

The application of the pedigree approach to the distributions foreseen in ecoinvent v3

Stéphanie Muller · Pascal Lesage · Andreas Ciroth ·
Christopher Mutel · Bo P. Weidema · Réjean Samson

Received: 31 October 2012 / Accepted: 16 May 2014 / Published online: 29 May 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Purpose Data used in life cycle inventories are uncertain (Ciroth et al. Int J Life Cycle Assess 9(4):216–226, 2004). The ecoinvent LCI database considers uncertainty on exchange values. The default approach applied to quantify uncertainty in ecoinvent is a semi-quantitative approach based on the use of a pedigree matrix; it considers two types of uncertainties: the basic uncertainty (the epistemic error) and the additional uncertainty (the uncertainty due to using imperfect data). This approach as implemented in ecoinvent v2 has several weaknesses or limitations, one being that uncertainty is always considered as following a lognormal distribution. The aim of this paper is to show how ecoinvent v3 will apply this approach to all types of distributions allowed by the ecoSpold v2 data format.

Methods A new methodology was developed to apply the semi-quantitative approach to distributions other than the lognormal. This methodology and the consequent formulas

were based on (1) how the basic and the additional uncertainties are combined for the lognormal distribution and on (2) the links between the lognormal and the normal distributions. These two points are summarized in four principles. In order to test the robustness of the proposed approach, the resulting parameters for all probability density functions (PDFs) are tested with those obtained through a Monte Carlo simulation. This comparison will validate the proposed approach.

Results and discussion In order to combine the basic and the additional uncertainties for the considered distributions, the coefficient of variation (CV) is used as a relative measure of dispersion. Formulas to express the definition parameters for each distribution modeling a flow with its total uncertainty are given. The obtained results are illustrated with default values; they agree with the results obtained through the Monte Carlo simulation. Some limitations of the proposed approach are cited.

Conclusions Providing formulas to apply the semi-quantitative pedigree approach to distributions other than the lognormal will allow the life cycle assessment (LCA) practitioner to select the appropriate distribution to model a datum with its total uncertainty. These data variability definition technique can be applied on all flow exchanges and also on parameters which play an important role in ecoinvent v3.

Responsible editor: Reinout Heijungs

Electronic supplementary material The online version of this article (doi:10.1007/s11367-014-0759-5) contains supplementary material, which is available to authorized users.

S. Muller (✉) · P. Lesage · R. Samson
CIRAIG, Department of Chemical Engineering, Polytechnique
Montréal, P.O. Box 6079, Stn. Centre-ville, Montréal, Québec H3C
3A7, Canada
e-mail: stephanie.muller@polymtl.ca

A. Ciroth
GreenDelta GmbH, Müllerstrasse 135, 13349 Berlin, Germany

C. Mutel
ETH Zurich, Institute of Environmental Engineering, 8093 Zurich,
Switzerland

B. P. Weidema
Aalborg University, Skibbrogade 5, 1, 9000 Aalborg, Denmark

Keywords Data quality · Life cycle inventory database ·
Pedigree matrix · Probability density functions · Uncertainty

1 Introduction

Life cycle assessment (LCA) is a quantitative model of the industrial ecosystem, and LCA results always have uncertainty, i.e., are not perfectly known. Sources and typologies of uncertainty in LCA have been well documented (Huijbregts et al. 2003; Lloyd and Ries 2007; Björklund 2002;

Sonnemann et al. 2003). Uncertainty can be divided into model, scenario, and parameter uncertainty. The focus of this paper is on the parameter uncertainty due to lack of knowledge associated to the true value of a quantity associated with the values of exchanges (elementary and intermediate flows) in the context of life cycle inventory (LCI) databases. Although we recognize the importance of uncertainty in life cycle impact assessment, this paper discusses only the inventory phase.

These parameters, the quantitative data needed to compile a LCI (e.g., exchange values), are uncertain for a number of reasons; for example, the real value could not be known (Björklund 2002), or the measured data do not accurately describe the “true” median or mean value (Ciroth et al. 2004). In real-world technological and environmental processes, there is also considerable variability. The concept of variability is often confused with uncertainty. Variability refers to the inherent variations of a parameter value in space, time, or across individuals (Huijbregts 1998) while uncertainty (specifically, parameter uncertainty) includes also the lack of confidence that a parameter value is truly representative of the true value. It is important to consider uncertainties in the interpretation phase of LCI and LCA studies in order to determine the confidence in results. Usually, a quantitative parameter is known with a certain level of uncertainty and a certain level of quality. Taking into account this uncertainty and quality of the inputs in a LCA study, especially on parameters, will reinforce the confidence in the results and help the decision-making process based on the results and their interpretation in a quantitative and qualitative way. Several approaches were also developed in order to take these uncertainties at a parameter level, see Lloyd and Ries (2007) for a review, including purely qualitative (Rousseaux et al. 2001), semi-quantitative (Kennedy et al. 1996; Maurice et al. 2000; Weidema and Wesnaes 1996), and quantitative (Tan et al. 2002; Giroth et al. 2004; Coulon et al. 1997) approaches. Different approaches were also proposed to consider how the parameter level uncertainty translates into uncertainty of actual LCA results (Hong et al. 2010; Giroth et al. 2004), although this topic will for the most part be outside the scope of this paper.

The LCI phase of an LCA aims to collect and compile all the heterogeneous data needed to quantify the exchanges within the studied product system and between the product system and the environment. Since product systems model complex supply chains, LCA practitioners and researchers normally revert to using generic LCI databases that contain exchange values for different processes. Since these databases are an important data source for LCA, how they describe uncertainty in their data is the key to subsequently evaluating the overall uncertainty of the LCA.

