Int J Life Cycle Assess (2016) 21:1185–1196
DOI 10.1007/s11367-016-1098-5

UNCERTAINTIES IN LCA

Giving a scientific basis for uncertainty factors used in global life cycle inventory databases: an algorithm to update factors using new information

Stéphanie Muller[1] · **Pascal Lesage**[1] · **Réjean Samson**[1]

Received: 12 November 2015 / Accepted: 14 March 2016 / Published online: 30 March 2016

Abstract
Purpose Life cycle inventory (LCI) databases provide generic data on exchange values associated with unit processes. The "ecoinvent" LCI database estimates the uncertainty of all exchange values through the application of the so-called pedigree approach. In the first release of the database, the used uncertainty factors were based on experts' judgments. In 2013, Ciroth et al. derived empirically based factors. These, however, assumed that the same uncertainty factors could be used for all industrial sectors and fell short of providing basic uncertainty factors. The work presented here aims to overcome these limitations.
Methods The proposed methodological framework is based on the assessment of more than 60 data sources (23,200 data points) and the use of Bayesian inference. Using Bayesian inference allows an update of uncertainty factors by systematically combining experts' judgments and other information we already have about the uncertainty factors with new data.
Results and discussion The implementation of the methodology over the data sources results in the definition of new uncertainty factors for all additional uncertainty indicators and for some specific industrial sectors. It also results in the definition of some basic uncertainty factors. In general, the factors obtained are higher than the ones obtained in previous work, which suggests that the experts had initially underestimated uncertainty. Furthermore, the presented methodology can be applied to update uncertainty factors as new data become available.
Conclusions In practice, these uncertainty factors can systematically be incorporated in LCI databases as estimates of exchange value uncertainty where more formal uncertainty information is not available. The use of Bayesian inference is applied here to update uncertainty factors but can also be used in other life cycle assessment developments in order to improve experts' judgments or to update parameter values when new data can be accessed.

Keywords Bayesian statistics · Ecoinvent database · Life cycle inventory · Pedigree approach · Uncertainty

1 Introduction

1.1 Uncertainty modeling in the ecoinvent database

Even if life cycle assessment (LCA) results are often presented as single-point values, it is well known that these results have uncertainty. Uncertainty can be divided into three classes: model, scenario, and parameter uncertainty (Lloyd and Ries 2007; Reap et al. 2008), whereas model and scenario uncertainties that arise during a LCA can be assessed through, for example, sensitivity analysis; the effect of parameter uncertainty on the final results can be assessed through uncertainty propagation methods (by numerical methods as Monte Carlo simulation or by analytical methods) (Heijungs and Lenzen 2014; Hong et al. 2010; Imbeault-Tétreault et al. 2013). These uncertainty propagation techniques can however only be used

Responsible editor: Roland Hischier

Electronic supplementary material The online version of this article (doi:10.1007/s11367-016-1098-5) contains supplementary material, which is available to authorized users.

✉ Stéphanie Muller
stephanie.muller@polymtl.ca

[1] CIRAIG, Department of Chemical Engineering, Polytechnique Montreal, P.O. Box 6079, Stn. Centre-Ville, Montréal, Québec H3C 3A7, Canada

if the uncertainty associated to the value of exchanges (elementary or intermediate flows) are quantified during the inventory stage of an LCA.

The ecoinvent database is the only LCI database that systematically includes uncertainty on data modeled in the database. It uses a semi-quantitative approach to estimate exchange value uncertainty. This approach is often referred to as the pedigree approach because it is based on the use of a pedigree matrix inspired by the NUSAP system (Funtowicz and Ravetz 1990). It was originally developed for LCA by Weidema and Wesnæs (1996) and has been used in the ecoinvent database since 2005 (Frischknecht et al. 2005).

For a detailed description of the pedigree approach and the way it is applied in the ecoinvent database, the reader can refer to several papers (Ciroth et al. 2013; Frischknecht et al. 2005; Muller et al. 2014). In short, the approach ascribes a *basic uncertainty* value to exchanges, representing the intrinsic variability and stochastic error of the parameter and increases this uncertainty using *additional uncertainty* factors that represent the use of imperfect data for the context of the study. By default, exchange values are assumed to be lognormally distributed, with the geometric mean defined as the deterministic value for the exchange and the geometric standard deviation estimated based on the basic and additional uncertainty factors that the pedigree approach provides. Distributions other than the lognormal can now also be used with the pedigree approach, see Muller et al. (2014).

If sufficient information can be accessed to calculate descriptive statistics and thus define the uncertainty, it should be used. If this is not the case, ecoinvent proposes default basic uncertainty values, expressed as GSD^2 and classified by type of exchange values and for three sectors: agriculture, combustion, and process.

The additional uncertainty factors, also by default expressed as GSD^2, are derived from a scoring of the data quality on five characteristics:

- *Reliability of the data source*, scoring the quality of the sources and acquisition methods of the data used to quantify the exchange
- *Completeness*, scoring the statistical representativeness of the data
- *Temporal*, *geographical*, *further technological correlation*, scoring the degree to which the data used is representative of the time, area, and technology of interest

Using the cells descriptions of the pedigree matrix, scores of 1 to 5 (where 5 is the worst score) are given to each exchange value. The pedigree matrix in use in the ecoinvent v3 database is presented in the Electronic Supplementary Material. These scores are then converted into so-called additional uncertainty factors, expressed as "contributors to the GSD^2" (Frischknecht et al. 2005; Muller et al. 2014). The basic uncertainty and the additional uncertainty factors are then compiled, leading to a measure of the total uncertainty expressed as GSD^2.

1.2 The need to develop new uncertainty factors

This semi-quantitative pedigree approach is convenient to model the uncertainty on a large number of exchanges where real uncertainty information is not available. It however has several limitations. Three important limitations are (1) the imposition of the lognormal to describe the uncertainty of exchange values, (2) the reliance on experts' judgments rather than on empirical data to quantify uncertainty factors, and (3) the use of additional uncertainty factors that ignore the type of exchange or industrial sector being assessed (Henriksson et al. 2014).

The first cited limitation was addressed by Muller et al. (2014), who extended the pedigree matrix approach to other types of distributions. The second limitation was partially addressed by Ciroth et al. (2013) who developed new additional uncertainty factors based on a statistical assessment of seven different data sources. The uncertainty factors they developed were slightly different from the ones used in ecoinvent v2. However, they did not address the case of the basic uncertainty; they assumed that the factors are valid for all industrial sectors and did not calculate factors for all "pedigree indicator-pedigree score" couples.

