Int J Life Cycle Assess (2016) 21:1338–1348
DOI 10.1007/s11367-013-0670-5

THE ECOINVENT DATABASE V3

Empirically based uncertainty factors for the pedigree matrix in ecoinvent

Andreas Ciroth · Stéphanie Muller · Bo Weidema · Pascal Lesage

Received: 20 November 2012 / Accepted: 13 November 2013 / Published online: 20 December 2013

Abstract
Purpose Ecoinvent applies a method for estimation of default standard deviations for flow data from characteristics of these flows and the respective processes that are turned into uncertainty factors in a pedigree matrix, starting from qualitative assessments. The uncertainty factors are aggregated to the standard deviation. This approach allows calculating uncertainties for all flows in the ecoinvent database. In ecoinvent 2 the uncertainty factors were provided based on expert judgment, without (documented) empirical foundation. This paper presents (1) a procedure to obtain an empirical foundation for the uncertainty factors that are used in the pedigree approach and (2) a proposal for new uncertainty factors, received by applying the developed procedure. Both the factors and the procedure are a result of a first phase of an ecoinvent project to refine the pedigree matrix approach. A separate paper in the same edition, also the result of the aforementioned project, deals with extending the developed approach to other probability distributions than lognormal (Muller et al.).

Methods Uncertainty is defined here simply as geometric standard deviation (GSD) of intermediate and elementary exchanges at the unit process level. This fits to the lognormal probability distribution that is assumed as default in ecoinvent 2, and helps to overcome scaling effects in the analysed data. In order to provide the required empirical basis, a broad portfolio of data sources is analysed; it is especially important to consider sources outside of the ecoinvent database to avoid circular reasoning. The ecoinvent pedigree matrix from version 2 is taken as a starting point, skipping the indicator "sample size" since it will not be used in ecoinvent 3. This leads to a pedigree matrix with five data quality indicators, each having five score values. The analysis is conducted as follows: for each matrix indicator and for each data source, indicator scores are set in relation to data sets, building groups of data sets that represent the different data quality indicator scores in the pedigree matrix. The uncertainty in each of the groups is calculated. The uncertainty obtained for the group with the ideal indicator score is set as a reference, and uncertainties for the other groups are set in relation to this reference uncertainty. The obtained ratio will be different from 1, it represents the unexplained uncertainty, additional uncertainty due to a lower data quality, and can be directly used as uncertainty factor candidates.
Results and discussion The developed approach was able to derive empirically based uncertainty factor candidates for the pedigree matrix in ecoinvent. Uncertainty factors were obtained for all data quality indicators and for almost all indicator scores in the matrix. The factors are the result of the first analysis of several data sources, further analyses and discussions should be used to strengthen their empirical basis. As a consequence, the provided uncertainty factors can change in future. Finally, a few of the qualitative score descriptions in the pedigree matrix left room for interpretation, making their application not ambiguous.

Responsible editor: Rainer Zah

A. Ciroth (✉)
GreenDelta GmbH, Müllerstrasse 135, 13349 Berlin, Germany
e-mail: ciroth@greendelta.com

S. Muller · P. Lesage
CIRAIG, Department of Chemical Engineering, Polytechnique Montréal, P.O. Box 6079, Stn. Centre-ville, Montréal, QC H3C 3A7, Canada

B. Weidema
Ecoinvent Centre, Lerchenfeldstrasse 5, 9014 St-Gallen, Switzerland

B. Weidema
Aalborg University, Fredrik Bajers Vej 5, P.O. Box 159, 9100 Aalborg, Denmark

Conclusions and perspectives An empirical foundation for the uncertainty factors in the pedigree matrix overcomes one main argument against their use, which in turn strengthens the whole pedigree approach for quantitative uncertainty assessment in ecoinvent. This paper provides an approach to obtain an empirical basis for the uncertainty factors, and it provides also empirically based uncertainty factors, for indicator scores in the pedigree matrix. Basic uncertainty factors are not provided, it is recommended to use the factors from ecoinvent 2 for the time being. In the developed procedure, using GSD as the uncertainty measure is essential to overcome scaling effects; it should therefore also be used if the analysed data do not follow a lognormal distribution. As a consequence, uncertainty factors obtained as GSD ratios need to be translated to range estimators relevant for these other distributions. Formulas for this step are provided in a separate paper (Muller et al.). The work presented in this paper could be the starting point for a much broader study to provide a better basis for input uncertainty in LCA, not only in ecoinvent.

Keywords Data quality · Ecoinvent database · Empirical · Pedigree approach · Uncertainty · Uncertainty factors

1 Introduction

The pedigree matrix was introduced to uncertainty analyses by Funtowicz and Ravetz in 1990, as a means to code qualitative expert judgement for a set of problem-specific "pedigree criteria" into a numerical scale, incorporating criteria such as columns of the table, the numerical codes as table lines, and linguistic descriptions for each value in each cell of the table. The basic aim originates from qualitative descriptions of relevant aspects of an object of study with quantitative figures assessing these aspects. The matrix thus is a tool for "coding" qualitative assessment descriptions. Both rating scale and criteria shall be selected according to the needs of the object of study. There is no further formal requirement on the structure of the matrix. For example, van der Sluijs et al. (2003) present three different applications with indicator scores from 0 to 4 and 0 to 2, and with 4, 39, and 7 criteria.