The ecoinvent database is one of a very few LCI databases that systematically includes explicit uncertainty data. A semi-

quantitative approach based on the use of a pedigree matrix is used. This approach, first proposed for the LCA by Weidema and Wesnaes (1996), is inspired from the NUSAP system developed in the early 1990s (Funtowicz and Ravetz 1990) and has been in use in the ecoinvent database since 2005 (Frischknecht et al. 2005). It considers two kinds of parameter uncertainty:

1. Intrinsic variability and stochastic error of the parameters, due to, e.g., measurement uncertainties, activity specific variability, temporal variability, etc. This uncertainty is captured in a so-called *basic uncertainty* factor;
2. Uncertainty due to the use of imperfect data, e.g., data resulting from estimates, lacking verification, or extrapolated from temporally, spatially and/or technologically different conditions, collectively grouped under the term *additional uncertainty*.

In other terms, the approach addresses two of the three categories of parameter uncertainty described in Huijbregts et al. (2001), namely, data inaccuracy and lack of representative data for the context of use. In the approach, a datum and its uncertainty (the basic and additional uncertainty) are modeled using a probability density functions (PDF). In the earlier version of the ecoinvent database, the lognormal distribution was used by default (Frischknecht et al. 2005). The lognormal distribution is represented by two definition parameters: the geometric mean (μ_g) and the geometric standard deviation (GSD) see Table 1. The geometric mean (also the median) is the deterministic value, and the GSD captures the information on the uncertainty. Uncertainty factors were also expressed in terms of GSD^2 (basic uncertainty factors) and “contributors to the GSD^2 ” (additional uncertainty factors) in ecoinvent v2. Having statistical information about the data, basic uncertainty factors were directly calculated as GSD^2 . When no statistical information was available, basic uncertainty factors could be defined by default. For this, ecoinvent provided a table with basic uncertainty factors differentiated by exchange type (total of 27 types, ranging from specific classes of air pollutants to different types of intermediate exchanges) and by class of process (combustion, process, or agricultural). The datum was then evaluated according to a “pedigree matrix” by which it was scored (1 to 5, where 5 is worse) on the following independent data quality characteristics: reliability (sampling methods and verification procedures); completeness (statistical representativeness of the datum and time periods for data collection); temporal, geographic, and further technological correlation (for data used outside its proper context) and sample size.

The semi-quantitative scores, based on the pedigree matrix, were then converted into uncertainty figures using additional uncertainty factors, expressed as a contribution to the square of the geometric standard deviation. For example, a reliability

Table 1 Definition of the distributions foreseen in ecoinvent v3 and how to express the CV according to the definition parameters

Name of the distribution	Parameters	Deterministic value	PDF	How to link the coefficient of variation (CV) with the parameters
Lognormal	μ_g : geometric mean σ_g : geometric standard deviation	Median: μ_g	$f(x, \mu_g, \sigma_g) = \frac{\exp\left(\frac{-(\ln x - \ln \mu_g)^2}{2 \ln^2 \sigma_g}\right)}{\sqrt{2\pi} \ln \sigma_g}$	$CV = \sqrt{\exp(\ln^2 \sigma_g) - 1}$
Normal	μ : arithmetical mean σ : arithmetical standard deviation	Mean: μ	$f(x, \mu, \sigma) = \frac{\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$	$CV = \frac{\sigma}{\mu}$
Uniform	a : minimum of the distribution b : maximum of the distribution	Mean 0.5 ($a+b$)	$f(x, a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}$	$CV = \frac{b-a}{\sqrt{3(b+a)}}$
Triangular	a : minimum of the distribution b : maximum of the distribution c : most likely value of the distribution	Most likely value: c	$f(x, a, b, c) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a < x < c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x < b \\ 0 & \text{otherwise} \end{cases}$	$CV = \frac{\sqrt{a^2+b^2+c^2-ab-ac-cb}}{\sqrt{2}(a+b+c)}$
Beta PERT	a : minimum of the distribution b : maximum of the distribution c : most likely value of the distribution $\alpha = 6 \frac{\mu-a}{b-a}$ and $\beta = 6 \frac{b-\mu}{b-a}$ and $\mu = \frac{a+4c+b}{6}$	Most likely value: c	$f(x, a, b) = \frac{(x-a)^{\alpha-1} (b-x)^{\beta-1}}{B(\alpha, \beta) (b-a)^{\alpha+\beta-1}}$	$CV = \frac{b-a}{a+4c+b}$
Gamma	k : shape parameter λ : scale parameter	Most likely value: λ ($k-1$)	$f(x, k, \lambda) = \frac{x^{k-1} \exp(-x/\lambda)}{\Gamma(k) \lambda^k}$	$CV = \frac{1}{\sqrt{k}}$
Binomial	k : number of successes n : number of trials p : probability of success		$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$	$CV = \sqrt{\frac{1-p}{np}}$

score of “5”, representing a value based on a non-qualified estimate, would be associated with a contributor to the GSD² of 1.50. These additional uncertainty factors were subsequently compiled with the basic uncertainty factors using a compilation formula, derived from the properties of the lognormal distribution, resulting in a measure of the total uncertainty. The total uncertainty was expressed as the square of the geometric standard deviation (GSD²). If in ecoinvent, the “pedigree matrix approach” refers to the determination of the additional uncertainty factors; the term “pedigree approach” will be used in this paper to name the whole semi-quantitative structure: from the determination of the additional uncertainty to the compilation of both types of uncertainty.

In the context of a database, this approach has many advantages. Uncertainty information can be made available for all types of exchanges, even if statistical information is not available. This ensures that the cumulative uncertainty of LCIs, calculated via, e.g., Monte Carlo or analytical propagation approaches, takes into account the uncertainty of every exchange in the model. The data quality indicators derived from the use of the pedigree matrix can also be used directly as a data quality management tool.

