

Mathematical analysis of ecoinvent data

Final report

Andreas Ciroth, June 2009

ciroth@greendeltatc.com

1	Aim & Objective	2
2	Background	2
3	Preparatory work: Creation of an access database with ecoinvent data.....	4
4	Analyses	6
4.1	Basic plausibility checks	7
4.1.1	Mass balance of a process	7
4.1.2	Input and output pattern	11
4.1.3	Negative flows.....	15
4.1.4	Carbon dioxide emissions per fuel consumption	17
4.2	Statistical analyses and procedures	20
4.2.1	Explorative data analysis (EDA).....	20
4.2.1.1	Boxplots: Transport effort in tkm.....	20
4.2.1.2	Scatter plots: CO ₂ emissions, climate change potential	24
4.2.1.3	Matrix scatter plots: Impact assessment categories.....	27
4.2.2	Statistical tests	30
4.2.3	Cluster analysis: Impact categories, and impact category results per process .	33
4.2.4	A heatmap of mean values per flow category and process category.....	36
4.2.5	Distribution tests: Fossil CO ₂ emissions, and radionuclides.....	38
4.2.6	Regression: CO ₂ vs. GWP.....	44
4.2.7	Principal Components Analysis: Underlying factors for flows of machinery processes?.....	46
5	A summary of analyses results, and of evaluation criteria.....	50
6	Towards a practical application for ecoinvent data.....	52
6.1	Towards a learning system of data mining and quality assurance	52
6.1.1	A general procedure for the development and maintenance of a data mining and quality assurance system	52
6.1.2	Elements of a learning system.....	54
6.1.3	Analyses and their results and use, in a learning system.....	56
6.2	Integration of mathematical analyses in a test suite?	57
6.2.1	Available statistical and data analysis software suites	57
6.2.2	Using R in Eclipse.....	59
7	Conclusions and outlook	61
8	References	62

1 Aim & Objective

The project “mathematical analysis of ecoinvent data” has three main objectives.

First, aim is to look for structures in data, and to possibly detect inconsistencies and errors in data. This aim can be summarised as data mining and automated quality check.

Second, aim is to explore different methods and tools available from multivariate statistics and elsewhere, in regards to their potential for detecting structures and for quality assurance. In the evaluation, different addresses of the methods and tools need to be considered.

And third, aim is, based on point 1 and 2, to develop and characterise one or several possible strategies for an automated quality assurance of ecoinvent data. These quality assurance strategies need to consider various application cases, such as a review of individual data sets, detection of possible flaws in the entire database, or of data sets that need to be replaced.

2 Background

Quality assurance of datasets to be used in Life Cycle Assessments and related applications is important. The ecoinvent database has since long paid high attention to it. Inter alia, each dataset has undergone an independent review, by external experts.

Many tasks that need to be performed in an individual review are, however, of a rather routine nature (completeness checks; comparison of values), and require additional calculations (mass balance calculations; calculation of single element balances such as C or metals; calculations of mean or of other indicators from raw data). Often, results benefit from a visualisation of structures in data that need so far to be done individually by each reviewer.

Some aspects can, further, only be seen on a larger scale, when looking not only at one but on several (similar or different) datasets. A statistical analysis of data, with the possibility to perform also tests that reveal whether, for example, it can be assumed that one group of data sets has indeed “significantly” higher values than another group, is only possible with access to these two groups of data sets¹.

On the other side, several powerful tools and methods for a mathematical and statistical analysis of data sets, and for a data mining, became rather easily available recently, including the open source software R (www.r-project.org). For specific applications, dedicated, powerful tools have emerged that allow a customised, individual analysis of data structures, also for non-statisticians. One example of the latter is GeneMaths², a tool designed for analyses of microarray data (in the analysis of genetic data).

¹ In statistics, the term „significant“ is defined as “probably caused by something other than mere chance” (www.merriam-webster.com/dictionary/significant) – having detected a significant relation in this sense is therefore a really strong finding. This contrasts the broad use of the term significant, also in LCA quality assurance and review, with the meaning of “quite clear”, or “important”, even if no uncertainty analysis is conducted.

² www.applied-maths.com/genemaths/genemaths.htm

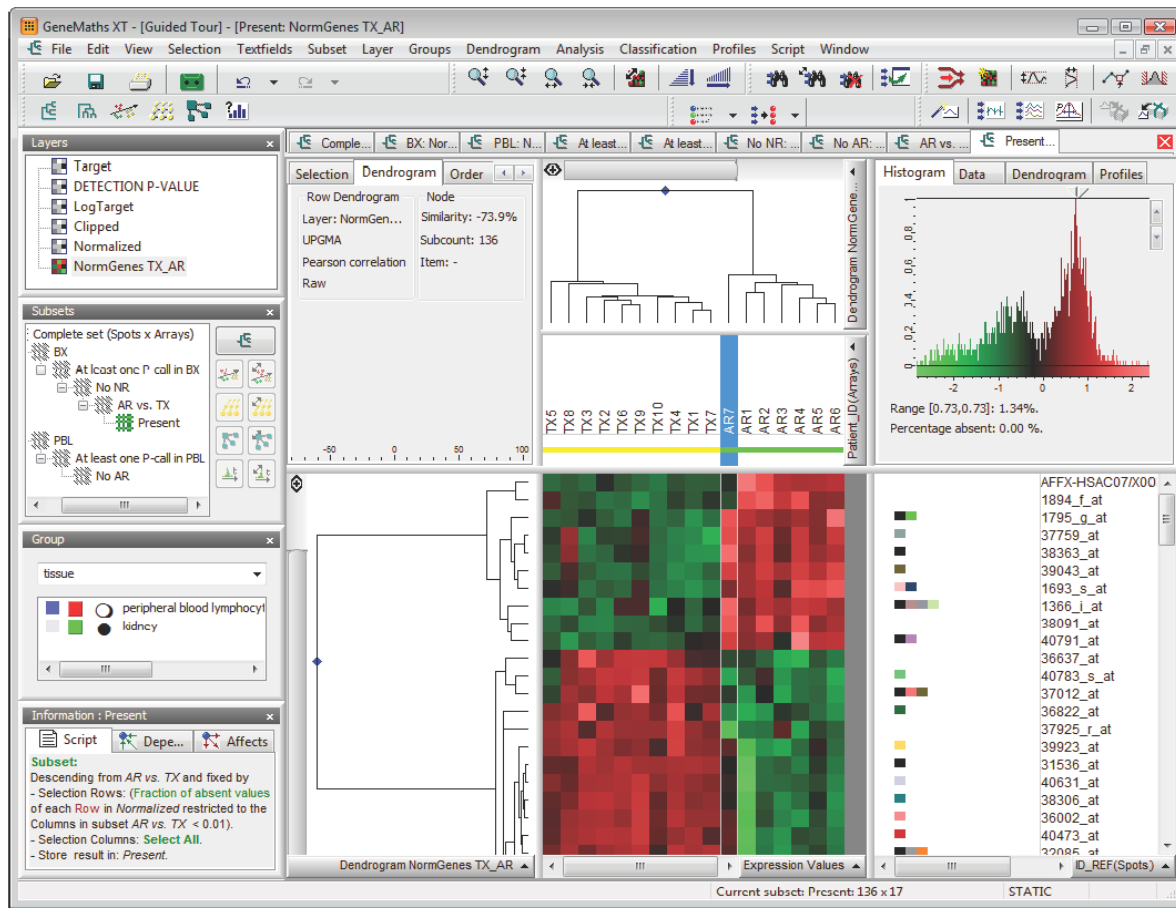


Figure 1: GeneMaths as example for a tool dedicated to statistical data analysis in a specific application field, screenshot²

For ecoinvent data, no “joint” database was available initially that would contain all to-be-analysed data sets. Data sets existed as separate XML files, for unit processes and for system processes, including information on material flows and on environmental impacts (fig. 2).

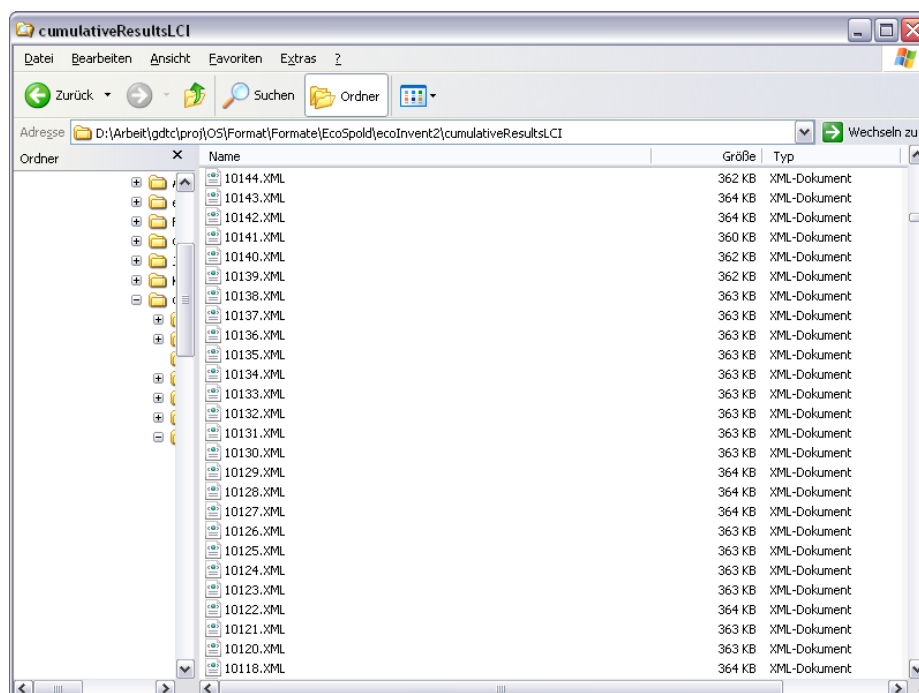


Figure 2: XML process files, in EcoSpold format, as starting point (here: system processes)

As an addition, the ecoinvent centre provided several Excel files with all process data and environmental data.

3 Preparatory work: Creation of an access database with ecoinvent data

As a first step, process data sets from ecoinvent version 2.0.1 were imported into an Access database. All LCI result processes (aggregated over the life cycle) and all unit processes (non-aggregated) were imported this way; multi-output processes were not imported.

The structure of the Access database is pretty straightforward:

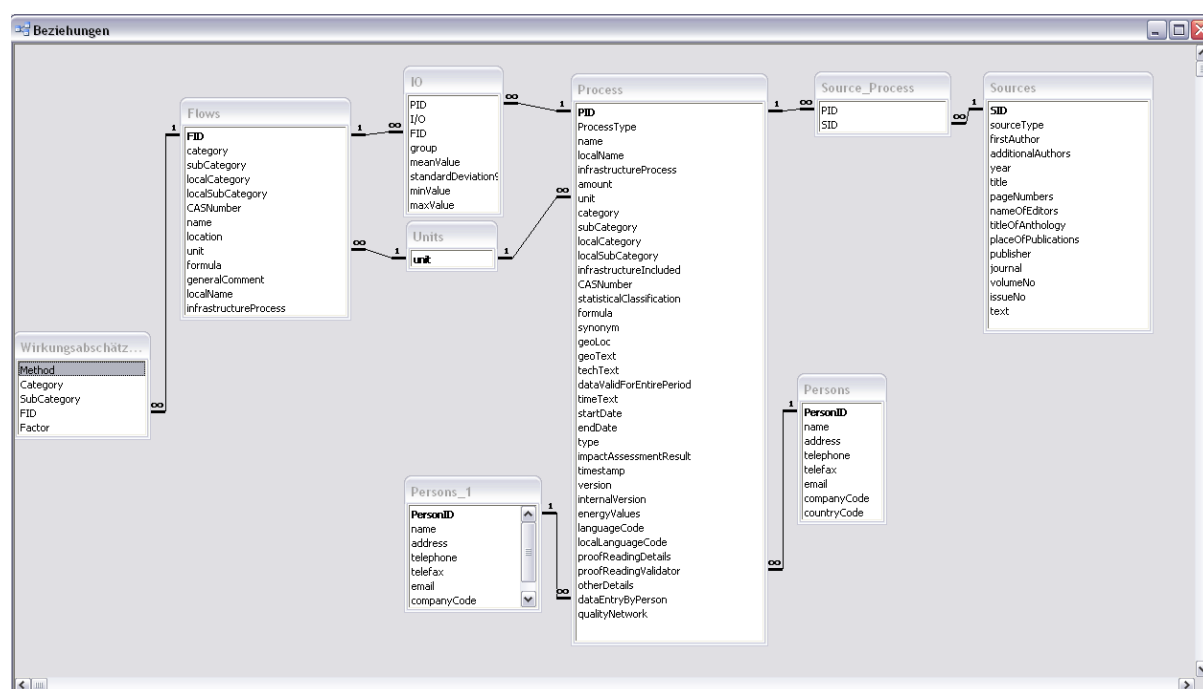


Figure 3: Structure of the created Access database “Ecoinvent.mdb”

With all data imported, the size of the database is around 370 MB, which is still manageable. Also the table sizes remain manageable, with the 3.5 million datasets in the “IO” table, for process exchanges, being the largest one.

The import tool references the EcoSpold Java archive (ecospold.jar) file which was created in the course of openLCA’s format converter project.

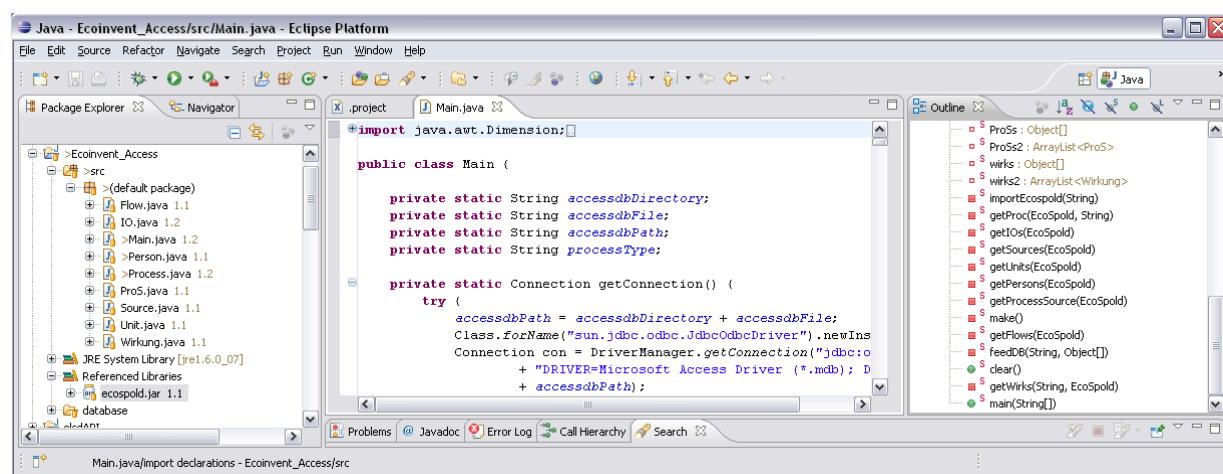


Figure 4: Screenshot of the Eclipse IDE with the “Ecoinvent_Access” project; under referenced libraries, the ecospold Jar file is visible

The Access database and the tool to import into the database are available on request.

Another option would have been the openLCA database (www.openlca.org), since openLCA is able to import EcoSpold files, and its MySQL database has a better performance than Access. However, openLCA does not fully implement EcoSpold, and therefore some fields would be lost; also, interaction with Excel, and creating queries, is slightly more convenient in Access, for this pilot project. A migration from Access to another relational database with better performance and with a specification that allows higher data loads (MS SQL Server, MySQL, postgresQL, and other database systems) is easily possible later on.

Analyses that are described in the following chapter were performed in part in Access, by queries, in Excel, and for a larger part in R, version 2.8.1. In R, data from Access was imported via an ODBC interface.

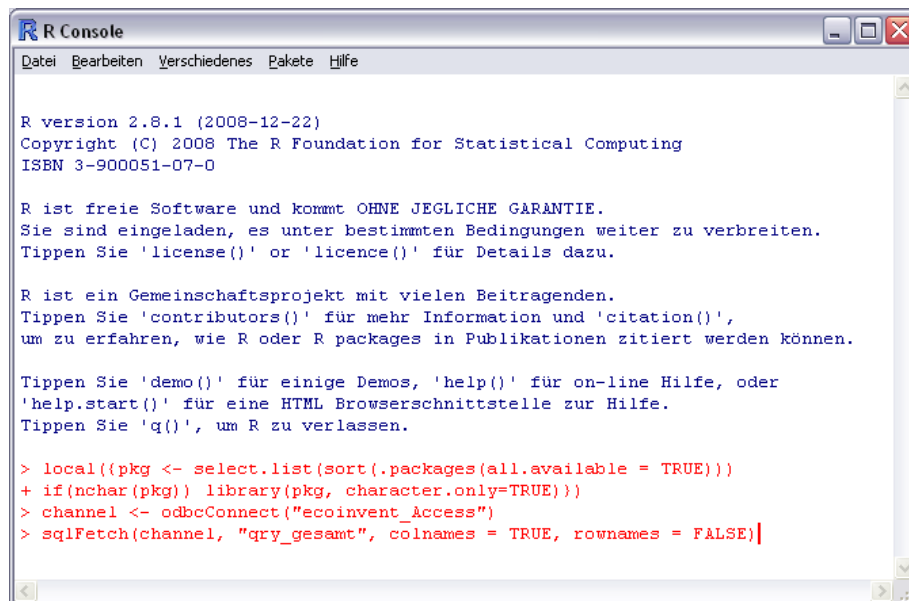


Figure 5: Start of a session in R, showing the connection to the Access database via ODBC

In order to be able to access all data in the database via ODBC, a query “qry_gesamt” in Access was created that basically “transports” all relevant data from Access to R, via the ODBC interface. For some tasks, it was useful to split this large data frame into one or several smaller data subsets.

Query 0: qry_gesamt

```
SELECT IO.*, Flows.*, Process.*, Source_Process.*, Sources.*, Persons.*,
Persons_1.*

FROM Sources INNER JOIN (((Persons AS Persons_1 INNER JOIN (Persons INNER
JOIN Process ON Persons.PersonID = Process.proofReadingValidator) ON
Persons_1.PersonID = Process.dataEntryByPerson) INNER JOIN (Flows INNER
JOIN IO ON Flows.FID = IO.FID) ON Process.PID = IO.PID) INNER JOIN
Source_Process ON Process.PID = Source_Process.PID) ON Sources.SID =
Source_Process.SID;
```

The following figure shows the very basic structure of the query in design view (which basically extracts all data from the Access database).

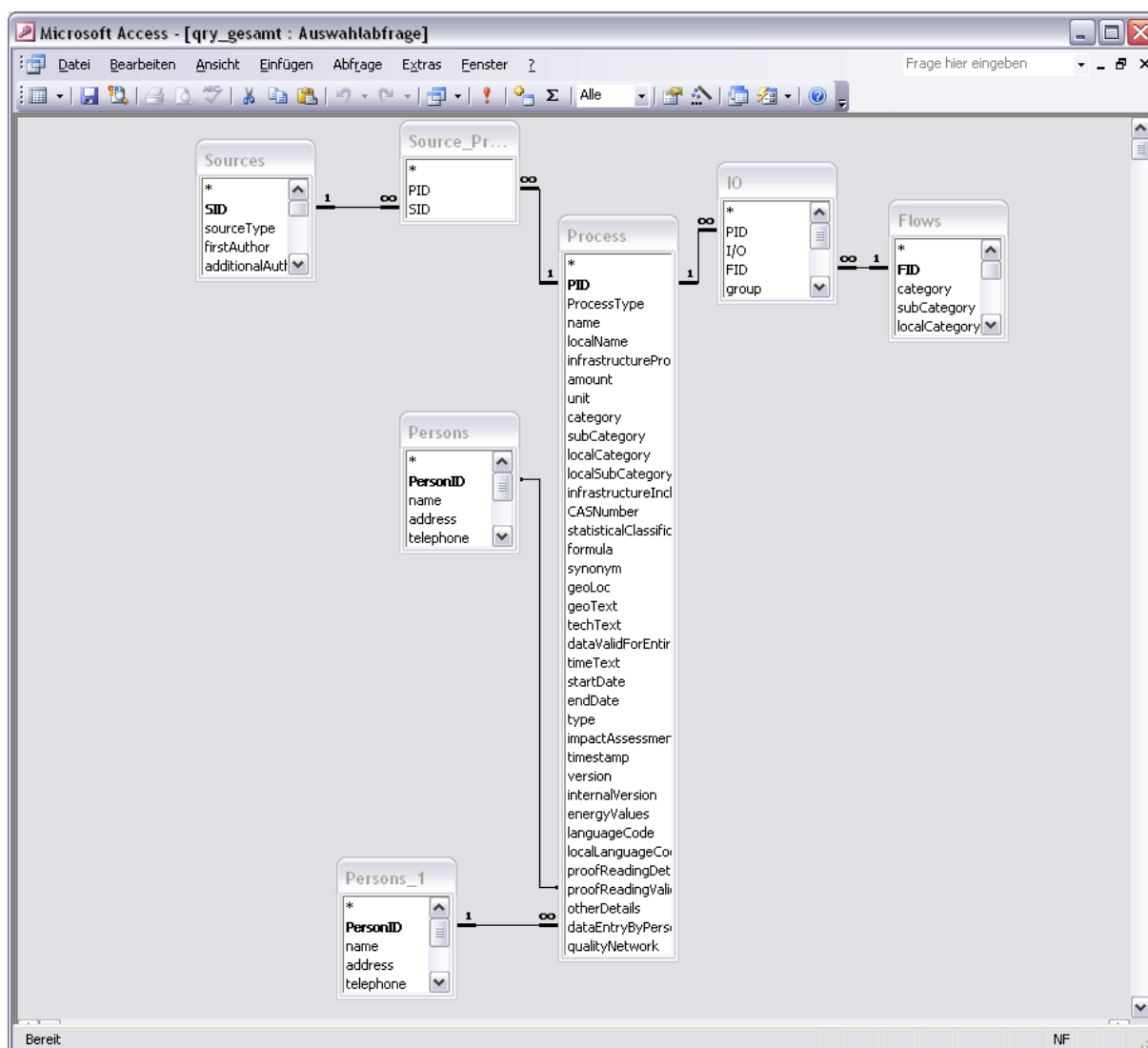


Figure 6: Query to extract all data from the Access database (qry_gesamt), design view

4 Analyses

In the following, a series of analysis are presented. In part, analyses were performed in the Access database, via queries, in part by using R, and in part also by pivot tables in Excel.

For performance reasons, often a “light version” of the Access database and of the overall ecoinvent data was used, with only unit processes. It is mentioned in each case where this limited data set was used. Motivation for this restriction was mere practicality. In consecutive queries, response time of Access is rather long with the full ecoinvent data. In another, better performing database environment, for example with stored procedures, these performance issues are likely to be solved.

The analyses described below are only a small subset of all possible analyses. Many roads lead to Rome. A first group of the analyses work with singular data sets (atomic checks, 4.1). A second group looks at “clusters” of data sets. Here, comparisons of one data set against other, similar or notably different data sets are performed; also, several datasets as a more or less unordered bunch are investigated in order to find possible relations and structures between data sets.

While the analyses look at the ecoinvent data in general, those analyses that are found to be of value should probably in future not be applied by everyone with access to ecoinvent data. Several “analysis roles” can be distinguished:

- database administration and maintenance (DB admin)
- database review (DB review)
- dataset review (dataset review)
- user (user)

Each of these roles has different rights and also different interests when accessing ecoinvent data.

4.1 Basic plausibility checks

In the following some examples of plausibility checks are reported. Ecoinvent error reports could be used to cover the field of potential errors in ecoinvent data more completely.

4.1.1 Mass balance of a process

Idea: Given that mass conservation holds for any of the modelled systems / processes in ecoinvent, and given further that no storage term is currently taken into account, the sum of all mass flows that enter a process should equal the sum of all mass flows that leave a process. The quotient of input to output should therefore equal one. Differences point either to methodological assumptions (e.g. water, biogenic C), to differing units between input and output (e.g. gas given in volume on the input side), or to flaws in data.