Weidema and Wesnæs (1996) transferred the pedigree matrix to Life Cycle Assessments; their matrix is square, with a rating scale from 1 to 5 and with five criteria. In 1998 Weidema published a slightly modified version based on a multi-user test of the initial matrix (Weidema 1998). It became widely acknowledged and was modified by some authors (e.g. Ciroth et al. 2002). One important application example is the ecoinvent database (yet in a slightly modified form, Frischknecht et al. 2005).

Table 1 shows the pedigree matrix that is proposed in ecoinvent version 3.0, which largely reverts to the Weidema (1998) version.[1]

In the original literature, the pedigree matrix "produces" numerical codes from expert judgement (Funtowicz and Ravetz 1990); in ecoinvent, the codes are numerical values, as indicator scores, ranging from 1 to 5 for each of the indicators in the matrix. For calculating the overall uncertainty, the pedigree matrix results in ecoinvent are also not taken directly, but after a transformation using the following Table 2—the values in this transformation table never exceed 2, and are mostly below 1.5. For ecoinvent 3.0, the same values are applied.

2 Methods

2.1 Uncertainty

Uncertainty in LCA can be defined in various ways. For our paper, we are interested in "translating" uncertainty that is specified on an ordinary scale, in the pedigree matrix, into quantitative uncertainty. In this context, uncertainty will be understood as follows:

Uncertainty means, basically, lack of certainty. The lack of certainty depends on the level of detail that is taken into account. Let us look at an LCA-related example, the amount of fertiliser used by farmers. With data sets for several farmers, and potentially also over a certain time interval, the amount will vary, and the exact amount used in a specific farm will be known precisely. The amount of fertiliser used is uncertain. This uncertainty will be lower, if we know in addition:

- the time interval covered
- the size of the farms
- the type of farm, their products
- the geographical area where the farm is located
- the (micro-)climate where the farm is located
- the management type of the farm (e.g. organic farming.)
- the farming background and expertise of the farmers
- etc.

The uncertainty in the amount of fertilizer thus can in part be "explained" by those details (the parameters listed above). This links directly to the concept of "explained variation" or "explained variance" in statistics (Kent 1983). The more

[1] This version is different from the version that was in use in ecoinvent 2.0 and 2.1 (Frischknecht et al. 2004, p. 45)—in the old version, several scores were not used, for example 2 for 'technological correlation', and the properties of aspects (the entries in the cells) were sometimes worded differently, and a sixth criteria "sample size" was introduced, which is now removed again, with the argument that the influence of the sample size is already included in the basic uncertainty.

Table 1 Ecoinvent 3.0 pedigree matrix

Indicator score	1	2	3	4	5 (default)
Reliability	Verified[a] data based on measurements[b]	Verified data partly based on assumptions *or* non-verified data based on measurements	Non-verified data partly based on qualified estimates	Qualified estimate (e.g. by industrial expert)	Non-qualified estimate
Completeness	Representative data from all sited relevant for the market considered, over an adequate period even out normal fluctuations	Representative data from >50 % of the sites relevant for the market considered, over an adequate period to even out normal fluctuations	Representative data from only some sited (≪50 %) relevant for the market considered *or* >50 % of sites but from shorter periods	Representative data from only one site relevant for the market considered *or* some sites but from shorter periods	Representativeness unknown or data from a small number of sites *and* from shorter periods
Temporal correlation	Less than 3 years of difference to the time period of the dataset	Less than 6 years of difference of the time period of the dataset	Less than 10 years of difference to the time period of the dataset	Less than 15 years of difference to the time period of the dataset	Age of data unknown or more than 15 years of difference to the time period of the dataset
Geographical correlation	Data from area under study	Average data from larger area in which the area under study is included	Data from area with similar production conditions	Data from area with slightly similar production conditions	Data from unknown *or* distinctly different area (North America instead of Middle East, OECD-Europe instead of Russia)
Further technological correlation	Data from enterprises, processes and materials under study	Data from processes and materials under study (i.e. identical technology) but from different enterprises	Data from processes and materials under study from different technology	Data on related processes or materials	Data on related processes on laboratory scale *or* from different technology

[a] Verification may take place in several ways, e.g. by on-site checking, by recalculation, through mass balances or cross-checks with other sources

[b] Includes calculated data (e.g. emissions calculated from inputs to an activity), when the basis for calculation is measurements (e.g. measured inputs). If the calculation is based partly on assumptions, the score would be 2 or 3.

Table 2 "Default uncertainty factors (contributing to the square of the geometric standard deviation) applied together with the pedigree matrix" (Frischknecht et al. 2004, p 46)

Indicator score	1	2	3	4	5
Reliability	1.00	1.05	1.10	1.20	1.50
Completeness	1.00	1.02	1.05	1.10	1.20
Temporal correlation	1.00	1.03	1.10	1.20	1.50
Geographical correlation	1.00	1.01	1.02		1.10
Further technological correlation	1.00		1.20	1.50	2.00
Sample size	1.00	1.02	1.05	1.10	1.20

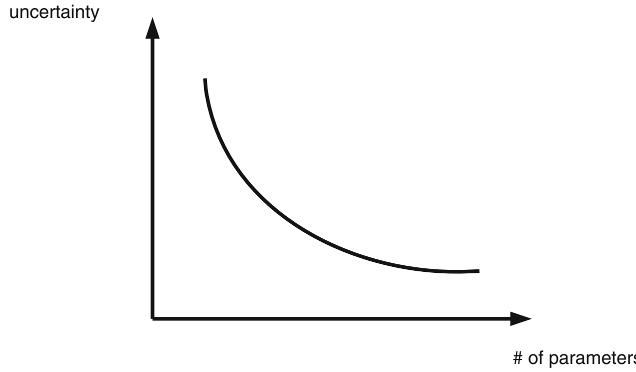


Fig. 1 General relation between uncertainty and the number of known parameters: the more parameters are known, the lower the uncertainty

parameters are precisely known, the lower the uncertainty (Fig. 1).