However, the pedigree approach as applied in ecoinvent v2 had several weaknesses. One of them was the fact that the pedigree approach was only valid for lognormal distributions,

while four different PDFs to model a datum’s uncertainty were already available in the previous version of the ecoinvent database. While the lognormal distribution has many nice properties for modeling physical parameters (e.g., it is strictly positive, positively skewed, and is the result of the multiplication of independent random variables (MacLeod et al. 2002; Slob 1994)), distributions other than the lognormal are more appropriate when they better represent the uncertainty associated with the datum. Most often, this will be the case when the basic uncertainty has been calculated based on available data. In these cases, the parameters of the basic uncertainty PDF will be directly known. In the cases where (1) the basic uncertainty cannot be calculated due to small sample size and (2) one nonetheless wishes to use a distribution other than the lognormal, the definition parameters of the PDF modeling a datum and its basic uncertainty can be determined using the default basic uncertainty factors used for the lognormal by converting the GSD² and the deterministic value into the definition parameters of the desired distribution.

The objective of this paper is to provide practical considerations on how to apply the pedigree approach for all distributions allowed in ecoinvent version 3. This specific objective treated in this paper belongs to a more global objective that is to improve the above approach and to legitimize the actual structure of uncertainty calculation in the ecoinvent database.

The ecoinvent v3 database and its associated ecoSpold v2 data format will include three changes to the way uncertainty is handled:

1. Three new distributions will be available, the gamma, the beta (PERT approximation), and the binomial distribution. The binomial distribution is not affected by additional uncertainty and is therefore not addressed further. The Erlang distribution is a special case of the gamma distribution and can therefore also be used. The definitions for these distributions are presented in Table 1. The uniform, triangular, normal, and lognormal distributions were already available in ecoinvent v2.
2. The data quality indicator “sample size” and its corresponding additional uncertainty factors will be removed from the pedigree approach, as sample size should already be captured in the basic uncertainty factor. This should be recognized as a simplification, since the generic basic uncertainty factors do not capture the large deviations of small sample sizes often encountered in LCA.
3. The pedigree approach will be expanded to cover all six probability distribution functions.

This paper deals specifically with point 3. The methodology developed to combine basic and additional uncertainties for all PDFs, and the resulting formulas are presented.

2 Materials and methods

2.1 Methodology applied to derive analytical formulas

In ecoinvent version 2, default basic and additional uncertainty factors were developed, and a method to combine basic and additional uncertainty was proposed only for the lognormal distribution, used by default in the database. This combination is based on the fact that the spread of the distribution increases when adding an uncertainty component (the additional uncertainty here). In case of the lognormal distribution, the spread can be expressed through the 95th confidence interval $[\mu_g/GSD^{1.96}; \mu_g \cdot GSD^{1.96}]$. Conventionally, the value 1.96 is rounded to 2, and the 95th confidence interval is commonly expressed as $[\mu_g/GSD^2; \mu_g \cdot GSD^2]$. It has been shown that in case of experts' judgments and scarce data (characteristics that apply to the pedigree approach), the GSD^2 is a good measure of the spread of the distribution (MacLeod et al. 2002; Slob 1994). Because the basic uncertainty was expressed in terms of GSD^2 , the additional uncertainty factors were also expressed as “a contribution to the GSD^2 ” (Frischknecht et al. 2005).

Concretely, the combination that aims to correct the initial uncertain value using different components of variability, represented by the pedigree criteria, is done as follows:

A quantitative datum needed to compile a LCI is expressed with its basic uncertainty as a random variable (D_b). In the pedigree approach as implemented to date, D_b follows a lognormal distribution whose deterministic value is set as the median (μ_g), and the basic uncertainty factor defines the $GSD^2_{D_b}$. Then, the different components of the pedigree matrix act as modifiers of this random variable D_b , leading to the random variable D_T . This parameter is also a lognormally distributed random variable and has the same deterministic value as D_b . However, the uncertainty factor, now represented by $GSD^2_{D_T}$, called the “total” uncertainty, represents the initial basic uncertainty whose spread was increased to account for additional uncertainty. This additional uncertainty expresses the variability between a perfect case and the case described by the related pedigree matrix cell, and this for each pedigree criterion. For example, for the pedigree criterion “temporal correlation” and for a pedigree score equals to 3, the cell's description is “less than 10 years of difference to the time period of the dataset” (Weidema et al. 2013). For this specific case, the additional uncertainty must express the variation that exists in a 0 from 10 years time interval. In other words, the variation that occurs when compiling different data measured in this time interval. It can be defined as a random variable. The same can be said about the 5 pedigree criteria. D_i is also the random variable representing the additional uncertainty (for i from 1 to 5, representing the five pedigree criteria—from *reliability to further technological correlation*). These random variables were initially defined by experts' judgments; they can also be determined using empirical sources as shown in the work by Ciroth et al. (2013). Experts' judgments are better represented using a lognormal distribution (MacLeod et al. 2002; Slob 1994); D_i are also represented as lognormal distributions with a geometric mean of 1 and with additional uncertainty factors as $GSD^2_{D_i}$. The lognormal distribution is also useful because, having a geometric mean of 1, D_i act as modifiers on D_b through Eq. 1 that results to an unmodified deterministic value. Moreover, the GSD^2 resulting from the multiplication of n independent lognormal distributions (D_b and D_i here) is expressed through Eq. 2 (Limpert et al. 2001). As the pedigree criteria are independent (Weidema and Wesnæs 1996) so are D_T , D_b , and D_i and Eq. 2 can be used to combine basic and additional uncertainty leading to the formula applied in ecoinvent v2.2 (Frischknecht et al. 2005). Combination of random variables to express uncertainties can be found in environmental science (through assessment factors) and in some risk assessment methods (MacLeod et al. 2002; Vermeire et al. 1998; Slob 1994).