Analyses conducted: In the Access database, a small series of interlinked queries was created. One query (“qry_Process_Input_per_kg”) aggregates all inputs in kg, per process, another all outputs (“qry_Process_Output_per_kg”), and a third (“qry_Process_Input_per_Output_kg”) combines both by calculating the relation of input to output³. The queries are straightforward since in ecoinvent, all mass flows are already converted into kg, with exception of enriched uranium which has the unit “kg SWU”.

unit	name	FID
kg SWU	uranium, enriched 3.8%, at URENCO enrichment plant	5957
kg SWU	uranium, enriched 3.8%, at TENEX enrichment plant	5958
kg SWU	uranium, enriched 3.8%, at EURODIF enrichment plant	5959
kg SWU	uranium, enriched 3.8%, at USEC enrichment plant	5960
kg SWU	uranium, enriched 3.9%, at URENCO enrichment plant	5961
kg SWU	uranium, enriched 3.9%, at TENEX enrichment plant	5962
kg SWU	uranium, enriched 3.9%, at EURODIF enrichment plant	5963
kg SWU	uranium, enriched 3.9%, at USEC enrichment plant	5964
kg SWU	uranium, enriched 4.0%, at URENCO enrichment plant	5965
kg SWU	uranium, enriched 4.0%, at TENEX enrichment plant	5966
kg SWU	uranium, enriched 4.0%, at EURODIF enrichment plant	5967
kg SWU	uranium, enriched 4.0%, at USEC enrichment plant	5968
kg SWU	uranium, enriched 4.2%, at URENCO enrichment plant	5969
kg SWU	uranium, enriched 4.2%, at TENEX enrichment plant	5970
kg SWU	uranium, enriched 4.2%, at EURODIF enrichment plant	5971
kg SWU	uranium, enriched 4.2%, at USEC enrichment plant	5972
kg SWU	uranium, enriched 4.2% for pressure water reactor	5973
kg SWU	uranium, enriched 4.0% for pressure water reactor	5975
kg SWU	uranium, enriched 3.8% for pressure water reactor	5976
kg SWU	uranium, enriched 3.9% for pressure water reactor	5977
kg SWU	uranium enriched 3.8%, for boiling water reactor	5978
kg SWU	uranium, enriched 4.0% for boiling water reactor	5979
kg SWU	uranium, enriched 4.0% for boiling water reactor	5980
kg SWU	uranium, enriched 4.2%, centrifugal enrichment, for pressure	6007
kg SWU	uranium, enriched 3.0%, at USEC enrichment plant	11082
kg SWU	uranium, enriched 3.0%, at EURODIF enrichment plant	11083
kg SWU	uranium, enriched 3.0% at URENCO enrichment plant	11084
kg SWU	uranium, enriched 3.0% at TENEX enrichment plant	11085
kg SWU	uranium, enriched 3.8%, at CNNC centrifuge enrichment plan	11104
kg SWU	uranium, enriched 3.0%, at CNNC centrifuge enrichment plan	11105
kg SWU	uranium, enriched 3.8% for pressure water reactor	11145
kg SWU	uranium, enriched 3.0% for boiling water reactor	11146

³ As can be seen from the typical Access-aggregation names, all queries were conducted on a German machine („SummevonmeanValue“). These will look differently when transferred to an English machine.

Figure 7: All flows with the unit “kg SWU” in ecoinvent, left, and all units in ecoinvent, right

Query 1: qry_Process_Input_per_kg

```
SELECT Process.PID, Process.name, IO.[Input?], Sum(IO.meanValue) AS
SummevonmeanValue, Sum(IO.minValue) AS SummevonminValue, Sum(IO.maxValue)
AS SummevonmaxValue, Flows.unit

FROM Process INNER JOIN (Flows INNER JOIN IO ON Flows.FID = IO.FID) ON
Process.PID = IO.PID

GROUP BY Process.PID, Process.name, IO.[Input?], Flows.unit

HAVING (((IO.[Input?])="Ja") AND ((Flows.unit)="kg"));
```

Query 2: qry_Process_Output_per_kg

```
SELECT Process.PID, Process.name, IO.[Input?], Sum(IO.meanValue) AS
SummevonmeanValue, Sum(IO.minValue) AS SummevonminValue, Sum(IO.maxValue)
AS SummevonmaxValue, Flows.unit

FROM Process INNER JOIN (Flows INNER JOIN IO ON Flows.FID = IO.FID) ON
Process.PID = IO.PID

GROUP BY Process.PID, Process.name, IO.[Input?], Flows.unit

HAVING (((IO.[Input?])=No) AND ((Flows.unit)="kg"));
```

Query 3: qry_Process_Input_per_Output_kg

```
SELECT
qry_Process_Input_per_kg.SummevonmeanValue/qry_Process_Output_per_kg.Summev
onmeanValue AS Input_per_Output_mean, qry_Process_Input_per_kg.PID,
qry_Process_Input_per_kg.name, qry_Process_Input_per_kg.SummevonmeanValue
AS [Input], qry_Process_Output_per_kg.SummevonmeanValue AS [Output],
qry_Process_Input_per_kg.unit

FROM qry_Process_Input_per_kg INNER JOIN qry_Process_Output_per_kg ON
qry_Process_Input_per_kg.PID = qry_Process_Output_per_kg.PID

ORDER BY
qry_Process_Input_per_kg.SummevonmeanValue/qry_Process_Output_per_kg.Summev
onmeanValue;
```

Results: This analysis is performed for every process. Results can therefore not be completely displayed here. Since the last query 3 orders all processes according to their ratio of input to output, a look at the lowest ratios, and at middle ranges, might suffice.

qry_Process_Input_per_Output_kg				
Input_per_Output_mean	name	Input	Output	unit
-46,9015003628549	ethanol, 95% in H2O, from whey, at fermentation plant	8,33707	-0,177757	kg
-1,07437265697117	sugar beet seed IP, at regional storehouse	-2,31262305672775	2,15253342657418	kg
-1	resource correction, PbZn, indium, negative	-1	1	kg

....

qry_Process_Input_per_Output_kg				
Input_per_Output_mean	name	Input	Output	unit
5437,83204653709	gold, from combined gold-silver production, at refinery	762045,8246	140,137800888	kg
6154,44144019366	selective coating, copper sheet, black majic	1,0220127	0,000166061	kg
7748,57603837605	gold, at refinery	176663,55303	22,799486274	kg
7965,93792584457	door, inner, glass-wood, at plant	3,9166127	0,00049167	kg
8280,47464940669	medium density fibreboard, at plant	76,76	0,00927	kg

qry_Process_Input_per_Output_kg				
Input_per_Output_mean	name	Input	Output	unit
8809,7350317192	indium, at regional storage	10000,08	1,135117	kg
9281,06595887485	door, inner, wood, at plant	4,5632217	0,00049167	kg
9346,45115952214	laser machining, metal, with YAG-laser, 40W power	0,266	0,00002846	kg
10950,8346074828	pulverised lignite, at plant	0,1116	0,000010191004065	kg
10962,3410550261	lignite briquettes, at plant	0,1115	0,0000101711851	kg
11135,2490704502	gold, at refinery	149708,390486	13,444547988	kg
12459,0163934426	laser machining, metal, with YAG-laser, 30W power	0,266	0,00002135	kg
12730,1455458047	pc-Si wafer, at plant	277,8575388	0,021826737	kg
12730,1455458047	mc-Si wafer, at plant	277,8575388	0,021826737	kg

...

Input_per_Output_mean	PID	name	Input	Output	unit
413729,206874368	5971	uranium, enriched 4.2%, at EUROTIF enrichment plant	3601,43	0,0087048	kg
413731,504457311	5967	uranium, enriched 4.0%, at EUROTIF enrichment plant	3601,45	0,0087048	kg
413733,802040254	5963	uranium, enriched 3.9%, at EUROTIF enrichment plant	3601,47	0,0087048	kg
413734,950831725	5959	uranium, enriched 3.8%, at EUROTIF enrichment plant	3601,48	0,0087048	kg
413752,182703796	11083	uranium, enriched 3.0%, at EUROTIF enrichment plant	3601,63	0,0087048	kg
2810786,73784771	10137	laser machining, metal, with YAG-laser, 500W power	1000,359	0,0003559	kg
3520795,21463758	10135	laser machining, metal, with YAG-laser, 200W power	500,305	0,0001421	kg
4265923,24093817	10136	laser machining, metal, with YAG-laser, 330W power	1000,359	0,0002345	kg
4692754,982415	10134	laser machining, metal, with YAG-laser, 120W power	400,292	0,0000853	kg
5627198,65130655	10132	laser machining, metal, with YAG-laser, 50W power	200,272	0,00003559	kg
5875399,06103286	10133	laser machining, metal, with YAG-laser, 60W power	250,292	0,0000426	kg
6187119,44444444	1185	selective coating, copper sheet, black chrome	11,136815	0,0000018	kg
1802045761,90476	6866	photovoltaic laminate, CIS, at plant	37,842961	0,000000021	kg
11971308328,5714	6864	photovoltaic panel, CdTe, at plant	251,3974749	0,000000021	kg

Figure 8: Ratio of input to output, in kg, for all processes (system and unit processes) in ecoinvent, detail

Quite interestingly, for only 1127 of overall 6329 processes the ratio of input to output lies within a range of 0.9 to 1.1 (PID is an internal database ID; system processes have an ID above 100.000, Figure 9).

Interpretation, conclusion: Mass conservation is a valid assumption for LCA datasets. It is therefore recommended

- To present the mass balance results to all kind of people that need to deal with an ecoinvent dataset⁴;
- to try to eliminate reasons for differing input and output mass, on the methodological side;
- to provide all input in mass units, or at least provide a possibility to convert other units to mass units (possibly also in an analysis tool)
- to flag differences that result from methodological and modelling conventions, so that remaining differences point to flaws in modelling (such as: omitted flows, incorrect values for flows).

⁴ It is understood that some methodological and measurement agreements are needed to come to an overall process balance, for example concerning water, waste, and environmentally not relevant emissions as air.

Negative values that exist due to system expansion / avoided products calculation interfere with this indicator. Negative values should therefore either be avoided, or these processes should not be considered when calculating the ratio, or they should get a flag that indicates that mass input/output ratio is not fully applicable.



Input_per_Output_mean	PID	name	Input	Output	unit
0,900767958549334	100046	lime, algae, at regional storehouse	1,1058692581	1,2276960427	kg
0,901119628201448	100926	heat pump, brine-water, 10kW	366,94136440	407,20605003	kg
0,901122139932689	100927	heat pump 30kW	1100,8245608	1221,6152639	kg
0,902335487484646	100033	agricultural machinery, tillage, production	5,0260852863	5,5700849141	kg
0,902980088021818	100510	lightweight concrete block, expanded clay, at plant	1,3028388001	1,4428211844	kg
0,903	72	atrazine, at regional storehouse	0,903	1	kg
0,903	73	atrazine, at regional storehouse	0,903	1	kg
0,90324	490	portland cement, strength class Z 42.5, at plant	0,90324	1	kg
0,903398970063191	7247	carbon disulfide, at plant	0,930751	1,0302768	kg
0,903802478778136	100032	agricultural machinery, general, production	4,3967015665	4,8646708432	kg
0,903922760180814	100434	pentane, at plant	1,7702156019	1,9583704271	kg
0,904239568432964	101868	heat, at flat plate collector, one-family house, for combir	0,0077790826	0,0086029	kg
0,90525785542943	101994	sodium metasilicate pentahydrate, 58%, powder, at plant	2,1218737275	2,3439440095	kg
0,905394904785637	7246	sodium carbonate from ammonium chloride production,	0,948686	1,0478146	kg
0,905394904785637	7238	ammonium chloride, at plant	0,948686	1,0478146	kg
0,905601151680702	109244	concrete roof tile, at plant	1,1005883224	1,2153124147	kg
0,905691143371202	100407	fatty alcohol, from palm kernel oil, at plant	8,9208512965	9,8497720352	kg
0,905906205952799	108266	milling, cast iron, small parts	4,5012321643	4,9687618152	kg
0,906160283625089	101085	copper, primary, at refinery	6,9458432733	7,6651376129	kg
0,906895046537763	100199	palm fruit bunches, at farm	1,2516843506	1,3801865557	kg
0,907215626577496	101637	crude oil, at production onshore	1,0320991959	1,1376558843	kg
0,907222270683791	110897	energy reduction, ventilation system, 6 x 120 m3/h, steel	0,0256579363	0,0282818634	kg
0,907224000407195	101023	electricity, lignite, at power plant	1,5946198307	1,7576914080	kg
0,907246773153456	110891	ventilation of dwellings, decentralized, 6 x 120 m3/h, steel	0,7995708911	0,8813157729	kg
0,907281268466007	101038	lignite, burned in power plant	0,1022203179	0,1126666244	kg
0,907708639679001	101656	pipeline, crude oil, offshore	1581421,7306	1742212,9321	kg
0,909066586826069	100135	grass seed IP, at regional storehouse	2,5792850855	2,8372895043	kg
0,909081239675575	7224	lithium, at plant	1,1111	1,222223	kg
0,909082727264545	7249	chlorine, gaseous, lithium chloride electrolysis, at plant	1,1111	1,222221	kg
0,909464771016583	107074	switch, toggle type, at plant	15,234059999	16,750577356	kg
0,90994090966466	106549	ethyl tert-butyl ether, from bioethanol, at plant	3,2493679165	3,5709658528	kg
0,910002807132212	100542	light mortar, at plant	1,3599103922	1,4944024145	kg
0,910055199597409	434	pentane, at plant	1,7779891972	1,9537157724	kg
0,910335303528643	101778	inverter, 2500W, at plant	187,20774135	205,64701888	kg
0,911852022044884	111094	hard coal supply mix	1,4245615482	1,5622727304	kg
0,91190548068076	110870	control and wiring, central unit, at plant	36,016581076	39,495958560	kg
0,912354487203511	101859	expansion vessel 25l, at plant	16,543830017	18,133116294	kg
0,91241	491	portland cement, strength class Z 52.5, at plant	0,91241	1	kg
0,912464577951268	6223	ethanol, 95% in H2O, from grass, at fermentation plant	1,603547	1,75738	kg
0,914948311306222	105898	fatty alcohol sulfate, palm oil, at plant	3,9473794886	4,3143196614	kg
0,915441566990497	105937	electricity, nuclear, at power plant boiling water reactor	0,0096698519	0,0105630466	kg
0,915475117684129	106551	glycerine, from rape oil, at esterification plant	7,9330393558	8,6654887747	kg
0,915763046511817	5694	heavy fuel oil, burned in power plant	0,072612609	0,0792919187	kg
0,916000366985447	109243	quarry tile, at plant	1,1293659385	1,2329317533	kg
0,917197814677341	100035	slurry tanker, production	4,0396007412	4,4042851788	kg
0,91719784130062	106221	electricity, at cogen with biogas engine, allocation exer	0,8435839838	0,9197404811	kg
0,917214666839404	6233	1,1-dimethylcyclopentane, from naphtha, at plant	1,0287	1,1215477	kg
0,91722006442742	6236	fraction 7, from naphtha, at plant	1,0287	1,1215411	kg
0,917293605818038	6230	2,3-dimethylbutan, from naphtha, at plant	1,0288	1,1215602	kg
0,917294423691589	6232	methylcyclopentane, from naphtha, at plant	1,0288	1,1215592	kg
0,917294587266475	6234	methylcyclohexane, from naphtha, at plant	1,0288	1,121559	kg
0,917294587266475	6229	heptane, at plant	1,0288	1,121559	kg
0,917295486929386	6235	fraction 1, from naphtha, at plant	1,0288	1,1215579	kg
0,917299985270416	6237	fraction 8, from naphtha, at plant	1,0288	1,1215524	kg
0,917379004761918	6231	2-methylpentane, from naphtha, at plant	1,0289	1,1215648	kg
0,917379986297701	6120	hexane, at plant	1,0289	1,1215636	kg
0,917656412437245	100964	electricity, hydropower, at pumped storage power plant	0,1535057537	0,1672802060	kg

Figure 9: Processes (system and unit processes) where the ratio of input to output lies between 0.9 and 1.1, ordered by ratio, detail (top of list)

4.1.2 Input and output pattern

Idea: Similar processes are likely to have a similar “pattern” of input and output flows. An incineration process without a specified output of CO₂ is suspicious, as is an incineration process that has as output (e.g.) a compost plant. For system processes, the life cycle will often blur the vision to an existing input and output pattern. Therefore, this analysis is most interesting on the level of unit processes.

Analyses conducted: It is assumed that all processes within one subcategory in ecoinvent can be seen as similar. One process needs to be selected to start the analysis (hence the analysis is meant to be conducted per process). In the example given below, this is the process “rice, at farm”. For this selected process, the issue is approached from two sides. For one, it is investigated which flows occur in this process but are not present in other processes of the same subcategory; and second, it is checked which flows occur in the subcategory and lack a representation at the selected process. In the Access database, another small series of interlinked queries was created. Query 4 (qry_Proc_Flow_cat_subCat_U) simply compiles all flows, categories and subcategories, and query 5 (qry_Flow_cat_subCat_U) selects all flows for processes in the subcategory that are not the selected process.

Query 4: qry_Proc_Flow_cat_subCat_U⁵

```
SELECT qry_ProcessIO.Process.name, qry_ProcessIO.Flows.name,
qry_ProcessIO.Process.category, qry_ProcessIO.Process.subCategory,
qry_ProcessIO.ProcessType
FROM qry_ProcessIO
GROUP BY qry_ProcessIO.Process.name, qry_ProcessIO.Flows.name,
qry_ProcessIO.Process.category, qry_ProcessIO.Process.subCategory,
qry_ProcessIO.ProcessType
HAVING (((qry_ProcessIO.Process.name)="rice, at farm") AND
((qry_ProcessIO.ProcessType)="UnitProcess"));
```

Query 5: qry_Flow_cat_subCat_U

```
SELECT qry_ProcessIO.Flows.name, qry_Proc_Flow_cat_subCat_U.category,
qry_Proc_Flow_cat_subCat_U.subCategory, qry_ProcessIO.ProcessType
FROM qry_Proc_Flow_cat_subCat_U INNER JOIN qry_ProcessIO ON
(qry_Proc_Flow_cat_subCat_U.subCategory =
qry_ProcessIO.Process.subCategory) AND (qry_Proc_Flow_cat_subCat_U.category
= qry_ProcessIO.Process.category)
WHERE
(((qry_ProcessIO.Process.name)<>[qry_Proc_Flow_cat_subCat_U].[Process.name]
))
GROUP BY qry_ProcessIO.Flows.name, qry_Proc_Flow_cat_subCat_U.category,
qry_Proc_Flow_cat_subCat_U.subCategory, qry_ProcessIO.ProcessType
HAVING (((qry_ProcessIO.ProcessType)="UnitProcess"));
```

⁵ “rice, at farm”, is only one possible process that can be selected

Process.name	Flows.name	category	subCategory
rice, at farm	[sulfonyl]urea-compounds, at re	agricultural production	plant production
rice, at farm	[thio]carbamate-compounds, at	agricultural production	plant production
rice, at farm	2,4-D	agricultural production	plant production
rice, at farm	acetamide-anillide-compounds,	agricultural production	plant production
rice, at farm	Ammonia	agricultural production	plant production
rice, at farm	ammonia, liquid, at regional stc	agricultural production	plant production
rice, at farm	ammonium nitrate, as N, at reg	agricultural production	plant production
rice, at farm	application of plant protection p	agricultural production	plant production
rice, at farm	Azoxystrobin	agricultural production	plant production

Figure 10: Result of query 5: Flows, category and subcategory for the example unit process “rice, at plant”

Based on these two queries, the following queries 6 and 7 list all flows that occur only for the selected process (query 6, qry_Flows_in_Process_not_in_CatSubCat_U), and all flows that occur only in other processes of the same category / subcategory, respectively (query 7, qry_Flows_inCatSubCat_not_in_Process_U).

Query 6: qry_Flows_in_Process_not_in_CatSubCat_U

```
SELECT qry_Proc_Flow_cat_subCat_U.Flows.name, qry_Flow_cat_subCat_U.name
FROM qry_Flow_cat_subCat_U RIGHT JOIN qry_Proc_Flow_cat_subCat_U ON
qry_Flow_cat_subCat_U.name = qry_Proc_Flow_cat_subCat_U.Flows.name
WHERE (((qry_Flow_cat_subCat_U.name) Is Null));
```

qry_Proc_Flow_cat_subCat_U.qry_ProcessIO.Flows.name	qry_Flow_cat_subCat_U.name
Bensulfuron methyl ester	
Halosulfuron-methyl	
Molinate	
Propanil	
Quinclorac	
rice seed, at regional storehouse	
rice, at farm	
Thiobencarb	
Triclopyr	

Figure 11: Result of query 6: Exclusive flows for the example unit process “rice, at plant”

While rice seed and rice, at farm are evident as specific flows for the process, the other flows seem to be specific pesticides for rice farming not used in other agricultural plant production processes in ecoinvent.

Query 7: qry_Flows_inCatSubCat_not_in_Process_U

```
SELECT qry_Proc_Flow_cat_subCat_U.Flows.name, qry_Flow_cat_subCat_U.name
FROM qry_Flow_cat_subCat_U LEFT JOIN qry_Proc_Flow_cat_subCat_U ON
qry_Flow_cat_subCat_U.name = qry_Proc_Flow_cat_subCat_U.Flows.name
WHERE (((qry_Proc_Flow_cat_subCat_U.Flows.name) Is Null));
```

Results for query 7 are presented below; overall 435 flows for the example are present only in other processes of agricultural production / plant production. Some of these are, again, evident (potato planting), others are again different kind of agricultural chemicals (Oxamyl), and others may simply have a different name for the rice process (planting).

qry_Proc_Flow_cat_subCat_U.qry_ProcessIO.Flows.name	qry_Flow_cat_subCat_U.name
	Nicosulfuron
	nitrile-compounds, at regional storehouse
	Norflurazon
	Occupation, arable, non-irrigated
	Occupation, forest, intensive, short rotation
	Occupation, pasture and meadow
	Occupation, pasture and meadow
	Occupation, permanent crop, fruit
	Oils, unspecified
	operation, van < 3,5t
	Orbencarb
	Oxamyl
	Oxydemeton-methyl
	Oxyfluorfen
	packaging box production unit
	palm fruit bunches, at farm
	parathion, at regional storehouse
	Particulates, > 10 um
	pea seed IP, at regional storehouse
	pea seed organic, at regional storehouse
	Permethrin
	Phenmedipham
	Phorate
	Phosmet
	phosphate rock, as P2O5, beneficiated
	phtalamide-compounds, at regional storehouse
	Picloram
	Picoxystrobin
	Piperonyl butoxide
	Pirimicarb
	planting
	potassium nitrate, as K2O, at regional storehouse
	potassium nitrate, as N, at regional storehouse
	potassium sulphate, as K2O, at regional storehouse
	potato grading
	potato haulm cutting
	potato planting
	potato seed IP, at regional storehouse
	potato seed organic, at regional storehouse
	potato starch at plant

Figure 12: Result of query 7 (detail): Flows not in the example unit process “rice, at plant” but in other unit processes of the same category / subcategory

It is to be noted that this analysis could also be performed by investigating an OLAP⁶ or matrix representation of ecoinvent. The example below is an OLAP cube of ecoinvent, created in Excel.

With all categories, subcategories, and process on e.g. the left side...

⁶ OLAP: Online analytical processing, a method to analyse data drawing typically from a multidimensional representation of a data stock (“OLAP cube”).

Microsoft Excel - olap_ei_versuch.xls

File | Bearbeiten | Ansicht | Einfügen | Format | Extras | Daten | Fenster | | Adgabe PDF

100% Arial

Frage hier eingeben

G20	A	B	C	
1				
2				
3	Summe von meanValue			UO
4				
5	gry_gesamt.Process.category	gry_gesamt.Process.subCategory	gry_gesamt.Process.name	gry_gesamt.Process.value
6	agricultural means of production	buildings	compost plant, open	
7			dried roughage store, air dried, solar	
8			dried roughage store, air dried, solar, operation	
9			dried roughage store, cold-air dried, conventional	
10			dried roughage store, cold-air dried, conventional, operation	
11			dried roughage store, non ventilated	
12			dried roughage store, non ventilated, operation	
13			dung slab	
14			housing system with fully-slatted floor, pig	
15			housing system with fully-slatted floor, pig, operation	
16			label housing system, pig	
17			label housing system, pig, operation	
18			loose housing system, cattle	
19			loose housing system, cattle, operation	
20			milking parlour	
21			shed	
22			slurry store and processing	
23			slurry store and processing, operation	
24			tied housing system, cattle	
25			tied housing system, cattle, operation	
26			tower silo, plastic	
27		feed	barley IP, at feed mill	
28			barley organic, at feed mill	
29			fava beans IP, at feed mill	
30			grain maize IP, at feed mill	
31			grain maize organic, at feed mill	
32			protein peas IP, at feed mill	
33			rye IP, at feed mill	
34			rye organic, at feed mill	
35			wheat IP, at feed mill	
36			wheat organic, at feed mill	
37		machinery	agricultural machinery, general, production	
38			agricultural machinery, tillage, production	
39			harvester, production	
40			slurry tanker, production	
41			tractor, production	
42			trailer, production	
43		mineral fertiliser	ammonium nitrate phosphate, as N, at regional storehouse	
44			ammonium nitrate phosphate, as P2O5, at regional storehouse	
45			ammonium nitrate, as N, at regional storehouse	
46			ammonium sulphate, as N, at regional storehouse	
47			calcium ammonium nitrate, as N, at regional storehouse	
48			calcium nitrate, as N, at regional storehouse	
49			diammonium phosphate, as N, at regional storehouse	
50			diammonium phosphate, as P2O5, at regional storehouse	
51			lime, algae, at regional storehouse	
52			lime, from carbonation, at regional storehouse	
53			monoammonium phosphate, as N, at regional storehouse	
54			monoammonium phosphate, as P2O5, at regional storehouse	
55			potassium chloride, as K2O, at regional storehouse	
56			potassium nitrate, as K2O, at regional storehouse	
57			potassium nitrate, as N, at regional storehouse	
58			potassium sulphate, as K2O, at regional storehouse	
59			potassium sulphate, as P2O5, at regional storehouse	

W | Tabelle1 | Tabelle2 | Tabelle3 |

Excel2EcoSpold | Excel2EcoSpold Impact | EcoSpold Access | EcoSpold Options | EcoSpoldExcel

... and all flows, divided by input and output, at the top...

The screenshot shows a Microsoft Excel window titled 'Microsoft Excel - olap_ei_versuch.xls'. The pivot table is structured as follows:

gry_gesamt.Process.name	gry_gesamt.Process.subCategory	gry_gesamt.Process.name	gry_gesamt.Process.value
compost plant, open			
dried roughage store, air dried, solar			
dried roughage store, air dried, solar, operation			
dried roughage store, cold-air dried, conventional			

The resulting table gives an overview of which flow occurs where (even including the quantitative amount).

