Anal Bioanal Chem (2004) 380: 467–474
DOI 10.1007/s00216-004-2794-8

ORIGINAL PAPER

Kristina Voigt · Rainer Brüggemann · Stefan Pudenz

Chemical databases evaluated by order theoretical tools

Received: 13 July 2004 / Revised: 29 July 2004 / Accepted: 29 July 2004 / Published online: 18 September 2004

Abstract Data on environmental chemicals are urgently needed to comply with the future chemicals policy in the European Union. The availability of data on parameters and chemicals can be evaluated by chemometrical and environmetrical methods. Different mathematical and statistical methods are taken into account in this paper. The emphasis is set on a new, discrete mathematical method called METEOR (method of evaluation by order theory). Application of the Hasse diagram technique (HDT) of the complete data-matrix comprising 12 objects (databases) × 27 attributes (parameters + chemicals) reveals that ECOTOX (ECO), environmental fate database (EFD) and extoxnet (EXT)—also called multi-database databases—are best. Most single databases which are specialised are found in a minimal position in the Hasse diagram; these are biocatalysis/biodegradation database (BID), pesticide database (PES) and UmweltInfo (UMW). The aggregation of environmental parameters and chemicals (equal weight) leads to a slimmer data-matrix on the attribute side. However, no significant differences are found in the ''best'' and ''worst'' objects. The whole approach indicates a rather bad situation in terms of the availability of data on existing chemicals and hence an alarming signal concerning the new and existing chemicals policies of the EEC.

Keywords Chemometrics · Environmetrics · Hasse diagram technique (HDT) · METEOR · Environmental chemicals · Environmental chemical databases

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00216-004-2794-8

K. Voigt (✉)
GSF - National Research Center for Environment and Health, Institute of Biomathematics and Biometry, 85764 Neuherberg, Germany
E-mail: kvoigt@gsf.de

R. Brüggemann
Leibniz-Institute of Freshwater Ecology and Inland Fisheries, 12587 Berlin, Germany

S. Pudenz
Criterion Evaluation and Information Management, 10999 Berlin, Germany

Environmental chemicals' data and databases

Achievement of sustainable development in the chemicals sector is the main objective of the Commission's white paper on a future chemicals strategy. The current number of existing substances marketed in volumes above 1 ton is estimated at 30,000 [1]. These substances amount to more than 99% of the total volume of all substances on the market. In the so-called white paper, the paper on the strategy for a future chemicals policy of the commission of the European communities, the testing and evaluation of a large number of existing substances in the coming 10 years is envisaged. Initiative was taken to collect data on chemicals for their risk assessment leading to, where necessary, risk reduction [2].

The gap in knowledge about intrinsic properties of existing substances should be closed to ensure that equivalent information to that on new substances is available. The available information on existing substances should be thoroughly examined and best use made of it in order to waive testing, wherever appropriate. Studies show significant gaps in publicly available knowledge of existing chemicals, especially in terms of their environmental fate and pathways as well as their ecotoxicity parameters [3].

It is evident that the topic ''environmental chemicals' data'' is strongly related to the subject of structuring and archiving them in environmental and chemical databases [4]. These databases are not only found in every medium (i.e. online, CD-ROM and on the Internet), but are quite varied in their type and contents. The interested community urgently needs support in finding relevant information and data about environmental chemicals. Within this context it is of the utmost importance to give

precise indications of the importance and quality of the databases. This means performing a comparative evaluation of databases with respect to several different evaluation criteria. Several approaches exist. Comparisons of eight large chemical structural databases have been performed [5]. Another evaluation approach for structural and reaction databases has been reported by Cooke and Schofield [6]. Approaches using the mathematical method of lattice theory called Hasse algebra may also be used to evaluate environmental and chemical databases. A comparative evaluation of data sources of online databases and databases on CD-ROM based on research results gained in the years 1996/1997 has been reported by Voigt et al. [7, 8]. An overview of evaluation approaches for chemical databases was also published recently [9].

The related scientific disciplines of environmetrics [10] and chemometrics [11, 12] aim to achieve a more comprehensive and objective evaluation and interpretation of environmental data (i.e. they pave the way to receiving knowledge out of data).

For an adequate use of environmental and chemical information the data analysis is of the utmost importance. The basic question tackled in this paper is: ''What kind of information density do commonly available databases encompass?'' We therefore investigate 12 numerical databases which focus on environmental fate and ecotoxicity.

