation 3 help assert that "Alternative j has a larger impact than alternative k in $100 \times f$ % of runs".

Impact category relevance

This approach evaluates trade-offs using the relevance parameter ($\Upsilon_{i,j,k}$), as introduced in Prado-Lopez et al (2014) and it is not intended to calculate statistical significance. It stems from the idea that similar impacts among alternatives do not influence the comparison of alternatives as much as impacts for which alternatives perform very different. It uses the mean (statistics of centrality, $\bar{X}_{i,j}$, $\bar{X}_{i,k}$) and standard deviation (statistic of dispersion, $S_{i,j}$, $S_{i,k}$) calculated from the obtained values for each impact ($X_{i,j,r}$ and $X_{i,k,r}$), thus not per MC run. The value of $\Upsilon_{i,j,k}$, has no meaning on its own, rather its purpose is to help explore the comparison of two alternatives by means of sorting according to the extent of the differences per impact. This approach is therefore exclusive to analysis with more than one impact. When uncertainties increase (as indicated by larger standard deviations) or the difference between the means of two alternatives gets closer to zero (as indicated by nearly equal means), it becomes harder to distinguish between the performance of two alternatives for an environmental impact and hence this aspect is deemed to have a lower relevance in the comparison. A higher relevance parameter for a specific impact indicates that this impact is more important to the comparison than others. The relevance parameter works as a pairwise analysis, as shown in Equation 4.

$$\Upsilon_{i,j,k} = \frac{\left|\bar{x}_{i,j} - \bar{x}_{i,k}\right|}{\frac{1}{2}\left(s_{i,j} + s_{i,k}\right)}$$

Eq.4

In this formula we interpret (in comparison to the original description of the method (Prado-Lopez et al. 2014)) μ as $\bar{x}$, because μ is unknown and only estimated by $\bar{x}$. Further, we interpreted the ambiguous SD in the original publication (Prado-Lopez et al. 2014), into s, which is an estimate of σ.

Overlap area of probability distributions

This method follows the same idea as the relevance parameter, but instead provides an indicator based on the overlap area of probability distribution functions (PDF). Similar to the relevance parameter, this method is not calculated per run and there is no significance threshold value in the overlap that defines statistical significance. The overlap area approach is exclusive to analysis with more than one impact (Prado-Lopez et al. 2016). It measures the common area between PDF of the stochastic impact results ($X_{i,j}$ *and* $X_{i,k}$) of two alternatives j and k, for a specific impact i. By doing this, the overlap area approach can technically apply to diverse types of distributions as opposed to assuming a normal distribution. The shared area between distributions ranges from one, when distributions are identical, to zero, when they are completely dissimilar. The smaller the overlap area, the more different two alternatives are in their performance for an impact. To compute the overlap area ($A_{i,j,k}$), two strategies can be followed. A conventional way is to assume a probability distribution for both $X_{i,j}$ and $X_{i,k}$ (for instance, a normal or lognormal distribution), to estimate the parameters ($\mu_{i,j}$, $\mu_{i,k}$, $\sigma_{i,j}$, $\sigma_{i,k}$) from the MC samples, and to find the overlap by integration. This is the approach followed by Prado-Lopez et al. (2016), using lognormal distributions. The second approach does not require an assumption on the distribution, but uses the information from the empirical histogram, using the Bhattacharyya coefficient (Kailath 1967). To our knowledge, the latter approach has not been used in the field of LCA. Here, we calculate the overlap area using the first approach. In our case, the statistic of centrality ($\bar{X}_{i,j}, \bar{X}_{i,k}$) and dispersion ($S_{i,j}$, $S_{i,k}$) of the assumed lognormally distributed stochastic impact results were calculated by means of the maximum likelihood estimation of parameters. The lower intercept (θ) and the upper intercept (ψ) of the two PDFs, are calculated using these parameters and used as a base to calculate the overlap area between two distributions (equation 5). Details on the calculation of θ and ψ, as well as the maximum likelihood estimation of parameters μ and σ, and the PDF Φ are described in the supporting information (SI, appendix II).

$$A_{i,j,k} = 1 - \left|\Phi\left(\theta; \mu_{i,j}, \sigma_{i,j}\right) - \Phi\left(\theta; \mu_{i,k}, \sigma_{i,k}\right)\right| - \left|\Phi\left(\psi; \mu_{i,j}, \sigma_{i,j}\right) - \Phi\left(\psi; \mu_{i,k}, \sigma_{i,k}\right)\right| \qquad \textit{Eq.5}$$

This method uses a pairwise analysis, yet when more than a pair of alternatives is compared, Prado-Lopez et al. (2016) proposed an averaging procedure for the overlap areas between all pairs. For reasons of comparability with the other methods, we did not pursue this extension and concentrate on the comparison per pair.