$$D_T = D_b \cdot \prod_{i=1}^5 D_i \quad (1)$$

$$GSD_{D_T}^2 = \exp \left(\sqrt{\ln^2 GSD_{D_b}^2 + \sum_{i=1}^5 \ln^2 GSD_{D_i}^2} \right) \quad (2)$$

The application of the pedigree approach to other PDFs will be directly based on how it was applied to the lognormal PDF and will follow the following four principles:

- The additional uncertainty must modify neither the deterministic value (i.e., the statistical parameter used to define the exchange values in ecoinvent (Weidema et al. 2013)) nor the type of distribution chosen to represent the data (with its basic uncertainty).
- The total uncertainty is equal to the basic uncertainty when no additional uncertainty is added, i.e., when the data quality is assumed to be perfect and hence scores “1” for all data quality indicators using the pedigree matrix.
- The additional uncertainty is dimensionless and expresses the datum’s relative dispersion, according to each pedigree criteria, due to the use of imperfect data.
- The value of the additional uncertainty factors used for the lognormal distribution are used to derive the additional uncertainty for other PDFs (see the Data Quality Guidelines of the ecoinvent database for the values of additional uncertainty factors (Weidema 2013)).

The lognormal and the normal distributions are intimately linked; if the random variable X is lognormally distributed, $\ln(X)$ is normally distributed and vice versa. Eq. 1 can also be translated into Eq. 3 where $\ln D_T$, $\ln D_b$, and $\ln D_i$ are normally distributed. That leads to Eq. 4.

$$\ln D_T = \ln D_b + \sum_{i=1}^5 \ln D_i \quad (3)$$

$$\sigma_{\ln D_T} = \sqrt{\sigma_{\ln D_b}^2 + \sum_{i=1}^5 \sigma_{\ln D_i}^2} \quad (4)$$

So, if a datum D_b is originally normally distributed, Eq. 4 can be expressed in terms of Eq. 5 based, moreover, on the widely used formula to determine the standard deviation of a function of random independent variables.

$$\sigma_{D_T} = \sqrt{\sigma_{D_b}^2 + \sigma_{D_i}^2} \quad (5)$$

where D_i represents the random variable modeling the additional uncertainty that combines the five pedigree criteria.

The correspondences between the normal and the lognormal distribution can be found in Table 2, and Eq. 5 can also be applied for the uniform, triangular, beta PERT, and gamma distributions.

The standard deviation expresses an absolute dispersion around the mean; in the development of generic factors, that capture uncertainty information for all type of PDFs, the measure of dispersion should be expressed in terms relative to the median or mode for all the PDFs. This is done by using the relations between the lognormal and the normal distributions (see Table 2).

The coefficient of variation (CV) is chosen as a universal measure of variability. It is defined as the ratio between the arithmetic standard deviation and mean. More concretely, the CV measures a relative dispersion within a specific sample, regardless the distribution type considered. As it is a dimensionless measure of dispersion, the CV allows the comparison of the dispersion from different samples. It is also the initial measure of variation chosen to express the uncertainty with the pedigree approach (Weidema and Wesnæs 1996). The total uncertainty will also be expressed as a CV. The uncertainty factors expressed as a GSD^2 applied in ecoinvent v2 can be easily translated into CV using equations presented in Table 2. In the same way, the expression of the additional uncertainty can easily be obtained using Eq. 2 and the definition of the CV using the GSD. The additional uncertainty expressed in term of CV (CV_i), is defined by Eq. 6 where CV_i are the additional uncertainty factors linked to the five pedigree criteria.

$$CV_i = \sqrt{\prod_{i=1}^5 (CV_i^2 + 1)} - 1 \quad (6)$$

The next step is also to express Eq. 5 in terms of CV. Secondly; the standard deviation expresses an absolute dispersion around the arithmetic mean value. This dispersion—as well as the confidence interval around the mean—increases when additional uncertainty is added. In the case of symmetric distributions (such as the normal and uniform distributions), the mean is the mode. Since, for symmetric distribution, the arithmetic mean should not be modified when adding the additional uncertainty, the absolute dispersion is equivalent to the relative dispersion, and Eq. 7 can be used for symmetric distributions.

$$CV_T = \sqrt{CV_D^2 + CV_I^2} \quad (7)$$

In the case of asymmetric distributions, the arithmetic mean differs from the mode. For these distributions, the arithmetic mean will be affected by the consideration of additional uncertainty. While Eq. 7 can still be used to calculate the relative dispersion parameter (CV_T), a new arithmetic mean (μ_T) that takes into account the effect of the additional uncertainty must be calculated using Eq. 8 where μ is the arithmetic mean of the datum with its basic uncertainty.

Table 2 Correspondences between the normal and the lognormal distributions

	Normal distribution	Lognormal distribution
Deterministic value	μ	μ_{\log}
Dimensionless measure of variability	Coefficient of variation $CV = \frac{\sigma}{\mu}$	$\sigma_g = \exp\left(\sqrt{\ln(CV^2 + 1)}\right)$ where σ_g is the GSD. $CV = \sqrt{\exp(\ln^2 \sigma_g) - 1}$ Alternatively:
Confidence interval (68 %)	$[\mu - \sigma; \mu + \sigma]$	$[\mu_g/\sigma_g; \mu_g\sigma_g]$

$$\mu_T CV_T = \mu \sqrt{CV_D^2 + CV_I^2} \quad (8)$$

Having Eq. 7 and Eq. 8, the definition parameters of each distribution can be expressed through CV_T and also be defined for the total uncertainty.