Environmetrical/chemometrical methods used: HDT and METEOR

Introductory remarks about Hasse diagram technique (HDT)

Well-known chemometrical and environmetrical methods are used to analyse environmental chemicals' data. A good overview of established methods of chemometrics and environmetrics is given in the literature [10, 12–15]. As the data situation in environmental sciences in combination with chemical substances becomes more and more complex, this poses a great challenge for establishing new data analysis methods. Einax has summarised important new chemometrical methods in a recent publication [11]. One of these challenging new chemometrical and environmetrical methods is the method of evaluation by order theory, based on the theory of partially ordered sets, and its specific application, known in the literature as the Hasse diagram technique (HDT). The HDT is well explained in a variety of different environmental, chemical and statistical journals. A rather comprehensive description can be found in refs. [16, 17], and a comparison of the HDT with multi-variate statistical methods is given by Voigt et al. [18]. Therefore only some aspects are highlighted here, which will be useful in the subsequent application.

Comparison of databases (or more generally, objects) needs an ordering. Thus, ordering of objects characterized by many attributes (scores or parameters) is the main focus of this paper: two objects, often in mathematical literature also called elements (if the aspect of belonging to sets is important), x, y of an object set are considered as being ordered (e.g. $x \le y$) if all scores of x are less than or equal to those of y. Order relations can be visualised by using Hasse diagrams, which are acyclic digraphs. Objects therein are often drawn as small circles together with an appropriate identifier. The edges of this graph are the cover-relations; that means, edges which express simply the transitivity of order relations are omitted, as they bear only redundant information. In our applications, the circles near the top of the Hasse diagram indicate objects that are ''better'' objects according to the criteria used to rank them: the objects not ''covered'' by other objects are called maximal objects. Objects which do not cover other objects are called minimal objects. In some Hasse diagrams, isolated objects also exist which can be considered as maximal and minimal objects at the same time. Sometimes it is useful to call objects which are not at the same time both maximal and minimal objects 'proper'. When there is exactly one maximal and one minimal element respectively, then these unique objects are called the greatest and least elements, respectively.

A brief technical information about the WHasse program, written in DELPHI, can be found in Brüggemann et al. [19]. Further theoretical developments concerning order theoretical tools in environmental sciences and their applications are discussed in regularly held workshops in Germany, Denmark and Italy [20–24]; for a commercial application, see ref. [25].

Hasse diagrams as mathematical objects

Attributes are—in the case of the object ''x''—denoted as $q(1,x)$, $q(2,x)$,...,$q(m,x)$ and often written as a tuple $\mathbf{q}(x)$. We consider attributes as a quantification of criteria related to more general aims for the evaluation. The set of properties is called an information base (IB). Technical remarks on partial order:

(a) Consider two objects x and y, then we say $y \ge x$ (with respect to the m properties of interest) if $q(i,x) \le q(i,y)$ for all $i = 1, 2,..,m$ and there is at least one i^*, for which $q(i^*,x) < q(i^*,y)$ (because of the demand ''for all'' this definition is denoted the ''generality principle''). If $q(i,x) \le q(i,y)$ for all $i = 1,..,m$ then the objects x and y are comparable. The mere fact that x is comparable with y is often denoted as $x \perp y$. If for two objects $x \le y$ there is no other object z such that $x \le z \le y$ then this neighbourhood of x, y is called a cover-relation.

(b) Often however one finds $q(i,x) < q(i,y)$ for one index set I' and $q(i,x) > q(i,y)$ for another index set I'' with $I' \cap I'' = \emptyset$. In that case, the objects x and y are incomparable and one writes: $x||y$. Note that the order relation defined here is known as product order.

(c) The main frame of HDT is the so-called four-point program:

1. Selecting a set of objects of interest which are to be compared, E. The so-called ground set.
2. Selecting a set of properties, by which the comparison is performed, called the information base IB.
3. Find a common orientation for all properties, according to the criteria they are assigned.
4. Analysing x, $y \in E$ whether one of the following relations is valid:
   - $x \sim y$ (equivalence, we call the corresponding equivalence relation R, the equality of two tuples $\mathbf{q}(x)$, $\mathbf{q}(y)$)
   - $x \leq y$ or $x \geq y$ (comparability), see (a)
   - $x || y$ (incomparability, there is a "contradiction" in the data of x and y"), see (b).

The relation defined above among all objects is indeed an order relation, because it fulfills the axioms of order, namely

- Reflexivity (one can compare each object with itself)
- Antisymmetry (if x is preferred to y then the reverse is only true if the two objects are equal or equivalent)
- Transitivity (if x is better than y, and y is better than z, then x is better than z).