Null hypothesis significance testing (NHST)

This method is delineated in Henriksson et al.(2015a) and applied in Henriksson et al. (2015b). It largely relies upon established null hypothesis significance tests. In comparative LCAs, a generally implicit null hypothesis presumes that two alternatives perform environmentally equal: $H_0: \mu_{i,j} = \mu_{i,k}$. This method's purpose is to show whether the centrality parameter (mean or median) of the relative impacts of two alternatives are statistically significantly different from each other. It builds on the quantification and propagation of overall dispersions in inventory data (Henriksson et al. 2014) to stochastic LCA results ($X_{i,j}$ and $X_{i,k}$). From the stochastic results per impact, the difference per pair of alternatives per MC run is calculated ($x_{i,j,r} - x_{i,k,r}$). This distribution of differences can then be statistically tested using the most appropriate statistical test with regards to the nature of the data, as proposed by Henriksson et.al (2015a). For instance, for normally distributed data, a paired t-test is appropriate to determine whether the mean of the distribution significantly differs from zero (the hypothesized mean). For non-parameterized data, more robust statistical tests, such as Wilcoxon's rank test, can be used. When three or more alternatives are compared, a two-way ANOVA can be used for normally distributed data, while a Friedman test can be used in more general cases. In both of these cases a post-hoc analysis is also required to establish significantly superior products. The null hypothesis of equal means (or medians) may then be rejected or not, depending on the *p*-value and the predefined significance level (α), e.g., α = 0.05. For our case, we apply a paired t-test to the distribution of the difference per pair of alternatives and MC run, because the mean is expected to be normally distributed as the number of runs is relatively large (1000 MC runs) (Agresti and Franklin 2007). We also explored a Bonferroni correction of the significance value from $\alpha = 0.05$ to $\alpha_b = 0.05/30 = 0.0016$ as the chance of false positives is rather high when multiple hypothesis tests are performed (Mittelhammer et al. 2000). The factor 30 is explained by the ten impacts and the three pairs of alternatives.

Modified NHST

Heijungs et al. (2016b) proposed this method as a way to deal with one of the major limitations encountered while applying NHST to data from simulation models: significance tests will theoretically always reject the null hypothesis of equality of means since propagated sample sizes are theoretically infinite. It is a method that attempts to cover significance (precision) and effect of size (relevance). Thus, from the classic H_0 in NHST that assumes "no difference" between the parameters ($\mu_{i,j} = \mu_i, k$), this method includes a "at least as different as" in the null hypothesis, which is stated as $H_0: S_{i,j,k} \leq \delta_0$ where $S_{i,j,k}$ is the standardized difference of means (also known as Cohen's d (Cohen 1988)) and δ_0 is a threshold value, conventionally set at 0.2 (Heijungs et al. 2016). So far the method has not been applied in the context of comparative LCA outside of Heijungs et al. (Heijungs et al. 2016). For its operationalization, the authors

proposed the following steps (Heijungs et al. 2016): 1) set a significance level (α); 2) set the difference threshold (δ_0); 3) define a test statistic D (see equation 6, which is a modification from the original proposal (Heijungs et al. 2016)); and 4) test the null hypothesis $H_0: \delta \leq \delta_0$ at the significance level α.

$$d_{i,j,k} = \frac{\bar{x}_{i,k} - \bar{x}_{i,j}}{s_{i,j,k}} \text{ that estimates } \delta_{i,j,k} = \frac{\mu_{i,k} - \mu_{i,j}}{\sigma_{i,j,k}}$$

$$s_{i,j,k} = \sqrt{\frac{1}{N-1}\sum_{r=1}^{N}\left(\left(x_{i,k} - x_{i,j}\right) - \left(\bar{x}_{i,k} - \bar{x}_{i,j}\right)\right)^2}$$

Eq.6

In equation 6, $s_{i,j,k}$ is the standard deviation of the difference between alternatives *j* and *k*. The t-value from the value of d as shown in equation 7. The t-value is a test statistic for t-tests that measures the difference between an observed sample statistic and its hypothesized population parameter in units of standard error.

$$t_{i,j,k} = \frac{d_{i,j,k} - \delta_0}{\sqrt{\frac{1}{N}}}$$

Eq.7

For our case, we consider the default values suggested by Heijungs et al. (2016b) where $\alpha = 0.05$ and $\delta_0 = 0.2$, and we calculate the test statistic D for the three pairs of alternatives (Equation 6 and 7). We also explored the significance with $\alpha_b = 0.0016$ as done for the NHST.