If, in theory, perfect uncertainty factors should be derived from empirical data only (i.e., data that are based on experiences, experiments, and observations), such data are not, in practice, widely available in LCI. New uncertainty factors will also be derived based on known data and information.

1.3 A way to update uncertainty information: the application of the Bayes theorem

Published uncertainty factors associated with the pedigree matrix approach were either based on experts' judgments (Frischknecht et al. 2005; Weidema and Wesnæs 1996) or on the assessment of a limited number of data sources (Ciroth et al. 2013). While useful as first estimates of basic or additional uncertainty of exchanges, the scientific basis for these factors can be improved using more extensive data sources. A way to update quantitative information using new data sources is to use Bayesian inference, coming from Bayes theorem as presented in a simple form in Eq. (1) (Qian et al. 2003).

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)} \quad (1)$$

with θ the random variable representing our variables of interest (in this case, the uncertainty factors) and d the vector

containing the new information. The elements of Eq. (1) can be interpreted as follows:

- $p(d|\theta)$ is the likelihood function: it describes the assumption that the data $\boldsymbol{d}$ were observed based on θ.
- $p(\theta)$ is the prior distribution: it represents the knowledge (often subjective) that is available on θ.
- $p(\theta|d)$ is the posterior distribution: it represents all the information we finally have on θ. The distribution's mean can become an estimate for θ and its confidence interval an estimate on θ's uncertainty.
- $p(d)$ is the partition function. This denominator can be considered as a standardization coefficient allowing the posterior probability distribution to take its value in the interval [0,1]. The Bayesian theorem can therefore be rewritten as $p(\theta|d) \propto p(d|\theta)p(\theta)$ or, in words, "the final information (*the posterior distribution*) is proportional to the prior belief (*the prior distribution*) modified by the observation (*the likelihood function*).

While the use of Bayesian inference in LCA is mentioned in several papers (Björklund 2002; Huijbregts 1998; Katz 2002), only few describe concrete applications. Ukidwe et al. (2004) and Miller et al. (2013) used the Bayesian approach to refine their LCI models; Lo et al. (2005) used it to better estimate the uncertainty of their systems by first identifying the greater contributors to the uncertainty and then applying the Bayes theorem to these greater contributors to refine their associated uncertainty.

The aim of this paper is to strengthen the pedigree approach by refining existing uncertainty factors through the consideration of new data sources, developing uncertainty factors that are specific to industrial sectors, and defining new basic uncertainty factors. In order to do so, Bayesian inference will be used to both combine information already known—by experts' judgments and data assessment—with new information obtained through the assessment of a large data set.

2 Methods

2.1 General methodology

Table 1 is a summary of the three-step methodology developed to obtain updated basic and additional uncertainty factors. These three steps are detailed in the following paragraphs.

First of all, a descriptor of the uncertainty factors should be chosen. Weidema and Wesnæs (1996) used the coefficient of variation (CV, ratio between the standard deviation and the mean) to describe the uncertainty factors. In the ecoinvent v2 database, the GSD^2 is used as a descriptor of the uncertainty factors; more specifically, the uncertainty factors are described as "contributors to the GSD^2" (Frischknecht et al. 2005). Then, in the third version of the ecoinvent database, they were expressed as "square of the standard deviation of the underlying normal distribution," i.e., the square of the standard deviations of the variable's logarithm (Weidema et al. 2013). In this paper, to remain consistent with the previous work done by Ciroth et al. (2013), the GSD^2 is chosen as a descriptor of the uncertainty factors. The GDS^2 can easily be transformed in the descriptor used in ecoinvent v3 (σ^2) using Eq. (2).

$$\sigma^2 = (\ln(GSD))^2 \tag{2}$$

2.2 Identify initial estimate of uncertainty factors

The basic and additional uncertainty factors used in ecoinvent v2 and v3 and as published in ecoinvent v3 data quality guidelines (Weidema et al. 2013) are coming from experts' judgments and expressed as the "square of the standard deviation of the underlying normal distribution" (σ^2). To express as "contributors to the GSD^2," the reverse transformation of the one performed between ecoinvent v2 and ecoinvent v3 can be applied. Equation (3) can so be used and the uncertainty factors (UF) can then be expressed by Eq. (4).

$$GSD = \exp(\sigma) \tag{3}$$

$$UF = (\exp(\sigma))^2 \tag{4}$$

where UF is the value of the uncertainty factor.

When the pedigree matrix was first developed in 1996, Weidema and Wesnaes also proposed uncertainty figures for the additional uncertainty based on estimates and expressed as CVs. These estimates are also used as prior information once translated into GSD^2. In order to perform this translation Eq. (5) can be used. This use is based on the following:

- Uncertainty factors are expressed as contributors to the GSD^2
- The pedigree approach was developed for lognormally distributed exchange values
- Equation (5) links the GSD and the CV for a lognormal distribution

$$GSD = \exp\left(\sqrt{\ln\left(CV^2 + 1\right)}\right) \tag{5}$$

Finally, the values developed by Ciroth et al. for the additional uncertainty, both in a published paper (2013) and in a report delivered to the ecoinvent Centre (2012), will also be used as prior information for this work.

These four data sources for prior information lead to one datum for each component of the basic uncertainty and to 2 to

Table 1 The three steps methodology used to develop updated uncertainty information

1. Identify initial estimate of uncertainty factors	1.1 Census of formerly published uncertainty figures
	1.2 Translation of these figures into uncertainty factors
2. Collect new information	2.1 Data harvesting and preparation
	-Data source selection
	-Data harvesting
	-Data classification into subgroups that isolate basic uncertainty and the different components of additional uncertainty
	2.2 Estimation of data-based uncertainty factors
	-GSD^2 calculations for each pedigree indicator, for specific industrial sectors and for the subgroups created for the basic uncertainty
	-GSD^2 transformation to "contributors to the GSD^2" (i.e., uncertainty factors)
3. Update uncertainty information	3.1 Estimation of the *prior distribution* based on the formerly published uncertainty factors
	3.2 Estimation of the *likelihood function* based on the obtained data-based uncertainty factors
	3.3 Application of Bayes theorem to obtain new uncertainty factors that combine expert judgements and data

26 for each component of the additional uncertainty. These data can be found in the Electronic Supplementary Material. In order to remain consistent, one outlier from the work of Ciroth et al. (2012) was removed from the analysis.