A set E equipped with an order relation $\leq$ is said to be an ordered set (or partially ordered set) or briefly "poset" and is denoted as $(E, \leq)$. Because the $\leq$-comparison depends on the selection of the information base (and of the data representation—classified or not, rounded, etc.) we also write (E, IB) to denote this important influence of the IB for any rankings [16].

Sometimes it is useful to refer to the quotient set, which is induced by the equivalence relation of equality, R (for details, see refs. [26, 27]). As usual, we write E/R for the quotient set, and $(E/R, \mathrm{IB})$ for the partially ordered quotient set.

Some structural details of the Hasse diagram

In order to interpret a Hasse diagram some further terms have to be introduced. A Hasse diagram can also be considered as a mathematical directed graph ("digraph"): the vertices of this graph are the objects of the ground set, and the arcs represent the cover-relations.

Articulation point

An articulation point is a vertex of the transitive hull of the digraph whose elimination would increase the number of hierarchies. A graphical scheme may be sufficient to explain why here a transitive hull is useful (see Fig. 1):

If there are articulation points, then the Hasse diagram can be almost separated in hierarchies. That means that the identification of articulation points helps to discover specific data structures within the data-matrix.

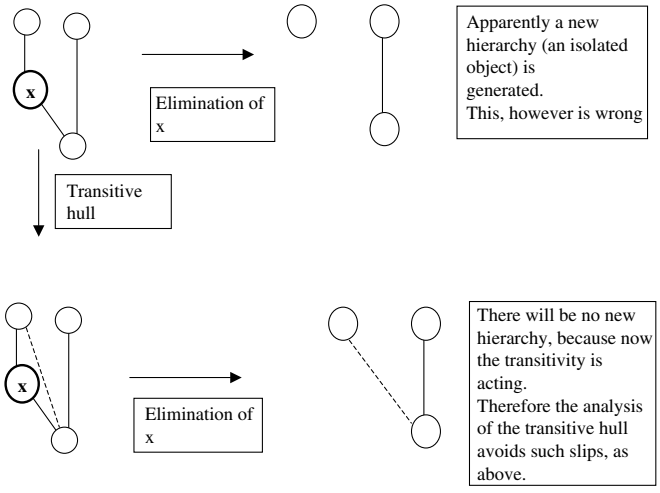

Fig. 1 The role of the transitive hull (each order relation is represented by an arc) in identifying articulation points

Levels

Levels represent a first screening and a partitioning of set E according to increasing values of the attributes. They are defined by the longest chain within the Hasse diagram (see below). Not unique from the point of view of order theory, but uniquely defined, if additional rules are introduced (e.g. conservativity, i.e. if HDT objects are assigned to the highest possible level). The set of levels together with the $\leq$-relation forms a new poset $(L, \leq)$, which represents a chain over all objects of L (i.e. a total order). Both the empirical poset $(E, \leq)$ and $(L, \leq)$ are related by an order-preserving map.

Chain

A subset of E, whose objects are all mutually comparable, is called a chain (see Fig. 2; the subset PES, EHC, CIV and ECO forms a chain PES $\leq$ EHC $\leq$ CIV $\leq$ ECO). Usually the Hasse diagram is more helpful for decision makers the larger the number of objects a chain contains. The special case that all objects of E form a chain allows a unique decision about prioritizing the objects of interest and is called a linear or synonymously total order.

METEOR (method of evaluation by order theory)

The basic idea of METEOR is "the principle of partial participation" (see ref. [17]) which is based on HDT. This means that subsets of the IB can be numerically combined (e.g. by weighted sums). Selecting weights is a matter of participation as demanded by the sustainability principle. The step-by-step aggregation procedures of the data-matrix will be performed by applying METEOR. However, in order to combine the attributes freely, a common scaling level must be assumed: it is not obvious to combine an ordinal attribute with an inter-

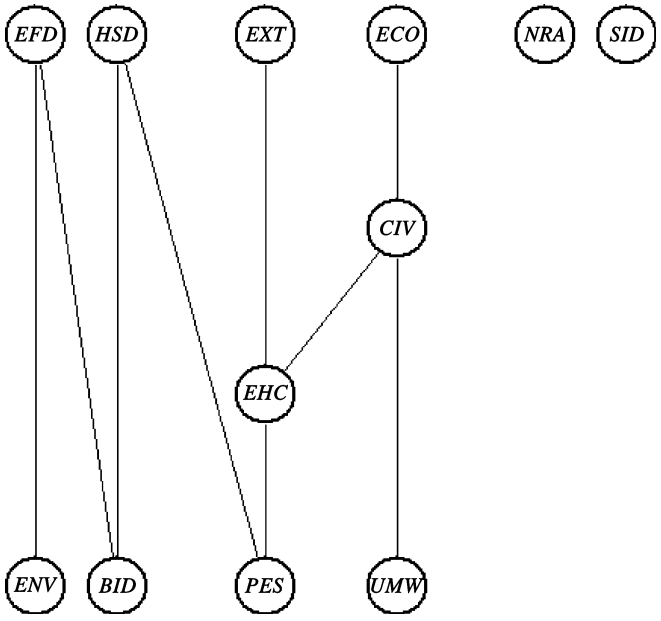

Fig. 2 Hasse diagram of a 12×27 data-matrix

val-scaled attribute! If this combination is to be done, one has to carefully justify this or (better) to stop the aggregation procedure.

