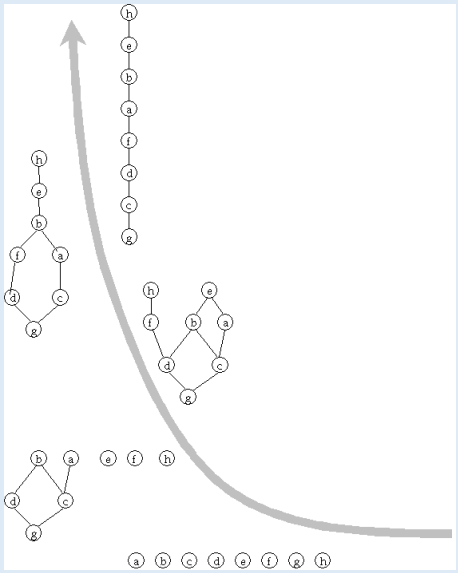
National Environmental Research Institute
Ministry of the Environment · Denmark

Order Theory in Environmental Sciences

Integrative approaches

The 5th workshop – held at the
Environmental Research Institute (NERI)

NERI
Technical Report
No. 479



[Blank page]

National Environmental Research Institute
Ministry of the Environment

Order Theory in Environmental Sciences

Integrative approaches

The 5th workshop – held at the
National Environmental Research Institute (NERI)

NERI Technical Report, No. 479
2004

P.B. Sørensen, D.B. Lerche, S. Gyldenkærne,
M. Thomsen, P. Fauser, B.B. Mogensen, B. Kronvang
National Environmental Research Institute

R. Brüggemann, U. Simon, M. Erfmann, M. Abs
Leibniz Institute of Freshwater Ecology
and Inland Fisheries

K. Voigt, G. Welzl
GSF – National Research Center for Environment
and Health

L. Carlsen
Awareness Center, Denmark

S. Pudenz
Criterion – Evaluation and Information Management,
Germany

Data sheet

Title:	Order Theory in Environmental Sciences
Subtitle:	Integrative approaches. The 5th workshop - held at the National Environmental Research Institute (NERI), Roskilde, Denmark, November 2002.
Authors:	P.B. Sørensen[1], R. Brüggemann[2], D.B. Lerche[1], K. Voigt[3], G. Welzl[3], U. Simon[2], M. Abs[2], M. Erfmann[2], L. Carlsen[4], S. Gyldenkærne[1], M. Thomsen[1], P. Fauser[1], B.B. Mogensen[1], S. Pudenz[5] and B. Kronvang[1].
Departments:	[1)]National Environmental Research Institute, Denmark [2)]Leibniz Institute of Freshwater Ecology and Inland Fisheries, Germany [3)]GSF - National Research Center for Environment and Health, Germany. [4)]Awareness Center, Denmark [5)]Criterion – Evaluation and Information Management, Germany.
Serial title and no.:	NERI Technical Report No. 479
Publisher:	National Environmental Research Institute © Ministry of the Environment
URL:	http://www.dmu.dk
Date of publication:	December 2003
Editing complete:	December 2003
Referee:	Hanne Bach
Financial support:	No external financing
Please cite as:	Sørensen, P.B., Brüggemann, R., Lerche, D.B., Voigt, K. Welzl, G., Simon, U., Abs, M., Erfmann, M., Carlsen, L., Gyldenkærne, S., Thomsen, M., Fauser, P., Mogensen, B.B., Pudenz, S. & Kronvang, B. 2003: Order Theory in Environmental Sciences. Integrative approaches. The 5th workshop held at the National Environmental Research Institute (NERI), Roskilde, Denmark, November 2002. National Environmental Research Institute, Denmark. 161 pp. - NERI Technical Report no. 479. http://technical-reports.dmu.dk

Abstract: This is a collection of proceedings from the fifth workshop in Order Theory in Environmental Science. This workshop series concern the development of the concept of Partial Order Theory is development in relation to practical application and the use is tested based on specific problems. The Partial Order Theory will have a potential use in cases where more than one criterion is included in a prioritisation problem both in relation to decision support and in relation to data-mining and interpretation. Especially the problems where a high degree of complexity results in considerable uncertainty are good candidates for application of Partial Order Theory.

Keywords:	Partial Order Theory, Ecological Modelling, Eco-toxicological Databases, GIS, Pesticides, Monitoring Data, QSAR.
Layout:	Ann-Katrine Holme Christoffersen
ISBN:	87-7772-783-5
ISSN (electronic):	1600-0048
Number of pages:	161
Internet-version:	The report is only available as a PDF file from NERI's homepage http://www.dmu.dk/1_viden/2_Publikationer/3_fagrapporter/rapporter/FR479.pdf
For sale at:	Ministry of the Environment Frontlinien Rentemestervej 8 DK-2400 Copenhagen NV Denmark Tel.: + 45 70 12 02 11 frontlinien@frontlinien.dk

Contents

Preface

This proceeding covers the major part of the presentations given at the 5th Workshop on *Order Theory in Environmental Sciences, Integrative approaches* held in November the 24-25 in year 2002 in Roskilde, Denmark. The National Environmental Research Institute (NERI) in Denmark hosted the workshop. This workshop continues a workshop series on application and methodological development of Partial Order Theory as a tool in environmental management. Actual information about the status of this ongoing series of workshops can be obtained by consulting either Dr. Rainer Brüggemann (brg@igb-berlin.de) or Dr. Peter B. Sørensen (email: pbs@dmu.dk).

The first workshop was held in Berlin November 16th, 1998 at the Institute of Freshwater Ecology and Inland Fisheries and a proceeding from this workshop is available as: "Proceeding of the workshop on Order Theoretical Tools in Environmental Sciences", Berichte des IGB 1998 (Berlin), Heft 6, Sonderheft I, ISSN-Nr. 1432-508X. The proceeding from the first workshop can be provided by contacting pbs@dmu.dk.

The second workshop was held in Roskilde October 21 st, 1999 at the National Environmental Research Institute (NERI) and the proceeding: Order Theoretical Tools in Environmental Science, NERI technical report No. 318, can be downloaded from the list of publications at www.dmu.dk.

The 3th workshop was held in Berlin November 6-7th 2000 at the Leibniz-Institute of Freshwater Ecology and Inland Fischeries, Germany and organised by Criterion with Dr. Stefan Pudenz. The proceeding: Order Theoretical Tools in Environmental Science and Decision Systems, Berichte des IGB, 2001, Heft 14, ISBN-No. 1432-508X.

The 4th workshop was held in Iffeldorf, Bavaria, Germany at the Limnologische Station der Technische Universität München, November 5-6th 2001. The proceeding: Kristina Voigt and Gerhard Welz (eds.), Order Theoretical Tools in Environmental Sciences, Order Theory (Hasse Diagram Technique) Meets Multivariate Statistics, Shaker Verlag 2002, ISBN No. 3-8322-0792-9.

First attempts to relate structures of Hasse diagrams with mutual probabilities

R. Brüggemann[*)], D. Lerche[1, 2], P.B.Sørensen[1]

[*)] Leibniz Institute of Freshwater Ecology and Inland Fisheries,
Department: Ecohydrology
Mueggelseedamm 310
D-12587 Berlin
Germany

[1)] National Environmental Research Institute
Department of Political Analysis
DK-4000 Roskilde
Denmark

[2)]University of Copenhagen
Institute of Chemistry
H.C. Ørsted Institute
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

Abstract

Both in an environmental as well as in many other contexts partial ordering is applied in order to rank objects. An example could be the ranking of chemical substances according to their environmental impact. However, sometimes the rank of all objects is not of primary interest, sometimes an estimation of the probability of one object to be "better" than another is necessary. Just as for the ranking of all objects, the partially ordered set can also be used to derive the mutual ranking probability of any two incomparable objects.

In order to arrive at a mutual rank for two objects the total set of linear extensions of the partial order is needed. This is however extremely time consuming and for data sets of more than 20 objects often not tractable. Therefore it is attractive to develop a methodology for estimating the mutual rank of two objects. This paper deals with first attempts of deriving such methodology.

In order to derive a semi-empirical formula, 27 different partial ordered sets are selected, which do not have too many objects (≤ 9). Additionally simple descriptors of the structure were defined, such as $N_u(x \circ y)$ and $N_d(x \circ y)$, which counts all objects respectively above and below x, which are **not** respectively above and below y at the same time. The statistical analysis shows that the expression

$prob_Q(x>y) := 1/[1+ Q(x_o y)/Q(y_o x)]$ is a fairly good estimator of the exact calculated probability over the variety of 27 partial ordered sets. The term $Q(x_o y)$ is $(N_u(x_o y)+1)/(N_d(x_o y) +1)$.

In order to understand this formula, a systematic study was carried out to see the influence of chains and antichains on the linear extensions and on the probability. Considerable effort has been done to derive a total order from a partially ordered set in order to facilitate decisions. In order to derive at an average rank again the total set of linear extensions of the partial order is needed. Therefore the concept of randomly chosen linear extensions was developed (Sørensen et al. 2001). This method was improved dramatically by the application of an estimation of the mutual probability $p_m(x>y)$ of any two incomparable objects in the partial ordered set. This further motivated the analysis of the mutual ranking probabilities.

1 Introduction

In environmental sciences models are widely applied in decision making. Usually additional assumptions are needed to derive a fitness function from the model outcomes or the original data set in order to support decisions. The plentiness of scoring models and of multicriteria assessment methods show that some consistency in the area is still missing. If one accepts that the role of all these attempts is to combine the a priori information in a more or less sophisticated way to get an estimate of the fitness function, then this can be reduced to the task of finding a monotonous function, which is interpreted as fitness and thus used as a ranking or ordering index. If on the other hand from the a priori knowledge a partial order is constructed, then its linear extensions (see Trotter, 1992, Brüggemann, R. et al., 1999) give a basis for estimation of a fitness function too. Additionally, information about the uncertainty of this ranking, which depends on the structure of the partial ordered set, is given from the linear extensions. This is done, without any arithmetical combinations of the variables. Because the information, which may be available if a certain conventional fitness model is used, is already found by any of these linear extensions, the construction of an averaged ranking by means of the set of linear extensions, encompasses all possibilities based on monotonous functions on the variables and is therefore called a GENERALIZED RANKING MODEL (GRM).

Now, what is the relation between any numerical/algebraic given fitness model and the GRM? If the conventional fitness function is known, it is represented by a certain linear extension. However, this fitness function is a priori not known therefore some uncertainty exists, which may be expressed by a Monte Carlo simulation of some parameters of this fitness function. In that case several, perhaps all linear extensions may be realised. However, how often one of these linear extensions will be met, depends clearly on the numerical/metric structure of the fitness function.

The partially ordered sets (P, ≤) are starting points to

- identify the incomparable objects (and to select eventually a subset of objects of interest)
- calculate rank probability distribution functions
- calculate average ranks and
- mutual probabilities $p_m(x>y)$ in GRM of incomparable objects $x||y$ in (P,).

Further, it has been shown that the application of an approximation of the mutual ranking probability improves the estimation of the ranking probability by up to 90% on average using the random linear extension methodology (Lerche et al., 2003). The importance of a systematic investigation of the estimation of the mutual ranking probability in this study is considered as an important starting point of the random linear extension methodology.

The paper is organised in the following manner:

1. A brief explanation of the theoretical concept to calculate mutual probabilities
2. The basis for almost all considerations is the linear extensions. Some remarks about counting linear extensions and about related topics are provided.
3. An empirical study is performed to come up with a formula for the estimations of the mutual probabilities. As a preliminary step a proposal is presented on how the counts of linear extensions can be stored.
4. A more theoretical approach for deriving a formula for the mutual ranking probability is performed and the model systems are used for evaluation.
5. All the approaches, mentioned in chapter 3-5 refer to some kind of "above-below"-calculations. In chapter 6 systems, which do not fit into this scheme are discussed.

The aim of chapter 3-4 is to find relations between mutual probabilities and structural information found in the Hasse diagram. It is not intended to compete with usual features maintained by the program WHASSE. Certainly this procedure makes the job, however it allows no insight into the governing structural information.

2 Methodological development

2.1 Theoretical concept

The true mutual ranking probability is calculated following a three-step procedure, which also can be performed by applying the WHASSE software, based on a data matrix. Subsequently all linear extensions are identified and from that the mutual probabilities can be calculated.

The standard algorithm is:

I) Find all linear extensions of the partial order, e(P).

II) Find the number of linear extensions, where an object, x, is larger than another object, y, n(x>y).

III) Calculate the mutual ranking probability, p_m, according to the following equation:

Let x||y in P, then p_m (x>y) = n(x>y)/e(P) Eq. 1

where x||y is an incomparable pair of objects and P is the partial order.

Note that such a statement only gives sense, if the concept of the GRM is taken into account: Only, if a total ranking is supposed to exist, the question on the mutual probability gives sense.

The linear extensions can be found in a rather systematic, but tedious way, by using the Atkinson scheme (Eq. 7). As the linear extensions play a central role, some statements, well known in mathematical literature are given.

2.2 Linear Extensions for theoretical deviations

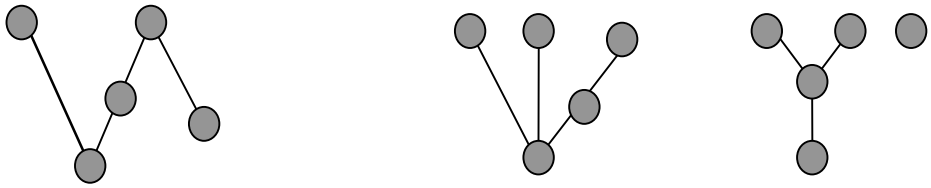
Counting

In order to estimate the mutual ranking probability in a more systematic and theoretical way the concept of linear extensions is investigated. Note that when p_m is calculated all linear extensions are applied. Therefore some remarks may be useful on how to estimate e(P), the number of linear extensions of the partially ordered set, P.

If a partial ordered set is characterized by a set of numbers, e.g. number of objects or classes (quotient set, see Brüggemann, R. & Bartel, H.-G., 1999) N, comparabilities, C, length of the longest chain, LC, length of the longest antichain, LAC, then up to now no formula is known to set :

e(P) = f(N,C,LC,LAC,....) Eq. 2

In terms of the most common parameters, like N and C it is easy to find partial ordered sets with the same N, C, LC and LAC but having different e(P)-values:



N = 5
C = 5
e(P) = 11

N = 5
C = 5
e(P) = 12

N = 5
C = 5
e(P) =10

Figure 1: Example of different partial ordered sets, having the same C=5, N=5, LC=3 and LAC=3 but nevertheless different e(P)-values.

One sees that e(P) can not be described only by easy accessible characteristcs.

However, for probability calculations where quotients are to be formed from Lx>y and LT (number of linear extension, where x>y, and the total number of linear extensions), it might be that some factors cancel out. Thus it may be realistic to look for an empirical equation of the mutual probabilities.

In WHASSE the upmost upper estimation is used to indicate what might be the order of magnitude:

$$e(P)_{max} = N! \qquad \text{Eq. 3}$$

Approximately one may use the number of incomparabilities U to roughly estimate the number of linear extensions:

$$e(P) \approx 2^U \qquad \text{Eq. 4}$$

Other approximations make use of random graph theory, for example Brightwell, et al. 1996 gave a formula for the asymptotic case $N \rightarrow \infty$.

$$e(P_{limit}) = (N/2)! \bullet [(N/4)!]^2 \qquad \text{Eq. 5}$$

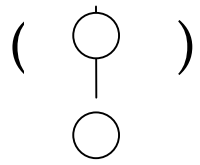
This follows from the limit where partial ordered sets will create a 3-leveled shape and a distribution of objects in the form N/4 , N/2 , N/4 (Brightwell 1996).

Finally for partial ordered sets with $N = 2^m$ objects, based on $\{0,1\}^m$ (Boolean partial ordered sets) Brightwell (personal communication) gives an upper bound depending on m:

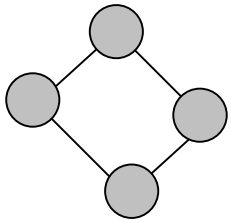
$$e(P, Boolean\,lattices) \leq \prod_{i=0}^{m} \binom{m}{i}^{\binom{m}{i}} \qquad \text{Eq. 6}$$

This equation is suitable for partial ordered sets, which may be constructed from:

m = 1



m=2:



m = 3

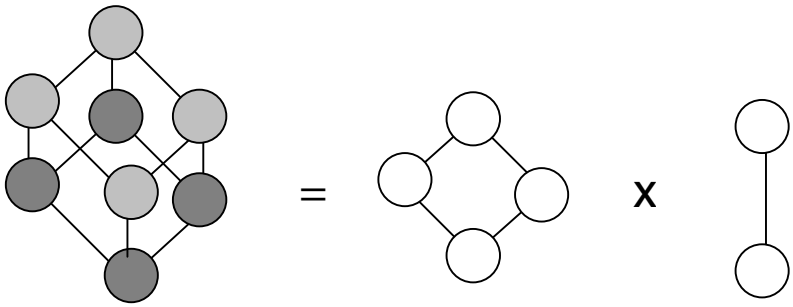


Figure 2: Boolean partial ordered sets

For example: m=3 :

$$\binom{3}{0}^{\binom{3}{0}} \cdot \binom{3}{1}^{\binom{3}{1}} \cdot \binom{3}{2}^{\binom{3}{2}} \cdot \binom{3}{3}^{\binom{3}{3}} = 1 \cdot 3^3 \cdot 3^3 \cdot 1 = 729$$

Compared with the real result e($\mathbf{2}^3$) = 48 this is quite bad, however compared with N! = 8! = 40320 this result is reasonable. Counting the incomparabilities, IC = 9, then 2^9 = 512 is, however, still somewhat better than the Brightwell estimation.

A recurrence relation

There are two powerful recurrence relations, which can be used to calculate e(P). The one is associated with Atkinson, the other recurrence relation is associated to Stanley, 1986. The equation 7 is of such importance that we would like to call the process behind the recurrence relation an Atkinson-scheme. For a simple example, see Figure 3:

Atkinson & Chang, 1986, Atkinson, 1989 and Edelman et al., 1989:

$e(P) = \Sigma\, e(P\text{-}x)$, Eq. 7

x taken from the set of maximal objects or
(exclusively from minimal objects)

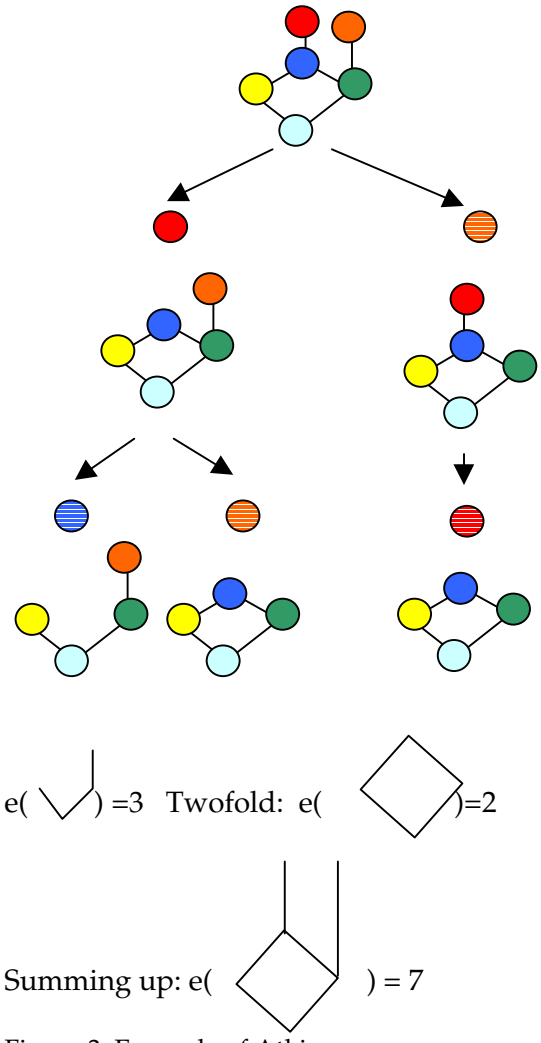


Figure 3: Example of Atkinson processes

As one can see, in the Atkinson process the same structure may appear several times. Thus knowing the number of linear extension of some typical appearing structures, like $\mathbf{2}^2$ would heavily facilitate the calculations (Atkinson & Chang 1986). Thus a systematic catalogue of a series of partial orders might be useful. This however would be far beyond the scope of this paper. If the Atkinson process leads to partial ordered sets with different hierachies, then one can clearly try to analyse the subpartial ordered sets separately, or one can use a relation, which is also generally of fundamental importance:

Let P be a partial ordered set and P_i disjoint subpartial ordered sets, so that it is valid:

$$P = \oplus P_i \quad and \quad P_i \cap P_j = \phi$$

Let N_i be the number of objects of P_i and $N=\Sigma N_i$, i=1,...,k then the following equation is given (Stanley, 1986):

$$e(P) = \left(\frac{N!}{N_1 \cdot N_2 \cdot \ldots \cdot N_k} \right) \cdot \prod_{i=1}^{k} e(P_i) \qquad \text{Eq. 8}$$

Using this formula and the catalogue the number of linear extensions of many partial ordered sets can be calculated. Especially Eq. 8 becomes very simple, if the subpartial ordered sets are only chains, because then the product of $e(P_i)$ can be replaced by 1. I.e. if the partial ordered set consists of two chains then:

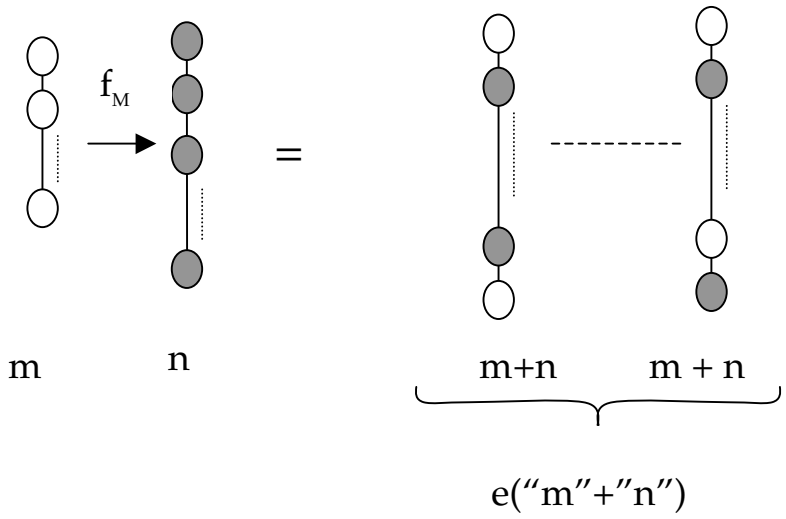


Figure 4: Notation used for a double chain system

$$e("m"+"n") = \left(\frac{(m+n)!}{m!n!} \right) \qquad \text{Eq. 9}$$

We will refer to this equation and the process behind it as "mixing equation" and "mixing process", respectively.

Spectrum of elements within a partial order

A more generalized kind of recurrence relation is to find information about linear extensions by examining parts only. Only one approach seems to be immediately applicable. This is the concept of a spectrum of an element. Looking for two - tree like posets (i.e. they have to be considered as an undirected graph with no cycles):

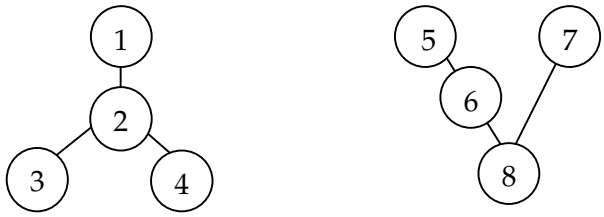


Figure 5: Two tree-posets to be fused by introduction of 7 covers 3 (see Figure 7).

The linear extensions are shown in Figure 6:

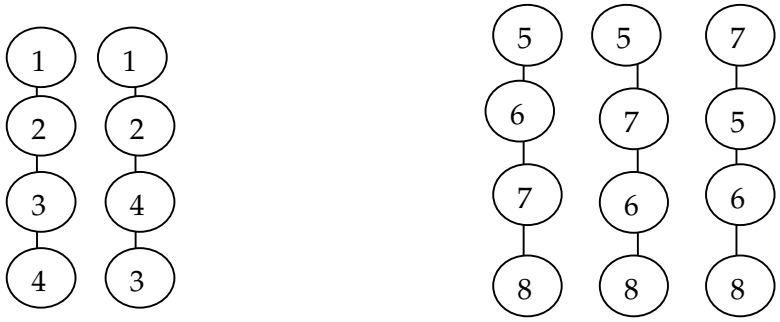


Figure 6: Linear extensions of the two posets shown in Fig 5.
(Note, that the number of objects is equal and is just arbitrary)

The spectrum of an element is the count on how often it occupies a certain rank within the set of linear extensions.