However in Excel, the whole cube gets difficult to manage (in Excel 2003, at least). The next figure shows a larger view on the OLAP cube in Excel (zoomed out to 10% - the tiny dots indicate quantitative entries)

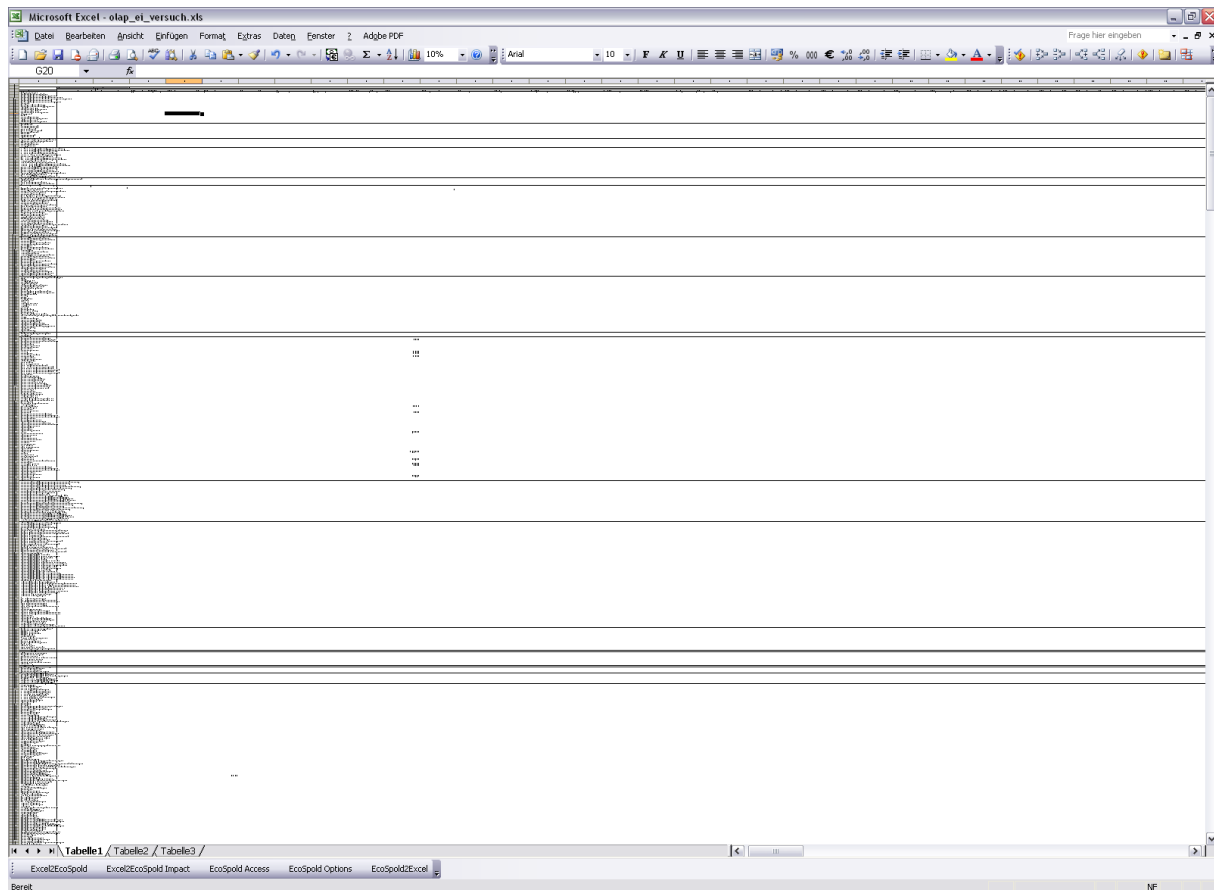


Figure 13: OLAP cube in Excel with link to ecoinvent, zoomed out to 10%. Processes, categories and subcategories are on the left, flows and input / output on top. The tiny dots in the table represent quantitative entries

Results: Results for the example process are given above, in the explanation of the developed queries.

Interpretation, conclusion: Results obviously depend on the criteria that determine “similar” processes. The present category / subcategory system acts only as a makeshift to this end. The whole analysis is iterative as it builds on knowledge about similarity of processes that is obtained from analysis, and even to some extent defined by all candidates for similarity, including new processes.

In longer terms, it is recommended to develop groups of similar processes, maybe under different perspectives, and use these in the analysis. This will render the analysis also independent from the currently selected category system. The further analysis of data sets will investigate if there can be identified, and characterised, suitable groups of processes. On the other side, it seems of value to also determine groups of flows, and to analyse whether a process is missing a certain group of flows, or has an additional group. This will turn the analysis somewhat independent from naming, and also provide a more stable result that requires less human interpretation and modification before being used at value. For example, all different seeds for planting can be considered, and it is then rather suspicious if an agricultural plant production process does not have any input of seed (while a rice plant will not have beans as input..).

4.1.3 Negative flows

Idea: Negative mass or energy flows point to either modelling specifics (negative system expansion may lead to negative flow values), or to errors.

Analyses conducted: For a process, it should simply be checked whether a flow is negative or not, and somebody inspecting or entering the data set should see a warning or a remark to why the amount is negative. A simple query was conducted to list all negative flows, for system and unit processes, that occur in ecoinvent. Process documentation, ideally the documentation that is distributed together with the data set, should explain negative values.

Query 8: qry_MeanBelowZero

```
SELECT Process.category, Process.subCategory, Process.ProcessType,
Process.name, Flows.category, Flows.subCategory, Flows.name, Flows.unit,
Flows.generalComment, IO.meanValue, IO.[Input?]

FROM Process INNER JOIN (Flows INNER JOIN IO ON Flows.FID = IO.FID) ON
Process.PID = IO.PID

WHERE (((IO.meanValue)<0))

ORDER BY Process.category, Process.subCategory, Process.ProcessType,
Process.name;
```

Results: Roughly 11,000 flows have negative mean values in ecoinvent (of a total of 4,671,855 input and output flows!). Many of them do not contain a general comment for explanation in the data set⁷. Negative values occur both on the input and on the output side, with a negative value on the input side being more frequent. A negative *output* flow is usually commented.

Process.category	Process.subCategory	ProcessType	Process.name	Flows.category	Flows.subCategory	Flows.name	unit	generalComment	meanValue	Input?
agricultural means of work processes	SystemProcess	haying, by rotary tedder	haying, by rotary tedder	resource	in ground	Cerium, 24% in bastnas	kg	null	-1,1273E-17	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	haying, by rotary tedder	haying, by rotary tedder	resource	in ground	Neodymium, 4% in bast	kg	null	-5,8384E-18	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Praseodymium, 0.42% i	kg	null	-7,2785E-19	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	water	ground-	Thorium-232	kBq	null	-1,607E-19	<input type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Gadolinium, 0.15% in b	kg	null	-2,2997E-19	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Samarium, 0.3% in bast	kg	null	-2,9844E-19	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Neodymium, 4% in bast	kg	null	-1,02E-17	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Lanthanum, 7.2% in bast	kg	null	-5,316E-18	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	hoeing	hoeing	resource	in ground	Cerium, 24% in bastnas	kg	null	-1,9913E-17	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	resource	in ground	Neodymium, 4% in bast	kg	null	-2,1206E-19	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	resource	in ground	Cerium, 24% in bastnas	kg	null	-1,0925E-16	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	water	ground-	Thorium-232	kBq	null	-1,0035E-18	<input type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	water	ground-	Suspended solids, unsp	kg	null	-3,4776E-19	<input type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	resource	in ground	Samarium, 0.3% in bast	kg	null	-2,5328E-18	<input checked="" type="checkbox"/>
agricultural means of work processes	SystemProcess	irrigating	irrigating	resource	in ground	Praseodymium, 0.42% i	kg	null	-3,4577E-18	<input checked="" type="checkbox"/>

Figure 14: Results of the query “qry_MeanBelowZero”, for unit and system processes (detail)

Most of the negative flows occur in system processes. Only 444 flows are negative for unit processes. This can be seen when query 8 is filtered to display only unit processes.

Process.category	Process.subCategory	ProcessType	Process.name	Flows.category	Flows.subCategory	Flows.name	unit	generalComment	meanValue	Input?
agricultural means of seed	UnitProcess	ammonium nitrate, as f	ammonium nitrate, as f	high population	Heat, waste	Heat, waste	MJ	(2,3,2,3,1,5)	-2,4	<input type="checkbox"/>
agricultural means of seed	UnitProcess	calcium ammonium nitr	calcium ammonium nitr	high population	Heat, waste	Heat, waste	MJ	(2,3,2,3,1,5)	-2,4	<input type="checkbox"/>
agricultural means of seed	UnitProcess	potassium nitrate, as N	potassium nitrate, as N	high population	Heat, waste	Heat, waste	MJ	(2,3,2,3,1,5)	-1,2643	<input type="checkbox"/>
agricultural means of seed	UnitProcess	potassium nitrate, as N	potassium nitrate, as N	high population	Heat, waste	Heat, waste	MJ	(2,3,2,3,1,5)	-6,3123	<input type="checkbox"/>
agricultural means of seed	UnitProcess	clover seed IP, at farm	clover seed IP, at farm	agricultural	Chromium	Chromium	kg	(2,2,1,1,1,na)	-5,3912E-06	<input type="checkbox"/>
agricultural means of seed	UnitProcess	clover seed IP, at farm	clover seed IP, at farm	agricultural	Copper	Copper	kg	(2,2,1,1,1,na)	-0,000039198	<input type="checkbox"/>
agricultural means of seed	UnitProcess	clover seed IP, at farm	clover seed IP, at farm	agricultural	Mercury	Mercury	kg	(2,2,1,1,1,na)	-3,5565E-08	<input type="checkbox"/>
agricultural means of seed	UnitProcess	clover seed IP, at farm	clover seed IP, at farm	agricultural	Nickel	Nickel	kg	(2,2,1,1,1,na)	-1,9587E-06	<input type="checkbox"/>
agricultural means of seed	UnitProcess	clover seed IP, at farm	clover seed IP, at farm	agricultural	Zinc	Zinc	kg	(2,2,1,1,1,na)	-0,000051224	<input type="checkbox"/>
agricultural means of seed	UnitProcess	grass seed IP, at farm	grass seed IP, at farm	agricultural	Chromium	Chromium	kg	(2,2,1,1,1,na)	-9,681E-07	<input type="checkbox"/>
agricultural means of seed	UnitProcess	grass seed IP, at farm	grass seed IP, at farm	agricultural	Copper	Copper	kg	(2,2,1,1,1,na)	-0,000016447	<input type="checkbox"/>
agricultural means of seed	UnitProcess	grass seed IP, at farm	grass seed IP, at farm	agricultural	Zinc	Zinc	kg	(2,2,1,1,1,na)	-0,000022536	<input type="checkbox"/>
agricultural means of seed	UnitProcess	maize seed IP, at farm	maize seed IP, at farm	agricultural	Copper	Copper	kg	(2,2,1,1,1,na)	-3,1582E-07	<input type="checkbox"/>
agricultural means of seed	UnitProcess	potato seed IP, at farm	potato seed IP, at farm	agricultural	Chromium	Chromium	kg	(2,2,1,1,1,na)	-2,8061E-07	<input type="checkbox"/>
agricultural means of seed	UnitProcess	potato seed organic, at	potato seed organic, at	agricultural	Chromium	Chromium	kg	(2,2,1,1,1,na)	-2,0362E-07	<input type="checkbox"/>

Figure 15: Results of the query “qry_MeanBelowZero”, only for unit processes (detail)

⁷ Explanations might be provided in the background reports; this was not checked for the present analysis.

Most of the negative values are small (1E-18, see Figure 14). Changing the order in query 8 (from categories to mean values) reveals that the absolute figures for some of the negative values are rather high; further, it indicates that for these negative values, non-mass units are more frequent (Figure 16)⁸.

Process.category	Process.subCate	ProcessType	Process.name	Flows.category	Flows.subCate	Flows.name	unit	generalComment	meanValue	Input?
wooden materials	extraction	UnitProcess	roundwood, azobe (SFM), under bark,	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-26487	✓
wooden materials	extraction	UnitProcess	roundwood, meranti (SFM), under bark,	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-18332	✓
wooden materials	extraction	UnitProcess	roundwood, paraná pine (SFM), under bark,	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-17168	✓
wooden materials	extraction	UnitProcess	roundwood, eucalyptus ssp. (SFM), u	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-4866	✓
wooden materials	processing	UnitProcess	roundwood, meranti (SFM), debarked,	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-1850	✓
wooden materials	extraction	UnitProcess	roundwood, azobe (SFM), debarked,	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-1312,2	✓
agricultural means of	seed	UnitProcess	rape seed organic, at regional storeho	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-22,186	✓
agricultural means of	seed	UnitProcess	rape seed IP, at regional storehouse	resource	biotic	Energy, gross calorific v	MJ	(2,2,1,1,1,na)	-18,668	✓
agricultural means of	seed	UnitProcess	sugar beet seed IP, at regional storeh	resource	in air	Carbon dioxide, in air	kg	(2,2,1,1,1,na)	-11,583	✓
construction process: machinery		UnitProcess	power sawing, with catalytic converter	transport system	ship	transport, transoceanic	tkm	(3,1,1,1,1,na)	-6,53	✓
construction process: machinery		UnitProcess	power sawing, without catalytic conver	transport system	ship	transport, transoceanic	tkm	(3,1,1,1,1,na)	-6,53	✓
agricultural means of	mineral fertiliser	UnitProcess	potassium nitrate, as N, at regional st	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-6,3123	✓
natural gas	power plants	UnitProcess	electricity, natural gas, at turbine, 10h	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-3,6	✓
wooden materials	processing	UnitProcess	sawn timber, paraná pine (SFM), kiln	electricity	production mix	electricity, low voltage,	kWh	null	-3,0664	✓
chemicals	inorganics	UnitProcess	urea ammonium nitrate, as N, at regio	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-2,58	✓
agricultural means of	mineral fertiliser	UnitProcess	ammonium nitrate, as N, at regional s	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-2,4	✓
agricultural means of	mineral fertiliser	UnitProcess	calcium ammonium nitrate, as N, at ri	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-2,4	✓
agricultural means of	seed	UnitProcess	rape seed organic, at regional storeho	resource	in air	Carbon dioxide, in air	kg	(2,2,1,1,1,na)	-2,0163	✓
agricultural means of	seed	UnitProcess	rape seed IP, at regional storehouse	resource	in air	Carbon dioxide, in air	kg	(2,2,1,1,1,na)	-2,0163	✓
wooden materials	extraction	UnitProcess	plywood, outdoor use, at plant	wood energy	fuels	wood chips, hardwood,	m3	null	-1,9297	✓
metals	extraction	UnitProcess	molybdenum, at regional storage	metals	extraction	zinc concentrate, at ber	kg	exact amount inv	-1,9028	✓
wooden materials	processing	UnitProcess	sawn timber, paraná pine (SFM), kiln	transport system	road	transport, lorry 3.5-20t,	t.km	(4,2,1,1,3,5); Vali	-1,6472	✓
insulation materials	production	UnitProcess	foam glass, at regional storage	electricity	supply mix	electricity, medium volta	kWh	(2,4,1,3,1,5); acc	-1,4969	✓
insulation materials	production	UnitProcess	foam glass, at regional storage	electricity	supply mix	electricity, medium volta	kWh	(2,4,1,3,1,5); acc	-1,4969	✓
wooden materials	processing	UnitProcess	sawn timber, paraná pine (SFM), kiln	wooden materials	processing	paraná pine, allocation c	m3	(3,4,1,3,1,5)	-1,3232	✓
wooden materials	extraction	UnitProcess	plywood, outdoor use, at plant	wooden materials	extraction	hardwood, allocation cor	m3	null	-1,32	✓
wooden materials	extraction	UnitProcess	plywood, indoor use, at plant	wooden materials	extraction	hardwood, allocation cor	m3	null	-1,32	✓
agricultural means of	mineral fertiliser	UnitProcess	potassium nitrate, as K2O, at regiona	air	high population	Heat, waste	MJ	(2,3,2,3,1,5)	-1,2643	✓
wooden materials	extraction	UnitProcess	three layered laminated board, at plan	wooden materials	extraction	softwood, allocation cor	m3	calculated correc	-1,23	✓
biomass	fuels	UnitProcess	ethanol, 95% in H2O, from whey, at fe	air	high population	Carbon dioxide, biogenic	kg	(2,3,1,1,1,5)	-1,1929	✓
wooden materials	extraction	UnitProcess	roundwood, paraná pine (SFM), under	resource	biotic	Wood, hard, standing	m3	null	-1,15	✓
wooden materials	extraction	UnitProcess	roundwood, meranti (SFM), under bark,	resource	biotic	Wood, hard, standing	m3	null	-1,11	✓
wooden materials	extraction	UnitProcess	roundwood, azobe (SFM), under bark,	resource	biotic	Wood, hard, standing	m3	null	-1,09	✓
metals	extraction	UnitProcess	resource correction, PbZn, lead, nega	resource	in ground	Lead, 5.0% in sulfide, P	kg	Uncertainty for LC	-1	✓
metals	extraction	UnitProcess	resource correction, PbZn, zinc, nega	resource	in ground	Zinc, 9.0% in sulfide, Zn	kg	Uncertainty for LC	-1	✓

Figure 16: Results of the query “qry_MeanBelowZero”, only for unit processes, ordered by meanValue (detail)

Interpretation, conclusion: Negative values can occur in flow data, depending on modelling conventions; they remain however in each case either modelling constructs, or errors, and therefore should be inspected in data quality checks, and explained.

A more refined analysis could take into account the type of each flow, since for some flows, negative values are more likely than for others.

4.1.4 Carbon dioxide emissions per fuel consumption

As an example for single substance and elementary tracing, kg carbon dioxide emissions per MJ of fuel consumed are calculated, for power plants.

Idea: Taking the fossil carbon dioxide emissions in kg per MJ of fuel consumed should provide a rather stable ratio, depending only, and only to some degree, on the specific fuel that is used. This analysis is related to the identification of outliers in single values; here, outliers in the relation of CO₂ to MJ consumed are identified⁹.

Analyses conducted: Similar to the previous analyses, a series of Access-queries was designed. In this case, the queries are rather straightforward. One collects all entries for power plant in units of MJ, on the input side (query 9); one that collects emissions of fossil carbon

⁸ Overall, there are 340 negative flows for unit processes with unit „kg“, so the majority of flows is in mass-units. For system processes, even more than 10,000 flows of the overall 11,000 are in mass units.

⁹ One could argue whether this analysis is a plausibility check or rather belongs to the field of exploratory data analysis which is detailed later in the text (section 4.2.1) – it is put here since it uses SQL queries, but its purpose is also to give insights into data and even to provide ideas for different explorative data analyses.

dioxide (query 10), and finally one that combines both queries 9 and 10 and calculates the ratio of CO₂ emissions per MJ input (query 11).

Query 9: qry_Powerplants_MJ

```
SELECT qry_ProcessIO_U.Process.category,
qry_ProcessIO_U.Process.subCategory, qry_ProcessIO_U.[Input?],
Sum(qry_ProcessIO_U.meanValue) AS SummevonmeanValue,
qry_ProcessIO_U.Flows.unit, qry_ProcessIO_U.Process.PID,
qry_ProcessIO_U.Process.name
FROM qry_ProcessIO_U
GROUP BY qry_ProcessIO_U.Process.category,
qry_ProcessIO_U.Process.subCategory, qry_ProcessIO_U.[Input?],
qry_ProcessIO_U.Flows.unit, qry_ProcessIO_U.Process.PID,
qry_ProcessIO_U.Process.name
HAVING (((qry_ProcessIO_U.Process.subCategory)="power plants") AND
((qry_ProcessIO_U.[Input?])=Yes) AND ((qry_ProcessIO_U.Flows.unit)="MJ"));
```

Query 10: qry_Powerplants_CO2

```
SELECT qry_ProcessIO_U.Process.category,
qry_ProcessIO_U.Process.subCategory, qry_ProcessIO_U.Flows.name,
qry_ProcessIO_U.[Input?], qry_ProcessIO_U.meanValue,
qry_ProcessIO_U.Flows.unit, qry_ProcessIO_U.Process.PID,
qry_ProcessIO_U.Process.name
FROM qry_ProcessIO_U
WHERE (((qry_ProcessIO_U.Process.subCategory)="power plants") AND
((qry_ProcessIO_U.Flows.name)="Carbon dioxide, fossil") AND
((qry_ProcessIO_U.[Input?])=No));
```

Query 11: qry_Powerplants_CO2perMJ

```
SELECT qry_ProcessIO_U.Process.category,
qry_ProcessIO_U.Process.subCategory, qry_ProcessIO_U.Flows.name,
qry_ProcessIO_U.[Input?], qry_ProcessIO_U.meanValue,
qry_ProcessIO_U.Flows.unit, qry_ProcessIO_U.Process.PID,
qry_ProcessIO_U.Process.name
FROM qry_ProcessIO_U
WHERE (((qry_ProcessIO_U.Process.subCategory)="power plants") AND
((qry_ProcessIO_U.Flows.name)="Carbon dioxide, fossil") AND
((qry_ProcessIO_U.[Input?])=No));
```

Results: For power plants and natural gas, the calculated ratio is stable in the area of 0.056. Most surprising is, however, that only gas power plants report CO₂ emissions in ecoinvent (with one coke oven gas as exception).

kgCO2perMJ	PID	qry_Powerplants_CO2.qry_ProcessIO_U.Process.name	qry_Powerplants_CO2.qry_ProcessIO_U.Flows.name
0,05	11031	natural gas, burned in power plant	Carbon dioxide, fossil
0,05	5876	coke oven gas, burned in power plant	Carbon dioxide, fossil
0,0506	11033	natural gas, burned in power plant	Carbon dioxide, fossil
0,0507	11039	natural gas, burned in power plant	Carbon dioxide, fossil
0,0507	11036	natural gas, burned in power plant	Carbon dioxide, fossil
0,0508	11040	natural gas, burned in power plant	Carbon dioxide, fossil
0,0508	11032	natural gas, burned in power plant	Carbon dioxide, fossil
0,0508	11034	natural gas, burned in power plant	Carbon dioxide, fossil
0,0508	11035	natural gas, burned in power plant	Carbon dioxide, fossil
0,0508	11037	natural gas, burned in power plant	Carbon dioxide, fossil
0,051	11038	natural gas, burned in power plant	Carbon dioxide, fossil
0,055	5862	natural gas, burned in power plant	Carbon dioxide, fossil
0,05599	7186	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5868	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5863	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5864	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5865	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5866	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5869	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5870	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5871	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5872	natural gas, burned in power plant	Carbon dioxide, fossil
0,056	5875	natural gas, burned in combined cycle plant, best technology	Carbon dioxide, fossil
0,056	5873	natural gas, burned in power plant	Carbon dioxide, fossil
0,0561	5867	natural gas, burned in power plant	Carbon dioxide, fossil

Figure 17: Results of the query “qry_Powerplants_CO2perMJ”, for unit processes.
KgCO2perMJ is the quotient of fossil CO₂ emissions and input of fuel in MJ, per process

Interpretation, conclusion: This type of analysis is only one example of possible “trace” analysis that rely on, and trace, natural science-based causalities (as: fossil fuel in input will lead to carbon dioxide in emissions if these are not specially treated; and the generated amount of CO₂ that is created per MJ of fuel depends on the energy – or here: C – content of the fuel which is rather independent from the specific process). In a practical application, a need for “just another” of these analyses might occur. For example, a simple replacement of power plants by heating systems gives the following result:

CO2perMJ	PID	qry_Heating_CO2.qry_ProcessIO_U.Process.name	qry_Heating_CO2.qry_ProcessIO_U.Flows.name
0,0825	846	hard coal briquette, burned in stove 5-15kW	Carbon dioxide, fossil
0,095	847	hard coal coke, burned in stove 5-15kW	Carbon dioxide, fossil
0,042424	6934	heat, natural gas, at diffusion absorption heat pump 4kW, future	Carbon dioxide, fossil
0,0915	1009	lignite briquette, burned in stove 5-15kW	Carbon dioxide, fossil
0,056	1354	natural gas, burned in boiler atm. low-NOx condensing non-modulating <100kW	Carbon dioxide, fossil
0,056	1355	natural gas, burned in boiler atmospheric burner non-modulating <100kW	Carbon dioxide, fossil
0,056	1356	natural gas, burned in boiler atmospheric low-NOx non-modulating <100kW	Carbon dioxide, fossil
0,056	1357	natural gas, burned in boiler condensing modulating <100kW	Carbon dioxide, fossil
0,056	1358	natural gas, burned in boiler condensing modulating >100kW	Carbon dioxide, fossil
0,056	1359	natural gas, burned in boiler fan burner low-NOx non-modulating <100kW	Carbon dioxide, fossil
0,056	1360	natural gas, burned in boiler fan burner non-modulating <100kW	Carbon dioxide, fossil
0,056	1361	natural gas, burned in boiler modulating <100kW	Carbon dioxide, fossil
0,056	1362	natural gas, burned in boiler modulating >100kW	Carbon dioxide, fossil
0,056	1363	natural gas, burned in industrial furnace >100kW	Carbon dioxide, fossil
0,056	1364	natural gas, burned in industrial furnace low-NOx >100kW	Carbon dioxide, fossil

Figure 18: Emitted fossil CO₂ per MJ input, heating systems

It seems therefore good to allow users a modification of provided means of analysis to the specific needs of their present analysis and question. And this, in turn, requires flexible and robust analyses that can ideally be combined and modified by users (of different user roles) while being robust against calculation or interpretation errors.

4.2 *Statistical analyses and procedures*

In the previous section, most analyses results could be understood on the basis of a single process. This is for many analyses too limiting. In the following, results will therefore be presented always for a group of processes. This comprises statistical analyses in a narrow sense, which are meaningless when applied on one single data set, but it covers also rather basic analyses where one process is for example plotted together with other processes. There is of course some overlap to the previous section, where values of one process were in part also set against those of other processes. However, focus in this “combinations” section is on groups of processes, while focus in the previous section was one single process. This distinction is merely meant for organisational purposes and has no further value.