Case study for passenger vehicles

A case study for a comparative LCA that evaluates the *environmental performance of powertrain alternatives for passenger cars in Europe* is used to illustrate the USMs. Comparative assertions are common among LCAs that test the environmental superiority of electric powertrains over conventional internal combustion engines(Hawkins et al. 2012). Several LCA studies have comparatively evaluated the environmental performance of hybrid, plug-in hybrid(Samaras and Meisterling 2008; Nordelöf et al. 2014), full battery electric (Notter et al. 2010; Majeau-Bettez et al. 2011), and hydrogen fuel cell vehicles(Granovskii et al. 2006; Font Vivanco et al. 2014). Many of these studies describe multiple trade-offs between environmental impacts: while electric powertrains notably reduce tailpipe emissions from fuel combustion, various other impacts may increase (e.g. toxic emissions from metal mining related to electric batteries(Hawkins

et al. 2013). Against this background, electric powertrains in passenger vehicles are an example of problem shifting and a sound case to test comparative methods in LCA.

Goal and Scope

The goal of this comparative LCA is to illustrate different USMs by applying these methods to the uncertainty analysis results for three powertrain alternatives for passenger cars in Europe: a full battery electric (FBE), a hydrogen fuel cell (HFC), and an internal combustion engine (ICE) passenger car. The functional unit for the three alternatives corresponds to a driving distance of 150,000 vehicle-kilometers (vkm). The scope includes production, operation, maintenance, and end of life. The flow diagram for the three alternatives can be found in the SI (Appendix III). The case has been implemented in version 5.2 of the CMLCA software (www.cmlca.eu), and the same software has been used to propagate uncertainty. The five USMs have been implemented in a Microsoft Excel (2010) workbook available in the SI.

Life Cycle Inventory

The foreground system was built using existing physical inventory data for a common glider as well as the FBE and ICE powertrains as described by Hawkins and colleagues (Hawkins et al. 2013), whereas the HFC power train data is based on Bartolozzi and colleagues (Bartolozzi et al. 2013). The background system contains process data from ecoinvent v2.2, following the concordances described by the original sources of data. A complete physical inventory is presented in the SI (Appendix IV). The uncertainty of the background inventory data corresponds to the pedigree matrix (Weidema and Wesnæs 1996) scores assigned in the ecoinvent v2.2 database. In addition, overall dispersions and probability distributions of the foreground inventory data have been estimated by means of the protocol for horizontal averaging of unit process data by Henriksson et al. (2014). Thus, the parameters are weighted averages with the inherent uncertainty, spread, and unrepresentativeness quantified. Specifically, unrepresentativeness was characterized in terms of reliability, completeness, temporal, geographical, technological correlation, and sample size (Frischknecht et al. 2007), to the extent possible based on the information provided in the original data sources. Further details of the implementation of parameter uncertainty are presented in the SI (appendix IV).

Life Cycle Impact Assessment (LCIA)

The environmental performance of the selected transport alternatives is assessed according to ten mid-point impact categories, namely: climate change, eutrophication, photochemical oxidation, depletion of abiotic resources, acidification, terrestrial ecotoxicity, ionizing radiation, freshwater ecotoxicity, stratospheric ozone depletion, and human toxicity. The characterization factors correspond to the CML-IA factors without long term effects (version 4.7) (CML - Department of Industrial Ecology 2016), and

exclude uncertainty. No normalization or weighting was performed and the results are presented at the characterized level.

Uncertainty calculations

Uncertainty parameters of background and foreground inventory data were propagated to the LCA results using 1000 MC iterations. We provide a convergence test for the results at the characterized level for all impacts and alternatives considered to show that this amount of MC runs is appropriate for this case study (SI, Appendix VI). Although other sources of uncertainty could be incorporated by means of various methods (Andrianandraina et al. 2015; Mendoza Beltran et al. 2016), we did not account for uncertainty due to methodological choices (such as allocation and impact assessment methods) or modeling uncertainties, neither due to data gaps that disallow the application of such methods. Also, correlations between input parameters was not accounted for (Groen and Heijungs 2017). In our experimental setup, the same technology and environmental matrix was used to calculate the results for the three alternatives for each MC run. Thus, dependent sampling underlies the calculations of paired samples. This experimental setup is important because it accounts for *common uncertainties* between alternatives (de Koning et al. 2010; Henriksson et al. 2015a) that are particularly important in the context of comparative LCAs (Henriksson et al. 2015a; Heijungs et al. 2017). Although the five statistical methods under study could be applied to independent sampled datasets, it would lack meaning as common uncertainties would then be disregarded. Thus, only dependently sampled MC runs were explored for the purpose of the present research. These MC runs per impact are available in the Microsoft Excel (2010) workbook in the SI.