2.2 Comparison between the analytical approach and a numerical approach

In order to test and validate the PDFs obtained through analytically derived formulas, they will be compared with that obtained using a numerical approach, the Monte Carlo simulation. The starting assumption that the additional uncertainty factors are described using GSD² remains unchanged, and distributions with total uncertainty are generated through a Monte Carlo simulation where the distribution modeling the data with its basic uncertainty is multiplied by the additional uncertainty represented by a lognormal distribution. The Monte Carlo simulation consists in randomly sampling the probability of each uncertain parameter (the data and its basic and additional uncertainty here) and then computing the sampled values using the model. The five distribution applied in ecoinvent v3 are tested, each with an arbitrary basic uncertainty factor (expressed as a CV), the same deterministic value and four different quintuplets giving a total of 20 tests. The simulation will be a 10,000-step analysis performed by the ORACLE Crystal Ball release fusion edition (v 11.11.2.0) (ORACLE 2010). As the basic uncertainty and the additional uncertainty components are considered as independent (Weidema and Wesnæs 1996), covariances will not be considered in the simulation.

The Monte Carlo simulation is a good approximation that tends to theoretical when the number of steps increases. So, in order to compare the analytical and the numerical approaches for each distribution, relative errors will be calculated on the most relevant parameters representing the distribution and its total uncertainty: the standard deviation and the CV for the normal distribution; the minimum, the maximum and the CV

for the uniform, triangular and beta PERT distributions; and the CV for the gamma distribution. In case of the distributions defined by location parameters (the uniform, triangular, and beta PERT ones), the percentage of the resulting values obtained through the Monte Carlo simulation and that belongs to the interval defined by the minimum and the maximum of the distribution with its total uncertainty will also be determined.

3 Results and discussion

This section presents the obtained results for all distributions and the starting assumptions to derive the resulting formulas. Throughout, the subscript T will represent a distribution or parameter that models a datum with its total uncertainty.

3.1 Symmetric distributions

The normal and the uniform distributions are both symmetrical. Eq. 7 is used, and the arithmetic mean is not modified by adding the additional uncertainty. Table 3 presents these assumptions in a mathematical form for both distributions and the resulting formulas to combine the basic and the additional uncertainties.

3.2 Asymmetric distributions

The triangular, beta PERT, and gamma distributions are all asymmetric. The starting points to derive the formulas to combine basic and additional uncertainties are slightly different as the ones cited in the previous section. First of all, Eq. 8 is used to determine CV_T and the deterministic value (the mode is not modified when the additional uncertainty is added). Furthermore, the shape of the distribution is also not modified by adding the additional uncertainty. This last assumption is translated into a shape parameter (γ) that is defined for the triangular and the Beta PERT distribution. Table 4 presents these assumptions and the resulting formulas to combine the basic and the additional uncertainties.

Table 3 Assumptions and resulting formulas for symmetric distributions

Distribution	Assumptions	Formulas to combine basic and additional uncertainties
Normal	$\begin{cases} \mu = \mu_T \\ CV_T = \sigma_T/\mu \end{cases}$	$\begin{aligned} \mu &= \mu_T \\ \sigma_T &= \mu CV_T \end{aligned}$
Uniform	$\begin{cases} \mu = \frac{a+b}{2} = \frac{a_T+b_T}{2} = \mu_T \\ CV_T = \frac{b_T-a_T}{\sqrt{3}(a_T+b_T)} \end{cases}$	$\begin{aligned} a_T &= a + b - b_T \\ b_T &= \mu (1 + \sqrt{3} CV_T) \end{aligned}$

A note can be added on the triangular distribution: the starting points lead to solve a quadratic equation. Although this quadratic equation has two solutions, only one is realistic here: the one that considers that b_T will be higher than b and a_T smaller than a (where b and a are respectively the maximum and the minimum of the probability density function).

3.3 Comparison with the results obtained through the Monte Carlo simulation

The total uncertainty resulting from the formulas available in Table 3 and Table 4 is compared with that obtained through a Monte Carlo simulation. The five new distributions with an arbitrary basic uncertainty factor, now expressed as a CV and the same most likely value (1.5), are tested with four different pedigree quintuplets ((2;2;2;2;2), (3;3;3;3;3), (4;4;4;4;4), (5;5;5;5;5)), giving a total of 20 tests. The definition parameters for the PDFs without additional uncertainty, as well as those calculated for the total uncertainty, are presented in Table 5. The graphics of the different resulting distributions are presented in Fig. 1 (for the basic uncertainty and pedigree scores (3;3;3;3;3) and (5;5;5;5;5)) both for the analytical approach and distributions obtained through the Monte Carlo

simulation (representing by the histograms on Fig. 1). The relative errors on the relevant parameters are calculated. All the results from this comparison can be found in the [Electronic Supplementary Material](#), and only the major conclusions are presented here.

In general, the CVs obtained through the Monte Carlo simulation are greater than the ones determined by the proposed formulas, except for three tests (the normal distributions with all pedigree scores equal to 2 and the beta PERT distributions with all pedigree scores equal to 2 and to 3). That can be explained by the fact that the additional uncertainty is modeled by a lognormal distribution which is unbounded and with longer tails when the additional uncertainty (and also the spread) increases. The resulting distributions from Monte Carlo simulations have also longer tails (see Fig. 1).