4.2.1 **Explorative data analysis (EDA)**

Explorative data analysis is a technique to literally explore data, without thinking of specific questions or application. Tukey, who wrote the standard book about EDA, motivates and defines EDA in a classical quote:

"exploratory data analysis is detective work. [...] restricting one's self to the planned analysis – failing to accompany it with exploration – loses sight of the most interesting results too frequently to be comfortable" [Tukey 1977, p. 3].

Thinking of ecoinvent data, there is of course quite a lot to explore. This has two consequences. First, in the context of this report, not all different possible “exploration tasks” can be conducted and described. Rather, different possible, and typical, examples for explorative data analysis will be presented. Second, in a practical application, it will make sense to allow some kind of unguided data analysis also for users. Typically, a series of analysis are performed, where one or several “exploration tracks” are followed. These different tracks are difficult to foresee. This idea will be picked up later, in chapter 6.2.

4.2.1.1 **Boxplots: Transport effort in tkm**

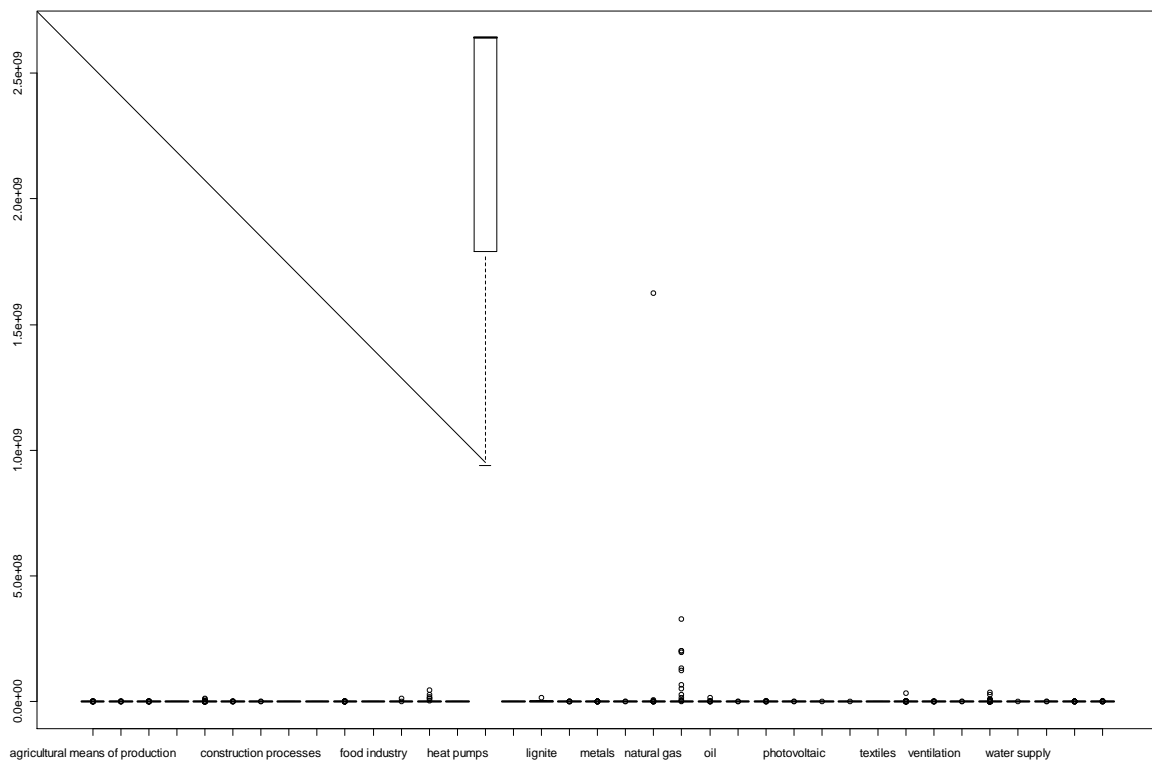
A first example of “data exploration paths” will be the analysis of the transport effort in tkm, per process¹⁰. To this end, the main query (qry_gesamt) was filtered to contain only unit processes, and only processes where the unit of an input flow is tkm. This new data frame is called tkm_U. These data are then displayed as boxplots, per main category.

The corresponding command line in R is...

```
boxplot(tkm_U[[1]]~tkm_U[[4]])
```

... which provides the following result:

¹⁰ It is understood that by “bluntly” comparing all processes, largely varying functional units may be the cause for large differences. Differentiating by functional unit therefore makes sense, but any such finetuning is typically done in a next step. The current report aims to follow the typical pathway of an analysis where little is known about the analysed data beforehand.

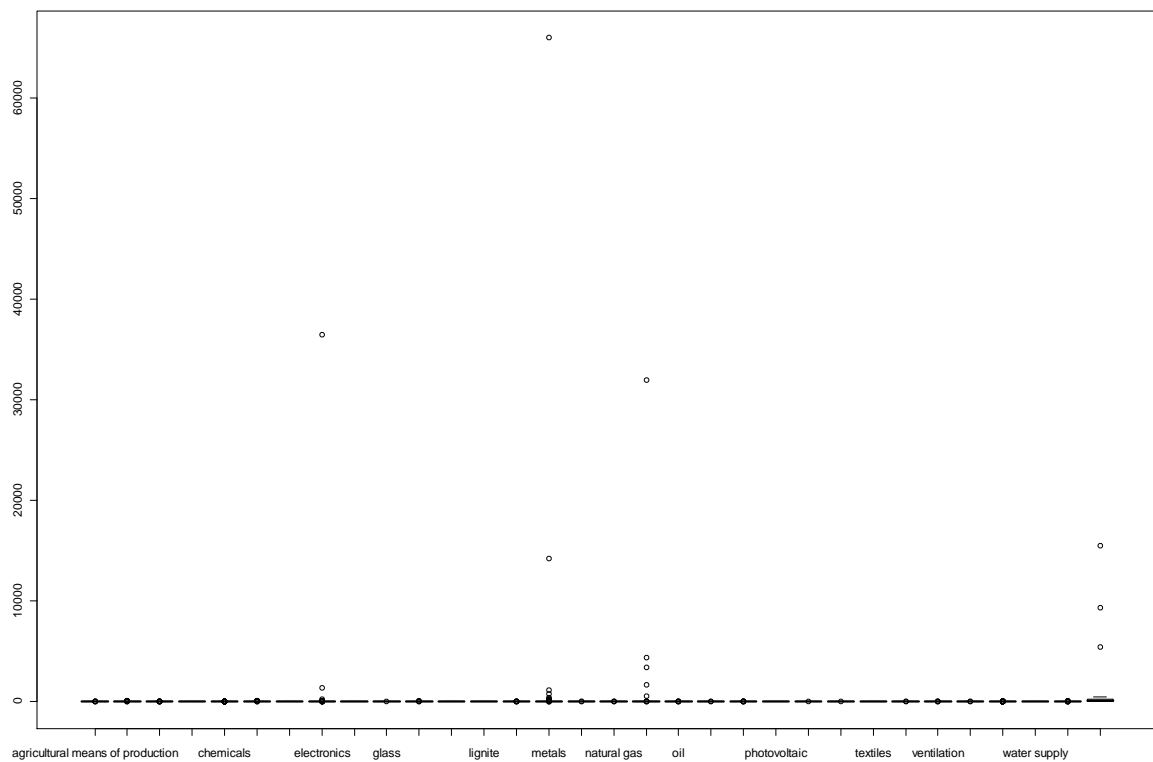


Most of the boxes are not visible due to category with extremely high values.

An inspection of the “raw data” (in Access) shows that the respective category is hydro power (plants). Other infrastructure processes (as uranium mill, nuclear power) have also high tkm values.

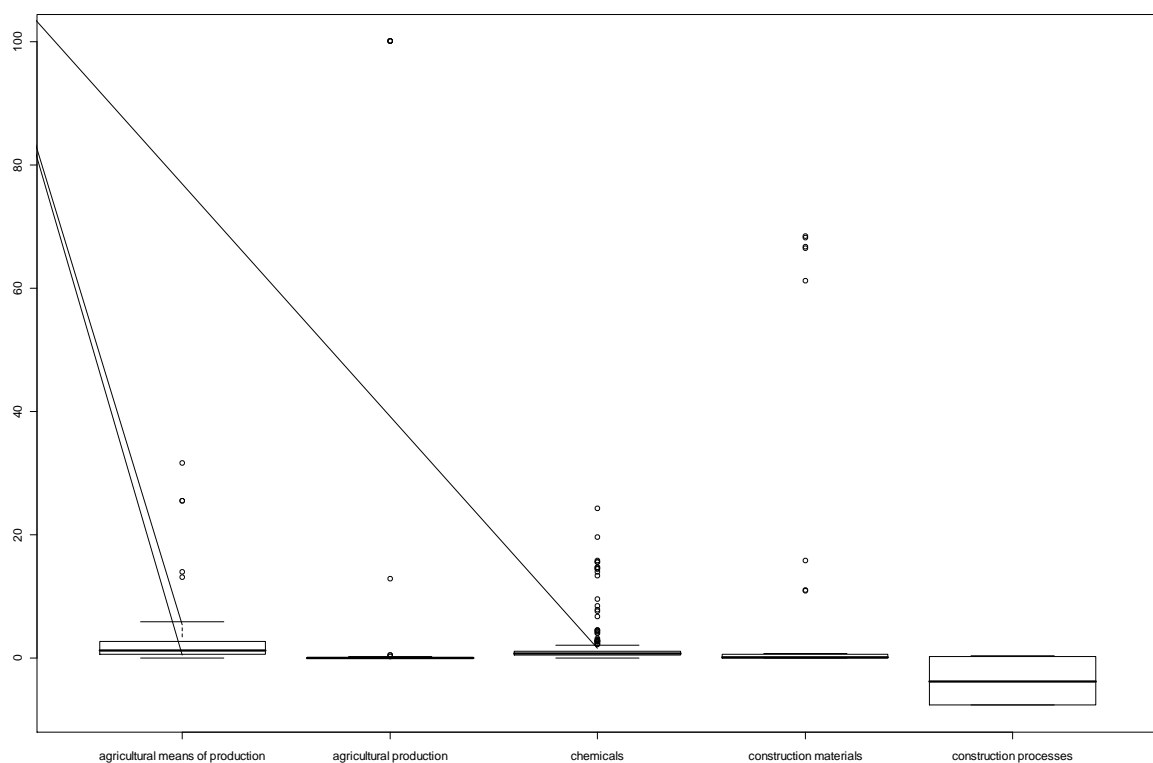
Summe von meanValue	Input?	Process_name	Process_category	Process_subCategory	infrastructureInc	Flows
2640000000	-1	reservoir hydropower plant, non alpine regions	hydro power	production of components	1	tkm
2640000000	-1	reservoir hydropower plant, alpine region	hydro power	production of components	1	tkm
2640000000	-1	reservoir hydropower plant	hydro power	production of components	1	tkm
1625000000	-1	production plant, natural gas	natural gas	production	1	tkm
940000000	-1	run-of-river hydropower plant	hydro power	production of components	1	tkm
328360000	-1	nuclear power plant, pressure water reactor 1000MW	nuclear power	power plants	1	tkm
204100000	-1	uranium mill	nuclear power	production	1	tkm
201900000	-1	nuclear power plant, boiling water reactor 1000MW	nuclear power	power plants	1	tkm
196200000	-1	uranium enrichment diffusion plant	nuclear power	production	1	tkm
134400000	-1	nuclear spent fuel reprocessing plant	nuclear power	waste treatment	1	tkm
124480000	-1	uranium enrichment centrifuge plant	nuclear power	production	1	tkm
67300000	-1	final repository for nuclear waste SF, HLW, and ILW	nuclear power	waste treatment	1	tkm
53760000	-1	nuclear fuel fabrication plant	nuclear power	production	1	tkm
47700000	-1	hard coal coke production plant	hard coal	production	1	tkm
36676000	-1	sanitary landfill facility	waste management	sanitary landfill	1	tkm
33100000	-1	port facilities	transport systems	ship	1	tkm
28900000	-1	final repository for nuclear waste LLW	nuclear power	waste treatment	1	tkm
28080000	-1	hard coal power plant	hard coal	power plants	1	tkm

One possible way forward could now be to exclude infrastructure processes This provides the following picture:



Evidently, the numbers are getting much smaller, but still there are outliers (in metals, with gold having more than 6000 tkm) which scale other boxplots to single lines.

So in addition, several similar categories are selected (and the others excluded). As example, agricultural production, chemicals, and construction processes are chosen.



This simple boxplot alone shows several things: Boxes for agricultural means of production and construction materials are comparatively narrow, meaning that all values for transport effort lie within a certain close range. For construction materials, there are only two small groups of outliers – by inspection of the Access data, these are concrete production and refractory...

Summe von meanValue	Input?	Process_name	Process_category	Process_subC
58,46	-1	concrete, exacting, with de-icing salt contact, at plant	construction materials	concrete
68,24	-1	concrete, exacting, at plant	construction materials	concrete
66,73	-1	concrete, sole plate and foundation, at plant	construction materials	concrete
66,458	-1	concrete, normal, at plant	construction materials	concrete
61,19	-1	poor concrete, at plant	construction materials	concrete
15,8027	-1	expanded vermiculite, at plant	construction materials	additives
11,02762	-1	refractory, basic, packed, at plant	construction materials	bricks
10,92737	-1	refractory, high aluminium oxide, packed, at plant	construction materials	bricks
0,6814	-1	sanitary ceramics, at regional storage	construction materials	others
0,661	-1	ceramic tiles, at regional storage	construction materials	coverings
0,632448	-1	fibre cement facing tile, at plant	construction materials	coverings
0,55212	-1	refractory, fireclay, packed, at plant	construction materials	bricks
0,473456	-1	fibre cement roof slate, at plant	construction materials	coverings
0,356344	-1	fibre cement corrugated slab, at plant	construction materials	coverings
0,3	-1	gypsum fibre board, at plant	construction materials	coverings
0,3	-1	gypsum plaster board, at plant	construction materials	coverings
0,3	-1	natural stone plate, cut, at regional storage	construction materials	others
0,121	-1	lightweight concrete block, pumice, at plant	construction materials	concrete

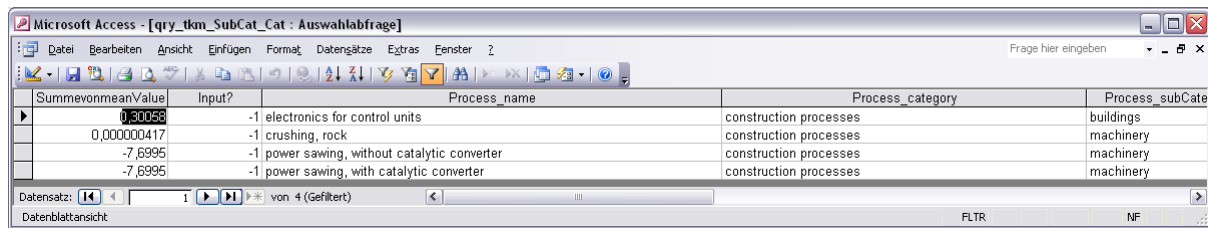
... and one reason for this is that the quantitative reference for the concrete unit processes is 1 m³, while it is 1 kg for other construction materials.

PID	ProcessType	name	localName	infrastructurePr	amount	unit	category	subCategory	localCategory	local
490	UnitProcess	portland cement, strength class Z 42	Portlandzement		1	kg	construction materials	binder	Mineralische B: Bind	
491	UnitProcess	portland cement, strength class Z 52	Portlandzement		1	kg	construction materials	binder	Mineralische B: Bind	
492	UnitProcess	portland slag sand cement, at plant	Portlandhüttenz		1	kg	construction materials	binder	Mineralische B: Bind	
493	UnitProcess	stucco, at plant	Stuckgips, ab v		1	kg	construction materials	binder	Mineralische B: Bind	
494	UnitProcess	autoclaved aerated concrete block, at plant	Gasbetonstein		1	kg	construction materials	bricks	Mineralische B: Mau	
495	UnitProcess	brick, at plant	Backstein, ab v		1	kg	construction materials	bricks	Mineralische B: Mau	
496	UnitProcess	light clay brick, at plant	Leichtlehmstei		1	kg	construction materials	bricks	Mineralische B: Mau	
497	UnitProcess	refractory, basic, packed, at plant	Feuerfeste Stei		1	kg	construction materials	bricks	Mineralische B: Mau	
498	UnitProcess	refractory, fireclay, packed, at plant	Feuerfeste Stei		1	kg	construction materials	bricks	Mineralische B: Mau	
499	UnitProcess	refractory, high aluminium oxide, pac	Feuerfeste Stei		1	kg	construction materials	bricks	Mineralische B: Mau	
500	UnitProcess	sand-lime brick, at plant	Kalksandstein		1	kg	construction materials	bricks	Mineralische B: Mau	
501	UnitProcess	cement cast plaster floor, at plant	Zement Unterla		1	kg	construction materials	concrete	Mineralische B: Beto	
502	UnitProcess	concrete, exacting, at plant	Beton, hohe An		1	m ³	construction materials	concrete	Mineralische B: Beto	
503	UnitProcess	concrete, exacting, with de-icing salt	Beton, hohe An		1	m ³	construction materials	concrete	Mineralische B: Beto	
504	UnitProcess	concrete, normal, at plant	Beton, normal		1	m ³	construction materials	concrete	Mineralische B: Beto	
505	UnitProcess	concrete, sole plate and foundation, at plant	Beton, Bodenpl		1	m ³	construction materials	concrete	Mineralische B: Beto	
506	UnitProcess	concrete block, at plant	Betonstein, ab		1	kg	construction materials	concrete	Mineralische B: Beto	
507	UnitProcess	lightweight concrete block, expanded	Leichtbetonstei		1	kg	construction materials	concrete	Mineralische B: Beto	

The group agricultural production has only two outliers, green manure – and these processes have again a different quantitative reference than the other processes in this category (ha instead of kg):

PID	ProcessType	name	localName	infrastructurePr	amount	unit	category	subCategory	localCategory	local
193	UnitProcess	barley straw extensive, at farm	Gerstenstroh E		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
194	UnitProcess	barley straw IP, at farm	Gerstenstroh IF		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
195	UnitProcess	barley straw organic, at farm	Gerstenstroh B		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
196	UnitProcess	fava beans IP, at farm	Ackerbohnen IF		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
197	UnitProcess	fava beans organic, at farm	Ackerbohnen B		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
198	UnitProcess	fodder beets IP, at farm	Futterrüben IP		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
199	UnitProcess	palm fruit bunches, at farm	Palmlfruchtstän		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
200	UnitProcess	grain maize IP, at farm	Körnermais IP		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
201	UnitProcess	grain maize organic, at farm	Körnermais Bio		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	
202	UnitProcess	green manure IP, until April	Gründüngung IF		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
203	UnitProcess	green manure IP, until February	Gründüngung IF		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
204	UnitProcess	green manure IP, until January	Gründüngung IF		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
205	UnitProcess	green manure IP, until march	Gründüngung IF		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
206	UnitProcess	green manure organic, until April	Gründüngung B		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
207	UnitProcess	green manure organic, until February	Gründüngung B		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
208	UnitProcess	green manure organic, until January	Gründüngung B		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
209	UnitProcess	green manure organic, until march	Gründüngung B		1	ha	agricultural production	plant production	Landwirtschaftli Pfan	
210	UnitProcess	hay extensive, at farm	Heu extensiv, a		1	kg	agricultural production	plant production	Landwirtschaftli Pfan	

Construction processes, on the far right, have a surprising boxplot, with mainly negative transport efforts. This category contains only four processes...



Summe von meanValue	Input?	Process_name	Process_category	Process_subCate
0.30058	-1	electronics for control units	construction processes	buildings
0.000000417	-1	crushing, rock	construction processes	machinery
-7.6995	-1	power sawing, without catalytic converter	construction processes	machinery
-7.6995	-1	power sawing, with catalytic converter	construction processes	machinery

where the negative values stem from the process “power sawing”. This needs further explanation and should be checked when looking at the detailed process documentation.

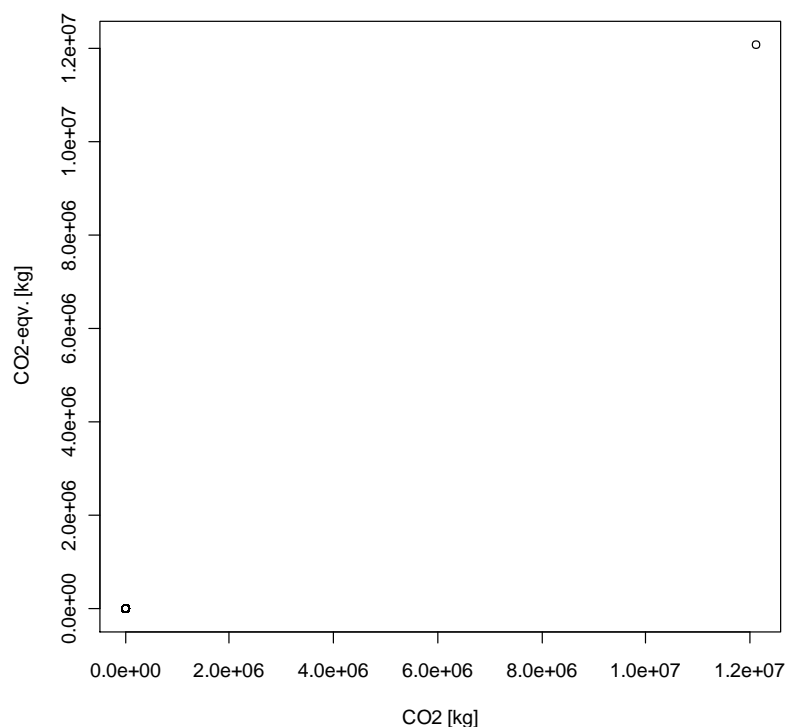
4.2.1.2 Scatter plots: CO₂ emissions, climate change potential

A scatter plot plots quantitative values as dots. As an example, a scatter plot is generated for all types of CO₂ emissions, vs. climate change potential contributions, for each process. First, the impact assessment result is calculated in Access.

The R command to create the basic plot is...

```
plot(gwp_CO2[[1]], gwp_CO2[[2]], ylab= "CO2-equiv. [kg]", xlab = "CO2 [kg]")
```

...and the result is (x-Axis: CO₂, y-axis: GWP acc. to IPCC 2001).



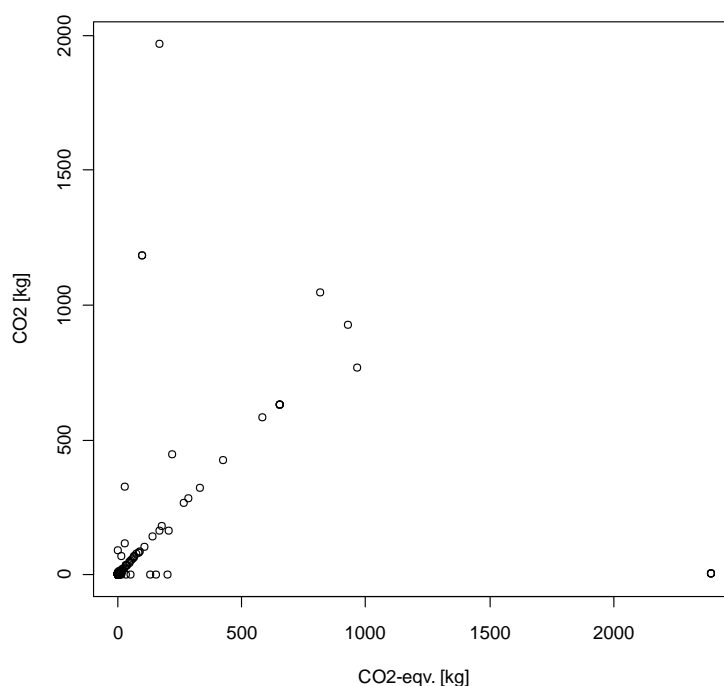
Again, for all processes, the scatter plot does not show much due to the extreme value of one process. From raw (Access) data it can be seen that the responsible process is “operation, maintenance, airport”, which has as quantitative reference 1 unit of airport (over the airport’s life time).

Microsoft Access - [qry_process_CO2 : Auswahlabfrage]

Summe von mea	Process_name	Process_PID
12100000	operation, maintenance, airport	1891
11400	radioactive waste, in interim storage conditioning	6000
1970.8	roundwood, azobe (SFM), under bark, u=30%, at	10200
1970	azobe (SFM), standing, under bark, in rain forest	10199
1808	azobe, allocation correction 1	10216
1808	azobe, allocation correction 2	10217
1640	eucalyptus ssp., standing, under bark, u=50%, in	10212

Datensatz: 1 von 1210
Datenblattansicht

Excluding this process, and also the also rather exotic process “radioactive waste in interim storage conditioning” yields the following plot.



This plot shows some correlation of CO₂ emissions of a process to GWP indicator values, but there are also processes with high CO₂-equivalent values and low CO₂ emissions, and vice versa. Both types of processes are related to biomass and agriculture. The processes with high CO₂-equivalent values are the “manure” processes discussed already above.