2.3 Collecting new information

2.3.1 Data harvesting and preparation

The selection of the data sources used in this assessment is based on the following points:

- The typologies of data that arise in LCI must be considered: the harvested data must represent both intermediate and elementary flows (together referred to as exchanges)
- It must be possible to normalize data to a reference flow, i.e., a unit output from an activity
- The data sources must cover a large spectrum of industrial sectors, years, geographical areas, and types of sources to have a sufficient set of data for each assessed "pedigree indicator-pedigree score" couple

Based on these points, data from publicly available LCA reports, published LCA papers, emissions factors databases, and sector-specific LCI databases were collected. The 68 different data sources used, listed in the Electronic Supplementary Material, yielded 23,200 data points for analysis.

These data are compiled in a single database that contains all the information needed to perform the assessment:

- The type and the name of the datum, its value, and its corresponding unit.
- The reference flow to which the datum is normalized, its value, and its corresponding unit.
- The industrial sector from which the data was generated. The industrial sectors classification used is based on the 2012 North American Industry Classification System (NAICS) and on the 1997 Selected Nomenclature for sources of Air Pollution (SNAP97) for the combustion sector.
- The other information useful to classify the data into a specific "pedigree indicator-pedigree sector" couple: the year and the geographic area where the data were generated and how it was generated (e.g., estimated or calculated).

The compiled data have, at this point, different units and are linked to different reference flows. All data were converted to SI units and normalized to a unit amount of reference flow.

Finally, this compilation into a single database allows the classification of the different data into specific subsets for the basic uncertainty and the different "pedigree indicator-pedigree score" couples for different industrial sectors (based on the NAICS—level 2) for the additional uncertainty. The basic uncertainty is defined as the uncertainty that remains when all pedigree scores equal 1; in each created subgroup, the data are coming from a same technology, a same year, and a same country, and their reliability score equals 1. The data classification for the additional uncertainty is based on the pedigree matrix cell description (as described in Weidema et al. (2013)). The way the pedigree matrix's cells were interpreted to create the different subgroups is available in the Electronic Supplementary Material.

In practical considerations, for all pedigree indicators (except for the completeness indicator which case is described in Section 2.3.3), the creation of subgroups is based on what is

named here the "Russian dolls" principle, as described in Fig. 1 and in Ciroth et al. (2013) (see Ciroth et al. (2013) for a description of the principle using the temporal correlation indicator as an example). This principle is here applied to remain consistent with the work done by Ciroth et al. (2013). Moreover, this principle of data classification into different subgroups followed a majority of the pedigree matrix cells' descriptions (especially for the completeness, the temporal correlation, and the further technological correlation indicators). In order to remain consistent in the whole methodology, this same principle is also followed for the reliability and the geographical correlation indicators. If the pedigree matrix cells' descriptions are literally followed for these two indicators, all the subgroups created (from a score 1 to a score 5) will totally be independent, leading to GSD possibly greater in the subgroup for a score 1 than the GSD in subgroups for other score (see Eq. (6)). The approach used in this paper allows us to overcome this possible problem.

2.3.2 Estimation of uncertainty factors derived from the data

The estimation of uncertainty factors derived from the data are based on the formulas proposed by Ciroth et al. (2013). In the first place, the GSD^2 for each created subgroup is calculated based on Eq. (6). For the additional uncertainty, these GSD^2 must be converted into uncertainty factors using Eq. (7). The basic uncertainty factor is directly defined as the calculated GSD^2.

$$GSD = \exp\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{x_i}{\overline{x_g}}\right)^2}\right) \quad (6)$$

,where n is the number of data in the sample and $\overline{x}_g$ the geometric mean of the sample.

$$UF_i = \begin{cases} \dfrac{GSD_{i,j}^2}{GSD_{i,1}^2} & if \quad GSD_{i,1}^2 \le GSD_{i,j}^2 \\ \dfrac{GSD_{i,1}^2}{GSD_{i,j}^2} & else \end{cases} \quad (7)$$

Fig. 1 Illustration of the "Russian dolls" principle used to classify data into subgroups for the additional uncertainty

where UF_i is the uncertainty factor for the ith pedigree indicator;

$GSD_{i,1}^2$ is the geometric standard deviation for the ith pedigree indicator and the score 1

$GSD_{i,j}^2$ is the geometric standard deviation for the ith pedigree indicator and a score j.

Example 1: how the uncertainty factors are calculated—case of the temporal correlation indicator

Following the Russian dolls principle illustrated in Fig. 1, subgroups are created based on the industrial sector (based on two-digit NAICS codes) and the type of exchange (e.g., specific emission, specific type of intermediary input). For example, all the CO_2 emissions for the primary metal manufacturing sector were classified into a specific subgroup.

In this subgroup, the GSD^2 of all data having a pedigree score of 1 is calculated (using a reference year of 2013). The same is done for the data having scores from 2 to 5. Then, Eq. (7) is used to calculate the uncertainty factors for CO_2 emissions of the primary metal manufacturing sector.

2.3.3 The completeness indicator case

For Weidema and Wesnæs (1996), the completeness indicator expresses the variability due to "the number of data collection points, the period of collection and the representativeness to the total population"; in ecoinvent v3, the completeness indicator expresses the representativeness of the data regarding the relevant sites of the considered market (Weidema et al. 2013). This indicator can also be directly linked to the representativeness of the used sample to describe the entire population (and so here, to produce a representative LCI datum). The question of the effect of sampling on the variability can be answered statistically by building a confidence interval for the population variance knowing the size of the sample regarding the whole population and under the assumption that the size of the entire population is known (as it is done for opinion surveys for example).

Confidence interval estimation for a population is well known when the population is assumed to be normally distributed. For normally distributed data, the ratio $n.s^2/\sigma^2$ (where n is the size of the population, s the sample standard deviation, and σ the standard deviation of the population) follows a chi-square distribution with $n-1$ degrees of freedom. For a 95 % confidence level, the confidence interval estimation for the population variance can be obtained through Eq. (8).

$$\frac{ns^2}{\chi^2_{\alpha_1}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{1-\alpha_1}} \quad (8)$$

As before, Eq. (3) can be used to transform σ^2 into GSD^2. Equation (7) is then used to derive uncertainty factors for the completeness indicator. In Eq. (8), $\chi^2{}_{\alpha 1}$ and $\chi^2{}_{1-\alpha 1}$ can be found in the table of the χ^2 distribution; α_1 is linked to the

confidence level that need to be reached; for a symmetric confidence interval estimation and a confidence level of 95 %, α_1 is then equal to 0.025.