The five USMs are applied to the same 1000 MC runs dependently sampled for each of the three alternatives and for each impact. As all methods are pairwise, we apply them for three pairs of alternatives: ICE/HFC, ICE/FBE, and FBE/HFC.

5.3 Results

Figure 15 shows the results for our comparative LCA following the classic visualization of deterministic characterization, in which results are directly superposed for comparison. All impacts considered are lower for the HFC except for depletion of abiotic resources. Both the ICE and FBE show various environmental trade-offs: the ICE performs worse than both the FBE and HFC in five impacts, while the FBE performs worse than the ICE and HFC in six impacts. Overall, the HFC performs better than both the FBE and ICE on most impacts considered. However, these results bear no information on their significance or likelihood, as no uncertainties are included.

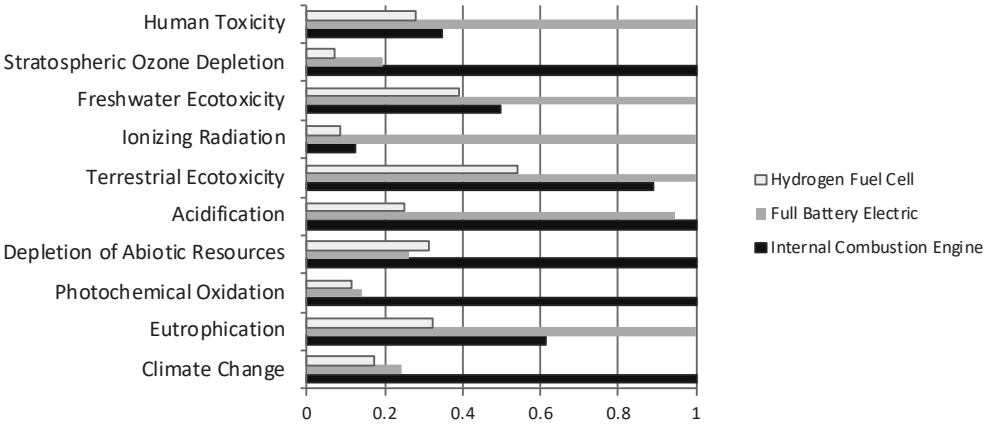

Figure 15. Deterministic results (scaled to the maximum results per impact) for comparative LCA of three alternatives of vehicles.

The complete set of results for the ten impacts considered and the five methods are found in the Microsoft Excel (2010) workbook in the SI. The *deterministic* LCA results shown in Table 12, correspond to those in Figure 15: HFC shows a better environmental performance than both the ICE and FBE for all impacts except for depletion of abiotic resources. In addition, Table 12 shows the results for the five statistical methods and for three selected impacts that display discrepant results.

For the *discernibility analysis,* and taking acidification as an example, the ICE and FBE vehicles have higher acidification results than HFC in 100% of the runs (Table 12, white cells under discernibility). Thus, the ICE and FBE are likely to be discernible alternatives from the HFC for acidification. For photochemical oxidation and acidification, there are pairs of alternatives that are not likely to be discernible as the percentage of runs in which one alternative is higher than the other is close to 50% (see Table 12 darker blue cells).

The *impact category relevance* results show the highest relevance parameter for acidification for the pairs ICE/HFC and FBE/HFC (Table 12, darker red cells). Thus, for the comparison between ICE, FBE and HFC vehicles, acidification is an impact that plays the most important role in the comparison. The lowest relevance parameter was obtained for the pair ICE/FBE for acidification as well as for the pair ICE/HFC for ionizing radiation these are impacts for which efforts to refine data would be most fruitful (Table 12, white cells under impact category relevance).

For the *overlap area,* the pair HFC/FBE has a large overlapping area for ionizing radiation and the pair FBE/ICE has a large overlap for acidification (Table 12, darker orange cells). Aspects contributing to the alternatives' performance in ionizing radiation and acidification would be areas to prioritize in data refinement. Other pairs have almost no overlapping area for instance HFC/ICE for photochemical oxidation and HFC/FBE for acidification (Table 12, white cells under overlap area). This means, that the choice

of an alternative between pairs, HFC/ICE and HFC/FBE, represents a greater effect on photochemical oxidation and acidification respectively.