This understanding of uncertainty is a bit broader than the distinction into systematic and random uncertainty, or systematic and random errors, that is quite common in LCA (e.g. Ciroth et al. 2004): Depending on the level of detail and knowledge, a certain deviation from a true value is either explained by additional parameters or not, and it is therefore either a systematic deviation (if the parameters responsible for the deviation are known) or not.

2.2 What does "empirical" mean?

Empirical is defined here as "derived from experiment and observation rather than theory and expert guesses", expanding thereby a definition given by the Princeton Wordnet database.[2]

Own experiments were not possible during this project; the aim is therefore to compare data to measurements available from literature or from own sources, where possible. Any parameters used in these measurements will need to be considered in this comparison, as explained in the uncertainty section, 2.1.

2.3 Analysed data sources

As a summary, the following sources were analysed, for the different indicators in the pedigree matrix:

Reliability: the German GEMIS database (www.gemis.de) and their investigation in a "validation" project (Ciroth 2009a, b; Ciroth and Weidema 2009), non-LCA sources (EPER 2010), measurement data (Lundie et al. 2004).

Completeness: sources about the representativeness of LCA data (and of related data outside of the LCA domain), e.g. again, (Ciroth and Srocka 2008), investigations about representative means of transport and energy systems (TREMOD 2010; ZSE 2010).

Temporal correlation: Emission inventories, such as the German ZSE system, eurostat and US statistics (Census 2010; Eurostat 2010), and also in part the PRTR system (PRTR 2010), have data sets over several years that allow time series analyses that were taken into account.[3] Several national statistics, including the North American Transport statistic, were also considered.

Geographical correlation: Comparison of transport emissions of the same or very similar transportation vehicles from different regions (Tremod database and again the North American Transport database); differences in electrical grids for different regions, in different databases.

Further technological correlation: Solar cell comparison from the GEMIS database as available in ProBas (2010), for transport data sets from the Tremod database and from the GREET model (GREET 2009).

Especially for geographical correlation, correlations (in the statistical sense) with other attributes need to be considered; the factor used for geographical correlation should only reflect those aspects that are indeed caused by geographical differences. Little influence of geography is expected by specifically described technical processes (emissions of a car with Euro4 emission category, for example, will barely depend on where it is operated). Differences will rather occur due to different technologies that are used and not specified, or a different geographical background—sulphur content in coal—that is not specified. Higher influence is therefore expected for average processes (average emissions for heavy truck transport, etc.). The uncertainty is applied at the level of individual exchanges, and therefore further uncertainty on aggregated process levels e.g. can (and should) be left out of consideration.

[2] Wordnet defines empirical as "derived from experiment and observation rather than theory", http://wordnetweb.princeton.edu/perl/webwn?s=empirical.

[3] See, e.g. http://www.epa.gov/ttn/chief/conference/ei11/datamgt/doring.pdf for an analysis of German ZSE data in this respect.

Figure 2 shows the different data sources that were used for analysing the pedigree indicator scores.

2.4 Dealing with scaling effects in data

For the computation of data ranges and for characterising the "spread" in data values, the standard deviation or the variance are often used. The standard deviation is the square root of the variance; it is the parameter in the normal probability distribution that describes the spread in underlying data, and it is commonly used in random error analyses.

For the analysis of uncertainty, the standard deviation seems therefore an ideal candidate. It has, however, the disadvantage of depending on the scale of data, in a linear manner. Recall that for the variance Var(X) holds, with X being a random variable, and a and b being constants:

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

For the standard deviation SD holds, respectively:

$$\mathrm{SD}(aX + b) = a\mathrm{SD}(X)$$

This means that a constant factor that is applied to all analysed data values changes the standard deviation by the same factor. This may happen if, for example, data is given in gramme instead of kilogramme; all values will be multiplied by a constant factor of 1,000, and the resulting standard deviation will also increase by a factor of 1,000. This is of course undesirable for the development of generic uncertainty factors by analyzing data from different sources, since data may have different units; it is also undesirable for the later application of the factors, which should as best as possible be independent from the measurement units of data. At best, the developed factors should represent only the spread in data, independent from the scale of data in one sample. The factors should therefore be independent from scaling effects.