Concerning the value of s, it can be directly linked to the basic uncertainty as it represents the value of the sample standard deviation that can be calculated by the person that defines the LCI datum with its uncertainty. For the rest of the paper, it will be assumed that s^2 equals 1. It must be here underlined that the value can be changed by the user if necessary.

The factors developed through this method will not directly fit the pedigree matrix cell descriptions for the completeness indicator but the ones in Table 2. Supposing that the whole population is represented by 100 sites, a perfect data set is obtained when the 100 sites are represented, so $n = 100$ for a score 1. For the score 2, 75 % of the sites must be represented, so $n = 75$ (respectively, 50 and 25 % of the sites lead to $n = 50$ and $n = 25$ for scores 3 and 4). The score 5 is linked to data for which the completeness is unknown, so in theory, n should be equal to 0, but in practice, $n = 0$ leads to an infinite uncertainty factor. In order to avoid this situation of an infinite uncertainty factor, n is here fixed to 5.

2.4 Updating uncertainty information

Theoretically, the data used to derive new uncertainty factors that can be applied to all LCI exchanges should cover all types of exchanges and all types of industrial sectors. In reality, the available sample is not representative of the whole technosphere. Rather than using only partial information to define new uncertainty factors, the initial estimates of uncertainty factors will be updated using the new, quantitative but partial information and the Bayes theorem as described in Eq. (1). Before applying the theorem, the parameter of interest (i.e., the information that will be refined thanks to the theorem) needs to be defined. Here, this parameter is the *UF*. Then, and in order to apply the Bayes theorem for basic uncertainty and each "pedigree indicator-pedigree score" couple, three more components are needed:

- The prior distribution of *UF*:*UF* needs to be defined as a random variable whose prior distribution represents all the information we have on the uncertainty factor. Defining *UF* as a random variable means that variability on the uncertainty factor itself is introduced.
- The data *d*: d is here the vector containing the new data-based estimation of the uncertainty factors based on the newly assessed data sources.
- The likelihood function *f*:*f* links the data to the parameter of interest. It is here assumed that the data can be described by *f(UF)* where the random variable *UF* is the mean of the likelihood function.

The application of the Bayes theorem depends on the nature of the prior distribution and the likelihood function. If the numerator in Eq. (1) can always be calculated analytically, it's not the same with the denominator that needs to be determined numerically using Bayesian Monte Carlo or Markov chain Monte Carlo when the prior distribution and the likelihood function are not conjugate (Ben Letham 2012; Qian et al. 2003).

In this application of the Bayes theorem, the information we have on the prior distribution of the uncertainty factors is very limited. In order to define the prior distribution, and also the likelihood function, the distribution of the newly obtained uncertainty factors is plotted. When graphically representing the distribution of the new obtained uncertainty factors for the basic uncertainty and all "pedigree indicator—pedigree score" couples, two observations can be noted:

- All the factors are positive (by definition).
- For a majority of the factors, their distributions are right-tailed.

Given these two characteristics, the lognormal distribution is chosen to represent both the prior distribution and the likelihood function. This assumption of lognormally distributed uncertainty factors is tested on the new data-based uncertainty factors (see Electronic Supplementary Material). A lognormal distribution is not directly conjugate to another lognormal distribution. However, two normal distributions are, and they result in a normally distributed posterior distribution. In order to perform the Bayesian application, we use the fact that, if X is a lognormally distributed random variable with parameters μ and σ, then $ln(X)$ is normally distributed with the same parameters (see Eqs. (9) and (10)). Figure 2 makes a census

Table 2 Representative data

1	2	3	4	5
Representative data from all sites relevant from the market considered.	Representative data from >75 % of the sites relevant from the market considered.	Representative data from >50 % of the sites relevant from the market considered.	Representative data from >25 % of the sites relevant from the market considered.	Unknown representativeness or representative data from less than 25 % of the sites relevant from the market considered

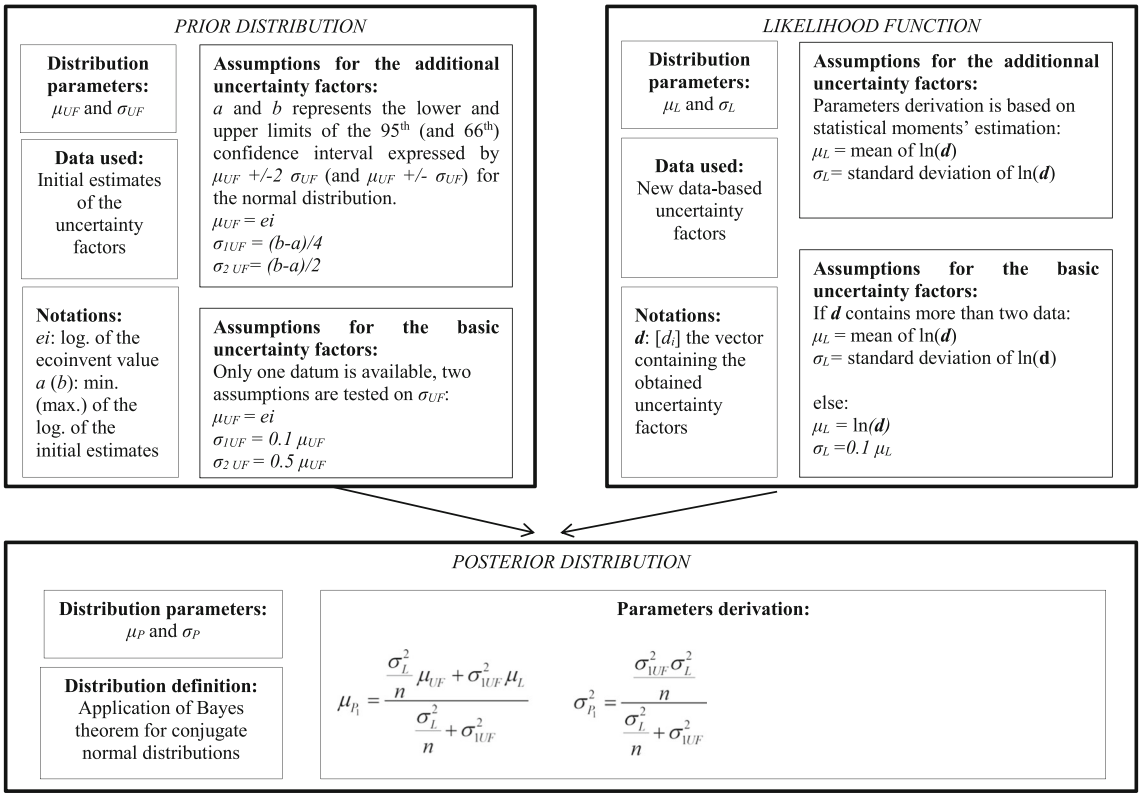

Fig. 2 Assumptions used to apply the Bayes theorem—*Notations log. (logarithm), min. (minimum), max. (maximum)*