The results for the *NHST* consist of the p-values for the paired t-test performed and the decision to reject (yes) or to fail to reject (no) the null hypothesis. This latter outcome has been included in Table 12. The p-values for all impacts and pairs of alternatives are < 0.0001, and thus the null hypothesis was rejected in all cases (See worksheet 'NHST' in the Microsoft Excel (2010) workbook in the SI). Therefore, results for all pairs of alternatives were significantly different for all impact categories (Table 12, purple cells). With the corrected significance level (α_b) we re-evaluated the null hypothesis but still rejected the null hypothesis in all comparisons.

For the modified NHST the comparison between the ICE and FBE for the acidification impact, cannot reject (no) the modified null hypothesis. Yet in the case of the NHST method it is rejected. Table 12 does not correspond to a mirror matrix for this method because the direction of the comparison matters. For acidification, we see that the pair FBE/ICE is not significantly different as well as the pair ICE/FBE. Thus, in both comparisons the scores of the first alternative are not at least δ_0 significantly higher than the scores of the second alternative. Therefore, the distance between the means of both alternatives is less than δ_0 i.e. 0.2 standard error units. With the corrected significance level (α_b) we re-evaluated the null hypothesis but found no changes in the outcomes.

Cross comparison of methods

Exploring the results across methods for the same impact shows consistent results for most impacts i.e. seven out of ten. A higher relevance parameter coincides with a smaller overlap area between distributions, and this generally coincides with well-discernible alternatives. Likewise, pairs of alternatives are more likely to have significantly different mean results when discernible. Below we focus our comparison of methods on three impacts (Table 12) that show discrepancies or conflicting results for some of the five methods.

For photochemical oxidation, the results for the five methods seem to agree to a large extent. Deterministic results show that HFC has the lowest characterized results among the three alternatives. However, according to the discernibility results, HFC is lower than FBE, for 83% of the runs. This shows that point-value results can be misleading, because there is a 17% likelihood that a point value would have given an opposite result. The overlap area results show a 0.63 overlap between the HFC and FBE on photochemical oxidation, indicating a mild difference (given the range of 0 to 1) in their performance. NHST and modified significance are in agreement with results from other methods and show significant different means for the two alternatives.

For acidification, results for some methods are consistent (Table 12). Discernibility of almost 100% along with a high relevance parameter and a low overlap

Method >	Deterministic LCA (point values)	Discernibility	Impact category relevance	Overlap area	NHST	Modified NHST
Meaning of result >	Does *j* have a lower impact than *k*?	% of total runs in which *j* has a lower impact than *k*	Which impact is important for the comparison of *j* and *k*?	Overlap area between distribution of impact of *j* and *k*	Are the mean impacts of *j* and *k* significantly different?	Is the mean impact of *j* at least 0.2 standard deviation units significantly lower than that of *k*?
	no / yes	0% / 50% / 100%	least / 0,24 / 6,68 / most	no overlap / 0,00 / 1,00 / full overlap	no / yes	no / yes
Impact						
Photochemical Oxidation	j↓ k→ ICE FBE HFC ICE — no no FBE yes — no HFC yes yes —	j↓ k→ ICE FBE HFC ICE — 0% 0% FBE 100% — 17% HFC 100% 83% —	j↓ k→ ICE FBE HFC ICE — 2,37 2,53 FBE 2,37 — 0,81 HFC 2,53 0,81 —	j↓ k→ ICE FBE HFC ICE — 0,01 0,00 FBE 0,01 — 0,63 HFC 0,00 0,63 —	j↓ k→ ICE FBE HFC ICE — yes yes FBE yes — yes HFC yes yes —	j↓ k→ ICE FBE HFC ICE — no no FBE yes — no HFC yes yes —
Acidification	j↓ k→ ICE FBE HFC ICE — no no FBE yes — no HFC yes yes —	j↓ k→ ICE FBE HFC ICE — 45% 0% FBE 54% — 0% HFC 100% 100% —	j↓ k→ ICE FBE HFC ICE — 0,24 5,44 FBE 0,24 — 6,68 HFC 5,44 6,68 —	j↓ k→ ICE FBE HFC ICE — 0,88 0,00 FBE 0,88 — 0,00 HFC 0,00 0,00 —	j↓ k→ ICE FBE HFC ICE — yes yes FBE yes — yes HFC yes yes —	j↓ k→ ICE FBE HFC ICE — no no FBE no — no HFC yes yes —
Ionizing Radiation	j↓ k→ ICE FBE HFC ICE — yes no FBE no — no HFC yes yes —	j↓ k→ ICE FBE HFC ICE — 100% 0% FBE 0% — 0% HFC 100% 100% —	j↓ k→ ICE FBE HFC ICE — 1,27 0,34 FBE 1,27 — 1,36 HFC 0,34 1,36 —	j↓ k→ ICE FBE HFC ICE — 0,13 0,79 FBE 0,13 — 0,08 HFC 0,79 0,08 —	j↓ k→ ICE FBE HFC ICE — yes yes FBE yes — yes HFC yes yes —	j↓ k→ ICE FBE HFC ICE — yes no FBE no — no HFC yes yes —

