Refining the pedigree matrix approach in ecoinvent

Andreas Ciroth

With contributions from Stéphanie Muller and Bo Weidema

May 2012

Version 7.1

ciroth@greendeltatc.com

Index

1 Aim & Objective

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 in a formula that is valid for lognormally distributed data only.

This report covers phase 0 of a three-phased project that will in detail refine this “pedigree approach” in ecoinvent, aiming to put it on an empirically better founded basis.

The phase 0 will work entirely with the existing approach and will

(1) derive empirically based, reasonable values for the uncertainty factors used in this approach, and

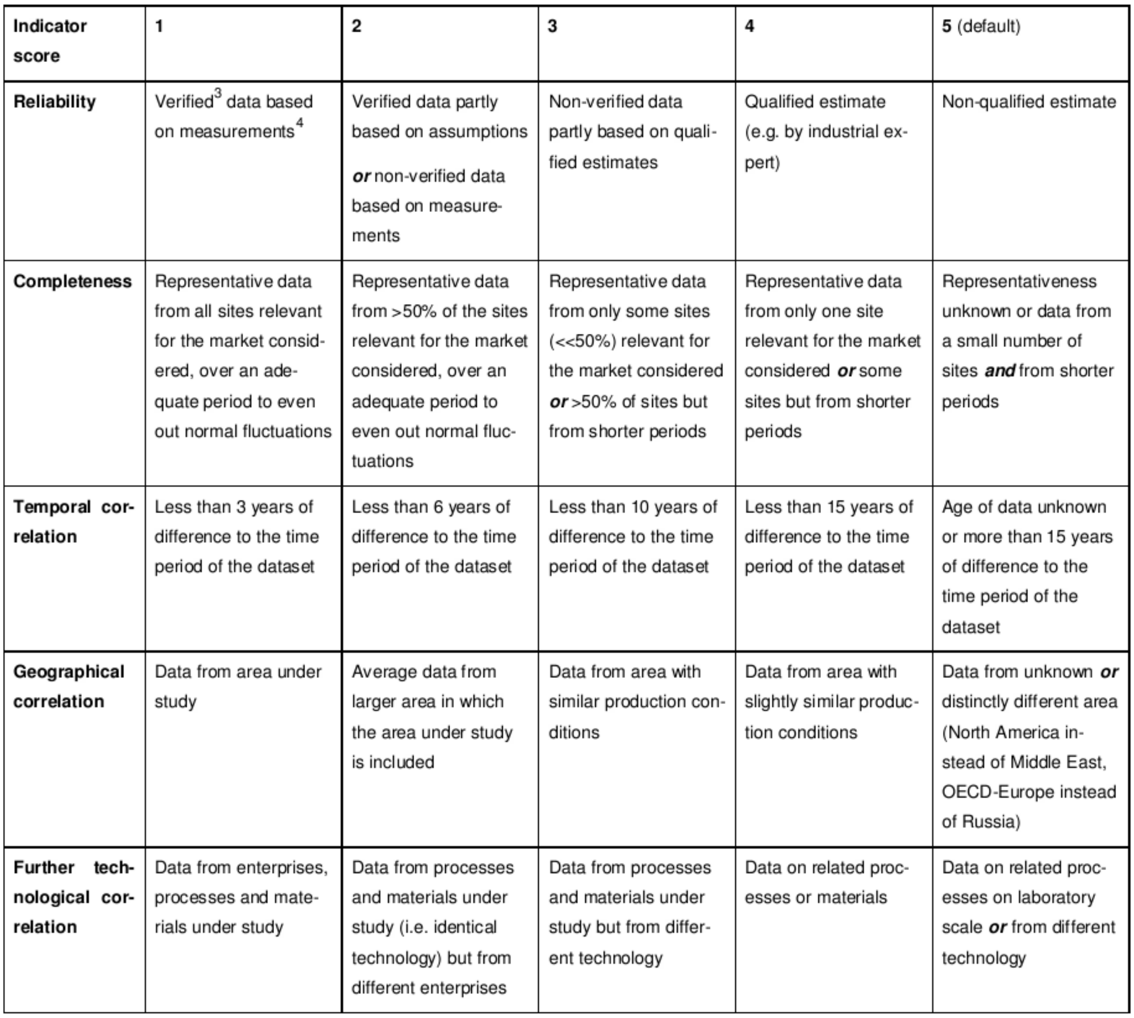
(2) provide practical considerations on how to apply the approach to other distributions.

2 Background

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, with criteria as columns of the table, the numerical codes as table lines, and linguistic descriptions for each value in each cell of the table. Basic aim is to come from qualitative description of relevant aspects of an object of study to quantitative figures assessing these aspects. The matrix thus is a tool for quantification of 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, 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 5 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. One important application example is the ecoinvent database (yet in a slightly modified form, Frischknecht (2005)).

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

[1] This version is different from the version that was in use in ecoinvent 2.0 and 2.1 (Frischknecht, Jungbluth 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, see ecoinvent 3.0 Draft Data Quality Guidelines v.0.14, p.75, with the argument that the influence of the sample size is already included in the basic uncertainty.



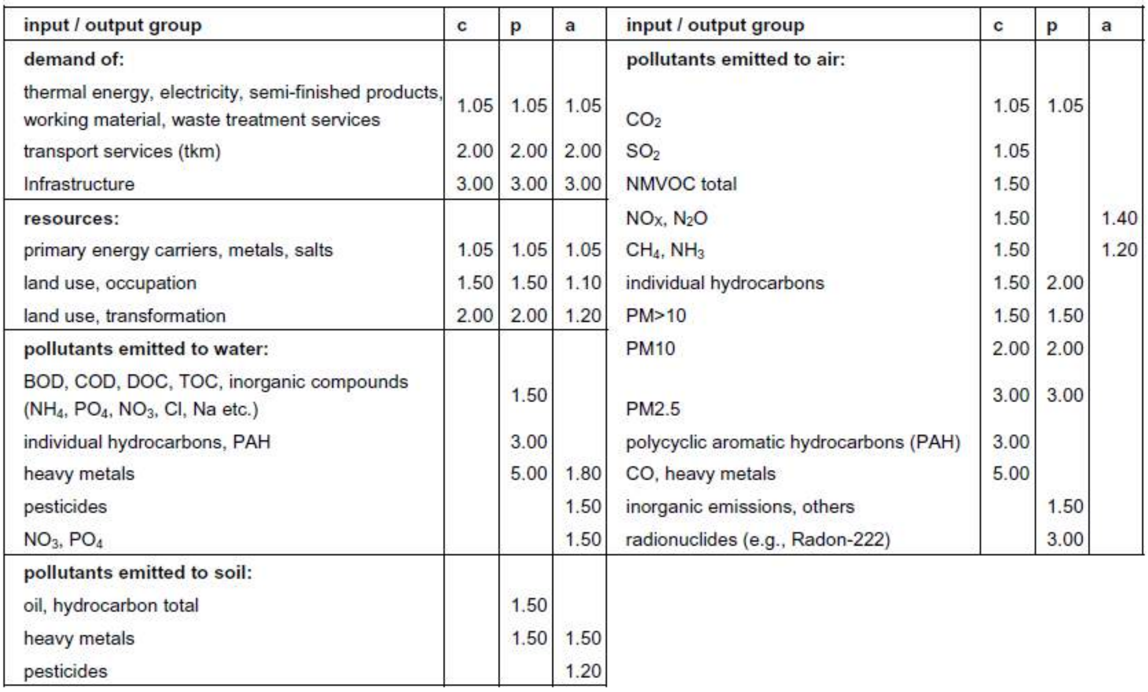
Indicator score	**1**	**2**	**3**	**4**	**5** (default)
Reliability	Verified[3] data based on measurements[4]	Verified data partly based on assumptions ***or*** non-verified data based on measurements	Non-verified data partly based on qualified estimates	Qualified estimate (e.g. by industrial expert)	Non-qualified estimate
Completeness	Representative data from all sites relevant for the market considered, over an adequate period to 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 sites (<<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 to 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 but from different technology	Data on related processes or materials	Data on related processes on laboratory scale ***or*** from different technology

Figure 1: ecoinvent 3.0 pedigree matrix

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

[4] 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.

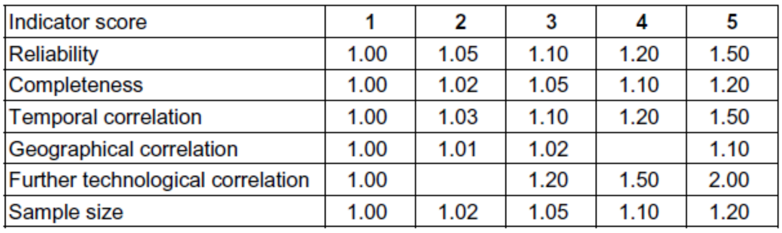
Depending on the type of exchange and the input or output “pathway”, figures for input and output data of flows will differ in their uncertainty. In order to take this into account, a basic uncertainty is attributed, again based on expert judgements, following the table shown in Figure 2. In contrast to the pedigree matrix, numerical values in this table are uncertainty factors (and not just numerical “scores”, as in the pedigree matrix). Figure 2 shows the uncertainty factors that are applied per type of exchange (working material, heavy metals, …) and per type of emission path (c, p, a for combustion, process, and agricultural emissions, respectively).



input / output group	c	p	a	input / output group	c	p	a
demand of:				**pollutants emitted to air:**			
thermal energy, electricity, semi-finished products, working material, waste treatment services	1.05	1.05	1.05	CO_2	1.05	1.05	
transport services (tkm)	2.00	2.00	2.00	SO_2	1.05		
Infrastructure	3.00	3.00	3.00	NMVOC total	1.50		
resources:				NO_X, N_2O	1.50		1.40
primary energy carriers, metals, salts	1.05	1.05	1.05	CH_4, NH_3	1.50		1.20
land use, occupation	1.50	1.50	1.10	individual hydrocarbons	1.50	2.00	
land use, transformation	2.00	2.00	1.20	PM>10	1.50	1.50	
pollutants emitted to water:				PM10	2.00	2.00	
BOD, COD, DOC, TOC, inorganic compounds (NH_4, PO_4, NO_3, Cl, Na etc.)		1.50		PM2.5	3.00	3.00	
individual hydrocarbons, PAH		3.00		polycyclic aromatic hydrocarbons (PAH)	3.00		
heavy metals		5.00	1.80	CO, heavy metals	5.00		
pesticides			1.50	inorganic emissions, others		1.50	
NO_3, PO_4			1.50	radionuclides (e.g., Radon-222)		3.00	
pollutants emitted to soil:							
oil, hydrocarbon total		1.50					
heavy metals		1.50	1.50				
pesticides			1.20				

Figure 2: Basic uncertainty factors per type of flow and per type of emission "pathway", in ecoinvent 2.0 (Frischknecht, Jungbluth 2004 p 44)

In the original literature, the pedigree matrix "produces" numerical values from expert judgement (Funtowicz and Ravetz 1990); in ecoinvent, the numerical values are the indicator scores 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 (Figure 3) – the values in this transformation table never exceed 2, and are mostly below 1.5. For ecoinvent 3.0, the same values are applied.



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

Figure 3: "Default uncertainty factors (contributing to the square of the geometric standard deviation) applied together with the pedigree matrix", (Frischknecht, Jungbluth 2004 p 46)

In order to combine both the pedigree matrix uncertainty and the basic uncertainty, the following formula is used (Frischknecht, Jungbluth 2004 p 44):

$$SD_{g95} := \sigma_g{}^2 = \exp^{\sqrt{[\ln(U_1)]^2+[\ln(U_2)]^2+[\ln(U_3)]^2+[\ln(U_4)]^2+[\ln(U_5)]^2+[\ln(U_6)]^2+[\ln(U_b)]^2}}$$

with :

U_1 : uncertainty factor of reliability

U_2 : uncertainty factor of completeness

U_3 : uncertainty factor of temporal correlation

U_4 : uncertainty factor of geographic correlation

U_5 : uncertainty factor of other technological correlation

U_6 : uncertainty factor of sample size

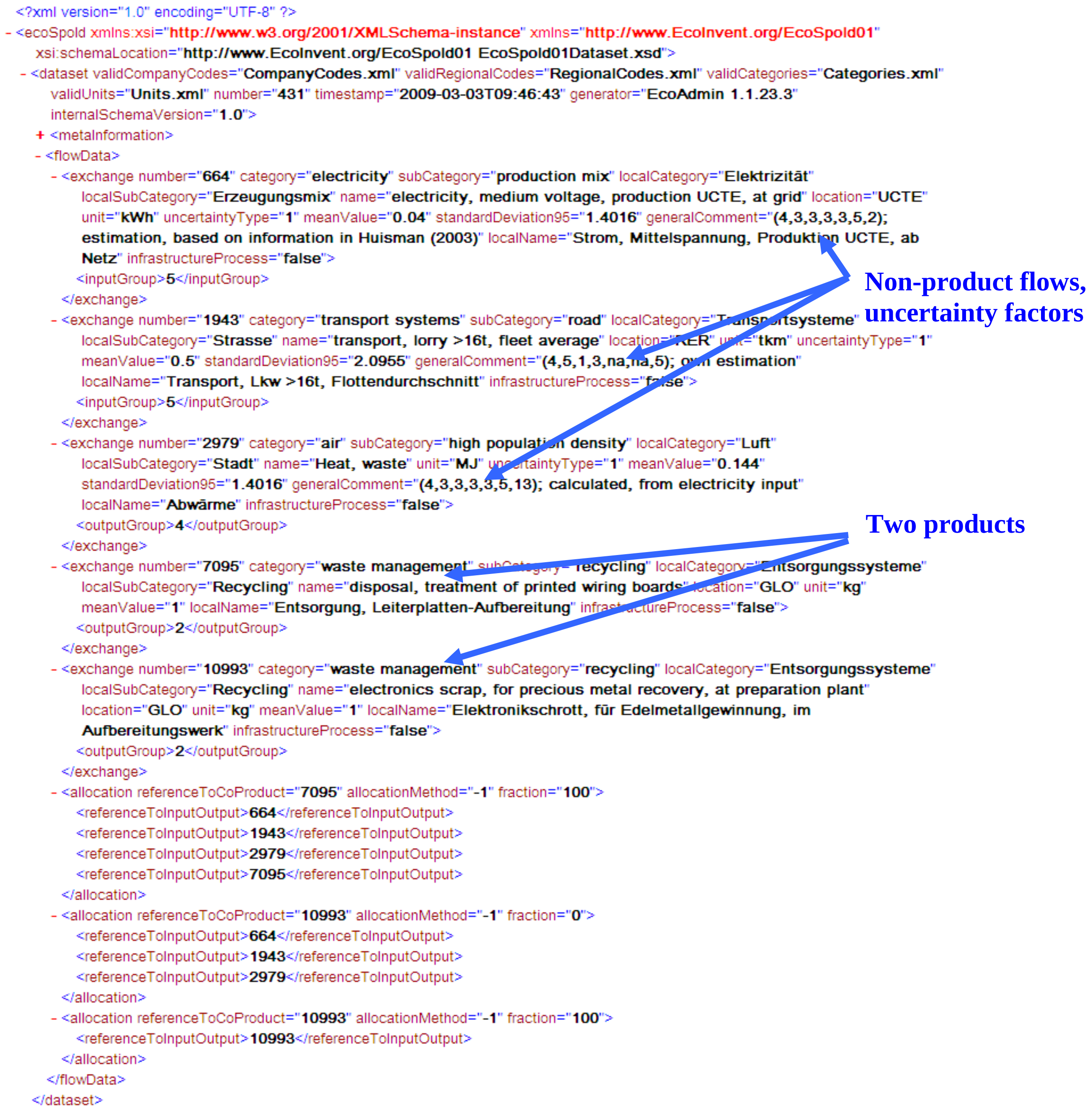
U_b : basic uncertainty factor

Equation 1: Calculating the standard deviation from the uncertainty factors

Note that in order to reflect the modified pedigree matrix in ecoinvent 3.0, the sample size factor U6 will be removed.

This formula calculates the geometric standard deviation from the uncertainty factors, and is heavily applied in ecoinvent under the assumption that elementary and intermediate exchanges are lognormally distributed.

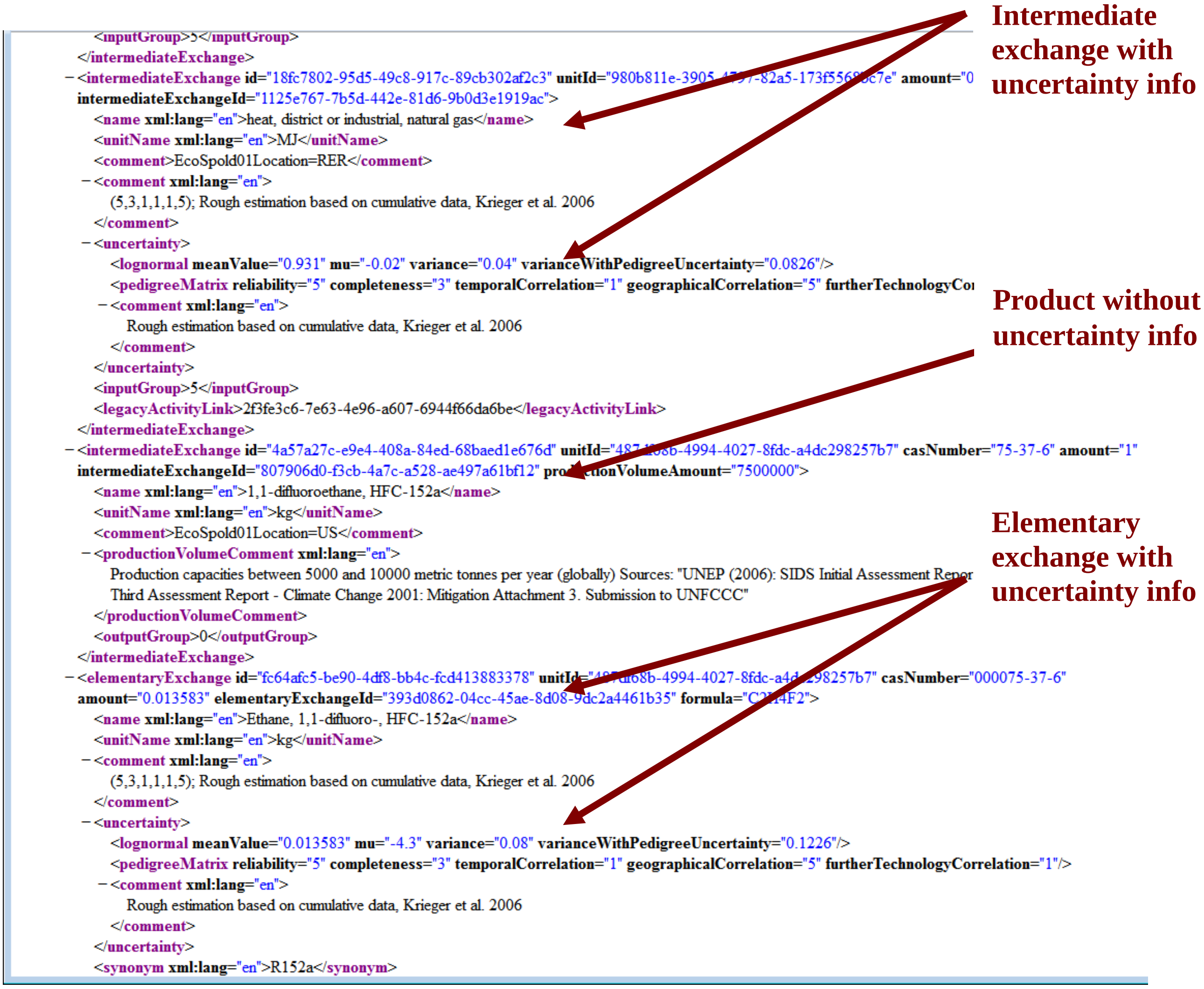
Figure 4 shows, finally, how the uncertainty information is currently displayed in ecoinvent. Process “0431.xml” is a multi-output process with two products, and three intermediate exchanges (inputs from technosphere). For the latter, the uncertainty is specified in two ways. First, the uncertainty is displayed as mean (meanValue) and twice the standard deviation (standardDeviation95). Second, the uncertainty scores from which the standardDeviation95 was calculated and obtained from the pedigree matrix and from the basic uncertainty table, are written in the general description field in brackets. For example, “(4,3,3,3,3,5,2)” for the flow nr. 664 (electricity production mix UCTE) are the scores for reliability, completeness and so on. Some values are surprising, e.g. waste heat, exchange 2979, has one uncertainty indicator score of 13, which is of course not provided directly by the pedigree matrix.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ecoSpold xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.EcoInvent.org/EcoSpold01"
    xsi:schemaLocation="http://www.EcoInvent.org/EcoSpold01 EcoSpold01Dataset.xsd">
  - <dataset validCompanyCodes="CompanyCodes.xml" validRegionalCodes="RegionalCodes.xml" validCategories="Categories.xml"
      validUnits="Units.xml" number="431" timestamp="2009-03-03T09:46:43" generator="EcoAdmin 1.1.23.3"
      internalSchemaVersion="1.0">
    + <metaInformation>
    - <flowData>
      - <exchange number="664" category="electricity" subCategory="production mix" localCategory="Elektrizität"
          localSubCategory="Erzeugungsmix" name="electricity, medium voltage, production UCTE, at grid" location="UCTE"
          unit="kWh" uncertaintyType="1" meanValue="0.04" standardDeviation95="1.4016" generalComment="(4,3,3,3,3,5,2);
          estimation, based on information in Huisman (2003)" localName="Strom, Mittelspannung, Produktion UCTE, ab
          Netz" infrastructureProcess="false">
          <inputGroup>5</inputGroup>
        </exchange>
      - <exchange number="1943" category="transport systems" subCategory="road" localCategory="Transportsysteme"
          localSubCategory="Strasse" name="transport, lorry >16t, fleet average" location="RER" unit="tkm" uncertaintyType="1"
          meanValue="0.5" standardDeviation95="2.0955" generalComment="(4,5,1,3,na,na,5); own estimation"
          localName="Transport, Lkw >16t, Flottendurchschnitt" infrastructureProcess="false">
          <inputGroup>5</inputGroup>
        </exchange>
      - <exchange number="2979" category="air" subCategory="high population density" localCategory="Luft"
          localSubCategory="Stadt" name="Heat, waste" unit="MJ" uncertaintyType="1" meanValue="0.144"
          standardDeviation95="1.4016" generalComment="(4,3,3,3,3,5,13); calculated, from electricity input"
          localName="Abwärme" infrastructureProcess="false">
          <outputGroup>4</outputGroup>
        </exchange>
      - <exchange number="7095" category="waste management" subCategory="recycling" localCategory="Entsorgungssysteme"
          localSubCategory="Recycling" name="disposal, treatment of printed wiring boards" location="GLO" unit="kg"
          meanValue="1" localName="Entsorgung, Leiterplatten-Aufbereitung" infrastructureProcess="false">
          <outputGroup>2</outputGroup>
        </exchange>
      - <exchange number="10993" category="waste management" subCategory="recycling" localCategory="Entsorgungssysteme"
          localSubCategory="Recycling" name="electronics scrap, for precious metal recovery, at preparation plant"
          location="GLO" unit="kg" meanValue="1" localName="Elektronikschrott, für Edelmetallgewinnung, im
          Aufbereitungswerk" infrastructureProcess="false">
          <outputGroup>2</outputGroup>
        </exchange>
      - <allocation referenceToCoProduct="7095" allocationMethod="-1" fraction="100">
          <referenceToInputOutput>664</referenceToInputOutput>
          <referenceToInputOutput>1943</referenceToInputOutput>
          <referenceToInputOutput>2979</referenceToInputOutput>
          <referenceToInputOutput>7095</referenceToInputOutput>
        </allocation>
      - <allocation referenceToCoProduct="10993" allocationMethod="-1" fraction="0">
          <referenceToInputOutput>664</referenceToInputOutput>
          <referenceToInputOutput>1943</referenceToInputOutput>
          <referenceToInputOutput>2979</referenceToInputOutput>
        </allocation>
      - <allocation referenceToCoProduct="10993" allocationMethod="-1" fraction="100">
          <referenceToInputOutput>10993</referenceToInputOutput>
        </allocation>
      </flowData>
    </dataset>
```

Figure 4: XML view of ecoinvent 2.0 multi-output process 0431.xml with uncertainty information

Figure 5 shows one example of a data set in the new EcoSpold 2.0 format. The uncertainty information in these examples does not provide the pedigree matrix scores (figure 5 – the entries in “general comment” are lacking). Concerning uncertainty, the basic structure in the data set is similar than in EcoSpold 1 format, with point and range estimators (mean, standard deviation) provided for each exchange that is not the product / quantitative reference.