The main advantage of the step-by-step procedure is that each positive monotonous combination of, say, two attributes, leading to a "super-attribute" corresponds order theoretically to an order-preserving map. Thus, a systematic sequence of Hasse diagrams arises and the role of weighting can be traced back, when the final result, a linear order, is found by a stepwise aggregation. Furthermore, it is possible to identify weight-sensitive and weight-insensitive objects of the ground set E and E/R, respectively [27]. Note that now the columns of the data-matrix (rows represent the objects; columns represent the attributes) can be considered as vectors of a linear space.

Formally all weighting schemes in METEOR, and especially those discussed in this paper can be written as follows:

$$\begin{pmatrix} q_1^{\text{new}} \\ \vdots \\ q_k^{\text{new}} \end{pmatrix} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,27} \\ \vdots & & & \vdots \\ w_{k,1} & w_{k,2} & \cdots & w_{k,27} \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ \vdots \\ q_{27} \end{pmatrix}$$

This matrix equation is valid for each object and relates the new attributes, the super-attributes, q_i^{new}, with the original ones by a weighted sum, $w_{i,j}$, which are the (normalized) weights. One motive to combine attributes is an abstraction process: for example, concentrations of heavy metals may be aggregated to one super-attribute and similarly those of organics to one other super-attribute if the general chemical loading is of interest and not the details. If, as in our example, the first 25 original attributes are aggregated which generates one new super-attribute, whereas the attributes q_{26}, q_{27} are left unchanged, then this weighting scheme can be written as follows:

$$\begin{pmatrix} q_1^{\text{new}} \\ q_2^{\text{new}} \\ q_3^{\text{new}} \end{pmatrix} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,25} & 0 & 0 \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ \vdots \\ q_{25} \\ q_{26} \\ q_{27} \end{pmatrix}$$

$$q_1^{\text{new}} = \sum_{i=1}^{25} w_{1,i} \cdot q_i$$

$$q_2^{\text{new}} = q_{26}$$

and

$$q_3^{\text{new}} = q_{27}$$

for all objects.

Selection of databases (objects) and environmental parameters and chemicals (attributes)

Environmental chemical databases (objects)

All the 12 evaluated non-fee databases are incorporated in the DAIN (metadatabase of internet resources for environmental chemicals) which can be found under http://www.wiz.uni-kassel.de/dain [28].

The chosen databases which are listed together with their later used abbreviations and their Internet address uniform resource locator (URL) are all available for free on the Internet. US, European, Japanese and Australian data sources are covered (see Table 1). Three different types of numerical databases can be distinguished:

- Single databases which cover only one data collection (BID, CIV, HSD, UMW)
- Multi-database databases which encompass several databases under the same name and search interface (ECO, ENV, EFD, EXT)
- Monograph databases which cover extensive reviews on very few chemicals (EHC, NRA, PES, SID).

Environmental parameters and high-production-volume chemicals

Data-matrix (overview)

The databases are looked upon with respect to some selected chemicals and environmental parameters (ecotoxicological and environmental fate and pathways). Twelve high-production-volume chemicals (see Table 3) and 15 environmental parameters (see Table 2) are chosen. Hence we have to cope with a 12×27 data-matrix, whose entries are the attribute values (columns) for each database (rows). The data-matrix is given in

Table 1 List of chosen numerical databases focussing on environmental chemicals

Database name	Abbreviation	URL
Biocatalysis/biodegradation database	BID	http://umbbd.ahc.umn.edu/
Chemicals information system for consumer-relevant substances (CIVS)	CIV	http://www.bgvv.de/cms/detail.php?template=internet_en_index_js
ECOTOX	ECO	http://www.epa.gov/ecotox/
Envirofacts	ENV	http://www.epa.gov/enviro/html/emci/chemref/index.html
Environmental fate database	EFD	http://esc.syrres.com/efdb.htm
Environmental health criteria monographs (EHCs)	EHC	http://www.inchem.org/pages/ehc.html
EXTOXNET	EXT	http://ace.ace.orst.edu/info/extoxnet/
HSDB	HSD	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB
NRA chemical review program	NRA	http://www.nra.gov.au/chemrev/chemrev.shtml
Pesticide database, Japan	PES	http://chrom.tutms.tut.ac.jp/JINNO/PESDATA/00alphabet.html
SIDS	SID	http://www.chem.unep.ch/irptc/sids/sidspub.html
UmweltInfo	UMW	http://www.umweltinfo.de/ui-such/ui-such.htm