of all the assumptions needed to derive the prior distribution and the likelihood function, as well as the calculation performed to obtain the posterior distribution. The most subjective assumptions are made on the parameters of the prior distribution due to the lack of available information. In order to derive these parameters, the mean of the prior distribution is set as the value of the uncertainty factor as found in ecoinvent v3. These uncertainty factors are the ones that need to be updated as they are the ones currently in use. We therefore supposed here that they are "the best" available prior information. The definition of the standard deviation for the prior distribution is based on the 95th interval if sufficient data are available (see Fig. 2).

$$LN(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \quad (9)$$

and

$$N(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (10)$$

,where LN(x,μ,σ) stands for the lognormal PDF and N(x,μ,σ) for the normal PDF.

Once the parameters of the posterior distribution have been derived, it is possible to determine the updated uncertainty factors. The updated uncertainty factor is defined as the mean of the posterior distribution that can be defined using Eq. (11) (which is the equation linking the mean of a lognormal distribution to the logarithmic parameters that defines the distribution).

$$UF_P = \exp\left(\mu_P + 0.5\sigma_P^2\right) \quad (11)$$

Example 2: how the information is updated—case of the temporal correlation indicator.

Take the temporal correlation indicator for a score 2. The uncertainty factor for this case is noted UF_{TC2}. To update the information, the prior distribution of UF_{TC2} and the likelihood function need to be defined.

For the prior distribution, all the uncertainty factors published in both previous papers and ecoinvent report (Ciroth et al. 2012, 2013; Weidema et al. 2013; Weidema and Wesnæs 1996) are used to define the distribution based on the assumptions presented in Fig. 2 and in paragraph 2.4.

The likelihood function is obtained by calculating the logarithmic mean and the logarithmic standard deviation of all uncertainty factors calculated by subgroups for a score 2 (uncertainty factors obtained in the case of the CO_2 emissions for the primary metal manufacturing sector are among them). This logarithmic mean and this logarithmic standard deviation define the lognormal distribution used as likelihood function.

3 Results

3.1 The case of the additional uncertainty factors

The results obtained for the additional uncertainty factors are presented in Fig. 3 and Table 3. Figure 3 shows the values of the generic factor for all "pedigree indicator-pedigree score" couples. These generic values can also be found in Table 3 that adds the additional uncertainty factors obtained by industrial sectors. Generally, the likelihood values are greater than the prior values. That can be explained by the large variability present in the subgroups of assessed data. Even if these data are grouped according to specific industrial sectors, these sectors remain large (for example, the manufacturing sector contains both cement production and pulp and paper production), explaining large variability and large additional uncertainty factors especially for pedigree scores equal to 5. The fact that the posterior values lay between the prior and the likelihood can be explained by the application of the Bayes theorem that gives a weighted compromise between the prior knowledge and the data.

Some of the calculated uncertainty factors are counter-intuitive. First, for some pedigree indicators, the likelihood value is the same for two consecutive pedigree scores (see for example the scores 4 and 5 for the geographical correlation indicator in the manufacturing sectors). These cases arise when the same data constitute the subgroup for each score, due to a lack of data. One exception is the equality of scores 4 and 5 for the generic factor in the further technological indicator, which is not a real equality and comes from the expression of the uncertainty factor using only three digits. Second, some uncertainty factors are greater for a score 4 than for a score 5 (see for example the temporal correlation indicator for the transportation sector or the geographical correlation indicator for the combustion and agricultural sectors. This arises when a subgroup with more data (for a score 5) is less variable than the subgroup with a score 4. More precisely, the subgroup for a score 5 (noted here S_5) contains all the data of the equivalent subgroup for a score 4 (here noted S_4) and some other data points. In certain cases, the variability in S_5 is smaller than the variability in S_4; this is due to the definition of the GSD (see Eq. (6)). If, in general, individual data points in S_5