```
        <inputGroup>5</inputGroup>
      </intermediateExchange>
    - <intermediateExchange id="18fc7802-95d5-49c8-917c-89cb302af2c3" unitId="980b811e-3905-4[illegible]7-82a5-173f556[illegible]c7e" amount="0
      intermediateExchangeId="1125e767-7b5d-442e-81d6-9b0d3e1919ac">
        <name xml:lang="en">heat, district or industrial, natural gas</name>
        <unitName xml:lang="en">MJ</unitName>
        <comment>EcoSpold01Location=RER</comment>
      - <comment xml:lang="en">
          (5,3,1,1,1,5); Rough estimation based on cumulative data, Krieger et al. 2006
        </comment>
      - <uncertainty>
          <lognormal meanValue="0.931" mu="-0.02" variance="0.04" varianceWithPedigreeUncertainty="0.0826"/>
          <pedigreeMatrix reliability="5" completeness="3" temporalCorrelation="1" geographicalCorrelation="5" furtherTechnologyCo
        - <comment xml:lang="en">
            Rough estimation based on cumulative data, Krieger et al. 2006
          </comment>
        </uncertainty>
        <inputGroup>5</inputGroup>
        <legacyActivityLink>2f3fe3c6-7e63-4e96-a607-6944f66da6be</legacyActivityLink>
      </intermediateExchange>
    - <intermediateExchange id="4a57a27c-e9e4-408a-84ed-68baed1e676d" unitId="487[illegible]86b-4994-4027-8fdc-a4dc298257b7" casNumber="75-37-6" amount="1"
      intermediateExchangeId="807906d0-f3cb-4a7c-a528-ae497a61bf12" pro[illegible]ctionVolumeAmount="7500000">
        <name xml:lang="en">1,1-difluoroethane, HFC-152a</name>
        <unitName xml:lang="en">kg</unitName>
        <comment>EcoSpold01Location=US</comment>
      - <productionVolumeComment xml:lang="en">
          Production capacities between 5000 and 10000 metric tonnes per year (globally) Sources: "UNEP (2006): SIDS Initial Assessment Repor
          Third Assessment Report - Climate Change 2001: Mitigation Attachment 3. Submission to UNFCCC"
        </productionVolumeComment>
        <outputGroup>0</outputGroup>
      </intermediateExchange>
    - <elementaryExchange id="fc64afc5-be90-4df8-bb4c-fcd413883378" unitId="[illegible]df68b-4994-4027-8fdc-a4d[illegible]98257b7" casNumber="000075-37-6"
      amount="0.013583" elementaryExchangeId="393d0862-04cc-45ae-8d08-9dc2a4461b35" formula="C[illegible]4F2">
        <name xml:lang="en">Ethane, 1,1-difluoro-, HFC-152a</name>
        <unitName xml:lang="en">kg</unitName>
      - <comment xml:lang="en">
          (5,3,1,1,1,5); Rough estimation based on cumulative data, Krieger et al. 2006
        </comment>
      - <uncertainty>
          <lognormal meanValue="0.013583" mu="-4.3" variance="0.08" varianceWithPedigreeUncertainty="0.1226"/>
          <pedigreeMatrix reliability="5" completeness="3" temporalCorrelation="1" geographicalCorrelation="5" furtherTechnologyCorrelation="1"/>
        - <comment xml:lang="en">
            Rough estimation based on cumulative data, Krieger et al. 2006
          </comment>
        </uncertainty>
        <synonym xml:lang="en">R152a</synonym>
```

Intermediate exchange with uncertainty info

Product without uncertainty info

Elementary exchange with uncertainty info

Figure 5: XML view of one example process provided with EcoSpold02

While the current ecoinvent approach of specifying uncertainty for non-product flows in all processes is certainly advanced, it has the following drawbacks and improvement options:

- the approach relies heavily on expert judgement in several steps; the basic uncertainty scores, and the transformation table for pedigree matrix indicator scores to uncertainty figures would be better founded by a broader empirical basis;
- the approach is at present only applicable for lognormally distributed values and further assumes that this type of distribution is relevant for the majority of elementary and intermediate exchanges;
- the approach does not cover parameters and other “model elements” besides non-product exchange amounts;
- I/O data and other recent modelling developments in LCA are not directly addressed by the approach.

The ecoinvent centre is interested in having the procedure for uncertainty modelling in ecoinvent scrutinised, bottlenecks and existing limitations removed where possible, and uncertainty information in ecoinvent data put on a broader empirical basis. The developed approach should improve the existing data basis at present, and at the same time be maintainable and allow further improvements over time.

3 Goal and scope for phase 0 of this project

Taking the structure of the pedigree matrix as given, and taking also the approach to derive quantitative figures from the properties of the attributes in the matrix as given, this phase will use the formula in equation 1 and seek to provide

i) empirical foundations for the values of the uncertainty factors used in the formula, and/or propose different values for these factors, based on empirical measurements as far as possible.

ii) practical considerations on how to apply the pedigree approach to other distributions than lognormal

The work considers only exchanges; for the moment, impact assessment questions and other parameters are excluded. Uncertainty is defined here simply as geometric standard deviation of intermediate and elementary exchanges at the individual (i.e. un-accumulated) unit process level. The calculation of uncertainty for accumulated "system" processes is not considered in this task.

→ i) empirical foundations for the values of the uncertainty factors used in the formula

This subtask will build on the existing pedigree matrix and basic uncertainty table, and on the existing approach to derive quantitative uncertainty figures from the matrix that are better grounded in available empirical data. The empirical data may suggest not only changing the values of the uncertainty factors as they relate to scores in the matrix, but also possibly re-wording and re-arranging the cells/cell contents and the differentiation of these factors/cell contents depending on data type or application area.

In order to provide this empirical basis, a combination of the following approaches will be used:

- meta analysis of existing studies, in the LCA domain
- analysis of existing data sets, in the LCA domain
- analysis of any other source available, recognising its relevance for the LCA domain
- analyses of specific measurements at industrial processes

The specific approach will develop during the study; it will be properly documented and allow further extension. This is important as there is currently very little experience about an empirical foundation.

The analysis must consider also sources outside of the ecoinvent database to avoid circular reasoning.

Analysis of specific measurements of industrial processes will be taken from sources from GreenDeltaTC, from the University of Wuppertal (Jutta Hildenbrand, e.g. surface coating and washing processes), and other sources (e.g. Lundie 2004). A broad range of LCA data will be accessed through the UNEP/SETAC Database Registry, (http://lca-data.org).

Other sources that do not belong traditionally to the LCA domain include IO data, EPER (EPER 2010) and PRTR (The European Pollutant Release and Transfer Register, PRTR 2010), ZSE/DeHSt (ZSE 2010, with data integrated in the ProBas database of the German ProBas 2010), a broad range of data available from eurostat (eurostat 2010), from US statistics / census (census 2010), and more.

These analyses will in the first place try to relate "shares" of the overall measured standard deviation to each of these attributes and to the indicator scores of each attribute.

On the other side, all factors together with the basic uncertainty factors need to be able to reproduce the overall uncertainty of the respective flow in terms of its standard deviation, and they might either omit relevant aspects, or, due to correlation, overestimate the uncertainty.

Therefore,

- correlation between different factors needs to be considered as well (as mentioned below in the footnote for the geographical "correlation" factor)
- an "uncertainty completeness check" is necessary in addition to the factor-specific analyses mentioned so far. In analogy to variance analysis, the uncertainty that is provided from the specific factors can be called the "explained uncertainty" – and the aggregated explained uncertainty can be lower or higher than the realistic uncertainty.

In order to relate the explained uncertainty to the realistic uncertainty, measurements and literature data will be analysed.

Finally, it is well possible that the analyses show either that the relation between indicator in the pedigree matrix and the empirically established uncertainty factors vary across process or flow type, or that the list of types of processes and exchange types that is used to differentiate basic uncertainty does not line up with emerging clusters of processes/exchanges. If this is the case, then the task will try to propose archetype processes, each of which will be associated to different uncertainty factors or even different pedigree matrix cell content.

This task has a number of results:

- an initial "seed" list of "archetype processes", with definition
- empirically based uncertainty factors, if needed distinguished by archetype, for the basic uncertainty and for the uncertainty factors from the pedigree matrix, for all exchanges in ecoinvent version 2.2, for each of the cells in the pedigree matrix
- a documented and tested approach for an empirical foundation of the uncertainty factors.

Due to the relatively short time frame, this task will not be able to produce very detailed factors for all unit processes in ecoinvent. It will, however, cover the variety of processes and flows in ecoinvent, and provide reasonable default values. These default values and the developed approach will be further refined in the following phases of the project and will provide a structure and "backbone" of the following analyses.

ii) practical considerations on how to apply the pedigree approach to other distributions than lognormal

The analysis will cover main practical aspects related to the (assumed) probability distribution of the data, including the aggregation formula (equation 1). Specifically, formulas for other distributions will be provided. All uncertainty distributions that are foreseen in ecoinvent (in the EcoSpold02 format) will be considered.

4 Starting points for an empirical foundation of uncertainty factors

4.1 Uncertainty

Long debates in the LCA community have not really provided a common understanding of uncertainty, nor have they yielded a commonly accepted definition of uncertainty.

In this project, uncertainty is understood as follows:

Uncertainty means, basically, lack of certainty. A quantitative figure for the emission of a flow is not exactly known; the correct allocation method for a multi output process is not exactly known; it is unclear whether electric arc furnace steel should be included in a product system, or converter steel: all these situations "contain" uncertainty.

Several authors emphasize that uncertainty is ubiquitous, or pervasive (Morgan Henrion 2000, p. 3), also for LCA (Heijungs Huijbregts 2004; Ciroth 2003). This can even be "traced back" to the Heisenberg Principle (!)[2].

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 (organic farming e.g.)
- the farming background and expertise of the farmers
- asf.

Uncertainty thus can in parts be "explained" by these details, the parameters listed above. This links directly to the concept of 'explained variation' or 'explained variance' in statistics, (Kent 1983).

Some authors distinguish variability from uncertainty, variability describing then variations in data that are "inherent", and not caused by measurement or perception errors[3]. A typical example is the temperature over the day, which cannot be controlled and completely explained by parameters.

However, the distinction between inherent variations and measurement errors is difficult in practice; it is always arbitrary to some extent, as the measurement procedure and technique has of course an influence on the parameters that can be controlled. In the fertiliser example, the time covered, and the size of the farms, are relatively easy to take into account, while the expertise of the farmers is much more difficult to operationalise and therefore to consider. But it is here (and often) rather a question of the effort spent on operationalising parameters that then determine the share of uncertainty that is considered as variability, and as uncertainty on the other side.

However, there will always be a remainder of unexplained uncertainty (i.e. variability, speaking with Huijbregts); therefore, the concept of variability is of interest.

Besides the specification of parameters that introduce variation in datasets, there is a next level of parameters that introduce variations in the specification of the parameters (so for example, how the size of the farms is determined) – and so on. So the understanding of uncertainty as the "remainder" of variations in data that cannot be explained by parameters is simple, Figure 6 shows this in principle: the more parameters are taken into account, the lower

[2] As a reminder, the Heisenberg uncertainty principle can be described as follows: "it is not possible to simultaneously determine the position and momentum of a particle. Moreover, the better position is known, the less well the momentum is known (and vice versa)" (Eric Weisstein's World of Physics, http://scienceworld.wolfram.com/physics/UncertaintyPrinciple.html, July 2010)

[3] For example Huijbregts 2001, p. 15: "Variability is understood here as stemming from inherent variations in the real world, while uncertainty comes from inaccurate measurements".

the uncertainty. This holds if the parameters itself are perfectly known, without any uncertainty.

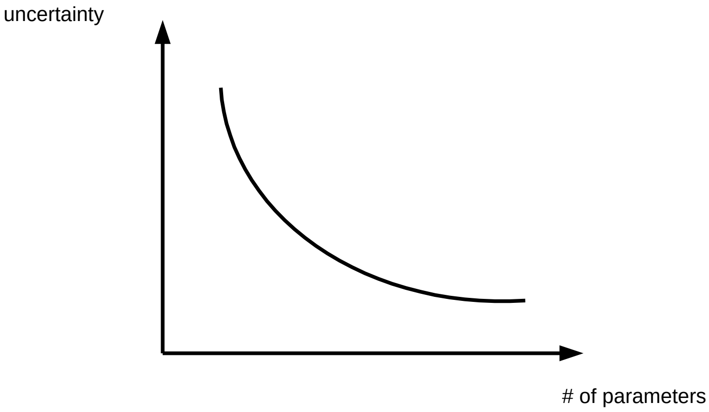


Figure 6: General relation between uncertainty and the number of known parameters: the more parameters are known, the lower the uncertainty. Further explanation see text

This model is very simple in principle. However, it applies also to the determination of the parameters, being therefore a recursive model that can be complex in a real situation; and as every parameter has in principle an own “network” of parameters that in turn determine it, each parameter introduces also an own uncertainty, therefore the overall uncertainty can even increase with an increase of the parameter number (and there will often be an optimum parameter set, with minimal overall uncertainty).

For the uncertainty analysis, therefore, a data range, or a “spread” in data, will be analysed; the data range is specified by identical values of “parameters” that describe the data. Example for parameters are the reference year of a data set, the geography, the specific technology – in short, the indicators used in the pedigree matrix, plus additional similar parameters where necessary. The data itself are always input or output values of exchanges for processes / activities, as the activities in the ecoinvent database.

4.2 What does “empirical” mean?

In the frame of this project, *empirical* will be defined as ‘derived from experiment and observation rather than theory and expert guesses’, expanding thereby a definition given by the Princeton Wordnet database[4].

Own experiments will be not possible during this project; aim is therefore to compare data to available measurements where possible. Any parameters used in these measurements will need to be considered in this comparison, as is explained in the uncertainty section, 4.1.

As a second option, other indirect sources will be used, especially those that are not directly linked to the ecoinvent database, and even not linked to the LCA context.

4.3 Analysed data sources

Following the idea of looking into many different, independent data sources for process inputs and outputs, a broad range of data sources has been investigated. Not every data source fits directly into the LCA context; often, data preparations were necessary before any analyses concerning uncertainty factors could be performed. Data preparations often included data transformation, in order to make data better comparable to LCA. And even after that, data

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

sources often contained limitations that need to be taken into account in the interpretation of analysis results. The annex explains the different data sources in detail, providing also information about any data transformation or other "preparatory work" that was performed, and about remaining limitations in 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 2009), non-LCA sources, measurement data.

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[5] 2010, ZSE)

Temporal correlation: Emission inventories, as the German ZSE system, eurostat and US statistics, and in parts also the PRTR system, have datasets over several years that allow time series analyses and will be taken into account[6]. ETH 96 – ecoinvent 2000 – ecoinvent 2007 – ecoinvent 2010 (all datasets) will be looked at as well, although ecoinvent data does of course not always reflect real process changes over the years. Also learning curve models and data will be considered (e.g. (Fuss Szolgayová 2009)).

Geographical correlation: Comparison of transport emissions of the same or very similar transportation vehicle from different regions[7]; differences in electrical grids for different regions, in different databases.

Further technological correlation: Solar cell comparison from the Gemis database and from ecoinvent, and for transport datasets from the Tremod database, from the GREET model (GREET 2009), and from ecoinvent.

Basic uncertainty: LCA databases (ecoinvent & others), literature, non-LCA sources, measurements.

4.4 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 characterises the spread in underlying data, and it is commonly used in random error analyses.

For the analysis of uncertainty in this project, the standard deviation seems therefore an ideal candidate. It has, however, the disadvantage to depend 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:

[5] http://www.ifeu.de/index.php?bereich=ver&seite=projekt_tremod

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

[7] Especially here, correlations with other attributes need to be considered; the factor used for geographical correlation should reflect only those aspects that are indeed caused by geographical differences. Little influence on 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 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, asf.). The uncertainty is applied at the level of individual exchanges, and therefore further uncertainty on aggregated process level e.g. can (and should) be left out of consideration.

$Var(aX + b) = a^2Var(X)$.

For the standard deviation SD holds, respectively: $SD(aX + b) = aSD(X)$

This means that a constant factor that is applied to all the analysed data values changes the standard deviation by the same factor. This may happen if for example data is given in g instead of kg; all values will be multiplied by a constant factor of 1000, and the resulting standard deviation will also increase by a factor of 1000, when values are given in g instead of kg.

There are three main reasons for scaling effects in the analysed data:

1. data may not be provided per functional unit at all; this requires a transformation of the data, for example from absolute emission figures of an industrial plant to “per kg product” emission figures
2. if a functional unit is given, the quantitative reference may differ (1000 m^2 for one data source or group of data; 1 m^2 for another)
3. data may simply be provided in different units (kg emissions vs. emissions in grams)

These scaling effects are undesirable as the uncertainty factors should be independent from the scale of underlying data; the factor should not change if data is given in kilogram or gram.

In order to try to overcome the scale dependency, there are several options. They are all in detail discussed in the annex, in Annexe B: Normalisation options, but can be summarised as follows:

- analyse the raw data as given, and analyse the uncertainty as standard deviation (ignore scaling effects).
- transform data in a linear way and analyse the standard deviation of the transformed data (linear transformation); possible specific transformations are division of all process flows by the mean of the flows per process, or by one common flow that is input or output of most of the analysed processes.
- perform a lognormal transformation of the data and analyse the standard deviation for the transformed data, which is the same as analysing the geometrical standard deviation of the raw data (geometrical standard deviation). This approach does not imply that data follows the lognormal probability distribution (!)