For example:

$\lambda(1) = (0,0,0,2)$ (2 linear extensions where object 1 is placed at rank 4)

$\lambda(4) = (1,1,0,0)$

$\lambda(3) = (1,1,0,0)$

$\mu(7) = (0,1,1,1)$

The symbols λ and μ refer to the right and left poset and their sets of linear extensions, respectively.

What will be the spectrum of (say) "7" if the two posets are pasted to form the new poset, where the object "3" is covering the object 7? Atkinson (1990) gave a (hardly readable) formula for spectra. To apply this formula

- the poset to be fused have to be trees,
- the bridging objects and their cover-relation have to be defined and
- the spectra of the two subtrees have to be known.

In the case discussed above the λ-spectrum of the object 3 and the μ-spectrum of the object 7 are known.

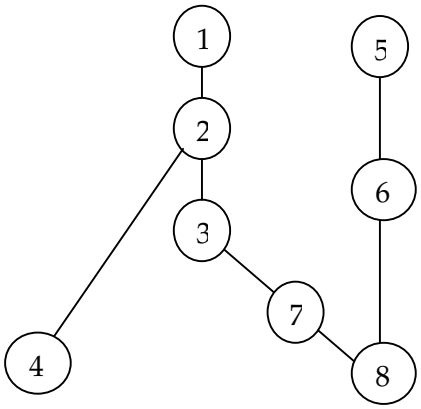


Figure 7: ´The fused posets forming a tree with 8 objects

Let κ(7) be the wanted spectrum of "7" in the combined poset and generally κ(x,r), the count of element x in position r (which means to have the rank r) then:

$$\kappa(x,r) = \sum_{i=z}^{w} \lambda_i(x) \cdot \binom{r-1}{i-1} \cdot \binom{u+v-r}{u-i} \cdot \sum_{j=r-i+1}^{v} \mu_j(y) \qquad \text{Eq.10}$$

with:

$u := card\ P\ (the\ \ poset\ containing\ x)$
$v := card\ Q\ (the\ poset\ containing\ y)$
$x : element\ of\ P\ \text{cov}ered\ in\ P \cup Q\ by\ y$
$y : element\ of\ Q\ \text{cov}ering\ x\ in\ P \cup Q$
$\lambda_i(x) : value\ \ of\ x\ in\ rank\ i\ in\ its\ spectrum\ \lambda\ of\ \ P$
$\mu_j(y) : value\ of\ \ y\ \ in\ \ its\ rank\ \ j\ in\ its\ spectrum\ \mu\ of\ \ Q$
$z := \max(1, r-v)$
$w := \min(u, r)$

λ

P and Q are the sets before the fusion is performed.

u/v : number of objects in P/Q

x/y: the objects, where both graphs will be connected

$\lambda_i(x)/\mu_j(y)$: the number how often x/y have the rank i/j within the set of linear extensions

z: if r-ν is greater than 1, then z = r-ν , else 1

w: if u is less r then w = u else w=r

The rank statistics of the fused system is shown in Table 1:

Table 1: Rank statistics of a "pasted tree-system"

obj.\rk	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	0	0	0	0	0,0513	0,2436	0,705
2	0	0	0	0	0,154	0,385	0,46154	0
3	0	0	0,128	0,333	0,346	0,192	0	0
4	0,192	0,192	0,192	0,192	0,154	0,077	0	0
5	0	0	0,0385	0,0897	0,141	0,192	0,244	0,295
6	0	0,231	0,256	0,205	0,154	0,10256	0,0513	0
7	0,808	0,192	0	0	0	0	0	0
8	0	0,385	0,385	0,1795	0,0513	0	0	0
∑	1	1	1	1	1	1	1	1

As only the elements 7 and 3 are of interest, their spectrum in the combined poset (above quantities multiplied with 78, the number of linear extensions) is given:

h:	0	30	30	14	4	0	0	0
c:	0	0	10	26	27	15	0	0

These quantities are available, directly through Eq. 10.

Let us perform the calculation for 7:

"7" is now inserted as "x" in equation 10.

u=4

v=4

We want to calculate, how often rank 4 is found. Therefore κ(7,4) is to be evaluated.

λ(7) = (0,1,1,1)
μ(3) =(1,1,0,0)

Let us begin with rank = 4

$$\kappa(7,4)=\sum_{i=z}^{w}\lambda_i(7)\cdot\binom{4-1}{i-1}\cdot\binom{u+v-4}{u-i}\cdot\sum_{j=r-i+1}^{v}\mu_j(y)$$

z = max(1,4-4) = 1, w = min(4,4) = 4, j runs from 4-i+1 to 4, i runs from 1 to 4

$$\kappa(7,4)=\sum_{i=1}^{4}\lambda_i(7)\cdot\binom{3}{i-1}\cdot\binom{4}{4-i}\cdot\sum_{j=4-i+1}^{4}\mu_j(y)$$

$$\kappa(7,4)=\lambda_1\cdot\binom{3}{0}\cdot\binom{4}{3}\cdot M_4+\lambda_2\cdot\binom{3}{1}\cdot\binom{4}{2}\cdot M_{34}+\lambda_3\cdot\binom{3}{2}\cdot\binom{4}{3}\cdot M_{234}+\lambda_4\cdot\binom{3}{3}\cdot\binom{4}{0}\cdot M_{1234}$$

$$M_4=\mu_4$$

$$M_{34} = \mu_3 + \mu_4 \qquad M_{234} = \mu_2 + \mu_3 + \mu_4$$

$$M_{1234} = \mu_1 + \mu_2 + \mu_3 + \mu_4$$

μ_3, μ_4, λ_1 are all 0, all other values are = 1

$$\kappa(7,4) = 1 \cdot \binom{3}{1} \cdot \binom{4}{2} \cdot (0) + 1 \cdot \binom{3}{2} \cdot \binom{4}{3} \cdot (1) + 1 \cdot \binom{3}{3} \cdot \binom{4}{0} \cdot 2 = \cdot (3 \cdot 4 + 1 \cdot 2\ ) = 14$$

Why is this formula given so much space? On the one hand it shows that even if the explicit formula is available, it is still hard to understand, how structures of the poset(s) influence the final result. On the other hand, even the "tree-pasting " equation, given by Atkinson, has a sense: Looking later at paste chains (i.e. looking for mutual probabilities of any element of chain 1 and of another element of chain 2) such a formula may be useful. Nevertheless no direct derivations could be made.

3 Empirical procedure

Catalogue of partial order structures and their number of linear extensions
As one can see in the Atkinson process the same structure may appear several times. Thus knowing the number of linear extensions of some typical appearing structures, like $\mathbf{2}^2$ would heavily facilitate the calculations. Thus a catalogue of a series of partial orders was made. For example let be $\mathbf{2}^2$ the basic form, one of several basic forms in the catalogue, then we first label it (Figure 8):

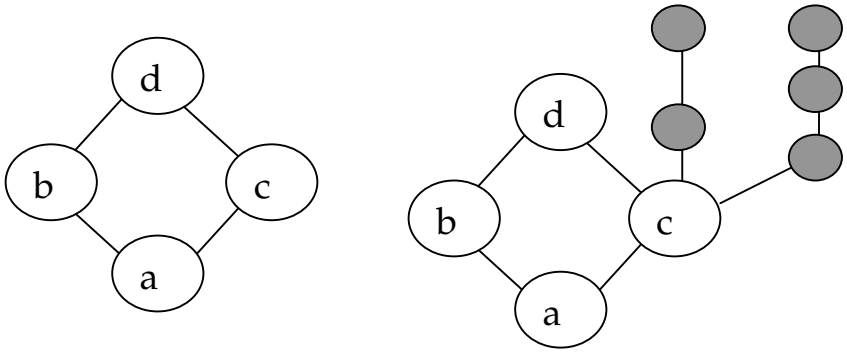


Figure 8: The $\mathbf{2}^2$-system, labelled, and a modified system

It is assumed that the modified poset (right side of Figure 8) arises from adding chains.

By labelling, the positions (vertices) are uniquely defined. Each chain is characterized by upward, "u" or downward "d" and by its length. If chains are thought of as being sticked on the vertices, then the label

of the vertex, the orientation and the length of the chain characterize the modification. If there is no additional chain, then a 0 indicates this. For example, in the case of Figure 8 C(c,u,2) and C(c,u,3) together with the basic poset , the labelled $\mathbf{2}^2$ is sufficient.

In total 61 structures are analysed and similar tables could be presented (Brüggemann, R., unpublished). 11 selected systems are shown in Table 2.

Table 2: Catalogue of linear extensions (extract). In the upper row: a, b, c and d refer to the objects.

a	b	c	d	e(P)
0	0	0	0	2
C(a,u,2)	0	0	0	20
0	C(b,u,1)	0	0	5
0	C(b,u,1), C(b,d,1)	0	0	12
0	C(b,u,1), C(b,u,1),C(b,d,1)	0	0	42
0	C(b,d,2)	C(c,d,3)	0	189
C(a,d,2), C(a,d,3)	0	0	0	20
0	C(b,u,2)	0	0	9
C(a,u,1)	0	0	0	8
0	C(b,u,1)	0	C(d,u,1)	7
0	0	C(c,u,2) C(c,u,3)		270

As can be easily understood, it is not sensible to produce thousands of such numbers, therefore the catalogue is simply thought of to be at hand for theoretical studies.

Empirical relation

In order to derive an estimation of the mutual ranking probability it does not make sense to look for intricate structural parameters, because then one tricky problem will replace another. For example, the dimension of posets may be such a characteristic, which might be useful; however, even if this number would be at hand all the time, it is still unclear whether this number would have a predictive power for that specific problem.

Instead, parameters should be used which are of general character for all partial orders and which can be easily calculated. Hence to predict the mutual probabilities for the two incomparable objects, x and y, $p_m(x>y)$, the following quantities are introduced: $N_u(x)$, $N_u(y)$, $N_u(x \circ y)$, $N_d(x)$, $N_d(y)$, $N_d(x \circ y)$, $Q(x)$ and $Q(x \circ y)$. The index "u" stands for up and the index "d" stands for "down". $N_u(x)$ is the number of comparable objects above x. Similarly, $N_d(x)$, is the number of objects below x. $N_u(x \circ y)$ is the number of objects above x, which are not at the same time above y. In $N_u(x \circ y)$ only the "netto-set" of objects which are above x is taken into account. The $N_d(x)$ quantities are related to the entries of the matrix D (Brüggemann & Halfon, 1995):

$$N_d(x) = D_{xx} \qquad \text{Eq. 11a}$$

$$N_d(x \circ y) = D_{xx} - D_{xy} \qquad \text{Eq. 11b}$$

The quantities Q are defined as follows:

$$Q(x) = (N_u(x)+1)/(N_d(x)+1) \qquad \text{Eq. 12}$$

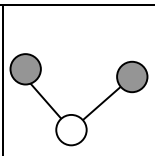
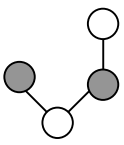
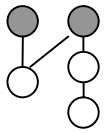
Q(y) and $Q(x_o y)$ are calculated in the same manner.

To establish a relation between $p_m(x>y)$ and the descriptors mentioned above, a training set of 27 posets was established. In Table 3 some examples are given. By the help of the training set of 27 objects a multiregression analysis was performed. As the number of descriptors should be below 27/5 no more than 3 descriptors were used in linear regression analysis at once. By correlation coefficients and F-tests a selection was performed. Equation 13 was most sufficient with respect to statistical tests.

$$prob_Q(x>y) := 1/(1+(Q(x_o y)/Q(y_o x))) \qquad \text{Eq. 13}$$

In the analysis it turned out that neither the number of local incomparabilities of x, U(x) or U(y) nor the length of the maximal chains up and down for x and y respectively nor the sum of $N_d(x)+N_u(x)$ and of $N_d(y)+N_u(y)$ were useful in the estimation of the mutual probabilities of x and y in the partial orders.

Table 3: The two marked objects are those, whose mutual probabilities are calculated. The probability will always be formulated as probability of the left located object being preferred to the right located object. We neglect sub- or superindices.

Hasse diagram						
$N_u(L)$	0	0	0	0	2	0
$N_u(LoR)$	0	0	0	0	1	0
$N_d(L)$	1	1	1	1	0	1
$N_d(LoR)$	0	0	0	0	0	0
Q(L)	0.5	0.5	0.5	0.5	3	0
Q(LoR)	1	1	1	1	2	0
$N_u(R)$	0	1	0	1	1	0
$N_u(RoL)$	0	1	0	1	0	0
$N_d(R)$	1	1	2	0	0	3
$N_d(RoL)$	0	0	1	0	0	2
Q(R)	0.5	1	0.333	2	2	0.25
Q(RoL)	1	2	0.5	2	1	0.333
probQ	0.5	0.666	0.333	0.666	0.333	
probQ+	0.5	0.666	0.4	0.80	0.4	
p_m	0.5	0.666	0.4	0.8	0.4	0.333

If the real mutual ranking probability $p_m(x>y)$ is put into a regression analysis with $prob_Q(x>y)$ as predictor,

$$p_m(x>y)_{estim} = prob_Q(x>y)*a + b \qquad \text{Eq. 14}$$

then a = 0.97 and b = 0.01. The correlation coefficient, $R_{DF}^2 = 0{,}95$, reveals that the variance of $p_m(x>y)$ is well explained by $prob_Q(x>y)$ but the coefficients, a and b, especially a, show that there is some bias. In Figure 9 $p_m(x>y)$ versus $prob_Q(x>y)$ is shown.

However, theoretically, b should be zero, because $prob_Q$ is intended to be p_m. Therefore the analysis can be repeated, excluding the constant, b, within the regression analysis. This exclusion is supported by the rather low value of b found in the statistical analysis, mentioned above. In this case the correlation coefficient, R^2, becomes, 0,99 and a becomes 0,99. It is thus a very good assumption to estimate p_m by using $prob_Q$. Note that R^2 cannot compared with R^2_{DF}, as b=0 is an additional constraint.

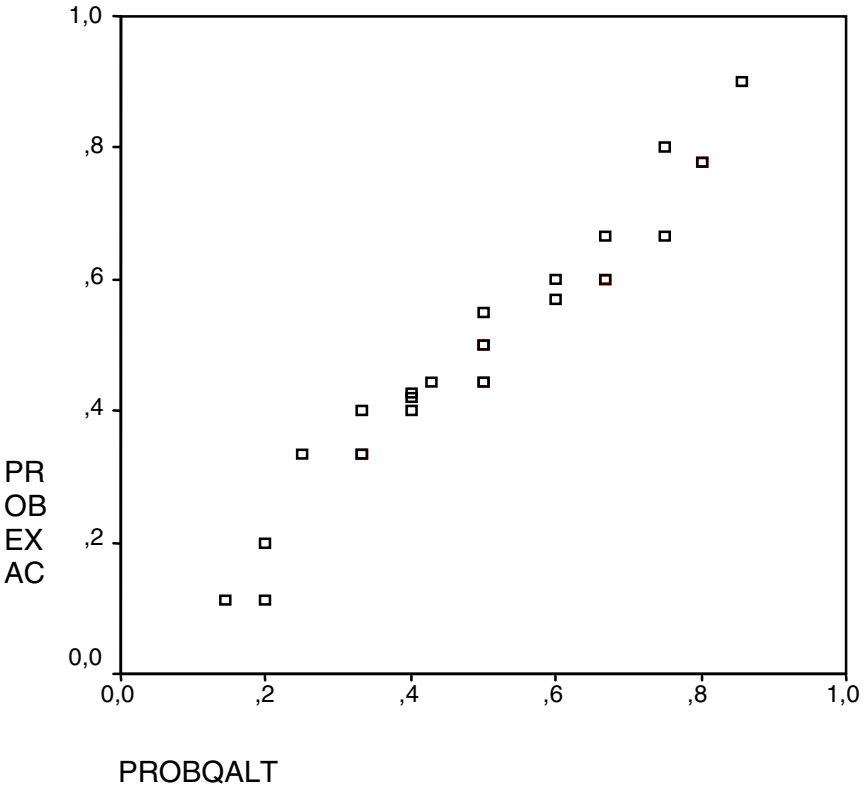


Figure 9: PROBQALT = 1/(1+(Q(xoy)/Q(yox))) vs mutual probabilities calculated by WHASSE software, corresponding to the procedure, described in the section: "Methodological development".

The equation 13 has Q(xoy) and Q(yox) as predicting quantities. It is quite obvious, also to look for Q(x) and Q(y) as descriptors, where the common objects above and below x,y are **included**:

$$prob_{Q+}(x>y) := 1/(1+(Q(x)/Q(y))) \qquad \text{Eq. 13'}$$

In this case the correlation coefficient, R^2, becomes 0.98 and a becomes 1.16. As one can see, the correlation is good, however the bias is worse than that of the former model. Note that the Q-quantities can be calculated by the help of the WHASSE software via matrix D.

4 Estimations of Mutual Ranking Probabilities by the ranking algorithm

A similar attempt to calculate mutual probabilities is described in Lerche et al., 2003. The algorithm can be derived as follows:

1. The maximal and minimal rank is derived from $N_u(x)$, $N_d(x)$, $N_u(y)$ and $N_d(y)$
2. It is counted how often it is possible to get a rank higher than that of the other element
3. Equal ranks are not counted.

The number of rank combinations where one element has higher ranks than the other divided by all possible combinations (excluding equality) leads to the desired quantity.

The four steps are best described by a schematic example. Let x and y be the incomparable elements, whose probability of x > y is to be calculated.

The maximal rank of ($R_u(x)$) is then:

$R_u(x) = N - N_u(x)$

where N is the total number of elements. Analogously: $R_u(y) = N - N_u(y)$

The minimal rank $R_d(x)$ and $R_d(y)$ respectively is:

$R_d(x) = 1 + N_d(x)$. Analogously $R_d(y) = 1 + N_d(y)$.

Now two rankings are to be compared, i.e. it will be counted, how often the rank of x is greater than that of y. Let us assume that $R_u(x) > R_u(y)$. Then two situations may appear:

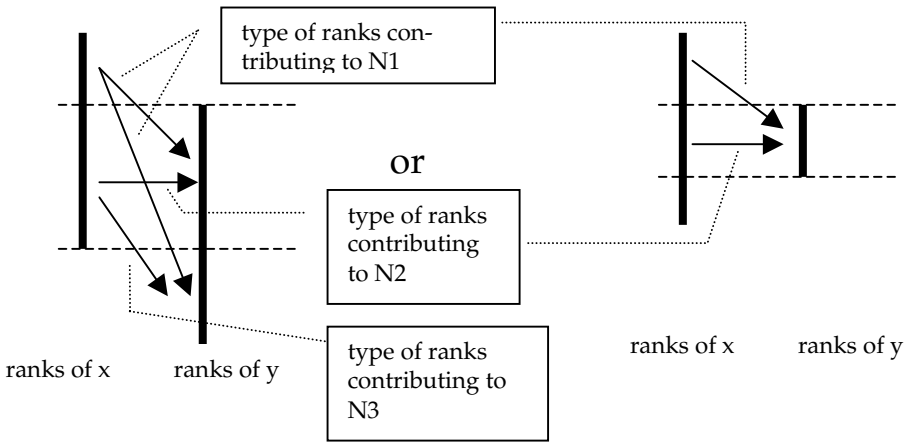


Figure 10: Comparison of ranks and their counts

N1: ranks of x are all greater than those of y

N2: ranks of x have the same range as the ranks of y

N3: ranks of y are all less than those of x.

If x > y is to be determined, then ranks of x lower than those of y are not counted. In the overlapping part, where both, x and y have the same range of ranks, then only those combinations are to be counted, where rank x is > rank y.

Therefore the number of ranks, where x > y is composed of three contributions (see Figure 10):

$$N1 = (R_u(x) - R_u(y))*(R_u(y) - R_d(y)+1)$$

$$N2 = C*(C-1)/2$$

with $C = (R_u(y) - \mathrm{Max}(R_d(x), R_d(y)) +1)$

$$N3 = (R_u(y) - \mathrm{Max}(R_d(x), R_d(y)) +1)*(\mathrm{Max}(R_d(x), R_d(y)) - R_d(y)) \quad \text{Eq. 15}$$

The sum leads to all rank combinations corresponding to x > y.

The sum N1+N2+N3 is to be compared with all possible combinations with the exception of the equalities, N_D.

$$N_D = (R_u(x) - R_d(x) +1)*(R_u(y) - R_d(y)+1) - C \quad \text{Eq. 16}$$

Therefore the heart of the algorithm is:

$$\text{prob}(x>y) = (N1 + N2 + N3)/N_D \quad \text{Eq. 17}$$

Consequently, the analysis by this algorithm needs the following information:

$N_u(x)$, $N_u(y)$, $N_d(x)$, $N_d(y)$ and N. However N cannot be totally independent of $N_u(x)$, $N_u(y)$, etc. Instead one may write:

$$N = N_u(x) + N_d(x) + N_u(y) + N_d(y) + 2 + N_s - X$$

$$X = N_u(x \cap y) + N_d(x \cap y) \quad \text{Eq. 18}$$

N_u: number of common elements upwards for both elements x and y, N_d: analogously, downwards; N_s: all other elements, which are not taken into regard by $N_u(x)$, $N_u(y)$, $N_d(x)$, $N_d(y)$, $N_u(x \cap y)$, $N_d(x \cap y)$. For example in the Hasse diagram as follows:

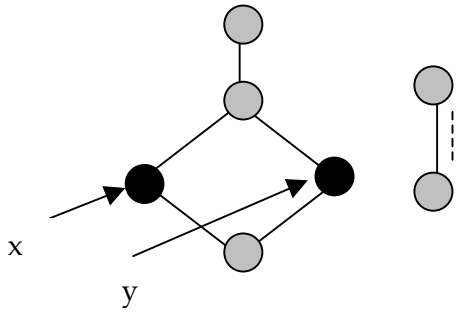


Figure 11: N = 7, varying the number of elements in the chain (right side), will change N and therefore according to Eq. 17 prob(x>y), but $p_m(x>y)$ will nevertheless be constant.

The equations, given above show that the crucial quantities are those which influence N but do not influence $R_u(x)$, $R_d(x)$, $R_u(y)$, $R_d(y)$. These are $N_u(x \cap y)$, $N_d(x \cap y)$ and N_s. In Table 4 the different situations are summarized.

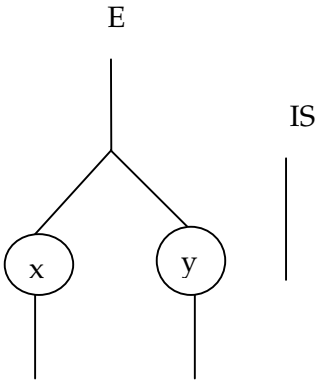


Figure 12: Model systems to analyze Eq. 17. Indeed if the quantity $N_u(x \cap y)$ has a non-negligible contribution, then the formalism above lead to a wrong dependency on N and the same is true, if N_s is varied, as is shown in Figure 11. Furthermore if simple model systems are analyzed, there are still some discrepancies, which show the approximate nature of the ranking approach (Figures 12, 13). In Figure 12: In all cases the number of objects above x and y is 3. The number of objects below x and y is held constant too (Below x: 3, below y: 1). However the number of **common** objects above x and y is varied. By changing the number of elements in the hierarchy formed by the chain, the total number of objects can be held constant.

Table 4: Analysis of Eq. 17 by means of the poset shown schematically in Figure 12. Second column, "E": common objects above x and y, third column "ISO": number of objects in the chain ISO. Fifth column: $prob_{D\_ISO}(x>y)$ without consideration of the objects of ISO. Sixth column: $prob_D(x>y)$ taking into regard the variation of the number of objects in ISO. $N_{\_ISO}$: number of objects without ISO, N_{ISO}: Number of objects with ISO. pm(x>y): exact values from explicit calculation by the WHASSE software.

Case	E	ISO	$p_m(x>y)$	$prob_{D\_ISO}(x>y)$	$prob_D(x>y)$	$N_{\_ISO}$	N_{ISO}
1	3	3	0.667	0.75	0.643	9	12
2	2	2	0.714	0.7	0.643	10	12
3	1	1	0.738	0.666	0.643	11	12
4	0	0	0.753	0.643	0.643	12	12

Figure 13 shows that neglection of -for example- $N_u(x \cap y)$ - , or of N_s will lead to severe deviations.