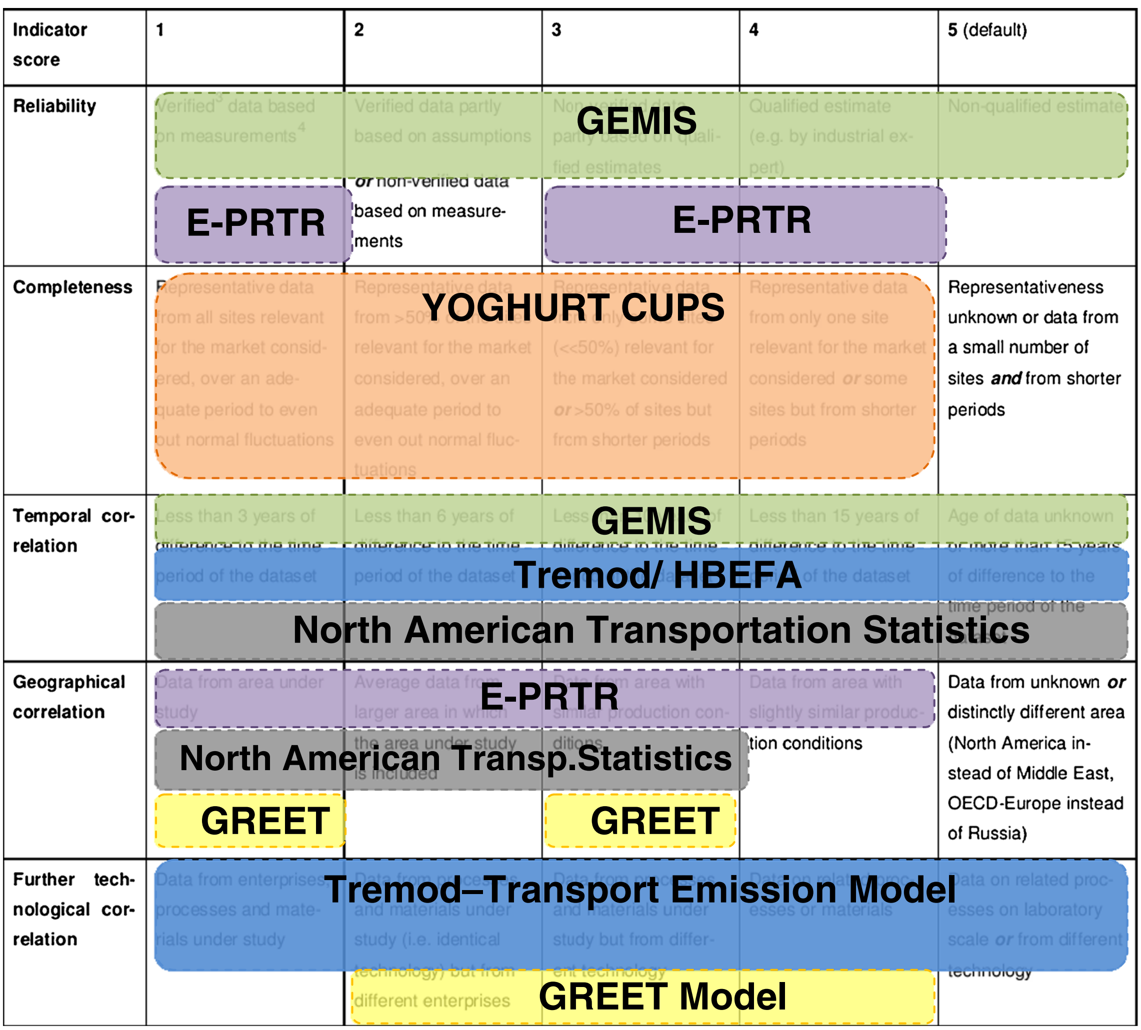


Fig. 2 Data sources taken into account for the different pedigree matrix indicator scores

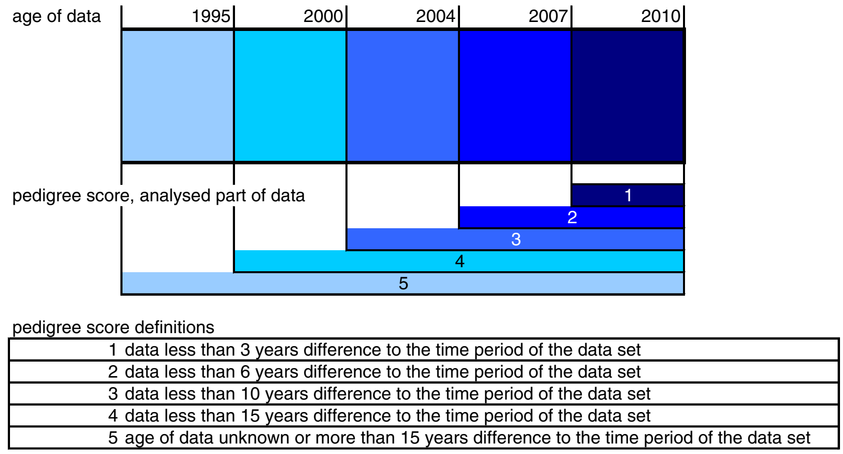


Fig. 3 Data groups for a data set that is supposed to relate to 2010

The current project investigated data sets that contain amounts of input and output flows for technical processes, similar or identical to data sets that are used for life cycle inventories. There are three main reasons for scaling effects in these data sets:

1. data may not be provided per quantitative reference, i.e. not per product unit; this requires a transformation of the data, for example from absolute emission figures of an industrial plant to "per kilogramme product" emission figures
2. if a quantitative reference is given, it may differ from one data set to another (1,000 m^2 for one data source or group of data; 1 m^2 for another)
3. data may simply be provided in different units (kilogramme emissions vs. emissions in grams)

In order to overcome the scale dependency, the geometrical standard deviation (GSD) is used as a measure for uncertainty. It is defined as

$$GSD(x) = \exp\left(\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n} ln\left(\frac{x_i}{\bar{x}_g}\right)^2}\right)$$

with $\bar{x}_g = \sqrt[n]{\Pi_{i=1}^{n} x_i}$, the geometric mean of x

GSD has the convenient property that linear factors in data "disappear":

GSD $(aX+b)$=GSD (X) with a, b being two constants. Therefore, by using GSD as an indicator for uncertainty, many reasons for scaling effects in the data can be overcome.