These options are analysed more in detail in the annex. In order to decide for one or the other option, three different aspects needs to be considered; first, how well the “original uncertainty” is preserved for the analysis, second, how well the scaling effect is managed, meaning removed, and finally, how well the result of the analysis fits into the pedigree scheme.

Ignoring scaling effects obviously is able to preserve the original uncertainty; if scaling effects are relevant, then these are not addressed and therefore still have a negative influence on the analysis, and third, the standard deviation does not fit into the current pedigree matrix[8].

Linear transformation “modify” the original variance and hence uncertainty in data, and, on the other side, do not really manage to overcome scaling effects, if the functional unit is not known.

[8] But as shown later in the text, the pedigree matrix works in principle also for other indicators than geometric standard deviation.

Lognormal transformations have the effect that linear factors, which are the same for all the analysed/transformed data, “disappear”[9].

Table 1 Evaluation summary of different options for managing scaling effects in data

Option	Preserving original uncertainty	Scaling effects mitigation	Results fits to pedigree approach
Ignore scaling	++	--	-
Linear transformation	O	o	-
Lognormal transformation	+	+	++

As the table shows, the lognormal transformation scores better than the other options, if the functional unit is not known which would allow removing the scaling effect.

As conclusion, the following approach will be taken in the data analysis in order to deal with scaling effects:

First, any scaling effect possible should be removed from the analysed data.

Units should be consistent per “unit group” (e.g., volume should be given always in litre, mass in kg, and the like)

If available, process data sets should be transformed to the same amount of quantitative reference (e.g., all processes should be transformed to represent 1 unit of product, avoiding mixtures of 1, 5 and 1000 product units)

Second, the geometric standard deviation, calculated as standard deviation of the log-transformed data, should be used for the analysis.

4.5 From data sources to uncertainty factors

This section describes how to arrive at the uncertainty factors in the pedigree matrix, starting from data sources. Data are transformed already to overcome scaling effects as best as possible, as explained in the previous section.

The analysis is always done targeting one specific indicator in the pedigree matrix (the lines in the matrix, time, geography, technology, asf.) or for the basic uncertainty factors. These indicators in the pedigree matrix are assumed as independent.

Data sources will be taken into account where at least one of the indicators in the pedigree matrix varies, or provide several different flows or processes so that they can be related to basic uncertainty factors.

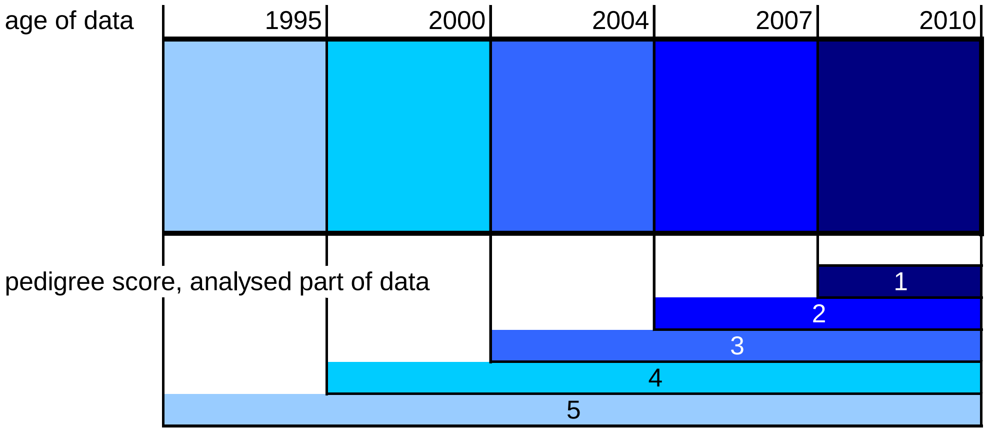
The basic approach is as follows:

For one indicator, the analysis will constrain its values stepwise, filtering out more and more data sets from the data source. The standard deviation will be calculated for each filtered sub-set. The constraints will set in a way that they reflect the thresholds foreseen in the pedigree matrix.

[9] Log(c*x) = log(c) + log (x); if c is a constant, the following holds: The variance var of a constant equals 0, therefore var(log(c*x)) = var(log(c)) + var(log(x)) = var(log(x)), see also the annex for more details.

As a result, the data sample will be more and more precise regarding the investigated indicator. The calculated (geometrical) standard deviation will reflect this, and be different for each of the “filter steps”.

Figure 7 below shows an example for the pedigree indicator **temporal correlation**, with 2010 being the time period of the data set.



pedigree score definitions

1	data less than 3 years difference to the time period of the data set
2	data less than 6 years difference to the time period of the data set
3	data less than 10 years difference to the time period of the data set
4	data less than 15 years difference to the time period of the data set
5	age of data unknown or more than 15 years difference to the time period of the data set

Figure 7: Relating data sources to the pedigree matrix and pedigree indicators, principal example for the indicator temporal correlation. Time period of the data set is 2010

When the analysis is done in 2010, then, in this example, data sets from data sources that do not match the time reference to the ideal data set will always be older than this data set. The difference to the data set can therefore only develop in one direction, in the figure from 2010 until 1995. For time, it seems reasonable to exclude prognostic data set from the analysis, or at least to treat them in a different way than data sets that refer to past times: Any data prognosis will involve additional uncertainty and make the data sets incomparable to data sets from the past.

If the time reference of the data set lies in the past, differences to the time reference “develop” in two directions. This is shown in the next figure, with 2004 as example time reference.

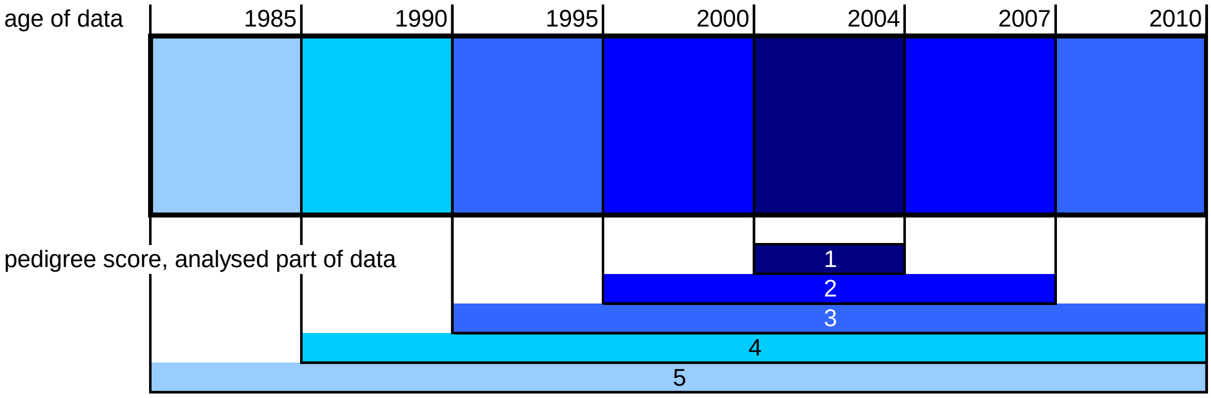


Figure 8: Relating data sources to the pedigree matrix and pedigree indicators, principal example for the indicator temporal correlation. Time period of the data set is 2004. Compare to Figure 7

The pedigree indicators **geographical correlation** and **further technological correlation** can be dealt with in a similar way. For the indicator **reliability**, the situation is easier as the scores in the matrix are obtained for absolute values that do not involve a comparison to a reference state.

For the indicator **completeness**, on the other side, the situation is more complex, as one data set cannot be assigned to a certain completeness level in an unambiguous manner. Whether a single data set belongs to the "<50%" size or not, for example, depends on the sample and not on the data set itself. And the data sets that are in the sample influence, of course, the calculated uncertainty and the uncertainty factors. Therefore, for completeness, several possible ways to building subgroups from a complete sample will be analysed.

As a complication for the analysis, there will, ideally, be **several independent sources for analysing one indicator**. This follows the concept of triangulation in measurement science: Regarding one specific aspect from several viewpoints, several sources, in order to get a more reliable image; following several lines of arguments, and several sources, will lead to slightly different estimates of the uncertainty factors; they are drawn as bullets in Figure 9.

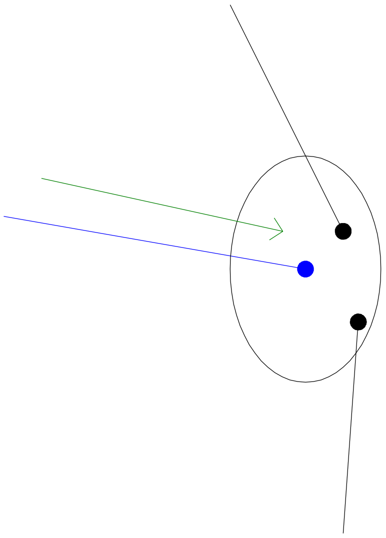


Figure 9: The principle of triangulation in the analysis; further explanation see text

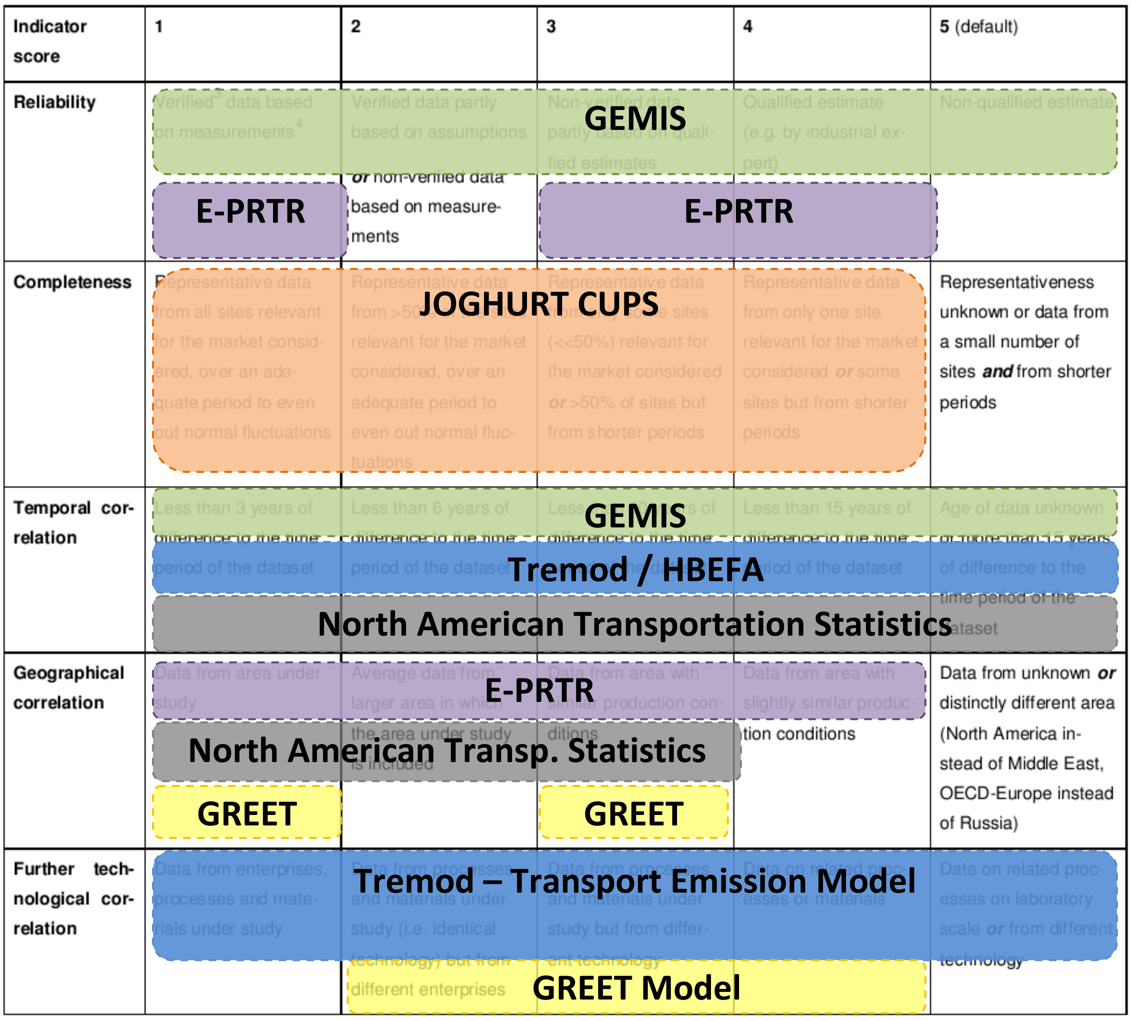
Considering different sources leads to an overall better founded estimate for the indicator values. Different results from different sources simply need to be compared and analysed. Due to the geometric standard deviation, the comparison is easier since constant factors per data source will disappear. However, the analysis cannot be done in a fully automated manner and will involve some expert judgement.

5 Analyses

5.1 Overview

Figure 10 shows the “coverage” of the different fields in the pedigree matrix by the various data sources that were taken into account. Several data sources are used for several indicators; also, following the triangulation idea, several data sources are used for the assessment of one indicator field.

The basic uncertainty factors are analysed separately; the data sources are identical to the ones listed for the pedigree matrix.



Indicator score	1	2	3	4	5 (default)
Reliability	Verified[3] data based on measurements[4]	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 sites relevant for the market considered, over an adequate period to 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 sites (<<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 to 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 but from different technology	Data on related processes or materials	Data on related processes on laboratory scale *or* from different technology

GEMIS
E-PRTR
E-PRTR
JOGHURT CUPS
GEMIS
Tremod / HBEFA
North American Transportation Statistics
E-PRTR
North American Transp. Statistics
GREET
GREET
Tremod – Transport Emission Model
GREET Model

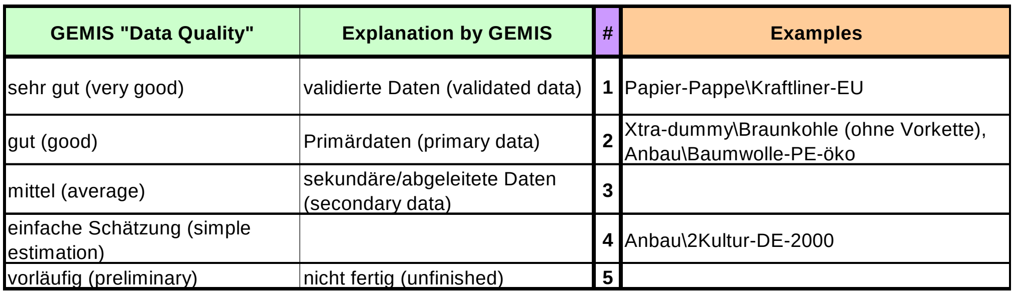
Figure 10: ecoinvent 3.0 pedigree matrix and coverage of databases analysed

5.2 Reliability

5.2.1 Data sources used

5.2.1.1 GEMIS

The GEMIS database provides, for each process data set, a “data quality indicator”, with the following possible entries:

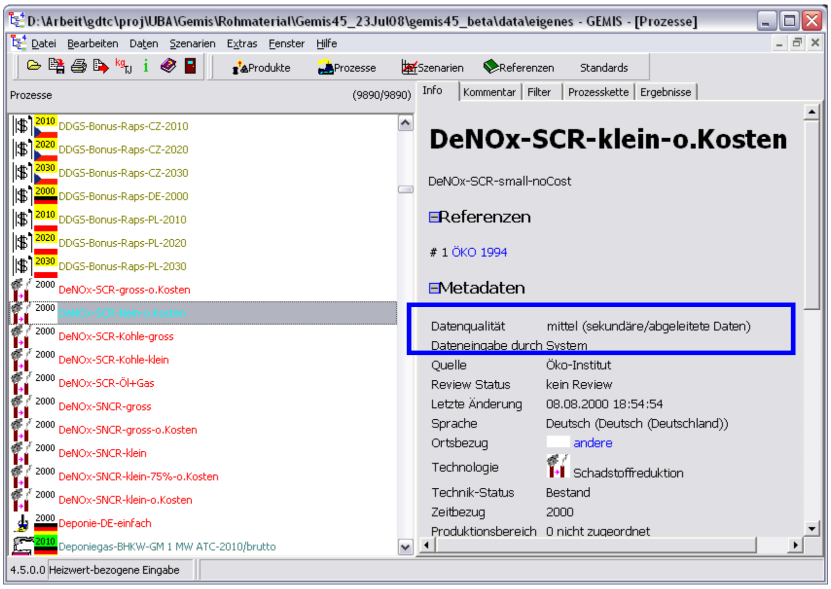


GEMIS "Data Quality"	Explanation by GEMIS	#	Examples
sehr gut (very good)	validierte Daten (validated data)	1	Papier-Pappe\Kraftliner-EU
gut (good)	Primärdaten (primary data)	2	Xtra-dummy\Braunkohle (ohne Vorkette), Anbau\Baumwolle-PE-öko
mittel (average)	sekundäre/abgeleitete Daten (secondary data)	3	
einfache Schätzung (simple estimation)		4	Anbau\2Kultur-DE-2000
vorläufig (preliminary)	nicht fertig (unfinished)	5	

Figure 11: GEMIS "data quality" indicators with explanation and examples

These indicators do not perfectly fit to the pedigree indicator scores; ecoinvent usually speaks of verification, while GEMIS uses validation, for the quality assurance process; often, ecoinvent is more precise. For example, "simple estimation", 4, in GEMIS links to "qualified estimate, e.g. by industrial expert", in ecoinvent. However, it is interesting to see whether there is different uncertainty in the data sets distinguished by these indicators in GEMIS.

From the examples given in the table above, it can be seen that "good data quality / primary data" is assigned also to the "xtra-dummy" processes. These are empty processes used to end a supply chain; they do not contain any specific process-related data. They are comparable to SimaPro dummy processes.



Prozesse (9890/9890)

2010 DDGS-Bonus-Raps-CZ-2010
2020 DDGS-Bonus-Raps-CZ-2020
2030 DDGS-Bonus-Raps-CZ-2030
2000 DDGS-Bonus-Raps-DE-2000
2010 DDGS-Bonus-Raps-PL-2010
2020 DDGS-Bonus-Raps-PL-2020
2030 DDGS-Bonus-Raps-PL-2030
2000 DeNOx-SCR-gross-o.Kosten
2000 DeNOx-SCR-klein-o.Kosten
2000 DeNOx-SCR-Kohle-gross
2000 DeNOx-SCR-Kohle-klein
2000 DeNOx-SCR-Öl+Gas
2000 DeNOx-SNCR-gross
2000 DeNOx-SNCR-gross-o.Kosten
2000 DeNOx-SNCR-klein
2000 DeNOx-SNCR-klein-75%-o.Kosten
2000 DeNOx-SNCR-klein-o.Kosten
2000 Deponie-DE-einfach
2010 Deponiegas-BHKW-GM 1 MW ATC-2010/brutto

DeNOx-SCR-klein-o.Kosten

DeNOx-SCR-small-noCost

⊟Referenzen

1 ÖKO 1994

⊟Metadaten

Datenqualität	mittel (sekundäre/abgeleitete Daten)
Dateneingabe durch	System
Quelle	Öko-Institut
Review Status	kein Review
Letzte Änderung	08.08.2000 18:54:54
Sprache	Deutsch (Deutsch (Deutschland))
Ortsbezug	andere
Technologie	Schadstoffreduktion
Technik-Status	Bestand
Zeitbezug	2000
Produktionsbereich	0 nicht zugeordnet

Figure 12: GEMIS "data quality" assignment for a process data set

In order to focus on the uncertainty related to the "data quality" indicator, prognostic data was excluded from the analysis; likewise, the analysis was done separately for processes with the quantitative reference given in kg and in TJ, to overcome scaling effects. All analysed process data sets have then the quantitative reference of 1 (kg or TJ, respectively).

As a first analysis, we checked the effect of the "reliability factor" (also called data quality in the database) on the standard deviation. This "data quality score" in GEMIS goes from 1 (reviewed data) to 5 (provisional). For the analysis, prognostic data was excluded; also, data was analysed per main functional unit (TJ or kg), and per economic sector (NACE code). Results are displayed in Figure 13.

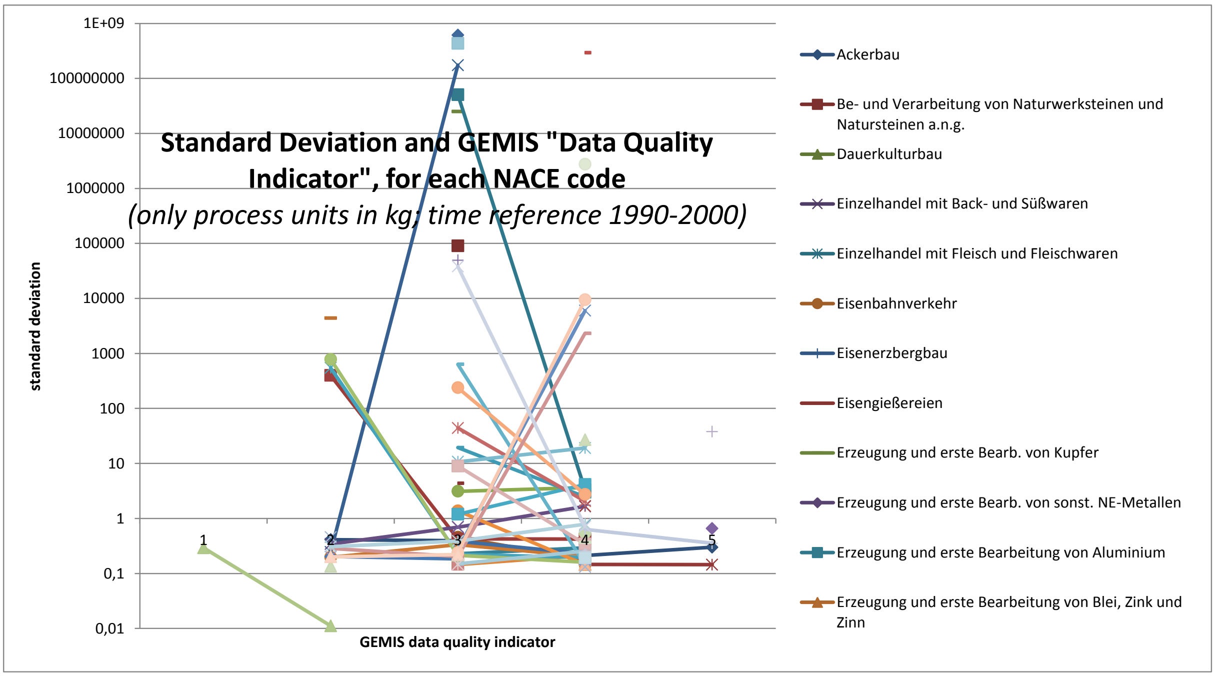


Figure 13: Standard deviation for process emissions in GEMIS in relation to the "quality indicator" in GEMIS, per NACE code (labelled in German language)

As the figure shows, there is no obvious link between the "quality indicator" and the standard deviation. For several sectors, the standard deviation seems to decrease when moving from secondary data (3) to estimates (4). Since scaling effects can appear, it seems reasonable to calculate the geometric standard deviation. This is done in Figure 14; also this chart, though, does not really show a clear structure in the uncertainty. The uncertainty is really high, with a GSD of 100 to 1000.

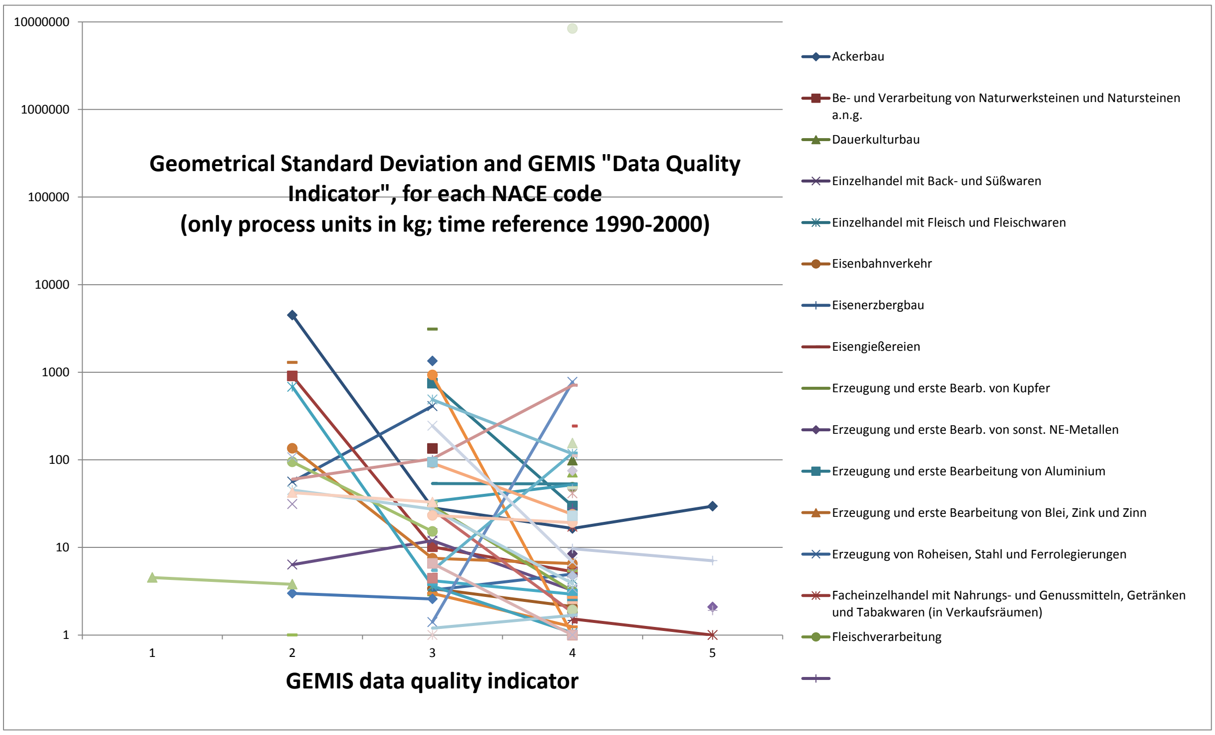


Figure 14: Geometrical standard deviation for process emissions in GEMIS in relation to the “quality indicator” in GEMIS, per NACE code (labelled in German language)

Since these results are not really satisfying, an analysis per flow (and reliability) is performed in addition to the previous analysis per sector. Motivation is that each flow has its special characteristics, and may therefore behave in a similar way even across different processes; these characteristics probably show also in the uncertainty obtained in different measurement procedures. Results are shown in Figure 15: The uncertainty across all flows is quite different. There are not so many flows that fall into the best (1) category, and also not many that fall into the worst, 5. For the categories 2, 3 and 4, the picture is quite obfuscated and unclear.

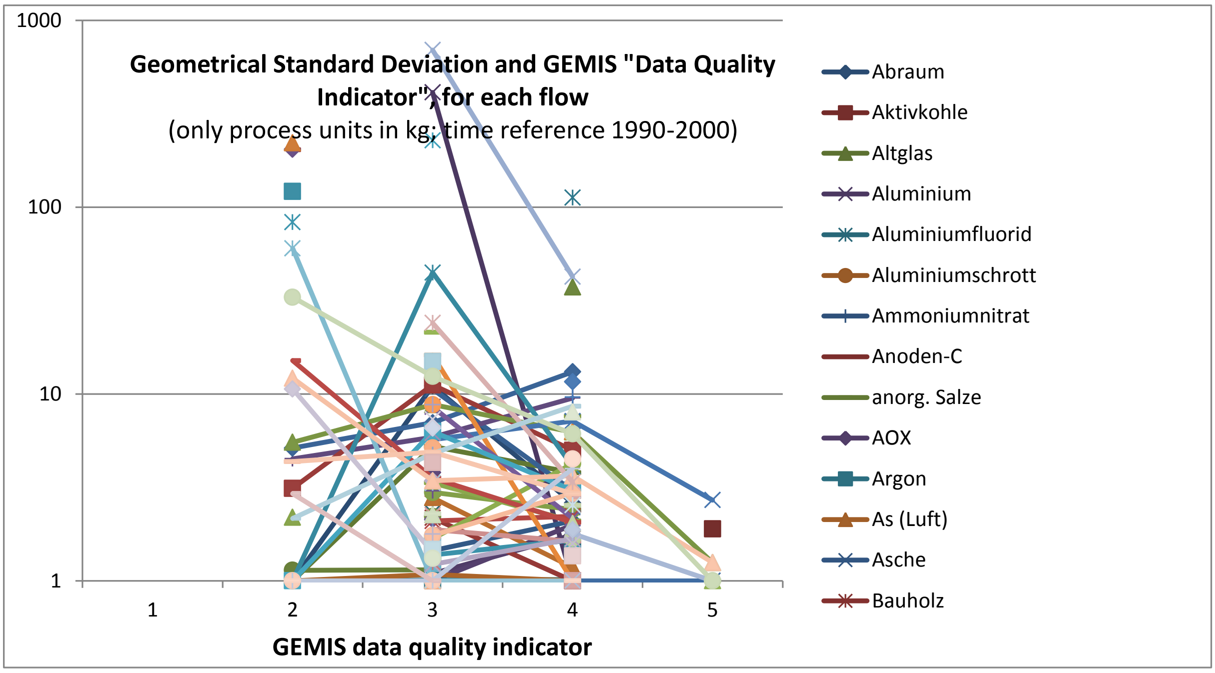


Figure 15: **Geometrical standard deviation for process emissions in GEMIS in relation to the "quality indicator" in GEMIS, per flow (labelled in German)**

In order to set these results in the perspective of the pedigree matrix and the uncertainty factors, the uncertainty with the best pedigree indicator score is set as a reference; other uncertainties are set in relation to this reference, and finally the uncertainties are expressed as geometrical standard deviation to overcome scaling effects. The ratio of an uncertainty to the reference uncertainty is then the additional uncertainty that can be considered as being caused by the respective other pedigree indicator value. They are the uncertainty factors, for the pedigree matrix.

Doing so produces, for the analysis per industrial sector, the following GSD ratios for the pedigree indicator reliability [10] that are shown in Figure 16 and Table 2. The analysis is hampered because the quality indicators in GEMIS are not available equally for all sectors; a missing value is displayed as a 1 in the figure since the figure displays the log-contributions. And yet, the GSD contributions over all sectors, the thick red line in the figure, shows a plausible behaviour.

[10] In short, values are: 1: verified data based on measurements; 2: verified data partly based on assumptions; 3: non verified data partly based on assumptions; 4: qualified estimate; 5: non-qualified estimate. For a more detailed explanation, see Figure 1.

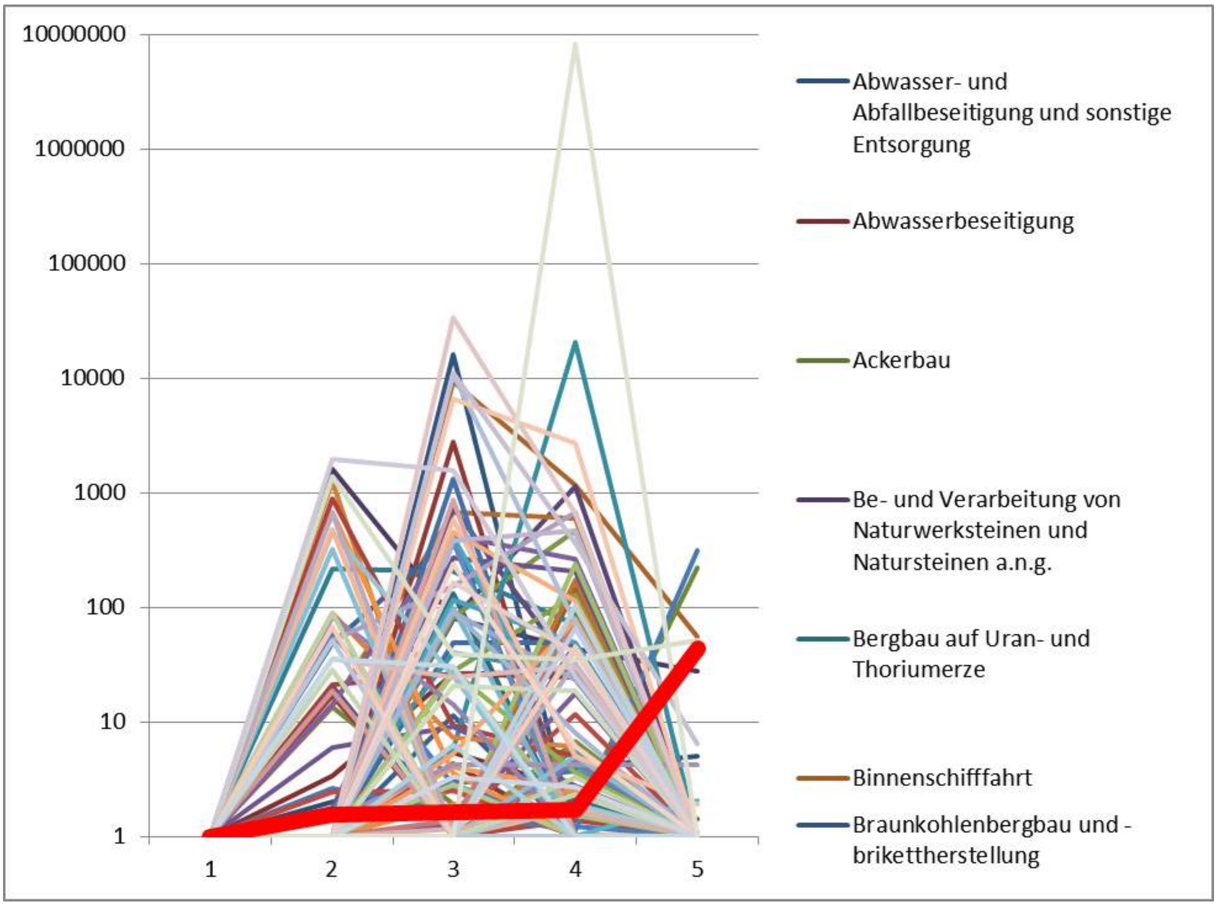


Figure 16: Geometrical standard deviation contributions for industrial sectors, in GEMIS for the indicator reliability in the pedigree matrix (x-axis), labelled in German; the thick red line is the GSD contribution over all flows

Table 2 GSD contributions for the indicator reliability in the pedigree matrix, from the GEMIS database

Indicator score “Reliability”	Uncertainty factor = GSD ratio
1	1
2	1,543529353
3	1,608154055
4	1,691120392
5	43,55207485

5.2.1.2 E-PRTR

The PRTR database contains information about the measurement approach for each process emission. Specifically, it describes if the resulting emission is estimated, calculated or measured. Figure 17 below displays the calculated geometric standard deviation (GSD) of the emissions, distinguished by measurement procedure.

Somewhat unexpected, the highest geometric standard deviation is obtained if emissions are *measured*. Second, it can be stated that the GSD is quite high for all measurement types.

The fact that estimated data has a lower uncertainty than measured data can (even!) be explained: Measured values reflect the natural variability in data. Estimated values are more often the same, varying less than real data. This seems to suggest that estimates are often done in a similar way, rather than being justified by reality[11].

The PRTR database does not contain a validated or verified data statement. As a replacement, the ‘measured’ value is taken as a reference for the indicator score 1 in the pedigree matrix. Figure 17 shows the data per country, as geometric standard deviation.

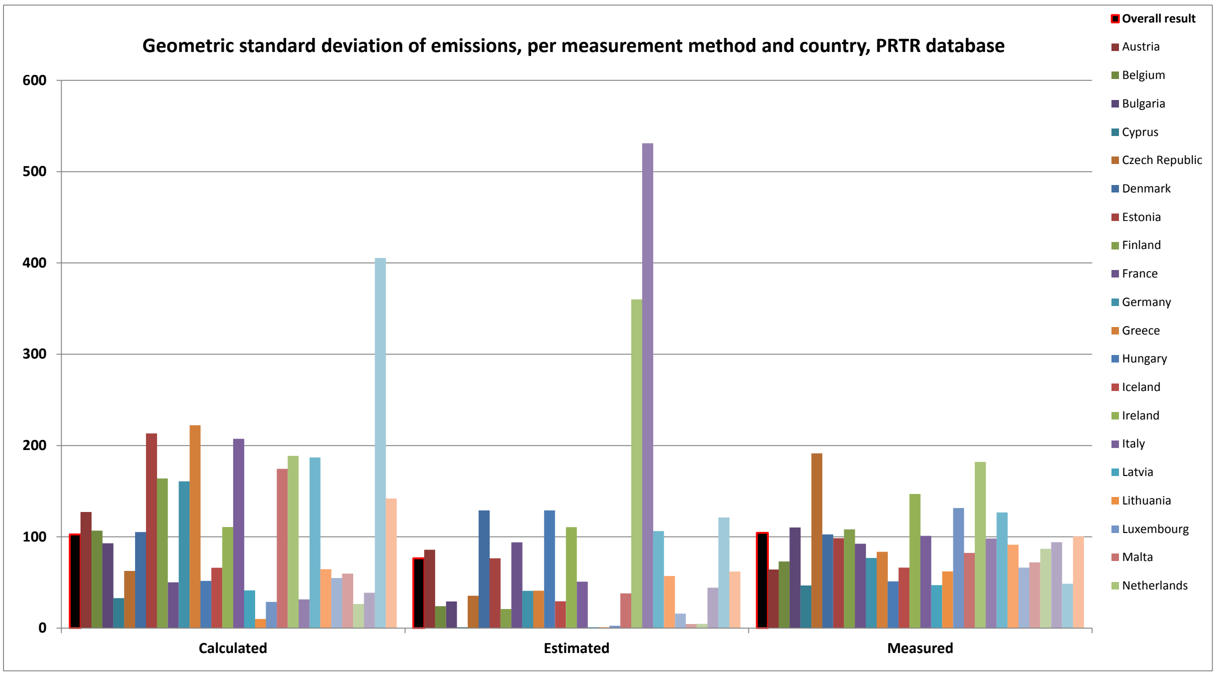


Figure 17: GSD of relative, per measurement method and country, from the PRTR database; Sweden is not shown due to negative emissions reported that cannot be represented in a log scale.

For setting this data again in perspective to the pedigree matrix, the measured GSD is taken as a reference, and all other GSD values are set in relation to this value. If GSD values are smaller than the reference, the inverse is used. The resulting ratio is then the uncertainty that can be attributed to the specific indicator value (e.g, calculated instead of measured).

Calculated is set as 3 (non verified data partly based on assumptions), and estimated is set as 4 (qualified estimate)[12].

As the above figure shows, values are quite divers, over different countries. The “overall result” value over all countries is used to come to an overall estimate.

The result is shown in Table 3.

[11] Take, as an example, the estimation of transport distances. One will rather estimate 50km or 75km distances than estimating a 48km or 79 km distance.

[12] See footnote 10 for an explanation of the indicator values. Recall that the PRTR database is an official database, therefore estimated can be assumed as qualified estimate.

Table 3 GSD contributions for the indicator reliability in the pedigree matrix, from the PRTR database

Indicator value	GSD contributions
1	1
2	(n.a.)
3	1,017149698
4	1,366327765
5	(n.a.)

5.2.2 Comparison of results

When comparing the GSD contributions calculated from the PRTR and the GEMIS database (Table 4), it is obvious that the estimates vary.

Table 4 Comparison of obtained GSD contributions for the indicator reliability in the pedigree matrix, from the PRTR and the GEMIS database

Indicator value	GSD contributions, PRTR	GSD contributions, GEMIS
1	1	1
2	(n.a.)	1,543529353
3	1,017149698	1,608154055
4	1,366327765	1,691120392
5	(n.a.)	43,55207485

5.2.3 Conclusions

GSD contributions obtained from two different data sources vary. This is not surprising, given the fact that the definition of measurement, of estimates asf. leaves room for interpretation, and that both databases do not fully reflect the descriptions foreseen in the pedigree matrix approach. Further, even within one database, variation is high; and finally, it is quite difficult to obtain a sound measurement of this indicator. A sound measurement of the reliability indicator would imply that identical information is once provided as obtained from measurement, and once as an estimate. It is likely that this is not the case in the analysed data, meaning that some data are always rather measured, and some are always rather estimated. In this case, variation in the analysed data would not only be caused by the reliability of the source, but also by other aspects inherent in the reported information.

To be on the safe side, the higher GSD contribution value is in principle recommended as an uncertainty factor. The factor for 5, unqualified estimate, is then extremely high. By expert judgement, this value is estimated to be “probably too high” and therefore not recommended. The respective indicator value is therefore set as “not available” (n.a.). Similarly, the drastic change from indicator score 1 to 2 (with GSD contributions of 1 to 1.6) is somewhat surprising. This value is therefore considered as explicitly “interim”[13].

[13] Please recall that all uncertainty factors provided in this text and in phase 0 of the pedigree project must not be considered as final, but will be tested and refined in further data analyses.

Table 5 Recommended uncertainty factors for the indicator reliability in the pedigree matrix

Indicator value	Uncertainty factor
1	1
2	1,54*
3	1,61
4	1,69
5	(n.a.)

*interim

5.3 Completeness

5.3.1 Data sources used

5.3.1.1 Yoghurt cup sampling study

Yoghurt cups were weighed in different supermarkets during a study. To analyse the completeness indicator, samples of these population are needed. Thus, for the analysis, 4 groups were created: 25%; 50%; 75%; 100% of each entire population. For these groups, the variance is calculated and compared to the variance of the entire population. This provides a distance of the less complete sample to the real variance.