Table 12. Results for selected impacts (those with discrepant outcomes between some methods) for the comparative LCA of full battery electric (FBE) vehicle, the hydrogen fuel cell (HFC) vehicle and the internal combustion engine (ICE) vehicle. Tables display different results for the comparison of alternatives j and k for the reviewed uncertainty-statistics methods (USMs). The meaning of results per method is shown in the second row of the table together with the color labels.

area are shown for two pairs of alternatives HFC/ICE and HFC/FBE. Nonetheless, for the pair FBE/ICE discernibility results show a close call (FBE scoring only higher than ICE on acidification results for 45% of the runs) suggesting similar performances in acidification for FBE and ICE. This outcome is confirmed by the results of the impact category relevance (0.24), the overlap area (0.88) and the modified NHST where the null hypothesis is accepted and therefore no statistical difference can be established. NHST results, however, show a rejection of the null hypothesis that FBE and ICE have significantly different means for acidification, confirming that this pair of alternatives has significantly different acidification impacts – thus opposing the outcome of the other methods. As the sample size is large (namely 1000 observations), so is the likelihood of significance in NHST (Heijungs et al. 2016). The extra feature of the modified NHST compared to NHST is that the null hypothesis in the modified NHST is evaluated with a minimum size of the difference ($\delta_0 = 0.2$). It then appears that the difference in mean acidification results is so small that the null hypothesis cannot be rejected and that the mean acidification results for the FBE/ICE pair are not significantly different. The modified NHST results show how a large number of observations can influence the outcome of results in a standard NHST. Thereby it is possible to change the conclusion of a study by sampling more MC runs. Given that LCA uncertainty data is simulated and does not represent actual samples, it is recommended to apply the modified NSHT.

Finally, for ionizing radiation we observe a discrepancy between the discernibility, NHST, and modified NHST results on the one hand, and the impact category relevance and overlap area results on the other hand. The HFC/ICE pair shows a low relevance parameter (0.34) with a high overlap area (0.79). However, the discernibility results show that ICE scores higher than HFC on ionizing radiation for 100% of the runs. NHST and modified NHST confirm these results and show that, despite the large overlap and a low relevance parameter, the alternatives are significantly different. Note that the results of the relevance parameter and the overall area is to be used relative to other impact categories for sorting purposes– it is not intended to provide a confirmation on the difference. Still, results for this impact show that such high overlap can correspond to significant differences. Opposing outcomes are due to the overall or per run set-up of the methods. The discernibility analysis, NHST and modified NHST perform the analysis on a per run basis (accounting for common uncertainties) and evaluate, per run, whether the performances fulfill a certain relationship. Alternatively, the overlap area and the relevance parameter look at the overall distribution of the two alternatives rather than the individual runs. They take into account the *extent* of the difference so that the output falls within a spectrum, e.g. from 0 to 1 for overlap area, as opposed to a binary type output, e.g. fail to reject or reject the null hypothesis for NHST and modified NHST. Figure 16 shows the histogram for the distribution of HFC and ICE outcomes as well as the discernibility in a scattered plot, for better understanding the contradicting results between overlap area and discernibility. Here we can see that while

the histograms overlap a considerable amount, the performance between the alternatives can still be considered statistically different since all the runs fall within one side of the diagonal in the scattered plot, which disregards the distance of each point to the diagonal.