Microsoft Access - [qry_process_gwp_vs_CO2 : Auswahlabfrage]

GWP_IPCC	CO2	Process_name
2389,7212	4,2888	green manure IP, until march
2389,7212	4,2888	green manure IP, until February
2389,7212	4,2888	green manure IP, until April
2389,7212	4,2888	green manure IP, until January
966,9501234	769	treatment, heat carrier liquid, 40% C3H8O2, to
928,67	928,67	rhodium, primary, at refinery
815,605325445225	1045,094167	particle board, cement bonded, at plant

Datensatz: 2 von 1115
Datenblattansicht

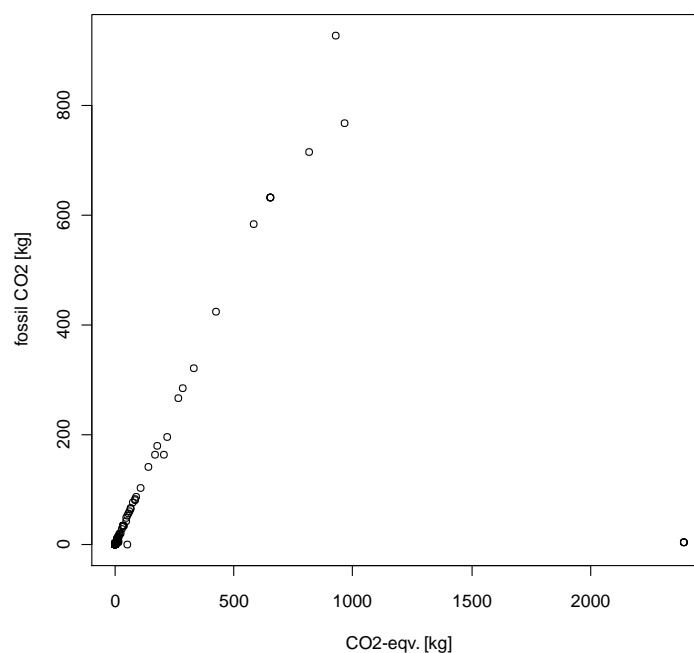
Processes with high CO₂ emissions and low GWP values are evidently emitting biogenic CO₂:

Microsoft Access - [qry_process_gwp_vs_CO2 : Auswahlabfrage]

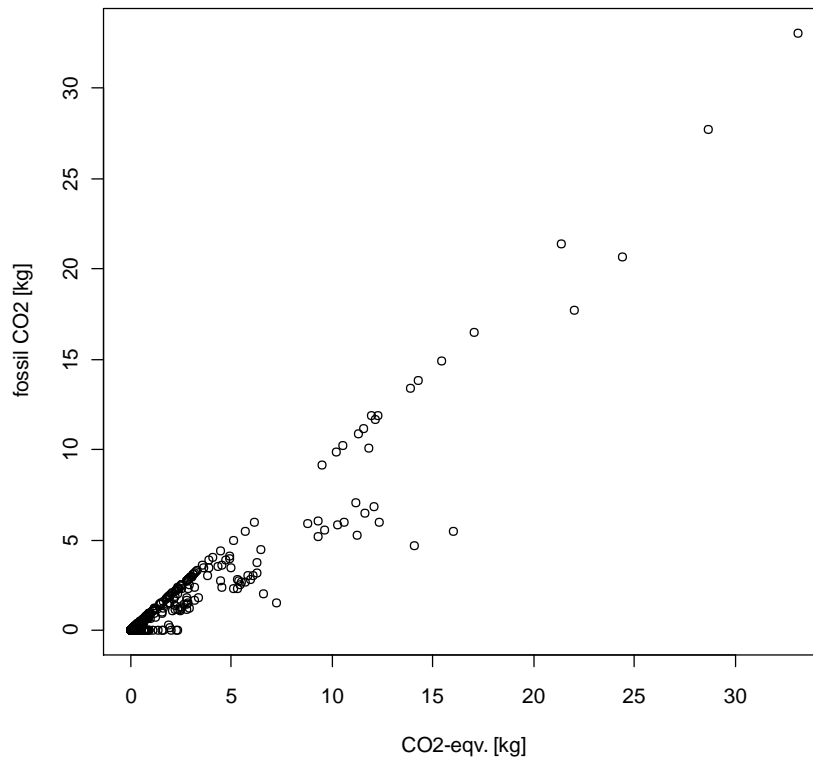
GWP_IPCC	CO2	Process_name
167,137242	1970,6	roundwood, azobe (SFM), under bark, u=30%,
100,2768979	1182,5	roundwood, meranti (SFM), under bark, u=70%,
100,2768979	1182,5	roundwood, paraná pine (SFM), under bark, u=E
815,605325445225	1045,094167	particle board, cement bonded, at plant
928,67	928,67	rhodium, primary, at refinery
966,9501234	769	treatment, heat carrier liquid, 40% C3H8O2, to v
652,2590722	633	lorry 40t

Datensatz: 14 von 1115
Datenblattansicht

Limiting the analysis to fossil CO₂ (by adding a filter to the Access query) provides a slightly different plot (again for values below 1500 kg):



Now there are no processes with higher CO₂-equivalents than CO₂ emissions, which is plausible. Limiting the analysis further to values below 35 kg shows the following plot.



4.2.1.3 Matrix scatter plots: Impact assessment categories

Next example in EDA is a matrix scatter plot of impact categories. A matrix scatter plot is a compilation of several x-y plots (as shown in the section above) in a matrix. It is intended to provide a good first overview “at one glance” over multivariate data. In the following it will be used to display results for selected impact categories, per process. For the matrix plot, it is necessary to transform the list-type data in the Access database into a pivot table.

A preparatory query selects all impact categories with eutrophication, acidification, and human health.

```
SELECT ProcessIO_U.Process_name, Sum([meanValue]*[Factor]) AS
indicatorValue, Wirkungsabschätzung.SubCategory,
Wirkungsabschätzung.Category, [SubCategory]+"_"+[Category] AS SubCatCat
FROM ProcessIO_U INNER JOIN Wirkungsabschätzung ON ProcessIO_U.Flows_FID =
Wirkungsabschätzung.FID
GROUP BY ProcessIO_U.Process_name, Wirkungsabschätzung.SubCategory,
Wirkungsabschätzung.Category, [SubCategory]+"_"+[Category]
HAVING ((Wirkungsabschätzung.SubCategory) Like "*eutrophication*" Or
(Wirkungsabschätzung.SubCategory) Like "*acidification*" Or
(Wirkungsabschätzung.SubCategory) Like "*human*");
```

Microsoft Access - [qry_WK_matrix_e_a_h : Auswahlabfrage]

Process_name	indicatorValue	SubCategory	Category	SubCatCat
uranium, enrich	0,5085	acidification	EDIP2003	acidification_EDIP2003
wood chips, froi	0,000961413	acidification	EDIP2003	acidification_EDIP2003
sodium perbora	0,0864106303	acidification	EDIP2003	acidification_EDIP2003
uranium, enrich	0,000008362	acidification	EDIP2003	acidification_EDIP2003
uranium, enrich	0,5085	acidification	EDIP2003	acidification_EDIP2003
uranium, enrich	0,000008362	acidification	EDIP2003	acidification_EDIP2003
sodium perbora	0,1940584353	acidification	EDIP2003	acidification_EDIP2003
uranium, enrich	0,000004181	acidification	EDIP2003	acidification_EDIP2003

Datensatz: 1 von 7247

In a following query, these results are transformed into a pivot structure. Missing values are filled with 0:

TRANSFORM

IIf(Avg([qry_WK_matrix_e_a_h].[indicatorValue])>0,Avg([qry_WK_matrix_e_a_h].[indicatorValue]),0) AS MittelwertvonindicatorValue

SELECT qry_WK_matrix_e_a_h.Process_name,
Avg(qry_WK_matrix_e_a_h.indicatorValue) AS [Gesamtsumme von indicatorValue]
FROM qry_WK_matrix_e_a_h

GROUP BY qry_WK_matrix_e_a_h.Process_name

PIVOT qry_WK_matrix_e_a_h.SubCatCat;

Microsoft Access - [qry_WK_matrix_e_a_h_pivot : Kreuztabellenabfrage]

Process_name	Gesamtsumme	acidification_E	eutrophication	human health	human toxicity
1,1-difluoroethane, HF	0,0131300076			0,0262600139	1,306046E-09
1,1-dimethylcyclopent	0,0005360541		0,0001072764	0,0014989444	1,941602E-06
1-butanol, propylene h	0,0002467554		0,000706926	3,326264E-05	7,760733E-08
1-pentanol, at plant	0,0017069192		0,0002141678	0,0049061218	4,681766E-07
1-propanol, at plant	0,0012393419			0,0024730736	5,610118E-06
2,3-dimethylbutan, fro	0,0005361065		0,0001072874	0,0014990904	1,941767E-06
2-butanol, at plant	0,0017679321		0,00046123	0,0048370224	5,543975E-06
2-methyl-1-butanol, at	0,0017069133		0,0002141656	0,0049061063	4,681408E-07

Datensatz: 1 von 1624

These results are transferred to R via the usual fetch procedure, into a new data frame, and are ready to be used for the matrix plots. In order to create the matrix plots, the command line in R is as follows:

```
WK_matrix <- sqlFetch(channel, "qry_WK_matrix_e_a_h_pivot") #to read the
pivot table
pairs(WK_matrix [3:9], labels(1:7))
```

... which provides the following result (Figure 19).

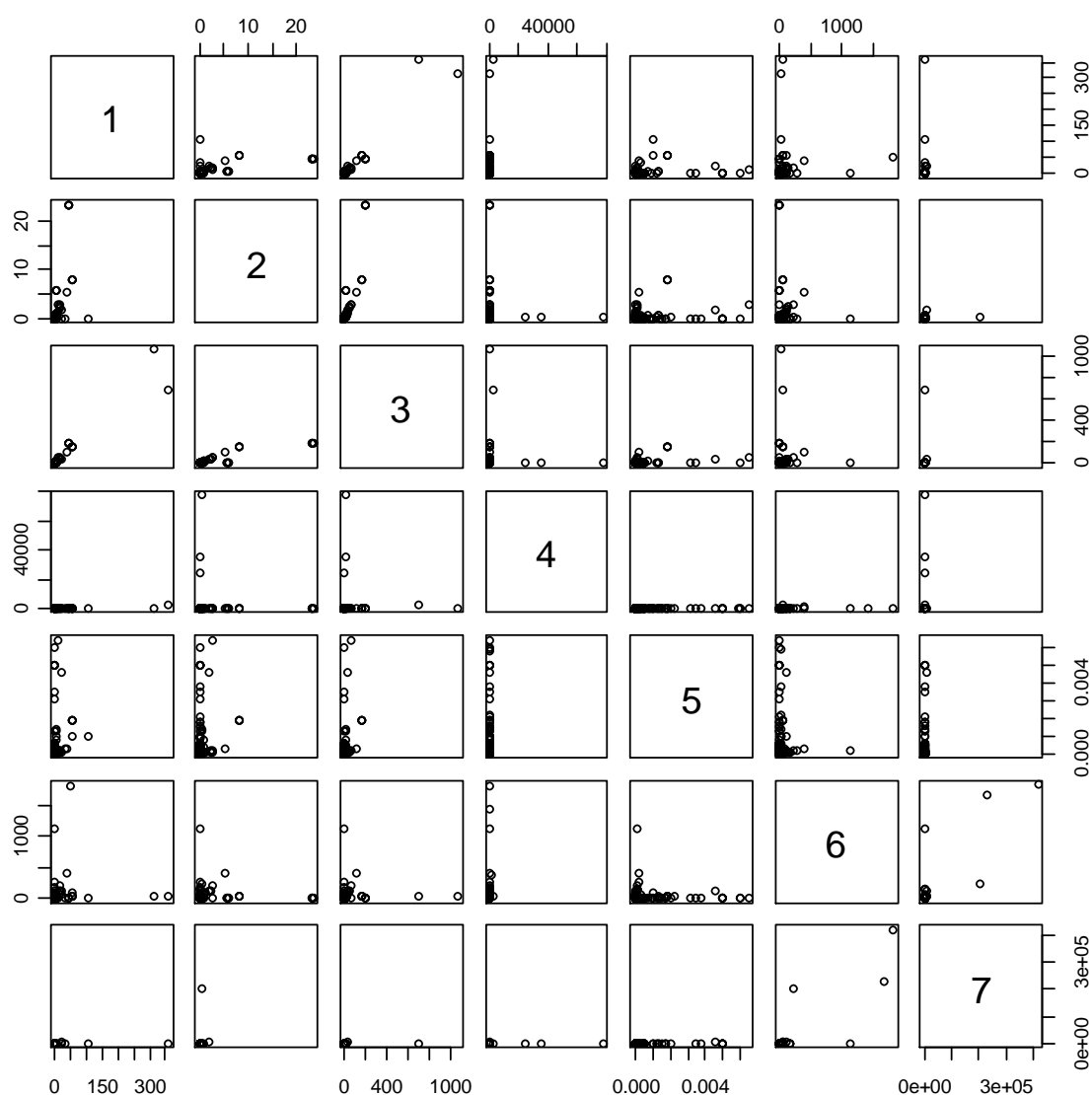


Figure 20: Matrix scatter plots of selected impact category results for all unit processes in ecoinvent, limited to amounts smaller than twice the mean;
1: Acidification acc. to EDIP 2003, 2: eutrophication acc. to CML 2001, 3: eutrophication acc. to EDIP 2003, 4: human health acc. to Ecoindicator 99 (I,I) 5: human health acc. to Impact2002+ 6: human toxicity acc. to CML 2001 7: human toxicity acc. to EDIP 2003

This plot allows more insight. Several things are interesting. For example, there seems to be correlation between the EDIP categories eutrophication and acidification (1 and 3 in the figure) - which is plausible. There seems no correlation at all between category results for human health in EcoIndicator and in EDIP (4 and 5). There seems also no correlation at all between human health in IMPACT 2002+ and human tox. in EDIP (5 and 7). This figure will motivate further analysis, later in this report.

4.2.2 Statistical tests

Statistical tests can be seen as a compliment to exploratory data analysis. While the latter openly explores “things” contained in data, tests are performed with an idea about something that might be in a data sample. Some, e.g. Sachs 1992, speak of ‘confirmatory data analysis’ when referring to tests.

Since statistical tests are in broad use for decades, it is not surprising that many different approaches exist. Providing an overall introduction in this context would lead too far; good introductions can be found in many statistical textbooks, such as (Sachs 1992, pp. 193, Hartung 1993, pp. 132).

The following paragraph will describe a so-called significance test. Basically, a significance test serves to investigate whether a hypothesis (a statement, an assumption) about certain data should be rejected or not.

An example for a hypothesis could be:

“Carbon dioxide emissions for agricultural production processes are not very different, per kg product produced, from carbon dioxide emissions in all other production processes, per kg product produced, in ecoinvent.”

In a test, this is then called the Null hypothesis, H_0 . H_0 is rejected if it is found that by doing so an error is very unlikely, according to a specific test statistic. Usually, how “very unlikely” is still acceptable, i.e. the probability threshold for rejecting H_0 , is specified for a test. This threshold is often called the significance level. Common significance levels are 5%, or 1%.

The decision about whether to reject H_0 or not has two potential errors: First, H_0 can be erroneously rejected; second, it can be erroneously accepted. Erroneously rejecting H_0 is often called a “type I” error, while erroneously accepting H_0 is called a “type II” error. Ideally, a test statistic has a very low type II error, and users usually select the type I error, which is often equated with the significance level of the test.

Nice thing is that based on the test results, one can indeed speak of significance (while the term significance, or that something may be significantly lower or higher than something else, is used broadly in engineering and managerial speak, without any such foundation).

For demonstration, the null hypothesis above shall be implemented in R. To this end, first database queries are created that aggregate all CO₂ flows (fossil and biogenic) per process, negative and positive values, for agricultural production processes and for all other production processes (with ‘kg’ as unit for the product):

Query 12: qry_process_CO2_agri

```
SELECT ProcessIO_U.Process_name, ProcessIO_U.Process_PID,
Sum(ProcessIO_U.meanValue) AS SummevonmeanValue,
ProcessIO_U.Process_category, ProcessIO_U.Process_unit
FROM ProcessIO_U
WHERE (((ProcessIO_U.Flows_name) Like "Carbon dioxide,*"))
GROUP BY ProcessIO_U.Process_name, ProcessIO_U.Process_PID,
ProcessIO_U.Process_category, ProcessIO_U.Process_unit
HAVING (((ProcessIO_U.Process_category)="agricultural production") AND
((ProcessIO_U.Process_unit)="kg"));
```

Query 13: qry_process_CO2_non_agri

```
SELECT ProcessIO_U.Process_name, ProcessIO_U.Process_PID,
Sum(ProcessIO_U.meanValue) AS SummevonmeanValue,
ProcessIO_U.Process_category, ProcessIO_U.Process_unit
FROM ProcessIO_U
WHERE (((ProcessIO_U.Flows_name) Like "Carbon dioxide,*"))
GROUP BY ProcessIO_U.Process_name, ProcessIO_U.Process_PID,
ProcessIO_U.Process_category, ProcessIO_U.Process_unit
HAVING ((Not (ProcessIO_U.Process_category)="agricultural production") AND
((ProcessIO_U.Process_unit)="kg"));
```

Before the test is run, a boxplot explores the emissions in each data group:

```
carbTestAgri <- sqlFetch(channel, "qry_process_CO2_agri", colnames = TRUE,
rownames = FALSE, max=0)

carbTestNonAgri <- sqlFetch(channel, "qry_process_CO2_non_agri", colnames =
TRUE, rownames = FALSE, max=0)

boxplot(carbTestAgri[[3]], xlab = "Agricultural Production Processes", ylab =
"CO2 emissions [kg/kg product]")

boxplot(carbTestNonAgri[[3]], xlab = "Production Processes, w/o
Agriculture", ylab = "CO2 emissions [kg/kg product]")
```

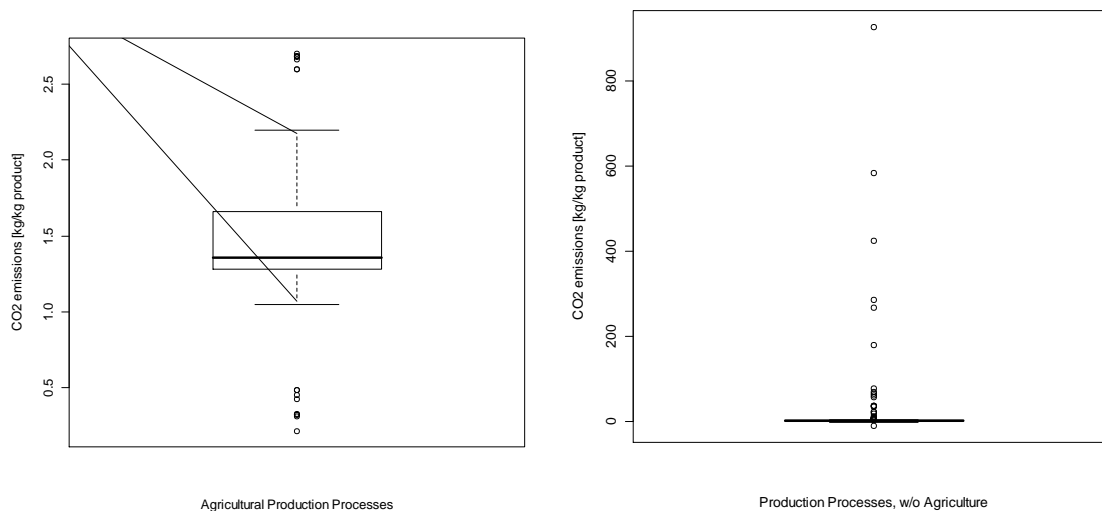


Figure 21: Boxplot of CO2 emissions of ecoinvent processes with a product in unit of ‘kg’; left: agricultural production processes; right: all other processes

These plots raise doubts about H0 (which is quite common: often tests aim at rejecting H0).

A common test for the task at hand is a t-test, which compares the means of two data samples. Basically there are two variants to be distinguished, one that assumes equal variances, and one that does not. The boxplots indicate that the variances are different¹¹. An unpaired t-test¹² is then conducted with the following R command...

```
> t.test(carbTestNonAgri[[3]], carbTestAgri[[3]], var.equal=FALSE)
```

...providing as result:

¹¹ In a more detailed analysis, this should be further investigated by another test, e.g. the F-test; also, one should check the probability distribution of the tested samples. The t-test assumes normal distribution of data. If it is found that this is probably violated by the data, data might be transformed to meet the normal distribution requirement.

¹² A „paired“ t-test would assume that data is obtained by sampling the same population twice which is not the case here.

```

Welch Two Sample t-test

data: carbTestNonAgri[[3]] and carbTestAgri[[3]]
t = 2.2976, df = 550.772, p-value = 0.02196
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7596372 9.7140700
sample estimates:
mean of x mean of y
 6.690909  1.454056

```

The “p-value” provides the calculated probability threshold. If it is lower than the stated significance level (e.g. 0.05 or 0.01), then H0 should be rejected, and otherwise, H0 cannot be rejected. So, assuming a significance level of 5%, one can safely reject H0:

This leads to the following interpretation: Carbon dioxide emissions for agricultural production processes are significantly different, per kg product produced, from carbon dioxide emissions in all other production processes, per kg product produced, in ecoinvent, with a significance level of 5%.

Interesting to note that with a more demanding significance level of 1%, it is not possible to reject H0, based on the test results.

Tests play an important role in checking, and validating, the insights gained from the structural analyses, for example the cluster analysis described in section 4.2.3.

4.2.3 Cluster analysis: Impact categories, and impact category results per process

A cluster analysis is a technique to classify data in groups. The groups should be homogenous, while the cluster should differ from one to another (e.g. [Schnell 1994, p. 291]). Many different ways to perform a cluster analysis exist; main differences are the ways the similarity / homogeneity is detected, and the specific selection and grouping algorithm.

All impact category results in ecoinvent are first aggregated, per unit process, in one pivot table in Access.

These data are then transferred to R, where the cluster analysis is called by the following statements.

```
WK_ges_clean <- WK_ges[2:26] #to eliminate Process-PIDs
tWK_ges <- t(WK_ges_clean)
# transpose the matrix, needed for cluster analysis

d_WK_ges <- dist(tWK_ges) #creation of a distance matrix

hc <- hclust(d_WK_ges, method = "average")
#perform hierarchical cluster analysis

plot(hc, hang = -1) #and plot the results
```

The result is as follows.

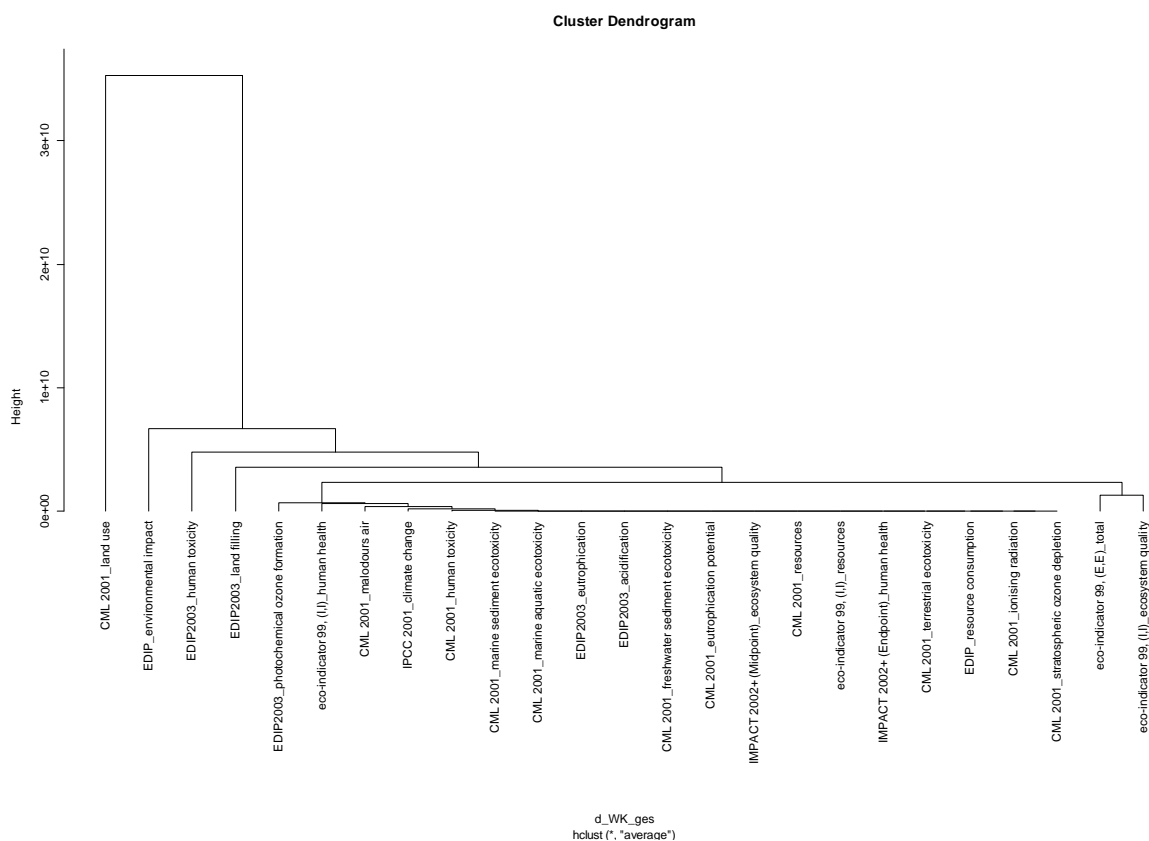


Figure 22: Dendrogram as result of a cluster analysis for impact categories provided in ecoinvent (not all categories are visible)

This dendrogram shows that land use according to CML is very different from all other categories; some EDIP, and some EcoIndicator results are also a bit different from other categories, but the remaining “lot” of categories can be barely distinguished in the graph.