Table 2 Environmental parameters

Parameter	Abbreviation	Super-attribute
Photodegradation	PHO	FATE
Stability in water	SWA	FATE
Stability in soil	SSO	FATE
Biodegradation	BDE	FATE
BOD5, COD or BOD5/COD ratio	BOD	FATE
Bioaccumulation	BAC	FATE
Acute/prolonged toxicity to fish	ATF	ETOX
Acute toxicity to aquatic invertebrates	ATD	ETOX
Toxicity to aquatic plants (e.g. Algae)	ATP	ETOX
Toxicity to microorganisms (e.g. Bacteria)	ATB	ETOX
Chronic toxicity to fish	CTF	ETOX
Chronic toxicity to aquatic invertebrates	CTD	ETOX
Toxicity to soil-dwelling organisms	TSO	ETOX
Toxicity to terrestrial plants	TTP	ETOX
Toxicity to other non-mammalian terrestrial species	TNT	ETOX

Table 3 List of chosen chemicals for the evaluation of environmental chemicals' databases

CAS number	Chemical name	Remarks	Super-attribute
100-00-5	1-Chloro-4-nitrobenzene	HPVC	CHID
100-01-6	4-Nitroaniline	HPVC	CHID
100-02-7	4-Nitrophenol	HPVC	CHID
1912-24-9	Atrazine	HPVC, ED	CHID
999-81-5	Chlormequat chloride	HPVC	CHID
333-41-5	Diazinon	HPVC	CHID
60-51-5	Dimethoate	HPVC	CHID
26761-40-0	Ethofumesate	HPVC	CHID
1071-83-6	Glyphosate	HPVC	CHID
34123-59-6	Isoproturon	HPVC	CHID
121-75-5	Malathion	HPVC, ED	CHID
137-26-8	Thiram	HPVC	CHID

HPVC High-production-volume chemical, *ED* endocrine disruptor

Annex 1. Furthermore, there are several subgroups of parameters which are candidates for an aggregation. For example, one may define a super-attribute "FATE" by aggregation of all fate descriptors or all ecotoxicity descriptors to "ETOX" and leave the 12 environmental chemicals unchanged. For our demonstration in this paper, we select an aggregation scheme which leads to "FATE", "ETOX" and "CHID" (see Tables 2 and 3).

Environmental parameters

The environmental fate and pathways and the ecotoxicity parameters implemented in the IUCLID database [3] will be looked upon. These are:

Environmental fate and pathways Photodegradation, stability in water, stability in soil, monitoring data (environment), transport between environmental compartments, distribution, mode of degradation in actual

use, biodegradation, BOD5, COD or BOD5/COD ratio, bioaccumulation.

Ecotoxicity Acute/prolonged toxicity to fish, acute toxicity to aquatic invertebrates, toxicity to aquatic plants (e.g. algae), toxicity to microorganisms, (e.g. bacteria), chronic toxicity to fish, chronic toxicity to aquatic invertebrates, toxicity to soil-dwelling organisms, toxicity to terrestrial plants, toxicity to other non-mammalian terrestrial species, biological effects monitoring, biotransformation and kinetics.

HPV environmental chemicals

The databases are not only looked upon with respect to their parameters but also with respect to some selected chemicals. The selection of a pragmatic number of existing chemical substances which are not only relevant in one aspect is difficult. The 12 high-production-volume (HPV) chemicals chosen are listed in Table 3. The chemicals were selected according to a ranking approach for chemical substances reported by Lerche et al. [29].

Search results

Application of Hasse diagram technique

The data-driven evaluation method, HDT, is applied to the 12×27 data-matrix (matrix, see Annex S-1 in Electronic Supplementary Material). Concerning the evaluation of environmental chemical databases with respect to environmental parameters and chemicals, we define that the orientation (point three of the four-point program) is as follows: the value 1 means available information, hence "good"; the value 0 means information unavailable, hence "bad".