As a sample becomes larger and closer to the full population, its standard deviation tends to be the same as the standard deviation of the entire population (Figure 18 – the different lines represent different supermarket chains). On the charts below, this leads to a difference that tends to zero. The figure shows also the 95% confidence intervals, which are largest for the 50% sample group.

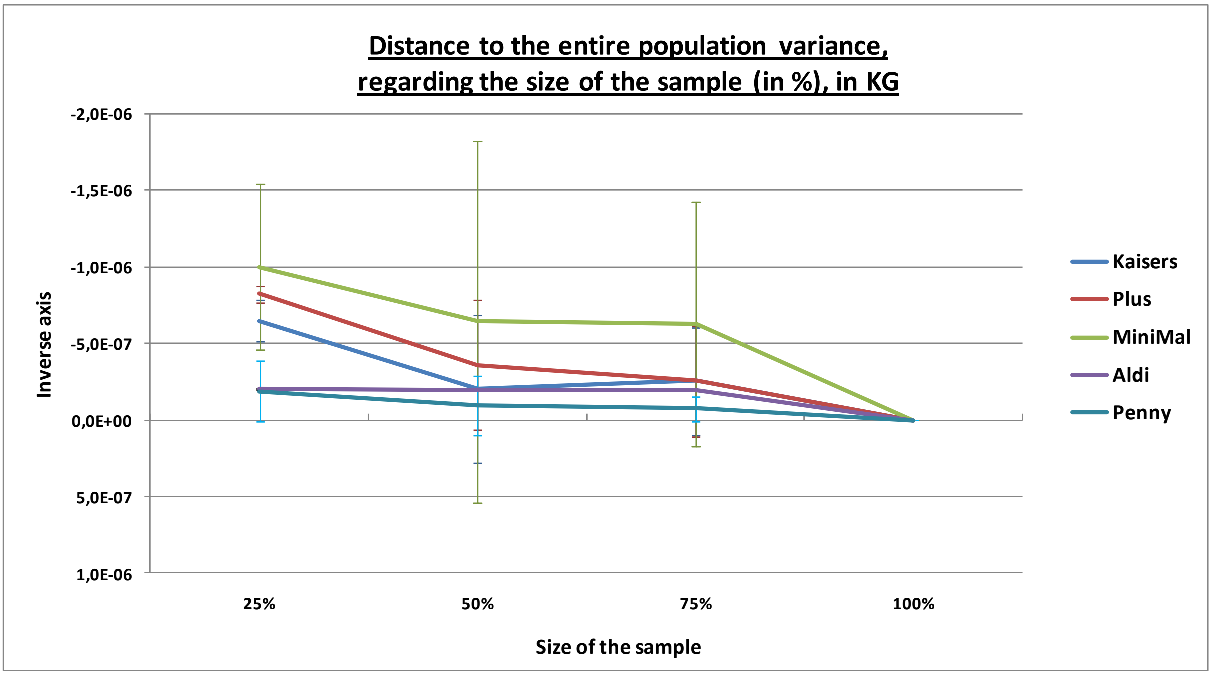


Figure 18: Difference to the variance of each population, depending on the sample size

For repeating the analysis with the indicator value for completeness in the pedigree matrix[14], 1 is taken for 100%, 2 for 75%, 3 for 50%, and 4 for only one sample, i.e. one supermarket. The result is shown in Figure 19. The higher share of 50% is taken for score 3 to compensate for a lack of representativeness in the considered supermarkets.

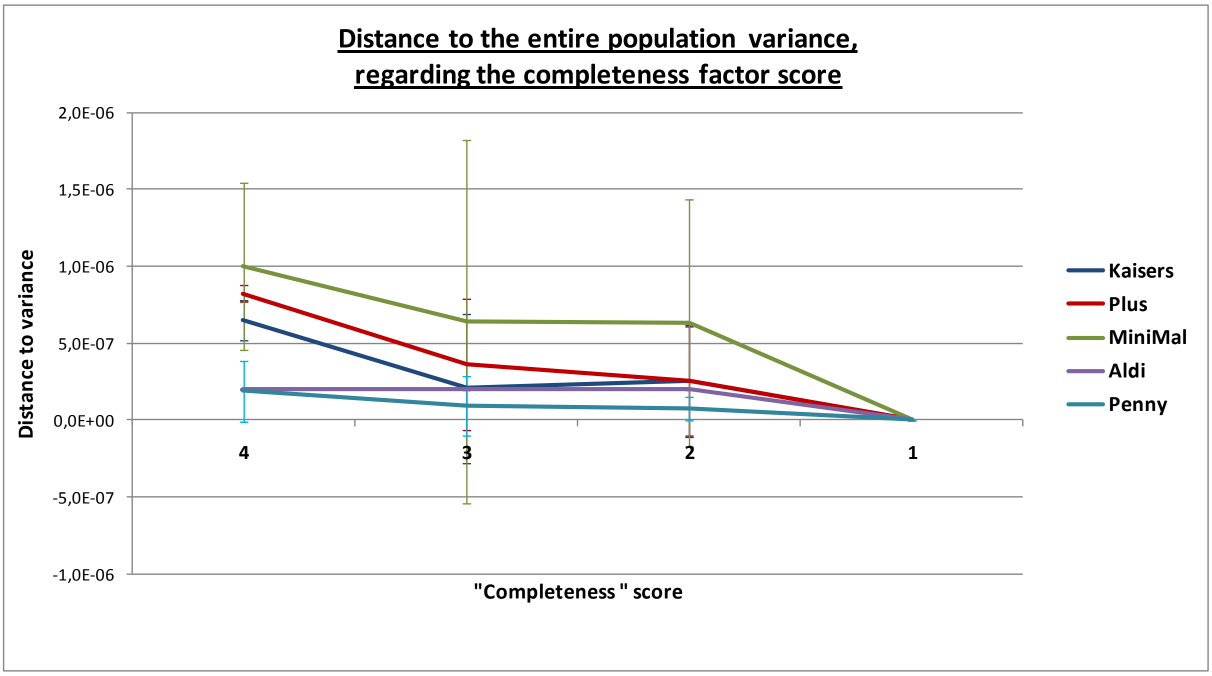


Figure 19: Effects of the "completeness indicator" on variance

Also, we have analysed the weight of the yoghurt cups using the geometric standard deviation (GSD). Considering the geometric standard deviation of each sample and the one from the

[14] Recall that he values in the pedigree matrix are:
1: 100%, 2: > 50% and assumed representative, 3: << 50 % and (assumed) representative, 4: one site but assumed representative

entire population, the "GSD$_{population}$ / GSD$_{sample}$" ratio is calculated and displayed. This allows understanding the contribution of the subsample to the overall uncertainty.

The results are fairly similar to those obtained from the analysis with the standard deviation (Figure 19). Again, there is almost no difference in the average of the values for 50% and 75% sample size, but the variation (in the GSD ratios!) increases when the sample size is reduced.

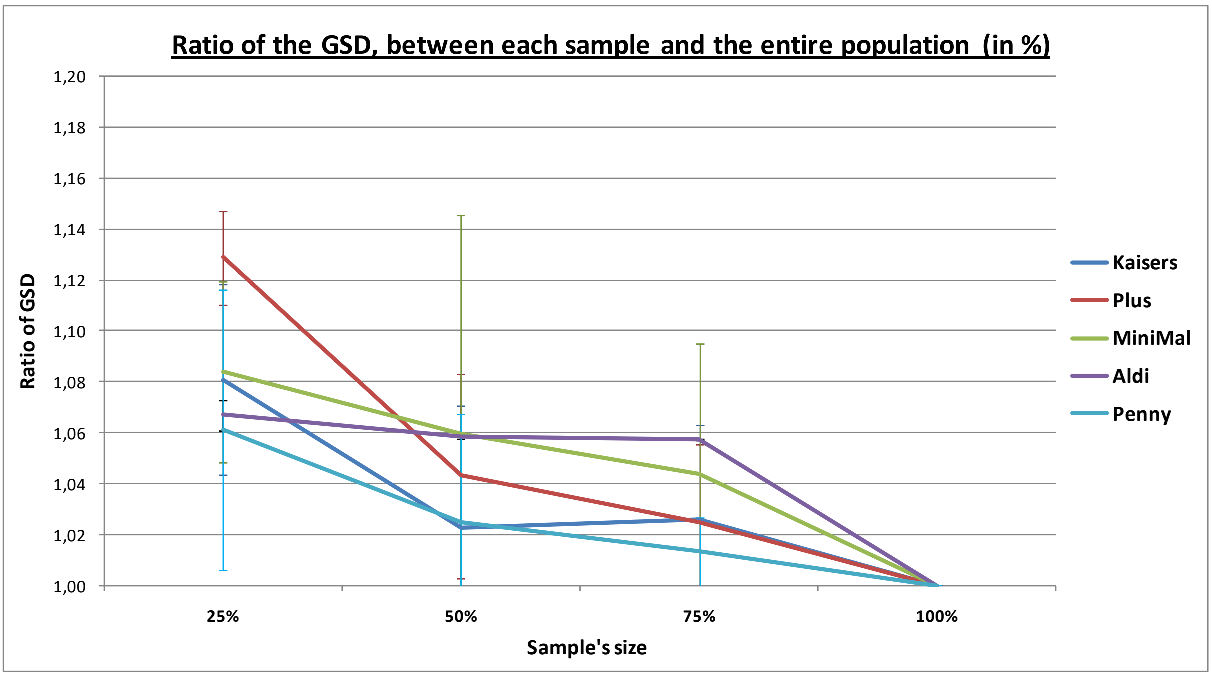


Figure 20: **Ratio of geometric standard deviations of each population, depending on the sample size**

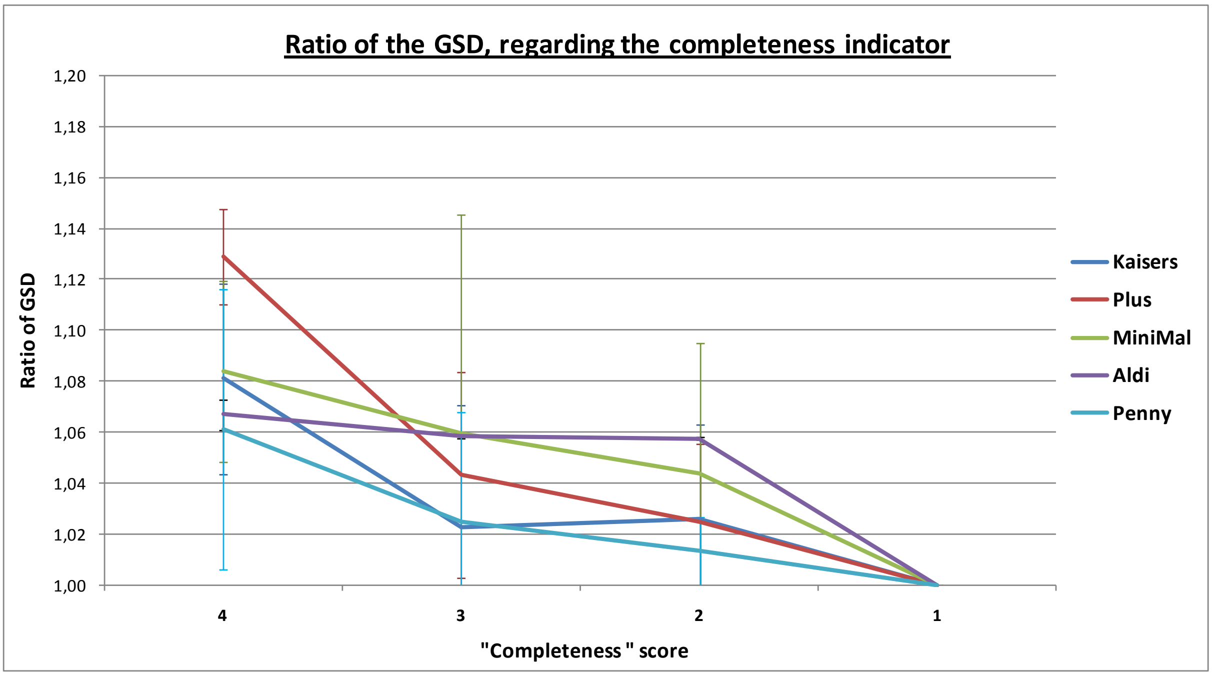


Figure 21: **Effects of the "completeness indicator" on the geometric standard deviation**

Transferring this result again to the pedigree indicator completeness score, with identical values assigned gives again a fairly similar picture as in the sample size figure (Figure 21).

Based on this figure, the following tentative uncertainty factors can be derived (Table 6), as GSD.

Table 6 Tentative uncertainty factors for the indicator completeness in the pedigree matrix, as GSD

Indicator value	Uncertainty factor
1	1
2	1,032982251
3	1,041623865
4	1,084355355
5	(n.a.)

5.3.2 Conclusions

As a sample becomes larger and closer to an entire population, the uncertainty gets smaller. This is also reflected in the uncertainty factors. The overall results seem therefore plausible, but of course, have been obtained only from one study, which analysed a really homogenous sample, 150g yoghurt cups in Berlin as they are available in different supermarkets. An interesting observation is that the uncertainty factors themselves have an underlying variation, which raises the question on how to deal with these “meta” uncertainties.

With some reservations, mainly because only one data sample has been analysed, the following tentative uncertainty factors can be proposed, for the indicator completeness, and expressed as geometric standard deviation, GSD.

Table 7 Tentative uncertainty factors for the indicator completeness in the pedigree matrix, as GSD

Indicator value	Uncertainty factor
1	1
2	1,03
3	1,04
4	1,08
5	(n.a.)

5.4 Temporal correlation

5.4.1 Data sources used

5.4.1.1 GEMIS

Because of the small amount of data older than the year 2000 in GEMIS, the analysis of the “temporal correlation” in GEMIS is mainly done with forecasted data. Figure 22 presents the effect of the temporal factor on the standard deviation.

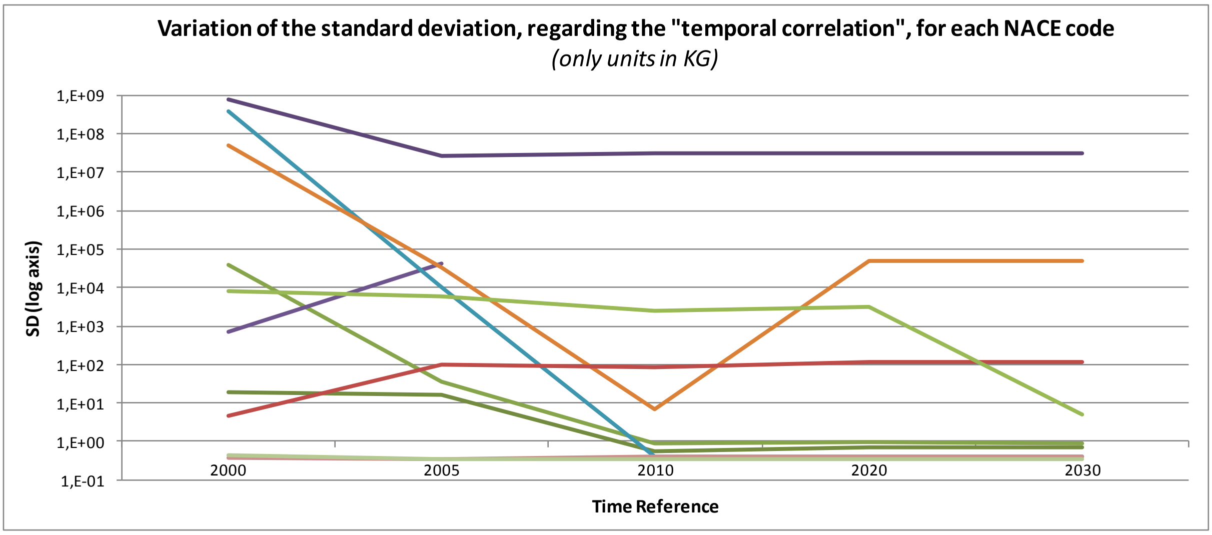


Figure 22: Variation of the standard deviation with the "temporal correlation factor"

Figure 23 considers another process unit (TJ, and not kg, leading to more homogenous processes, from energy production), and the geometric standard deviation.

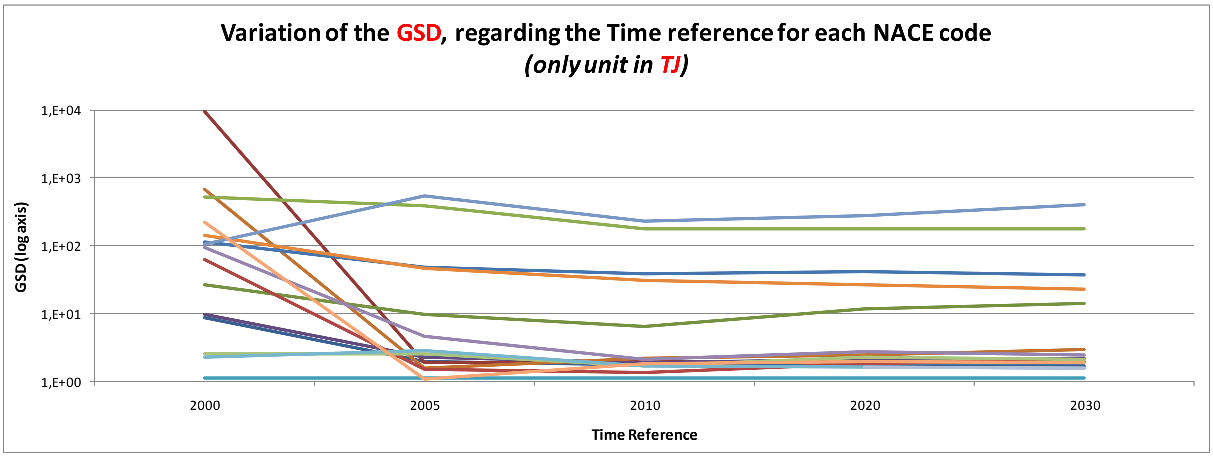


Figure 23: Variation of the geometric standard deviation GSD with the "temporal correlation factor"

There seems no obvious correlation between the "temporal correlation factor" and the variation of the (arithmetic and geometric) standard deviation.

Nevertheless, one can see that geometric standard deviations from years 2005 to 2030 are almost the same. This can be due to the forecasting calculations which are often based on the same model in GEMIS, resulting in the same deviation. This, in turn, makes the data source less valuable for the analysis of the temporal correlation factor.

5.4.1.2 North American Transportation statistics

The analysis of the "temporal correlation factor" in the NATS database shows that the standard deviation of relative pollutant emissions (yearly emissions divided by the emissions of each country in one year) varies with the year considered. This variation depends also a lot on the country and on the transport mode (Figure 24).

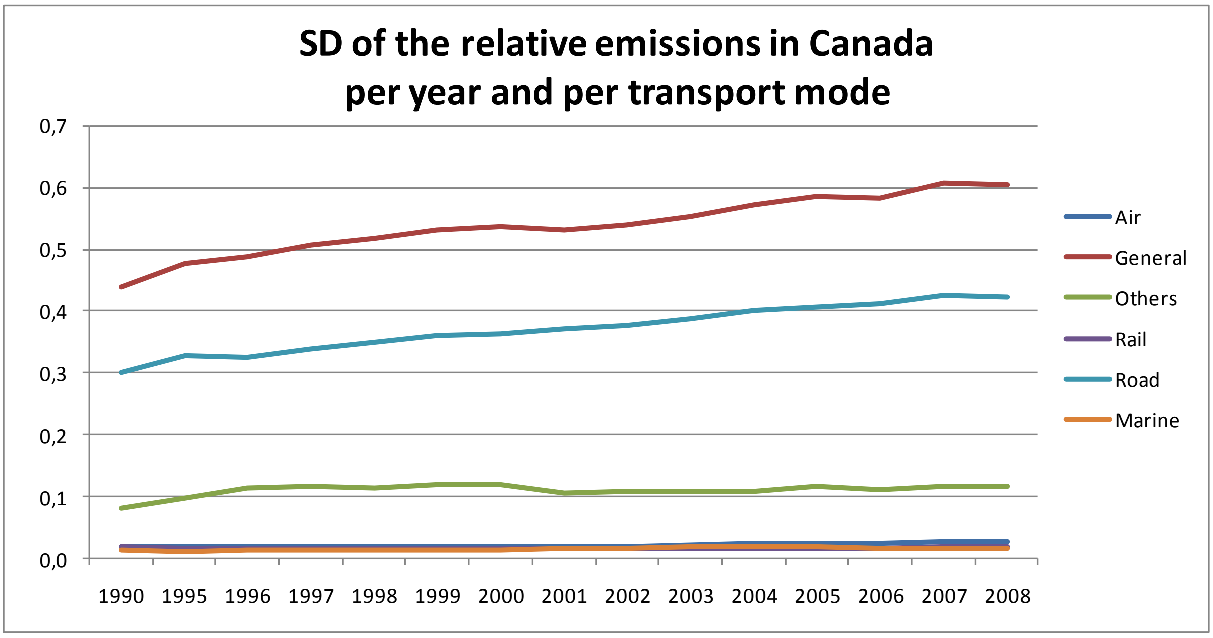


Figure 24: Standard deviation of "temporal correlation factor", in Canada

A similar chart, with the analysis of the geometric standard deviation, shows values that lead to the same conclusion.

For further analysis, the absolute emission figures are considered (and not the relative emissions, as above); further, the emission values are related to the value of the indicator temporal correlation in the pedigree matrix[15], and finally, similar to the previous analysis of the completeness indicator, the GSD is related to the reference GSD, which is the one where the indicator value is 1. Note that, since the analysis starts from the year 2008, the initial variance in the data groups is higher for later years, and lower for earlier years; therefore, the GSD ratio may be below one. In these cases, the reverse of the ratio is used, since the GSD for the larger group fails to represent a certain amount of existing uncertainty (Figure 26).

[15] 1: < 3 years, 2: < 6 years, 3: < 10 years, 4: < 15 years, 5: unkown

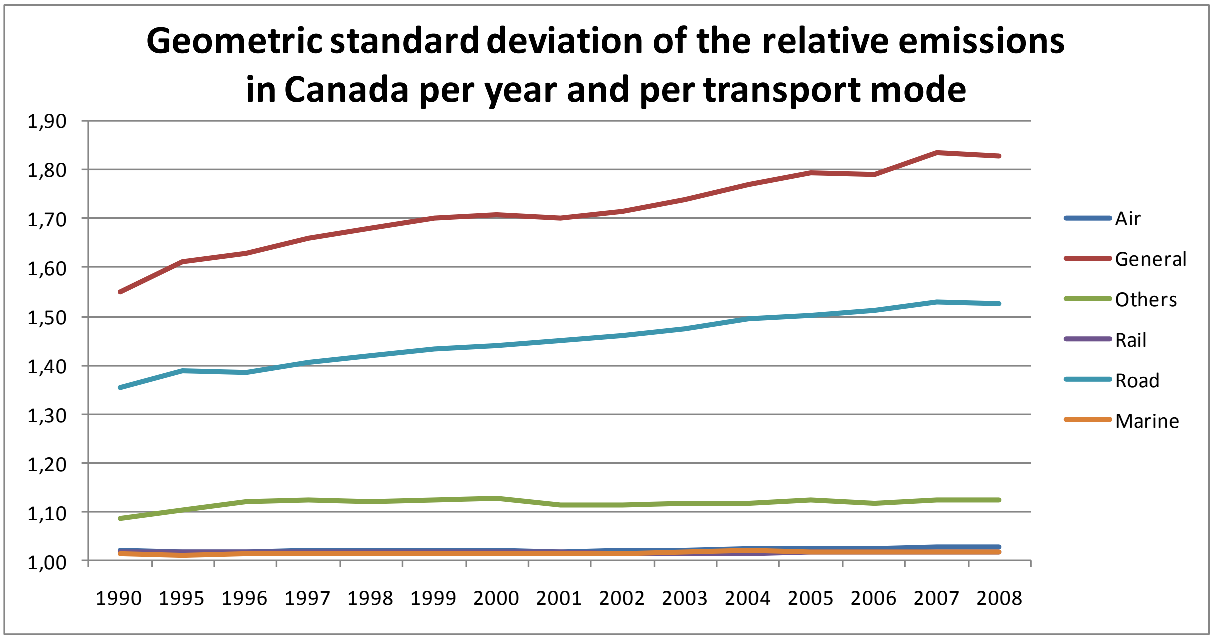


Figure 25: Geometric standard deviation of "temporal correlation factor", in Canada

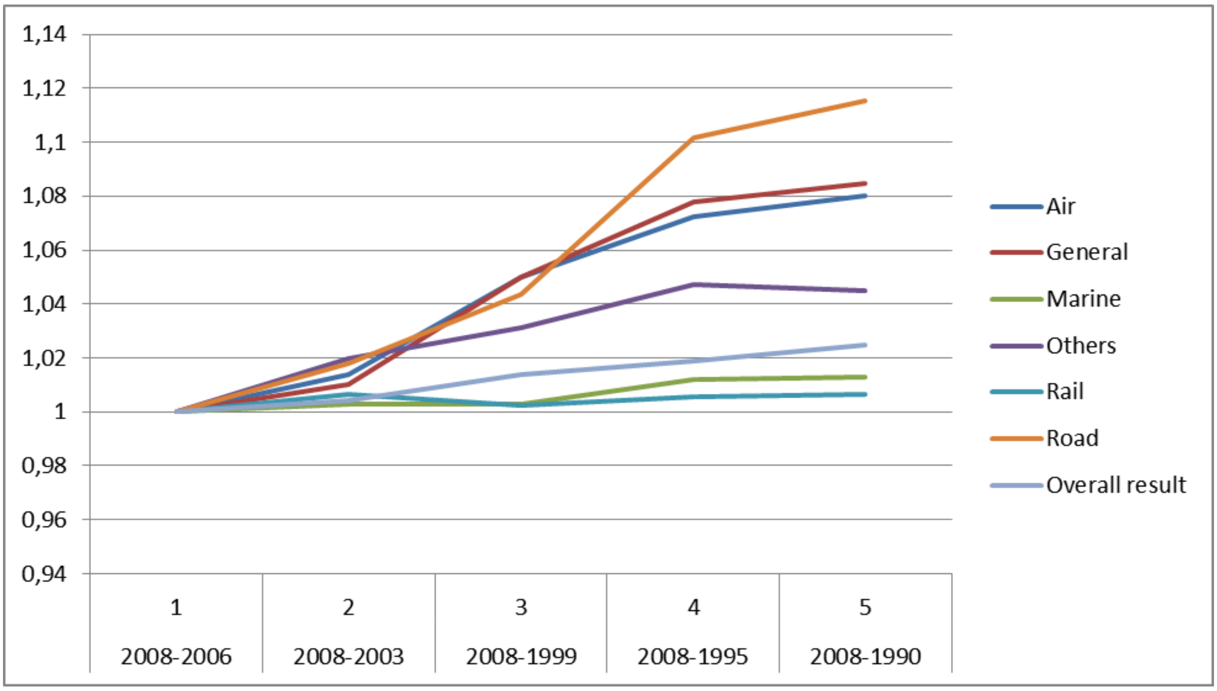


Figure 26: Relative geometric standard deviation of the "temporal correlation factor", in Canada, for different ways of transport, from the North American Transport Statistics

The figures are also represented in the table below. They show that temporal correlation depends also on the technology.

Table 8 Relative GSD per temporal correlation indicator, for different ways of transport, from the North American Transport Statistics, for Canada

temporal correlation				relative GSD = GSD contribution				
indicator		Air	General	Marine	Others	Rail	Road	Overall result
2008-2006	1	1	1	1	1	1	1	1
2008-2003	2	1,01393294	1,01038669	1,00266982	1,01970999	1,00628962	1,01787221	1,004181921
2008-1999	3	1,0501336	1,05012577	1,00290506	1,03117235	1,00218088	1,04340714	1,014037518
2008-1995	4	1,07232953	1,07792393	1,01210961	1,04717815	1,00566427	1,10154783	1,018719227
2008-1990	5	1,08035763	1,08486721	1,01269524	1,0449874	1,00634116	1,11551967	1,024956798

Results show that variation is much higher for road than for marine emissions. This is a clear indication that temporal correlation is always to some extent related to technological or other change; if everything remains identical, then also a time change does not change anything. Obviously, emissions patterns have changed less for marine transport than for road transport in Canada, from 1990 – 2008.

On the other side, this is also an indication that road transport technology might have been simply more innovative, in the regarded time span, than marine transport technology, at least regarding caused emissions.

So, in the end, the question remains which part of the change that shows in the temporal correlation should rather be considered in indicator technological correlation than in the indicator temporal correlation.

5.4.1.3 Tremod / HBEFA

The HBEFA database contains vehicle emissions (in g/km) for 5 countries, from 1990 to 2010. Figure 27 below displays the standard deviations of these emissions, per year, and per emission. The values are quite plausible; for example, N_2O uncertainties increase from 2000 onwards to a higher share of diesel cars, leading to a more heterogeneous sample. On the other side, uncertainties for lead decrease and are really low from 2000 onwards, since leaded fuel disappeared from the market for road transport.

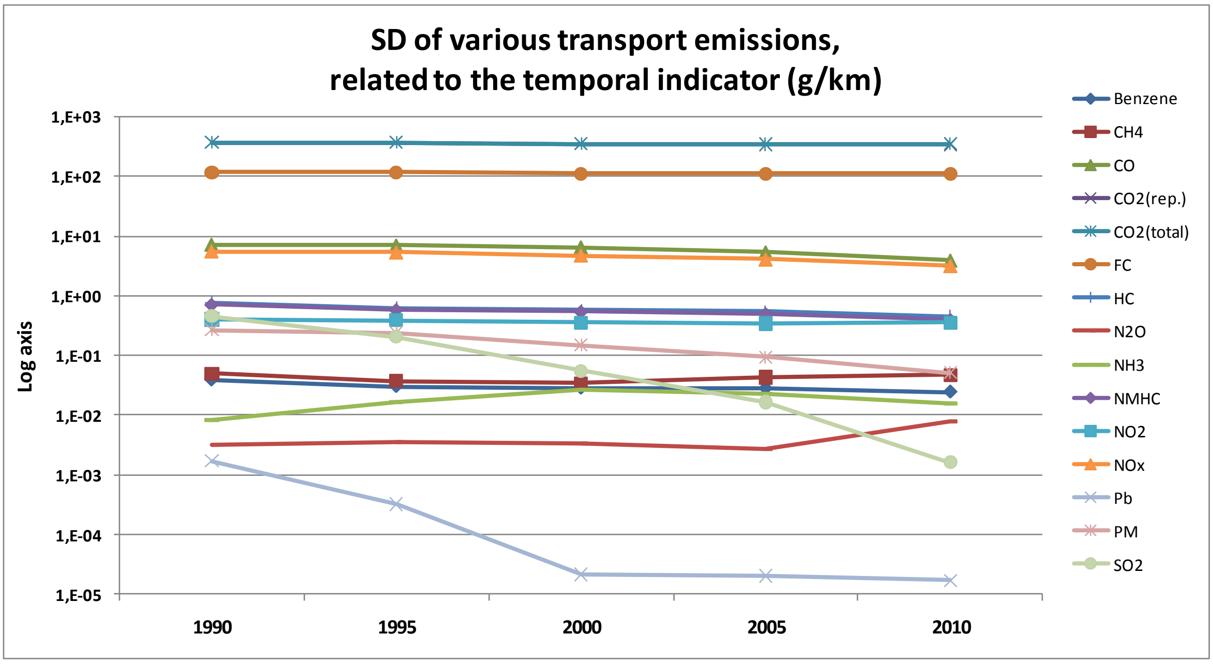


Figure 27: Analyse of the temporal correlation (Standard deviation)

Considering the geometric standard deviation of these emissions, the same chart is obtained (Figure 28). Both figures show that the (arithmetic and geometric) standard deviations vary over years, but depend also a lot on the pollutant.

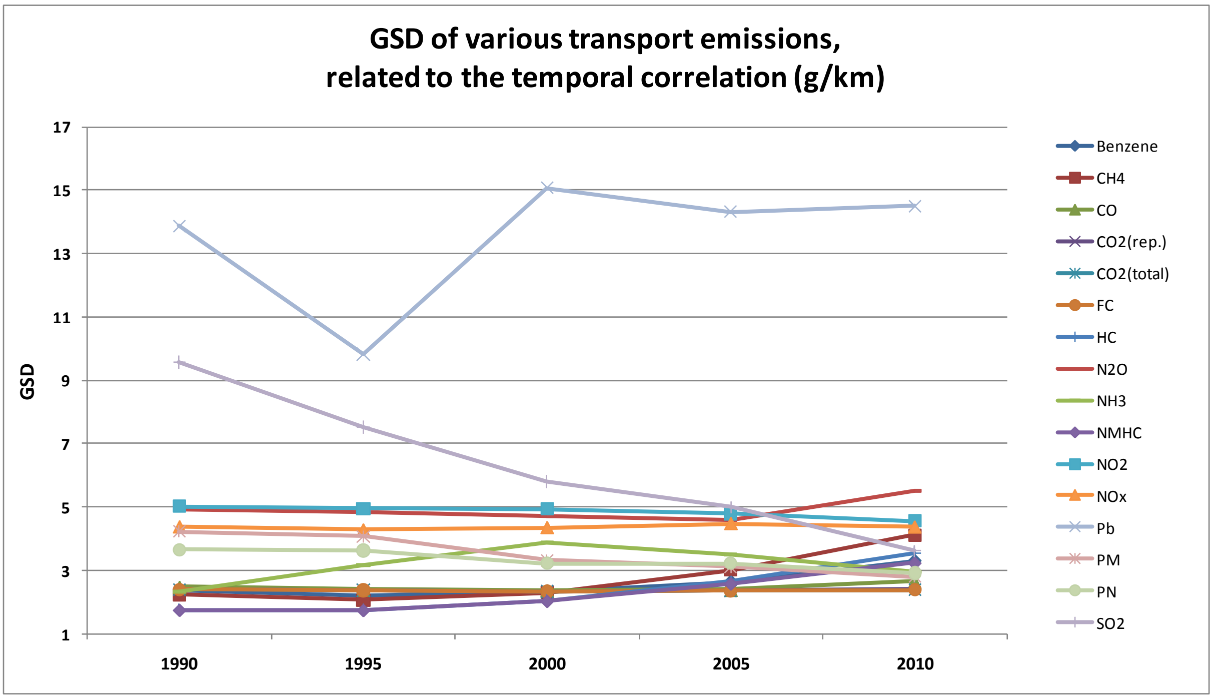


Figure 28: Analyse of the temporal correlation (Geometric standard deviation)

Considering year 2010 as reference, and, as explained in Figure 7, building groups of years depending on the difference to 2010 produces a (Figure 29). As previously, the standard deviation depends on the pollutant: SO_2 and lead change much more over the years than other emissions.

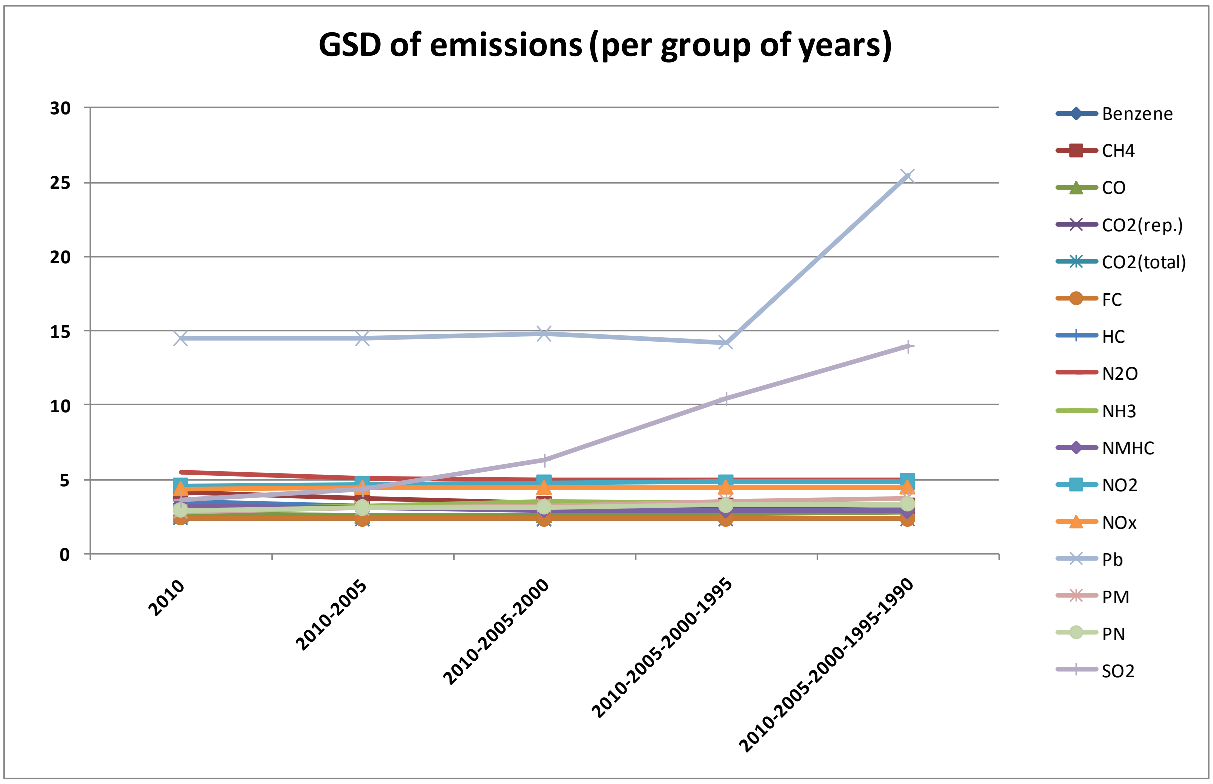


Figure 29: GSD of emissions, with 2010 as a reference

Building again a ratio of these geometric standard deviations produces the respective contributions to the overall uncertainty that can be used as input for the uncertainty factors (Figure 30).

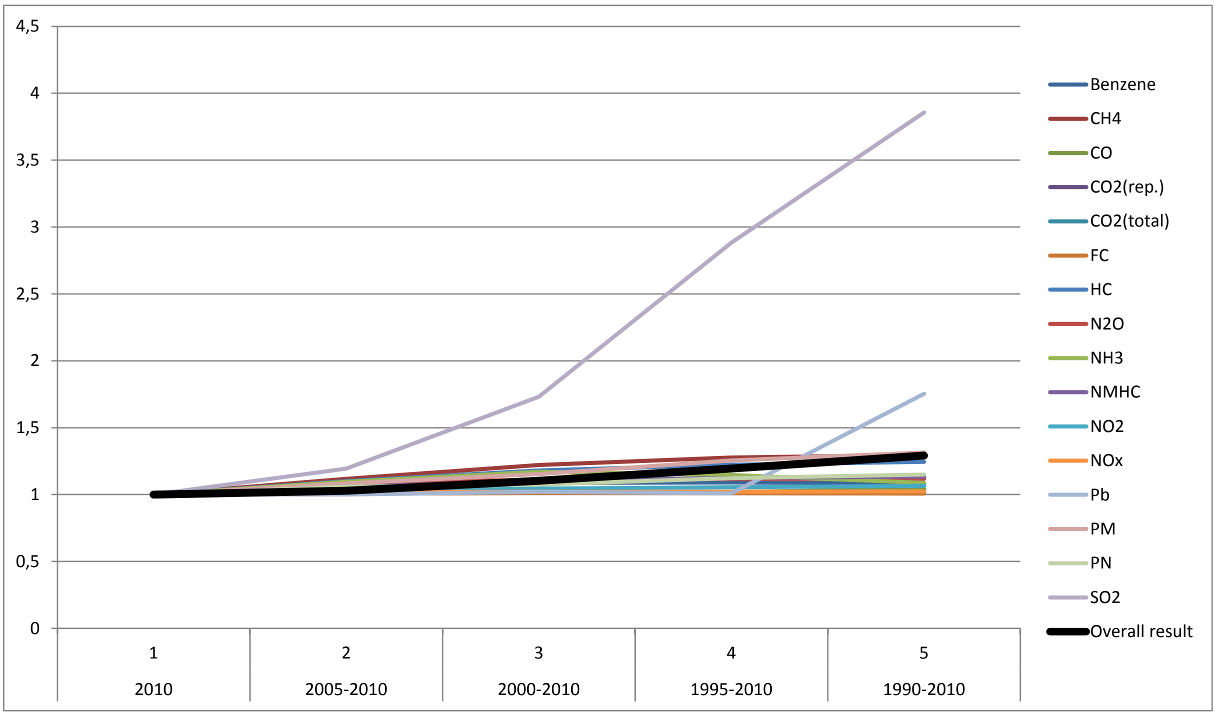


Figure 30: Ratio of GSD of emissions, with 2010 as a reference, for different values of the indicator temporal correlation

Table 9 Relative GSD per temporal correlation indicator, for different ways of transport, for different emissions, per ton-km and person-km, from the Tremod database

temporal correlation indicator		relative GSD = GSD contribution																
		Benzene	CH4	CO	CO2(rep.)	CO2(total)	FC	HC	N2O	NH3	NMHC	NO2	NOx	Pb	PM	PN	SO2	Overall result
2010	1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
2005-2010	2	1,06	1,12	1,03	1,01	1,01	1,01	1,10	1,09	1,10	1,06	1,03	1,02	1,00	1,08	1,06	1,19	1,03
2000-2010	3	1,09	1,22	1,01	1,02	1,01	1,01	1,18	1,11	1,17	1,10	1,05	1,02	1,02	1,15	1,07	1,73	1,10
1995-2010	4	1,09	1,28	1,02	1,01	1,01	1,01	1,22	1,12	1,15	1,12	1,06	1,02	1,01	1,25	1,12	2,88	1,19
1990-2010	5	1,08	1,30	1,04	1,01	1,01	1,01	1,24	1,12	1,09	1,13	1,06	1,02	1,75	1,32	1,15	3,86	1,29

5.4.2 Comparison of results, and conclusions

GEMIS does not provide meaningful uncertainty factors; factors for Tremod and for the North American transport statistics vary, with some outliers for pollutants that have been regulated in the considered time span (SO_2, lead).

Temporal correlation will in any case be a “backup” uncertainty factor, expressing the uncertainty not yet covered by further technological correlation.

As a proposal, the overall result from the analysis of the Tremod database is used, leading to the following results (Table 10).

The factors should especially be applied for situations where a variation over time can be expected that is not related to technology, i.e. variation related to geobiophysical variations or variations in population density etc., in the same place, or where changes are probably not covered by any of the other indicators in the matrix.

Table 10 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

5.5 Geographical correlation

5.5.1 Data sources used

5.5.1.1 Tremod vs. GREET

Tremod and GREET are both databases of emissions of transport means; Tremod has a European / German background, while GREET is from the US. Both databases will be used for the analysis of the pedigree indicator geographical correlation. Both report only some few emissions. Figure 31 shows the standard deviation for the emissions that can be found in both databases, for passenger fuel cars.

In order to analyse both data sources for the indicator geographical correlation, data from one of the databases can be compared with data from both databases combined; “translated” into pedigree indicator scores, this means comparing the pedigree scores, 1 to 3[16]. For the analysis, Tremod data is filtered to include only average emission concepts (instead of the variety of Euro1, Euro2 and so forth, which are also not provided in GREET), and to contain only average street conditions, similarly to GREET.

[16] Indicator geographical correlation, 1: data from area under study; 3: data from area with similar production conditions, see Figure 1.

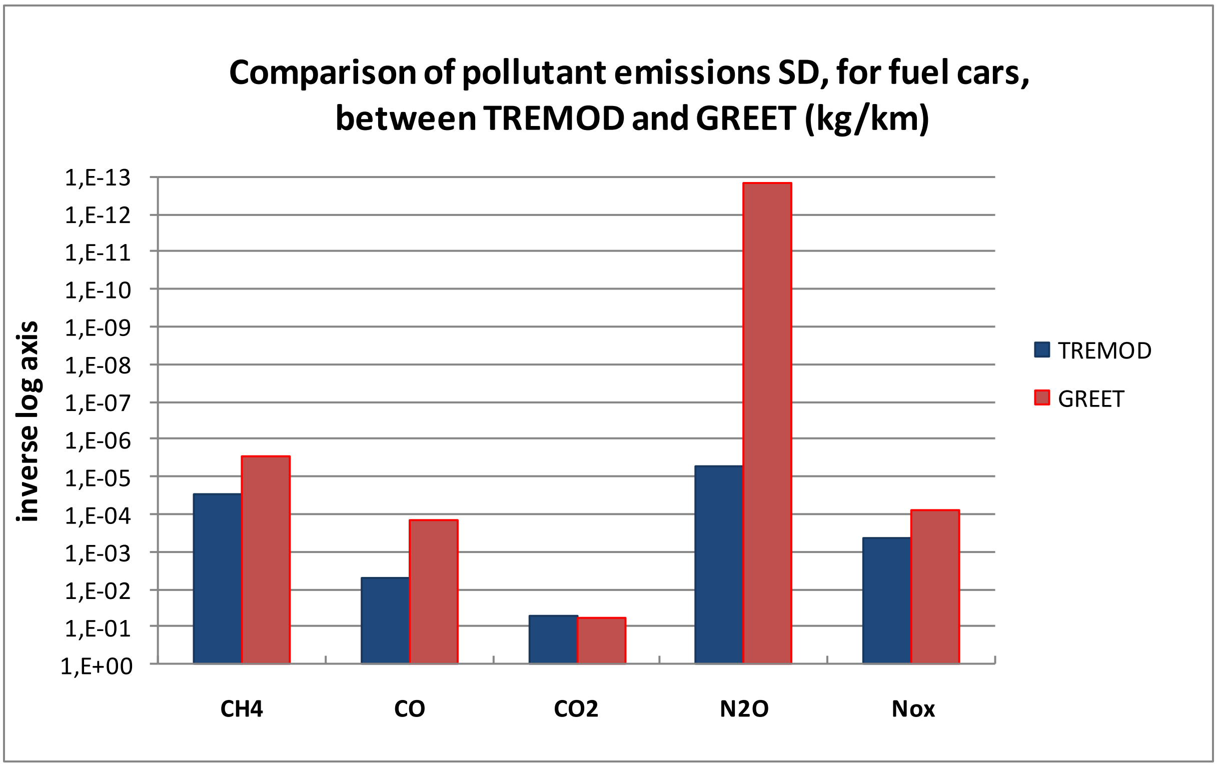


Figure 31: Comparison of German and American emission data on fuel vehicles

Still, Tremod contains many more datasets than GREET. This causes a bias in the analysis, as Tremod data “outweighs” data from GREET. As a result, with Tremod (=Europe) as a reference, the relative GSD values are a bit smaller than with GREET as a reference (Table 11).

Table 11 Relative GSD values for the indicator geographical correlation, from the analysis of GREET and Tremod

Indicator value	Relative GSD values
1	1
3	1,020439873*
3	1,032117664**

* with Tremod as reference

**with GREET as reference

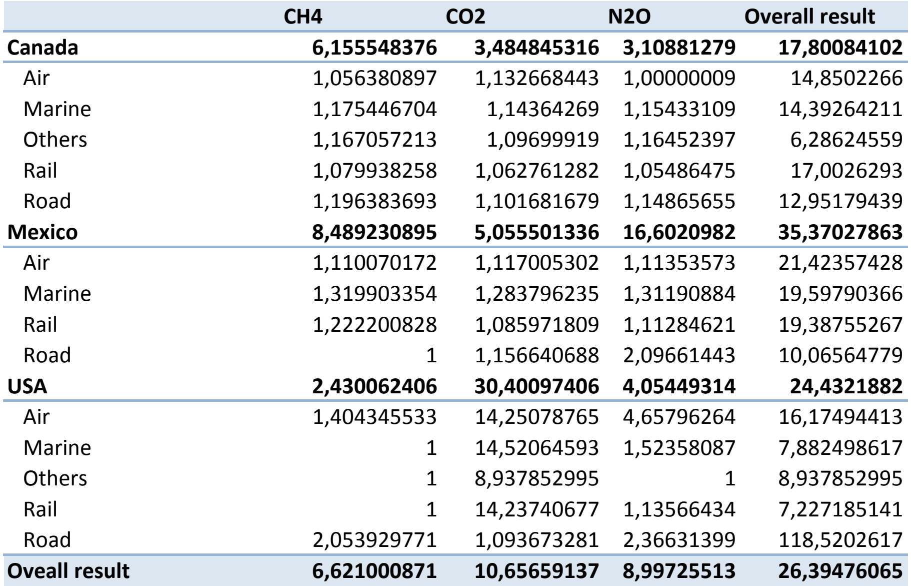
5.5.1.2 North American Transportation Statistics

The North American Transportation Statistics contains data from the USA, Canada and Mexico, and is therefore also suited for analysing the geographical correlation indicator.

The database contains total emissions per country, per mode of transport, per year, and for several airborne emissions: CH_4, CO2, N_2O.

Table 12 shows an overview of the GSD values over all years.

Table 12 GSD values from the North American Transport Statistics database



	CH4	CO2	N2O	Overall result
Canada	**6,155548376**	**3,484845316**	**3,10881279**	**17,80084102**
Air	1,056380897	1,132668443	1,00000009	14,8502266
Marine	1,175446704	1,14364269	1,15433109	14,39264211
Others	1,167057213	1,09699919	1,16452397	6,28624559
Rail	1,079938258	1,062761282	1,05486475	17,0026293
Road	1,196383693	1,101681679	1,14865655	12,95179439
Mexico	**8,489230895**	**5,055501336**	**16,6020982**	**35,37027863**
Air	1,110070172	1,117005302	1,11353573	21,42357428
Marine	1,319903354	1,283796235	1,31190884	19,59790366
Rail	1,222200828	1,085971809	1,11284621	19,38755267
Road	1	1,156640688	2,09661443	10,06564779
USA	**2,430062406**	**30,40097406**	**4,05449314**	**24,4321882**
Air	1,404345533	14,25078765	4,65796264	16,17494413
Marine	1	14,52064593	1,52358087	7,882498617
Others	1	8,937852995	1	8,937852995
Rail	1	14,23740677	1,13566434	7,227185141
Road	2,053929771	1,093673281	2,36631399	118,5202617
Oveall result	**6,621000871**	**10,65659137**	**8,99725513**	**26,39476065**

Setting Canada as a reference country (geographical correlation = 1), USA & Canada together as larger area including the area under study (geographical correlation = 2), and, finally, Canada, USA and Mexico together as area with similar production conditions (geographical correlation = 3) allows us to calculate the following relative GSD contribution values (Table 13).

Table 13 Relative GSD values for the indicator geographical correlation, from the analysis of the North American Transport Statistics Database, with Canada as reference

Indicator value	Relative GSD values
1	1
2	1,159084043
3	1,482781663

These values are much higher than the one from Tremod. However, the data source here is less suited for the analysis, as it reports total emissions over all means of transport, over many years, and therefore contains also additional uncertainty due to changes in transport patterns and emission regulations that are not be considered in the geographical correlation indicator.

5.5.1.3 E-PRTR

Emissions included in the E-PRTR database are defined per year and per plant, for European countries. All the countries contained in the database are listed in Table 14.

Table 14 Relative GSD values for the indicator geographical correlation, from the analysis of the North American Transport Statistics Database, with Canada as reference

Country	Indicator score group			
	1	2	3	4
Austria			X	x
Belgium		x	X	x
Bulgaria				x
Cyprus				x
Czech Republic			X	x

Country	Indicator score group			
	1	2	3	4
Denmark			X	x
Estonia				x
Finland			X	x
France		x	X	x
Germany	x	x	X	x
Greece				x
Hungary				x
Iceland				x
Ireland				x
Italy			X	x
Latvia				x
Lithuania				x
Luxembourg			X	x
Malta				x
Netherlands			X	x
Norway			X	x
Poland		x	X	x
Portugal			X	x
Romania				x
Slovakia				x
Slovenia				x
Spain			X	x
Sweden			X	x
Switzerland			X	x
United Kingdom			X	x

In order to analyse the geographical correlation indicator, countries are grouped; Germany is taken as reference and hence the only member of group 1; countries covering a larger area including Germany are group 2 (Belgium, France, Poland); slightly similar production conditions provide many different countries, as group 3, and finally, some countries such as Malta, Latvia, Iceland, Bulgaria are only member of group 4[17].

For this grouping, the following relative GSD values, calculated as in the previous sections, are obtained (Table 15).

[17] This grouping is of course to some extent arbitrary.

Table 15 Relative GSD values for the indicator geographical correlation, from the analysis of the PRTR Database, with Germany as reference

Indicator value	Relative GSD values
1	1
2	1,043919013
3	1,082233009
4	1,105217922

The analysis ignores any existing differing production patterns in the countries; also, differing scales in countries (caused by different sizes of plants) are not considered. The sample size is much larger than in the case of the North American Transport database, where only one reported figure exists per year and mode of transport and pollutant. In the analysed PRTR database, there are more 100,000 data sets. It is somewhat surprising that although many different industrial sectors are considered, the resulting GSD contributions are smaller than in the case of the transport database.

5.5.2 Comparison of results

The results obtained from the analysed sources are quite different (Table 16). Especially the North American Transport Database proposes a really large GSD contribution for an indicator of 3, data from similar production conditions. This seems not really logical. The figure from the Tremod and Greet comparison, on the other side, for the same indicator, is rather small, with 1,03 as a maximum.

Table 16 Comparison of obtained GSD contributions for the indicator geographical correlation in the pedigree matrix, from the analysed sources

Indicator value	Tremod / GREET	North American Transport Statistics Database	PRTR
1	1	1	1
2	(n.a.)	1,159084043	1,043919013
3	1,020439873* / 1,032117664**	1,482781663	1,082233009
4	(n.a.)	(n.a.)	1,105217922
5	(n.a.)	(n.a.)	(n.a.)

* with Tremod as reference
**with GREET as reference

5.5.3 Conclusions

The values from the analysis of the PRTR database are proposed as tentative uncertainty factor values (Table 17). However, the indicator value of 5 (data from unknown area) could not be analysed with the available data so far. Also, correlations in data seem to ask for a more refined, multivariate analysis, which is possible work for later.

Table 17 Tentative uncertainty factors for the indicator ‘geographical correlation’ in the pedigree matrix, as GSD

Indicator value	Uncertainty factor
1	1
2	1,04
3	1,08
4	1,11
5	(n.a.)

5.6 *Further technological correlation*

5.6.1 Data sources used

5.6.1.1 Tremod

The Tremod database contains information on the technology of each of means of transport that is part of the database, which makes it suitable for analysing the “further technological correlation” indicator.

As an overview, Figure 32 shows the change in standard deviation, over all transport vehicles in the database, and for all provided pollutants that are available, when the considered sample is narrowed down more and more, by specifying more and more of the technology aspects of these vehicles. Note that “technology” includes here also use patterns, for example the type of road that is used (motorway, city street, …). The figure shows that the average standard deviation (and therefore the uncertainty) decreases, the more precise the sample is specified. This is in line with the uncertainty concept introduced in the introduction.

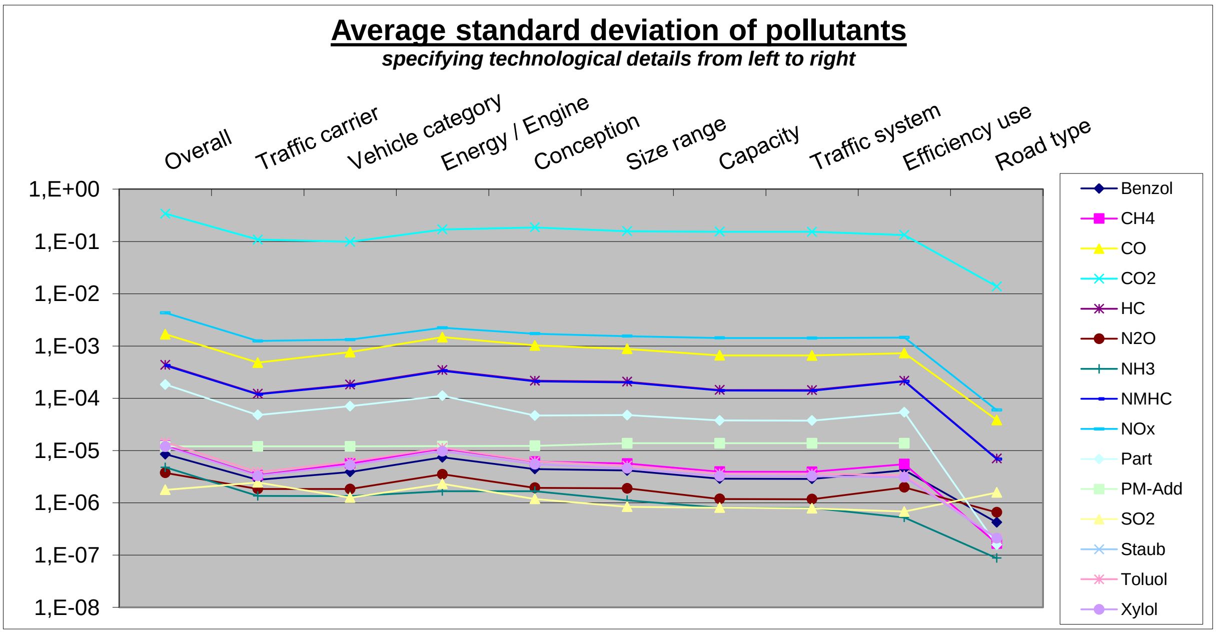


Figure 32: Standard deviation of pollutants

In order to analyse the contribution to the technological correlation indicator of the pedigree matrix, differences in the data sets need to be mapped to the indicator scores (Table 18).

Table 18 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, 1.4-2l 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 (0-1.4l, 2-9l), 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)