2.5 Building on the uncertainty factor concept in ecoinvent 2

Ecoinvent version 2 assumes a lognormal probability distribution for all uncertain values. Default basic and additional uncertainty factors are combined to an overall, total uncertainty using the following formula (Frischknecht et al. 2005):

$$U_T = \exp\left(\sqrt{(\ln U_b)^2 + \sum_i (ln U_i)^2}\right) \quad (1)$$

where U_b and U_i are the basic and additional uncertainty factors, respectively, and U_T (SD_{g95} following the Frischknecht et al. notation (2005) is the total uncertainty, all expressed as the square of a geometric standard deviation (GSD^2).

Since the geometric standard deviation is one of the two parameters of the lognormal probability distribution, the calculated GSD can, in ecoinvent 2, be directly used to build the probability distribution and to calculate confidence intervals.

The newly developed uncertainty factors build on this concept, even though data does not need to follow a lognormal distribution.

Also in the new approach, Eq. 1 will be used to determine the total uncertainty of an uncertain flow, in order to overcome scaling effects in the data as well as possible. If the underlying data does not follow a lognormal distribution, range estimates will be obtained by transforming the GSDs to the respective range estimates for these other distributions. This will be explained in more detail in the article by Muller et al. in the same Journal issue.

Table 3 Values and respective scores for the indicator temporal correlation in the pedigree matrix

Indicator value: years of difference to desired time	<3	<6	<10	<15	Unknown
Indicator score	1	2	3	4	5

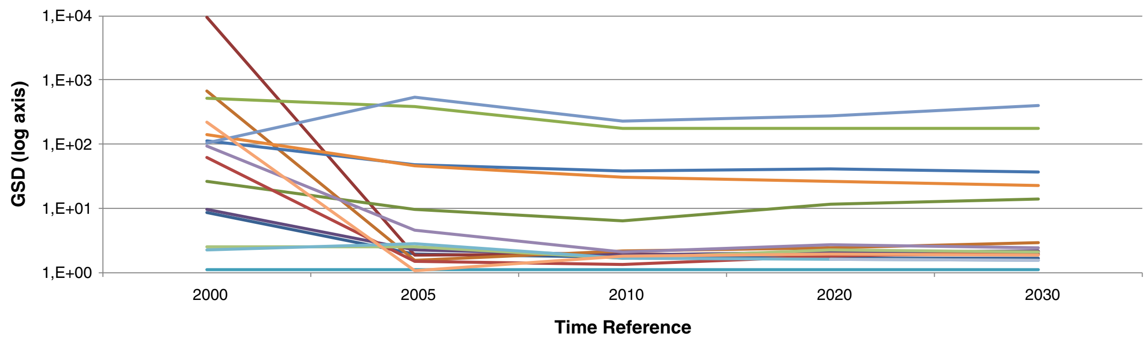


Fig. 4 Calculated GSD for all data sets in the GEMIS database with a product in energy units, over time. Each *line* represents one NACE code

2.6 Approach for obtaining the uncertainty factors

From the analysed data sources, the uncertainty factors are obtained as follows:

- Each uncertainty factor is analysed separately.
- For each analysis, one or several data sources are considered. Each source is analysed independently. The geometric standard deviation for the data sample is calculated, for several groups of data.
- The data groups are formed according to the specification in the pedigree matrix. An ideal data group is identified in the data sample. This group satisfies all criteria for the best indicator score 1 in the pedigree matrix. For example, for the "reliability" indicator, data must be verified based on measurements. Then, further groups are defined that are less ideal, according to the pedigree matrix, and the geometric standard deviation is calculated for each group. The GSD of the ideal group will not necessarily be 1, due to effects not related to the specific uncertainty factor (think of the initial definition that uncertainty is what is not known or ignored). It can even be larger than the GSD for another indicator value. The GSD for the other groups is then set in relation to the ideal GSD of the ideal group, so that the resulting ratio is equal or larger than 1. This ratio is the additional uncertainty in the less ideal group:

$$U_i = \begin{cases} \dfrac{\mathrm{GSD}_{i,j}}{\mathrm{GSD}_{i,1}}, & \mathrm{GSD}_{i,1} < \mathrm{GSD}_{i,j} \\ \dfrac{\mathrm{GSD}_{i,1}}{\mathrm{GSD}_{i,j}}, & \mathrm{GSD}_{i,1} \geq \mathrm{GSD}_{i,j} \end{cases}, U_i \geq 1$$

with:

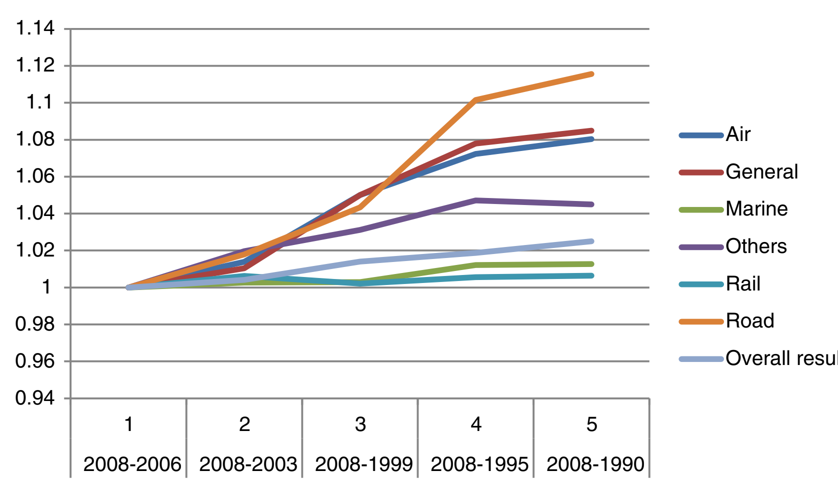
U_i uncertainty factor for pedigree indicator *i*