For the normal distribution, relative errors on CV_T were less than 5 %, which can be considered acceptable. These relative errors on CV_T are higher for the other types of distributions. In particular, the approach does not give satisfying results for the gamma distribution, for which the relative errors on CV_T were between 24 and 30 %. Moreover, the developed formulas for the gamma distribution are only valid when it admits 0 as location parameter. If the definition parameters of

Table 4 Assumptions and resulting formulas for asymmetric distributions

Distribution	Assumptions	Formulas to combine basic and additional uncertainties
Triangular	$\mu = \frac{a+b+c}{3} \quad \text{and} \quad \mu_T = \frac{a_T+b_T+c}{3}$ $\begin{cases} \gamma = \frac{c-a}{b-c} = \frac{c-a_T}{b_T-c} \\ CV_T^2 = 0.5 \frac{a_T^2 + b_T^2 + c^2 - a_T b_T - a_T c - b_T c}{(a_T + b_T + c)^2} \end{cases}$	$a_T = c(1 + \gamma) - \gamma b_T$ $b_T = c + 3\mu \sqrt{CV_D^2 + CV_I^2} \sqrt{\frac{2}{1 + \gamma + \gamma^2}}$
Beta PERT	$\mu = \frac{a+4c+b}{6} \quad \text{and} \quad \mu_T = \frac{a_T+4c+b_T}{6}$ $\begin{cases} \gamma = \frac{c-a}{b-c} = \frac{c-a_T}{b_T-c} \\ CV_T = \frac{b_T-a_T}{a_T+4c+b_T} \end{cases}$	$a_T = c(1 + \gamma) - \gamma b_T$ $b_T = c + \frac{\sqrt{CV_D^2 + CV_I^2}}{1 + \gamma} (a + 4c + b)$
Gamma	$\begin{cases} m = \lambda(k-1) = \lambda_T(k_T-1) \\ CV_T = 1/\sqrt{k_T} \\ \mu = k\lambda \quad \text{and} \quad \mu_T = k_T\lambda_T \end{cases}$ <p>where m is the most likely value and μ the arithmetic mean</p>	$k_T = 1 + \frac{m^2 + \sqrt{(2\mu m \sqrt{CV_D^2 + CV_I^2})^2 + m^4}}{4\mu^2 (CV_D^2 + CV_I^2)}$ $\lambda_T = \frac{m}{k_T-1}$

Table 5 Definition parameters used to define the tested distribution that modeled a datum with its basic uncertainty and resulting calculated parameters when additional uncertainty is added. The basic and additional uncertainty combination is made using the proposed formulas foreseen in Tables 3 and 4

PDF	Parameters	With basic uncertainty	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
Lognormal	μ_g	1.5	1.5	1.5	1.5	1.5
	σ_g	1.279	1.289	1.313	1.416	1.690
Normal	μ	1.5	1.5	1.5	1.5	1.5
	σ	0.375	0.380	0.414	0.530	0.821
Gamma	k	16	15.66	13.47	8.92	4.72
	λ	0.1	0.102	0.120	0.189	0.403
Uniform	a	1	0.991	0.921	0.677	0.0386
	b	3	3.009	3.079	3.323	3.961
Triangular	a	1	0.993	0.940	0.765	0.336
	b	3	3.021	3.180	3.706	4.991
	c	1.5	1.5	1.5	1.5	1.5
Beta PERT	a	1	0.991	0.921	0.700	0.184
	b	3	3.028	3.237	3.901	5.450
	c	1.5	1.5	1.5	1.5	1.5

a Gamma distribution (k and λ) are not perfectly known, a lognormal distribution should be preferred.

In the specific case of bounded distribution defined by location parameters (uniform, triangular, and beta PERT distributions), the percentage of the vales resulting from the 10,000-step Monte Carlo simulation and that belongs to the interval $[a_T, b_T]$ was also determined. For the three distributions, from 100 to 90.79 % of the values from the simulation belong to the associate interval giving also satisfying results. For these bounded distributions, the higher the additional uncertainty, the higher the relative error on the location parameters. Again, this can be explained by the fact that the lognormal distribution modeling the additional uncertainty is an unbounded distribution with higher spread when the value of additional uncertainty increases. As a_{MC} and b_{MC} are the minimum and the maximum resulting from the Monte Carlo simulation, they take into account this larger spread. Finally, for the uniform and the triangular distributions, the greater the additional uncertainty, the less values resulting from the Monte Carlo simulations lying in the interval $[a_T, b_T]$. For the beta PERT, the opposite is observed. The Monte Carlo simulation being only a numerical approximation, in this case (and also for the other tests), the differences between the obtained percentages (from 99.52 to 100 %) cannot be considered as significant. It could also be noted that the value 100 % (obtained from the beta PERT distribution with all pedigree scores equal to 5) is a rounded value as some outliers are lying outside the interval $[a_T, b_T]$.

3.4 Application of the formulas

The proposed formulas can be applied as well to define the parameters of the desired distribution, keeping in mind that the used deterministic value should remain the same

after adding additional uncertainty and should be the same as the one used as assumption in this proposed framework. Table 1 identifies the deterministic value for each distribution.

The formulas allow determining the CV of the data with its total uncertainty. If ones choose to express the uncertainty in a more usual fashion, for example with standard deviation σ_T , it can be retrieved from the following relation $\sigma_T = \mu_T CV_T$, where μ_T and CV_T are defined in Table 5.