*see Figure 1; obviously the “under study” specifications do not help a lot in the approach used here, since the more precise data “under study” is always, in the approach used in this text, part of the larger, less precise data sample.

The analysis uses as reference a 1.4-2l capacity gasoline (diesel) personal car, with emissions in line with the EURO 4 standard, and operated for inner city personal transport.

As for the other indicators, specification of this data set are more and more relaxed, leading to a less precisely defined group of data sets, with increased uncertainty.

It may be problematic to determine the specific indicator levels in practice; for example, for the analysed Tremod database, and a personal car, are busses also “related” processes? In the end, they provide also personal transport, but of course are rather different from a car. But a gas-driven car is also different from a car that runs on diesel.

Different ways to use a product (driving a car inner city or on a motorway) are not covered in any of the pedigree indicators but can of course influence results; they are addressed, in this text, under the technological correlation indicator, but this use of the indicator is not explicitly mentioned or “authorised” by ecoinvent.

With the distinctions for the different indicator levels given in Table 18, the following results are obtained (Table 19).

Table 19 Relative GSD values for the indicator further technological correlation, from the analysis of the Tremod Database

Indicator value	Relative GSD values
1	1
2	1,177388503
3	1,653132597
4	2,081454983
5	2,799827129

5.6.1.2 GREET

The GREET database contains also information about passenger cars technology and can therefore be used for the analysis of the ‘further technological correlation’ indicator. Compared to the Tremod database, GREET contains fewer datasets, with fewer technological differentiation.

Table 20 Mapping of indicator scores for “technological correlation” to differences in data sets in the GREET 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	Baseline LDT (light duty truck) Vehicle: CG and RFG
2	Data from processes and materials under study (i.e. identical technology) but from different enterprises	LDT (light duty truck) Vehicle: all gasoline-driven LDTs
3	Data from processes and materials under study but from different technology	LDT (light duty truck) Vehicle: all LDTs (including diesel-powered LDTs)

*see also the comment under Table 18.

For personal cars, there are fewer data in the data base (or rather, the reported emission factors are often all equal, independent from the type of car, which makes the data sets not suitable for an analysis).

The following table gives an overview of the results, obtained from the analysis of the following emission factors in the database: CO, PM10: brake and tire wear, PM10: exhaust, PM2.5: brake and tire wear, VOC: evaporation, VOC: exhaust, CH4, CO2, N2O, NOx, SOx. The database contains several additional indicators, but these strongly correlate with the analysed indicators in the list, e.g. GHG emissions, and “CO2 (w/ C in VOC & CO)”.

Table 21 Relative GSD values for the indicator further technological correlation, from the analysis of the GREET Database

Indicator value	Relative GSD values
1	1
2	1,019808156
3	1,069030753

5.6.2 Comparison of results

A comparison of relative GSD contributions from GREET and Tremod shows that GREET values are smaller; this can, at least in parts, be explained by GREET emission factors that are constant over all types of vehicles in the database. GSD contributions from Tremod, on the other side, are very high, especially for higher indicator values. This is, however, also logical, since data from processes with “different technologies”, as foreseen for the indicator value 5, can indeed vary a lot from the targeted data set.

Table 22 Comparison of obtained GSD contributions for the indicator further technological correlation in the pedigree matrix, from the analysed sources

Indicator value	Tremod	GREET
1	1	1
2	1,177388503	1,019808156
3	1,653132597	1,069030753
4	2,081454983	(n.a.)
5	2,799827129	(n.a.)

5.6.3 Conclusions

The values from the analysis of the Tremod database are proposed as tentative uncertainty factor values (Table 17), since the lower GREET values are to some extent caused by emission factors identical over all data sets in the database, which seem not realistic.

Defining the thresholds for the indicator value can be discussed, it less clear than for the other indicators. More guidance would be helpful.

Also, further analysis, with data from other sources, would help to better found the determined uncertainty factor.

Table 23 Tentative uncertainty factors for the indicator 'further technological correlation' in the pedigree matrix, as GSD

Indicator value	Uncertainty factor
1	1
2	1,18
3	1,65
4	2,08
5	2,80

5.7 Basic uncertainty factors

Basic uncertainty factors express uncertainty that is not dependent on one or several of the indicators in the pedigree matrix, but can be seen as "inherent" in a specific flow. Carbon emissions usually are known more precisely than specific toxic emissions, for example, even if all the indicator scores for the pedigree matrix are the same.

The basic uncertainty factors currently used by ecoinvent are given in Figure 2, above.

For deriving empirically based basic uncertainty factors, a similar approach should be followed as used in the previous sections, for pedigree indicator scores. However, here, the approach is more demanding to apply. For the pedigree scores, the score 1 can be used as a reference, and uncertainty for the other scores can be expressed in relation to this reference. Following this approach for basic uncertainty factors, a reference would consist of an "as-best-as-possible" measurement of the specific flow. These values and their inherent uncertainty would then need to be compared to values and uncertainty obtained for the best score in the pedigree matrix, and this comparison would, in turn, provide an estimate for the basic uncertainty factor for the analysed flow.

Such "as-best-as-possible" measurement data are hardly available today. The basic uncertainty factors as they are currently used in ecoinvent, on the other side, are in the range of the empirically derived uncertainty factors for pedigree scores, and seem not completely out of range.

Therefore, as intermediate solution, it is recommended to keep the currently used basic uncertainty factors, and, at the same time, to identify suitable data sources for applying the approach developed for the pedigree score indicators. This work could be done as part of an uncertainty project commissioned by The Sustainability Consortium, to the UNEP/SETAC working group on uncertainty, which started in February 2012. One of three tasks in this project is to find ways for quantifying uncertainty "on the input side" of LCA[18].

6 Archetype processes seed list

The analyses conducted in this report focus on transport processes and on plants from a large European emission database. They did not show typical differences between processes, that would allow distinguishing them by type, and to group them into "process archetypes".

A broader analysis of more, different data bases could reveal that such a distinction makes sense; if this is the case, then the uncertainty factors should be distinguished by type of

[18] For details see http://lca-data.org:8080/web/uncertainty-working-group/home; the input related task of the project is lead by the author.

processes. The current analysis did not indicate that this makes sense, for the selection of investigated data sources.

7 Summary and conclusions for the uncertainty factors in ecoinvent v3

In summary, the following uncertainty factors have been identified in the analysis (Table 24):

Table 24 Summary of tentative uncertainty factors for all pedigree matrix indicators, as GSD

Indicator score	1	2	3	4	5
Reliability	1	1,54*	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

*interim

Compared to the previously used uncertainty factors, these factors are not so different, with some exceptions: Reliability with a score of 2 has a higher factor, but us seen as “interim” already in this preliminary list of factors that will anyhow be further investigated. The new factor for further technological correlation is also considerably higher than the “old” one. Factors for completeness are surprisingly similar to the old ones.

These factors have been obtained from a rather small data base of several few data sources, mostly from the transport sector. It is recommended to further expand the analysis to other data sources. Correlations were not explicitly addressed in the analysis, but rather, selected analysis steps tried to avoid cases of visible, strong correlation in data, between two different pedigree indicator scores. A more refined correlation analysis would be an interesting future task.

Some of the indicator score definitions were difficult to determine in data sets; especially the different score values for the ‘further technological correlation’ indicator should be formulated more clearly, and supported by examples (when is a process ‘similar’? when is a technology different?).

Basic uncertainty factors could not be investigated in the course of this phase 0 project. The previously used basic uncertainty factors seem not unlikely. Therefore, before empirically better founded basic uncertainty factors are available, it is recommended to use the basic uncertainty factors from ecoinvent version 2 (Figure 2).

The developed factors need to be translated to the different possible and existing probability distributions in ecoinvent data. The geometric standard deviation used for deriving the factors is useful due to its stability against different scales in the analysed data, but is not available for distributions other than the lognormal.

8 Mathematical formula for calculating uncertainty for non-lognormal distributions

Stéphanie Muller

8.1 Notations

U_b: the basic uncertainty factor expressed as the square of a geometric standard deviation

U_i: the additional uncertainty factors expressed as the square of a geometric standard deviation

X_P: the parameter or the distribution representing the additional uncertainty

X_{wP}: the parameter or the distribution representing the total uncertainty

X_{MC}: the parameter obtained through a Monte Carlo simulation

σ_g: the geometric standard deviation

CV: the coefficient of variation (ratio between the standard deviation and the mean)

PDF: probability density function

a: the minimum of a PDF

b: the maximum of a PDF

m: the most likely value

μ: the arithmetical mean

σ: the arithmetical standard deviation

ε: the relative error

8.2 Census of the distribution that are foreseen in ecoinvent v3

Seven different probability distributions can be chosen to model a flow with its uncertainty in ecoEditor2 (plus one “Undefined probability function”). The uncertainty of each flow is defined by a specification of the assumed distribution function, required parameters that give information on the central tendency, the so-called “basic uncertainty” and its pedigree scores, which are transformed to “additional uncertainty factors”.

In ecoinvent v2, the lognormal distribution is the distribution used “by default”. To take a census of the data that are not lognormally distributed, an export from the ecoinvent database has been attempted (via the Simapro v.7.2 software). 121,152 economical and elementary flows were scrutinized. The results are confined inTable 25, 70% of the flows are modeled through a lognormal distribution.

No commonality was found between the flows for which a normal distribution had been assumed.
For flows that had been assigned a triangular distribution, the pedigree scores had not been filled in. The minimum, maximum and most likely were known, pedigree scores are 5 by default but they are not considered in the determination of the total uncertainty.

Table 25 Number of flows for each distribution types

Distribution type	Number of flows
Lognormal	85,631
Normal	28
Triangle	5
Undefined	35,488

Table 26 Definition of each probability density function foreseen in ecoinvent v2 and their corresponding parameters

Distribution type	Required parameters	PDF
Lognormal	Geometric mean value (μ_g) Basic uncertainty ($SD_{g95}=\sigma_g^2$)	$f(x,\mu_g,\sigma_g)=\dfrac{\exp(\dfrac{-(\ln x-\ln\mu_g)^2}{2\ln^2\sigma_g})}{\sqrt{2\pi}\ln\sigma_g}$
Normal	Arithmetic mean value (μ) Basic uncertainty (2σ)	$f(x,\mu,\sigma)=\dfrac{\exp(\dfrac{-(x-\mu)^2}{2\sigma^2})}{\sigma\sqrt{2\pi}}$
Triangular	Minimal value (a) Maximal value (b) Most likely value (c)	$\begin{cases} f(x,a,b,c)=\dfrac{2(x-a)}{(b-a)(c-a)} & for \quad a<x<c \\ f(x,a,b,c)=\dfrac{2(b-x)}{(b-a)(c-b)} & for \quad c<x<b \\ f(x,a,b,c)=0 \quad otherwise \end{cases}$
Uniform	Minimal value (a) Maximal value (b)	$\begin{cases} f(x,a,b)=\dfrac{1}{b-a} & for \quad a<x<b \\ f(x,a,b)=0 \quad otherwise \end{cases}$
Beta **PERT**	Minimal value (a) Maximal value (b) Most frequent value (c)	$f(x,a,b)=\dfrac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)(b-a)^{\alpha+\beta-1}}$ with $\alpha=6\dfrac{\mu-a}{b-a} \quad and \quad \beta=6\dfrac{b-\mu}{b-a}$ $\mu=\dfrac{a+4c+b}{6}$
Gamma	Shape (k) Scale (θ)	$f(x,k,\theta)=\dfrac{x^{k-1}\exp(-k/\theta)}{\Gamma(k)\theta^k}$
Erlang	Mean value (kλ) Order (λ)	$f(x,k,\lambda)=\dfrac{\lambda^k x^{k-1}e^{-\lambda x}}{(k-1)!}$
Undefined	Minimal value Maximal value	*Defined as a range*

Table 26 takes a census of all distributions that can be chosen to model a flow with its uncertainty in ecoinvent v2; it shows the required parameters to define each probability density function.

Seeing this table, a first remark can be made:

We observe that the PDFs of the gamma and the Erlang distribution seem not be so different. If k is an integer, $\Gamma(k)=(k-1)!$. If, moreover, $\theta=\lambda^{-1}$ in the gamma distribution, we have defined the Erlang PDF. The only difference between the two PDFs is also in the characteristic of the k parameter.

We must therefore question the advantage of keeping these two PDFs in the ecoEditor?

The advantage of the Erlang distribution used to be computational time, which can hardly be relevant in the case of the ecoinvent database.

Moreover, a search was done in Inspec and Compendex databases with "Erlang distribution" as a title term. Only 25 papers were found, and the topic of these papers was mainly queue systems (Wireless, stocks control…).

Considering these two aspects, we suggest that it is not relevant to keep both distributions for the ecoinvent database, and that the Erlang distribution should be the one that is eliminated.

The objective of the following sections is to develop a way to model a flow – not lognormally distributed – with its total uncertainty based on what happen for the lognormal distribution.

8.3 Assumptions derived from the use of the pedigree approach with a lognormal distribution

It has been previously seen that to determine the total uncertainty for a flow lognormally distributed, the following formula is applied:

Equation 8-1

$$SD_{g95wP} = \exp\left(\sqrt{\ln^2 U_b + \ln^2 U_1 + \ln^2 U_2 + \ln^2 U_3 + \ln^2 U_4 + \ln^2 U_5}\right).$$

This formula is linked to the determination of the geometric standard deviation of the multiplication of independent lognormally distributed variables. Equation 8-1 shows that the additional uncertainty increases the dispersion of the original data with its basic uncertainty.

In fact, suppose we have R and S two random variables logrnormally distributed, their product T=R*S is also lognormally distributed, the geometric mean is given by the product of R and S geometric means and the geometric standard deviation is obtained through the following formula:

Equation 8-2

$$\sigma_{g_T} = \exp\left(\sqrt{\ln^2\sigma_{g_R} + \ln^2\sigma_{g_S}}\right).$$

The application if the pedigree approach to other PDFs will be directly based on how it is applied to the lognormal distribution and will follow the four rules:

- The additional uncertainty must modify neither the median (or mode where applicable) value nor the type of distribution chosen to represent the data;

- 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 expresses a relative dispersion;
- The additional uncertainty factors used for the lognormal distribution are used to derive the additional uncertainty for other PDFs.

The development of the formulas for the other distributions will be based on these rules.

8.4 Dimensionless measure of variability and total uncertainty compilation

8.4.1 From the multiplicative to the additive world

Considering X, a lognormally distributed random variable, the random variable Y=ln(X) is normally distributed and we have: X ↪ LN ($\mu_{\log}$, $\sigma_{\log}$) and Y↪ N ($\mu_{\log}$, $\sigma_{\log}$).

Inversely, if Y↪ N (μ, σ), the random variable X=exp(Y) is lognormally distributed and we have X↪ LN (μ, σ). μ and σ are also here the logarithmic parameters for the random variable X.

As seen in the previous section, the product of n independent lognormally distributed random variables is lognormally distributed. In the same way, the sum of n independent normally distributed random variables $\{X_1,..., X_n\}$ is normally distributed and the resulting standard deviation is given by the following formula:

Equation 8-3

$$\sigma_Y = \sqrt{\sum_{i=1}^{n} \sigma_{X_i}^2}.$$

And more generally, if we consider a random variable Y, that is a function of n independent random variables $\{X_1,..., X_n\}$, it has been shown that the standard deviation of Y is given by:

Equation 8-4

$$\sigma_Y = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial Y}{\partial X_i}\right)^2 \sigma_{X_i}^2}$$

Knowing these links between the lognormal and the normal distribution and remembering some of the conclusions drawn for the lognormal distribution – the mode is the deterministic value expressed in the unit of the modeled datum and the geometric standard deviation is a dimensionless measure of variability – we can build some links between the “additive world” and the “multiplicative world” (Table 27).

For the normal distribution the most famous measure of variability is the variance and also the standard deviation. However, this measure of variability has a unit, the same as the data; it cannot also directly be employed to express the uncertainty. The coefficient of variation (ratio between the standard deviation and the mean) is a dimensionless measure of variability. It’s in fact a parameter that defines the relative dispersion of a sample. As it can be seen in Table 27, the geometric standard deviation and the standard deviation can be expressed through the CV.

In order to have a dimensionless measure of variability for both the “additive” and the “multiplicative” world, the **CV is chosen to express uncertainty factors**.

**Table 27	From the additive to the multiplicative world**

	“Additive world” (Normal distribution)	“Multiplicative world” (Lognormal distrbution)
Mode	μ	μ_g
Dimensionless measure of variability	CV	$\sigma_g = e^{\sqrt{\ln(CV^2+1)}}$
Confidence interval (68%)	[μ-σ ; μ+σ]	[μ_g/σ_g ; $\mu_g\sigma_g$]

8.4.2 Uncertainty factors expressed as coefficient of variations

Applying the translation formula between CV and geometric standard deviation (see Table 27):

Equation 8-5

$$CV = \sqrt{\exp(\ln^2\sigma_g) - 1} \text{ with } \sigma_g = \sqrt{U_i}$$

allow us to express the uncertainty factors as coefficient of variations. Table 28 takes a census of the basic uncertainty factors express in terms of CV and Table 29 shows the new developed additional uncertainty factors expressed in terms of CV.

Table 28 Basic uncertainty factors expressed in terms of CV

Input/Output group	*c*	*p*	*a*	*Input/Output group*	*c*	*p*	*a*
Demand of				***Pollutants emitted to air***			
thermal energy, electricity, semi-finished products, working material, waste treatment service	0,02	0,02	0,02	CO_2	0,02	0,02	
transport services (tkm)	0,36	0,36	0,36	SO_2	0,02		
Infrastructure	0,59	0,59	0,59	NMVOC total	0,20		
Resources				NO_x, N_2O	0,20		0,17
primary energy carriers, metals, salts	0,02	0,02	0,02	CH_4, NH_3	0,20		0,09
land use, occupation	0,20	0,20	0,05	individual hydrocarbons	0,20	0,36	
land use, transformation	0,36	0,36	0,09	PM>10	0,20	0,20	
Pollutants emitted to water				PM10	0,36	0,36	
BOD, COD, DOC, TOC, inorganic compounds		0,20		PM2,5	0,59	0,59	
individual hydrocarbons, PAH		0,59		PAH	0,59		
heavy metals		0,95	0,30	CO, heavy metals	0,95		
Pesticides			0,20	inorganic emissions, others		0,20	
NO_3, PO_4			0,20	radionuclides		0,59	
Pollutants emitted to soil							
oil, hydrocarbon total		0,20					
heavy metals		0,20	0,20				
Pesticides			0,09				

Table 29 New developed additional uncertainty factors expressed in terms of CV

Indicator score	1	2	3	4	5
Reliability	0,00	0,45	0,50	0,56	1237,93
Completeness	0,00	0,03	0,04	0,08	(n.a.)
Temporal correlation	0,00	0,03	0,10	0,18	0,26
Geographical correlation	0,00	0,04	0,08	0,10	(n.a.)