It was found that ECOTOX (ECO), EXTONNET (EXT), environmental fate database (EFD) and hazardous substances database (HSD) showed the best results. These objects are proper maximal objects: there are no other databases which are better in all aspects than these proper maximal objects. UmweltInfo (UMW), envirofacts databases (ENV), biocatalysis/biodegradation database (BID) and pesticide database (PES) give bad results in comparison to most of the other databases. These are the proper minimal objects. Following the definition of minimal objects, there are no worse databases. The databases chemical review programme (NRA) and screening information datasets (SID) from the OECD are so-called isolated objects. They cannot be compared to any other object (see above). Hence four proper maximal, four proper minimal and two isolated objects are found in Fig. 2.

Differences concerning the numbers of successors can be detected in the diagram: e.g. the maximal object EFD is only connected and hence comparable with two other objects, namely ENV and BID, whereas the maximal object ECO is connected with four other objects CIV, EHC, PES and UMV.

There are four levels:

Level 1: {ENV, BID, PES, UMW}

Level 2: {EHC}

Level 3: {CIV}

Level 4: {EFD, HSD, EXT, ECO, NRA, SID}

Level 1<Level 2<Level 3<Level 4
"bad" "good"

Furthermore, PES is one of the articulation points (others are EFD, BID and HSD). By elimination of the object PES, two rather big hierarchies would appear. The subset of ENV, BID, HSD, EFD should have a peculiar data structure by which it is separated from the other databases. A full analysis is beyond the scope of this paper and still needs additional order theoretical background material.

As we have to cope with a broad data-matrix, that is to say more attributes than objects, data reduction procedures on the attributes' side seem to be appropriate. The next logical step is therefore to perform other data reduction procedures (e.g. logical aggregations of attributes), that is to perform METEOR.

Application of METEOR

The original data-matrix of 12 databases (objects) and 27 parameters (attributes) will be subject to a step-by-step aggregation procedure. The aim of this aggregation, which can be performed by applying the HDT program [27], is to get a unique prioritisation scheme after several steps. In this paper we will only apply a single weighting scheme, namely the one in which an aggregation is calculated by equal weights. The criteria (attributes) encompass ecotoxicity and environmental fate. As all the chemical substances are high-production-volume chemicals and used as pesticides, we aggregate them into one group. Hence we cope with three aggregation groups where each of the following super indicators "FATE", "ETOX" and "CHID" (see Tables 2 and 3 last column) are calculated by a sum with equal weights. For example:

$$\text{Fate} = \sum_{i=1}^{i=6} w_i \cdot q_i \quad q_i \in IB_{\text{environmentalfate}}$$

$$= \{PHO, SWA, \ldots, BAC\} \subset IB \quad w_i = 1/6$$

The other two criteria are aggregated analogously.

Aggregation of six environmental fate attributes: FATE

Aggregation of nine ecotoxicity parameters: ETOX

Aggregation of twelve chemicals: CHID

The result of this aggregation procedure in given in the Hasse diagram of Fig. 3. The methodological step is that now no single parameter is responsible for an order relation; groups of similar parameters are responsible instead. In comparison to Fig. 2 comparability is not a matter of all 27 parameters being synchronously ordered but only of the group parameters (i.e. of the super-attributes).

Some visible changes have taken place if one compares the original Hasse diagram of the 12×27 data-matrix (Fig. 2) with the reduced 12×3 data-matrix (Fig. 3). NRA is now a maximal object and no longer an isolated object. The only isolated object in this approach is SID. ENV is no longer a minimal object but one level above. Further differences are found in Table 4, in which the two Hasse diagrams are compared in detail. Furthermore, PES is no longer an articulation point.

Both diagrams show a lot of similarities. They both have four levels and no equivalent objects. The number of comparabilities decreases enormously with the number of attributes. In other words, the more attributes the data-matrix encompasses, the less comparabilities are found. Clearly the process of aggregation can be continued: if now—in a second step—FATE, ETOX and CHID are aggregated into one single "hyper-attribute" the following series is found:

$$\text{UMW} < \text{BID} = \text{PES} < \text{SID} < \text{EHC} < \text{NRA} < \text{EXT} < \text{ENV} < \text{HSD} < \text{CIV} < \text{EFD} < \text{ECO}$$