An analysis with some more similar categories, as used also in the matrix scatter plot, might provide a different, finer resolved picture. Interesting, on the other side, might also be a cluster analysis of processes instead of impact categories. This is shown in the next figure (Figure 23). The figure can be read from bottom to top. At the bottom, every process is one single vertical line. Moving upwards, similar lines are grouped together with a horizontal connecting line.

With several thousand processes, the labels at the bottom become unreadable; interesting is though the structure that is appearing in the graph. There is one small group, at the far right, which is very different from all other processes; there is another small group, at the far left, which is also quite different. There is further a very large group of processes that seem rather similar, with some smaller subgroups in between. Results will depend though on the composition of the overall process group that is investigated; allowing only processes with “non-extreme” impact assessment results served already to that purpose.

These results show that it is possible to differentiate processes in ecoinvent by their impact assessment results. This can be helpful additional information in the evaluation of processes¹³.

¹³ Of course, the difference matrix that is basis for cluster analysis, contains Euclidian distances for all categories – which implies that all categories are treated equally, and therefore a weighting is applied. Therefore, these cluster analysis results should be used similar to a proxy and not used to overwrite specific considerations for processes.

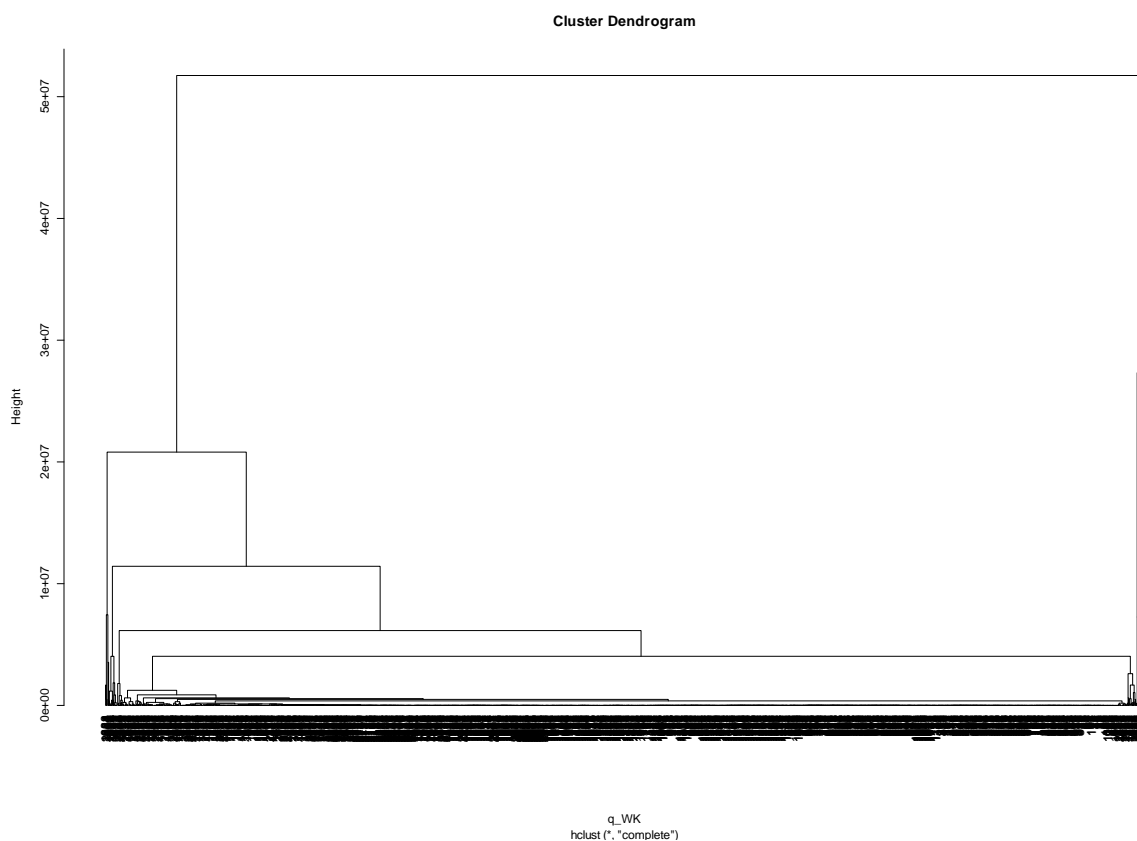


Figure 23: Dendrogram as result of a cluster analysis for impact category results for all unit processes in ecoinvent with results below twice the mean in each impact category (process labels at bottom are not readable)

4.2.4 A heatmap of mean values per flow category and process category

A heatmap is a tool to for refined EDA of two groups with one common (shared) numerical variable. In a two-dimensional matrix, differences of the variable values are calculated and presented on a colour scale (usually from red – low – to white – very high); rows and columns of the matrix are given by the different group values. In addition, for row and columns a cluster analysis dendrogram is calculated and put at the row and column.

As a test, a heatmap is produced for all flow means as numerical values, with the process categories and the flow categories building the two groups.

A simple Access query is created to extract data from the database:

```
TRANSFORM Sum(ProcessIO_U.meanValue) AS SummevonmeanValue
SELECT ProcessIO_U.Flows_category
FROM ProcessIO_U
GROUP BY ProcessIO_U.Flows_category
PIVOT ProcessIO_U.Process_category;
```

In R, these data first need to be converted into a numerical matrix, and afterwards the heatmap procedure can be called:

```
dm_ProcCat_Flow_Mean_pivot <- data.matrix(ProcCat_Flow_Mean_pivot,
rownames.force=TRUE)

# transform into a numerical matrix

heatmap(dm_ProcCat_Flow_Mean_pivot, main = "heatmap of flow categories over
process categories")
```


The graphical output is shown in the next figure.

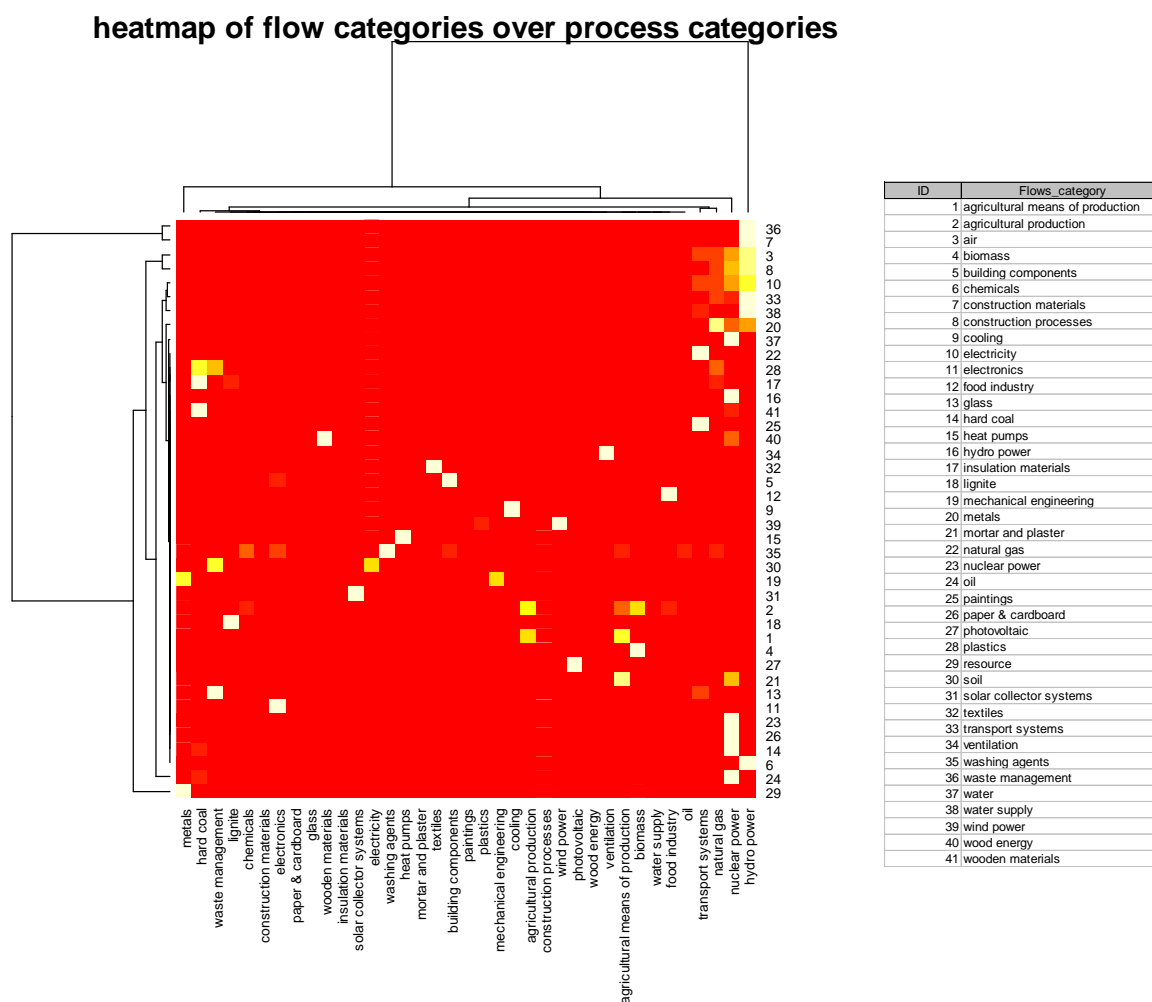


Figure 24: Heat map, missing values replaced by 0 (further explanation see text)

The output looks plausible; natural gas, nuclear power, and hydropower are different from all other process categories (on the right side – this is shown in the dendrogram and can also be seen by the colours); many missing values (non-occurring or non-reported emissions) result in red space. Excluding the three categories natural gas, nuclear power, and hydro power with their extreme values provides the following results:

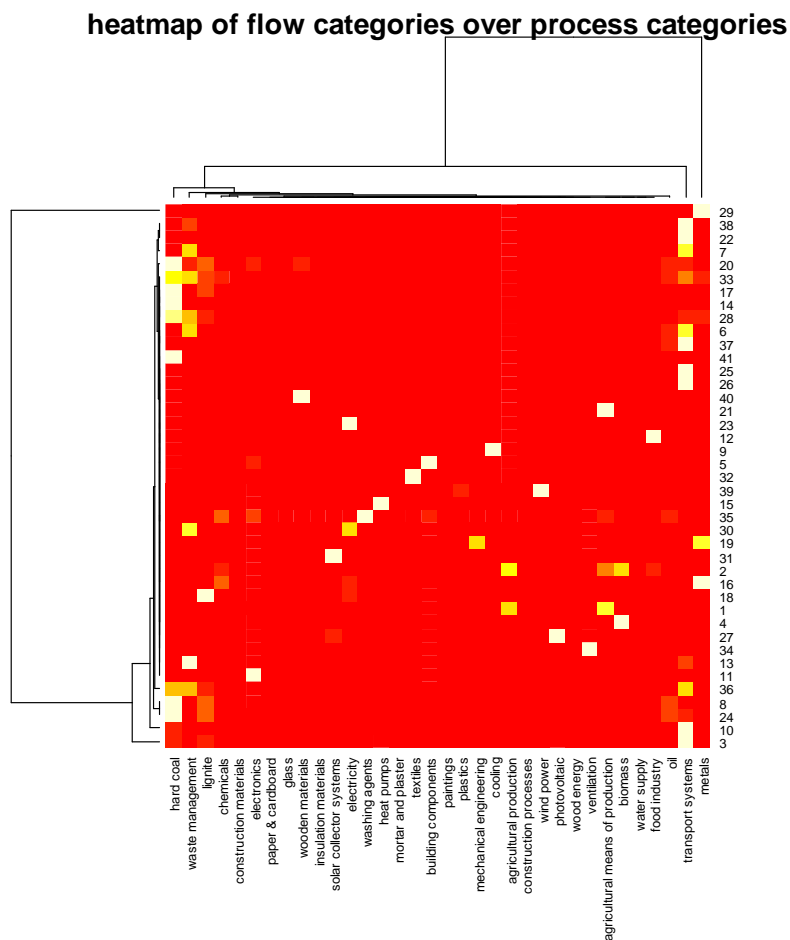


Figure 25: Heat map, missing values replaced by 0, extreme categories excluded (further explanation see text)

In the new picture, transport systems, waste management, and also hard coal are differing from the other categories. For transport and waste management, this is again plausible.

Although both flow categories and process categories are defined in general, at the moment in ecoinvent, based on user convenience and not on their flow pattern, some patterns evidently exist. A more refined, data management type investigation could classify processes based on their input output properties, or also based on their results in several impact assessment methods. This would then allow benchmarking processes in several respects: Is a new process put – automatically – in a category that is plausible? Does an updated process move out of a category that is plausible? These questions will become more important once ecoinvent decides to distinguish between processes and flows.

4.2.5 Distribution tests: Fossil CO₂ emissions, and radionuclides

Regarding the different values for single variables, such as the emission of a certain substance, in the ecoinvent database one may wonder about their probability distribution. This implies that the variables are treated as random variables which is a common assumption in statistics.

In the (not so abundant) literature on probability distributions for LCA data, it is often assumed that emission flows are distributed following a log normal distribution. The type of probability distribution that is “obeyed” by the data in ecoinvent is needed, for example, for the calculation of confidence intervals for the respective variables.

In the following, emissions for CO₂ and for Cesium-137, one of the radionuclide emissions reported by ecoinvent, will be tested against a log normal distribution. As a third variable, the aggregated EcoIndicator99 results will be taken in the analysis as well. The analysis will consider the system processes only.

In order to do so, the following steps are performed:

First, process balance data is filtered (in Access) to provide only emissions of fossil CO₂ and Cesium-137, respectively. EcoIndicator values are calculated in a query:

```
SELECT Qry_ProcessIO_S.Process.PID, Sum([meanValue]*[Factor]) AS
indicatorValue, Wirkungsabschätzung.SubCategory,
Wirkungsabschätzung.Category, Qry_ProcessIO_S.Process.name
FROM Qry_ProcessIO_S INNER JOIN Wirkungsabschätzung ON
Qry_ProcessIO_S.Flows.FID = Wirkungsabschätzung.FID
GROUP BY Qry_ProcessIO_S.Process.PID, Wirkungsabschätzung.SubCategory,
Wirkungsabschätzung.Category, Qry_ProcessIO_S.Process.name
HAVING (((Wirkungsabschätzung.SubCategory)="total") AND
((Wirkungsabschätzung.Category)="eco-indicator 99, (E,E)"));
```

These data are, separately, read into R with the usual sqlfetch procedure, e.g. for Cesium-137:

```
d_Cesium <- sqlFetch(channel, "qry_ProcessIO_S_Cesium", colnames = TRUE,
rownames = FALSE, max=0)
```

Afterwards, a QQ-Plot is created to compare the given data (e.g. Cesium-137 emissions) to those of a theoretical, ideal distribution. Since the plot compares normally distributed values, the logarithm needs to be taken into account by taking also the logarithm of the probabilities¹⁴:

```
> qqnorm(log(d_Cesium[[5]]), main = "Normal Q-Q- Plot, Cesium-137,
ecoinvent System Processes")
Warning message:
In log(d_Cesium[[5]]) : NaNs wurden erzeugt
```

Some negative values for the emissions cause errors.

The resulting plot is shown in the next figure. If the data followed a perfect lognormal distribution, the dots would follow the diagonal. This is almost fulfilled (Figure 26).

¹⁴ “d_Cesium[[5]]” refers to the fifth element in the data frame, which is the mean of the emission.

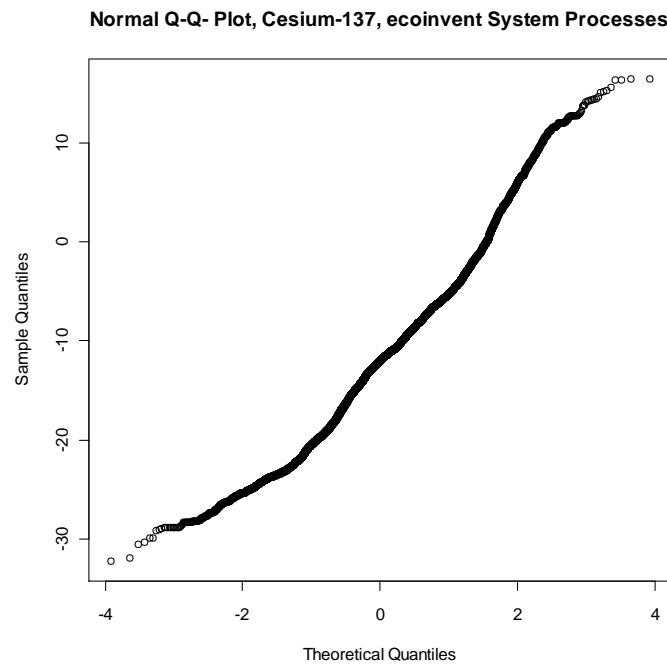


Figure 26: Q-Q- plot of Cesium-137 emissions of all ecoinvent system process, perfectly lognormally distributed data would follow the diagonal

Repeating the exercise for fossil CO₂ generates the following plot:

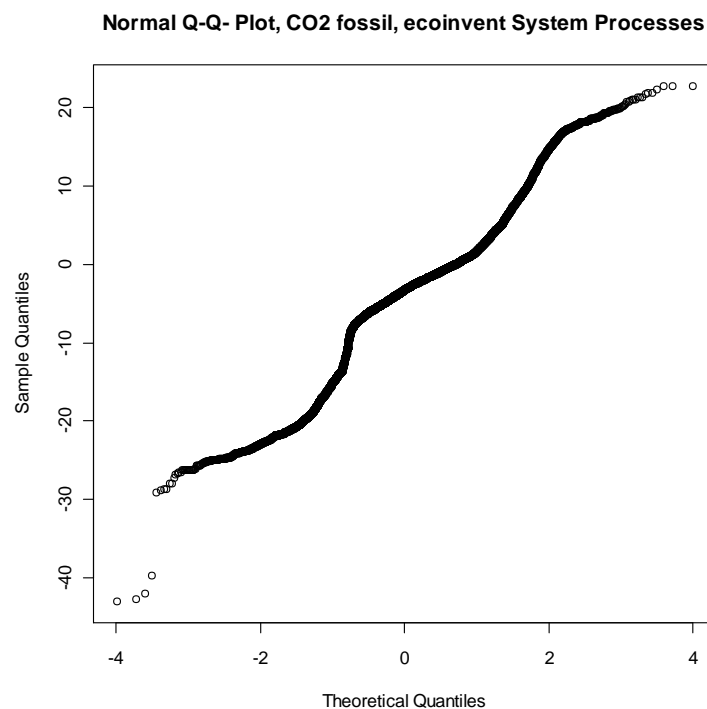


Figure 27: Q-Q- plot of fossil CO₂ of all ecoinvent system processes, perfectly lognormally distributed data would follow the diagonal

Fossil CO₂ emissions in ecoinvent system processes follow to a lesser degree the diagonal in the Q-Q- plot (Figure 27).

To further query the distribution of these two emission flows, the Kolmogorov-Smirnov test for lognormal distribution is performed. It is called by...

```
ks.test(d_Cesium[[5]], "dlnorm")
```

...which produces as output:

```
One-sample Kolmogorov-Smirnov test

data:  d_Cesium[[5]]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(d_Cesium[[5]], "dlnorm") :
kann bei Bindungen nicht die korrekten p-Werte berechnen
```

„Bindungen“ (english: ties) mean that the data contain several times exactly the same values, which causes problems in the ranking procedure of the test. This is a problem once many identical values are in the analysed data. According to the plot above, this does not seem to be the case – and yet according to the test statistics, with an extremely low probability, the data does *not* follow a lognormal distribution.

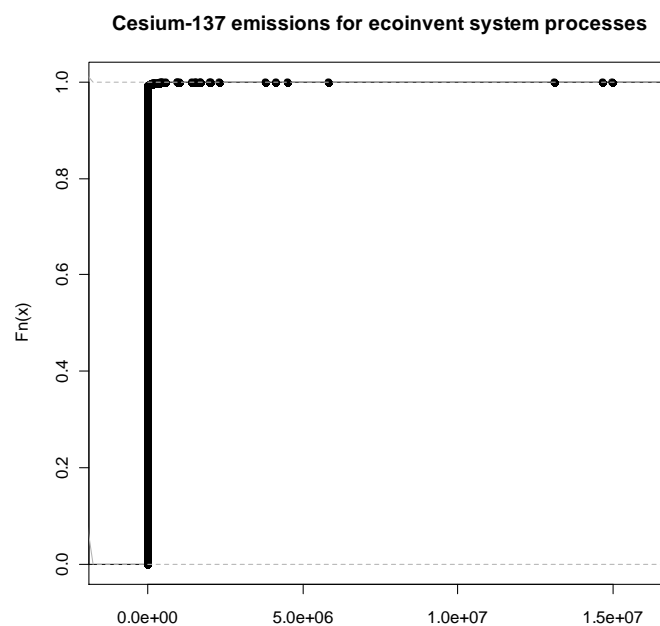


Figure 28: Empirical cumulated distribution function of all ecoinvent system processes, emissions of Cesium-137

Looking at the empirical distribution plot for these data shows several extreme values; for lower values, the distribution function seems one vertical line. The histogram (Figure 29) does not reveal more information, indicating only that the vast majority of values is rather small.

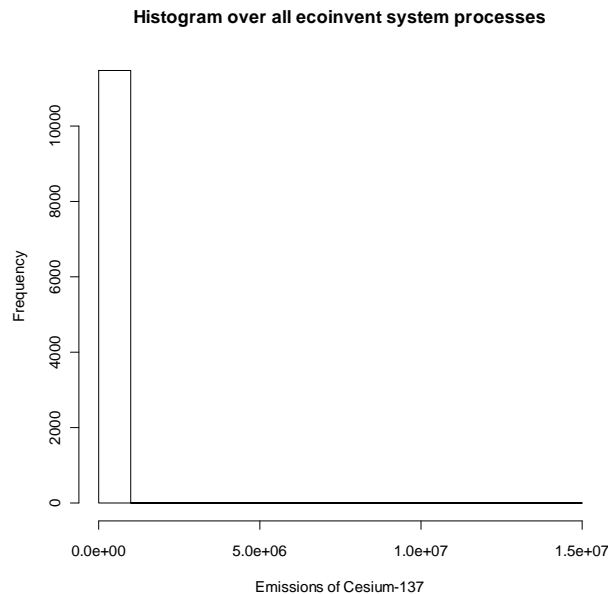


Figure 29: Histogram of the emissions of Cesium-137 of all ecoinvent system processes

The simple stem & leaf plot, however, is able to give a somewhat better impression of the distribution:

```
stem(d_Cesium[[5]])
```

The decimal point is 6 digit(s) to the right of the |

```
-0 | 0000  
0 | 000000000000000000000000000000000000000000000000000000000000+11393  
1 | 00045677  
2 | 003  
3 | 8  
4 | 15  
5 | 8  
6 |  
7 |  
8 |  
9 |  
10 |  
11 |  
12 |  
13 | 11  
14 | 7  
15 | 0
```

It can be seen that only some very few extreme values exist. The highest values are in the area of 1E+7 since they are in the figure at the exponent of 13, and the diagram states that “the decimal point is 6 digit(s) to the right of the |”. By inspection of the raw data, these can be identified as the processes “production plant, natural gas”, “reservoir hydropower plant”, and some others (see the ordered query below).

4.2.6 Regression: CO₂ vs. GWP

In this section, the perception of a strong correlation between CO₂ and GWP, which appeared in the explorative data analysis and which is also logical, is further questioned in a regression.

The general regression in R is called by ‘glm’ (for general linear model). GWP_CO2 is the same data that was used already in section 4.2.1.2, all fossil CO₂ emissions for unit processes process vs. all GWP values according to IPCC. The complete call:

```
regr <- glm(GWP_CO2[[1]] ~ GWP_CO2[[2]])
summary(regr)
```

gives the following output (GWP_CO2[[1]] are the GWP values, GWP_CO2[[2]] are emissions of fossil CO₂):

```
Call:
glm(formula = GWP_CO2[[1]] ~ GWP_CO2[[2]])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-757.78  -12.33  -12.29  -11.96  2375.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.33609    4.55283     2.71  0.00684 **
GWP_CO2[[2]]  0.46305    0.04365    10.61 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 22694.77)

    Null deviance: 27790866  on 1113  degrees of freedom
Residual deviance: 25236583  on 1112  degrees of freedom
AIC: 14339

Number of Fisher Scoring iterations: 2  1 | 2233
```

So, the linear equation is estimated as $GWP [kg] = 12.33609 + 0.46305 * CO_2 [kg]$. ‘t-value’ is the coefficient divided by its standard error. ‘Pr’ is the significance level (the probability that the null-hypothesis: ‘both variables have no effect on each other’ is satisfied). In this case, the null hypothesis can be safely rejected.

We can also have a look at the residuals; from the summary, it is clear that they are not symmetrical. This motivates plotting a histogram:

```
res <- residuals(regr)
hist(res, main = "Histogram of residuals", xlab = "linear regression CO2
vs. GWP, unit processes")
```

It is shown in the next figure. Ideally, residuals should distribute evenly across 0, with a symmetrical, normal distribution. This is obviously not fulfilled, since most of the residuals are negative (Figure 30). This can be further investigated by comparing the predicted to the actual residuals. In an ideal case, residuals and predicted residuals should distribute randomly, and evenly – which is clearly not the case (Figure 31).

Looking at the EDA plot, and at raw data, this is again plausible. CO₂ is emitted by many processes and often dominates GWP results, but other substances contribute also to GWP which are not emitted together with CO₂, but rather separately (N₂O and CH₄ from agriculture, e.g.).