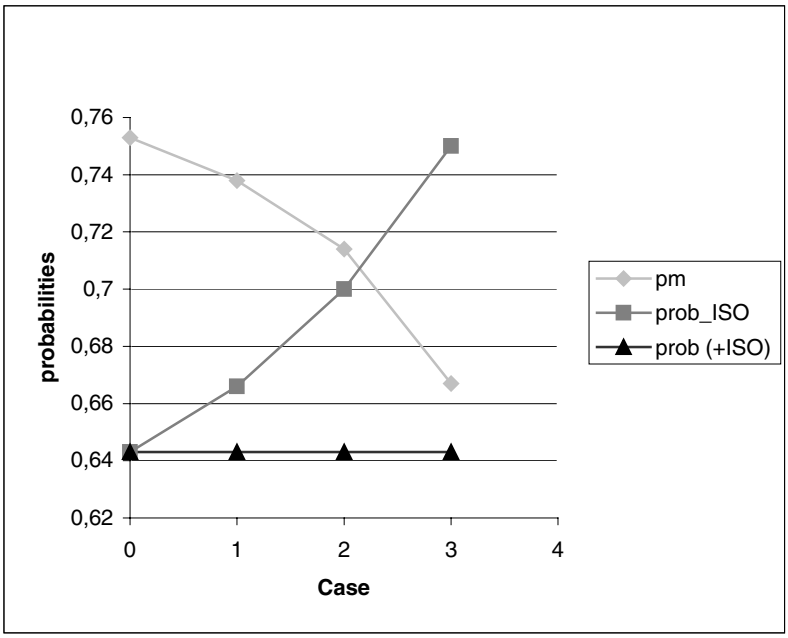


Figure 13: Analysis of Eq. 17, see Table 4.

5 Theoretical deviations of Estimations of Mutual Ranking Probabilities, Investigations in model-partial ordered sets

Within this section some more intricated formulas are derived. If the notation $N_d(xoy)$ etc. is maintained, some of the formulas are hardly readable. Therefore the actual notation is introduced for each model-system separately. Firstly several model systems are discussed. Afterwards a comparison between exact results and the semi empirical method (section 3) is performed.

5.1 Double chain systems (also called: "double sausage system" or CC-system)

Double chain partial ordered sets are the first, which are investigated, because by equation 8 at least the total number of linear extensions is easily obtained.

It is more difficult to find the number of linear extensions where one element in one chain is supposed to be higher ranked than another element in the other chain.

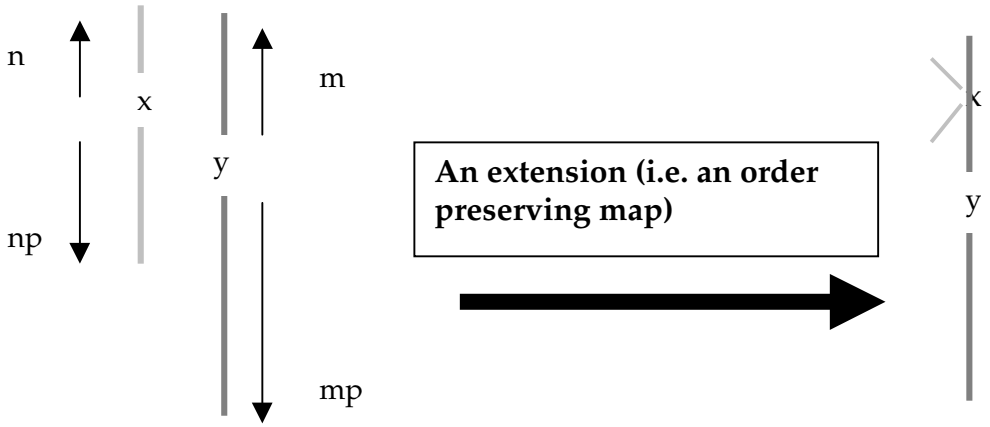


Figure 14: Double chain system, and analysis of the mutual ranking probability of x>y. Left side: An intermediate step is shown. Object x is located somewhere above object y. Now the chains downward and upward, respectively, of x are to be mixed with the objects of the y-containing chain.

If the object x is located somewhere above y, then the objects originally below x and the objects originally above x have to mix appropriately with the objects of the right chain. To examine this the notation should be changed (see Figure 15):

If x is located at the i, the position above y, the Eq. 8 can be applied for respectively the double chain above x and below x.

Thus for the lower part the $N_d(x) = np$ objects below x are going to be mixed with $N_d(y) + i - 1 = mp + i - 1$ objects below x (we write instead of $N_d(y)$: mp), and for the upper part, once again Eq. 8 can be applied, mixing n objects above x in the original left chain with m-i objects in the right chain. We simplify the notation in order to get readable equations.

Therefore the number of linear extensions with x > y ($L_{CC}xy$) is given by:

$$L_{CC}xy = \sum_{i=0}^{m} \frac{(np + mp + 1 + i)!(n + m - i)!}{np!n!(mp + 1 + i)!\cdot(m - i)!} \qquad \text{Eq. 19}$$

As for the total number of linear extensions it can be found (now using the notation of Figure 15)

$$LT_{CC} = \frac{((n + 1 + np) + (m + 1 + mp))!}{(n + 1 + np)!\cdot(m + 1 + mp)!} \qquad \text{Eq. 20}$$

the problem for the double chain – system is solved.

Note that we use L as symbols instead of e(P) because we are referring to very specific systems. The specific probability, calculated by this model, will be called

$$\text{prob}_{CC}(x>y) = L_{CC}xy / LT_{CC} \qquad \text{Eq. 21}$$

Equation 21 can be applied in several ways:

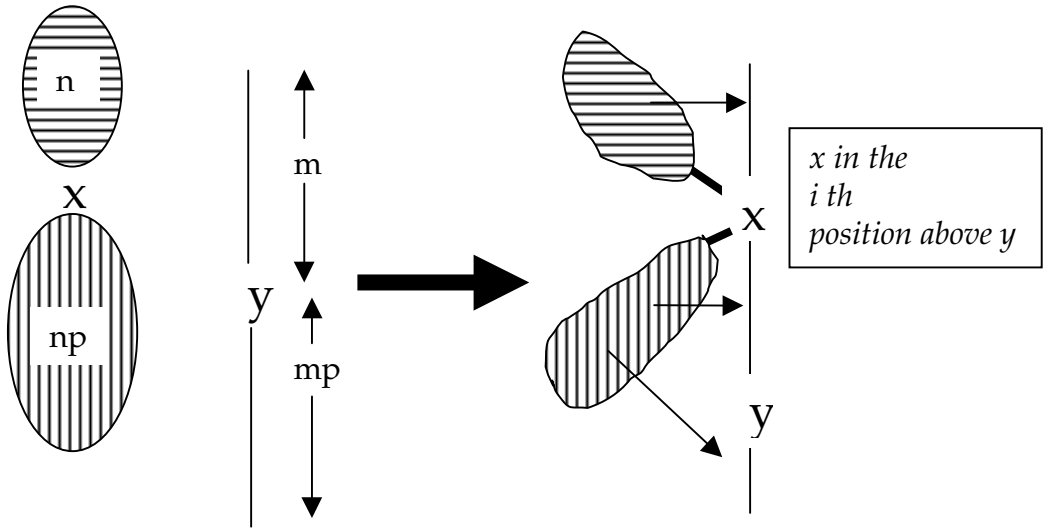


Figure 15: Scheme to explain the "mixing effects". An interim extension of the poset, shown on the left side.

1. $Prob_{CC}(x>y)$ could be a substitute of $prob_Q$, because it does not derived empirically. However, its application may only be restricted, as the model system the "double chain" may only - in a restricted way - represent the variety of Hasse diagrams.
2. The equations 19-21 may be considered as a starting point to perform some simplifications,

Finally the equation itself is not the main use, but the way to derive such kind of equations is of interest.

First of all it is of interest, how well $prob_{CC}$ can be used to estimate the exact mutual ranking probability $p_m(x>y)$ in systems shown in Figure 16. We call these systems the AW-systems.

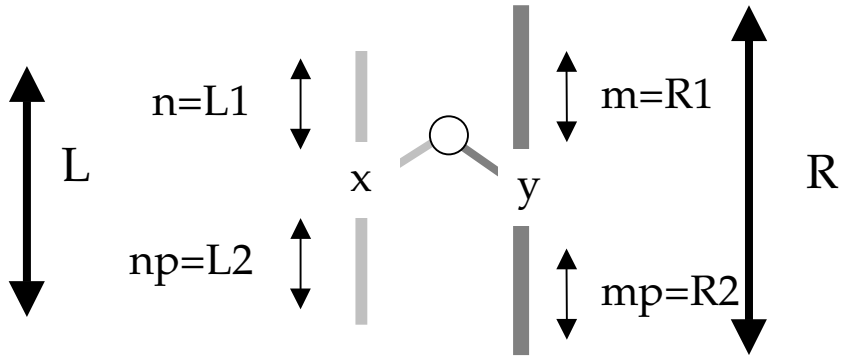


Figure 16: The AW-system. L=L1+1+L2, R=R1+1+R2

By a statistical analysis, where $prob_{CC}$ was a predictor for p_m, the equation

$p_m(estim) = a + b^* prob_{CC}$

was found, with:

$R^2_{DF} = 0.997$, $F = 3200$, $N_{records} = 11$

$a = 0.057 \pm 0.008$

$b = 0.888 \pm 0.016$

See also figure 17.

Thus $prob_{CC}$ seems to be a fairly good estimator, even for systems like the AW-systems, for which it was not derived.

The results and inputs of the AW-system are shown in Table 5.

Table 5: Inputs and results of the AW-system

L1	L2	L	R1	R2	R	$prob_{CC}$ p_m	$prob_Q$	
1.00	3.00	5.00	2.00	1.00	4.00	.8330	.8000	.7500
4.00	1.00	6.00	4.00	2.00	7.00	.3430	.3500	.4000
3.00	1.00	6.00	3.00	2.00	6.00	.3480	.3560	.4000
2.00	.00	3.00	1.00	1.00	3.00	.2000	.2310	.2500
2.00	.00	3.00	2.00	1.00	4.00	.2860	.2940	.3330
4.00	4.00	9.00	4.00	3.00	8.00	.6010	.5950	.5560
1.00	2.00	4.00	2.00	1.00	4.00	.7570	.7200	.6920
4.00	2.00	7.00	1.00	2.00	4.00	.1970	.2556	.2860
1.00	3.00	5.00	3.00	1.00	5.00	.8970	.8600	.8000
4.00	2.00	7.00	2.00	2.00	5.00	.3110	.3470	.3750
1.00	1.00	3.00	1.00	1.00	3.00	.5000	.5000	.5000

The variances are quite well explained. However, there is still a bias as the coefficients a and b deviate remarkably from 0 and 1 respectively.

Note that in the study of the AW-system the common objects are **not** considered.

In the following SPSS-results $prob_{Q+}$ is calculated too. Here the common objects are taken into account. The aim is to decide whether $prob_{CC}$, $prob_Q$ or $prob_{Q+}$ is a good predictor for p_m if the AW-system is considered.

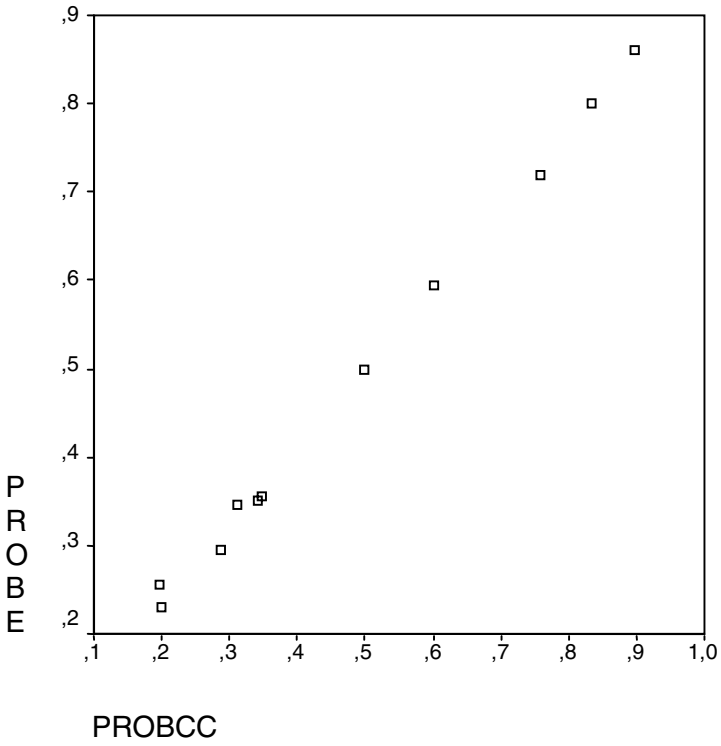


Figure 17: $Prob_{CC}$ applied on the AW-system. PROBEXACT = p_m

The Table 6 summarizes the main results.

Table 6: Main statistical results of testing the AW-system

model	R^2_{DF}	F	a	b	comments
$prob_{CC}$	0.997	3200	0.006	0.888	bridging object not included
$prob_{Q}$	0.993	1329	-0.094	1.189	xoy, yox
$prob_{Q+}$	0.994	1469	-0.164	1.328	x,y

All models have a bias. It turns out that the variance is better explained by $prob_{Q+}$, however, the bias is increased in comparison to $prob_{Q}$. The estimation by the CC-system, i.e. by $prob_{CC}$ seems to be the best.

5.2 The "double-sandwich-system" (ACAC)

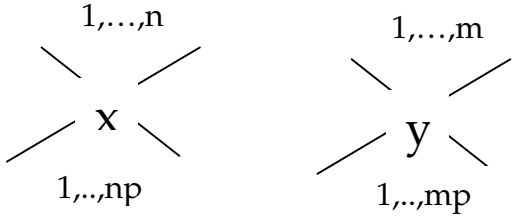


Figure 18: Double - sandwich-system also called ACAC-system.

The ACAC-system is called a double sandwich system because it consists of 4 antichains (two above and below x, two above and below y). Object x has n incomparable covering objects and np incomparable objects as covered. Similarly, y is covered by m incomparable objects and covers mp other mutually incomparable objects.

Once again, LT_{ACAC} and $L_{ACAC}xy$ are to be determined:

$$LT_{ACAC} = n!np!m!mp! \cdot \frac{(n+np+m+mp+2)!}{(n+np+1)!\cdot(m+mp+1)!} \qquad \text{Eq. 22}$$

It is useful to write as an abbreviation:

$$F := n!\cdot np!\cdot m!\cdot mp! \qquad \text{Eq. 23}$$

$$L_{ACAC}xy = F \cdot \sum_{i=0}^{m} \frac{(m-i+n)!\cdot(mp+1+i+np)!}{(m-i)!\cdot(mp+1+i)!\cdot n!\cdot np!} \qquad \text{Eq. 24}$$

In order to calculate the $prob_{ACAC}(x>y)$ the quotient $L_{ACAC}xy/LT_{ACAC}$ is to be formed:

$$prob_{ACAC}(x > y) = \frac{L_{ACAC}xy}{LT_{ACAC}} \qquad \text{Eq. 25}$$

The factor F is cancelled out. Therefore the same equation is found for $prob_{ACAC}(x>y)$ as for $prob_{CC}(x>y)$.

Therefore we come to the following intuitively obvious conjecture:

To estimate mutual probabilities it seems that the dominant factor is the number of objects above and below. It plays no role whether these objects are forming chains or antichains.

This is especially of importance for the algorithm to generate randomly formed linear extensions (Lerche et al. 2003), because it does not seem important what the relations will be among the objects above and below. Note that this result does not include cases where the objects x and y are incomparable but indirectly coupled by common other objects. This will be seen immediately, if the third general system is analysed:

5.3 Down- double-chain-System

The elements x and y have each a downward chain. The number of chain-elements below x is np, that of y: mp.

$$LT_{CCdown} = \frac{(np + mp + 2)!}{(np + 1)!(mp + 1)!} \quad \text{Eq. 26}$$

$$L_{CCdown}xy = \frac{(np + mp + 1)!}{np!(mp + 1)!} \quad \text{Eq. 27}$$

$$prob_{CCdown}(x > y) = \frac{np + 1}{np + mp + 2} \quad \text{Eq. 28}$$

In the corresponding "up-System" the equivalent similar formula is obtained, just replacing mp by n and np by m.

5.4 Up- and down-ACAC-system

The same formulas as in section 5.3 can be found if the double sandwich system is correspondingly simplified. The resulting system contains either a covering antichain or (exclusively) a covered antichain (Figure 19):

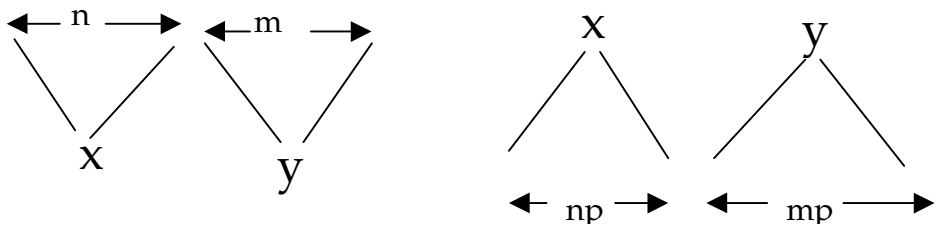


Figure 19: The "up-ACAC-system" (left side) and the "down-ACAC-system (right side)

5.5 The h-system (Figure 20):

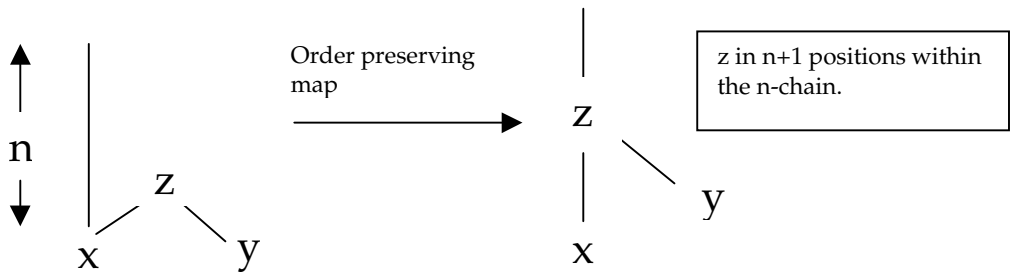


Figure 20: Left side: x and y are still incomparable but they cannot freely interchange their places in the GRM because there is a common object z covering both. Above x there are n additional objects in a chain. Right side: Extensions of the h-system.

$$LT_h = \frac{(n + 1) \cdot (n + 4)}{2} \quad \text{Eq. 29}$$

$$L_h xy = n + 1 \quad \text{Eq. 30}$$

Eq. 30 is easily understandable because the bridging object z has n+1 position in the n-chain. Therefore there are n+1 extensions. In each of them only one is a realization of x>y. Therefore the number of realisations only depend on n, the length of the chain above x.

5.6 The "M-system" (Figure 21):

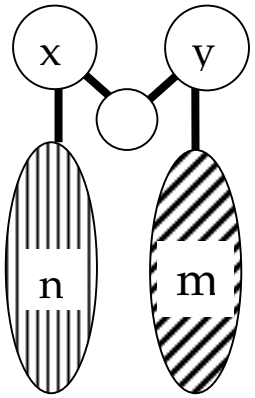


Figure 21: The M-system. The objects x and y are incomparable but they cannot move independently because of the bridging object. Below x there is a chain with n objects, below y there is a chain with m objects. The objects x and y cover the bridging object.

By applying the same technique as for the double-chain – system, together with the Atkinson-process one comes up to:

$$LT_M = \frac{(m+n+4)\cdot(n+2+m)!}{m!\cdot n!\cdot(n+2)\cdot(m+2)} \qquad \text{Eq. 31}$$

Note that faculties are **not** applied for the terms (m+n+4) and (n+2). (m+2)!.

The linear extensions for x>y are found to obey the following formula:

$$L_M xy = (m+1)\cdot\frac{(n+2+m)!}{(m+2)!\cdot n!} \qquad \text{Eq. 32}$$

This equation can be rearranged to give it a more symmetrical form:

$$L_M xy = \frac{(n+2+m)!}{n!\cdot m!}\cdot\frac{1}{m+2} \qquad \text{Eq. 33}$$

The expression cannot be completely symmetric in m, n because x>y is demanded. I.e. exchanging the chains will influence the number of linear extensions.

The $prob_M(x>y)$ is once again easily found by forming the quotients from Eq. 31 and 33:

$$prob_M(x > y) = \frac{n+2}{m+n+4} \qquad \text{Eq. 34}$$

5.7 A-system

Its graphical representation is shown in Figure 22.

By applying the Atkinson process (Eq. 7) one can easily show that:

$$prob_A(x > y) = prob_{CC}(x > y) \qquad \text{Eq. 35}$$

Therefore, there are no own formulas for this system.

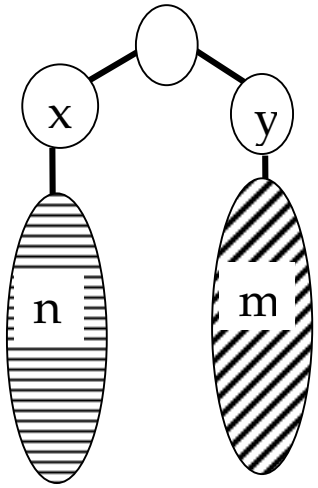


Figure 22: The A-system. The incomparable objects x and y are bridged by another object, but in that way that this object covers the objects x and y respectively.

6 Systems which cannot be handled by "above-below" calculations

6.1 ISO-Systems

In ISO-systems we are interested in studying the role of isolated objects.

Let us look for a partial ordered set, which contains some isolated objects i_1, i_2, i_3, i_n (which we gather in an ISO set) and the residual part, C, which is either consisting of several nontrivial hierarchies or of exactly one hierarchy.

If one analyses the mutual probabilities of proper maximal objects (i.e. maximal objects being in C) one would like to know whether these mutual probabilities are affected by the presence of the set ISO. A numerical study with the partial ordered sets was performed. The corresponding Hasse diagram is shown in Figure 23.

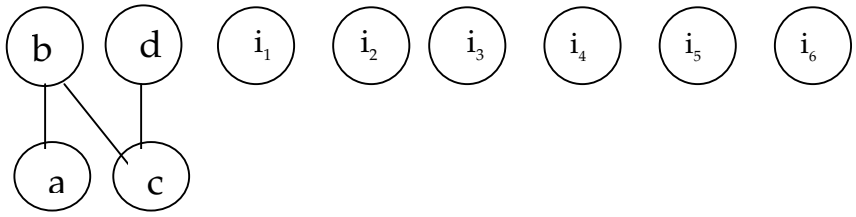


Figure 23: A partial ordered set consisting of a system of objects, which are connected and some isolated objects "i_1, i_2,i_3...."

In Figure 23 four objects are forming the connected part of the partial order denoted c and are added. The numerical study shows that the presence of ISO does not affect the mutual probabilities among objects of C. This is quite clear, because each isolated object can be below or above the objects of interest and their effect cancels out.

Therefore:

$$prob_C\,(x>y) = prob_{C\cup ISO}(x>y) \qquad \text{Eq. 36}$$

Now the next question is: What is the mutual ranking probability of an object x of ISO and y of C, if one knows the system C?

$$p_m(\,x > y)\ (x \text{ of ISO, } y \text{ of C}) = 1 - Rk^{av}(y)/(N_C+1) \qquad \text{Eq. 37}$$

$Rk^{av}(y)$ is the averaged rank of y in GRM. N_C is the number of objects in C.

It seems that this can be generalized, but here additional work should be done.

6.2 Generalization

So far we have studied several cases, in which the number of objects has conceptually influenced the mutual ranking probability. I.e. thought of as formula depending on $N_u(x)$, $N_d(x)$, $N_u(y)$, $N_d(y)$ the variations in these descriptive numbers are more or less well represented by all the equations given above. We call this kind of calculations an "above-below" calculation, because it is purely based on counting those objects, which are above or below the two objects of interest.

The isolated objects, however, show that one has to be careful with respect to those partial ordered sets where a number of objects can

influence the mutual ranking probability, which are not included in the descriptors like $N_u(x)$, $N_d(x)$, $N_u(y)$, $N_d(y)$.