Further technological correlation	0,00	0,17	0,53	0,84	1,37

8.4.3 Uncertainty compilation

As seen before, if the product is the operation used in the “multiplicative world”; to compile the basic and the additional uncertainties, in the “additive world” this operation is the sum.

Consider D_{wP} a random variable that models a datum D with its total uncertainty (I), the widely used formula to determine the standard deviation of a function of random variables can be applied to express the standard deviation of datum with its total uncertainty

Equation 8-6

$$\sigma_{D_{wP}} = \sqrt{\left(\frac{\partial D_{wP}}{\partial D}\right)^2 \sigma_D^2 + \left(\frac{\partial D_{wP}}{\partial I}\right)^2 \sigma_I^2}$$

As the variables are independent we have:

$$\frac{\partial D_{wP}}{\partial D} = \frac{\partial D_{wP}}{\partial I} = 1$$

and also:

Equation 8-7

$$\sigma_{D_{wP}} = \sqrt{\sigma_D^2 + \sigma_I^2}$$

Equation 8-7 is valid whatever the distribution used.

As seen in Table 27, the standard deviation expressed an absolute dispersion around a central value, which is most of the time the mean. σ_D and σ_I express this dispersion around the central value. More concretely, the additional uncertainty increases the confidence interval around the central value.

Moreover, as the additional uncertainty is lognormally distributed and combine thanks to Equation 8-1, we can express the whole additional uncertainty through a CV. Combining Equation 8-1 and Equation 8-5 we obtained Equation 8-8 that expresses the coefficient of variation of the additional uncertainty and where CV_i are the different additional uncertainty factors expressed in terms of CVs.

Equation 8-8

$$\mathrm{CV_p} = \sqrt{\Pi_{\mathrm{i}=1}^{5}(\mathrm{CV_i}+1)^2 - 1}$$

8.5 Development of the formulas for the other distributions

8.5.1 Symmetric distributions

In the case where the mean is the central value, we also have:

$$\mu_{D_{wP}} = \mu_D = \mu_I$$

and then, Equation 8-7 becomes:

Equation 8-9

$$CV_{D_{wP}} = \sqrt{CV_D^2 + CV_I^2}$$

The total uncertainty is also here expressed as a relative measure of variability.

The normal and the uniform distributions are both symmetric distributions, i.e. the most likely value is also the mean. Equation 8-9 is also directly applicable and the parameters of the distribution modeling a datum with its total uncertainty can be expressed through CV_{wP}.

- **Normal distribution**

$$\begin{Bmatrix} \mu = \mu_{wP} \\ CV_{wP} = \dfrac{\sigma_{wP}}{\mu} \end{Bmatrix} \Rightarrow \begin{Bmatrix} \mu \\ \sigma_{wP} = \mu CV_{wP} \end{Bmatrix}$$

- **Uniform distribution**

$$\left\{\begin{array}{l} \mu = \dfrac{a+b}{2} = \dfrac{a_{wP}+b_{wP}}{2} \\ CV_{wP} = \dfrac{b_{wP}-a_{wP}}{\sqrt{3}(b_{wP}+a_{wP})} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} a_{wP} = a+b-b_{wP} \\ b_{wP} = \mu(1+\sqrt{3}CV_{wP}) \end{array}\right\}$$

8.5.2 Asymmetric distributions

In the case of asymmetric distributions, the mean differs from the most likely value. For these distributions, the mean will be affected by the consideration of additional uncertainty. While Equation 8-9 can still be used to calculate the relative dispersion parameter (CV_{wP}), a new mean (μ_{wP}) that takes into account the effect of the additional uncertainty must be calculated using Equation 8-10 where μ is the mean of the datum with its basic uncertainty.

Equation 8-10

$$\mu_{wP}CV_{wP} = \mu\sqrt{CV_D^2 + CV_P^2}$$

Considering Equation 8-10 and the different properties of each distribution, the parameters of the distribution modeling a datum with its total uncertainty can be expressed.

- **Triangular distribution**

Definition of the mean and the coefficient of variation:

$$\left\{\begin{array}{c} \mu = \dfrac{a+b+c}{3} \\ CV^2 = 0.5\dfrac{a^2+b^2+c^2-ab-ac-bc}{(a+b+c)^2} \end{array}\right\}$$

The mean and the coefficient of variation for the data with its total uncertainty, i.e. μ_{wP} and CV_{wP} can be expressed by the same two formulas by replacing a and b by a_{wP} and b_{wP} respectively.

Considering Equation 8-10 and the previously system, we have:

$$\left\{\begin{array}{c} a_{wP} = c(1+\gamma) - \gamma b_{wP} \\ b_{wP} = c + 3\mu\sqrt{CV_D^2 + CV_P^2}\sqrt{\dfrac{2}{1+\gamma+\gamma^2}} \end{array}\right\}$$

where $\gamma = \frac{c-a}{b-c} = \frac{c-a_{wP}}{b_{wP}-c}$ and expresses the asymmetry of the distribution.

- **Beta PERT distribution**

Definition of the mean and the coefficient of variation

$$\left\{\begin{array}{l} \mu = \dfrac{a+4c+b}{6} \\ CV = \dfrac{b-a}{a+4c+b} \end{array}\right\}$$

As for the triangular distribution, μ_{wP} and CV_{wP} can be expressed by replacing a and b by a_{wP} and b_{wP} respectively.

Considering Equation 8-10 and the previously system, we have:

$$\left\{\begin{array}{c} a_{wP} = c(1+\gamma) - \gamma b_{wP} \\ b_{wP} = c + \dfrac{\sqrt{CV_D^2 + CV_P^2}}{1+\gamma}(a + 4c + b) \end{array}\right\}$$

where $\gamma = \frac{c-a}{b-c} = \frac{c-a_{wP}}{b_{wP}-c}$ and expresses the asymmetry of the distribution.

- **Gamma distribution**

The here developed formula for the gamma distribution is only valid when the location parameter is 0.

Definition of the mean, the coefficient of variation and the most likely value (m), which is not modified by adding of the additional uncertainty:

$$\left\{\begin{array}{c} \mu = \lambda k \\ CV^2 = \dfrac{1}{k} \\ m = \lambda(k-1) = \lambda_{wP}(k_{wP}-1) \end{array}\right\}$$

As for the triangular distribution, μ_{wP} and CV_{wP} can be expressed by replacing λ and k by λ_{wP} and k_{wP} respectively.

Considering Equation 8-10 and the previously system, we have:

$$\left\{\begin{array}{c} k_{wP} = 1 + \dfrac{m^2 + \sqrt{\left(2\mu m\sqrt{CV_D^2 + CV_P^2}\right)^2 + m^4}}{4\mu^2(CV_D^2 + CV_I^2)} \\ \lambda_{wP} = \dfrac{m}{k_{wP} - 1} \end{array}\right\}$$

8.6 Illustrations and verification based on a Monte Carlo analysis

Figure 33 and Figure 34 are an illustrative example of Table 29. In this table, the parameters of each distribution modeling data with their total uncertainty are determined for various pedigree scores.

Table 30 Definition of the parameters of the different PDFs (with basic and total uncertainty

PDF and acronym	Parameters	With basic uncertainty	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
Log normal LN	**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 N	**μ**	1.5	1.5	1.5	1.5	1.5
	σ	0.375	0.380	0.414	0.530	0.821
Gamma G	**k**	16	15.66	13.47	8.92	4.72
	λ	0.1	0.102	0.120	0.189	0.403
Uniform U	**a**	1	0.991	0.921	0.677	0.0386
	b	3	3.009	3.079	3.323	3.961

PDF and acronym	Parameters	With basic uncertainty	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
Triangular T	**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 B	**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

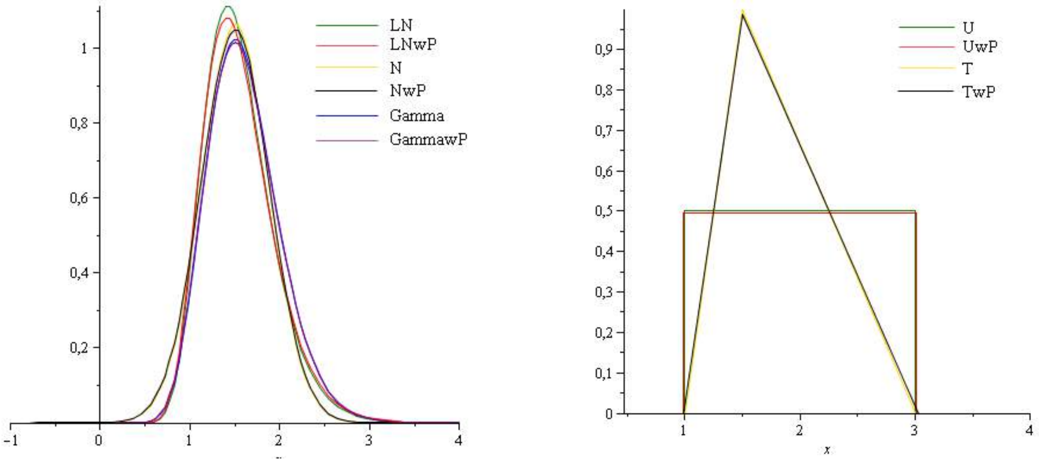


Figure 33: Illustration of Table 30– pedigree scores: (2;2;2;2;2)

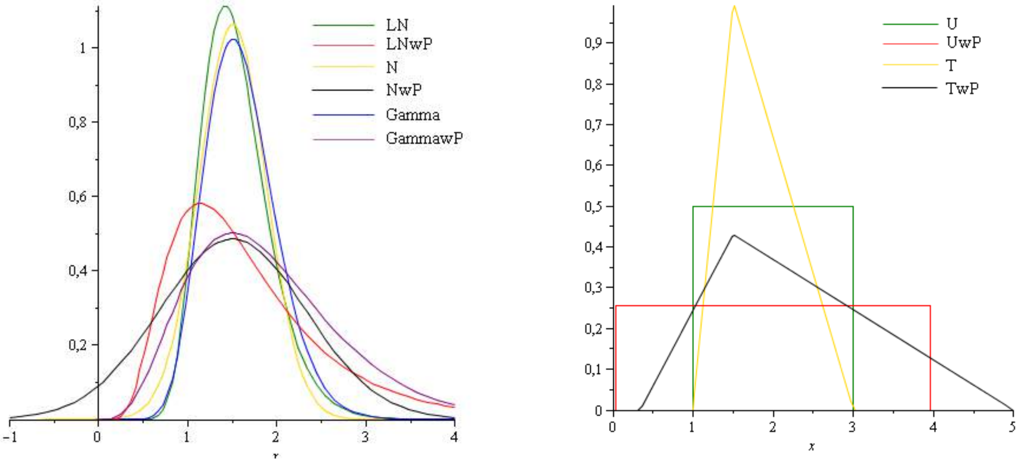


Figure 34: Illustration of Table 30– pedigree scores (5;5;5;5;5)

In order to provide evidence that the developed formulas are appropriate conversion equations for each distribution, the obtained results were compared to the one obtained through a Monte Carlo analysis.

Monte Carlo analyses were run on the sum of a lognormal distribution representing the additional uncertainty and each distribution representing the basic uncertainty. The analyses were performed by the ORACLE Crystal Ball release, fusion edition (v 11.1.2.0), 10 000 steps were performed in the simulation.

Table 30 to Table 34 provide the different results comparisons. For each assessed parameter, the relative error ε is calculated. The field “% of values in the interval” stands for the percentage of the values obtained through the Monte Carlo simulation that lie in the interval [a_{wP}; b_{wP}].

For the normal distribution, the relative errors on the coefficient of variation are less than 5% which is an acceptable level. For the triangular, uniform and BetaPERT distribution, if the relative errors are higher on every parameter, the percentage of the values obtained through the Monte Carlo simulation that lie in [a_{wP}; b_{wP}] are higher than 90% which is also acceptable here.

Concerning the gamma distribution, the relative errors on CV are too high to conclude that the proposed formulas are appropriate. Moreover, these formulas are valid only for the distributions that have 0 as a location parameter. Considering these limits, the gamma distribution has to be chosen to model a data with its total uncertainty only when the shape and the scale parameters are perfectly known. If there is not the case, the lognormal distribution must rather be chosen.

Table 31 Results comparison for the normal distribution

	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
σ_{wP}	0,38	0,414	0,53	0,821
CV_{wP}	0,253	0,276	0,353	0,547
σ_{MC}	0,38	0,42	0,56	0,93
CV_{MC}	0,252	0,277	0,366	0,554
ε_{σ}	0,00%	1,43%	5,36%	11,72%
ε_{CV}	0,40%	0,36%	3,55%	1,26%

Table 32 Results comparison for the uniform distribution

	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
a_{wp}	0,991	0,921	0,677	0,039
b_{wp}	3,009	3,079	3,323	3,961
CV_{wP}	0,291	0,311	0,382	0,566
a_{MC}	0,93	0,75	0,42	0,2
b_{MC}	3,37	4,23	7,33	12,38
CV_{MC}	0,293	0,311	0,394	0,578
ε_{a}	6,16%	18,57%	37,96%	80,50%
ε_{b}	10,71%	27,21%	54,67%	68,00%
ε_{CV}	0,68%	0,00%	3,05%	2,08%
% of values in the interval	97,10%	94,07%	92,15%	90,79%

Table 33 Results comparison for the triangular distribution

	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
a_{wp}	0,993	0,94	0,765	0,336
b_{wp}	3,021	3,18	3,706	4,991
CV_{wP}	0,234	0,254	0,314	0,435
a_{MC}	0,95	0,78	0,56	0,24
b_{MC}	3,18	3,83	4,87	12,61
CV_{MC}	0,236	0,261	0,349	0,543
ε_a	4,33%	17,02%	26,80%	28,57%
ε_b	5,00%	16,97%	23,90%	60,42%
ε_{CV}	0,85%	2,68%	10,03%	19,89%
% of values in the interval	99,60%	98,93%	97,90%	97,84%

Table 34 Results comparison for the beta Pert distribution

	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
a_{wp}	0,991	0,921	0,7	0,184
b_{wp}	3,028	3,237	3,9	5,45
CV_{wP}	0,235	0,255	0,322	0,464
a_{MC}	0,96	0,76	0,52	0,26
b_{MC}	2,98	3,41	5,32	10,68
CV_{MC}	0,216	0,244	0,333	0,54
ε_a	3,13%	17,48%	25,71%	29,23%
ε_b	1,59%	5,07%	26,69%	48,97%
ε_{CV}	8,09%	4,31%	3,30%	14,07%
% of values in the interval	99,93%	99,62%	99,52%	100,00%

Table 35 Results comparison for the gamma distribution

	(2;2;2;2;2)	(3;3;3;3;3)	(4;4;4;4;4)	(5;5;5;5;5)
CV_{wP}	0,253	0,272	0,335	0,46
CV_{MC}	0,36	0,38	0,444	0,633
ε_{CV}	29,72%	28,42%	24,55%	27,33%

8.7 Limits

8.7.1 How to treat negative values?

While negative values cannot be defined through a lognormal distribution, this is not the case for a normal, uniform, triangular, beta PERT or a gamma distribution. Most of the flows which are scrutinised in LCA are physical quantities: they can by definition not be negative.

Adding the pedigree uncertainty modifies the parameters of the distribution and also the minimal value which can "become" negative. These negative values can be considered as follow:

- Consider them in the uncertainty analysis (keeping in mind that these values are fictive).
- Define a threshold value or a location value (in this case 0) in the definition of the distribution. A certain percentage of values are, with this method, not considered.
- Transform all the negative values into a null value. The probability to have a null value will also be more important.

In these three cases, a supplementary error is created. Depending on the percentage of the values which are negative, this error can be more or less important. Here, these three solutions are given in a descending order of relevance.

8.7.2 Results are based on many assumptions

In this section, the additional uncertainty factors are considered as lognormally distributed - with one as the geometric mean and the square root of U_i as the geometric standard deviation.

The whole additional uncertainty is also lognormally distributed too. This assumption can be checked in the next phases of the project, when the uncertainty factors will be determined specifically by archetypes.

8.8 Other possible approaches

8.8.1 The computational solution

This section deals with the development of analytical solutions to the application of the pedigree matrix approach to distributions other than the lognormal distribution. However, there may be many advantages to forego such an analytical approach and rather opting for a computational solution, specifically one based on Monte Carlo simulation.

A computational approach would have two advantages:

- It would be possible to use different distribution functions for different additional uncertainty types. For example, it is possible that the present exercise will uncover that, e.g., the basic uncertainty is best described by a triangular distribution, the uncertainty associated with geographical correlation is best described by a normal distribution and that the uncertainty associated with for temporal correlation is best described by a lognormal distribution. The analytical solutions above do not allow a mix of distribution types in the calculation of the total uncertainty, as the additional uncertainty is assumed to be lognormally distributed.
- It would dispense us of making some of the assumptions described in the previous sections.
- It would be possible to introduce corrections of the mean, useful e.g. when using industrialized country data for an activity in a developing country (translate central

tendency to left) or when using older data to represent actual situations (translate central tendency to right).

This said, one major assumption would have to be made, that we are in the “multiplicative” world, i.e. that the total uncertainty can be represented by the multiplication of the deterministic value by probability distribution functions with central tendencies equal to 1 (except when a correction is required). The approach would then be simply to carry out a Monte Carlo simulation of the following relation:

PDF(total uncertainty) = Monte Carlo (Deterministic Value * X * U1 * U2* U3* U4* U5)

Where:

- X = Probability distribution function representing the basic uncertainty, with central tendency equal to 1
- Ui = Probability distribution function representing the additional uncertainty, with central tendency equal to 1 unless a correction is required.

An ostensible disadvantage of this approach is the computational time required. Tests carried out using the Crystal Ball software package indicate that about 0.2 seconds are required per flow for 5000 iterations. There is no reason to believe that the ecoEditor could not be as efficient in this calculation, meaning that an average data provider would have to wait 6.13 seconds (using the average 9.06 intermediate exchanges and 21.59 elementary flows per unit process in ecoinvent 2.0).

8.8.2 The “perfect” way : the convolution product

In order to calculate the total uncertainty of a flow representing by the random variable F, we have to determine twice the standard deviation of:
$F+P_1+ P_2+ P_3+ P_4+ P_5$ (if we do not consider here a lognormal distribution).
The theoretical mathematical way to determine the variance (and also the standard deviation) is to determine the PDF of $F+P_1+ P_2+ P_3+ P_4+ P_5$ using the convolution product.

Considering X and Y two independant random variables, their PDF are respectively f and g. We define Z as Z=X+Y, h is the PDF of the random variable Z.
By definition, in statistics, we have:

$$\begin{aligned} h(t) = (f * g)(t) &= \int f(\tau)g(t-\tau)d\tau \\ &= \int f(t-\tau)g(\tau)d\tau. \end{aligned}$$

Once the PDF h is determined, the variance can be calculated using the following formula:

$$\mathrm{Var}(Z) = \int (z-\mu)^2 h(z)dz$$

$$\text{where} \qquad \mu = \int zh(z)dz.$$

With this theoretical approach, it is not always possible to express the variance in a simple way (that can be easily implemented in ecoEditor). Moreover, depending on the type of analyzed distribution, the expression of the variance could only be obtained through a numerical approximation.

9 References

Census 2010: US statistics and census office, http://www.census.gov/

Ciroth 2008: Ciroth, A., Cost data quality considerations for eco-efficiency measures, Ecol. Econ. (2008), doi:10.1016/j.ecolecon.2008.08.005.

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

Ciroth Srocka 2008: Ciroth, A., Srocka, M.: How to Obtain a Precise and Representative Estimate for Parameters in LCA: A case study for the functional unit, 13 LCA (3) 265-277 (2008).

Ciroth Weidema 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., Fleischer, G., Steinbach, J.: Uncertainty Calculation in Life Cycle Assessments - A Combined Model of Simulation and Approximation, Int J LCA 9 (4) 216 - 226 (2004).

EPER 2010, European Pollutant Emission Register, www.eper.ec.europa.eu/

Eurostat 2010: eurostat central database, http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

Frischknecht 2005: Frischknecht, R., et al.: The ecoinvent Database: Overview and Methodological Framework, Int J LCA 10 (1) 2005, 3-9.

Funtowicz and Ravetz 1990: Funtowicz S., and Ravetz J.R.: Uncertainty and Quality in Science for Policy, Kluwer, Dordrecht, 1990.

Fuss Szolgayová 2009: Fuss, S., Szolgayová, J.: Fuel price and technological uncertainty in a real options model for electricity planning, Applied Energy, submitted and accepted.

GREET 2009: http://www.transportation.anl.gov/modeling_simulation/GREET/index.html, GREET model fuel cycle, version 1.8c.0, 2009

Grimm 2010: Grimm, German EPA: personal communication, October 2010

Heijungs Huijbregts 2004: Heijungs, R., Huijbregts, M.A.J., A review of approaches to treat uncertainty in LCA, iEMS Conference Proceedings: Complexity and Integrated Resources Management, Osnabrueck, Germany, 2004.

Kent 1983: J. T. Kent: Information gain and a general measure of correlation, Biometrika, London 70.1983,1, 163-173.

Lloyd and Ries 2007: Lloyd, Sh., Ries, R.: Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment, A Survey of Quantitative Approaches, Journal of Industrial Ecology, Volume 11, Number 1, pp 161-179, 2007.

Lundie 2004: Lundie, S., et al: Australian dairy farm data, 2004, unpublished.

ProBas 2010: ProBas database of the German EPA, www.probas.umweltbundesamt.de

PRTR 2010: The European Pollutant Release and Transfer Register, http://prtr.ec.europa.eu/

Sluijs, J.R., et al. 2003: van der Sluijs, J.R.; Kloprogge, PJ; Risbey, J.; Ravetz, J.: 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, 2003.

Tremod 2010: European transport emission model, http://www.umweltbundesamt.de/verkehr/index-daten.htm

Weidema B.P. 1998: Multi-User Test of the Data Quality Matrix for Product Life Cycle Inventory Data. Int.J LCA 3(5):259-265.

Weidema, B.P., Wesnæs, M.S 1996.: Data quality management for life cycle inventories – an example of using data quality indicators, Journal of Cleaner Production, Vol. 4 pp. 167-174 (1996).

ZSE 2010: German CO_2 Emission trading institution, http://www.dehst.de/

Annexe A: Analysed Data Sources

1 E-PRTR

1.1 *Description*

The original name of the database is ‘European Pollutant Release and Transfer Register’ (E-PRTR). It is based on the UNECE PRTR Protocol (Source: http://prtr.ec.europa.eu). About 25,000 industrial facilities, covering 65 economic activities across 31 countries in Europe register their yearly pollutant releases to air and water and the amount of waste. Only facilities exceeding specified emission threshold need to provide data. Not all emitted substances need to be provided. Altogether, 91 different substances are recorded, mainly in absolute terms, per industrial facility, and mainly in kilogram per year.

For each facility, the following items are usually provided: facility name and activity, name of pollutants, measurement or calculation method, total amount, release medium, unit.

The information is interesting for this project because it is created completely outside of the field of LCA, and often obtained via measurement and monitoring procedures; hence it is often indeed empirically based.

Although the database claims to be reporting for the years 2007 and 2008 only, the last update is from the 18th of October 2010. The next report is due on 31/03/2011 (covering 2009).

The database copyright holder is the Directorate-General for Environment. Re-use of content for commercial or non-commercial purposes is permitted free of charge, provided that the source is acknowledged (http://www.eea.europa.eu/legal/copyright).

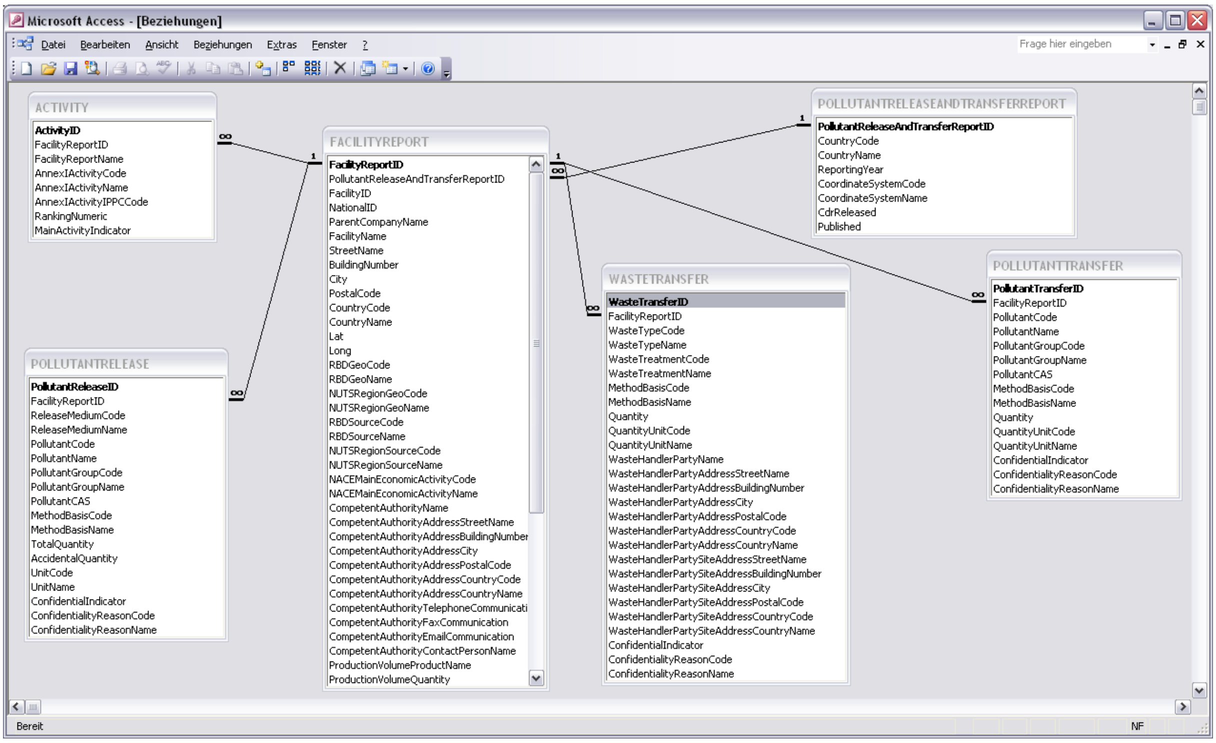


Figure 35: PRTR database screenshot, overview of the database structure

1.2 *Limitations*

Main limitation is that PRTR almost only reports absolute figures, on a “per plant” or “per facility” level. The amount of product produced is usually lacking, which makes it difficult to use the data without further preparation.

Then, for an in-depth analysis, it should be kept in mind that the database is reflecting European conditions only. Further, it is reflecting large emitters only, and only a number of regulated substances.