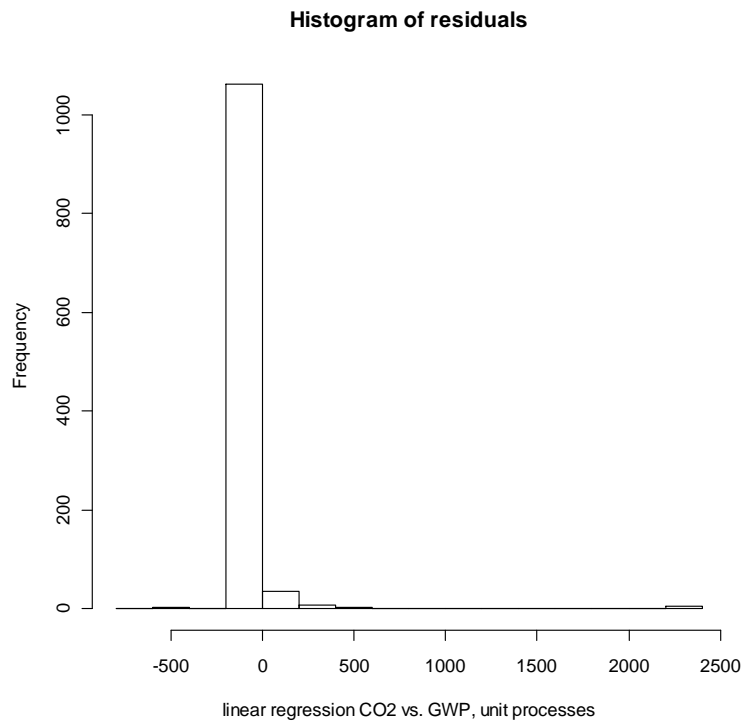


Figure 30: Histogram of residuals of the linear regression between CO₂ and GWP, ecoinvent unit processes

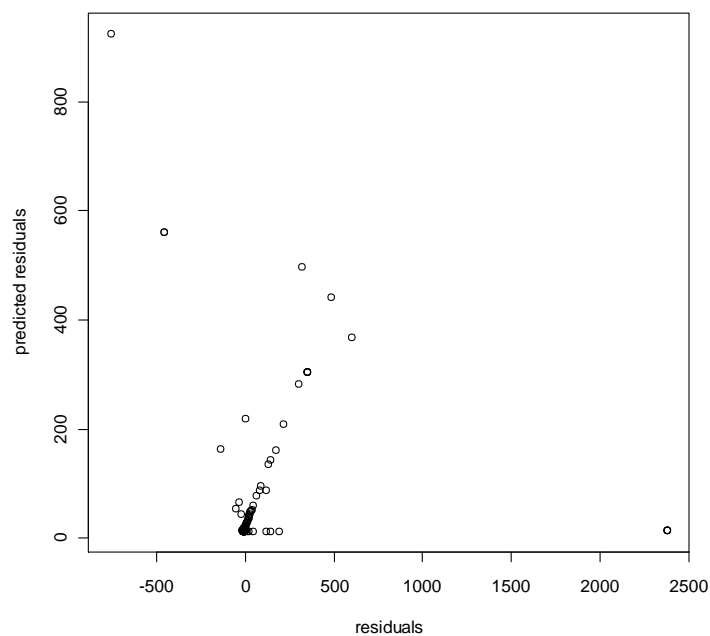


Figure 31: Predicted residuals and actual residuals from a linear regression between CO₂ and GWP, ecoinvent unit processes

4.2.7 Principal Components Analysis: Underlying factors for flows of machinery processes?

A Principal Components Analysis (German: “Hauptkomponentenanalyse”) is a form of a factor analysis; aim is to find factors that explain several given variables, and thereby allow to summarise data, and also to better understand data with the detection of central, “underlying” factors (e.g. [Backhaus 1994, pp. 221]). These underlying factors are typically not measured, either because they are simply not present in the data set, or because they cannot be measured. An example is “health” of an individual, which influences among others weight, productivity, and happiness.

The aim to find factors underlying a data set is common for any factor analysis. The principal components analysis is a special kind of factor analysis; it assumes that all given variables can be fully explained by underlying factors, the principal components. These are independent.

For the ecoinvent data, two analyses are conducted. First, unit process emissions are investigated; and second, it is checked whether impact assessment results may lead to some few impact assessment factors.

Problem for the analysis are the sparsely populated matrices. For example, limiting the analysis to processes with subcategory machinery yields the following matrix:

Process name	LangName	32	33	34	35	36	37	416	444	505	556	557	558
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionCH34-1kgDkg												
slurry tanker, production	slurry tanker, productionCH35-1kgDkg												
tractor, production	tractor, productionCH36-1kgDkg-1C												
trailer, production	trailer, productionCH37-1kgDkg-1C												
conveyor belt, at plant	conveyor belt, at plantCH557-1kgDkg												
diesel, burned in building machine	diesel, burned in building machineCH556-1kgDkg												
industrial machine, heavy, unspec	industrial machine, heavy, unspec												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionCH34-1kgDkg												
slurry tanker, production	slurry tanker, productionCH35-1kgDkg												
tractor, production	tractor, productionCH36-1kgDkg-1C												
trailer, production	trailer, productionCH37-1kgDkg-1C												
conveyor belt, at plant	conveyor belt, at plantCH557-1kgDkg												
agricultural machinery, general, pr	agricultural machinery, general, pr									0,01			
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionCH34-1kgDkg												
slurry tanker, production	slurry tanker, productionCH35-1kgDkg												
tractor, production	tractor, productionCH36-1kgDkg-1C												
trailer, production	trailer, productionCH37-1kgDkg-1C												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionCH34-1kgDkg												
slurry tanker, production	slurry tanker, productionCH35-1kgDkg												
tractor, production	tractor, productionCH36-1kgDkg-1C												
trailer, production	trailer, productionCH37-1kgDkg-1C												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionCH34-1kgDkg												
slurry tanker, production	slurry tanker, productionCH35-1kgDkg												
tractor, production	tractor, productionCH36-1kgDkg-1C												
trailer, production	trailer, productionCH37-1kgDkg-1C												
building machine	building machineCH556-1kgDkg												
power saw, without catalytic conver	power saw, without catalytic conver												
power sawing, with catalytic conver	power sawing, with catalytic conver												
power sawing, without catalytic conver	power sawing, without catalytic conver												
power sawing, with catalytic conver	power sawing, with catalytic conver												
power sawing, without catalytic conver	power sawing, without catalytic conver												
crushing, rock	crushing, rockRER558-1kgDkg												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionRER34-1kgDkg							0,013					
slurry tanker, production	slurry tanker, productionRER35-1kgDkg							0,206					
tractor, production	tractor, productionRER36-1kgDkg-1C												
trailer, production	trailer, productionRER37-1kgDkg-1C												
building machine	building machineRER556-1kgDkg												
conveyor belt, at plant	conveyor belt, at plantRER557-1kgDkg												
crushing, rock	crushing, rockRER558-1kgDkg							0,00000045					
diesel, burned in building machine	diesel, burned in building machineCH556-1kgDkg							0,000514					
industrial machine, heavy, unspec	industrial machine, heavy, unspec												
power saw, with catalytic conver	power saw, with catalytic conver												
power saw, without catalytic conver	power saw, without catalytic conver												
power sawing, with catalytic conver	power sawing, with catalytic conver												
power sawing, without catalytic conver	power sawing, without catalytic conver												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionRER34-1kgDkg												
slurry tanker, production	slurry tanker, productionRER35-1kgDkg												
tractor, production	tractor, productionRER36-1kgDkg-1C												
trailer, production	trailer, productionRER37-1kgDkg-1C												
building machine	building machineRER556-1kgDkg												
conveyor belt, at plant	conveyor belt, at plantRER557-1kgDkg												
crushing, rock	crushing, rockRER558-1kgDkg												
diesel, burned in building machine	diesel, burned in building machineCH556-1kgDkg												
industrial machine, heavy, unspec	industrial machine, heavy, unspec												
power saw, with catalytic conver	power saw, with catalytic conver												
power saw, without catalytic conver	power saw, without catalytic conver												
power sawing, with catalytic conver	power sawing, with catalytic conver												
power sawing, without catalytic conver	power sawing, without catalytic conver												
agricultural machinery, general, pr	agricultural machinery, general, pr												
agricultural machinery, tillage, pro	agricultural machinery, tillage, pro												
harvester, production	harvester, productionRER34-1kgDkg												
slurry tanker, production	slurry tanker, productionRER35-1kgDkg												
tractor, production	tractor, productionRER36-1kgDkg-1C												
trailer, production	trailer, productionRER37-1kgDkg-1C												
building machine	building machineRER556-1kgDkg												

Figure 32: Sparse matrix example: Processes with subcategory machinery and their input and output flows; column labels 32 and following are the IDs of flows. An empty cell means that the respective process data set, in one line, has no entry for this flow

After the above query is read into the workspace of R, the following lines start a principal components analysis, and produce some graphics output:

```
machinery_pca <- prcomp(pca_test_emissions_machinery[11:105])
# rows 1:10 are row headers and need to be excluded
```

```
plot(machinery_pca)
# plots the explained variance

summary(machinery_pca)
# print output statistics

biplot(machinery_pca)
# "factor loading plot" -> compare the identified factors / principal
components with the original data
```

The summary statistics show that many different principal components (PC) remain (95!), but that almost all of the variance is covered by the first two components, PC1 and PC2, with a share of the overall variance of 0.89 and 0.097 respectively. This can again be seen from the plot of the explained variances (Figure 33).

Importance of components:																									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	
PC25																									
Standard deviation	9534.68	3146.297	8.74e+02	7.26e+02	1.91e+02	95.32415	66.59979	5.29	3.8	2.03	1.58	1.01	0.474	0.432	0.375	0.245	0.224	0.134	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	
Proportion of Variance	0.89	0.097	7.48e-03	5.16e-03	3.60e-04	0.00009	0.00004	0.00	0.0	0.00	0.00	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
Cumulative Proportion	0.89	0.987	9.94e-01	1.00e+00	1.00e+00	0.99996	1.00000	1.00	1.0	1.00	1.00	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
PC48																									
Standard deviation	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.094	0.0845	0.0332	0.0317	0.0202	0.0186	0.0176	0.00947	0.00692	0.00647	0.0047	0.00391	0.00222	0.00207	0.00182	0.0014			
Proportion of Variance	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.00000	0.00000	0.00000	0.0000	0.00000	0.00000	0.00000	0.00000	0.00000	0.0000	0.0000	
Cumulative Proportion	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.00000	1.00000	1.00000	1.0000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.0000	
PC67																									
Standard deviation	0.00094	4.5e-05	1.54e-05	8.53e-06	1.27e-06	4.16e-07	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	
Proportion of Variance	0.00000	0.0e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	
Cumulative Proportion	1.00000	1.0e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	
PC86																									
Standard deviation	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	
Proportion of Variance	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	
Cumulative Proportion	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	
PC87																									
Standard deviation	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.00e-12	1.42e-15	7.42e-16	1.13e-16																
Proportion of Variance	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	
Cumulative Proportion	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	

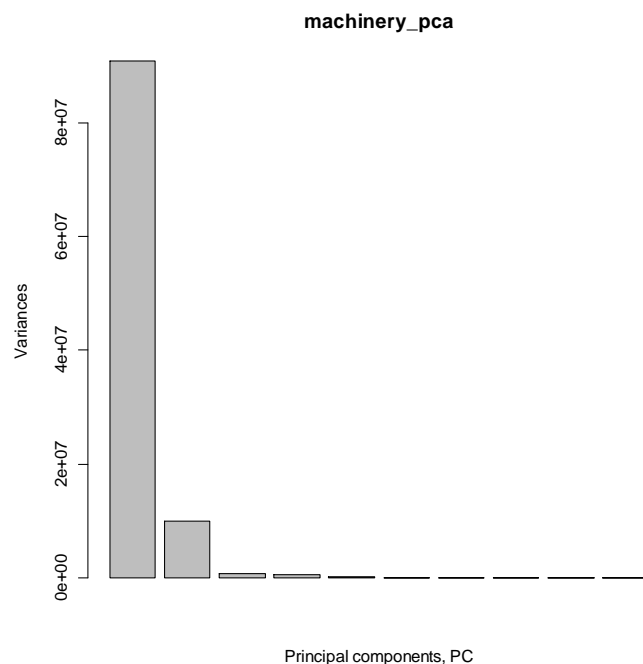


Figure 33: Explained variance of the different principal components of the Principle Components Analysis, for emissions of all unit processes in the subcategory “machinery”

A plot of the principal components against the real data, the process emissions, shows that – somewhat surprisingly – only two emissions dominate the factors, namely flows with ID 2983 and 1564. They correlate almost directly with the PCs (Figure 34).

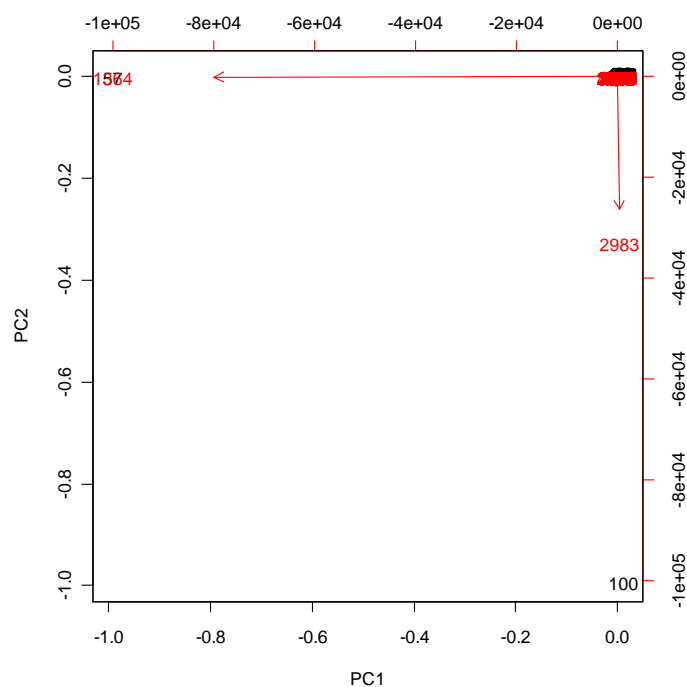


Figure 34: “Biplot”, plot of the detected principal components (in black) against real values (in red)

Other PCs are not important; a warning indicates that they cannot be drawn in the plot since their arrows have no length (which fits to the finding that two PCs explain almost all variance).

```
Es gab 50 oder mehr Warnungen (Anzeige der ersten 50 mit warnings())

> warnings()

Warnmeldungen:

1: In arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = col[2], ... :
  Pfeil ohne Länge hat keine Richtung und wird ignoriert
2: In arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = col[2], ... :
  Pfeil ohne Länge hat keine Richtung und wird ignoriert
3: In arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = col[2], ... :
  Pfeil ohne Länge hat keine Richtung und wird ignoriert
4: In arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = col[2], ... :
  Pfeil ohne Länge hat keine Richtung und wird ignoriert
5: In arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = col[2], ... :
  Pfeil ohne Länge hat keine Richtung und wird ignoriert
[...]
```

The identified flows are waste heat and petrol, at refinery:

FID	category	subCategory	localCategory	localSubCategory	CASNumber	name	location	unit	formula	generalComment	localName	infrastructureP
2564	oil	fuels	Erdöl	Brenn- und Treibstoffe	null	petrol, low-sulphur, at refinery	CH	kg	null	null	Benzin, schwef	
2983	air	unspecified	Luft	allgemein	null	Heat, waste	null	MJ	null	(1,1,2,1,1,3);	Abwärme	

While this is a strong and also plausible result, the analysis might be obfuscated by the different functional units of the processes. “1 kg” is predominant, but also “1 unit” (building machine), “1 m” (conveyor belt), “1 h” (power sawing), and “1 MJ” (diesel, burnt in machine) exist.

If indeed strong underlying factors are detected as in the example here, these can be used for further analysis in various forms, from plausibility checks (e.g. ratio of one component to another) to other statistical analysis.

Process_unit	infrastructureInc	geoLoc	Process_name	LangName	32
kg	1	CH	harvester, production	[harvester, productionRER34-1tkm0	
kg	1	CH	slurry tanker, production	[slurry tanker, productionRER35-1tk	
kg	1	CH	tractor, production	[tractor, productionRER36-1tkm0kg	
kg	1	CH	trailer, production	[trailer, productionRER37-1tkm0kg	
unit	1	RER	building machine	[building machineRER556-1tkm0un	
m	1	RER	conveyor belt, at plant	[conveyor belt, at plantRER557-1tk	
kg	1	RER	crushing, rock	[crushing, rockRER558-1tkm0kg0R	
kg	1	RER	industrial machine, heavy, unspec	[industrial machine, heavy, unspeci	
unit	1	RER	power saw, without catalytic conve	[power saw, without catalytic conve	
h	1	RER	power sawing, with catalytic conve	[power sawing, with catalytic conve	
h	1	RER	power sawing, without catalytic co	[power sawing, without catalytic co	
MJ	1	GLO	diesel, burned in building machine	[diesel, burned in building machinef	
unit	1	RER	power saw, with catalytic converte	[power saw, with catalytic converter	
h	1	RER	power sawing, with catalytic conve	[power sawing, with catalytic conve	
h	1	RER	power sawing, without catalytic co	[power sawing, without catalytic co	
unit	1	RER	building machine	[building machineUCTE556-1kWWh0	
unit	1	RER	power saw, without catalytic conve	[power saw, without catalytic conve	
kg	1	CH	agricultural machinery, general, pr	[agricultural machinery, general, pr	1
kg	1	CH	agricultural machinery, tillage, pro	[agricultural machinery, tillage, pro	
kg	1	CH	harvester, production	[harvester, productionCH340kg-1kg-	
kg	1	CH	slurry tanker, production	[slurry tanker, productionCH350kg-	
kg	1	CH	tractor, production	[tractor, productionCH360kg-1kg-1C	
kg	1	CH	trailer, production	[trailer, productionCH370kg-1kg-1C	
MJ	1	GLO	diesel, burned in building machine	[diesel, burned in building machineC	
kg	1	CH	agricultural machinery, general, pr	[agricultural machinery, general, pr	
kg	1	CH	agricultural machinery, tillage, pro	[agricultural machinery, tillage, pro	
kg	1	CH	harvester, production	[harvester, productionnull340kg0kg-	
kg	1	CH	slurry tanker, production	[slurry tanker, productionnull350kgC	
kg	1	CH	tractor, production	[tractor, productionnull360kg0kg-1C	
kg	1	CH	trailer, production	[trailer, productionnull370kg0kg-1C	

Figure 35: Differing functional units (left column) for processes of subcategory machinery as problem for the analysis

5 A summary of analyses results, and of evaluation criteria

This report has so far described many different statistical analyses, and some basic plausibility checks. Implementation in R and / or Access queries is documented as well.

The plausibility checks and their *possible* use¹⁵:

Plausibility check	Intention, result
Process mass balances	Processes where input does not equal output are suspicious
Input output flow comparison against other processes	Check for omitted flows
Negative flow amounts	Negative flow amounts are suspicious
Calculation of basic indicators, such as the emissions CO ₂ per amount of fuel consumed	Per fuel, the amount of emitted CO ₂ should constant

Although these are basic checks, their application showed several possible flaws in data (as the negative transport for the chain saw).

¹⁵ Plausibility checks should be applied for aspects that a) can be checked rather easily and do not require detailed analyses with many prerequisites and assumptions, and b) are in line with the methodological conventions for ecoinvent. The latter is of course not point for this report, the intentions listed in the table (as the intention to have for every process a complete mass balance) are suggestions only.

The statistical analyses and their uses, divided into the assessment of a group of datasets, and of a single data set:

Analysis, procedure	Intention, result, group of datasets	Intention, result, single dataset
Boxplots	Show the distribution of values, including distribution parameters	Show the values of the single data set in comparison to a “peer group”
Scatter plots, matrix scatter plots	Show the distribution of values	Show the values of the single data set in comparison to a “peer group”
Cluster analysis	Find groups of datasets that are within the group homogenous, and differ largely from one group to another	Compare if the dataset belongs to the group that is expected
Heatmap	Combination of two cluster analyses, for classification of data into two types of groups	Compare if the dataset belongs to the group that is expected
Distribution tests	Knowing the probability distribution of data is a pre-requisite for several statistical tests; it is necessary for uncertainty calculation via Monte Carlo simulation and also via approximation	-
Regression	Extrapolation and interpolation of missing / non-existing values in general	Compare values of a specific dataset to a general regression function valid for a group of datasets
Principal components analysis	Find “principal components” that are each independent and that explain the information provided in the datasets better than the original variables	-

The analyses and investigations that are called ‘statistical’ here belong mostly to the field of exploratory data analysis, EDA. This has several reasons. First, EDA is for every analysis of data a natural first start. More sophisticated analyses can be tailored to the properties of the data as discovered by EDA. It can then also be checked in the EDA whether assumptions and requirements of these analyses are fulfilled. Further, during the EDA questions occur that motivate further analyses. Second, and more important for the present report, is that any statistical analysis profits from questions that are to be answered. These questions need to be developed, with the ecoinvent centre, and possibly with users. Finally, and third, life cycle inventory data are peculiar; many statistical procedures apply a weighting across different realisations of variables. This needs to be carefully considered in LCA. Few data are at present obtained in a truly statistical sampling procedure. Creation of “the one” fitting data set for one specific case is more common. As consequence, there is not really a random sample of data, but a collection of carefully created individual data spots, with many missing values when put together in a matrix, which is not straightforward to analyse. Finally, some of the desirable analyses, such as an elementary balance, are barely doable for many elements, with the current data structure.

To sum up, the field of statistical and mathematical analysis is certainly wide; it takes considerable effort to find a procedure that fits best for the analysis of life cycle assessment data in a large database as ecoinvent. The analyses conducted show, however, that with rather basic procedures, flaws in data can be detected. Statistical analyses provided plausible results,

fostered also the detection of possible flaws in data and thereby data quality, and allowed to group data into clusters, to detect and test probability distributions, and to find relations between data.

6 Towards a practical application for ecoinvent data

Procedures and queries applied here in the report are documented and also separately available. The combination of an Access database and R proved to be flexible and able to solve all kinds of analyses requirements.

However, for practical application, two additional points are of importance. First, it will be necessary to specify a workflow for mathematical analyses of ecoinvent data that comprises an understanding of which procedures need to be taken, by whom, for data sets that are new for ecoinvent, and also about maintenance of data that exists already in ecoinvent. Second, it will be important to put an easy-to-use, and to some degree fault-tolerant, system in place, that fits to the needs of ecoinvent.

6.1 *Towards a learning system of data mining and quality assurance*

The different checks and analyses described so far are obviously related. To define a basic check, insight about “how the checked data should be” is needed (e.g., can it be expected that mass input equals mass output, for a process? And is this required?). This knowledge is either “given” externally (e.g. by methodological conventions), or it may be obtained from other analyses (e.g. by statistical confirmatory tests), or it may in parts use results from other analyses (e.g., process categories that are used in a check that are obtained in turn from cluster analysis). Analyses build upon each other, or results from one analysis call for other analyses (e.g., some analyses assume probability distributions that need to be checked).

A workflow should further consider different users and people with different backgrounds and different access rights. Also, maintenance for the system put in place should be foreseen.

6.1.1 **A general procedure for the development and maintenance of a data mining and quality assurance system**

The question of how to bring the different procedures together in a meaningful and smart system, how to develop it, and how to integrate such a system into an environment of data administration, collection, and data use, is not at all unique to ecoinvent, but rather bobs up in any practical data application.

Many different ideas for a workflow and structure of such a system exist; books on data mining and analysis frequently contain (differing) proposals (e.g. Peterson 2005, p. 13; Olson Delen 2008, pp. 8-22; Jarke et al. 2000, p. 53).

Among the different proposals, the “CRISP-DM” process seems to fit best for the development and maintenance of a data quality assurance and analysis system for ecoinvent (Olson Delen 2008, p. 9; CRISP 2000). The process consists of the following main steps (Figure 36):

1. Business understanding: Goal and scope of the analysis
2. Data understanding: knowledge about properties of the given data set (structures in data, interesting subsets, possible errors, intended applications, ...)
3. Data preparation: Preparation and possible transformation of data according to requirements of foreseen data analyses
4. Modelling: Modelling can be seen as central for the whole process; different types of analyses and procedures are selected, tailored to the problem (as specified in 1 and 2,

Business and data understanding), integrated into an overall “composition”, the model, which is also assessed from a data analysis perspective, and modified if needed.

5. **Evaluation:** this step comprises a more thorough evaluation of the model, specifically it is investigated whether goal and scope of the analysis are met by the model.
6. **Deployment:** this step is about the “delivery” of the model to users of potentially varying interest, and may include simple report generation or also specifically tailored data mining maintenance procedures.