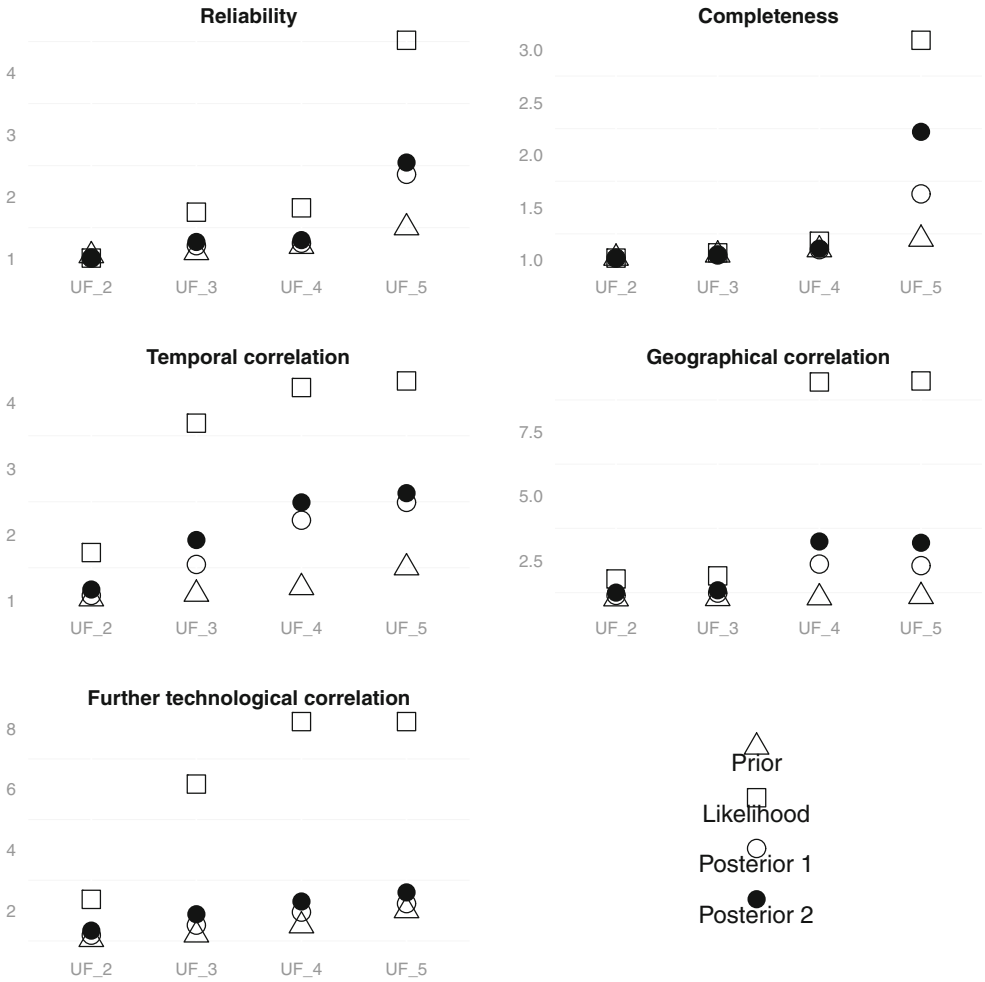

Fig. 3 Additional uncertainty factors representation for the five pedigree indicators and the four pedigree scores. The value of Posterior 1 is obtained using σ_{1UF} for the prior distribution, the value of Posterior 2 is obtained using σ_{2UF} (see Fig. 2)

Table 3 Updated additional uncertainty factors obtained by using σ_{1UF} for the prior distribution (see Fig. 2), the results for σ_{2UF} can be found in the aupporting information

Pedigree indicator		Prior	Generic		Agriculture		Combustion		Utilities		Manufacturing (other)		Chemical products manufacturing		Metal manufacturing		Transportation	
			L	*P*	*L*	*P*	*L*	*P*	*L*	*P*	*L*	*P*	*L*	*P*	*L*	*P*	*L*	*P*
Reliability	UF2	1.05	1.01	1.01	1.00	1.00	1.06	1.06			1.00	1.00			1.00	1.00		
	UF3	1.10	1.75	1.21	1.05	1.05	1.18	1.12			2.26	1.20			1.00	1.00		
	UF4	1.20	1.82	1.25	1.05	1.07	1.18	1.18			2.26	1.25			1.40	1.21		
	UF5	1.50	4.52	2.36	1.98	1.60	4.61	1.69			6.20	2.60			1.64	1.51		
Completeness	UF2	1.02	1.02	1.02														
	UF3	1.05	1.07	1.05														
	UF4	1.10	1.18	1.10														
	UF5	1.20	3.09	1.63														
Temporal correlation	UF2	1.03	1.73	1.08	3.41	1.03	1.47	1.08	1.00	1.00	3.58	1.05			1.18	1.07	1.20	1.16
	UF3	1.10	3.69	1.55	3.41	1.13	2.33	1.27	5.66	1.22	7.60	1.38			1.44	1.22	1.32	1.26
	UF4	1.20	4.23	2.22	5.03	1.80	2.68	1.72	9.52	1.87	6.34	2.12			1.44	1.32	1.27	1.26
	UF5	1.50	4.33	2.49	5.03	4.07	3.12	1.75	9.53	2.52	6.12	2.47			1.58	1.49	1.15	1.15
Geographical correlation	UF2	1.01	1.78	1.14	1.00	1.00	1.29	1.08			3.12	1.09						
	UF3	1.02	1.90	1.23	1.72	1.10	1.29	1.11			3.12	1.16						
	UF4	1.05	9.45	2.36	4.14	1.57	4.73	1.66			23.60	1.98						
	UF5	1.10	9.49	2.30	4.38	1.55	4.73	1.65			23.60	1.92						
Further technological correlation	UF2	1.05	2.37	1.19	1.00	1.00	1.04	1.04	1.00	1.00			8.58	1.16				
	UF3	1.20	6.17	1.52	3.33	1.32	1.03	1.04	4.62	1.29			14.47	1.44				
	UF4	1.50	8.23	1.95	3.73	1.59	1.54	1.50	8.13	1.62			14.69	2.04				
	UF5	2.00	8.23	2.23	3.72	2.03	1.54	1.89	3.03	2.00			35.02	2.21				

The prior is the ecoinvent value; L stands for the likelihood value and P for the posterior value

are closer to the geometric mean of S_5 than the data points in S_4 are from the geometric mean of S_4, then the GSD for S_5 will be smaller than the GSD of S_4. Ideally, with a better access to more data, this situation would not arise. In this specific case (where the factor for a score 4 is greater than the one for a score 5), it is recommended to use the generic factors rather than the sector-specific factors.

One should also note that the results are sensitive to the assumption made on the prior standard deviation (see Fig. 2). The effect on the assumption for the additional uncertainty factors expressed by industrial sector can be found in the Electronic Supplementary Material.

The values calculated using the Bayesian approach are more complete than those calculated by Ciroth et al. (2013), i.e., there are uncertainty factors for each "pedigree indicator-pedigree score" couple. They are also generally greater for the higher scores.

3.2 The case of the basic uncertainty factors

The subgroups created for the basic uncertainty factors leads to the development of fewer factors than the ones originally present in the basic uncertainty table in the ecoinvent database. The results can be found in Table 4. The obtained likelihood values and the uncertainty values for all obtained subgroups are available in the Electronic Supplementary Material. As for the additional uncertainty factors, the posterior values are sensitive to the assumptions made on the standard deviation of the prior distributions.

While for most values, the posterior values are only slightly different from their priors, differences are more marked for the particulates emissions (PM10 and PM2.5). This can be explained by the large number of subgroups (respectively, 9, and 19, on 64 subgroups assessed to derive basic uncertainty factors) were used to determine the likelihood function, subgroups that represent different type of technologies (from battery manufacturing to steel foundries, see Electronic Supplementary Material).