1.3 Use in this project

The database provides comprehensive information for emissions in European industrial facilities. This can be used to investigate and cover several fields of the pedigree matrix. Especially, the "reliability", the "geographical correlation" and the "further technological correlation" factors can be analysed. Limitations of the data source regarding the provided pollutants, and the focus on large volume emitters, for a European background, need to be taken into account.

2 Tremod - Transport Emission Model

2.1 Description

The Tremod database has been set up by IFEU ("the Institute for Energy and Environmental Research" in Heidelberg, Germany) in 1993, and has been updated since then. It is not available to the public[19], but a few institutions can use it, like the Umweltbundesamt which gives a public access to results through its PROBAS database: http://www.probas.umweltbundesamt.de/php/themen.php?id=12884901888&step=2& .

The access was realised by a joint project between IFEU and GreenDeltaTC. The database delivers energy use and pollutant emissions from various transport systems in Germany. There are up to 12 parameters to define the exact transport system (traffic carrier, energy, size, load factor, road type, efficiency, gradient of road…), and only up to 15 emitted pollutants, plus the fuel demand. Data are given per quantitative reference. The quantitative reference varies, possible values are "person kilometre", "ton kilometre", or also, simply, "km".

The data sets in Tremod are created via a mix of empirical measurements, modelling assumptions, and estimates. Overall, the database reflects European transport systems. Especially road transport is modelled in a very detailed way. For example, short distance travel with cars, where the catalysts are not fully heated up to operation temperature, and where a fatter fuel mix is usually used in cars, is distinguished from longer distances.

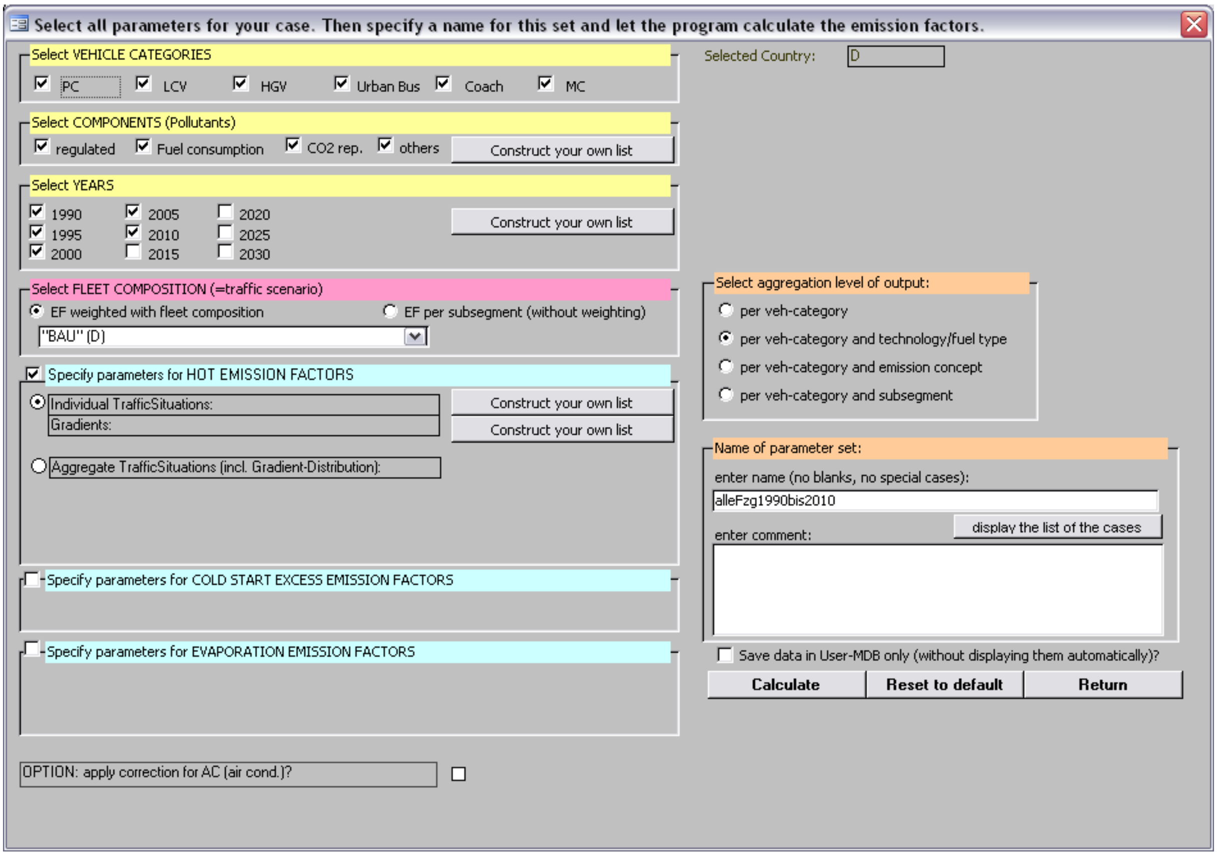
More information is available here:
http://www.ifeu.de/english/index.php?bereich=ver&seite=projekt_tremod

The database we used to analyse is based on Tremod 2005, reflecting the year 2005.

A related data source is the "HBEFA"[20] database; it contains the road transport part of Tremod, and is publicly available, distributed by the INFRAS institute in Switzerland, http://www.hbefa.net/e/index.html.

[19] "Due to its volume and complexness, Tremod is not available to the public." : http://www.ifeu.de/english/index.php?bereich=ver&seite=projekt_tremod

[20] HBEFA is the abbreviation for "Handbuch für Emissionsfaktoren" / "handbook for emission factors".



Select all parameters for your case. Then specify a name for this set and let the program calculate the emission factors.

Select VEHICLE CATEGORIES

☑ PC ☑ LCV ☑ HGV ☑ Urban Bus ☑ Coach ☑ MC

Selected Country: D

Select COMPONENTS (Pollutants)

☑ regulated ☑ Fuel consumption ☑ CO2 rep. ☑ others

Select YEARS

☑ 1990 ☑ 2005 □ 2020
☑ 1995 ☑ 2010 □ 2025
☑ 2000 □ 2015 □ 2030

Select FLEET COMPOSITION (=traffic scenario)

● EF weighted with fleet composition ○ EF per subsegment (without weighting)

"BAU" (D)

☑ Specify parameters for HOT EMISSION FACTORS

● Individual TrafficSituations:
Gradients:

○ Aggregate TrafficSituations (incl. Gradient-Distribution):

□ Specify parameters for COLD START EXCESS EMISSION FACTORS

□ Specify parameters for EVAPORATION EMISSION FACTORS

OPTION: apply correction for AC (air cond.)? □

Select aggregation level of output:

○ per veh-category
● per veh-category and technology/fuel type
○ per veh-category and emission concept
○ per veh-category and subsegment

Name of parameter set:

enter name (no blanks, no special cases):
alleFzg1990bis2010

enter comment:

□ Save data in User-MDB only (without displaying them automatically)?

Figure 36: HBEFA application screenshot, showing the different parameters and options for defining calculations in the model

2.2 *Modifications before data analysis*

Before the analysis, datasets need to be transformed so that they have a comparable quantitative reference. This was done by converting quantitative references given in km to the load-specific references person-km or ton-km, wherever possible. Then, in order to allow an analysis of technology levels, data must first be grouped into technology levels. Suitable grouping levels need to be entered into the data source to this end.

2.3 *Limitations*

Each transport system is described only once with its 15 pollutant emissions. The closer we look at technological details, the fewer data is available. This could result in a smaller standard deviation and could interfere with any variation of the standard deviation due to the "technology level".

2.4 *Use in this project*

Tremod can be useful for understanding parts of the pedigree matrix. Especially, the "further technological correlation" indicator can be analysed, by defining the level of detail for different transport systems. The other pedigree indicators "reliability", "completeness", "temporal correlation" and "geographical correlation" are almost the same and can therefore not be analysed.

The HBEFA database contains data on transport since 1990, see Figure 36, and can therefore be used to analyse temporal correlation.

Both data sources will therefore be used.

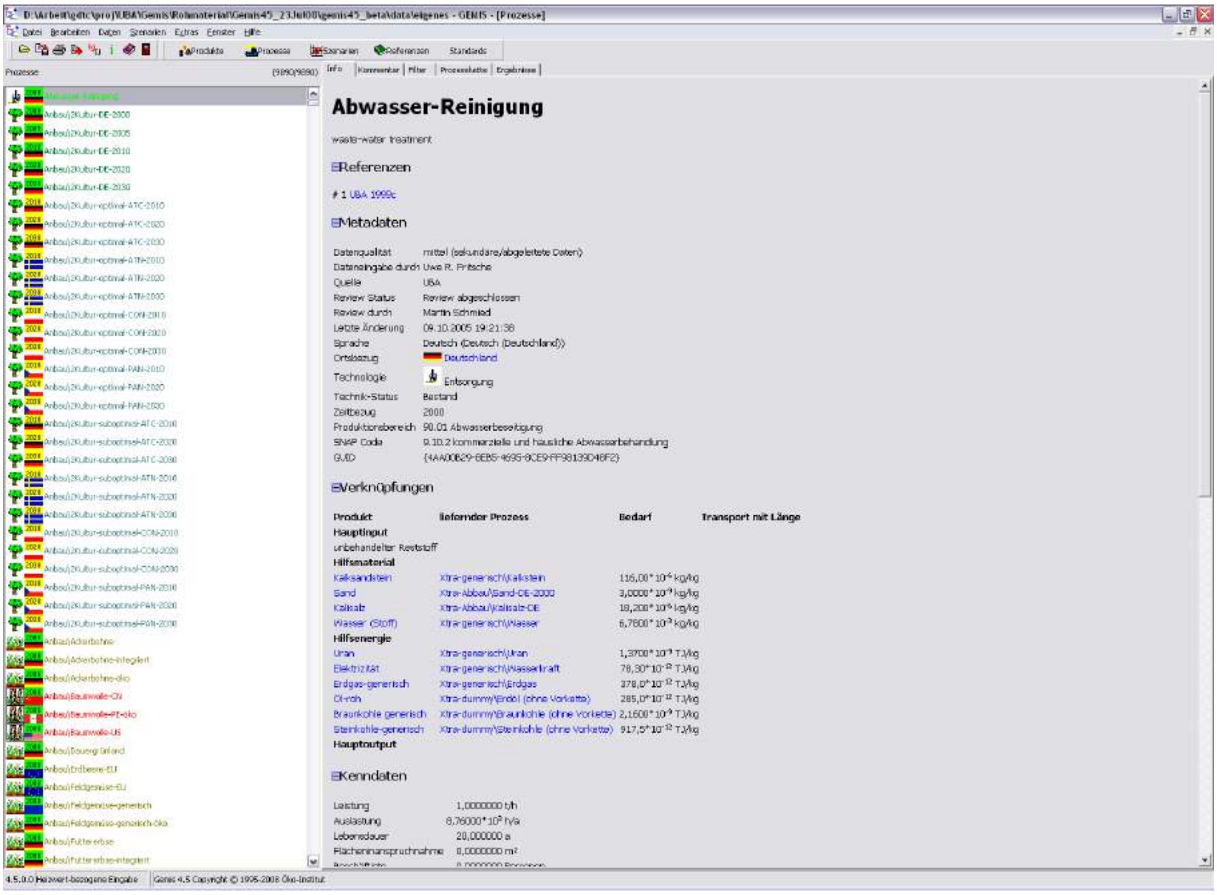
3 GEMIS

3.1 *Description*

The GEMIS acronym stands for “Global Emission Model for Integrated Systems”. It is a life-cycle analysis program and database covering energy, material, process and transport systems. It describes all kind of processes, their main product and emissions. Each process with its parameters (economic sector, technological group, technical status, time reference, source, data quality, etc) is only described once. The software also provides future potential emissions according to scenarios.

The first development started in 1987 by the “Öko-Institut”. Since then, the model was upgraded and updated. The last version (4.6) is from August 2010. More information can be found here: www.gemis.de .

Data analysed is from the GEMIS 4.5 version, which was transformed into an “easy to analyse” Access database in a previous project (Ciroth 2009).



Anbau\2Kultur-DE-2000
Anbau\2Kultur-DE-2005
Anbau\2Kultur-DE-2010
Anbau\2Kultur-DE-2020
Anbau\2Kultur-DE-2030
Anbau\2Kultur-optimal-ATC-2010
Anbau\2Kultur-optimal-ATC-2020
Anbau\2Kultur-optimal-ATC-2030
Anbau\2Kultur-optimal-ATN-2010
Anbau\2Kultur-optimal-ATN-2020
Anbau\2Kultur-optimal-ATN-2030
Anbau\2Kultur-optimal-CON-2010
Anbau\2Kultur-optimal-CON-2020
Anbau\2Kultur-optimal-CON-2030
Anbau\2Kultur-optimal-PAN-2010
Anbau\2Kultur-optimal-PAN-2020
Anbau\2Kultur-optimal-PAN-2030
Anbau\2Kultur-suboptimal-ATC-2010
Anbau\2Kultur-suboptimal-ATC-2020
Anbau\2Kultur-suboptimal-ATC-2030
Anbau\2Kultur-suboptimal-ATN-2010
Anbau\2Kultur-suboptimal-ATN-2020
Anbau\2Kultur-suboptimal-ATN-2030
Anbau\2Kultur-suboptimal-CON-2010
Anbau\2Kultur-suboptimal-CON-2020
Anbau\2Kultur-suboptimal-CON-2030
Anbau\2Kultur-suboptimal-PAN-2010
Anbau\2Kultur-suboptimal-PAN-2020
Anbau\2Kultur-suboptimal-PAN-2030
Anbau\Ackerbohne
Anbau\Ackerbohne-integriert
Anbau\Ackerbohne-öko
Anbau\Baumwolle-CN
Anbau\Baumwolle-PE-öko
Anbau\Baumwolle-US
Anbau\Dauergrünland
Anbau\Erdbeere-EU
Anbau\Feldgemüse-EU
Anbau\Feldgemüse-generisch
Anbau\Feldgemüse-generisch-öko
Anbau\Futtererbse
Anbau\Futtererbse-integriert

Abwasser-Reinigung

waste-water treatment

⊟Referenzen

1 UBA 1999c

⊟Metadaten

Datenqualität	mittel (sekundäre/abgeleitete Daten)
Dateneingabe durch	Uwe R. Fritsche
Quelle	UBA
Review Status	Review abgeschlossen
Review durch	Martin Schmied
Letzte Änderung	09.10.2005 19:21:38
Sprache	Deutsch (Deutsch (Deutschland))
Ortsbezug	Deutschland
Technologie	Entsorgung
Technik-Status	Bestand
Zeitbezug	2000
Produktionsbereich	90.01 Abwasserbeseitigung
SNAP Code	9.10.2 kommerzielle und häusliche Abwasserbehandlung
GUID	{4AA00B29-6EB5-4695-8CE9-FF98139D48F2}

⊟Verknüpfungen

Produkt	liefernder Prozess	Bedarf	Transport mit Länge
Hauptinput			
unbehandelter Reststoff			
Hilfsmaterial			
Kalksandstein	Xtra-generisch\Kalkstein	116,00 * 10^{-6} kg/kg	
Sand	Xtra-Abbau\Sand-DE-2000	3,0000 * 10^{-9} kg/kg	
Kalisalz	Xtra-Abbau\Kalisalz-DE	18,200 * 10^{-6} kg/kg	
Wasser (Stoff)	Xtra-generisch\Wasser	6,7800 * 10^{-3} kg/kg	
Hilfsenergie			
Uran	Xtra-generisch\Uran	1,3700 * 10^{-9} TJ/kg	
Elektrizität	Xtra-generisch\Wasserkraft	78,30 * 10^{-12} TJ/kg	
Erdgas-generisch	Xtra-generisch\Erdgas	378,0 * 10^{-12} TJ/kg	
Öl-roh	Xtra-dummy\Erdöl (ohne Vorkette)	285,0 * 10^{-12} TJ/kg	
Braunkohle generisch	Xtra-dummy\Braunkohle (ohne Vorkette)	2,1600 * 10^{-9} TJ/kg	
Steinkohle-generisch	Xtra-dummy\Steinkohle (ohne Vorkette)	917,5 * 10^{-12} TJ/kg	
Hauptoutput			

⊟Kenndaten

Leistung	1,0000000 t/h
Auslastung	8,76000 * 10^{3} h/a
Lebensdauer	20,000000 a
Flächeninanspruchnahme	0,0000000 m²
Beschäftigte	0,0000000 Personen

4.5.0.0 Heizwert-bezogene Eingabe | Gemis 4.5 Copyright © 1995-2008 Öko-Institut

Figure 37: GEMIS application screenshot

3.2 *Modifications before data analysis*

Input and output flows of processes are given for one kg or one TJ of resulting main product, i.e. per quantitative reference; therefore, if processes with both TJ and kg should be analysed, the quantitative reference should be aligned. This was rarely possible, and therefore, the analyses focused on processes with the same unit of quantitative reference. Processes needed to be first classified per economic sector (NACE code) and per unit.

3.3 *Limitations*

Regarding the "temporal correlation factor", many data are from the year 2000, and a few before 2000. Processes with a reference year later than 2000 often come from forecast calculations. These forecasted data sets do not seem useful for providing an empirical basis and are therefore excluded from the analysis. Data are quite often not empirically based, but rather expert guesses and adaptations from LCA data – but this varies.

3.4 *Use in this project*

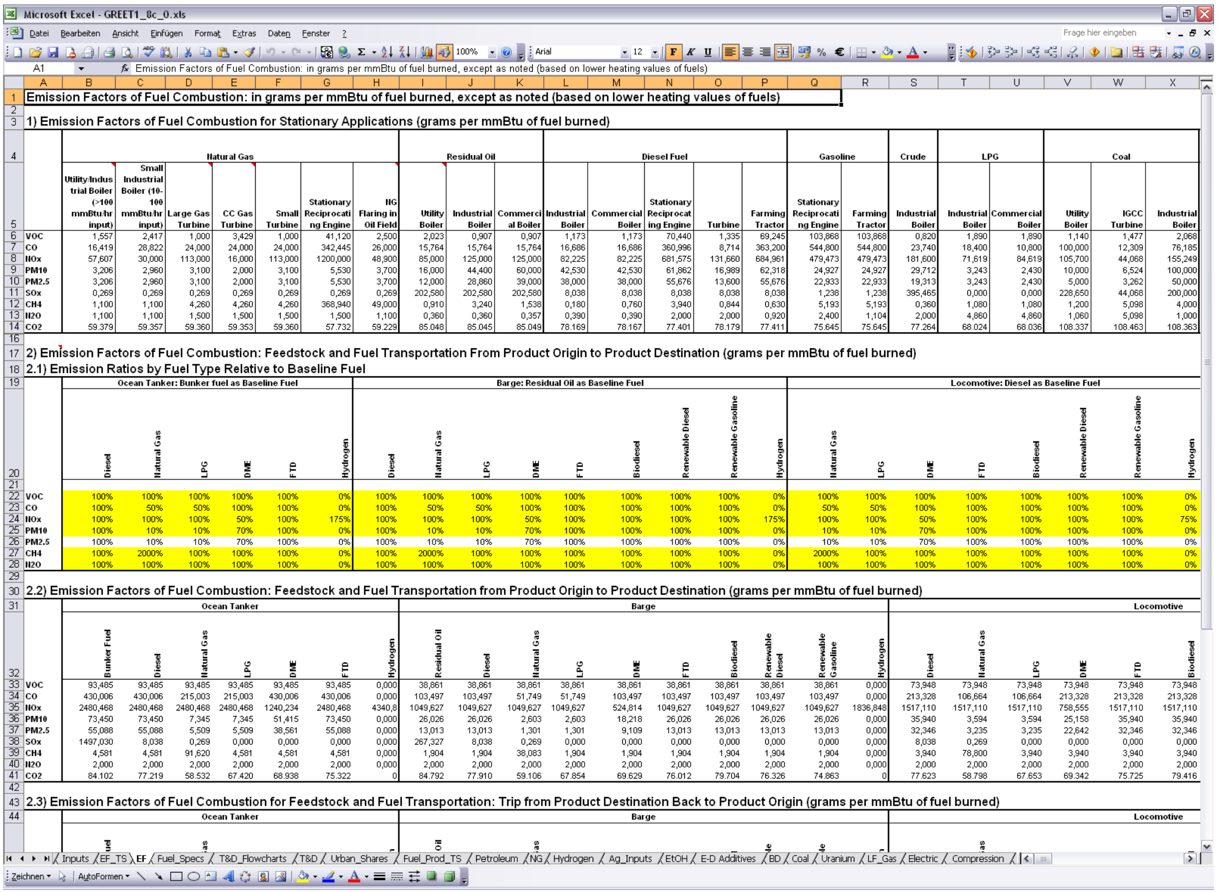
Data come from different sources in GEMIS. GEMIS assigns an indicator "data quality" to each data set, with values from "very good" to "basic calculation". This can be used for assessing the "reliability" indicator of the pedigree matrix. Also, the temporal correlation" indicator can be investigated.

4 GREET Model

4.1 *Description*

The GREET Model (The Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation Model) contains data sets on American vehicles. A multidimensional spreadsheet model in Excel is available free of charge. The first version of GREET was released in 1996. Since then, Argonne, a national laboratory of the U.S department of energy, is responsible for the update of the model. More information is available here: http://greet.es.anl.gov/ .

The database shows information on vehicle emissions (CO_2, VOCs, NOx…), when specified vehicles (3 vehicle classes) and fuels (8 different fuels) are selected.



Emission Factors of Fuel Combustion: in grams per mmBtu of fuel burned, except as noted (based on lower heating values of fuels)

1) Emission Factors of Fuel Combustion for Stationary Applications (grams per mmBtu of fuel burned)

	Natural Gas							Residual Oil			Diesel Fuel					Gasoline		Crude	LPG		Coal		
	Utility/Industrial Boiler (>100 mmBtu/hr input)	Small Industrial Boiler (10-100 mmBtu/hr input)	Large Gas Turbine	CC Gas Turbine	Small Turbine	Stationary Reciprocating Engine	NG Flaring in Oil Field	Utility Boiler	Industrial Boiler	Commercial Boiler	Industrial Boiler	Commercial Boiler	Stationary Reciprocating Engine	Turbine	Farming Tractor	Stationary Reciprocating Engine	Farming Tractor	Industrial Boiler	Industrial Boiler	Commercial Boiler	Utility Boiler	IGCC Turbine	Industrial Boiler
VOC	1,557	2,417	1,000	3,429	1,000	41,120	2,500	2,023	0,907	0,907	1,173	1,173	70,440	1,335	69,245	103,868	103,868	0,820	1,890	1,890	1,140	1,477	2,068
CO	16,419	28,822	24,000	24,000	24,000	342,445	26,000	15,764	15,764	15,764	16,686	16,686	360,996	8,714	363,200	544,800	544,800	23,740	18,400	10,800	100,000	12,309	76,185
NOx	57,607	30,000	113,000	16,000	113,000	1200,000	48,900	85,000	125,000	125,000	82,225	82,225	681,575	131,660	684,961	479,473	479,473	181,600	71,619	84,619	105,700	44,068	155,249
PM10	3,206	2,960	3,100	2,000	3,100	5,530	3,700	16,000	44,400	60,000	42,530	42,530	61,862	16,989	62,318	24,927	24,927	29,712	3,243	2,430	10,000	6,524	100,000
PM2.5	3,206	2,960	3,100	2,000	3,100	5,530	3,700	12,000	28,860	39,000	38,000	38,000	55,676	13,600	55,676	22,933	22,933	19,313	3,243	2,430	5,000	3,262	50,000
SOx	0,269	0,269	0,269	0,269	0,269	0,269	0,269	202,580	202,580	202,580	8,038	8,038	8,038	8,038	8,038	1,238	1,238	395,465	0,000	0,000	228,650	44,068	200,000
CH4	1,100	1,100	4,260	4,260	4,260	368,940	49,000	0,910	3,240	1,538	0,180	0,760	3,940	0,844	0,630	5,193	5,193	0,360	1,080	1,080	1,200	5,098	4,000
N2O	1,100	1,100	1,500	1,500	1,500	1,500	1,100	0,360	0,360	0,357	0,390	0,390	2,000	2,000	0,920	2,400	1,104	2,000	4,860	4,860	1,060	5,098	1,000
CO2	59.379	59.357	59.360	59.353	59.360	57.732	59.229	85.048	85.045	85.049	78.169	78.167	77.401	78.179	77.411	75.645	75.645	77.264	68.024	68.036	108.337	108.463	108.363

2) Emission Factors of Fuel Combustion: Feedstock and Fuel Transportation From Product Origin to Product Destination (grams per mmBtu of fuel burned)

2.1) Emission Ratios by Fuel Type Relative to Baseline Fuel

	Ocean Tanker: Bunker fuel as Baseline Fuel						Barge: Residual Oil as Baseline Fuel									Locomotive: Diesel as Baseline Fuel							
	Diesel	Natural Gas	LPG	DME	FTD	Hydrogen	Diesel	Natural Gas	LPG	DME	FTD	Biodiesel	Renewable Diesel	Renewable Gasoline	Hydrogen	Natural Gas	LPG	DME	FTD	Biodiesel	Renewable Diesel	Renewable Gasoline	Hydrogen
VOC	100%	100%	100%	100%	100%	0%	100%	100%	100%	100%	100%	100%	100%	100%	0%	100%	100%	100%	100%	100%	100%	100%	0%
CO	100%	50%	50%	100%	100%	0%	100%	50%	50%	100%	100%	100%	100%	100%	0%	50%	50%	100%	100%	100%	100%	100%	0%
NOx	100%	100%	100%	50%	100%	175%	100%	100%	100%	50%	100%	100%	100%	100%	175%	100%	100%	50%	100%	100%	100%	100%	75%
PM10	100%	10%	10%	70%	100%	0%	100%	10%	10%	70%	100%	100%	100%	100%	0%	10%	10%	70%	100%	100%	100%	100%	0%
PM2.5	100%	10%	10%	70%	100%	0%	100%	10%	10%	70%	100%	100%	100%	100%	0%	10%	10%	70%	100%	100%	100%	100%	0%
CH4	100%	2000%	100%	100%	100%	0%	100%	2000%	100%	100%	100%	100%	100%	100%	0%	2000%	100%	100%	100%	100%	100%	100%	0%
N2O	100%	100%	100%	100%	100%	0%	100%	100%	100%	100%	100%	100%	100%	100%	0%	100%	100%	100%	100%	100%	100%	100%	0%

2.2) Emission Factors of Fuel Combustion: Feedstock and Fuel Transportation from Product Origin to Product Destination (grams per mmBtu of fuel burned)

	Ocean Tanker							Barge										Locomotive					
	Bunker Fuel	Diesel	Natural Gas	LPG	DME	FTD	Hydrogen	Residual Oil	Diesel	Natural Gas	LPG	DME	FTD	Biodiesel	Renewable Diesel	Renewable Gasoline	Hydrogen	Diesel	Natural Gas	LPG	DME	FTD	Biodiesel
VOC	93,485	93,485	93,485	93,485	93,485	93,485	0,000	38,861	38,861	38,861	38,861	38,861	38,861	38,861	38,861	38,861	0,000	73,948	73,948	73,948	73,948	73,948	73,948
CO	430,006	430,006	215,003	215,003	430,006	430,006	0,000	103,497	103,497	51,749	51,749	103,497	103,497	103,497	103,497	103,497	0,000	213,328	106,664	106,664	213,328	213,328	213,328
NOx	2480,468	2480,468	2480,468	2480,468	1240,234	2480,468	4340,8	1049,627	1049,627	1049,627	1049,627	524,814	1049,627	1049,627	1049,627	1049,627	1836,848	1517,110	1517,110	1517,110	758,555	1517,110	1517,110
PM10	73,450	73,450	7,345	7,345	51,415	73,450	0,000	26,026	26,026	2,603	2,603	18,218	26,026	26,026	26,026	26,026	0,000	35,940	3,594	3,594	25,158	35,940	35,940
PM2.5	55,088	55,088	5,509	5,509	38,561	55,088	0,000	13,013	13,013	1,301	1,301	9,109	13,013	13,013	13,013	13,013	0,000	32,346	3,235	3,235	22,642	32,346	32,346
SOx	1497,030	8,038	0,269	0,000	0,000	0,000	0,000	267,327	8,038	0,269	0,000	0,000	0,000	0,000	0,000	0,000	0,000	8,038	0,269	0,000	0,000	0,000	0,000
CH4	4,581	4,581	91,620	4,581	4,581	4,581	0,000	1,904	1,904	38,083	1,904	1,904	1,904	1,904	1,904	1,904	0,000	3,940	78,800	3,940	3,940	3,940	3,940
N2O	2,000	2,000	2,000	2,000	2,000	2,000	0,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	0,000	2,000	2,000	2,000	2,000	2,000	2,000
CO2	84.102	77.219	58.532	67.420	68.938	75.322	0	84.792	77.910	59.106	67.854	69.629	76.012	79.704	76.326	74.863	0	77.623	58.798	67.653	69.342	75.725	79.416

2.3) Emission Factors of Fuel Combustion for Feedstock and Fuel Transportation: Trip from Product Destination Back to Product Origin (grams per mmBtu of fuel burned)

Ocean Tanker | Barge | Locomotive

Figure 38: GREET screenshot, showing the different emission factors of fuel combustion for various technologies

4.2 Modifications before data analysis

First, data was manually sorted from the multidimensional spreadsheet model in Excel. This provides a small database, available for analysis. Then, data was put into a pivot-table in order to calculate the standard deviation of a “technology level”. This has to be done for each level.

For the comparison to the German emissions, data must be first converted into the same unit (km instead of mile), and then selected and rearranged, as both databases do not have the same vocabulary or description of technologies.

4.3 Limitations

GREET is a very specific database, dealing only with different vehicles’ technologies available in the U.S.A.

4.4 Use in this project

The factor “further technological correlation” can be studied, while looking at the different technologies covered in the GREET database. Also, a comparison can be made with the German/European Tremod database to analyse the geographical indicator.

5 Yoghurt cup sampling study

5.1 Description

The study “How to Obtain a Precise and Representative Estimate for Parameters in LCA. A case study for the functional unit” was held by Andreas Ciroth and Michael Srocka in 2006 in Berlin on yoghurt cups. Aim was to investigate how to best determine the weight of yoghurt cups at point of sale, in an empirical statistical sampling study; the yoghurt cups were functional units in Life Cycle Assessment studies.

Many types of yoghurt cups from supermarkets in Berlin were sampled and weighed, considering different parameters (type of plastic, label, supermarket origin, brand…). Results are available in an internal database that easily allows comparison and analysis.

More information can be found here:
http://www.springerlink.com/content/y7168373353k3431/

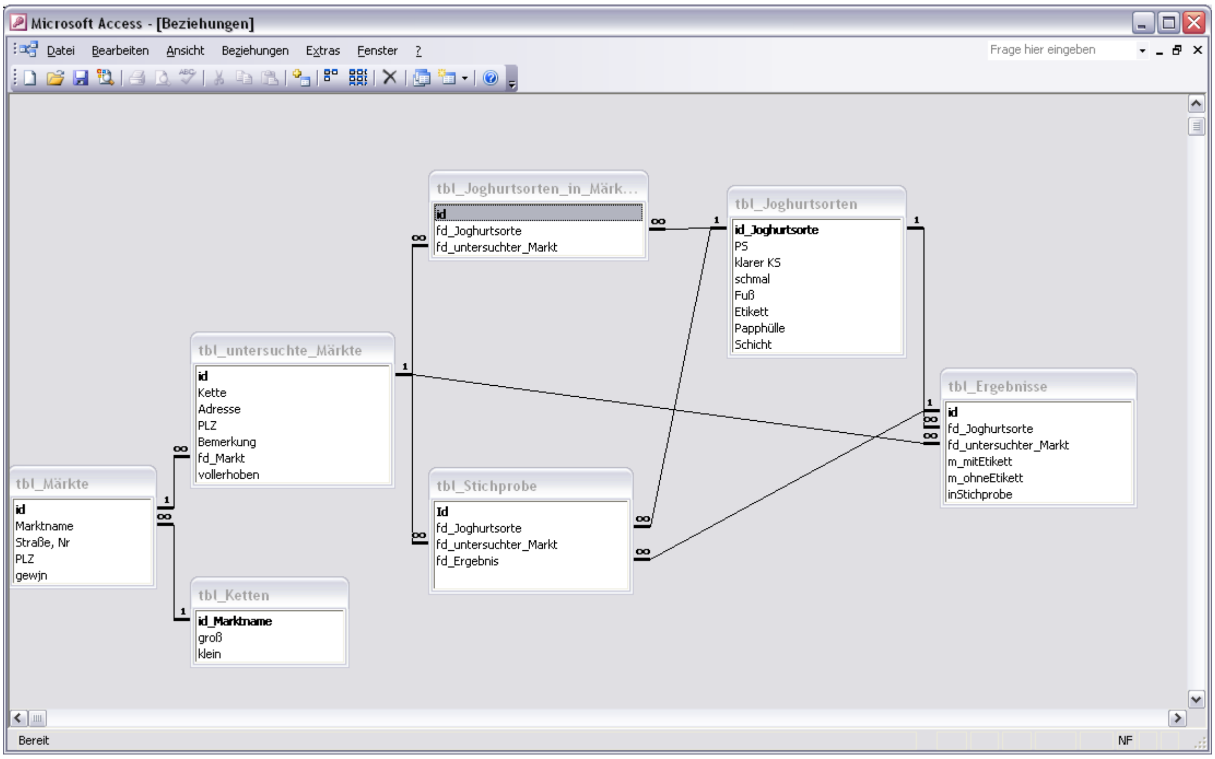


Figure 39: Yoghurt cup sampling study database screenshot, overview of the database structure

5.2 *Modifications before data analysis*

Regarding the database, some calculations must be done, but no major change is needed. For analysing the "completeness" indicator, we need to have different sample sizes of a complete population. And this in turn means we need first to define methods for ordering the data, and for identifying subsets. There is of course no "natural" way to identify subsets, since data can be selected in various ways. Therefore, a number of different approaches for ordering data and for selecting subsets was used, in order to get an overall picture.

5.3 *Limitations*

The database is somewhat small; the complete population available in the database is already a sample of the real population that can be found in markets in Berlin.

5.4 *Use in this project*

Considering the pedigree matrix, the indicator "completeness" can be analysed.

6 North American Transportation Statistics

6.1 Description

"North American Transportation Statistics" is a working group presenting information on transportation and transportation-related activities among Canada, the United States and Mexico. The group gathers and analyses information from the "Bureau of Transportation Statistics" of each country since 1991. More information can be found here: http://nats.sct.gob.mx/.

Three tables are especially interesting as they deal with the main pollutants (CH_4, CO_2, N_2O) emitted from transport modes (rail, road, air, marine, general, others) in North America (Mexico, Canada, USA) since 1990.

6.2 Modifications before data analysis

Emissions are given for each year in "Thousands of metric tons of CO_2 Equivalents", per country, in absolute figures. As an effect, the USA as larger country have much higher emissions than Canada for example. To be comparable, data needs first to become a "relative emission". This can be achieved by dividing each value by the total emission amount of each country in 1990, as a reference. This approach was used in this project.

6.3 Limitations

Dividing the reported values in the database per country by the reported values for the year 1990 is much simpler than, e.g., dividing the reported emissions by the amount of traffic; since the amount of traffic per means of transport is not available in the transport statistics database, this information would need to be provided from other sources, and is not always available. Our approach avoids mixing data from different data sources, and is able to transform the reported absolute values in relative information that can be compared.