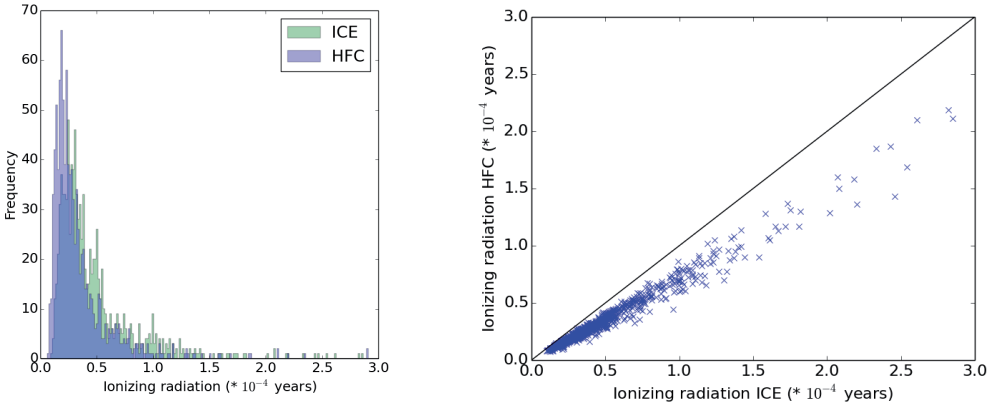

Figure 16. Histograms (left) and scatter plot (right) for 1000 MC runs for the hydrogen fuel cell (HFC) vehicle and the internal combustion engine (ICE) vehicle for ionizing radiation. The performances of ICE and HFC show great similarities in the histogram, and thus a large overlap area (i.e. 0.79). However, the scatter plot shows that for each MC run, the difference between HFC and ICE ≠ 0 (the diagonal line in the scattered plot represents equal values for both alternatives). Hence, alternatives are discernible in 100% of the runs.

5.4 Discussion and conclusions

We have reviewed, applied and compared different methods for uncertainty-statistics in comparative LCA. We showed how deterministic LCA can lead to oversimplified results that lack information on significance and likelihood, and that these results do not constitute a robust basis for decision-making. In addition, we found that, while in most instances (seven out of ten impacts), the five methods concur with each other, we identified instances where the methods produce conflicting results. Discrepancies are due to differences in the setup of the analysis (i.e. overall or per run) which accounts or not for common uncertainties and due to accounting or not for the magnitude of the differences in performances. We identify two groups of methods according to the type of analysis they entail: *exploratory* and *confirmatory* methods. This division corresponds with the statistical theories by Tukey (1973), in which data analysis initially requires an exploratory phase without probability theory, so without determining significance levels or confidence intervals, followed by a confirmatory phase determining the level of significance of the appearances identified in the exploratory phase. Exploratory statistics help delve into the results from uncertainty analysis and confirmatory methods evaluate hypotheses and identify environmental differences deemed statistically significant.

The NHST and modified NHST methods belong to the confirmatory group. Confirmatory methods are calculated per MC run, account for common uncertainties between alternatives and provide an absolute measure of statistical significance of the difference (Heijungs et al. 2017). These methods are appropriate for both single impact and multiple impact assessments and support statistical significance confirmation. NHST was shown to detect irrelevant differences of the means and to label them nevertheless as significant, while alternatives are considered to be indiscernible by modified NHST whenever the difference is small. The modified NHST approach is therefore recommended for confirmatory purposes and for all propagated LCA results, where the sample size in theory is indefinite and in practice is very large.

The impact category relevance and the overlap area methods belong to the exploratory group, as they help to identify some characteristics of uncertainty results among alternatives and impacts. These methods account for the magnitude of the difference per impact but do not consider common uncertainties or provide a measure of confidence or significance of the difference. These two methods are exclusively for exploring the uncertainty results in comparative LCAs with multiple impacts. Because the calculations are not per MC run, common uncertainties are disregarded and they do not serve confirmatory purposes. Disregarding common uncertainties can lead to instances where alternatives appear to be similar, while they actually perform different (like in ionizing radiation between ICE and HFC, Figure 16). Overcoming the fact that they do not account for common uncertainties would require generalization of the methods to "per run" calculations and could lead to a method similar to modified NHST accounting for the distance between means and common uncertainties.

Discernibility belongs to both groups. It accounts for common uncertainties, but it does not account for the magnitude of the difference per impact. It can be complimented with a *p*-value calculation, to develop its confirmatory potential, that would generate statistical significance based on the counts of the sign tests per pair. A proposal for such a procedure can be found in the SI (appendix V) and involves the use of the binomial distribution. As it stands now, we consider it to serve an exploratory purpose similar to the impact category relevance or the overlap area, but with a different mechanism.

Both exploratory and confirmatory methods are valuable and synergistic in data-driven research (Tukey 1980), yet the specific choice of method is not straightforward for LCA practitioners given the discrepancies and characteristics previously discussed. Figure 17 provides guidance on which statistical methods LCA practitioners should use to interpret the results of a comparative LCA in light of its goal and scope, and when considering uncertainty. Figure 17 is in line with the main findings of this chapter. That is, exploratory methods facilitate the decision-making process by identifying differences and trade-offs in impacts between alternatives as well as by pointing to places where data refinement could benefit the assessment. Moreover, confirmatory methods effectively

aid in making complex decisions from comparative assessments but should be used with statistical significance. For instance, carbon footprints, product environmental declarations, and LCAs aiming for comparative assertions disclosed to the public, should use confirmatory methods supporting conclusions with statistical significance calculations and accounting for common uncertainties.