3.3 What if new data are available?

Using Bayesian inference allows updating the uncertainty factors when new data are available. Three different cases for the update can be cited. All of them use the posterior values found

Table 4 Obtained posterior value for the basic uncertainty factors

	Combustion			Process			Agriculture		
	Prior	Posterior 1	Posterior 2	Prior	Posterior 1	Posterior 2	Prior	Posterior 1	Posterior 2
Thermal energy, electricity, semi-finished products, material, waste	1.05	1.05	1.11	1.05	1.05	1.06	1.05	1.05	1.05
Transport	2.00			2.00			2.00		
Infrastructure	3.00			3.00			3.00		
Primary energy carriers, metals, salt	1.05			1.05			1.05		
land use occupation	1.50			1.50			1.10		
land use transformation	2.00			2.00			1.20		
Water									
BOD, COD, TOC, DOC, inorganic compounds				1.50	2.13	2.13			
Individual hydrocarbons, PAH				3.00	3.25	3.57			
Heavy metals				5.00			1.80		
Pesticides							1.50		
NO3, PO4							1.50		
land use occupation									
land use transformation									
SOIL									
Oil, hydrocarbons				1.50					
Heavy metals				1.50			1.50		
Pesticides							1.50		
AIR									
Carbon dioxide	1.05	1.06	1.15	1.05					
SO2	1.05								
NMVOC total	1.50								
Nox, N2O	1.50						1.40		
CH4, NH3	1.50						1.20	1.20	1.26
Individuals hydrocarbons	1.50			2.00					
PM > 10	1.50			1.50					
PM10	2.00			2.00	2.12	5.46			
PM2.5	3.00			3.00	3.74	18.74			
PAH	3.00								
CO, Heavy metals	5.00								
Inorganic emissions				1.50	1.56	3.32			
Radionucleides				3.00					

The posterior 1 values were obtained using σ_{1UF} (see Fig. 2) and the posterior values using σ_{2UF}

in the previous section and the posterior distributions that are described in the Electronic Supplementary Material. For an update, the posterior distribution becomes the new prior distributions.

Case 1 Data are available to update generic factors; the data are used to define the likelihood function that, combined to the new generic prior distribution, leads to an updated generic posterior distribution of the uncertainty factor.

Case 2 Data are available for a specific industrial sector *i* for which a new prior distribution is available. The data are used to define the likelihood function for the sector *i* that, combined to the prior distribution for the sector *i*, lead to an update posterior distribution of the uncertainty factor for the specific industrial sector *i*.

Case 3 Data are available for a specific industrial sector *j* for which there is no specific new prior distribution available. The data are used to define the likelihood function for the sector *j*, and this likelihood function is combined to the generic prior distribution, leading to an updated posterior distribution for the specific industrial sector *j*. These data can also be added to the original data (as the sector *j* was not represented): this

Table 5 Updated additional uncertainty factors for the further technological correlation indicator for the manufacturing sector using a new data source

	UF_2	UF_3	UF_4	UF_5
Prior value	1.19	1.52	1.95	2.23
Likelihood value	1.95	2.22	2.40	NA
Posterior value	1.26	1.69	2.07	NA

new group of data allows one to define a new generic likelihood function that, combined to the generic prior distribution, leads to an updated generic posterior distribution of the uncertainty factor.

The following example illustrates a *case 3* situation. Using the data of the INIES database (a French database for the residential sector (INIES 2013)) and, more specifically, the data regarding the manufacturing of boards for walls and ceiling, a likelihood function for the manufacturing sector and the further technological indicator can be defined. The data used to derive the likelihood function can be found in the Electronic Supplementary Material. The new obtained uncertainty factors for the manufacturing sector can be found in Table 5 and is illustrated in Fig. 4 for the pedigree score 4. The data used here do not allow the calculation of an uncertainty factor for a score 5 (these factors developed here, in this specific case, should also be only considered as an example.)

4 Discussion and conclusion

4.1 Limitations due to the methodology

The results described in the preceding section permit to have updated additional uncertainty factors for some specific sectors and new updated based basic uncertainty factors. These results rely on some assumptions, some of which are directly linked to the application of the Bayes theorem.

A first assumption is the definition of the fifth pedigree score for the completeness indicator (see Section 2.3.3). In order to assess the effect of this choice, the case where $n=1$ was assessed leading to an uncertainty factor greater than 3×10^9. To avoid this situation, the value $n=5$ is here kept.

Another assumption is the choice of the logarithm distribution as the distribution for modeling the uncertainty factors (for both the prior distribution and the likelihood distribution). This choice is only based on data representation and was tested (the assumptions of lognormally distributed uncertainty factors is at 95 % true for one third of the assessed uncertainty factors, see Electronic Supplementary Material). This choice is also made to simplify the use of the Bayes theorem by using conjugate distributions. Nonetheless, the specific distribution of the uncertainty factors is not, per se, the parameter of interest here; the mean of the posterior distribution is here the parameter used to describe the uncertainty factors.

The definition of the distribution parameters, especially for the prior distribution for the basic uncertainty factors, are also based on assumptions that affect the results, since the standard deviation of the prior distribution is used to derive the mean of the posterior distribution (see Fig. 2). Ideally, these values should have been derived when the experts defined the uncertainty factors themselves (as the standard deviation can be linked to the experts' level of confidences when defining the uncertainty factors).

This last assumption can be overcome by only using the uncertainty factors coming from the data assessment to derive specific and generic uncertainty factors and so without applying the Bayes theorem. This can be done by simply determining the mean of the uncertainty factors for each subgroup. However, the data harvested for this study don't represent the entire technosphere and this lack of representativeness isn't assessed. Using the Bayes theorem permit to insert a generic information (the experts' judgments); by not using the Bayes theorem, specific information (the one coming from the data assessment) are used to derive generic uncertainty factors that raises the question of representativeness.

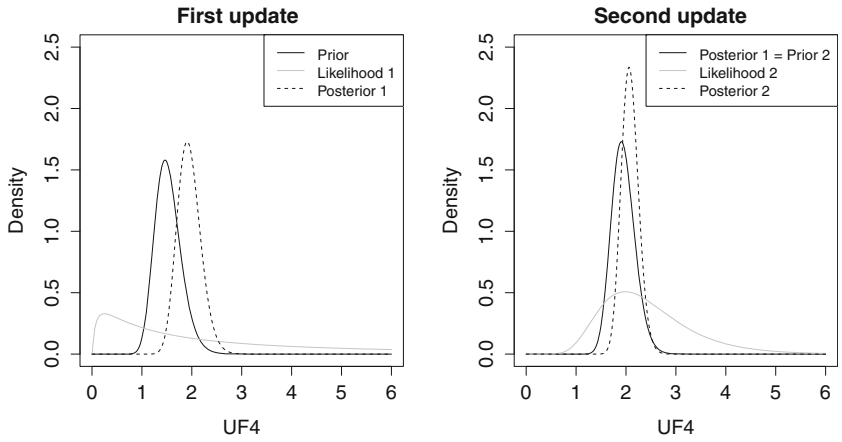

Fig. 4 Representation of the further technological correlation indicator for the manufacturing sector updated two times: once using the master data table used in this study, the second by accessing a new data sources from the INIES database

Finally, if an access to more data is possible, it will allow the calculation of more refined uncertainty factors, i.e., uncertainty factors more representative of a given sector and for more specific industrial sectors.