Example of partial orders where the $N_u(x)$, $N_d(x)$, $N_u(y)$, $N_d(y)$-formalism cannot work are shown in figure 24. These systems are called the CCC-systems because there are several chains, two of them include x and y respectively, for which the mutual probability should be estimated. The other chains are incomparable to x and y respectively.

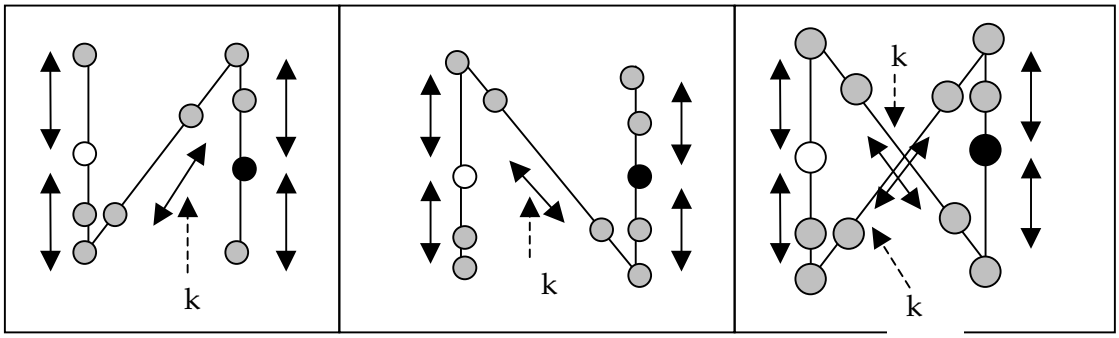


Figure 24: Examples of CCC-systems, where k (and k' respectively) objects are influencing the mutual probabilities but are nevertheless not accounted for in any "above-below" process or using the descriptors $N_u(x)$, $N_d(x)$, $N_u(y)$, $N_d(y)$. Clearly there are many and much more complicated configurations.

6.3 CCC-systems

Remark: There is - up until now - no general structural concept. Therefore this section will be subdivided to find a better way through all cases discussed so far.

6.3.1 CCC-G-system (Figure 25)

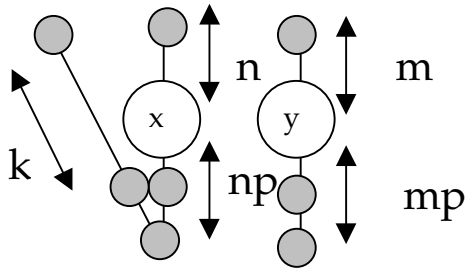


Figure 25: Example of a partial ordered set, where k objects may influence the mutual ranking probability of x vs y.

The system shown in Figure 25 is a member of a CCC-system because it consists of three chains: One with k elements, one containing x and one -isolated- containing y.

Note that the k objects are not contained in any $N_u(x)$, $N_d(x)$, $N_u(y)$, $N_d(y)$ -formalism discussed so far.

If the chain of k objects has no connection to either the chain of x or to that of y (i.e. is an isolated hierarchy) then clearly

$p_m(k) = p_m(0) = prob_{CC}$ Eq. 38

By means of WHASSE - software p_m was calculated by varying k and the environment of object x. The results are summarized in figure 26.

It is found that k indeed influences the p_m-values. That means that none of all algorithms mentioned above will model this k-dependence. Furthermore it seems as if the environment around the objects of interest (here object x) determines, how sensitive the variation due to the number of elements in the side chain, i.e. k, is.

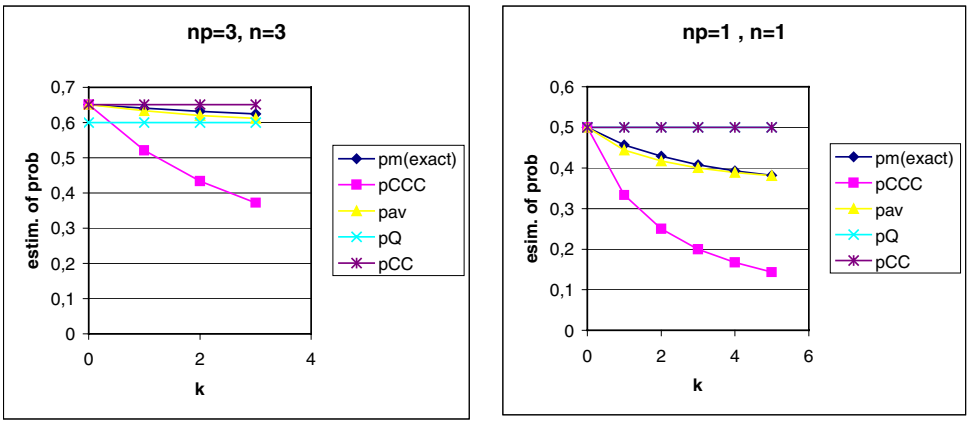


Figure 26: Estimations of mutual probabilities by means of different approaches (see text below) p_m(exact) is calculated explicitly by WHASSE software; it has nothing to do with p_m of equation 38. The Hasse diagram of the corresponding partial order is shown in Figure 25.

In order to understand how the lengths of chains containing x or y influence the p_m(k)-function a simple model is more closely studied.

As an idea one could try to estimate the mutual ranking probability of the CCC-G-system as follows:

First the k objects are to be mixed with those of the x- chain, after that check those linear extensions, where x > y.

The k-chain can as a whole get n+1 sites (above x) and np sites (below x). Thus formally the x-chain will be elongated and then the mixing with the y-chain is to be calculated.

If one performs this algorithm, the kind of distribution of the k objects between the upper and the lower part of the x-chain is not regarded. Therefore clearly the above mentioned algorithm is only an approximation.

The resulting formula:

$$p_m(k) \approx p_m(0) \cdot \frac{(np+k)!\cdot(n+1)!}{np!\cdot(n+k+1)!} \qquad \text{Eq. 39}$$

Taking into account that the length of the original x chain has an influence on the effect of the k-chain, one finally comes to the following first draft:

$$p_m(k) = p_m(0) \cdot (f(n, np, m, mp) + (n + np)) / (n + np + 1) \qquad \text{Eq. 40}$$

$$f(n, np, m, mp) = \frac{(np + k)! \cdot (n + 1)!}{np! \cdot (n + k + 1)!} \qquad \text{Eq. 41}$$

The motivation for this formula is just a weighting: $p_m(0)$ is weighted dependent on the chain length. The estimation by the formula 39 is called p_{CCC} (see Figure 26). The estimation by equations 40 and 41 is called p_{av}. All these estimations together with the applications by the CC- or Q-formalism are shown in Figure 26.

6.3.2 CCC-1-system (Figure 27)

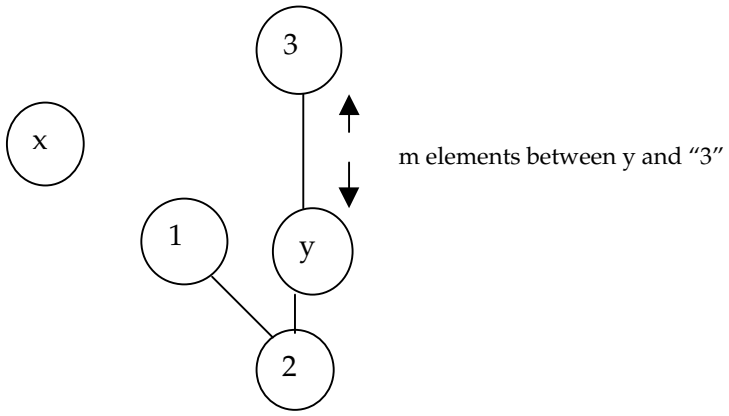


Figure 27: Model-system (CCC-1-system)

With object 1 (p_{1CCC}) and without object 1 (p_{CCC}) are found to be:

$$p_{1CCC} = \frac{(m+2) * (m+3)}{(m+2) * (m+6) + 3} \qquad p_{CCC} = \frac{m+2}{m+4} \qquad \text{Eq 42}$$

The figure 28 shows that with increasing m the mutual probability p_m increases.

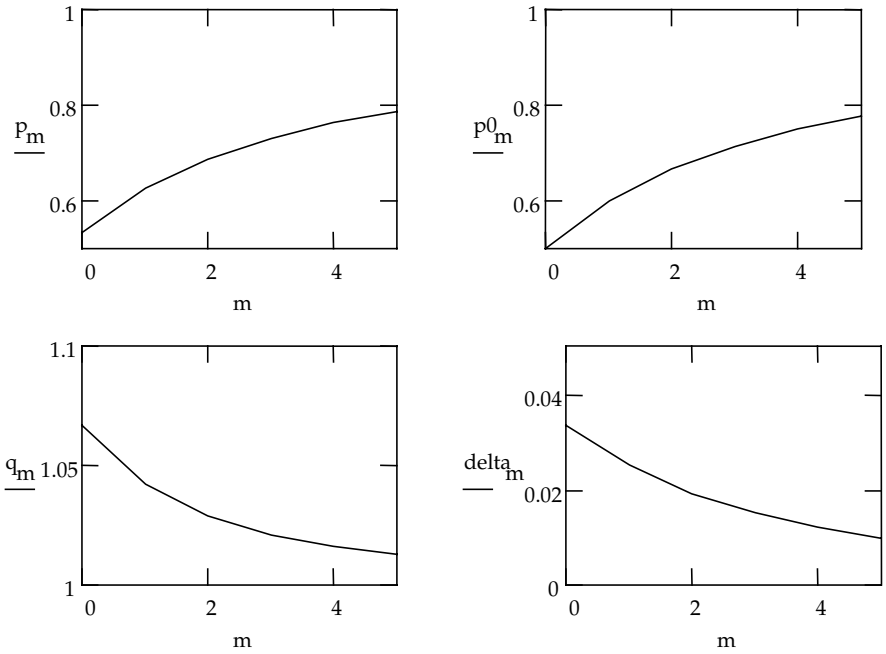


Figure 28: In the graphic: $p_m = p_{1CCC}$, $p0_m = p_{CCC}$, $q_m = p_{1CCC}/p_{CCC}$ and $delta_m = p_{1CCC}-p_{CCC}$

The effect of the object „1“, i.e. of a side chain can be more concisely described by the formula:

$$p_{1CCC} = p_{CCC} \cdot \frac{1}{1-\frac{1}{(m+4)^2}} \qquad \text{Eq. 43}$$

The presence of the sidechain (here realized by only one object) increases the mutual probability.

It may be useful to generalize the CCC-1- system to a CCC-k-system. This is described in the next section.

6.3.3 CCC-k-system (Figure 29)

Instead of only one element in the sidechain, now k elements are located. Figure 29 shows the configuration:

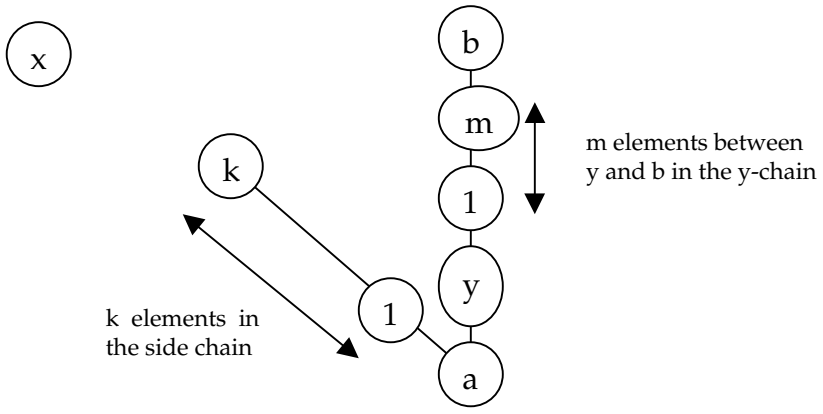


Figure 29:.The parameters of the CCC-k-system.

With the help of equation 7 and the mixing extension technique (Eq. 9) it can be derived:

$$LT(k,m) = (k+m+4) * \frac{(k+m+2)!}{k!(m+2)!} \qquad \text{Eq. 44}$$

It is slightly simpler to calculate first the number of linear extensions, where x is below y (Lyx) and after that deriving $p_m(x>y)$.

$$Lyx(k,m) = \sum_{i=0}^{k} (i+2) \cdot \frac{(k-i+m+1)!}{(k-i)!(m+1)!} \qquad \text{Eq.45}$$

From both equations, first $p_m(y<x)$ can be derived:

$p_m(y>x) = Lyx(k,m)/LT(k,m)$

$p_m(x>y) = 1-p_m(y>x)$

As the figure 30 shows, there is an increasing effect by k and by m. As in the more simple system CCC-1 the influence of k is decreased as the number of elements in the y-chain, m, is increased.

Table 7: Dependence of $p_m(x>y)$ as function of k and m

k	m=0	m=1	m=2	m=3	m=4
0	0,5	0,6	0,667	0,714	0,75
1	0,533	0,625	0,686	0,729	0,762
2	0,556	0,643	0,7	0,741	0,771
3	0,571	0,656	0,711	0,75	0,779
4	0,583	0,667	0,72	0,758	0,786

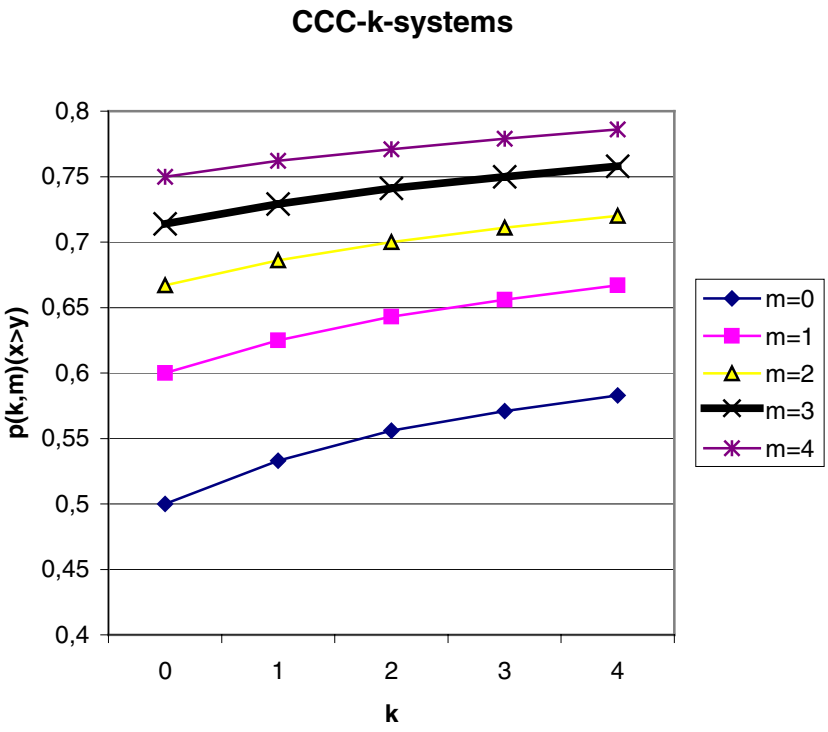


Figure 30: Mutual probability of the CCC-k-system as function of k and m. k as abcissa.

By the formulas given above 49 cases were calculated with k varying from 0 to 6 and m varying from 0 to 6.

A statistical analysis shows that a rather good estimation of $p_m(x>y)$ is possible:

$$\left(\frac{1}{p_{m,est}}\right) = a + \frac{b}{m+1} + \frac{c}{k+1} + \frac{d}{(k+1)\cdot(m+1)} + e\cdot(m+1) + f\cdot(k+1) \qquad \text{Eq 46}$$

a=1.426±0.016, b=0.328±0.015 , c=-0.0657±0.015 , d=0.379±0.023,

e=-0.0294±0.002, f=-0.01288±0.002

The statistical parameters: R^2_{adj} = 0.996 (stepwise inclusion of parameters by F-tests)

F=2289.7 , N_{rec} = 49

It is somewhat difficult to generalize these results. Numerical experiences together with the theoretical results discussed above may be summarized as follows.

Given a CCC-k-system, then

1. Increasing the number of objects above x: prob(x>y) will decrease.
2. Increasing the number of objects below x: prob(x>y) will increase.
3. Locating y in the y-chain in higher positions: prob(x>y) will decrease.
4. The effects, mentioned in 1-3 may be estimated by algorithms 3.-5. These effects seem to be the dominant ones.
5. The effect of the k-sidechain in the CCC-k-system is subdominant. Increasing k will lead to an enhancement of approximately 10% of the probability of the k-less system.

6.3.4 The CCC-N-system (Figure 31)

If the k-chain is connected to both, the x- and the y-chain, then a system arises, which is called CCC-N-system.

As model system the partial ordered set, visualized by the following Hasse diagram was used. In Table 8 the data, which were all calculated with the WHASSE software are shown.

Table 8: Exact mutual probabilities of the system, shown in Figure 33

k	$N_d(x)=3$	$N_d(x)=2$	$N_d(x)=1$
0	0.3992	0.2858	0.1677
1	0.3748	0.2588	0.1394
2	0.3563	0.2395	0.1211
3	0.3417	0.225	0.1084
4	0.3299	0.2125	0.0991

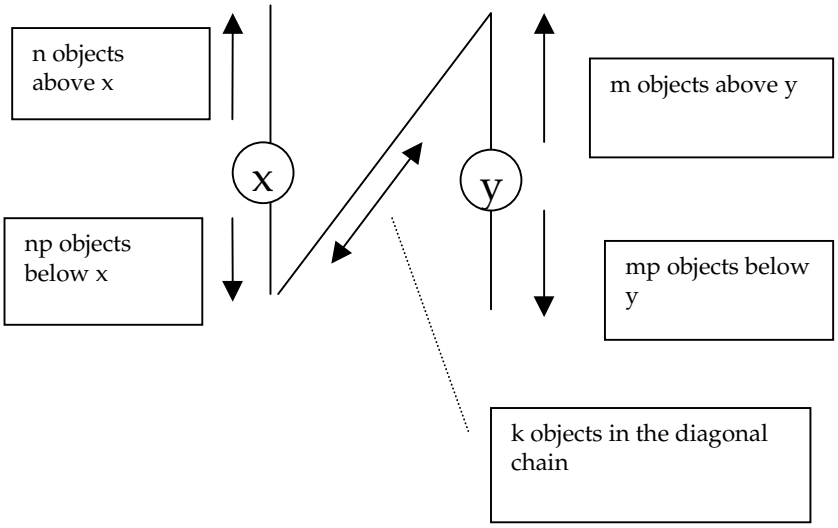


Figure 31: Model system studied (CCC-N-system).np = $N_d(x)$, mp= $N_d(y)$ n = 4 , m = mp = 4, np = $N_d(x)$ is varied. Thus in the maximum, the system contains 21 objects.

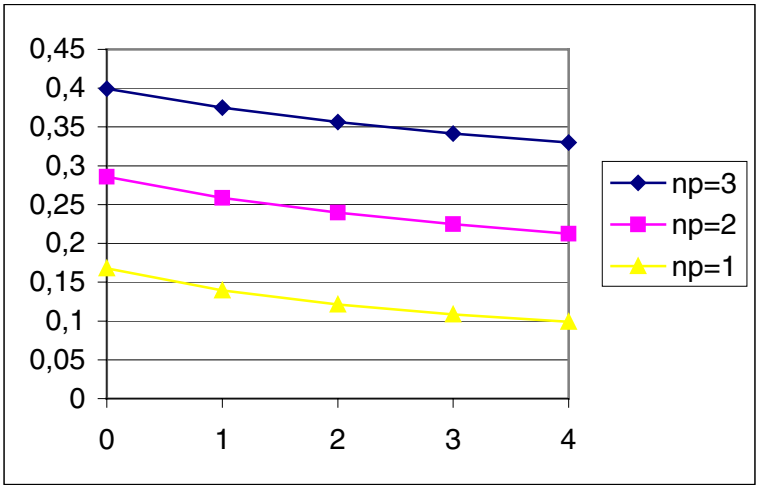


Figure 32: Dependence of the mutual ranking probability $p_m(k)(x>y)$ on np ($N_d(x)$) and on k, which is the abscissa.

Here obviously the effect of increasing for example the chain, containing x does not have a remarkable influence on the dependence of $p_m(x>y)$ on k. The empirical equation found for the CCC-G-system does not work well, albeit it may give a first impression on the dependencies.

Empirically, i.e. by a statistical regression analysis another relation

$$p_m(k) = p_m(0)*f \qquad \text{Eq. 47}$$

was tested. Here by a trial and error procedure f was given the form:

$$f(n, np, m, mp) = [(np/n)*(m/mp)]^k \qquad \text{Eq. 48}$$

f estimated = 0.705 + 0.318 * f(n, np,m,mp) , R^2_{DF} = 0.859, F = 87, $N_{Records}$: 15 , n=4 , m = mp = 4 Eq. 49

7 Comparison of $prob_Q$ with probabilities of model systems

The formulas found for the model systems are not only useful to calculate explicitly mutual probabilities but also to test the $prob_Q$-approximation.

7.1 $prob_Q$ vs $prob_h$

For the h-system as shown in Figure 20 the mutual probability is given as:

$$prob_h(x > y) = \frac{2}{n+4}$$

It is interesting that prob Q(xoy) lead to another equation thus showing that indeed a slight systematic error will be behind the $prob_Q$-estimation:

$$prob_{Q;h}(x > y) = \frac{1}{n+2}$$

If, however, the common objects of x and y are included, then the formula

$$prob_{Q+,h}(x > y) = \frac{2}{n+4}$$

is obtained. This formula gives the correct expression.

7.2 $prob_Q$ vs $prob_M$

For the M-system as shown in Figure 21 the following relationship is seen between $prob_Q$ and the true mutual probability $Prob_M$.

If $prob_Q$ is calculated on the basis of (xoy) and (yox) respectively one gets:

$prob_{QjM}$ (x>y) = (n+1)/(n+m+2) Eq. 50

whereas: $prob_M$(x>y) = (n+2)/(n+m+2).

Obviously the approximation by Q(xoy) is wrong.

If, however, the Q-values are calculated taking into account the common objects, then after:

$N_u(x) = 0$, $N_d(x) = n+1$, $N_u(y) = 0$, $N_d(y) = m+1$

Q(x) = 1/(n+2). Q(y) = 1/(m+2)

and

$prob_{Q+.M}(x>y) = (n+2)/(n+m+4)$

it therefore seems once again as if the inclusion of the common elements leads to the correct result within the $prob_Q$-formalism.

7.3 $prob_Q$ vs $prob_{CC}$

In this section the double chain system shown in section 5.3 will be used for evaluation of the $prob_Q$ estimates by comparison with the true probability $Prob_M$.

In order to test $prob_Q$ vs $prob_{CC}$ a double chain system with the total length of each chain = 5 was selected and the variation of $prob_{CC}$ and $prob_Q$ as function of the parts above x and y are collected:

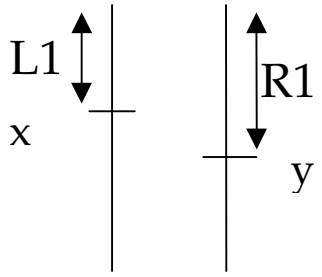


Figure 33: Test of $prob_Q$ vs $prob_{CC}$

Table 9 summarizes the result:

Table 9: Comparison of $prob_{CC}$ and $prob_Q$-values in double chain systems. In parentheses: the $prob_Q$-values.