$\mathrm{GSD}_{i,j}$ GSD for pedigree indicator i for pedigree matrix score $j, j \in [1; 5]$

$\mathrm{GSD}_{i,1}$ GSD for pedigree indicator i for pedigree matrix score 1 (ideal case)

Following this formula, the uncertainty factor for the ideal data group with indicator score 1 always will be 1 even if the

Fig. 5 Calculated GSD for data set groups and the pedigree indicator values, in the NATS database, by mode of transport and overall



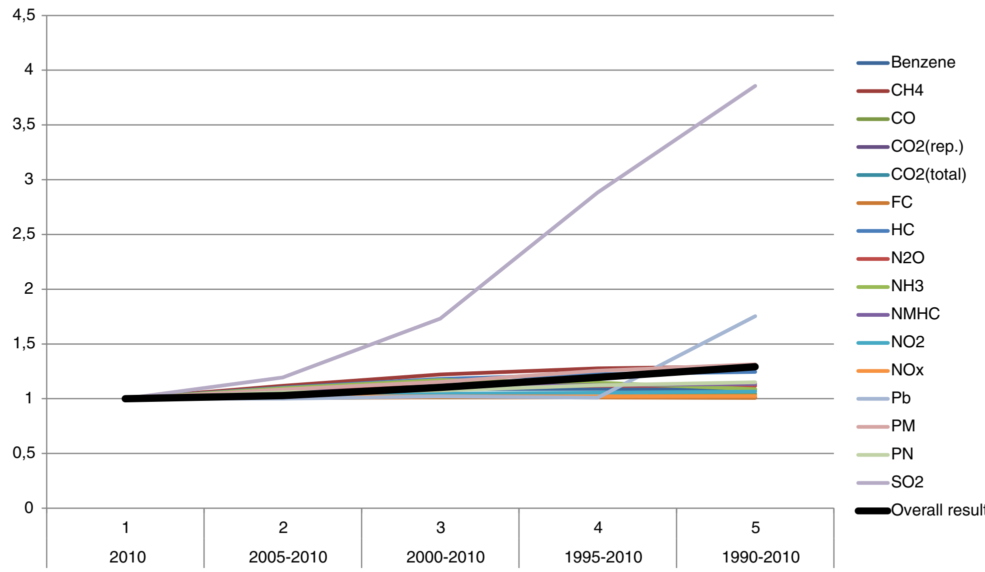


Fig. 6 Calculated GSD for data set groups and the pedigree indicator values, in the TREMOD database, by flow emission and overall

calculated ideal GSD might be different from 1. This ratio can be seen as the uncertainty contribution of the less ideal indicator value.

Let us illustrate this with an example, for the pedigree indicator "temporal correlation". Assume we want to have a data set for the year 2010 and have a group of data sets from a broad range of time periods, the newest from 2010, but otherwise comparable. The ideal group is then formed by all data sets younger than 2007 (score 1, "less than 3 years of difference between the time periods of the data sets). The second best group is formed by all data sets from 2004 until 2010 (less than 6 years difference between the time periods of the data sets), and so forth (Fig. 3).

For the ideal group with score 1 and for the other groups, the GSD values are calculated, and the ratio of group-GSD $GSD_{temp\ j}$ to ideal $GSD_{temp\ 1}$ is the uncertainty factor for temporal correlation and the respective indicator score. For example, let's say we obtain for score 4 a GSD of $GSD_{temp\ 4}$= 3.2 and for score 1 a $GSD_{temp\ 1}$=1.3. Then, an "uncertainty factor candidate" for temporal correlation of score 4 is $\frac{3.2}{1.3} = 2.46$. Different similar analyses from different data sources provide different candidate values that are then used to determine a final proposal for the uncertainty factor.

3 Results and discussion

Results will not be presented for all the analysed indicators nor for all the analysed data sources. These can be found in the project report (Ciroth et al. 2012). As an example, the indicator temporal correlation will be looked at in more detail. At the end of this section, the full list of the proposed uncertainty factors will be provided.

3.1 The analysis of the indicator temporal correlation as an example

As also shown in Table 1, the pedigree scores for the indicator temporal correlation are as follows (Table 3):

For this indicator, different data sources were analysed, see also section 2.3. Results for the following sources will be shown here:

- GEMIS, a German, free life cycle database that contains prognostic data up until 2030 (GEMIS 2010)
- The North American Transport statistics (NATS) database with annual data from 1990 until 2008 (NATS 2011)
- The TREMOD database, a European transport emission database with transport data sets from 1990 until 2010 (TREMOD 2010).

GEMIS contains very few data sets older than 2000, therefore the analyses were performed with prognostic data only, covering a time span from 2000 until 2030. Also these prognostic data were not useful for the analysis. Differences in data uncertainty were caused by other reasons than time; for a time difference of 30 years, the GSD remains almost the same, if it is calculated within one industrial sector. Figure 4 shows the result of the calculated GSD per NACE code,[4] restricted to processes with a product in energy units to narrow down the side effects due to different technology and so forth.