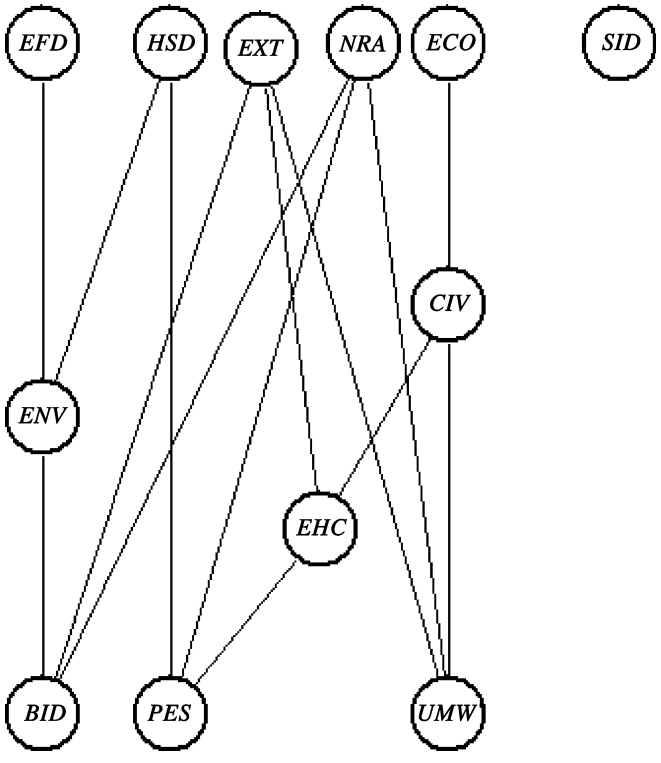

Fig. 3 Hasse diagram of IB = {FATE, ETOX, CHID}, 12×3 data-matrix

Table 4 Comparison of two Hasse diagrams

Data-matrix	12×27 (initial data-matrix)	12×3 (weighting to three super-attributes)
Number of levels	4	4
Objects in largest level	6	6
Equivalent objects	0	0
Maximal objects	4	5
Minimal objects	4	3
Isolated Objects	SID, NRA	SID

The aggregation was performed by considering the chemicals to be twice as important as the fate and ecotoxicity parameters. This sequence is with respect to the top and bottom objects being rather stable against different weights. All objects near the bottom of the sequence above are the minimal objects found in Fig. 3.

Discussion and outlook

The availability of environmentally relevant parameters (environmental fate and ecotoxicity) and data on 12 high-production-volume chemicals in 12 well-known international databases (available free on the Internet) was analysed. The general deficiency of data availability on chemical substances can be seen from the data-matrix alone without applying any sophisticated mathematical method. However, order theoretical methods were taken into account to deduce more and detailed information from a data-matrix with 12×27 entries. The simple approach by Hasse diagrams might not be as informative, as it is not of interest to detect data-gaps in single attributes and in single databases. Therefore the emphasis is set on the METEOR, a discrete mathematical method. METEOR allows one to aggregate groups of attributes of similar meaning in order to detect data-gaps referring to specific attribute groups and not necessarily to one single attribute. The Hasse diagram of the complete data-matrix 12 objects (databases)×27 attributes (parameters + chemicals) revealed that the databases ECO, EFD and EXT, also called multi-database databases, performed best. One has to consider the fact that these databases comprise different sizes of data: EFD comprises approximately 20,000 chemicals, ECO encompasses data on approximately 8,000 chemicals, and EXT only on 400 chemicals. EXT has extended profiles on high-production-volume chemicals. HSD has a broad data collection on 4,500 chemicals. Most single databases which are specialised (BID, PES, and UMW) are found in a minimal position in the Hasse diagram.

The monograph databases EHC, ENV, EFD and EXT are distributed over the whole sequence (linear order) found in the previous section. The reason is that the main focus of these databases is not a large number of chemicals but detailed information (many parameters) on few chemicals. As the sequence was found by just weighting the chemicals as being twice as impor-

tant as the parameters, this non-prominent location is understandable. Controversially, if the parameters were weighted twice as important as the chemicals, the monograph databases would be located in top positions.

The whole approach indicates a rather bad situation in terms of the data availability on existing chemicals and hence an alarming signal concerning the new and existing chemicals policies of the EEC.

For future steps concerning the data availability of chemicals five ways should be taken into account:

- Foster many data-sources (timeliness)
- Foster new publications and enter the data into the numeric databases
- Estimate data by well-established methods (QSAR) and fill up data gaps indicating that the data are estimated ones
- Test chemicals in the described way by the EEC according to the white paper [1]
- Dynamically evaluate the actually best databases

Only by improving the data situation can the amount of extensive testing be reduced to a pragmatic size. A final remark should be added: it is an obvious choice that 12 free databases have been investigated, as this is what we all will have access to. However, since commercial databases are often claimed to be better, a forthcoming investigation by order theory should include those databases, to check whether it is really worthwhile to pay for their information.