L1\R1	0	1	2	3	4
0	0.5 (0.5)	0.778 (0.714)	0.917 (0.833)	0.976 (0.909)	0.996 (0.962)
1	0.222 (0.286)	0.5 (0.5)	0.738 (0.667)	0.897 (0.8)	0.976 (0.909)
2	0.083 (0.167)	0.262 (0.333)	0.5 (0.5)	0.738 (0.667)	0.917 (0.833)
3	0.024 (0.091)	0.103 (0.2)	0.262 (0.333)	0.5 (0.5)	0.778 (0.714)
4	0.004 (0.038)	0.024 (0.091)	0.083 (0.167)	0.222 (0.286)	0.5 (0.5)

Schematically the degree of approximation may be shown as:

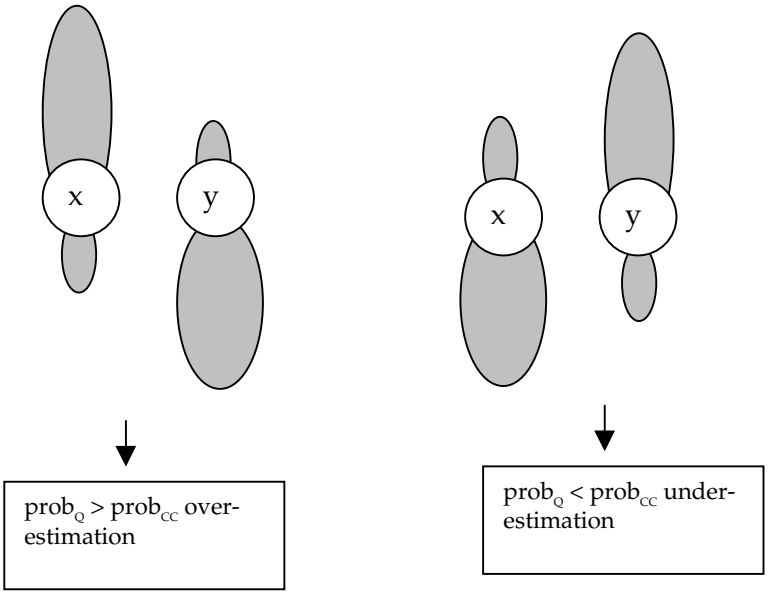


Figure 34: The cases of over (left side) and under (right side) estimation. This might be an explanation why in the average $prob_Q$ is working quite well.

It can be hypothesized that in the case of pure down- or pure up-systems (i.e. x,y are either maximal or minimal elements) the $prob_Q$-formalism works well.

Indeed an explicit calculation shows with help of the combinatorial relation (Stanley, 1986, p. 44):

$$\sum_{i=0}^{n}\binom{x+i}{i}=\binom{x+n+1}{n} \qquad \text{Eq. 51}$$

that surprising simple formulas for the pure up, or pure down systems can be found and that these formulas are exactly the same as those derived for the $prob_Q$-formalism. More relations like that of eq. 51 can also be found in Matousek & Nesetril, 2002.

Therefore, $prob_Q$ leads to exact values in double chain systems, if the system is a purely up or a purely down-system.

Furthermore: If there are common objects then $prob_{Q+}$ seems to be a better approximation than $prob_Q$.

We summarize the findings by:

$prob_{Q+}$ can be used as estimator of mutual probabilities especially in those cases where the objects are maximal or minimal objects.

7.4 probCC-variants vs different probQ-variants

After performing all the tests shown above a check of $prob_{CC}$ and (taking the results of the above sections into account) $prob_{CC+}$ was done.

The sample contained 41 different partial ordered sets. This set included the original 27 ones used to derive $prob_Q$. However, additional partial orders of the specific systems h-, M-, A-, AW-systems, double chains and double sandwiches are included.

The Table 10 summarizes the results number of testing records: Nrec = 41:

$p_m$estim = a + b*prob-model

Table 10: Different models for mutual ranking probability (probQ+, prob CC+ refer to those approximations, where the common elements are included.)

model	R^2_{DF}	F	a	b
probQ	0.913	421	-0.016	1.021
probQ+	0.96	948	-0.108	1.19
probCC+	0.97	1251.5	0.007	0.964

Therefore the $prob_{CC+}$-model might be satisfactory. However, the formula is somewhat tedious and we do not learn directly the influence of structure from it.

8 Outlook

One can still do much more work: For example, formulas for more or less regular systems may be developed. By this the estimation via numerical generation of linear extensions and calculation of probability can be bypassed. If formulas are not too complicated, then even a structural insight will be possible to gain.

With all experiences shown here and made by many other calculations with $prob_Q$ it seems promising to rely purely on this concept. Thus even the need to evaluate complicated combinatorial formulas will not be necessary. Then the only information, which is needed, is reduced to what is above and what is below the two objects of interest. The underlying partial ordered set is then just a background information, but it is still needed because comparabilities leading to $prob_Q = 1$ or $prob_Q=0$ will never be found in that kind of analysis. Thus one has to extend the formula as follows:

$$prob_Q(x,y)=\begin{cases} prob_Q(x>y) \; if \; x \| y \\ 1 \qquad if \; x=y \; or \; x>y \\ 0 \qquad if \; x<y \end{cases} \qquad \text{Eq. 52}$$

This fact again makes it evident that we are speaking in different order theoretical sets: The decision whether two objects are incomparable, equivalent (or equal) or comparable has to be done within the partial ordered set, deduced from those attributes, which are worth to be considered (which already needs some ideas about the structure of

the model). If a mutual ranking probability is calculated, then one refers to a setting within the GRM.

Therefore one may ask whether some estimations on GRM could be made directly. Indeed this seems to be the case, following the paper of Mallows 1957 and the idea therein, reported as Bradley-Terry – model. The steps are:

- consider the $prob_Q$ as starting point,
- take into regard that the structure (x/(y+x)) assumed by Bradley-Terry guarantees that transitivity is fullfilled
- identify with each $prob_Q$ (x>y) > 0.5 a line from y to x to construct a directed graph and
- check whether this graph is acyclic.

If the graph is acyclic, then it represents a linear order. From this linear order a ranking, namely that of GRM can be derived. Finally a probability for this ranking can be given.

Furthermore it seems as if the p_m-values can be directly used to estimate the averaged rank of GRM: Let p be the matrix of mutual probabilities with 1 in the diagonal and let e be a vector having a 1 as its component. Furthermore let prk be the rank probabilities and rk the vector of ranks (1.2.....N) (N: the number of objects)

Then numerical observation shows that

$$p \cdot e = prk \cdot rk = R \qquad \text{Eq. 53}$$

with R the vector of averaged ranks of GRM.

Therefore estimations of p_m by means of $prob_Q$ may be a useful step to obtain an approximated GRM.

A simple example might be helpful:

Let us consider four objects. We would like to know the fitness of these objects. In order to learn something about the fitness some attributes, which may describe these objects with regard to their expected benefit are gathered.

As there is no deterministic model at hand it remains to analyze all possible positive monotonous functions just to see what may be the order among the four objects. To do this in reality would be wasting time and effort. Because a partially ordered set, equipped with the product order (because the attributes are considered as relevant) would do the same:

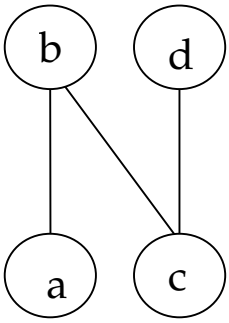


Figure 33: Hasse diagram of the four objects.

Now it is immediately clear that any positive monotonous function would keep the order among a < b , c < b, c < d.

What happens is that each of these functions will differently affect the positions of b relative to d, of object a relative to d, and of object a relative to c. What would be probable: If no other information is at hand, the fact, that a is below b and c below two objects would indicate that object a might more often have a higher position than c.

In the concept of GRM this means: A total order exists; and with respect to this total order we could clarify

a) which rank an object might have and

b) what is the probability that a is less than c.

Now calculating the mutual probabilities will lead to the following matrix:

$$p_m = \begin{pmatrix} 1 & 0 & x & y \\ 1 & 1 & 1 & z \\ (1-x) & 0 & 1 & 0 \\ (1-y) & (1-z) & 1 & 1 \end{pmatrix}$$

The 1 and 0 can be immediately put into the matrix, because of the known comparabilities of the above mentioned partial ordered set.

To get x,y,z one may apply the full formalism. This means that one has to go through all linear extensions, counting the number of linear extensions, where for example a > c and after that calculate the needed fraction. This means that one must:

- firstly assume the existence of GRM (because otherwise mutual probabilities make no sense)
- secondly realize the GRM by finding the linear extensions (or a statistical sound fraction of these) and
- thirdly derive a total order and
- find out what the mutual probabilities are for an appropriate arrangement of these objects.

Instead we can first calculate the mutual probabilities by a model function (for example probQ+)

$$p_m = \begin{pmatrix} 1 & 0 & 0.6 & 0.2 \\ 1 & 1 & 1 & 0.4 \\ 0.4 & 0 & 1 & 0 \\ 0.8 & 0.6 & 1 & 1 \end{pmatrix}$$

The probability of this ranking might be estimated, see Mallows. (1957).

To illustrate the methodology a data set of 12 POP candidates to the UNECE CLRTAP POP Protocol selected in a former study is applied (Lerche et al. 2002). The chemicals and the descriptors are listed in Table 11.

Table 11: Chemicals and their data

Id	Name	Cas no.	log K_{ow}	Bio-degradation*	Vapour pressure (Pa)	Atmospheric half-life (days)	Toxicity category: Human	Toxicity category: Ecotox
Df	Dicofol	115-32-2	5.0	2	$1.6 * 10^{-6}$	3.1	2	3
cp	1.3-Cyclopentadiene. 1.2.3.4.5.5-hexachloro-	77-47	5.0	1	2.82	27	1	3
pcp	phenol. pentachloro-	87-86-5	5.1	1	$7.0 * 10^{-3}$	19	1	2
bz5	Benzene. pentachloro-	608-93-5	5.2	1	0.67	190	3	3
py4	Pyridine. 2.3.4.5-tetrachloro-6-(trichloromethyl)-	1134-04-9	5.3	2	0.011	3700	3	**
na6	Naphthalene. hexachloro-	1335-87-1	7.0	1	$4.4 * 10^{-4}$	57	**	3
pm	Phenol. 4.4'-1-methylethylidenebis 2.6-dibromo-	79-94-7	7.2	1	$2.3 * 10^{-9}$	3.6	**	3
p3	Phenol. 2.2'-methylenebis 3.4.6-trichloro-	70-30-4	7.5	1	$1.4 * 10^{-8}$	4.9	2	3
ib	Isobenzan	297-78-9	5.2	2	$1.0 * 10^{-3}$	2.3	1	2
ap	Ammonium perfluorooctanoate	3825-26-1	6.3	2	990	21	2	**
dn	Decanoic acid. nonadecafluoro-	335-76-2	8.2	2	770	21	2	**
nh	Nonanoic acid. 2.2.3.3.4.4.5.5.6.6.7.7.8.8. 9.9-hexadecafluoro	76-21-1	6.7	1	690	21	3	**

*2 = More than month and 1 = months
** unknown

By rounding and supplying missing data conservatively a Hasse diagram as follows results (Figure 34):

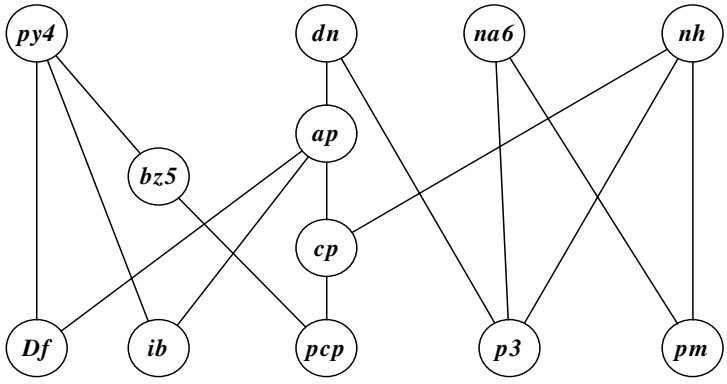


Figure 34: Example of a real Hasse diagram

In GRM it gives sense to ask for example for p_m(py4>dn) or for p_m(bz5>ap).

The total number of linear extensions is LT = 266464.

The number of linear extensions, where py4 > dn is 101728, and bz5>ap: 59232. Therefore the mutual probabilities are: p_m(py4>dn) = 0.382 and p_m(bz5>ap)=0.222

In Table 12 the calculation results are summarized:

Table 12: n, np, n+,np+ refer to the object py4 and bz5 respectively, whereas m,mp, m+,mp+ refer to the objects dn and ap respectively The +-sign indicates that common elements are counted.

	py4	dn	bz5	ap
n	0	-	1	-
n+	0	-	1	-
np	1	-	0	-
np+	3	-	1	-
m	-	0	-	1
m+	-	0	-	1
mp	-	3	-	3
mp+	-	6	-	4
$prob_{CC}$	0.333		0.143	
$prob_{Q}$	0.333		0.25	
$prob_{Q+}$	0.417		0.286	

There is a considerable part, which is not accounted for by the "above-below"-formalism, therefore a considerable uncertainty for these results should be taken into account.

References

Atkinson, M.D. & Chang, H.W. 1986. Extensions of Partial Orders of bounded width. Congressus Numerantium 52, 21-35.

Atkinson, M.D. 1989. The complexity of Orders. In: *Algorithms and Order NATO ASI series, Series C: Mathematical and Physical Sciences, Vol. 255 ISBN 0-7923-0007-6* (ed I. Rival) pp. 195-230. Kluwer Academic Publishers, Dordrecht.

Atkinson, M.D. 1990. On the computing the Number of Linear Extensions of a Tree. Order 7, 23-25.

Brightwell, G.R., Prömel, H.J., & Steger, A. 1996. The average number of linear extensions of a partial order. J.Combin.Theory (A) 73, 193-206.

Brüggemann, R. & Halfon, E. 1995. Theoretical Base of the Program "Hasse". GSF-Bericht 20/95, München-Neuherberg.

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.

Brüggemann, R. & Bartel, H.-G. 1999. A Theoretical Concept to Rank Environmentally Significant Chemicals. J.Chem.Inf.Comp.Sc. 39, 211-217.

Brüggemann R., Halfon, E., Luther, B., & Simon, U. 2000. New Tools in Hasse diagram technique - Example: Comparative Evaluation of Near - Shore - Sediments by a battery of tests. In: *Order Theoretical Tools in Environmental Sciences - Proceedings of the Second Workshop , October 21st, 1999 in Roskilde, Denmark* (eds P. B. Soerensen, L. Carlsen, B. B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U. Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, and F. Rediske) pp. 73-94. National Environmental Research Institute, Roskilde.

Brüggemann, R. & Halfon, E. 2000. Introduction to the General Principles of Partial Order Ranking Theory. In: *Order Theoretical Tools in Environmental Sciences - Proceedings of the Second Workshop , October 21st, 1999 in Roskilde, Denmark* (eds P. B. Soerensen, L. Carlsen, B. B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U. Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, and F. Rediske) pp. 7-43. National Environmental Research Institute, Roskilde.

Edelman, P., Hibi, T., & Stanley, R.P. 1989. A Recurrence for Linear Extensions. Order 6, 15-18.

Lerche, D.B., Van de Plassche, E., Schwegler, A., Balk, F. 2003. Selecting chemical substances for the UN-ECE POP Protocol. Chemosphere 47, 617-630.

Lerche, D., Sørensen, P. & Brüggemann, R., 2003. Improved Estimation of Ranking Probabilities in Partial Orders using Random Linear Extensions by Approximation of the Mutual Ranking Probability, J. Chem. Inf. Comput. Sci, Vol. 43, pp. 1471-1480.

Mallows, C.L. 1957. Non-Null Ranking Models I. Biometrika 44, 114-129.

Matousek, J. & Nesetril, J. 2002. Diskrete Mathematik - Eine Entdeckungsreise. Springer, Berlin.

Sørensen, P. B., D. Lerche, L. Carlsen and R. Brüggemann 2001. Statistically approach for estimating the total set of linear orders A possible way for analysing larger partial order sets. In Pudenz, S., Brüggemann, R. and Lühr, H.-P. (2001). Order Theoretical Tools in Environmental Sciences and Decision Systems. Proceeding of the Third Workshop November 6th-7th, 2000 in Berlin, Germany, Leibniz Institute of Fresh Water Ecology and Inland Fisheries, Berlin, Germany. –pp. 222- heft 14, Sonderheft IV.

Sørensen, P.B., Lerche, D.B. 2002. Quantification of the uncertainty related to the use of a limited number of random linear extensions. In K.Voigt, G.Welzl: Oder Theoretical Tools in Environmental Sciences, 2002. p. 65-72.

Stanley, R.P. 1986. Enumerative Combinatorics Volume I. Wadsworth&Brooks/cole, Monterey.

Trotter, W.T. 1992. Combinatorics and Partially Ordered Sets Dimension Theory. The Johns Hopkins University Press, Baltimore,Maryland.

Voigt, K. & Welzl, G. (Eds.) 2002. Order Theoretical Tools in Environmental Sciences. Order Theory (Hasse diagram technique) Meets Multivariate Statistics. Shaker-Verlag. Aachen.. pp. 206.

Data availability on existing substances in publicly available databases - a data-analysis approach

Kristina Voigt, Gerhard Welzl[1]

GSF - National Research Center for Environment and Health,
Institute of Biomathematics and Biometry,
Department of Biostatistics,
85764 Neuherberg, Germany,
[1]GSF - Institute of Development Genetics,
85764 Neuherberg, Germany
kvoigt@gsf.de

Abstract

The gap in knowledge about intrinsic properties for 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 made use of in order to waive testing, wherever appropriate. In this paper the emphasis lies on the evaluation of numerical databases, which focus on environmental fate and ecotoxicity data. In total 12 important international data-sources and 15 environmental fate and ecotoxicity parameters are chosen for examination.

Different data-analysis methods are chosen: the Correspondence Analysis, the Hasse Diagram Technique and the Partially Ordered Scalogram Analysis with Coordinates method. These three methods are applied on the 12x15 data matrix (12 databases, 15 environmental parameters). The aim of the Correspondence Analysis is to order the data matrix and to visualise the new ordering to detect the data-gaps and data fillings immediately. SID (Screening Information Data Sets from the OECD), NRA (NRA Chemical Review Program), and EXT (EXTOXNET) hold a lot of parameters. Performing the Hasse diagram analysis it is found that the ECO (ECOTOX), EFD (Environmental Fate Database), EXT (EXTOXNET) and NRA (NRA Chemical Review Program) are maximal objects. These are profile databases (NRA, SID) and multi-database databases (ECO, EFD, EXT). The POSAC method reduces the data matrix in plotting it in a two-dimensional space. The proportion of the order relations, which are correctly represented, is 95,6 %. This means that the Hasse diagram can be approximated by using the two latent order variables (LOVs) instead of the 15 initial variables.

As it is also of great interest to know the ranking of the parameters, the data matrix 12x15 is transformed to a 15x12 data matrix. The objects are now the ecotoxicological parameters and the attributes are the databases. The Hasse diagram of this data matrix shows two iso-

lated non-trivial hierachies, one for the ecotoxicity parameters and one for the environmental fate parameters.

The whole data-analysis approach shows that considerable data-gaps in ecotoxicological parameters exist in publicly available Internet databases.

1 Chemicals of Environmental Concern

The global production of chemicals increased from 1 million in 1930 to 400 million tones today. The number of existing substances reported in 1981 and published in EINECS (European Inventory of Existing Chemical Substances) was 100,106, the current number of existing substances marketed in volumes above 1 ton is estimated at 30.000 (Commission 2001). Existing substances, i.e substances, which have not been examined according to the present demand, 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.

The gap in knowledge about intrinsic properties for 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 made use of in order to waive testing, wherever appropriate.

Significant gaps in publicly available knowledge of existing chemicals were revealed (Allanou, 1999). The contents of the IUCLID (International Uniform Chemical Information Database) were evaluated in a recent study. The study of Allanou showed considerable data-gaps in environmental fate and fate pathways as well as in ecotoxicity parameters. The same finding might be true for other databases. Therefore, in this paper the emphasis lies on the evaluation of numerical databases, which focus on environmental fate and ecotoxicity.

2 Selection of Databases with Environmental Parameter Focus

The following list of databases (Table 1) was selected from the 746 databases found in the DAIN - Metadatabase of Internet Resources for Environmental Chemicals [http://www.wiz.uni-kassel.de/dain/]. This metadatabase encompasses data-sources on the free Internet with information on environmental data with respect

to chemical substances (Voigt 2002a). Different types of data-sources are incorporated in DAIN, e.g. bibliographic databases, full-text databases, research project databases, chemical dictionaries, structural databases, reaction databases and numerical databases. In the selection only numerical databases are considered. We tried to incorporate in the set of objects not only U.S. resources but also European and Japanese databases. The list of databases with their abbreviations and URLs is given in Table 1.

Table 1: List of selected numerical databases focusing on environmental chemicals

Name	Abb.	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

3 Environmental Parameters

The content of the numerical databases was evaluated according to the data-fields implemented in the IUCLID database. However, only the environmental fate and pathways and the ecotoxicity parameters will be looked upon (see Table 2). 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, BOD_5, COD or BOD_5/COD ratio, bioaccumulation.

Ecotoxicity: acute/prolonged toxicity to fish, acute toxicity to aquatic invertebrates, toxicity to aquatic plants, e.g. algae; toxicity to micro organisms, 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, biotransformation and kinetics.

Table 2: List of selected environmental parameters

Parameter	Abbreviation
photodegradation	PHO
stability in water	SWA
stability in soil	SSO
biodegradation	BDE
BOD_5, COD or BOD_5/COD ratio	BOD
bioaccumulation	BAC
acute/prolonged toxicity to fish	ATF
acute toxicity to aquatic invertebrates	ATD
toxicity to aquatic plants e.g. algae	ATP
toxicity to micro organisms, e.g. bacteria	ATB
chronic toxicity to fish	CTF
chronic toxicity to aquatic invertebrates	CTD
toxicity to soil dwelling organisms	TSO
toxicity to terrestrial plants	TTP
toxicity to other non-mammalian terrestrial species	TNT

4 Applied Data-Analysis Methods

Different data-analysis methods will be applied, the Hasse diagram technique (HDT), a method derived from discrete mathematics and the Partially Ordered Scalogram Analysis with Coordinates (POSAC) method, a multivariate statistics' approach. Furthermore the correspondence analysis (CA) is applied. These methods are compared and advantages and disadvantages will be elaborated.

The aim of the whole data-analysis procedure is the evaluation of the publicly available numerical databases comprising ecotoxicological information on chemical substances. This will show data-gaps as well as data richness for some parameters or for some chemicals. Furthermore our evaluation approach will show which database is more useful with regard to finding information on ecotoxicological parameters. The approach might also provide some recommendations for the improvement of numerical databases. A brief characterization of the methods used here follows:

CA
The correspondence analysis is a method, which rearranges the order of rows and that of columns to detect homogenous fields in the array of data. Thus, relationships between categorical variables in a so-called matrix-plot are displayed. In the resulting diagram Euclidian distances approximate chi-squared distances between rows and column categories. The coordinates are analogous to those resulting from a PCA (Principle Component Analysis) of continuous variables, except that they involve a partition of a chi-squared statistic rather than the total variance. Such an analysis of a contingency table allows a visual examination of any structure of pattern in the data (Everitt 1998).

HDT
The componentwise order of tuples is graphically displayed. Details, see Brüggemann (2000, 2002).

POSAC
There are many statistical approaches of condensing a data matrix by creation of new variables. This process – called ordination – is often used to visualize relationships in two dimensions based on the first two variables. These new variables, which are derived from the original variables, are constructed to optimize some specific criteria. For example, principal component analysis creates new axes to explain as much as possible of the variance of the data matrix. This idea can be applied when order relations (comparability as well as incomparability) are considered as the essential aspect of the data to be preserved in the analysis. This method – construction of new axes, which presents correctly as many as possible of the order relations – is called Partially Ordered Scalogram Analysis with Coordinates (POSAC). POSAC is integrated in the program package SYSTAT 10 (SPSS Science 2001) under the feature of statistics, data reduction. In POSAC, order relations (comparability as well as incomparability of the structuples) are considered as the essential empirical-substantive aspect of the data to be preserved in the data-analysis (Borg 1995). For a better interpretation of the new axes correlations between old and new variables can be calculated (Borg 1995). The background of the POSAC method as well as the mathematics in it, is described in a textbook, entitled "Multiple Scaling" by Shye (1985). Note that contrary to the concept of order preserving maps in HDT the map between the order relations based on the original data and that of POSAC is not necessarily order preserving.

The POSAC method has already been applied on data-matrices in environmental sciences and chemistry: Welzl examined regions polluted with metals (Welzl 1998). Pesticide Internet resources were analyzed with chemical and environmental evaluation criteria by Voigt et al. (2000), environmental and chemical search engines were ranked by Glander-Höbel (2001) and Voigt & Welzl (2001), drinking water analysis systems were evaluated by Voigt & Welzl (2002b) and preferred habitats of fish communities were detected by Brüggemann et al., this issue.

5 Search Results for Environmental Parameters and Chemicals

The following Table 3 shows the search results for the environmental parameters. The number 0 indicates that no data are available whereas the number 1 means that data are available at least for one chemical.