[4] NACE is the Statistical Classification of Economic Activities in the European Community coding system (in French: Nomenclature statistique des activités économiques dans la Communauté européenne), NACE (2010)

Table 4 Tentative uncertainty factors for the indicator temporal correlation in the pedigree matrix, as GSD

Indicator value	Uncertainty factor
1	1
2	1.03
3	1.10
4	1.19
5	1.29

A possible reason for the almost constant GSD from 2010 onwards is the forecast models that are used in GEMIS. They do not seem to reflect true changes related to time. For this reason, GEMIS is excluded from the further analyses concerning the time indicator.

The NATS database can be grouped into five different groups; since emissions have changed differently over recent years per transport mode (ship, rail, road etc.), the analysis is done per transport mode; a total value is calculated as well, for all data sets.

Figure 5 shows the result, with the year 2008 as a reference.

The figure shows that rail and marine emissions have changed less than road and air emissions over the years.

For the TREMOD database, a similar analysis is conducted. Figure 6 shows the results, again for five groups of data sets; this time, results are provided per emission flow. While many pollutants are more or less stable over time, two, lead (Pb) and sulphur dioxide (SO_2), show very high uncertainty with a growing time span. This can be explained by regulatory measures that took place from 2000 onwards in Europe for leaded fuels (due to the emergence of catalysts) and regarding the sulphur content in fuels (EU 1998).

For the overall results of the temporal correlation factor, changes in technology should not be considered since these will be dealt with by the technological correlation score. Therefore, the TREMOD data are more meaningful than the NATS database. This leads to the following results for the indicator temporal correlation (Table 4):

Table 6 Summary of tentative uncertainty factors for all pedigree matrix indicators, as GSD

Indicator score	1	2	3	4	5
Reliability	1	1.54[a]	1.61	1.69	(n.a.)
Completeness	1	1.03	1.04	1.08	(n.a.)
Temporal correlation	1	1.03	1.10	1.19	1.29
Geographical correlation	1	1.04	1.08	1.11	(n.a.)
Further technological correlation	1	1.18	1.65	2.08	2.80

[a] Interim

3.2 Remarks on the analyses of other pedigree indicators

For the other indicators, similar analyses were performed. Details can be found in the project report. Some specific remarks about the analyses and about the indicators could also be interesting here, however.

Reliability of the data source For this indicator, some sources showed a decrease in uncertainty when moving from 3 to 4, i.e. from non-verified data to qualified estimates. This even seems reasonable, one would rather estimate a transport distance as 50 km than 69.4 km, therefore estimated samples might have fewer spread. For the indicator score 5, unqualified estimate, a very high uncertainty factor candidate was obtained (higher than 50) which was excluded from the final uncertainty factor result list.

The *completeness* indicator could be well analysed, based on a fully empirical study. For the indicator score 5, however ("representativity unknown"), no analysis data was available, or, taking any other possibly not representative data, would also lead to a very high uncertainty factor here.

Geographical correlation Also for this indicator, the worst score of 5, "data from distinctly different area", could not really be analysed, or would have led to an extremely high uncertainty factor. For the analysis, sector specific databases were used or databases were investigated per sector,

Table 5 Mapping of indicator scores for "technological correlation" to differences in data sets in the TREMOD database

Indicator score	Meaning of the indicator score	Differences in data sets relevant for this indicator score
1	Data from enterprises, processes and materials under study	Personal car, EURO 4 emission type, 1.4–2-l capacity, inner city use, diesel
2	Data from processes and materials under study (i.e. identical technology) but from different enterprises	For personal car: different use, inner city use vs. other use types
3	Data from processes and materials under study but from different technology	For personal car: different size (0–1.4 l, 2–9 l), different emission category (EURO 1, 2, 3 and 5 in addition to 4)
4	Data on related processes or materials	For personal car: also old cars (pre Euro 1)
5	Data on related processes on laboratory scale or from different technology	For personal car: different fuel (gasoline)

nevertheless some correlations with technology might exist that deserve a further analysis in order to arrive at more specific uncertainty factors, e.g. per industrial sector.

Further technological correlation For this indicator, the definition of different indicator scores needed to be developed. For example, for the (transport-related) TREMOD database, the indicator scores were understood as shown in Table 5. Ideal data set here is a Personal car, EURO 4 emission type, 1.4–2-l capacity, for inner city use, with diesel fuel.

3.3 Summary of the proposed new uncertainty factors for pedigree indicators

Overall, the following generic uncertainty factors are recommended as a result of the analyses (Table 6).

For reliability, results for "measurements partly based on assumptions" (score 2) were obtained from a rather small data sample, therefore the uncertainty factor must be regarded as interim. For several indicators, no (meaningful) analyses could be performed for the worst data quality, score 5, or obtained results showed extremely high uncertainty factors which seem not really useful. Therefore for these scores, the uncertainty factors are considered as unavailable.

4 Conclusions

It was indeed possible to obtain uncertainty factors based on empirical data for the pedigree matrix concept in ecoinvent, for the first time. With these factors, the whole generic uncertainty assessment in ecoinvent is put on a better founded basis. The identified factors are different, but not very different, to ecoinvent factors previously used in ecoinvent 2. They can already now be used for calculating generic uncertainty based on a qualitative assessment of data quality.

However, several aspects deserve further attention. As they are, the uncertainty factors represent a geometric standard deviation which can be used to calculate confidence intervals for data that follows a lognormal probability distribution. For other distributions, however, the uncertainty factors need, firstly, to be "migrated". This is explained in more detail in the paper by Muller et al. in this issue of the journal.