However, since 1990, traffic intensity has changed in all countries, probably to a different degree. A geographical comparison for one year, from one country to another, then always comprises the comparison of the "transport intensity development" since 1990, in the compared countries. Therefore, a correlation between geography and time can be assumed. This will be checked in the analyses.

6.4 Use in this project

The "geography correlation" and the "temporal correlation" can be analysed.

7 Mexican cement

7.1 Description

This database has been established by CADIS (Centre for Life Cycle Analysis and Sustainable Design). It deals with the environmental impacts of cement plants in Mexico, between 1993 and 2004. Data have to be kept confidential, but we can dispose of the results from our analyses.

The database lists some Mexican cement plants, their products and pollutants to water, air, as well as their raw materials and energy consumptions. Data come from different environmental audits.

More information about the CADIS can be found here: www.centroacv.com.mx.

7.2 Use

This database can be useful to compare results from other databases and also to analyse the “basic uncertainty factor”, as it contains pollutant emissions to water and to air.

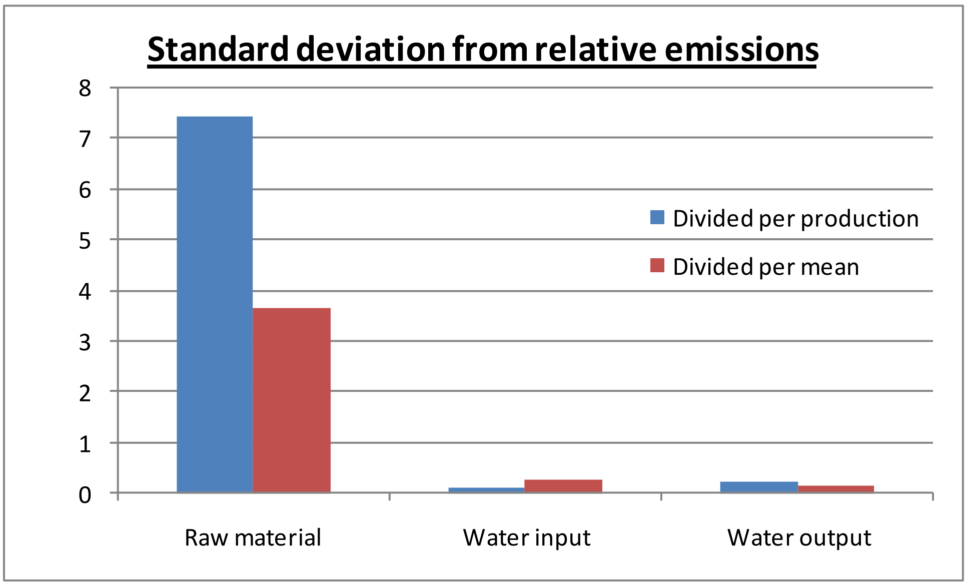
7.3 Modifications before data analysis

Data have usually different units and are not expressed per production amount. This required a transformation into relative emissions (per production or per mean), and also transformation, when possible, into the same unit.

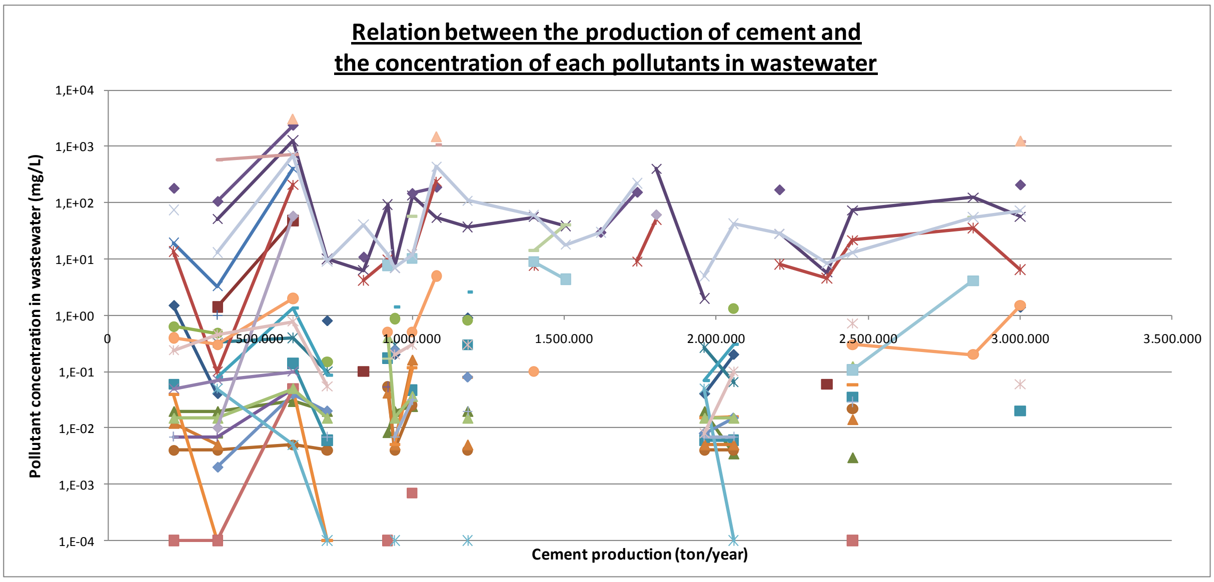
7.4 Results

Here are different charts obtained with this database. Different analyse were performed in order to make the most of it.

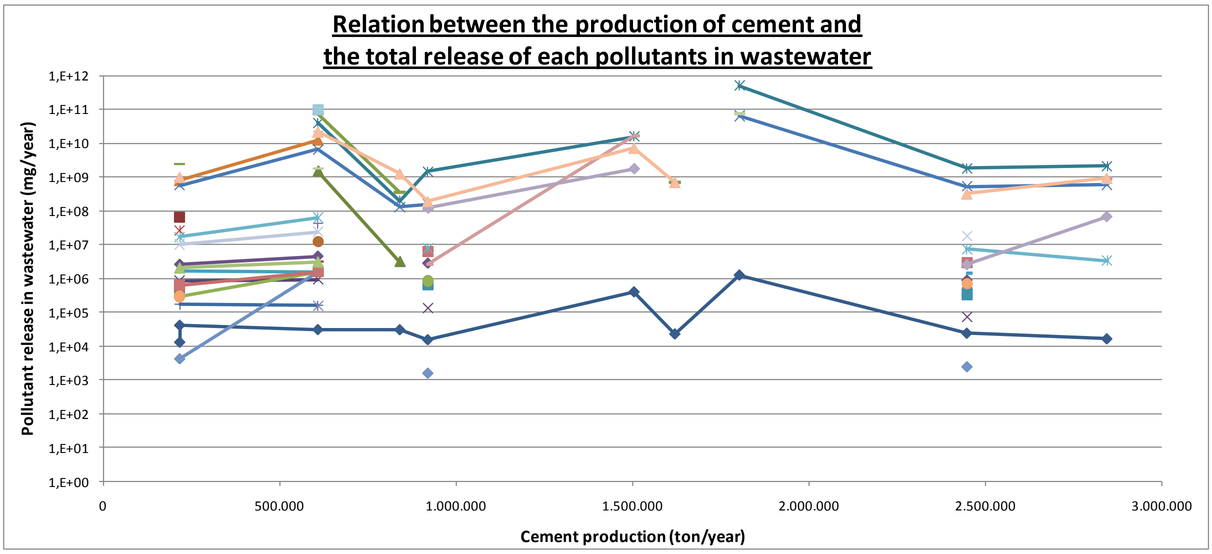
As the production amount per cement factory is not always known, it is important to define if we can obtain relative emissions by dividing by the mean of emissions. The chart below shows that both relative values are close to each other.



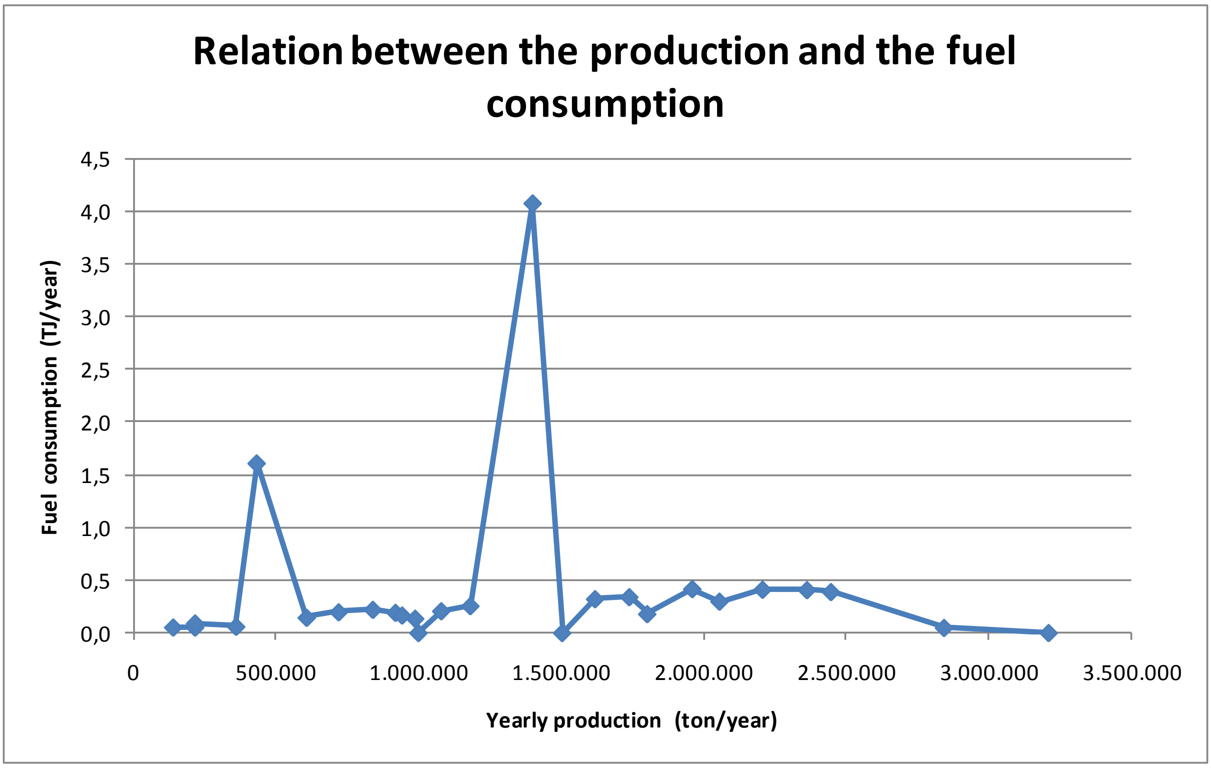
Pollutions to water are often given in mg/L. the following chart shows that there is no link between the production of cement and the concentration of pollutants in wastewater:



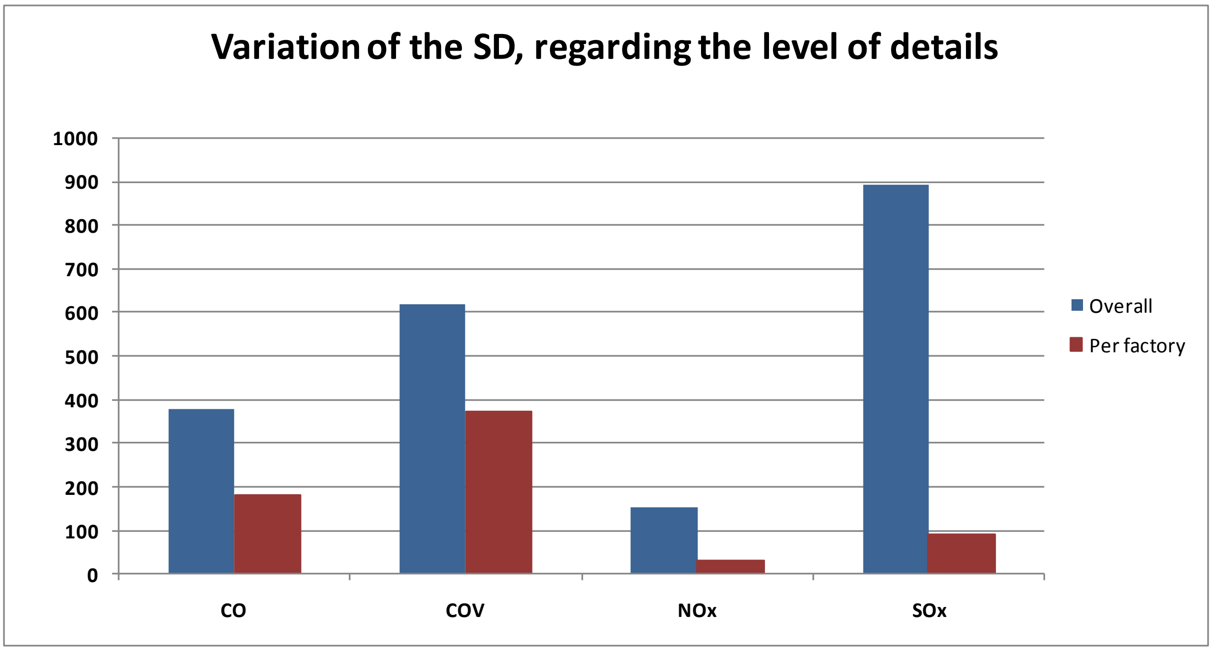
When multiplying these concentrations by the amount of waste water, one can see that there also is no obvious relation between the production of cement and the total quantity of pollutants released to the water.



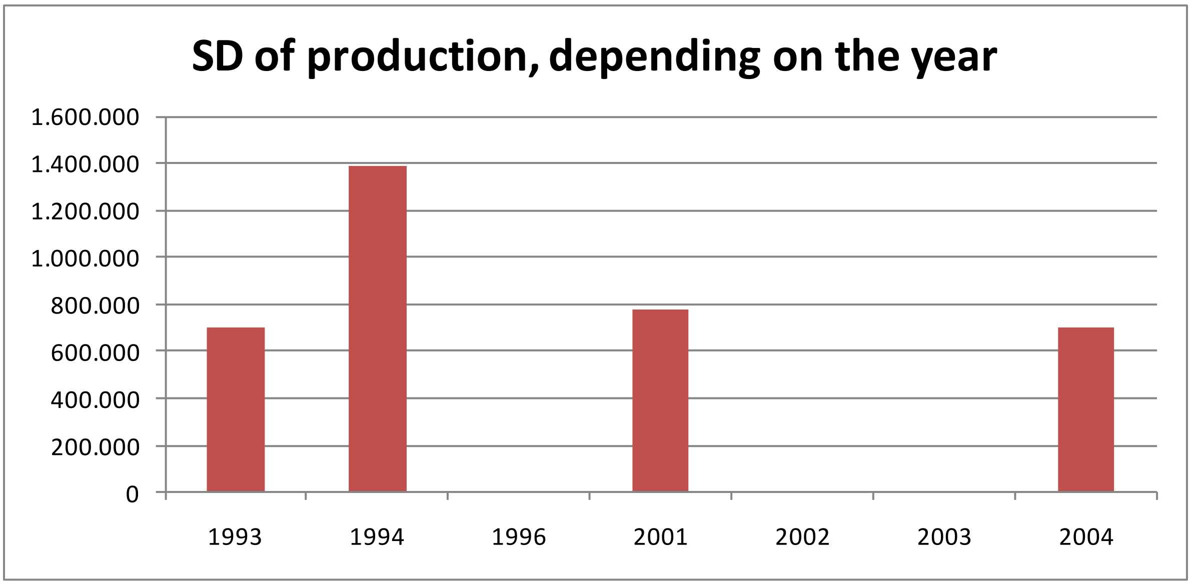
The following chart is also interesting as it displays the relation between the cement production and the fuel consumption. It seems also that there is no such relation.



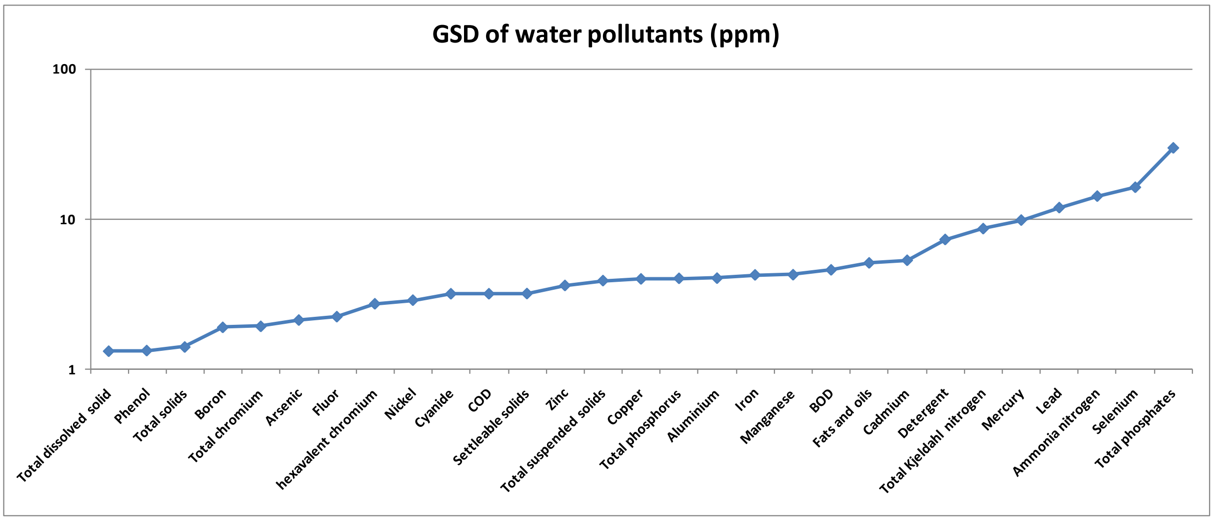
To analyse the “further technological correlation”, it is possible to look at two different levels: overall level and factory level. The standard deviation of 4 air pollutant concentrations has been calculated. The following chart shows clearly that results of chapter 5.6 are confirmed.



Unfortunately, the analysis of the “temporal correlation” does not show any obvious relation.



While looking at the “basic uncertainty factor”, geometric standard deviation of water pollutants is possible. The chart below shows that this geometric standard deviation depends a lot on the pollutant itself. This confirms the results of the chapter 5.7.



7.5 Conclusion

This database does not seem to be of a major interest, but it is still useful to double check some results already gained with deeper analyses.

Annexe B: Normalisation options

Normalisation options

Andreas Ciroth

January 5, 2011

ciroth@greendeltatc.com

1 Motivation & background

For the analysis of uncertainty in many different data sources, as it is necessary for the empirical foundation of uncertainty factors that is currently performed in the ecoinvent pedigree project, the scale of the data needs to be considered. In the pedigree project, uncertainty is understood as the variance in data, and the variance of data depends on the scaling of the data. For example, multiplying all data with a constant factor yields a variance that is multiplied with the square of this factor.

When the variance in different data sources is analysed, the variance should, ideally, be fully comparable. This is not the case if the scaling of data is different.

For LCA related data, there are three main reasons for different scales in data:

4. data may not be provided per functional unit at all; this requires a transformation of the data, for example from absolute emission figures of an industrial plant to “per kg product” emission figures
5. if a functional unit is given, the quantitative reference may differ (1000 m² for one data source or group of data; 1 m² for another)
6. data may simply be provided in different units (kg emissions vs. emissions in grams)

These different scales must not be mixed with true differences in data.

2 Possible and proposed approaches

There are two principally different ways to deal with differently scaled data, with the aim to overcome the scaling effect, as good as possible. These two different approaches are

a. take the arithmetical variance / arithmetical standard deviation directly as uncertainty measures, and transform the data where necessary, to remove possible scaling effects. For example, transform all process data sets so that the quantitative reference is 1 kg, for each data set.
b. take the geometric variance / geometric standard deviation instead; this measure is not directly a measure for uncertainty, but a factor that is itself dimensionless. It needs to be multiplied by the geometric mean to provide a measure for the spread in the analysed data. Since the geometrical standard deviation is dimensionless, data needs not be transformed to remove scaling effects (!)[21]. As a downside, the calculated uncertainty measure (geometric mean times geometric standard deviation) is less straightforward to interpret.

2.1 Arithmetical variance / arithmetical standard deviation as uncertainty measures

The arithmetical variance is the variance that is commonly used, defined as #. The corresponding standard deviation is one of the two parameters of the normal probability distribution.

[21] http://www.thinkingapplied.com/means_folder/deceptive_means.htm

2.1.1 Standard deviation as uncertainty measure for the pedigree factors

The original pedigree matrix formula, presented e.g. in the ecoinvent 1 methodology report, was meant to be used for lognormally distributed data only. The formula is now available also for data that follow other distributions (see chapter 8).

The formula does not consider the scale of the data; therefore, the uncertainty factors need to be independent from scale.

As stated in the introduction, the standard deviation depends on the scale of the analysed data. If the standard deviation is used in the pedigree matrix, this scaling effect needs to be removed as good as possible before the analysis results can be used in the pedigree formula; removing the scaling effect is the goal of the data transformations.

2.1.2 Data transformations

For the variance and standard deviation as measure for the uncertainty, reasons for different scales in data will be addressed as follows.

2.1.2.1 Data not given per amount of product

Process data sets in Life Cycle Assessments have always a functional unit that is based on a unit of product (be it kg product, MJ, or vehicle kilometre for example). If raw data is given without reference to a product, then the reference to the product amount needs to be established.

Specifically, emission data bases (PRTR in Europe, TRI in the US) provide information of emissions per production facility. The production volume of each facility is usually not available.

The data transformation is performed as follows.

First, a linear relationship between the absolute emissions E and the amount of the produced product x is assumed:

$$E_{ij} = e_{ij} * x_j \quad for\ i = 1,..., n,\ j = 1,..., m$$

With E_{ij}: absolute emissions reported per plant j, for emission type i;
e_{ij}: specific emissions per amount of product per plant j, for emission type i

This fits to the usual process modelling in LCA, as linear processes.

The average of the absolute emissions for process j is

$$\overline{E}_j = \frac{1}{n} * \sum_{i=1}^{n} E_{ij} = \frac{1}{n} * \sum_{i=1}^{n} (e_{ij} * x_j) = x_j * \frac{1}{n} * \sum_{i=1}^{n} e_{ij} = x_j * \overline{e}_j$$

With $\overline{e}_j$: mean of the specific emissions

Dividing the absolute emissions by the mean of the absolute emissions yields the specific emissions, divided by the mean of the specific emissions:

$$E_{ij} / \overline{E}_j = e_{ij} / \overline{e}_j$$

Dividing by the average of the absolute emissions has the advantage that the product amount is indeed removed; further, the average can always be calculated and is therefore available for every process / facility.

It has also two disadvantages. First, the specific emissions are (still) not available after the calculation; and second, the remaining factor "$1/\overline{e}_j$ " makes the variance for the calculation result smaller, by a factor of $(1/\overline{e}_j)^2$. The factor is the average emission per amount of product, squared. This needs to be considered in the interpretation of the variance / uncertainty results.

2.1.2.2 Differing quantitative reference

Process data sets with different quantitative references need to be transformed to identical quantitative references wherever possible. A factor of thousand, for example, in quantitative references yields a difference in variance in 1000^2, i.e. 1E+6.

The transformation is straightforward, one common quantitative reference $q_{reference}$ needs to be selected, e.g. 1 (amount of product, in suitable unit), and process data sets with differing quantitative references q_j need to be scaled accordingly, by multiplying all their input and output with the factor $q_{reference}$ / $q_{j.}$

2.1.2.3 Differing units

Datasets with different units should be transformed to the same unit wherever possible. This is especially relevant when the units differ by several orders of magnitude (as with kg and tonne for example – in this and many similar cases, a unit transformation is easily possible). It will not always be possible to achieve consistent units, especially if units are provided in different unit groups as energy content and mass. In these cases, units should be made consistent as good as possible.

2.2 Geometric variance / geometric standard deviation

The geometric standard deviation σ_{gj} per data set j is calculated as follows:

$$\sigma_{gj}(E_j) = \exp\left(\sqrt{\frac{\sum_{i=1}^{n}(\ln E_{ij} - \ln \mu_{gj})^2}{n}}\right).$$

There is an interesting relation to the (arithmetical) standard deviation[22]:

The geometric mean μ_g is given by

$$\mu_g = \sqrt[n]{E_1 * E_2 * ... * E_n}$$

Applying the logarithm to both sides of the equation yields

$$\ln(\mu_g) = \frac{1}{n}\ln(E_1 * E_2 * ... * E_n) = \frac{1}{n}\sum_{i=1}^{n}\ln(E_i) = \mu(\ln E)$$

So, the logarithm of the geometric mean is the arithmetical mean of the logarithm of the analysed data.

Therefore, the (arithmetical) standard deviation of the data is

$$\sigma(\ln E_j) = \sqrt{\frac{\sum_{i=1}^{n}(\ln E_{ij} - \ln \mu_{gj})^2}{n}}.$$

Comparing to the formula provided for the geometrical standard deviation above gives

[22] To my shame, I found this on wikipedia only so far: http://en.wikipedia.org/wiki/Geometric_standard_deviation

$$\ln(\sigma_{gj}(E_j)) = \sqrt{\frac{\sum_{i=1}^{n}(\ln E_{ij} - \ln \mu_{gj})^2}{n}} = \sigma(\ln E_j) \qquad (*)$$

The logarithm of the geometric standard deviation equals the arithmetical standard deviation of the logarithm of a data sample.

One could also say

$$\sigma_{gj}(E_j) = \exp(\sigma(\ln E_j))$$

This relationship explains an interesting property of the geometric standard deviation about scaling effects in data: The geometric standard deviation is not affected by constant factors in analysed data.

From $\ln(E_j) = \ln(e_j{*}x_j)$ follows

$$\ln(e_j{*}x_j) = \ln(e_j) + \ln(x_j).$$

If, similar to the section about the arithmetical standard deviation above, e_j are the different emissions for a process j, and x_j the production volume of the process, then x_j is constant for all emissions of the process. The arithmetical standard deviation of the logarithm is

$$\sigma(\ln E_j) = \sigma(\ln e_j + \ln x_j)$$

Since ln (xj) is constant, it holds

$$\sigma(\ln e_j + \ln x_j) = \sigma(\ln e_j)$$

And since $\sigma_{gj}(E_j) = \exp(\sigma(\ln E_j))$

it holds that

$$\sigma_{gj}(E_j) = \sigma_{gj}(x_j * e_j) = \exp(\sigma(\ln E_j)) = \exp(\sigma(\ln x_j * \ln e_j)) = \exp(\sigma(\ln e_j)) = \sigma_{gj}(e_j)$$

Constant factors in the analysed data do not influence the value of the geometric standard deviation. The geometric standard deviation is independent from scaling effects in data.

2.2.1 Geometric standard deviation as uncertainty measure for the pedigree factors

The geometric standard deviation is dimensionless and therefore not directly applicable as indicator for the uncertainty, in contrast to the arithmetical standard deviation.

However, a range can be calculated, similar to the confidence interval for the normal distribution; instead of

$CI_{upper} = \mu + \sigma$; $CI_{lower} = \mu - \sigma$, as it is valid for the normal probability distribution,

the relation for the geometric standard deviation is

$range_{upper} = \mu_g * \sigma_g$; $range_{lower} = \mu_g / \sigma_g$.

Note that this relation is not directly linked to the lognormal probability distribution. The calculated range is not symmetric.

The calculated geometric standard deviations can, however, be directly inserted into the calculation formula for the calculation of the overall geometric standard deviation (reference to formula).

2.2.2 Data transformations

Constant scaling factors in the analysed data do not pose a problem for the geometric standard deviation, as explained above. Therefore, data transformations will often be skipped completely (or, calculating the geometric standard deviation is already a log-transformation of data, and a second data transformation step is not required).

However, a scaling effect can only be ignored if a constant factor applies for the whole set of the analysed data; this is true on the process level, if processes can be assumed as linear. It is not true if σ_g is calculated for example for two processes with different quantitative references. In this case, the differing quantitative references contribute to the calculated standard deviation.

Therefore, also for the geometric standard deviation (additional) data transformations are necessary, depending on the analysis and the data sample.

In principle, the same transformations apply as described for the arithmetical standard deviation, above:

Differing units in data sets should be aligned, differing quantitative references should be made consistent, and specific emissions should be analysed in stead of absolute emissions that are given per production volume.

2.3 z-transformation

In statistical data analysis, scaling effects in data are often addressed by applying a z-transformation (Fisher z-transformation). This transformation subtracts the mean from all data sets in the sample, and divides this difference by the standard deviation. The result has a mean of 0, and a standard deviation of 1, which makes an analysis very convenient. It removes, however, differences in variance, and is therefore in principle not applicable in the course of this project.

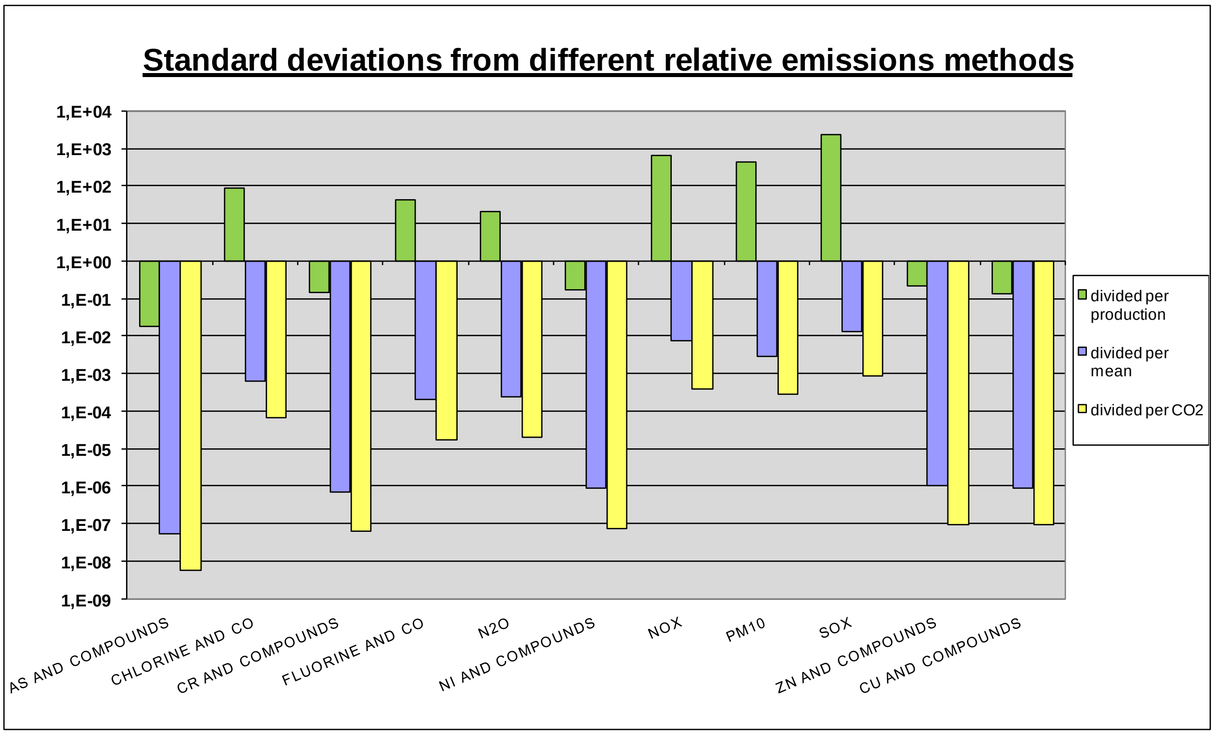
There are exceptions, though. For example, if the z-transformation can be applied to a whole data source/sample, then differences of data within the data source are still visible after the transformation, and results from the analysis of the data are independent from the scale of data in the sample. However, a disadvantage remains, because the z-transformation does not distinguish between differences in data due to different scales and differences due to “uncertainty differences” – all contribute to the calculated standard deviation that is used as the denominator.

3 Analyses

The differences between arithmetical and geometrical standard deviation are analysed on behalf of a small case study: For four coal power plants contained in the PRTR data the yearly production amount was available from other sources. This allows a comparison of the relative and the absolute values in PRTR compared to the true values.

Dividing for each of the power plant the PRTR-reported emissions by the production amount yields the relative data we are looking for. Based on that, we calculated other relative data (divided by the average release, by the CO_2 quantity) and compared them to the previous results. On the chart below, the data “emission divided by production volume” in green is our reference; it is the true relative value, per pollutant. The other relative emissions (per mean and per CO_2) are both rather close to the reference, but not always.

The following chart displays the standard deviations for three “relative emission” methods:



It is also possible to calculate a coefficient of variation (= Standard deviation / mean). This brings all values to a number between 0 and 1 and it allows fair comparison between the three standard deviations. In our case, we can see that the standard deviation "per mean" is close to the standard deviation "per production".

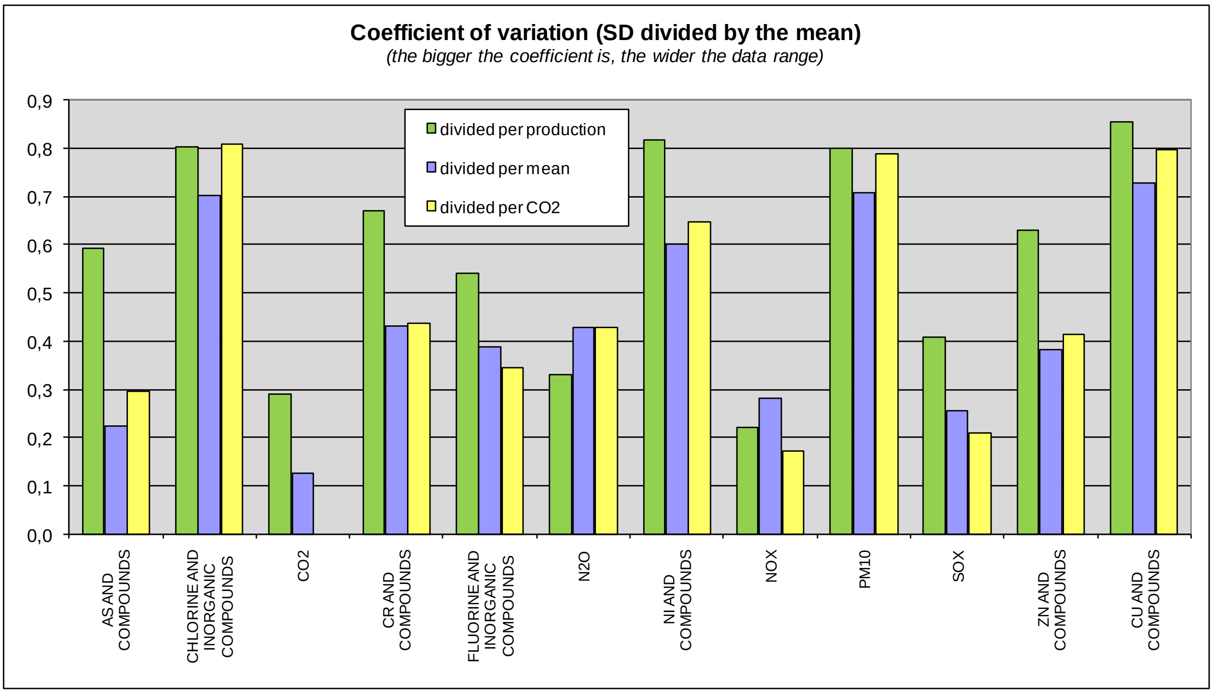


Figure 40: Comparison of options to obtain relative data, for four coal-power plants

Building the ratio of calculated mean vs. reference (true relative value) gives an overview of how well the data transformation preserves the original standard deviation. The result is shown in Figure 41.

An ideal data transformation would yield always a "1" for this ratio; the calculated standard deviation of transformed data would be the same as the true standard deviation of the relative, per-product amount, emissions. This is never the case; for CO2, the largest emission flow for the coal power plants, the ratio is around 7 if data is divided by the mean, and of course 1 if data is divided by CO2 emissions. For all other flows, the standard deviation of the transformed data is much smaller than the true standard deviation, up to a factor of 1E+8.

Dividing by the mean produces a slightly better result than a division by CO2 quantity; however, both transformation approaches overall fail to reproduce and "preserve" the standard deviation in the original data sets.

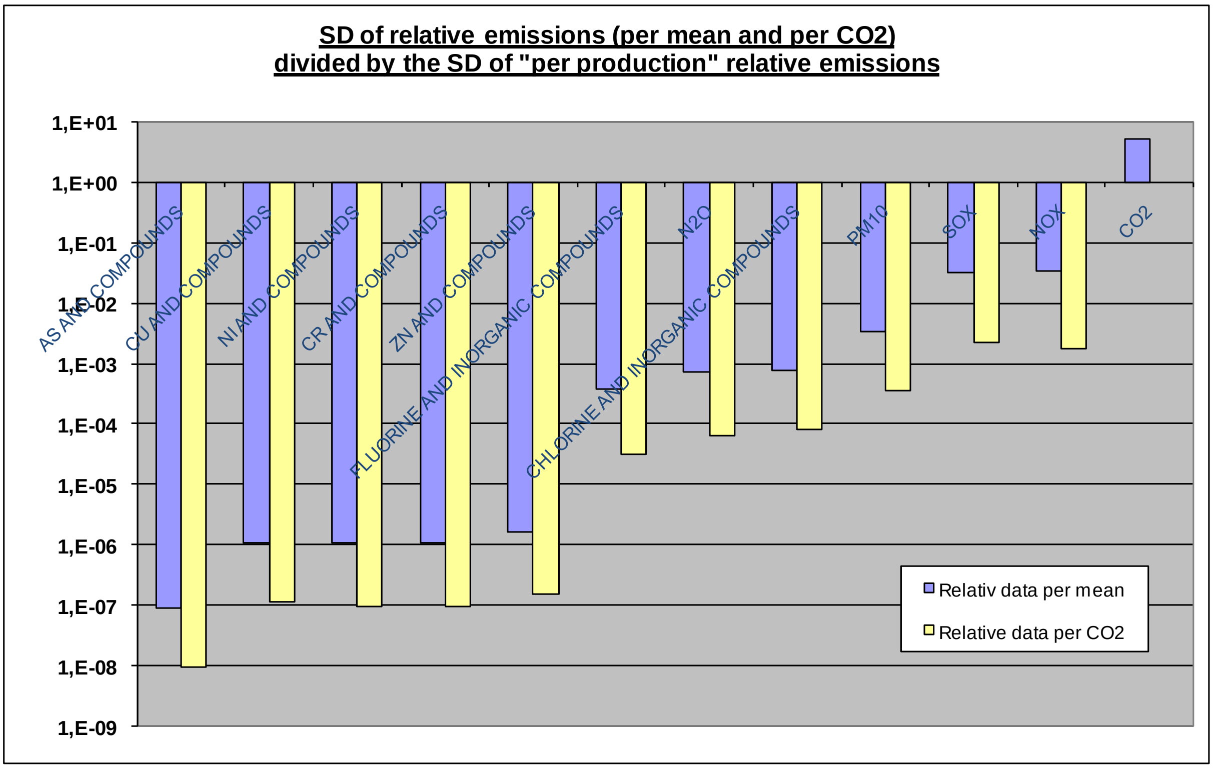


Figure 41: Comparison between 2 relative data, for four coal-power plants

Using the same dataset, we have compared different ways to analyse the direct emissions. First, the chart below compares, for the absolute emissions per plant, the (arithmetical) standard deviation with the geometric standard deviation, which is multiplied by the geometric mean to make it comparable. The figure shows that both options provide rather similar values (cf. Figure 42).

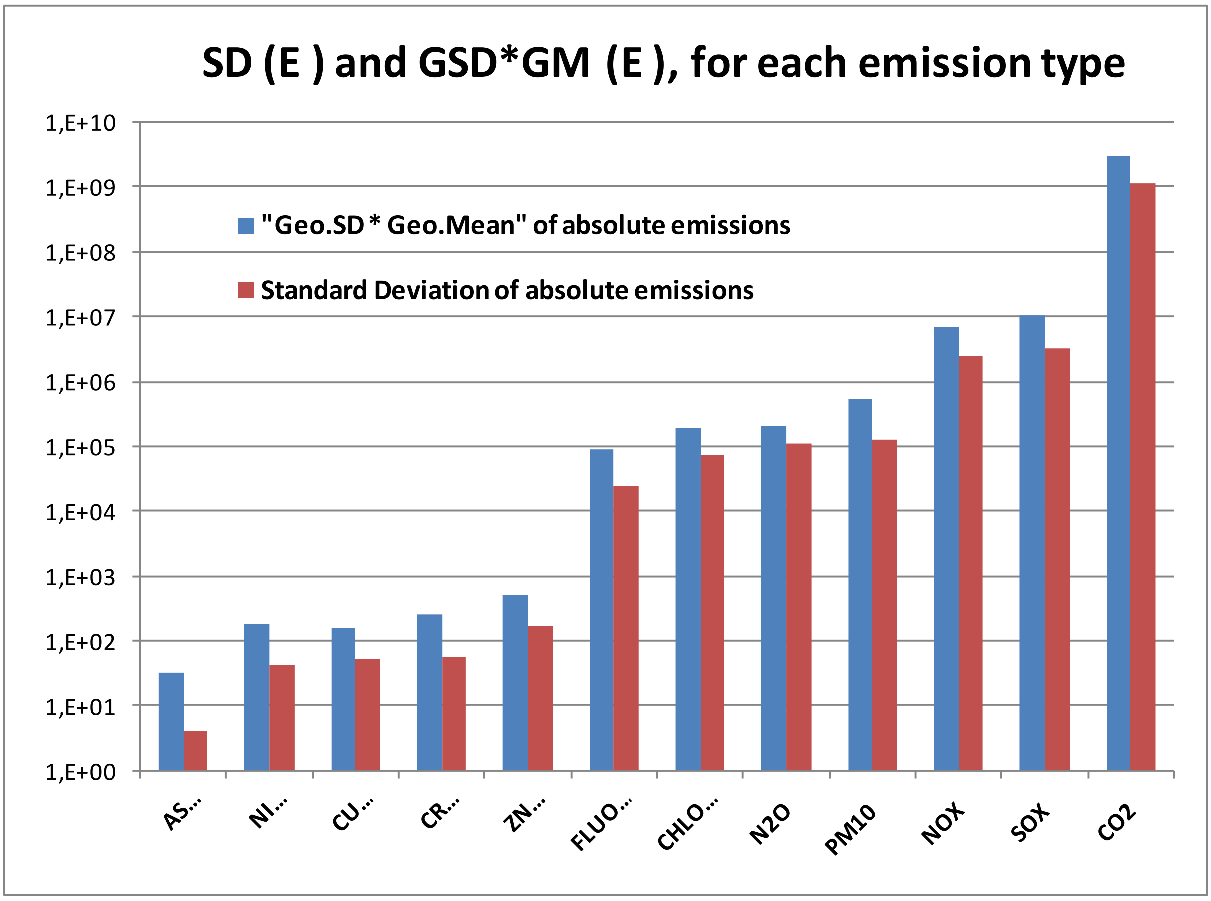


Figure 42: Comparison of analysis with "standard deviation" and "geometric standard deviation", for absolute emissions

The second chart provides the figures for the relative emissions, i.e. the absolute emissions per plant divided by the true production. Again, arithmetical and geometric standard deviations are compared, and again, the geometric standard deviation is multiplied by the geometric mean to make it comparable. Also in this figure, both options are also closely related.

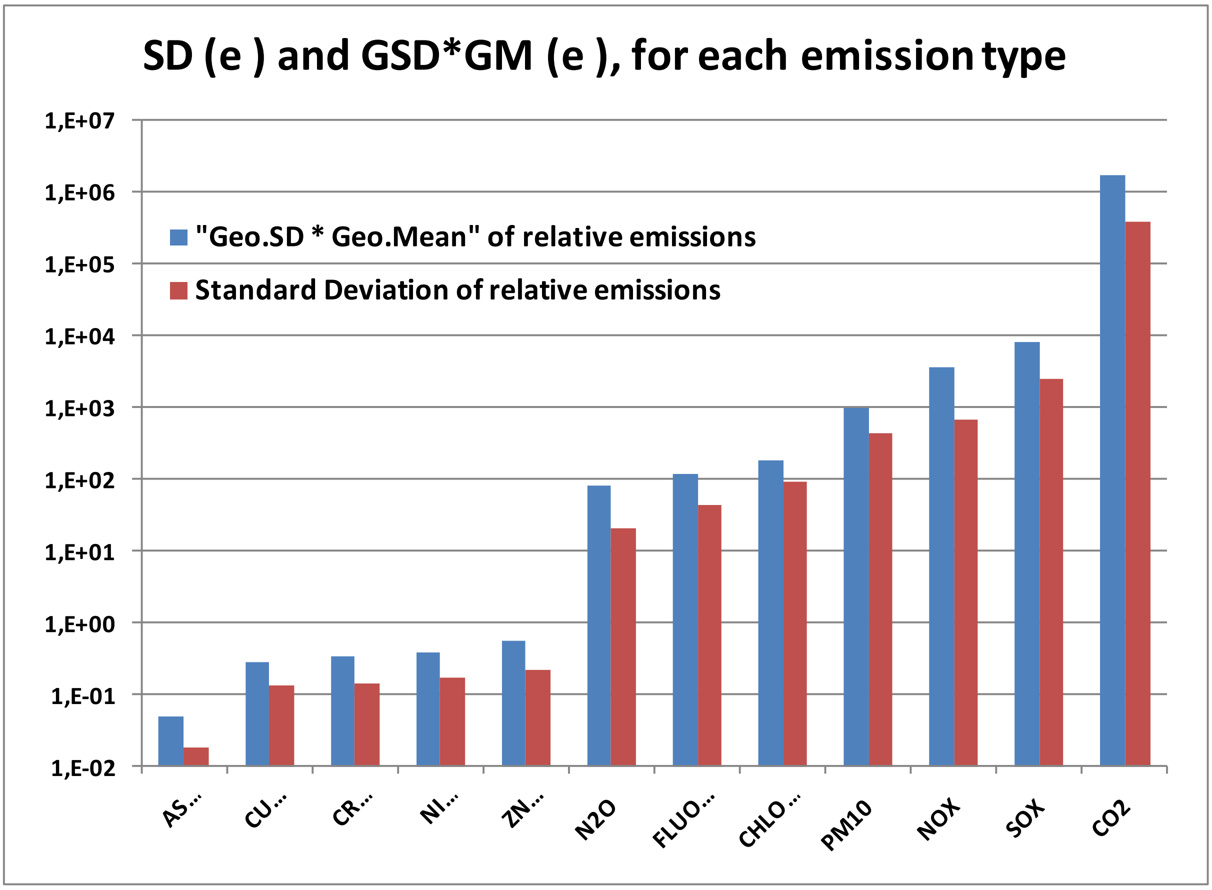


Figure 43: Comparison of analysis with "standard deviation" and "geometric standard deviation" for relative emissions

4 Conclusions

The z-transformation is as such not suitable for the analysis. The geometric standard deviation has advantages over the (arithmetical, usual) standard deviation, because constant factors in data do not contribute to the calculated uncertainty. It fits directly to the pedigree formula that is used in the current ecoinvent methodology reports (2.2 and older).

On the other side, results of the geometric standard deviation are more difficult to interpret, although a formula exists to calculate similar results as from the (arithmetical) standard deviation.

Conclusion is therefore to prefer the geometric over the arithmetical standard deviation in the analysis. In some instances, the arithmetical standard deviation may be useful in addition for better understanding the characteristics of the analysed data.

If the geometric standard deviation is used, then no data transformation is necessary.