Table 3: 15 Environmental Parameters Found in 12 Ecotoxicological Databases (15x12 Data-Matrix, D, or 12x15 Data-Matrix D (transposed) = D^t)

DB / PA	PHO	SWA	SSO	BDE	BOD	BAC	ATF	ATD	ATP	ATB	CTF	CTD	TSO	TTP	TNT
BID	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CIV	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
ECO	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
ENV	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
EFD	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
EHC	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
EXT	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0
HSD	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0
NRA	1	0	1	1	0	1	1	1	1	1	0	0	1	1	0
PES	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
SID	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1
UMW	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0

1= parameter available, 0 = parameter not available

6 Applying Different Data-Analysis Methods on the Data-Matrix

As we have two different data-matrices, namely D and D^t and several possibilities to order the objects and attributes the following Figure 1 outlines the procedure by which the data-analyses are performed.

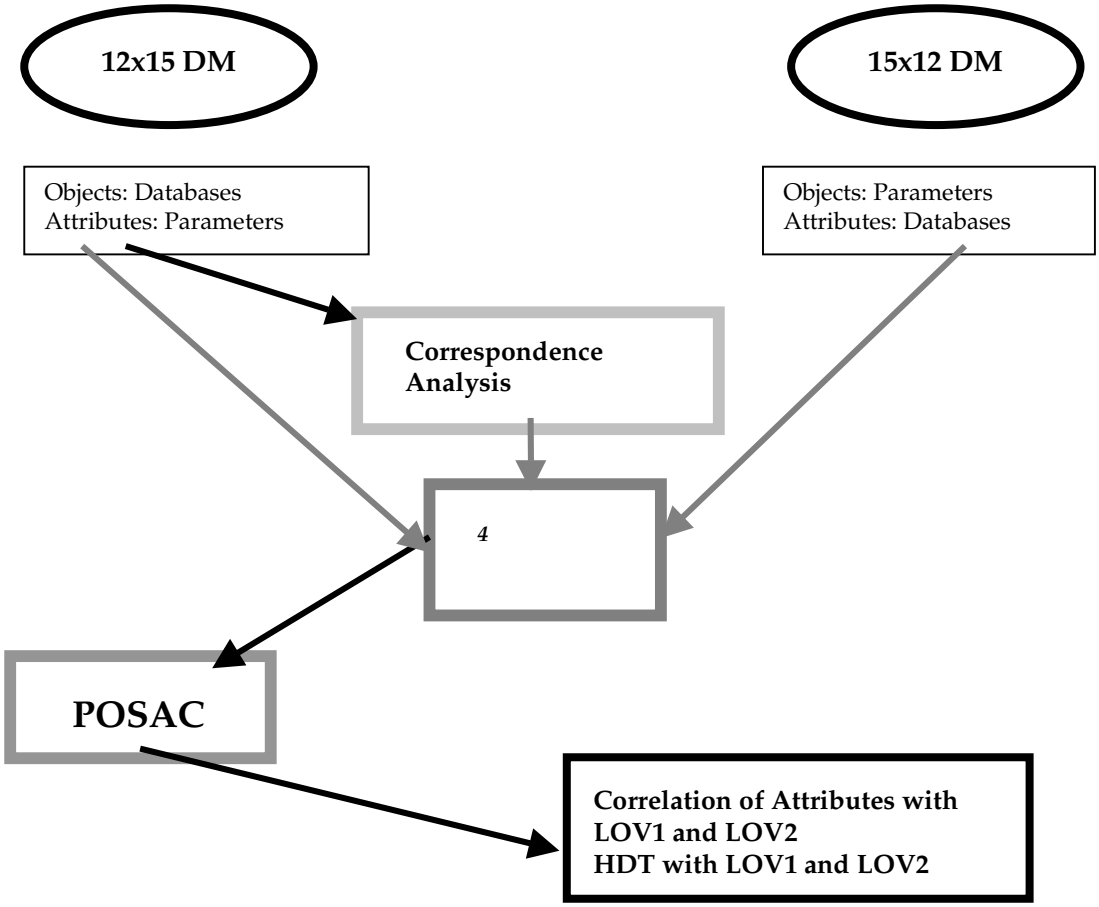


Figure 1: Schematic Outline of the Performed Data-Analyses

6.1 Visualisation by rearrangement of data matrix

Under the generic term Bertin strategies [Sawitzki 1996] there are many ways to rearrange the data matrix in order to detect patterns or structures in the data. For this purpose we apply the Correspondence Analysis on the data-matrix 12x15 (12 databases, 15 ecotoxicological parameters). The aim of this data-analysis method is to order the data-matrix and to visualise the new ordering detecting the data-gaps and data fillings immediately. The result is shown in Figure 2.

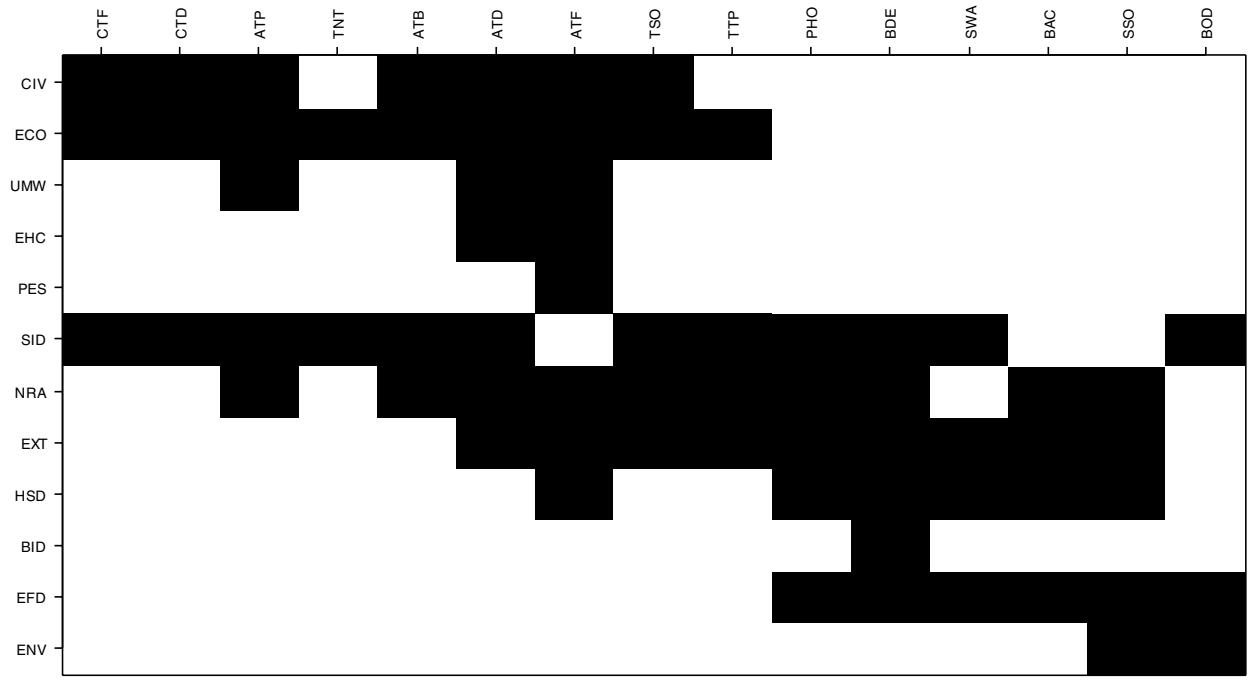


Figure 2: Shaded Plot Parameters (12 Databases x 15 Parameters) (light area data-gaps, dark area data fillings)

The light areas in Figure 2 indicate the data-gaps whereas the dark areas are the data fillings. Three large data-gap areas can easily be detected: One in the upper right side, one in the lower left side and one slightly smaller in the upper left side. The data-gap in the upper right side shows all the databases, which do not encompass data on environmental fate and pathways, whereas the light area in the left lower area demonstrates those databases, which do not cover data concerning ecotoxicity parameters. The third area (upper right) shows those databases where data-gaps in chronic ecotoxicity data are encountered. Taking a look at the databases, which show a lot of parameters, the SID (Screening Information Data Sets from the OECD), NRA (NRA Chemical Review Program), and EXT (EXTOX-NET) must be mentioned. SID and NRA are so-called profile databases, which means that they offer comprehensive data-profiles for each chemical. The disadvantage of these databases is that they only cover very few chemicals. EXT is a multi-database database. Several databases are searched simultaneously.

Having now some insight about the data-gaps and data fillings it is of importance to know which database (or parameter) is better or worse than others.

6.2 Application of Hasse Diagram Technique on Two Data-Matrices

The Hasse diagram technique will now be applied on the following two data-matrices:

12 x 15 (12 databases = objects; 15 parameters = attributes)

15 x 12 (15 parameters = objects; 12 databases = attributes)

6.2.1 Hasse Diagram Technique Applied on Databases (Parameters) (12x15 Data-matrix)

In order to get an idea of the ranking of the databases with respect to the environmental parameters we generate a Hasse diagram from the 12 databases x 15 parameters data-matrix.

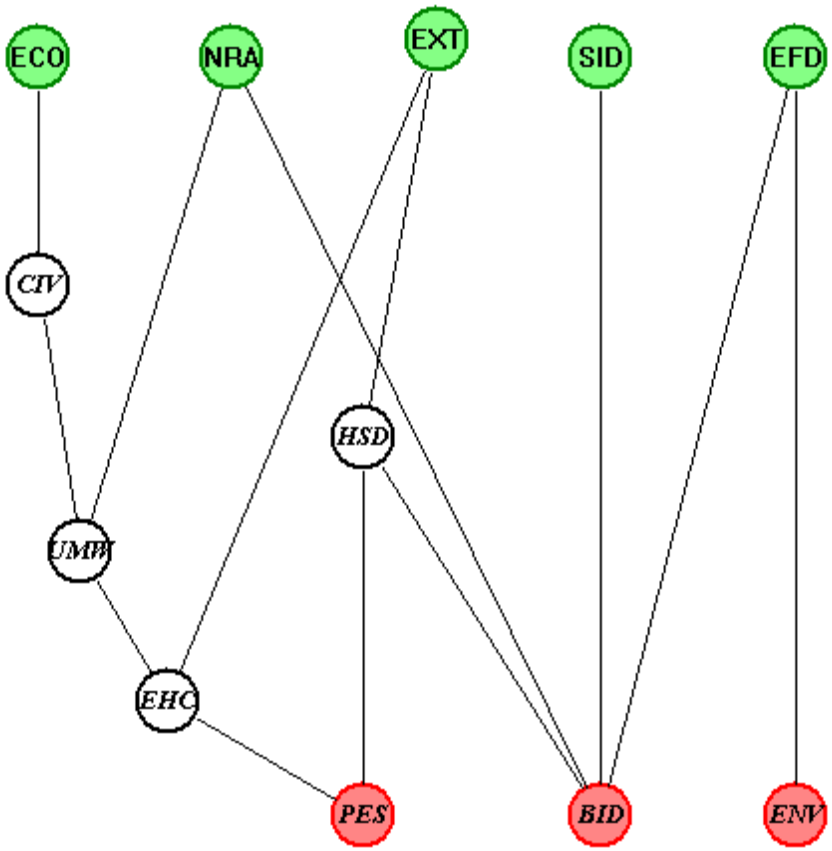


Figure 3: Hasse Diagram for 12 Databases (Objects) x 15 Parameters (Attributes)

We can state that we have 5 maximal objects: {ECO}, {EFD}, {EXT}, {NRA}, and {SID}. Where ECO, NRA and EXT show several succes-

sors, the objects SID and EFD only have respectively one and two successors. The number of minimal objects is 3: {BID}, {ENV}, {PES}. No equivalent objects are found in this data set. The number of comparabilities is 23 and the number of incomparabilities 43. The number of levels is 5.

The relation to Figure 2 is easily established: The dark areas, taken in their horizontal direction of a database x must contain that of y, if in the poset (visualized in Figure 3) x≥y. As an example: Compare the chain PES ≤ HSD ≤ EXT: database PES comprises only data on ATF (acute toxicity to fish), database HSD: ATF and environmental fate and pathway parameters and finally database EXT, those of HSD however, additionally ATD (acute toxicity to aquatic invertebrates). Figure 3 delivers some kind of vertical view along the matrix-plot.

6.2.2 Hasse Diagram Technique Applied on Parameters (15x12 Data-matrix)

As it is also of great interest to know the ranking of the parameters, the data-matrix 12x15 is transformed to a 15x12 data-matrix. The objects are now the ecotoxicological parameters and the attributes are the databases. Figure 4 shows the Hasse diagram of this approach.

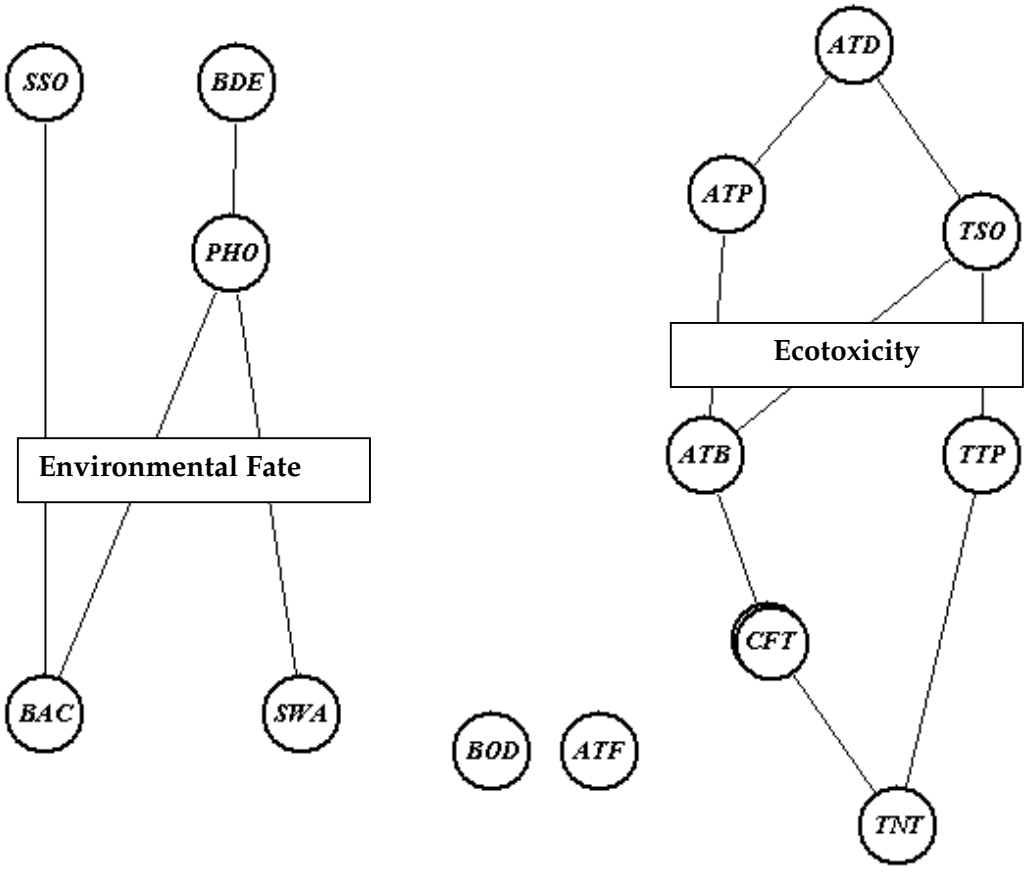


Figure 4: Hasse Diagram of 15 Parameters (Objects) x 12 Databases (Attributes) Equivalence classes: CFT, CTD

This Hasse diagram shows two isolated nontrivial hierarchies. The hierarchy on the left hand side comprises the environmental fate pa-

rameters: SSO (stability in soil), BDE (biodegradation), PHO (photo-degradation), BAC (bioaccumulation), SWA (stability in water). The hierarchy on the right hand side covers the following ecotoxicity parameters: ATD (acute toxicity to aquatic invertebrates), ATP (toxicity to aquatic plants, e.g. algae), ATB (toxicity to micro organisms, e.g. bacteria), CFT (chronic toxicity to fish), CTD (chronic toxicity to aquatic invertebrates), TNT (toxicity to other non-mammalian terrestrial species) TSO (toxicity to soil dwelling organisms), TTP (toxicity to terrestrial plants). The objects BOD (BOD_5, COD or BOD_5/COD ratio) and ATF (acute toxicity to fish) are isolated objects. The acute toxicity parameters are higher ranked than the chronic ecotoxicity parameters and the terrestrial parameters.

Concerning the maximal objects, the following 3 are given: {SSO}, {BDE}, {ATD}.

Minimal objects are: {SWA}, {BAC}, {TNT}.

The isolated objects {ATF} and {BOD} can be counted as maximal and minimal objects.

As in section 6.2.1 there is a close relation to Figure 2. If in Figure 4 a parameter "P" ≤ "Q", the dark grey areas of Q, taken now in vertical direction must contain those of P. The Hasse diagram shows somewhat like a horizontal gradient through the matrix-plot.

6.3 Application of the POSAC Method on the 12x15 Data-matrix

6.3.1 Dimension Reduction with POSAC

Now we perform the POSAC (Partially Ordered Scalogram Analysis with Coordinates) analysis on the given data-matrix. The Partially Ordered Scalogram Analysis with Coordinates (POSAC) method reduces the data-matrix in plotting it in a two-dimensional space. A given percentage of information is lost by this method. In POSAC, order relations (comparability as well as incomparability of the structuples) are considered as the essential empirical-substantive aspect of the data to be preserved in the data-analysis. The objects with the maximal score tuple 111111111111 as well as with the minimal score tuple 000000000000 are automatically added by the POSAC program

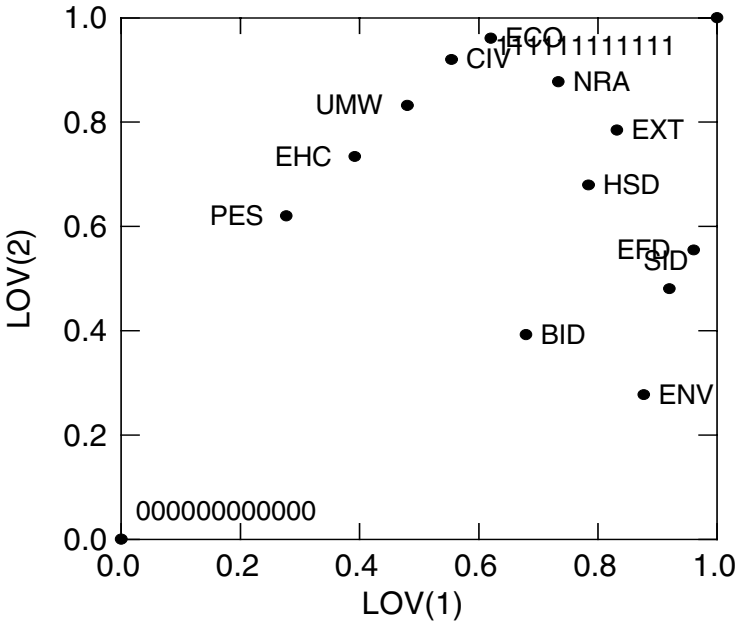


Figure 5: POSAC Plot for 12 Databases (Objects) x 15 Parameters (Attributes)

The proportion of the order relations, which are correctly represented, is 95,6 %. This means that the Hasse diagram can be approximated by using the two latent order variables (LOVs) instead of the 15 initial variables.

Two areas can be visualized: Those which show high values for LOV(1) and those which have high values for LOV(2). To the second category the objects: PES, ECH, UMW, CIV and ECO belong. High values for LOV(1) give NRA, EXT, HSD, EFD, SID, BID, ENV.

Figure 6 compares the diagrams.

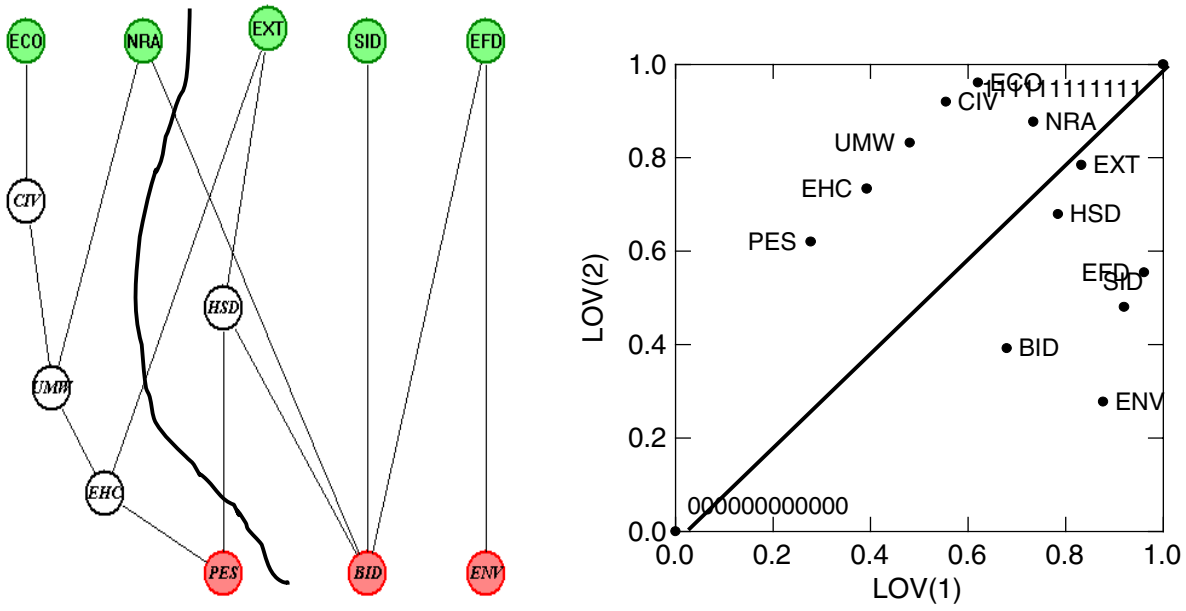


Figure 6: Comparison of Hasse Diagram with POSAC Plot for 15x12 Data-Matrix

6.3.2 Interpretation of the Latent Order Variables

In order to explore the influence of the attributes on the whole analysis, we perform a correlation analysis of the two latent order variables given in the POSAC plot. This is done by applying the analysis of variance (ANOVA).

The following F-statistics are calculated:

LOV (1): PHO: 9,272, BDE: 8,909, BOD: 8,477, SWA: 8,412

LOV (2): ATF: 26,422

The latent order variable 1 is mainly described by PHO (photodegradation), BDE (biodegradation), BOD, whereas the latent order variable 2 is influenced by ATF (acute toxicity to fish). Again LOV(1) is dominated by an ecotoxicity parameter whereas LOV(2) is described by an environmental fate parameter.

The databases in which data exists for respectively PHO and ATF are indicated in the LOV(1) and LOV(2) scatter plot using the symbol x, see Figure 7.

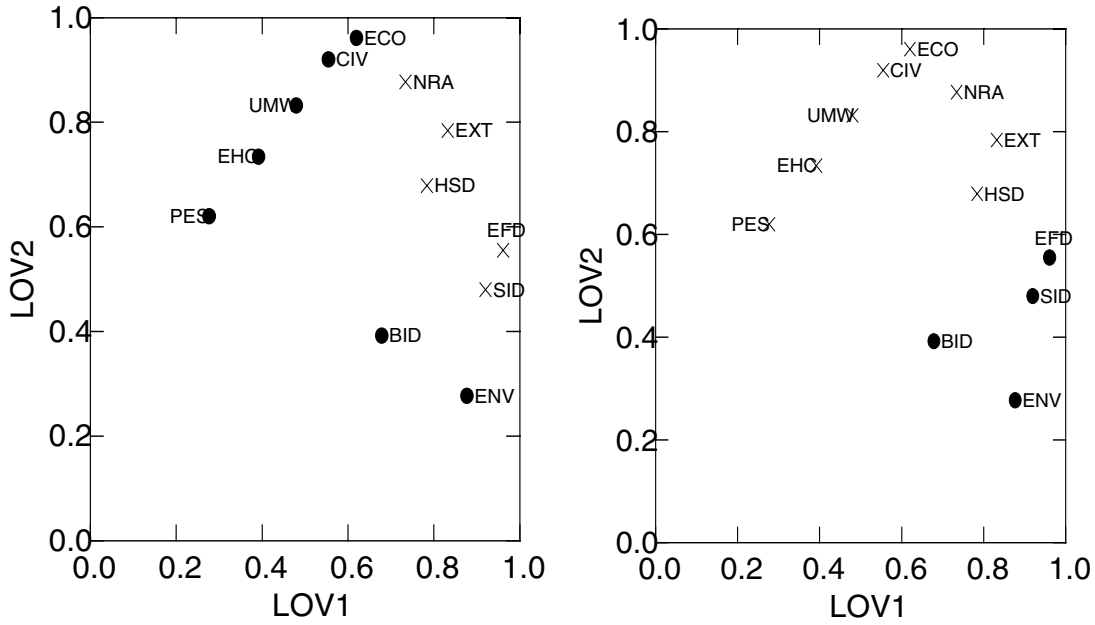


Figure 7: Scatter Plots of LOV(1) and LOV(2), where databases having data for PHO are shown to the left using the symbol X and the same procedure is applied for ATF at the right side plot.