References

1. EEC (2001) Commission of the European communities, white paper, strategy for a future chemicals policy, COM 88 final. http://www.europa.eu.int/comm/environment/chemicals/index.htm
2. Heidorn C, Rasmussen K, Hansen BG, Norager O, Allanou R (2003) J Chem Inf Comput Sci 43(3):779–786
3. Allanou R, Hansen BG, van der Bilt Y (1999) Public availability of data on EU high production volume chemicals, EUR 18996EN. http://ecb.jrc.it/
4. Page B, Voigt K (2003) Online Inf Rev 27(1):37–50
5. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) J Chem Inf Comput Sci 41:702–712
6. Cooke F, Schofield H (2001) J Chem Inf Comput Sci 41:1131–1140
7. Voigt K (1997) Erstellung von Metadatenbanken zu Umweltchemikalien und vergleichende Bewertung von Online Datenbanken und CD-ROMs. http://vermeer.organik.uni-erlangen.de/dissertationen/data/dissertation/Kristina_Voigt/html/
8. Voigt K, Gasteiger J, Brüggemann R (2000) J Chem Inf Comput Sci 40:44–49
9. Voigt K, Welzl G (2002) Online Inf Rev 26:172–192
10. El-Shaarawi AH, Hunter JS (2002) Environmetrics, overview. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of Environmetrics. Wiley, Chichester, pp 698–702
11. Einax JW (2003) In: Gnauck A, Heinrich R (eds) The Information society and the enlargement of the European Union. Metropolis Verlag, Marburg, pp 51–57
12. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J (1997) Handbook of chemometrics and qualimetrics. Elsevier, Amsterdam
13. Einax JW, Zwanzinger HW, Geiss S (1997) Chemometrics in environmental analysis. VCH, Weinheim
14. Stoyan D, Stoyan H, Jansen U (1997) In: Teubner BG (ed) Umweltstatistik, Statistische Verarbeitung und Analyse von Umweltdaten. Verlagsgesellschaft, Stuttgart
15. Welzl G, Faus-Kessler T, Scherb H, Voigt K (2004) Biostatistics. In: Sydow A (ed) EOLSS Encyclopedia of Life Support Systems. EOLSS Publishers Co. Ltd., Oxford. http://www.eolss.net
16. Brüggemann R, Welzl G (2002) Order theory meets statistics—Hasse Diagram technique. In: Voigt K, Welzl G (eds) Order theoretical tools in environmental sciences, order theory (Hasse diagram technique) meets multivariate statistics. Shaker-Verlag, Aachen, pp 9–40
17. Brüggemann R, Drescher-Kaden U (2003) Einführung in die modellgestützte Bewertung von Umweltchemikalien. Springer, Berlin Heidelberg New York
18. Voigt K, Welzl G, Brüggemann R (2004) Environmetrics 15: 577 - 596
19. Brüggemann R, Bücherl C, Pudenz S, Steinberg C (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim Hydrobiol 27:170–178
20. Brüggemann R (ed) (1998) Order theoretical tools in environmental sciences held on November 16, 1998 in Berlin, Berichte des IGB 1998, Heft 6, Sonderheft I, Institut für Gewässerökologie und Binnenfischerei, Berlin
21. Sørensen PB, Carlsen L, Mogensen BB, Brüggemann R, Luther B, Pudenz S, Simon U, Halfon E, Voigt K, Welzl G, Rediske G (eds) (2000) Order theoretical tools in environmental sciences, NERI—Technical report No. 318, National environmental research institute, Roskilde
22. Pudenz S, Brüggemann R, Lühr H-P (eds) (2001) Order theoretical tools on environmental science and decision systems. Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, Heft 14, Sonderheft IV, p 222
23. Voigt K, Welzl G (eds) (2002) Order theoretical tools in environmental sciences, order theory (Hasse diagram technique) meets multivariate statistics. Shaker-Verlag, Aachen
24. Sørensen P, Brüggemann R, Lerche DB, Voigt K (2003) Order theory in environmental sciences, NERI Technical Report No. 479, National Environmental Research Institute, Ministry of the Environment, Copenhagen, pp 68–95
25. Criterion (2004) WHasse Software. http://www.criteri-on.de/hdtws-site/index.html
26. Brüggemann R, Bartel HG (1999) J Chem Inf Comput Sci 39:211–217
27. Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (2001) J Chem Inf Comput Sci 41:918–925
28. Voigt K, Welzl G (2002) In: Voigt K, Welzl G (eds) Order theoretical tools in environmental sciences, order theory (Hasse diagram technique) meets multivariate statistics. Shaker-Verlag, Aachen, pp 113–128
29. Lerche D, Brüggemann R, Sørensen PB, Carlsen L, Nielsen OJ (2002) J Chem Inf Comput Sci 42:1086–1098