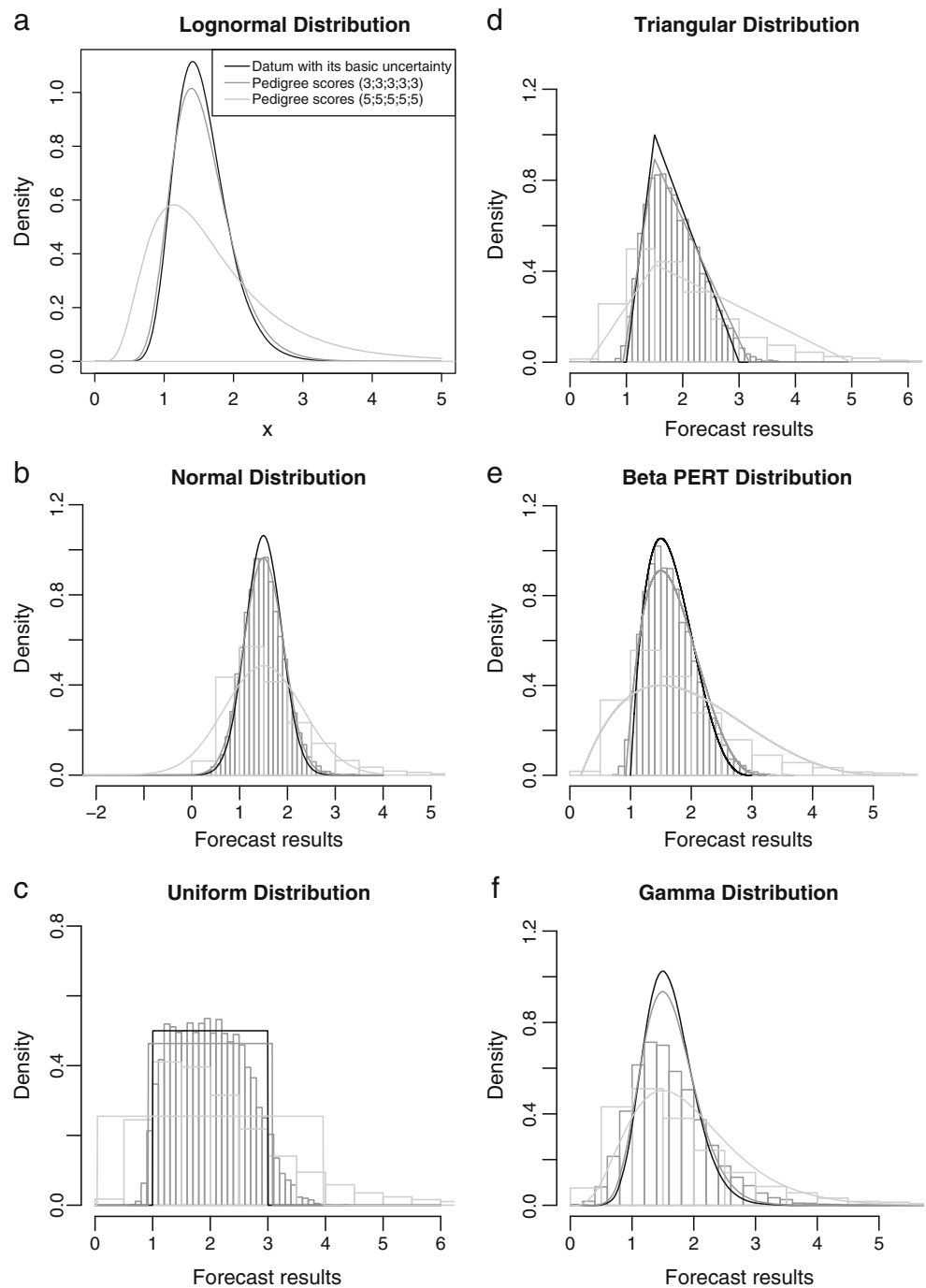
Furthermore, these formulas, as well as the generic pedigree framework, should be applied keeping in mind that it is a semi-quantitative structure that aims to convert a codification of quality and lack of knowledge into uncertainty figures. If it is a good approximation when data characteristics are not known, the approach cannot be substituted by the use of raw data and statistics when these are available.

3.5 Results and limitations

As seen for the gamma distribution, the proposed approach has several limits. Some other limits coming from the applied methodology and the obtained results can be cited. First of all, although the used distributions served to model physical quantities, negative values can be obtained through the developed formulas (it happens in the given example: for the normal distribution when the additional uncertainty is determined through pedigree scores (5;5;5;5;5) the probability to obtain negative values is 3.10^{-2}). No sustainable way for avoiding this is proposed here; nevertheless, three ways to consider negative values can be here cited:

- Consider them in uncertainty analysis, keeping in mind that they are fictive values;

Fig. 1 Illustrations of the different distributions foreseen in Table 5. The histograms represent the results obtained through the Monte Carlo simulation



- Define a threshold or a limit value (in this case 0) in the distribution definition. A certain percentage of possible values is, with this method, not considered;
- Transform all the negative values into a null value. The probability to get 0 will be nonetheless more important.

Another limit comes from the assumption that the additional uncertainty is best described using a lognormal distribution. This aspect will be checked in further work.

Another possible way to apply the pedigree approach to the other distributions will be to develop a computational solution based on Monte Carlo simulation. This would have some advantages. It would be possible to use different PDFs for different additional uncertainty types. For example, it is possible that, e.g., the basic uncertainty is best described by a lognormal distribution, and the uncertainty associated with the temporal correlation (one of the indicator from the pedigree matrix) is best described by a triangular distribution. The here

proposed approach does not permit to create a mix of distribution types as the additional uncertainty is assumed to be lognormally distributed. The computational approach based on Monte Carlo simulation will also dispense us of making some assumptions described in the previous section.

Last but not least, it should be noted that a mathematically more rigorous, but onerous, approach to determine the variance of a function of random variables is the convolution product. While convolution product theoretically yields exact results, it is difficult to implement in software such as the ecoEditor and, depending on the assessed distribution, may actually lead to numerical approximations of the variance.