The high values (x) are found on the upper right side of the scatter plot that is to say high values for LOV(1) and for LOV(2).

6.3.3 Combination of Results from POSAC with Hasse Diagram Technique: Hasse Diagram of 12x2 Data-Matrix

In the next step of the evaluation procedure for this data-matrix we calculate a Hasse diagram for the two latent order variables found by the POSAC method. The result of this analysis is given Figure 8. This Hasse diagram represents approximately 96 % of the original dia-

gram given in Figure 2 and 6. This means that most order relations are correctly represented in this diagram.

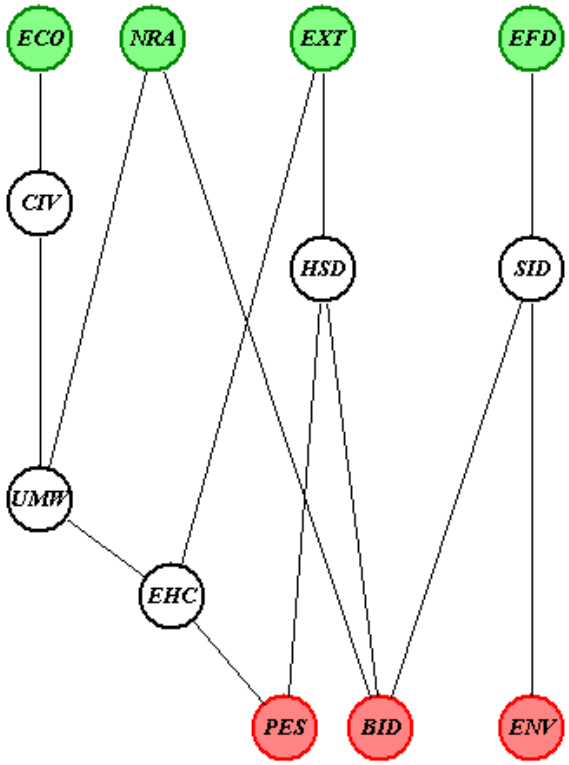


Figure 8: Hasse Diagram of Reduced Data-matrix 12 x 2 (LOVs from POSAC)

The only visible change is the position of the object SID that was a maximal object in the initial Hasse diagram of 12x15 (Figure 2) and is found at the second level in the diagram of the reduced data-matrix of 12x2. All other maximal objects are the same, namely {ECO}, {EFD}, {EXT}, and {NRA}. The minimal objects are the same in both diagrams: {BID}, {ENV}, {PES}. Taking a look at the incomparabilities and comparabilities, the numbers have only slightly changed from 25 to 23 at the comparability side and from 43 to 41 at the incomparability side. The overall impression is that the given data-matrix of 12x15 can easily be described by the reduced data-matrix of 12x2.

9 Discussion and Outlook

- The data-analysis shows considerable data-gaps for ecotoxicity and environmental fate parameters in publicly available Internet databases. Data-gaps are especially found in chronic ecotoxicity data and soil parameters. Taking a deeper look at the databases, several types are tested in the data-set: Profile

databases which give comprehensive data-profiles for a small number of chemicals. The data width is very broad.

- Multi-database databases: several databases are found in a kind of database cluster. Here many chemicals are treated and in some case broad data sets are provided.

- "Normal" databases: only one small, medium or large database is offered.

First we apply the Correspondence Analysis on the data-matrix 12x15 (12 databases, 15 ecotoxicological parameters). The aim of this data-analysis method is to order the data-matrix and to visualise the new ordering in order to detect the data-gaps and data fillings immediately. Taking a look at the databases, which cover a lot of parameters, the SID (Screening Information Data Sets from the OECD), NRA (NRA Chemical Review Program), and EXT (EXTOXNET) must be mentioned.

Performing the Hasse diagram analysis with the 12x15 data-matrix, it was found that the ECO (ECOTOX), EFD (Environmental Fate Database) EXT (EXTOXNET) and NRA (NRA Chemical Review Program) are maximal objects. These are profile databases (NRA, SID) and multi-database databases (ECO, EFD, EXT).

Exchanging the objects and the attributes we evaluated a data-matrix of 15x12, that means 15 parameters as objects and 12 databases as criteria. Here a clear distinction between environmental fate parameters and ecotoxicity parameters was found. Furthermore the acute ecotoxicity parameters were better than the chronic ecotoxicity parameters.

Concerning the data-analysis aspects the following can be stated:

The application of the POSAC method is not giving considerably different results compared with the Hasse diagram technique. By the POSAC method the initial data-matrix 12 x 15 can be reduced to 12 x 2 (latent order variables). The Hasse diagram for this reduced data-matrix shows only marginal changes compared with the original Hasse diagram. On the basis of the two latent order variables a correlation analysis is performed. The variables PHO (photodegradation), BDE (biodegradation) and ATF (acute toxicity to fish) showed an important meaning in the data-analysis.

Although this whole data-analysis approach already shows considerable data-gaps, further studies must be performed not only taking environmental parameters but also environmental chemical properties into account. Further studies in this direction have already been initiated.

However, all studies showed and will most probably underlie the fact that the data material available for existing chemicals in the commonly offered and used data-sources is too limited.

That is the reason why to our mind further actions have to be taken into account:

- Foster 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 [Commission, 2001].

According to the experiences of the authors it is not prudent to build up a new database but put every effort in cooperating with existing database producers in the EEC.

References:

Allanou, R., Hansen, B.G. & van der Bilt, Y. 1999. Public Availability of Data on EU High Production Volume Chemicals, EUR 18996EN, http://ecb.jrc.it/.

Borg, I. & Shye, S. 1995. Facet Theory, Form and Content, p. 111-112, Sage Publications, Thousand Oaks.

Brüggemann, R. & Halfon, E. 2000. Introduction to the General Principles of the Partial Order Ranking Theory, in: Sorensen P.B., Carlsen L., Mogensen B.B., Brüggemann R., Luther B., Pudenz S., Simon U., Halfon E., Voigt K., Welzl G., Rediske G., Order Theoretical Tools in Environmental Sciences, Proceedings of the Second Workshop October 21st, 1999 held in Roskilde, Denmark, NERI – Technical Report No. 318, p. 7-44, National Environmental Research Institute, Roskilde, Denmark.

Brüggemann, R. & Welzl, G. 2002. Order Theory Meets Statistics, in: Voigt K., Welzl G., Rediske G., 4. Hasse Workshop 2001, Order Theoretical Tools in Environmental Sciences, 05. – 06.11.2001 Iffeldorf, Germany, pp., Shaker-Verlag, Aachen.

Commission of the European Communities, White Paper, Strategy for a Future Chemicals Policy, COM (2001) 88 final, http://www.europa.eu.int/comm/environment/chemicals/index.htm

Everitt, B.S. 1998. The Cambridge Dictionary of Statistics, Cambridge University Press, Cambridge,

Glander-Höbel, C., Voigt, K. & Welzl, G. 2001. Comparing Search Engines with Respect to Environmental Chemistry Using Statistical and Mathematical Evaluation Methods, in: Graham C., Online Information 2001, Proceedings, 4-6 December 2001 London, pp. 187-195, Learned Information Europe Ltd, Oxford.

Sawithki, G. 1996. Extensible Statistical Software: On a Voyage to Oberon, Journal of Computational and Graphical Statistics, 5, 3, pp. 263-283.

SPSS Science, 2001. Systat 10.
http://www.spssscience.com/SYSTAT/index.html

Shye, S. 1985. Multiple Scaling, Elsevier Publishers, Amsterdam.

Voigt, K., Welzl, G. & Rediske, G. 2000. Environmetrical Approaches to Evaluate Internet Databases, in: Sorensen, P.B., Carlsen, L., Mogensen, B.B., Brüggemann, R., Luther, B., Pudenz, S., Simon, U., Halfon, E., Voigt, K., Welzl, G. & Rediske, G. Order Theoretical Tools in Environmental Sciences, Proceedings of the Second Workshop October 21st, 1999 held in Roskilde, Denmark, NERI – Technical Report No. 318, pp. 135-144, National Environmental Research Institute, Roskilde, Denmark.

Voigt, K. & Welzl, G. 2001. Evaluation of Search Engines Concerning Environmental Terms, in: Hilty L.M., Gilgen P.W., Sustainability in the Information Society, 15th International Symposium Informatics for Environmental Protection, Zürich 2001, pp. 683-690, Metropolis-Verlag, Marburg.

Voigt, K. & Welzl, G. 2002a. Chemical Databases: An Overview of Selected Databases and Evaluation Methods, Online Information Review, 26,3, pp. 172-192.

Voigt, K. & Welzl, G. 2002b. Drinking Water Analysis Systems in German Cities: An Evaluation Approach Combining Hasse Diagram Technique with Multivariate Statistics, in: Voigt K., Welzl. G. (Eds.), Order Theoretical Tools in Environmental Sciences, Order Theory (Hasse Diagram Technique) Meets Multivatiate Statistics, pp. 113-140, Shaker-Verlag, Aachen.

Welzl, G., Voigt, K. & Rediske, G. 1998. Visualisation of environmental pollution - Hasse diagram technique and explorative statistical methods, pp. 101-110, Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences held on November 16th, 1998 in Berlin, Berichte des IGB 1998, Heft 6, Sonderheft I, Institut für Gewässerökologie und Binnenfischerei, Berlin.

Description of fish communities with help of partially ordered sets

Rainer Brüggemann[1] **, KristinaVoigt** [2]**,**
Gerhard Welzl [3]**, Peter B. Sørensen**[4]

1: Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Department, Ecohydrology, Berlin, Germany

2: GSF National Research Center for Environment and Health, Institute of Biomathematics and Biometry,
Neuherberg, Germany

3: GSF National Research Center for Environment and Health, Institute of Development Genetics, Neuherberg, Germany

4: National Environmental Research Institute,
Department of Policy Analysis, Roskilde, Denmark

Abstract

The concept of partially ordered sets is a useful tool to derive a ranking of objects of interest.

This paper is a first attempt to show what can be said about fish habitats derived from abundance measurements in the region around Berlin.

a) Tributaries of the creek Buckau (South west of Berlin). It is shown that an evaluation by diversity index will lead to contradictions. The concept of "Karamata-order" will be helpful for the way foward. In the Karamata order their partial sums are compared instead of comparing attributes directly.

b) Often the primary information (individuals number) is more valuable than the condensed information by diversity indices. An example will be provided by means of fish communities in the South east of Berlin. The corresponding Hasse diagram shows that it is possible to divide the habitates into two main types and that the site preference of fish species depends on the appearance of the species "*crucian carp*". Furthermore the dimension of the corresponding poset is 2. Therefore two latent attributes should suffice to explain the abundance tuples of 9 different fish species within the 12 different creeks in the wetland zone. In that context a Partially Ordered Scalogram Analysis with Coordinates (POSAC) will also be applied. The latent variables may be related to abiotic factors like water depth, or light climate. However, abundance may not completely follow abiotic

factors, because competition and carrying capacities also influence the population.

1 Introduction

Decisions are often needed on the value of a certain region, river or creek. Such decisions are especially relevant when maintenance of natural objects is associated with costs.

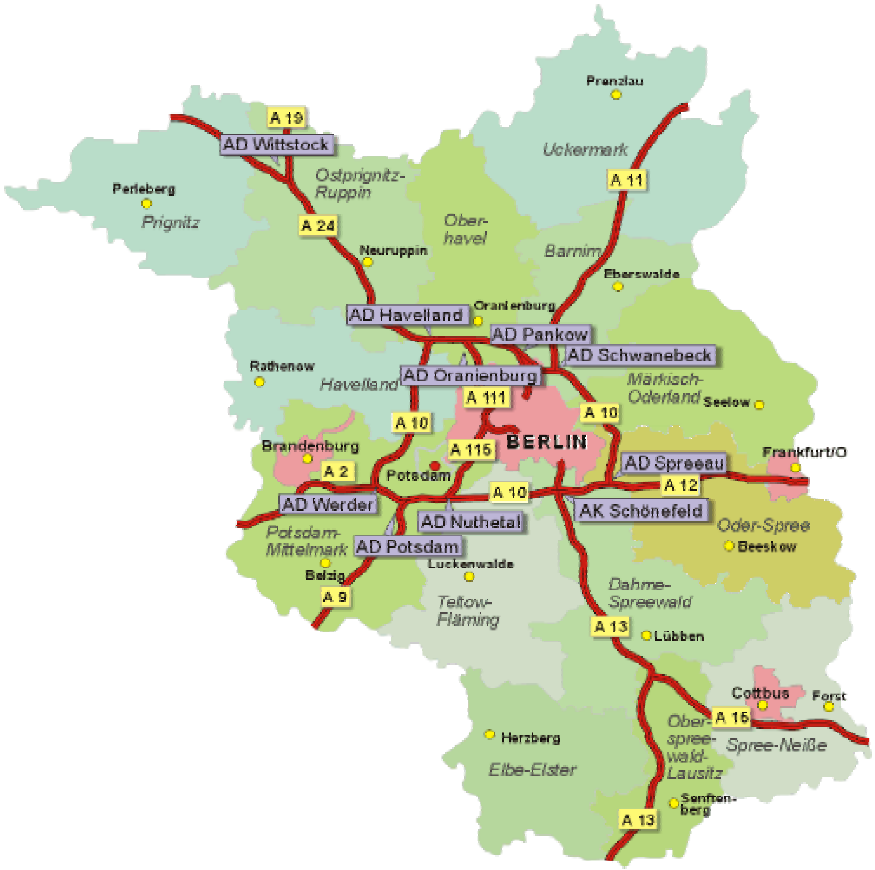


Figure 1: Map of state Brandenburg/Germany. Geographical information about both the creek system which is studied: Buckau river system: Teltow-Fläming; wetlands of Gosen: South east of Berlin, near "AD (German: Autobahndreieck) Spreeau".

This is for example the case in a wetland region in the south east of Berlin (see Figure 1) where creeks and the area between creeks have some importance because of

- landscape value
- the only habitat for a local pair of cranes
- rare animals and plants especially: rare waterplants, (German: Krebsschere)
- fish communities

On the other hand the maintenance of these creeks is associated with certain costs, because for example the water bulk flow velocity is to be regulated and the meadows are to be treated (at least if the character of an open landscape should be preserved).

Therefore in order to have a scientific basis for decision making, besides others, the fish communities are examined.

This presentation therefore will look for the interplay of diversity indices, abundancies and ecological classifications. As an example the rivers of the hilly region of the "Hohe Fläming" (south west of Berlin) will be studied (the Buckau-system). After that we will return to the wetland of Gosen where especially the role of some fish species are examined.

2 Two examples of fish communities

2.1 Buckau

2.1.1 Classification

The Buckau is a small river in the hilly region of the "Hohe Fläming" with several tributaries.

The fish communities in the "Hohe Fläming" were examined in order to derive a classification scheme for lowland rivers by fish abundancies. The fish species are classified according to the preference of bulk flow velocity and of spawning substrates. Table 1 shows the classification of the fish species found in the Buckau and its tributaries.

Table 1: Names and properties of fish species in the Buckau river system

No	Fish English name	Fish German name	Abbreviation	Preference bulk flow velocity (abbreviation)	Preference spawning (abbreviation)
1	brook lamprey	Bachneunauge	na	rheophil (R)	lithophil LALI
2	brown trout	Bachforelle	bf	rheophil (R)	lithophil LALI
3	pike	Hecht	ht	eurytop (E)	phytophil LAPH
4	gudgeon	Gründling	gr	rheophil (R)	psammophil LAPS
5	roach	Ploetze	p	eurytop (E)	phyto-lithophil LAPL
6	tench	Schleie	sc	limnophil (L)	phytophil LAPH
7	gibel carp	Giebel	gi	eurytop (E)	phytophil LAPH
8	stone loach	Bachschmerle	sm	rheophil (R)	psammophil LAPS
9	burbot	Quappe	qp	rheophil (R)	litho-pelagophil LALP
10	European eel	Flussaal	aa	eurytop (E)	pelagophil LAPE
11	three spined stickleback	Dreist. Stichling	ds	eurytop (E)	ariadnophil LAAR
12	dwarf form of 11	Zwergstichling	zs	eurytop (E)	ariadnophil LAAR

Applying Formal Concept Analysis one gets a mathematical graph, a lattice (Figure 2), (Ganter and Wille, 1996; Brüggemann and Drescher-Kaden, 2003). In brief terms this kind of graph is read as follows: Classifications (E,L, LALP, LAPS, etc.) are true for all those fish species which belong to the same concept (example B12: LALI is a classi-

fication characteristic which is true for the fish species na and bf) and to all other fish species, which can be reached downwards following the edges of the graph (example: LAPH is true for sc (B3) and ht, gi (B8)).

In the reverse direction: fish species have the classification characteristics, which can be reached following the edges of the graph upwards (for example: aa has the classification characteristic LAPE (B5) and E (B4)).

One can see that E and R are not suitable for classification (at least within the sample discussed here), because nearly all fish species belong either to E or to R. The spawning behaviour allows a finer classification: LAPE, LAAR and LAPL classify the E- fish species additionally, whereas LALP, LAPS, LALI classify the R-fish species. So for the subgroup of fish species found in Buckau and its tributaries, the property R is common for all those fish species, which need a sediment of sand or stones (psammophilic, lithophilic or litho/phytophilic). Phytophilic spawning behavior, however, may be found for some eurytopic fish species like ht, gi (concept no. 8) but also for limnophilic fish species, like sc.

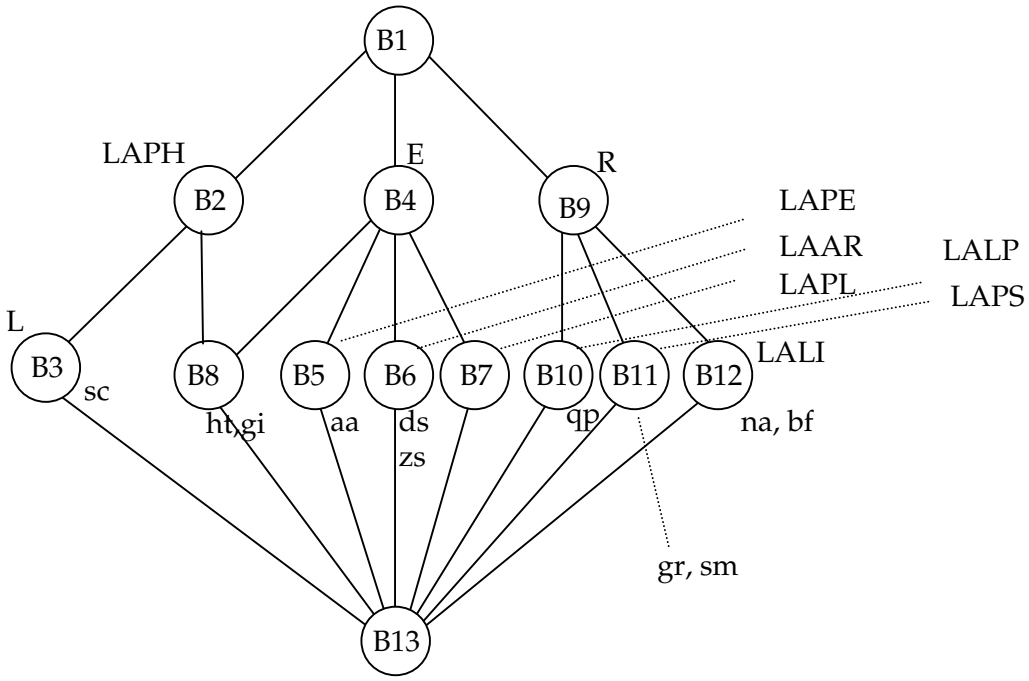


Figure 2: Lattice of formal concepts (B1,..,B13 from German: Begriffe). Fish abbreviation on the right side below; classifications: near the top of each circle.

The *tench* (sc) is (following the diagram (Figure 2)) a fish, characterized as limnophilic and phytophilic. The tributaries of the Buckau in general do not match these characteristics. The reason is that the *tench* was found in an artificial pond which gets the water from the Buckau tributaries and therefore examined. But this site is rather atypical compared with all other sites of Buckau.

2.1.2 Analysis of the rivers

The data matrix, found by field studies is shown in Table 2.

Table 2: Abundancies of 12 fish species in 12 rivers and river sections respectively

	bf	sm	gr	ds	zs	ht	aa	na	sc	gi	qp	p
H[a)]	194	19	4	62	36	4	3	0	0	0	0	0
Sp[a)]	0	0	0	47	66	0	0	0	0	0	0	0
K[a)]	0	13	0	30	7	0	0	0	0	0	0	0
Kr[a)]	75	0	0	26	34	0	0	6	0	0	0	0
Ka[a)]	100	0	0	11	0	0	0	0	0	0	0	0
G1[a)]	14	98	0	97	14	0	0	0	0	0	0	0
G2[a)]	0	0	0	74	20	0	0	0	1	7	0	0
Gr[a)]	13	21	0	86	16	0	0	0	0	0	0	0
Vw[a)]	563	33	0	61	8	0	0	4	0	0	1	0
R[a)]	850	1	0	137	0	0	0	14	4	0	0	0
L[a)]	102	2	2	47	2	0	0	0	0	0	0	1
Sb[a)]	0	0	1	58	60	0	0	1	0	0	0	0

[a)] The names of the rivers are german names: H: Herrenmühlengraben, Sp: Strepenbach, K: Kirchhainer Bach, Kr: Krumme Bache, Ka: Kalte Bache, G1: Geuenbach 1, G2: Geuenbach 2, Gr: Großbriesener Bach, Vw: Verlorenwasser, R: Riembach, L: Litzenbach, Sb: Strebenbach.

The Hasse diagram (Hasse diagram technique, see e.g. Brüggemann et al, 2001 and Brüggemann and Drescher-Kaden, 2003) based on raw data is very poor (Figure 3). The river K is mainly populated by the "*Three spined stickleback*", ds and by its dwarf form, zs, which indicates a monotonuous, rhithral river (i.e. rivers with low temperatures in the summer in contrary to potamal rivers, where temperatures get rather high). Rhithral rivers are mainly found in mountain regions. Those rivers are also populated by the *brown trouts*. However, here in the river K *brown trouts* are not found (compare with Table 2). Thus the dwarf form of ds is indicating rhithral lowland rivers, which are not preferred by *brown trouts* (perhaps because they do not contain the natural variety of alpine rivers).

The river Ka, is once again a rhithral river and in a natural state. Here the *brown trouts*, bf, is found in high abundancies. Thus the comparable rivers located at higher positions within the Hasse diagram might be classified as zs- and as bf-rivers. The exclusivity in the habitat preferences of both fish species, bf and zs, is also found by rather high values of the W-matrix (Figure 4). The W-matrix can be used as a measure of sensitivity, see details in Brüggemann and Halfon, 2000, or Brüggemann et al., 2001. In this part of the analysis the W-matrix result can be interpreted as the number of additional comparisons arising in Hasse diagram (Figure 3) when the specific fish specie is omitted in the ranking. E.g. the value of 10 for zs in Figure 4 shows that 10 extra comparisons are added to Figure 3 if zs is excluded as ranking parameter.