Not for all indicator scores are now factors available, especially for some extreme values, factors are lacking. It is recommended to decide about suitable uncertainty factors for these cases on a case by case basis. More investigations would obviously be useful here to determine suitable factors, and/or to revise the pedigree matrix concept for these, since it will probably always be difficult to obtain meaningful uncertainty factors for, e.g. unqualified estimates. In the meantime, an expert guess for the so far uncovered indicator scores should be provided, in coordination with the specific database, i.e. ecoinvent, in order to allow the application of the uncertainty factors for the database.

The analyses show clearly that the uncertainty factors are generic and should be used as default values, however they are not suitable for the consideration of specific effects. For example, for temporal correlation, two specific emissions, lead and sulphur dioxide had a very different uncertainty due to emission regulations over the years. In that sense, case specific uncertainty analyses are preferable to using the generic factors.

A more detailed analysis therefore might make sense. Such an analysis could recognize different types of processes and industrial sectors, per pedigree indicator, to distinguish for example marine transport from personal car transport, over time, broadening the information basis used in this initial project, identifying possibly also more, or less, representative empirical data sources.

But finally, it is also worth applying these newly developed factors, in combination with case specific uncertainty factors.

References

Census (2010) US statistics and census office, http://www.census.gov/. Accessed 7 Oct 2010

Ciroth A (2009a) Cost data quality considerations for eco-efficiency measures. Ecol Econ 68(6):1583–1590

Ciroth A (2009b) Validierung der Emissionsfaktoren ausgewählter erneuerbarer Energiebereitstellungsketten, Endbericht (final report), commissioned by Umweltbundesamt, Berlin Dessau 2009, www.greendelta.com/Emissionsfaktoren-erneuerbarer.97.0.html?&L=1. Accessed 2 Nov 2011

Ciroth A, Srocka M (2008) How to obtain a precise and representative estimate for parameters in LCA: a case study for the functional unit. Int J Life Cycle Assess 13(3):265–277

Ciroth A, Weidema B (2009) Mathematical analysis of the ecoinvent LCI database with the purpose of developing new validation tools for the database, presentation, Boston LCA IX, October 2009

Ciroth A, Hagelüken M, Sonnemann GW, Castells F, Fleischer G (2002) Geographical and technological differences in life cycle inventories shown by the use of process models for waste incinerators. Part I: technological and geographical differences. Int J Life Cycle Assess 7(5):295–300, Part II: Int J Life Cycle Assess 7(6):363-368

Ciroth A, Fleischer G, Steinbach J (2004) Uncertainty calculation in life cycle assessments—a combined model of simulation and approximation. Int J Life Cycle Assess 9(4):216–226

Ciroth A, Muller St, Weidema B (2012) Refining the pedigree matrix approach in ecoinvent, project report for ecoinvent, May 2012, final draft report

EPER (2010) European pollutant emission register. www.eper.ec.europa.eu/. Accessed 7 Oct 2010

EU (1998) Directive 98/70/EC of the European Parliament and of the Council of 13 October 1998 relating to the quality of petrol and diesel fuels and amending Council Directive 93/12/EEC, Official Journal L 350, 28/12/1998 P. 0058–0068

Eurostat (2010) Eurostat central database, http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database. Accessed 7 Oct 2010

Frischknecht R, Jungbluth N, Althaus H-J, Doka G, Dones R, Hischier R, Hellweg S, Nemecek T, Rebitzer G, Spielmann M (2004) Overview and Methodology. Final report ecoinvent 2000 No. 1, Swiss Centre for Life Cycle Inventories, Dübendorf, Switzerland

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 JR (1990) Uncertainty and quality in science for policy. Kluwer, Dordrecht

GEMIS (2010) GEMIS, Global Emission Model for Integrated Systems, version 4.6, www.gemis.de. Accessed 7 Oct 2010

GREET (2009) http://www.transportation.anl.gov/modeling_simulation/GREET/index.html, GREET model fuel cycle, version 1.8c.0, 2009. Accessed 2 Nov 2011

Kent JT (1983) Information gain and a general measure of correlation. Biometrika 70(1):163–173

Lundie S et al (2004) Australian dairy farm data, 2004, unpublished

NACE (2010) http://ec.europa.eu/competition/mergers/cases/index/nace_all.html. Accessed 7 Oct 2010

NATS (2011) North American Transportation Statistics database, http://nats.sct.gob.mx/. Accessed 2 Nov 2011

ProBas (2010) ProBas database of the German EPA, www.probas.umweltbundesamt.de. Accessed 7 Oct 2010

PRTR (2010) The European Pollutant Release and Transfer Register, http://prtr.ec.europa.eu/. Accessed 7 Oct 2010

TREMOD (2010) Transport Emission Model, www.ifeu.de/english/index.php?bereich=ver&seite=projekt_tremod, data access via www.probas.umweltbundesamt.de/php/themen.php?id=12884901888&step=2&. Accessed 7 Oct 2010

van der Sluijs JR, Kloprogge PJ, Risbey J, Ravetz J (2003) Towards a synthesis of qualitative and quantitative uncertainty assessment: application of the numeral, unit, spread, assessment, pedigree (NUSAP) system, International Workshop on Uncertainty, Sensitivity, and Parameter Estimation for Multimedia Environmental Modeling, Rockville, USA

Weidema BP (1998) Multi-user test of the data quality matrix for product life cycle inventory data. Int J Life Cycle Assess 3(5):259–265

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

ZSE (2010) German CO_2 Emission trading institution, http://www.dehst.de/. Accessed 7 Oct 2010