4 Conclusions

In order to apply the pedigree approach to all distributions foreseen in ecoinvent v3, it is first proposed to consider the CV as a dimensionless measure of variability. The proposed analytical formulas to combine basic and additional uncertainties for distributions other than lognormal are based on how this combination is made for the lognormal PDF and on the links between the multiplicative and the additive models. The different formulas are also based on the CV definition for each distribution. These obtained formulas fit, often quite well, with the combination of basic and additional uncertainties obtained through a Monte Carlo simulation. However, this proposed approach has some limitations. Some of them, particularly the starting assumption that the additional uncertainty is lognormally distributed, will be treated in future work. Indeed, despite its limits, based on the fact that it is a generic approach, the pedigree approach permits to consider the two elements that affect parameter uncertainty: data inaccuracy (through basic uncertainty) and lack of representative data (through additional uncertainty) in a semi-automatic way, explaining its use in a generic LCI database (see Huijbregts et al. for further details (2001)).

These data variability assessment techniques will be applied on all flow exchanges. They will also be applied to parameters which play an important role in ecoinvent v3 to calculate different exchange values. In order to be consistent in the whole ecoinvent database, the same approach—the pedigree approach—will be applied to define and quantify uncertainty on parameters.

Acknowledgments The authors would like to acknowledge the financial support of the industrial partners of the International Chair in Life Cycle Assessment and the International Life Cycle Chair (research units of CIRAI): ArcelorMittal, Bell Canada, Bombardier, Cascades, Eco Entreprises Québec, Groupe EDF, GDF-SUEZ, Hydro-Québec, Johnson&Johnson, LVMH, Michelin, Mouvement Desjardins, Nestlé, Rio Tinto Alcan, RECYC-QUÉBEC, RONA, SAQ, Solvay, Total, Umicore, and Veolia Environnement. The industrial partners were in no

way involved with the study design, the collection, analysis, and interpretation of the data, the writing of the paper, or the decision to submit the paper for publication.

The authors would also like to acknowledge the work of two anonymous reviewers that helped clarify and improve this paper.

References

- Björklund A (2002) Survey of approaches to improve reliability in LCA. *Int J Life Cycle Assess* 7(2):64–72
- Ciroth A, Fleischer G, Steinbach J (2004) Uncertainty calculation in life cycle assessments. *Int J Life Cycle Assess* 9(4):216–226
- Ciroth A, Muller S, Weidema B, Leage P (2013) Empirically based uncertainty factors for the pedigree matrix in ecoinvent. *Int J Life Cycle Assess*. doi:10.1007/s11367-013-0670-5
- Coulon R, Camobreco V, Teulon H, Besnainou J (1997) Data quality and uncertainty in LCI. *Int J Life Cycle Assess* 2(3):178–182
- Frischknecht R, Jungbluth N, Althaus H-J, Doka G, Dones R, Heck T, Hellweg S, Hirschler 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 SO, Ravetz JR (1990) Uncertainty and quality in science for policy. Philosophy and methodology of the social sciences. Kluwer Academic Publishers, Dordrecht
- 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
- Huijbregts M, Norris G, Bretz R, Giroth A, Maurice B, von Bahr B, Weidema B, de Beaufort A (2001) Framework for modelling data uncertainty in life cycle inventories. *Int J Life Cycle Assess* 6(3):127–132
- Huijbregts MAJ, Gilijamse W, Ragas AMJ, Reijnders L (2003) Evaluating uncertainty in environmental life-cycle assessment. A case study comparing two insulation options for a Dutch one-family dwelling. *Environ Sci Technol* 37(11):2600–2608
- Kennedy D, Montgomery D, Quay B (1996) Data quality: Stochastic environmental life cycle assessment modeling. *Int J Life Cycle Assess* 1(4):199–207
- Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the science: keys and clues. *Biosci* 51(5):341–351
- 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
- MacLeod M, Fraser AJ, Mackay D (2002) Evaluating and expressing the propagation of uncertainty in chemical fate and bioaccumulation models. *Environ Toxicol Chem* 21(4):700–709
- Maurice B, Frischknecht R, Coelho-Schwartz V, Hungerbühler K (2000) Uncertainty analysis in life cycle inventory. Application to the production of electricity with French coal power plants. *J Clean Prod* 8(2):95–108
- ORACLE (2010) Oracle crystal ball for enterprise performance management. Fusion Edition Release 11(1):2.0.00
- Rousseaux P, Labouze E, Suh Y, Blanc I, Gaveglia V, Navarro A (2001) An overall assessment of life cycle inventory quality. *Int J Life Cycle Assess* 6(5):299–306
- Slob W (1994) Uncertainty analysis in multiplicative models. *Risk Anal* 14(4):571–576
- Sonnemann GW, Schuhmacher M, Castells F (2003) Uncertainty assessment by a Monte Carlo simulation in a life cycle inventory of

- electricity produced by a waste incinerator. *J Clean Prod* 11(3):279–292
- Tan RR, Culaba AB, Purvis MRI (2002) Application of possibility theory in the life-cycle inventory assessment of biofuels. *Int J Energ Res* 26(8):737–745
- Vermeire TG, Stevenson H, Pieters MN, Rennen M, Slob W, Hakkert BC (1998) Assessment factors for human health risk assessment: a discussion paper. The Netherlands: National Institute of Public Health and the Environment (RIMVN) and Netherlands Organization for Applied Scientific Research
- 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, Hischer R, Mutel C, Nemecek T, Reinhard J, Vadenbo CO, Wernet G (2013) Overview and methodology. Data quality guideline for the ecoinvent database version 3. Ecoinvent Report 1(v3). St. Gallen, The ecoinvent Centre