4.2 How to use these new uncertainty factors?

These updated uncertainty factors can directly be used in studies in which uncertainties on LCI parameters need to be defined. Specifically, when sector-generic basic uncertainty factors and sector-specific additional uncertainty factors are available, they should be preferred to ecoinvent default factors. If they are not available, then the ecoinvent default factors should be used.

Finally and as mentioned in Section 3.3, these updated uncertainty factors (and their specific distribution, see Electronic Supplementary Material) can also be used to develop new uncertainty factors. Depending on the available data, existing factors can be updated, or new sector-specific factors can be defined using the presented methodology (see Table 1 and Fig. 2) and the different underlying assumptions presented above.

4.3 Updating data other than uncertainty factors

Despite the limitations, the developed methodology allowed the calculation of new scientific based uncertainty factors for both the basic and the additional uncertainties that can be directly used or that can be updated as new data become available.

Bayesian inference could also be used in the models used in LCA (whether it is a model to obtain new input parameters for the inventory or models to derive life cycle impact assessment characterization factors). The inference can especially be used when the parameters are based on temporal series. Updating parameters using inference rather that starting over again the derivation of new parameters allows developers or practitioners to save both computing space (only the posterior information needs to be kept rather than all initial data) and time once the Bayesian model is defined and ready to use, even if Bayesian Monte Carlo or Markov chain Monte Carlo need to be performed to apply the theorem. Depending on the models, a 5000-step Monte Carlo simulation takes only few seconds.

Acknowledgements The authors would like to acknowledge the financial support of the industrial partners of the International Life Cycle Chair, a research unit of the CIRAIG: ArcelorMittal, Bombardier, Desjardins Group, Hydro-Québec, LVMH, Michelin, Nestlé, RECYC-QUÉBEC, SAQ, Solvay, Total, Umicore, and Veolia. The authors would also like to acknowledge the support of the two Quebec ministries involved in the project (Ministère du développement durable, de l'environnement et des parcs—now MDELCC—and the Ministère du développement économique, de l'innovation et de l'exportation).

References

Ben Letham CR (2012) 15.097 probabilistic modeling and Bayesian analysis 15.097 prediction: machine learning and statistics - spring 2012: MIT OpenCourseWare.

Björklund A (2002) Survey of approaches to improve reliability in LCA. Int J Life Cycle Assess 7(2):64–72

Ciroth A, Muller S, Weidema B, Lesage P (2012) Refinig the pedigree matrix approach in ecoinvent (Version 7.1). Report for the ecoinvent database

Ciroth A, Muller S, Weidema B, Lesage P (2013) Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int J Life Cycle Assess. doi:10.1007/s11367-013-0670-5, **1–11**

Frischknecht R, Jungbluth N, Althaus H-J, Doka G, Dones R, Heck T, Hellweg S, Hischier R, Nemecek T, Rebitzer G, Spielmann M (2005) The ecoinvent database: overview and methodological framework. Int J Life Cycle Assess 10(1):3–9

Funtowicz S, Ravetz, J (1990) Uncertainty and quality in science for policy. Kluwer Academic Publishers, Dordrecht

Heijungs R, Lenzen M (2014) Error propagation methods for LCA—a comparison. Int J Life Cycle Assess 19(7):1445–1461

Henriksson P, Guinée J, Heijungs R, de Koning A, Green D (2014) A protocol for horizontal averaging of unit process data—including estimates for uncertainty. Int J Life Cycle Assess 19(2):429–436

Hong J, Shaked S, Rosenbaum R, Jolliet O (2010) Analytical uncertainty propagation in life cycle inventory and impact assessment: application to an automobile front panel. Int J Life Cycle Assess 15(5):499–510

Huijbregts M (1998) Application of uncertainty and variability in LCA. Int J Life Cycle Assess 3(5):273–280

Imbeault-Tétreault H, Jolliet O, Deschênes L, Rosenbaum RK (2013) Analytical propagation of uncertainty in life cycle assessment using matrix formulation. J Ind Ecol 17(4):485–492

INIES (2013) Base nationale de référence sur les impacts environnementaux et sanitaires des produits, équipements et services pour l'évaluation de la performance des ouvrages. INIES, France

Katz RW (2002) Techniques for estimating uncertainty in climate change scenarios and impact studies. Clim Res 20:167–185

Lloyd SM, Ries R (2007) Characterizing, propagating, and analyzing uncertainty in life-cycle assessment: a survey of quantitative approaches. J Ind Ecol 11(1):161–179

Lo S-C, Ma H-W, Lo S-L (2005) Quantifying and reducing uncertainty in life cycle assessment using the Bayesian Monte Carlo method. Sci Total Environ 340(1–3):23–33. doi:10.1016/j.scitotenv.2004.08.020

Miller SA, Moysey S, Sharp B, Alfaro J (2013) A stochastic approach to model dynamic systems in life cycle assessment. J Ind Ecol 17(3): 352–362

Muller S, Lesage P, Ciroth A, Mutel C, Weidema B, Samson R (2014) The application of the pedigree approach to the distributions foreseen in ecoinvent v3. Int J Life Cycle Assess. doi:10.1007/s11367-014-0759-5

Qian SS, Stow CA, Borsuk ME (2003) On Monte Carlo methods for Bayesian inference. Ecol Model 159(2–3):269–277

Reap J, Roman F, Duncan S, Bras B (2008) A survey of unresolved problems in life cycle assessment. Int J Life Cycle Assess 13(5):374–388

Ukidwe N, Bakshi BR, Parthasarathy G (2004) A multiscale Bayesian framework for designing efficient and sustainable industrial systems. Paper presented at the AIChE sustainability engineering conference proceedings, Austin

Weidema BP, Wesnæs MS (1996) Data quality management for life cycle inventories—an example of using data quality indicators. J Clean Prod 4(3–4):167–174

Weidema BP, Bauer C, Hischier R, Mutel C, Nemecek T, Vadenbo CO, Wernet G (2013) Overview and methodology. Data quality guidelines for the Ecoinvent database version 3. Ecoinvent report 1 (v3). The ecoinvent centre, St Gallen