Moreover, modified NHST appears to be the most well-developed method for confirmatory purposes. For exploratory purposes, however, we do not find a method that considers both core aspects: accounting for common uncertainties and for the extent of the differences per impact. Between these two aspects, common uncertainties are the most crucial aspect to address in a comparative context. Therefore, we recommend discernibility as the most suitable method for exploratory purposes while recognizing areas for improvement. Namely, we recommend that discernibility is further developed by adding a threshold of acceptable difference (as done in modified NHST) that, despite of being arbitrary, can better inform the exploration of trade-offs. We also recommend practitioners to exercise caution when applying overlap area and impact category relevance, and we recommend further developments of both methods to account for common uncertainties. Lastly, we call for caution when applying NHST regarding the sample size as it has been conceived for real samples (Henriksson et al. 2015a) and not for propagating uncertainty estimates where the sample size is in theory indefinite.
We encourage practitioners to use the excel workbook provided in the SI with the calculations made for the five methods in this paper which can aid them in delivering a more robust basis for decision-making.

As the use of statistical methods is becoming more frequent and increasingly important in environmental decision support (Hellweg and Canals 2014), the definition of thresholds to determine the acceptable uncertainty demands attention. Arbitrarily set thresholds, such as p-value = 0.05, should be carefully used accounting for basic principles addressing misinterpretation and misuse of the p-value, as recently proposed by the American Statistical Association (Wassertein and Lazar 2016). In the field of LCA, we need practical guidelines to establish meaningful uncertainty thresholds for different applications. Methods like modified NHST and extended discernibility (see appendix V), require such threshold levels to calculate statistical significance. We depart from the premise that various sources of uncertainties of the inputs have been adequately quantified and propagated to uncertainty results. The effects of the quality of uncertainty quantification and propagation on the interpretation of uncertainty results in comparative LCAs requires further study (Mila i Canals et al. 2011). Any outcome of any test is only as good as the quality of the input data, which for all studied methods corresponds to the results of an uncertainty analysis.

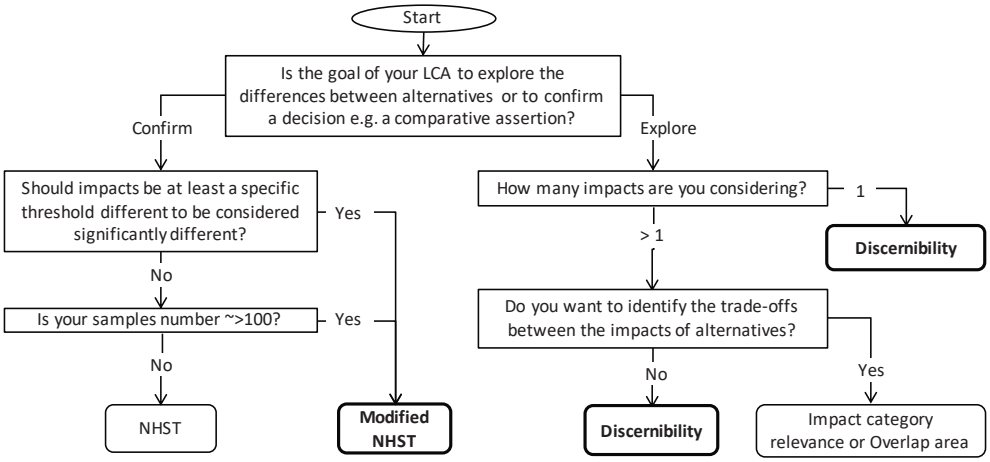

Figure 17. Decision tree to guide LCA practitioners on which uncertainty-statistics method (USM) to use for the interpretation of propagated LCA uncertainty outcomes in comparative LCAs. Thicker lines indicate recommended methods for confirmatory and exploratory purposes as per the considerations described in the main text. The type of information available from the uncertainty analysis results (in the following parenthesis) determines the choice between impact category relevance (statistical parameters of the distributions) or overlap area (MC runs).

Acknowledgements

Authors would like to acknowledge the ISIE Americas 2016 conference that took place in Bogota, Colombia, for providing the environment to shape the ideas further developed in this research. We also thank Sebastiaan Deetman and Sidney Niccolson for their insightful comments on visualizations.

Supporting information

Supporting information of this chapter may be found in the online version of the original article: https://pubs.acs.org/doi/abs/10.1021/acs.est.7b06365