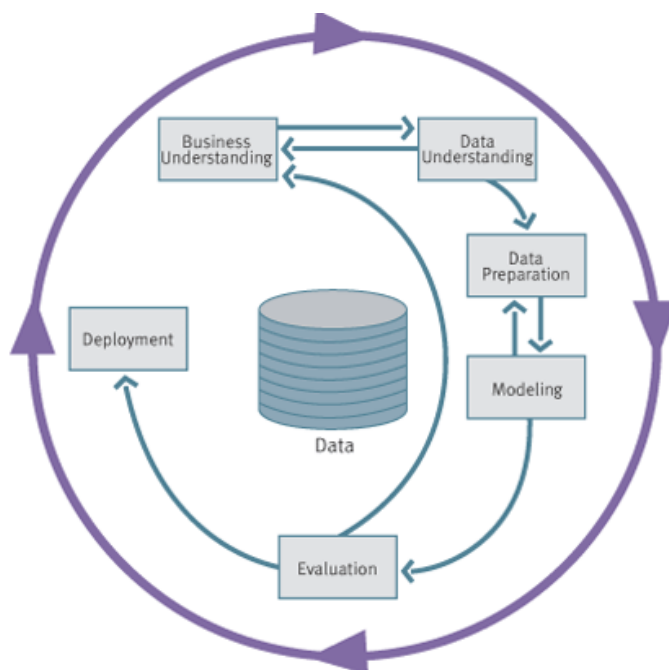


Figure 36: CRISP-DM process (Olson Delen 2008, p. 9; CRISP 2000)

The whole process is not a linear sequence but rather a continuous improvement and learning process. The different steps are related, and linked in feedback loops. Only main connections are shown in the figure, others exist as well (e.g., there will also be a feedback from Evaluation to Modelling, and from Modelling to Data Understanding).

The CRISP-DM process does not distinguish between different types of analysis and data evaluation tasks (basic checks – more refined analyses). However, its steps can be “translated” into the following steps for ecoinvent:

1. **Business understanding:** Goal and scope of the analysis (also for ecoinvent!) – goal and scope should be put forward by the ecoinvent management and based on methodical guidance handbooks; also, possibilities of data analysis, e.g. based on findings in this report, should be consulted.
2. **Data understanding:** Also this point remains unchanged for ecoinvent – knowledge about properties of the given data set (structures in data, interesting subsets, possible errors, intended applications, ...), obtained from methodological guidelines, expert knowledge, and from evaluation of the model (5.).
3. **Data preparation:** As an experience from the mathematical analyses, it will be useful to transform the functional unit of processes so that the functional units are comparable, as far as possible (convert every functional unit into 1 kg where possible, and flag the remaining processes so that they are excluded from the analysis). Further

- preparations might be needed depending on goal and scope, e.g. to meet inheritance requirements or tests for the analysis whether inheritance requirements are met.
4. **Modelling** for ecoinvent: Based on goal and scope, compose a set of methods and analysis, for different addresses and use cases. This will be detailed later in this chapter.
 5. **Evaluation**: Analysis of the whole model, from different viewpoints, and according to goal and scope.
 6. **Deployment**: this step is about the “delivery” of model results to users of potentially varying interest, and may include simple report generation or also specifically tailored data mining maintenance procedures. Also this step will be detailed later in this section.

The CRISP-DM process is interesting because it describes the approach to arrive at a data analysis tool implemented in practice. While this report cannot detail how exactly a model and learning system put in place will look like, it seems at least of value to look into the procedure to arrive at such a system. In addition *some* aspects of a likely architecture will be proposed, based on current understanding. In order to do so, the system that is implemented at present inecoinvent needs to be considered (Figure 37).

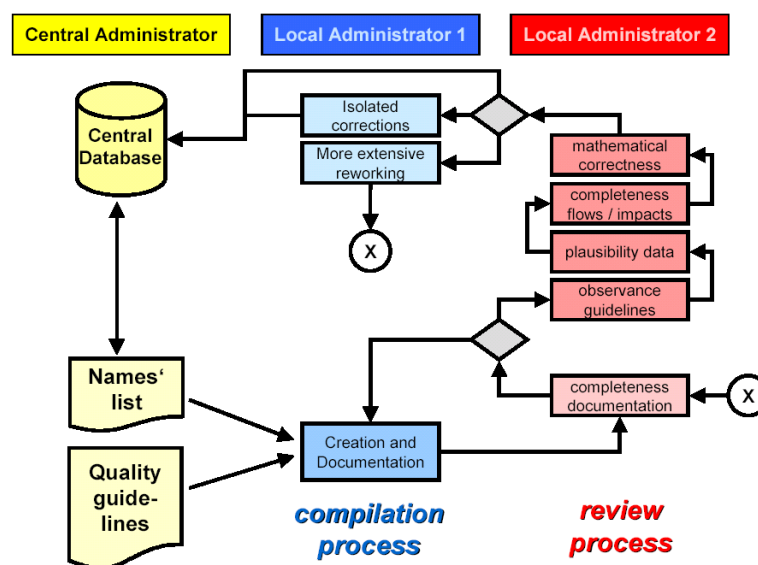


Figure 37: Overview of the internal review and data quality control within the ecoinvent project (Frischknecht and Jungbluth 2003, p. 54)

The ecoinvent system view is completely different from the “CRISP” view; it shows the procedures and different actors in a developed system instead of the development and maintenance of such a system.

Questions for ecoinvent: What are goal and scope for the model that is to be implemented (methodical conventions in data, applications for data, users of the system, principle maintenance) ?

6.1.2 Elements of a learning system

How could a learning system, as mentioned in the last paragraph of section 7, look like? In order to explore this question, let's first look at elements of such a system.

The following elements might exist:

Mathematical analyses and procedures:

- basic checks (as described in section 4.1)
- explorative data analyses (as described in section 4.2.1)
- in-depth statistical analyses (section 4.2)

Knowledge:

- data pattern and data structure knowledge (as result from tests, from other statistical analyses, and also from general expert knowledge)

Data:

- ecoinvent internal data pool with incoming and changing data sets, probably with version numbers
- ecoinvent released data pool

People:

- users with feedback concerning released data
- data generators with new data sets for ecoinvent
- “data management unit”, responsible for ecoinvent releases and for error corrections
- data reviewers and validation experts
- ecoinvent resellers (?)

In principle there are also mathematical procedures that are “inherently” learning, as for example neural networks, genetic algorithms, or other approaches. It would be an interesting research option to look into them (see e.g. Liu 2007), but for the present task they seem somewhat out of scope.

Therefore, it is proposed that the learning system should be a system that is an advanced “data management system” for ecoinvent, where mathematical analyses, in various forms, are an important element. The current review and quality assurance system in place for ecoinvent should be used as a starting point.

In order to compose, from the mentioned elements, an overall sound advanced data management system, the following principles should be followed:

- efficiency: Detect errors in data fast, and with least effort possible. This will often mean that errors should be detected early, for example while the data set is entered into an editor instead of after it is merged with all other data in ecoinvent.
- double checks: For important decisions, ideally more than one single analysis should be conducted. This does not hold for basic checks as e.g. balance checks, which provide an unambiguous result.
- “expertise”: Analyses should provide their results in a context where the results can be understood. LCA practitioners, for example, might not feel comfortable if they interpret certain multi-variate statistical analyses. Of course, training can help alleviate this to some extent.

The following use cases can be distinguished:

- Small review: a small amount of new data as candidate for ecoinvent data
- Large review: a large amount of new data, from one source, as candidate for ecoinvent data
- Stock-take: thorough investigation of the complete data stock
- Screening: light investigation of the complete data stock, e.g. after integration with new data

- Single error detection: by user feedback, an error in a data set is discovered (and it is not sure whether this error is also relevant for other datasets).
- Data background change: A new database or format structure
- Knowledge change: refinement of data structure knowledge and of the tests and analyses

Questions for ecoinvent:

Are these the elements, especially people / users, and data, that need to be taken into account (according to goal and scope)?

Are these the foreseen use cases? How often occur these?

6.1.3 Analyses and their results and use, in a learning system

How are the elements as outlined in the previous section related? Especially interesting is the relation of basic checks, explorative data analysis, and (confirmatory) tests.

This report starts with basic checks, and continues to explorative data analysis, and to further statistical tests. This can be criticised as basic checks often base on results found by the other types of analyses, described later; on the other side, basic checks follow method knowledge that is most likely be clear beforehand. Therefore there is no natural order of analysis and checks.

On the other hand, certain analysis will be used “en bloc” and will therefore build a sequence. And for several of these sequences, specific addressees can be assumed.

For example, a principal components analysis could be used to find underlying factors in data, that then are put into a cluster analysis, which yields groups of processes.

For a later application, more refined sequences will need to be defined. As a start, tentative analysis groups, their results, and uses for these results, are proposed in the table below.

Analysis	Result, used for	Applied by
Data transformation: adaption of functional unit; solving other known issues to enable further analysis (as might be: adding material flows as air or water that were omitted on purpose); preparation of data to enable further checks (as might be: re-calculating an allocation)	Analysis data basis, used for subsequent analyses	Admin
EDA: Explorative Data Analysis (various different investigations, from scatter plots to tentative cluster analyses)	Insight into data properties and structures, used for subsequent analyses / hypothesis formulation for tests / definition of plausibility checks	Admin
Group building analyses (cluster analysis)	Groups obtained by data properties, used for checks (is a process in the cluster where it was expected?), and for data display	Admin
Tests for detecting functional relationships	Functional relation (e.g., linear curve), used for plausibility checks and for inter- and extrapolation	Admin
“EDA light”: Selected and pre-defined EDA; also combinations	Display of data also in relation to other data, used for visualisation	Reviewers, data users

Analysis	Result, used for	Applied by
with identified data groups and functional relations (various different investigations, from scatter plots to tentative cluster analyses)	and plausibility checks and other purposes	
Simpler “service type” tests (type of probability distribution, independence of two data samples, ...)	Test result, used usually in order to check the assumptions of another analysis	Reviewers (selection, e.g. probability distribution), data users, admin
Basic checks (of various forms: negative mass flows; correct input or output group of elementary flows, mass balance, ...)	Results of the basic checks used for quality assurance (warning if check fails)	Reviewers, data providers, admin

Questions for ecoinvent:

Principal question is of course whether these analysis, result, and addressee combinations make sense from an ecoinvent viewpoint. It remains to be answered in the course of developing the system though which analyses in detail should be carried out, for which specific questions (as e.g. CO₂ emissions per MJ fuel consumed). These questions can better be formulated once goal and scope and users and the data stock are known, i.e. once the answers to questions from the previous two sections have been found.

6.2 Integration of mathematical analyses in a test suite?

The requirement for a fault tolerant system is not easily met by the combination of R and Access as it is used for the analyses in this report. Especially the R command-line language, with its smart object oriented concept, may be difficult to use for non-experts. Mistakes are possible¹⁶, not only due to the complexity of the language but also due to possible pitfalls inherent in statistical tests and analyses.

It is therefore highly recommended to integrate those analyses, and “analysis pathways” as combination of different kinds of analyses that build on each other, in an easy to use tool or test suite. The tool should be accompanied with documentation and explanation, of the analyses’ goal and scope, and also of results that will be produced by the analyses. This suite could be provided to users in a similar way as the current EcoSpold tool. It could also be provided LCA software developers, in order to enable an integration in LCA software that uses ecoinvent and also other LCA databases.

6.2.1 Available statistical and data analysis software suites

Several tools for statistical analyses of data exist that promise a similar integration of statistical analyses. Talend, kettle pentaho, rapidminer, and an example from biostatistics, will be briefly discussed.

Talend is a powerful tool for business-modelling and ETL¹⁷. Although it is sometimes announced as being (also) a data mining tool, this is clearly not the main aim. Statistical functions are lacking¹⁸.

¹⁶ One example: The code: `„WK_ges[is.na(WK_ges)] <- 0“` replaces all missing values in matrix WK_ges by 0

¹⁷ Extract, Transform, Load

¹⁸ It would however be interesting to model review schemes, and database and dataset update and maintenance schemes, with Talend.

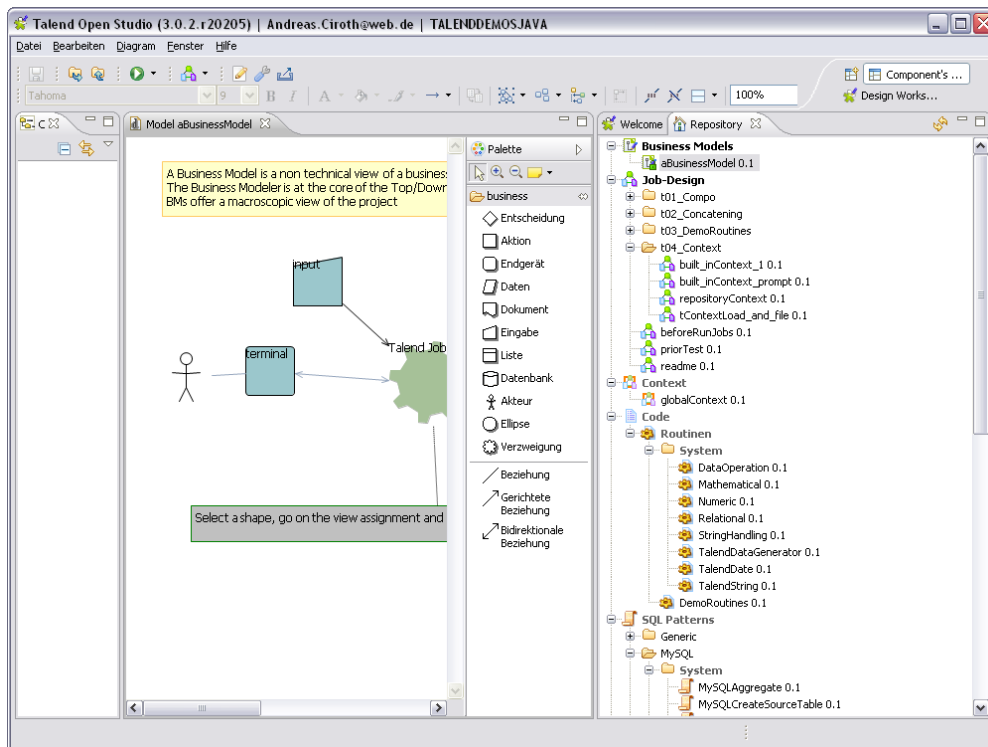


Figure 38: Talend Open Studio, screenshot

Kettle pentaho (www.pentaho.com) is also an ETL tool. In contrast to Talend, it offers basic data mining capabilities.

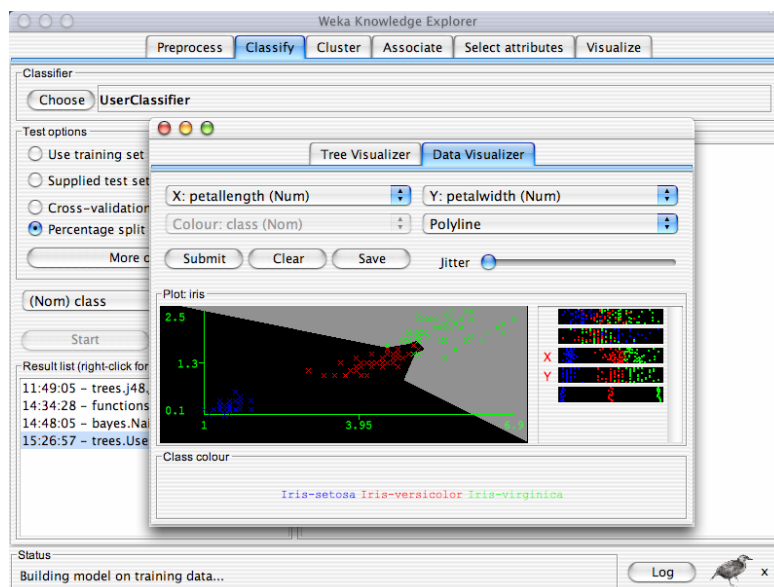


Figure 39: kettle pentaho, screenshot

Rapidminer (www.rapidminer.com) is a tool specifically developed for statistical analyses and data mining. The “operator tree” is a visualisation of the pathway of an analysis; single tests and analyses can be dragged and dropped in this tree, and conveniently deleted. Some of the single elements are highly complex – genetic algorithms and neural networks are available via mouseclick.

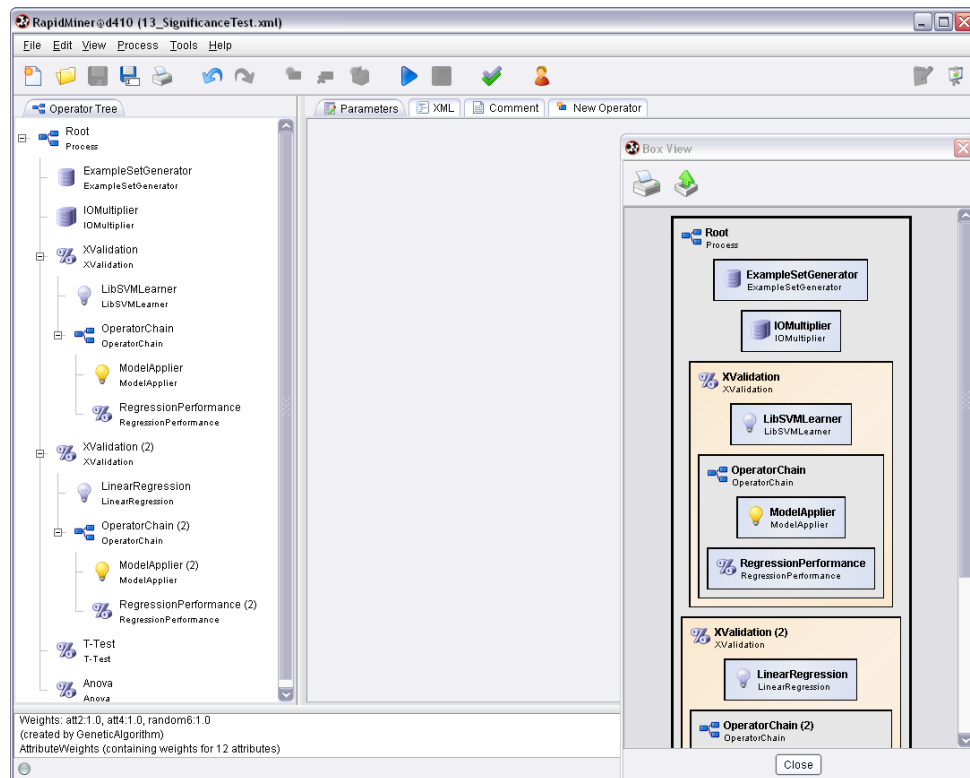


Figure 40: rapidminer, screenshot

An older example from biostatistics [Gentleman et al. 1994] shows yet a different way. The so-called “Bioconductor” project provided open software for a new field of science, with abundant example data files, and with example programs in R; it is often cited as a reference for application of a certain statistical method. What is interesting is the focus on research and method development *and* tool development.

Since LCA data are also different from other scientific data used in statistical examples, the latter approach is interesting for LCA: Also in LCA, method development should be first. Implementing in a powerful software as R will provide the tools for learning, teaching and application purposes.

All software packages discussed here are free and open source; all are based on the copyleft GPL licence which makes a distribution with commercial, closed source software difficult.

6.2.2 Using R in Eclipse

R can also be used from within Eclipse, via the so-called “StatET” plugin (www.walware.de/goto/statet). An additional R implementation is needed. Difference to using R directly is a more comfortable editor, with syntax highlighting, an outline, and variable recognition.

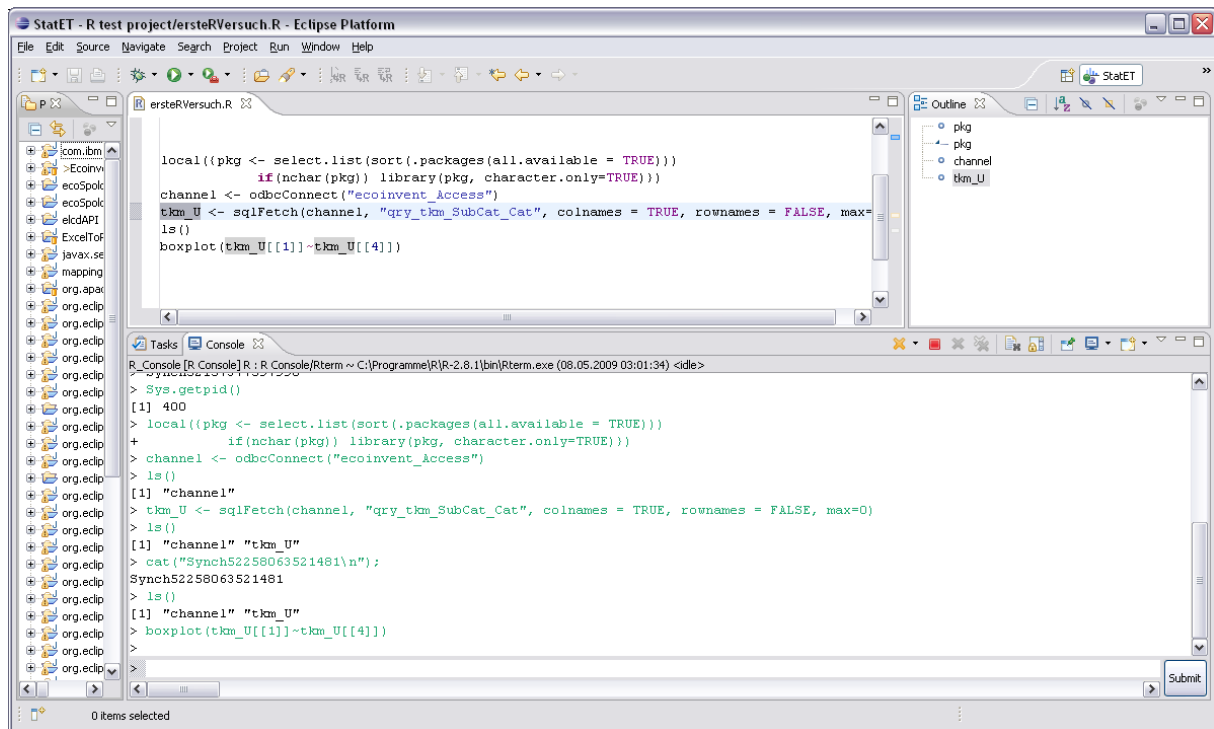


Figure 41: using R via Eclipse, screenshot; ‘ErsteRVersuch.R’: Editor for entering R code; ‘Outline’: Outline of all commands in the editor; ‘Console’: Text output of R

Also, R-files can be created and run from within Eclipse applications. This opens interesting possibilities for combining R with openLCA (www.openlca.org), which is an LCA software also written in Eclipse. The open source licence of the StatET Plugin (Eclipse Public Licence) fits to Eclipse and is compatible with the MPL licence of openLCA:

Any graphical output of R is displayed in separate windows, similar to the R application.



Figure 42: Using R via Eclipse, screenshot; Boxplot figure opens in a separate window

7 Conclusions and outlook

The field of statistical and mathematical analysis is certainly wide; it takes considerable effort to find a procedure that fits best for the analysis of life cycle assessment data in a large database as ecoinvent. The analyses conducted show, however, that with rather basic procedures, flaws in data can be detected. Statistical analyses provided plausible results, fostered also the detection of possible flaws in data and thereby data quality, and allowed to group data into clusters, to find relations between data, to detect and test probability distributions.

For a future application, several “use cases” should be distinguished. The result summary (section 5) distinguishes between focus on one single data set (this could be user or reviewer) and someone with focus on the whole database or on important parts of it. It is recommended to further think about, and discuss, possible use cases and to tailor data quality assessment tools for these cases.

New data sets and modifications and extensions of the information provided with each data set will require a follow-up also for data quality assurance tools. Some of the newly discussed features put much more emphasis on mathematical tools.

Finding a way to group processes together in clusters seems an important step. It can even be that an industrial-sector oriented system (as ISIC, or NACE) and a result-oriented classification system are used in parallel, allowing to switch from one to another.

In the long run, a learning system can be imagined that can cope better and better with detection of data flaws, and with user requests. Chapter 6 deals with some of the questions

that need to be answered in order to arrive at such a system, including an overall procedure for its development and maintenance.

As practical addition to this general outlook, tests and analyses should be applied to the new ecoinvent 2 format database once it is available.

8 References

- CRISP 2000: CRISP-DM 1.0, Step-by-step data mining guide, 2000, retrieved from <http://www.crisp-dm.org/CRISPWP-0800.pdf>, June 2009
- Drobics 2005: Drobics, M.: Data Analysis Using Fuzzy Expressions, Schriften der Johannes Kepler Universität Linz, Reihe C: Technik und Naturwissenschaften, 2005
- Gentleman et al. 2004: Gentleman, R., et al.: Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology* 2004, **5**:R80doi:10.1186/gb-2004-5-10-r80, <http://genomebiology.com/2004/5/10/R80>
- Frischknecht, R., N. Jungbluth, H.-J. Althaus, G. Doka, R. Dones, R. Hischer, S. Hellweg, T. Nemecek, G. Rebitzer, and M. Spielmann: Overview and Methodology. CD-ROM Final report ecoinvent 2000 No. 1, Swiss Centre for Life Cycle Inventories, Dübendorf (Switzerland), 2004
- Hartung 1993: Hartung, J.: Statistik, Oldenbourg, 1993
- Jarke et al. 2000: Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses, Springer, 2000
- Liu 2007: Liu, B.: Web Data Mining, Springer, 2007
- Olson Delen 2008: Olson, D.L.; Delen, D.: Advanced Data Mining Techniques, Springer, 2008
- Petersohn 2005: Petersohn, H.: Data Mining, Oldenbourg, 2005
- Sachs 1992: Sachs, L.: Angewandte Statistik, Springer 1992.
- Schnell 1994: Schnell, R.: Graphisch gestützte Datenanalyse, Oldenbourg, 1994.
- Scime 2005: Scime, A.: Web Mining, Idea Group Publishing, 2005
- Tukey 1977: Tukey JW: Exploratory Data Analysis, Addison Wesley, Reading
- Venables Ripley 1994: Venables, W.N.; Ripley, B.D.: Modern Applied Statistics with S-Plus, Springer, 1994.
- Wang Fu 2005: Wang, L.; Fu, X.: Data Mining with Computational Intelligence, Springer, 2005