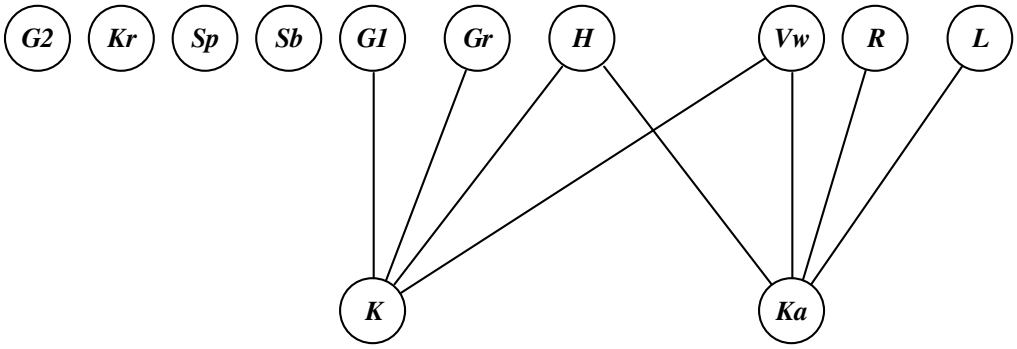


Figure 3: Hasse diagram based on 12 fish species mentioned above and 12 creeks/rivers/ponds of the Buckau system

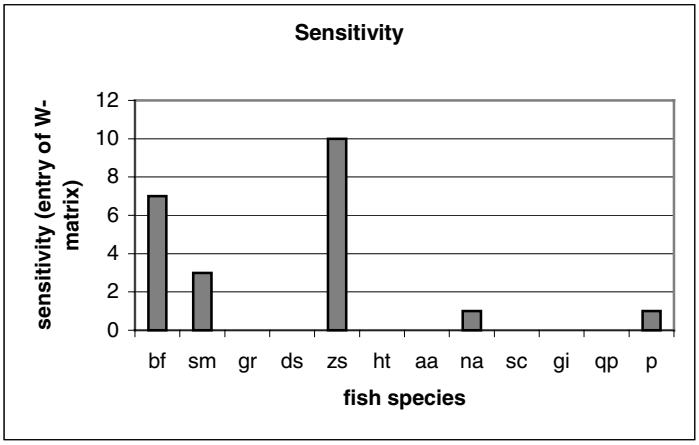


Figure 4: Sensitivity of the ranking of rivers due to their fish species composition and measured using the W-matrix

Because of the antagonism, exerted by bf and zs (either zs or bf) the ranking is very sensitive with respect to the abundancies of these two fish species.

Besides the two groups, which are predecessors of K and Ka respectively, there are four isolated elements, Kr, G2, Sp and Sb. Can we analyze what makes them different from the rest of the creeks? Using the computer programme WHASSE (Brüggemann et al, 1999) this is easily possible. Select: Calculate, Quotient set, Structure info, then an analysis for antagonisms is performed. About 81,3 % of the separation can be explained by just two attributes: sm and zs.

Kr, G2, Sp, Sb: sm low (indeed 0), zs rather high (66)

K, Ka, Gr, G1, H. Vw, R, L: zs low values, sm high values.

Examining the role of zs, sm alone leads to a Hasse diagram, shown in Figure 5. The separation is still not complete, which is seen by the connections to H and Ka.

A complete separation is achieved if 5 parameters are taken into account. A considerable improvement is the *brown trouts*, by which the degree of antagonism increases to 93.8 %. This is rather satisfactory

because we know that the fish species *Three spined stickleback* and its dwarf form resp and *brown trout* are somewhat antagonistic to each other.

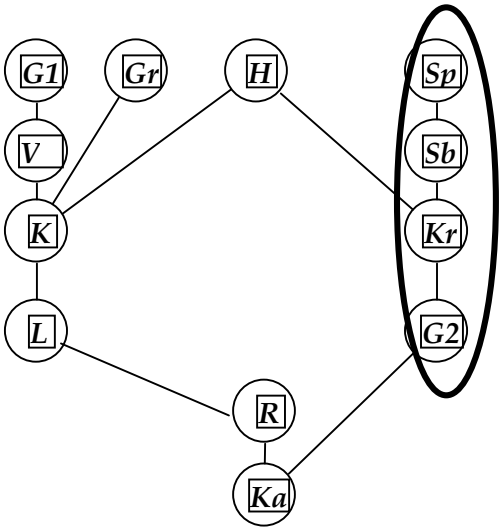


Figure 5: Role of the attributes zs and sm. The separation of the isolated elements is rather well reproduced (that the elements are themselves now comparable, is clearly not astonishing: We asked only for the separation of the two groups, not taking into account, that each member of ISO is isolated!)

The development of the degree of antagonism depending on the number of attributes is shown in Figure 6.

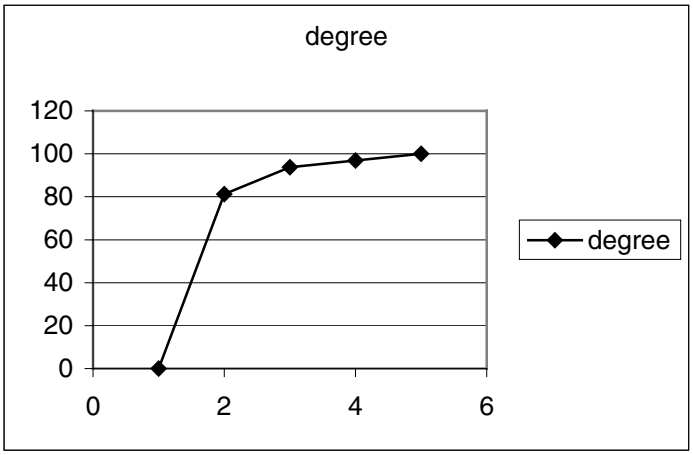


Figure 6: Degree of antagonism (in %) as a function of attributes. One can find several attribute sets, fulfilling the same degree of antagonism.

It is known that the *stone loach* (sm) prefers quickly flowing rivers, whereas it is rather tolerant with respect to O_2 deficits.

The Hasse diagram with four attributes, namely abundancies of *brown trout*, *stone loach*, *Three spined stickleback* and its dwarf form, i.e. bf, sm, ds, zs is shown in the following diagram (Figure 7). Four fish species may explain the variety of Buckau rivers quite well, because

this Hasse diagram is rather similar to the real one (with all fish species).

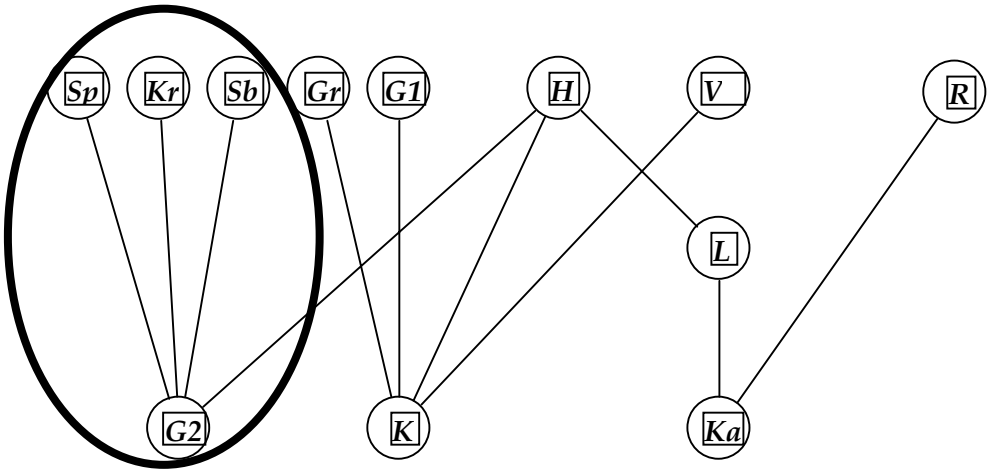


Figure 7: Attributes 1,2,4,5, i.e. abundances of bf, sm, ds, zs, which explain 96,9 % of the degree of the separation of the four rivers.

2.1.3 Buckau, the role of diversity

It is well known that the diversity is an open field of research. Worldwide attempts are made to establish a network of diversity observations (Ade, 2003). One consequence is that new diversity measures are invented and there may be more than 100 different types in wide use. Typically many landscape-planning processes rely on the diversity measure, preferring the habitats with a high diversity value. However, this implies that diversity indices can be used to rank the different options. Therefore the basic question is: Do different diversity indices lead to the same ranking?

Let us imagine that for some reason the rivers of Buckau has to be considered in relation to highway planning or some other activities. Among other questions, it may be worth to examine the rivers with regard to their diversity in order to be able to decide, which river is to exhibiting the highest diversity.

In the master thesis of C. Fieseler (Humboldt University of Berlin, Germany, 2002, unpublished) this has been done. Three diversity indices were calculated: the Eveness (E), the Shannon (Sh) and the Simpson (Sp) index (see Lande, 1996; Washington, 1984; Brüggemann et al., 2000a and Fromm & Brüggemann, 2001).

$Sh = - \sum p_i \ln(p_i)$ where p_i is the relative abundance of the i th species
$Sh' = 1+Sh$
$E = Sh' / \ln(S)$ where S is the number of species (not of the individuals)
$Sp = 1 - \sum p_i^2$

The Hasse diagram (not shown) based on these three parameters contains many contradictions, which means that if such highly aggregated numbers are used, the decision based on diversity measures will still be ambiguous. Relying on the two most typical diversity indices, a Hasse diagram as shown in Figure 8 is obtained. There are still incomparabilities! This means that a ranking, and the decision

based on the ranking may depend on the kind of diversity index used.

It is clear that it makes no sense to invent yet another index. It is a better strategy to develop an instrument by which we can decide, whether incomparabilities can be expected. Salomon (1979) derived a formalism for that. He states: "If the order is not based on individuals but on their cumulative distribution function we get a more robust result and we can still decide when a conflict as mentioned above will happen. The cumulative distribution function is found by ordering the tuples of each habitat, such that the largest normalized number of individuals appears first, then the next etc.". Clearly a normalization of each tuple to a constant, e.g. 1 will lead to an antichain if different tuples are compaired component-wise. Still worse: By ordering **within** the tuple according to the relative abundance, a component-wise comparison gives no sense, because the jth component of different rivers refer to different fish species.

Therefore another dominance relation must be introduced, which we call the 'Karamata order' after the mathematician Karamata, who has provided very useful results about this kind of ordering (cited in Beckenbach & Bellman, 1971). Note, that another mathematician, Muirhead introduces this kind of analysis already in the early 20th century (Muirhead, 1905) and that this kind of order is successfully applied in the field of Quantitative Structure Activity Relationships (QSAR) (see for example Randic, 1992 or, considering Wiener-indices Gutman et al., 2000).

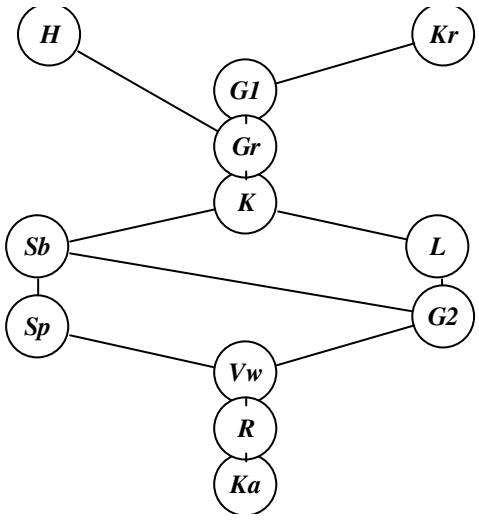


Figure 8: Hasse diagram of the 12 rivers of the Buckau system, using the two diversity parameters: Shannon (Sh) and Simpson (Sp).

In the following, we will explain the Karamata order step by step.

(1) We give to the ordered and normalized tuple a name, e.g. O_s . The index s refers to the specific habitat from which the relative abundancies were derived.

$O_s = (\sigma_1(n_i/N), \sigma_2(n_i/N), \ldots \sigma_N(n_i/N))$

$\sigma_i(n_i/N) \geq \sigma_j(n_i/N)$ if and only if i > j

Example: the abundancies of a fictitious habitat may be (1,5,1,0).

We normalize: N = 7

(1/7, 5/7, 1/7, 0/7)

We order within the tuple, i.e. $\sigma_1(n_i/N)$ will be 5/7 , $\sigma_2(n_i/N)$ = 1/7, $\sigma_3(n_i/N)$ = 1/7, $\sigma_4(n_i/N)$ = 0/7:

O_s = (1/7) * (5,1,1,0)

(2) Let us call π_k the partial sum from 1 to k formed from the components of the tuple O. :

$\pi_k = \Sigma\, \sigma_t(n_i/N)$, i = 1,…,k , k= 1,…N

where n_i is the number of individuals of the i th species, N the total number and σ refers to the ordering process mentioned above.

Continuing the example:

π_1 = 5/7 , π_2 = 6/7 , π_3 = 7/7 , π_4 = 7/7.

(3) By the partial sums generated for each river (or more general: for each habitat) a new tuple P_s of each river s can be generated:

P_s:=(π_1, π_2,…., π_N),

where the partial sums are generated from the normalized abundancies of fish species of the river s.

Continuing the example:

P_S = (1/7) * (5, 6, 7, 7)

(4) The Karamata order is now the component-wise order based on the tuples P.

Extending our example:

Let us imagine that the artificial example belongs to habitat A, i.e. :

P_A = (1/7)*(5, 6, 7, 7)

and let us add some other P-tuples (for the sake of simplicity we assume the same N for all other habitats):

O_B = (1/7) * (4, 3, 0, 0) $\Rightarrow P_B$ = (1/7) * (4, 7, 7, 7)
O_C = (1/7) * (4, 2, 1, 0) $\Rightarrow P_C$ = (1/7) * (4, 6, 7, 7)
O_D = (1/7) * (4, 1, 1, 1) $\Rightarrow P_D$ = (1/7) * (4, 5, 6, 7)

(5) Salomon (1979) shows that if $P_r < P_t$ then a broad class $\mathbb{C}$ of diversity indices will lead to the same but dual ranking:

$P_r < P_t$ implies $S_r > S_t$ for any diversity index $S \in \mathbb{C}$.

Thus, by those five steps the Karamata order and its application related to diversity is outlined. Note that Shannon-, Simpson- and many other diversity measures belong to C.

This result is promising as it generalizes the insights into the concept of diversity indices. With the help of YOUNG diagrams it is possible:

1. to relate the outcome of step 5 to the kind of distribution function, i.e. to the well known rank-abundancy diagrams (Begon et al, 1996).
2. to test the dominance (as in step 5) but also

Ad 1.:
The rank abundancy diagram uses the primary information, based on normalized and ordered abundancies (i.e. the tuple O) too. The curvature of this graph is of interest, or in its discrete form, the partitioning. The corresponding histogram (or bar diagram) is nothing else than the YOUNG diagram, see below.

Ad 2.:
The YOUNG diagrams help to find out whether the Karamata order will lead to an incomparability and therefore to dependence on the kind of diversity indices. YOUNG diagrams (see for example Brüggemann and Drescher-Kaden, 2003) are used in statistical mechanics and in quantum mechanics mostly when it is necessary to analyze the partitioning of integers (for example in quantum chemistry of the angular momentum).

Before we follow the way outlined in 2. the Karamata order, based on the P_k k=1,..,12 should be shown (Figure 9).

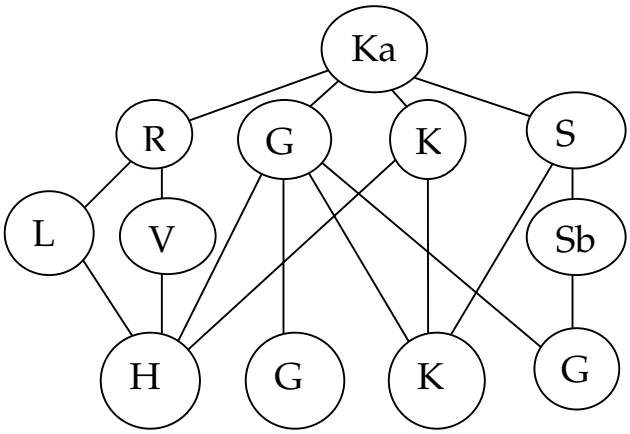


Figure 9: Hasse diagram of the P–tuples formed for each river of the Buckau system

One can easily check that each comparability in the diagram above is reproduced in that of the Sh', Sp– based diagram.

As a 6th step, the use of YOUNG-diagrams will be outlined.

The P – tuples can be considered as a partition of any integer. If one looks for example for the partitioning of the number 7 then one find beside others:

$7 = 5 + 1 + 1$, $7 = 4 + 3$, $7 = 4 + 2 + 1$, $7 = 4 + 1 + 1 + 1$

A partitioning dominates the other if the partial sums are dominating. We repeat the steps discussed before:
A: O_A=(5,1,1,0) P_A = (5,5+1, 5+1+1,7) = (5,6,7,7)
Similarly the other O- and P-tuples:
B: O_B=(4,3,0,0) P_B = (4,7,7,7)
C: O_C=(4,2,1,0) P_C = (4,6,7,7)
D: O_D=(4,1,1,1) P_D = (4,5,6,7)
Obviously: A>C>D and B>C>D but A || B

Instead of examining the P-tuple (i.e. calculating all the needed partial sums) the YOUNG diagram is based on the components of an O-tuple. Partitionings O_s like those presented above can be visualized by YOUNG diagrams (Figure 10).

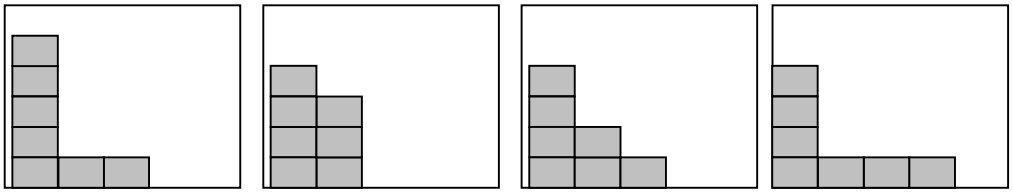


Figure 10: YOUNG diagrams visualizing some of the partitionings of the number 7. (See also: Brüggemann and Drescher-Kaden, 2003)

These partitionings drawn as YOUNG diagrams are the discrete form of the rank-abundancy diagrams, and relate therefore the primary information about abundancies to YOUNG-diagrams and these in turn to the diversity, by means of step 5. Therefore instead of testing the Karamata order by calculating the series of partial sums and performing then a component-wise comparison of each partial sum, it simply suffices to test the components of the tuple O for the four habitats A, B, C, D. The reason is that there is a theorem about YOUNG diagrams:

One diagram (partitioning) dominates the other, if and only if the two diagrams can be transformed by transfering units exclusively from the left side to the neighboring right side or in reverse direction.

This kind of transfer is not possible for A and B, but possible for the pairs A, C and A, D and C, D.

For the Buckau-river system (e.g. H and Kr, see Table 2) (normalized to 100 and rounded and truncated after the first 4 components for the sake of demonstration):
O_H = (60, 20, 10, 10)
O_{Kr}=(55, 25, 20, 0)

To draw a YOUNG diagram both tuples are divided by 5, thus we come up to:
O_H = (1/5) * (12, 4, 2, 2) and O_{Kr} = (1/5) * (11, 5, 4, 0) (see Figure 11).

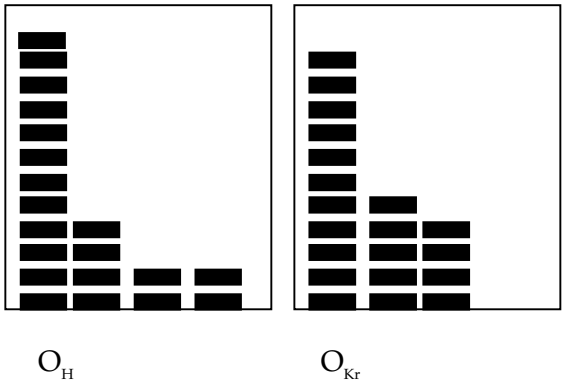


Figure 11: YOUNG diagrams of the rivers H and Kr

A transfer of the partitioning of H to that of Kr would only require that one unit is taken away from H. This has to be located to the right neighbour. But to get the lower part of Kr (.,.,4,0) one must transfer 2 units from the left tail of H to get 4 in Kr. Therefore two opposite transfers are needed, and therefore H and Kr are not Karamata-comparable, which in turn means that the diversity indices may differ (which is actually the case).

The interpretation is that diversity indices cannot differentiate distributions with a long even part, but one big exception, or distributions with two rather long even parts but no exception. This kind of analysis can systematize the use of aggregated numbers in ecology with this presentation as a first step.

2.2 Wetland of Gosen

2.2.1 Application of some HDT techniques

The system of creeks is shown in Figure 12. The creeks may in general terms be described as follows (Table 3).

Table 3: Two kinds of creeks in the wetlands of Gosen

Great Creeks	Meadow Creeks
long succesional history	-
reduced sun light	full sunshine
	macrophytes
slick	no slick
	O_2-deficits

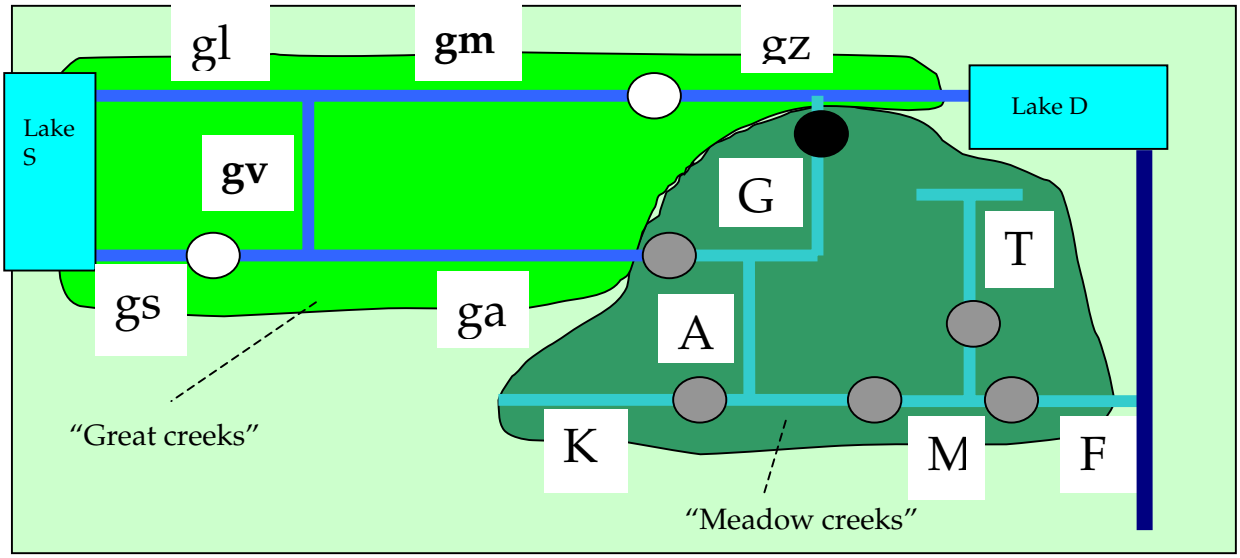


Figure 12: Topology of creeks in the wetlands of Gosen. Black circles: closed weirs, white circles: end of a section, grey circles: weirs temporarily closed to regulate the water current.

During May 1994 the creeks were examined with regard to fish communities by electro shock fishing (Table 4). The data are normalized according to the length of each section. The following fish species were at least found once: *roach*, p; *rudd*, r; *tench*, sc; *bleak*, u; *sunbleak*, m; *bream*, b; *crucian carp*, k; *pike*, ht; and *perch*, f.

Table 4: Fish abundancies (electro shock fishing, May 1994)

creek	p	r	sc	u	m	b	k	ht	f
K	0	0	26	0	0	0	6	4	3
M	1	1	41	1	41	1	0	35	50
A	0	0	30	0	0	0	21	7	4
G	0	0	17	0	0	0	18	6	10
T	0	0	72	0	0	0	21	17	9
F	0	0	0	0	0	0	0	0	0
gz	197	0	0	1	0	7	0	2	194
gm	94	0	0	1	0	7	0	1	24
gl	226	36	6	2	0	37	0	4	78
gv	99	27	10	2	1	7	0	4	46
ga	4	0	3	0	0	10	3	2	17
gs	124	25	2	0	0	12	0	2	72
max	226	36	72	2	41	37	21	35	194

Figure 13 shows the corresponding Hasse diagram. If connecting lines are interpreted as some kind of similar patterns of fish communities then it is comfortable to see that the partial order does not contradict the historical and morphometrical classification into the two creek systems. That means: If any connection is found then only creeks of the same type are comparable (the exception with the "0-element" F and with the "1-element" (max) do not contradict this finding). More details about morphometry and the Hasse diagram can be found in Brüggemann et al, 2002. The original task, which creek is to be protected, is now easily answered: 7 of 12 creeks might be important, because of their optima in certain fish communities.

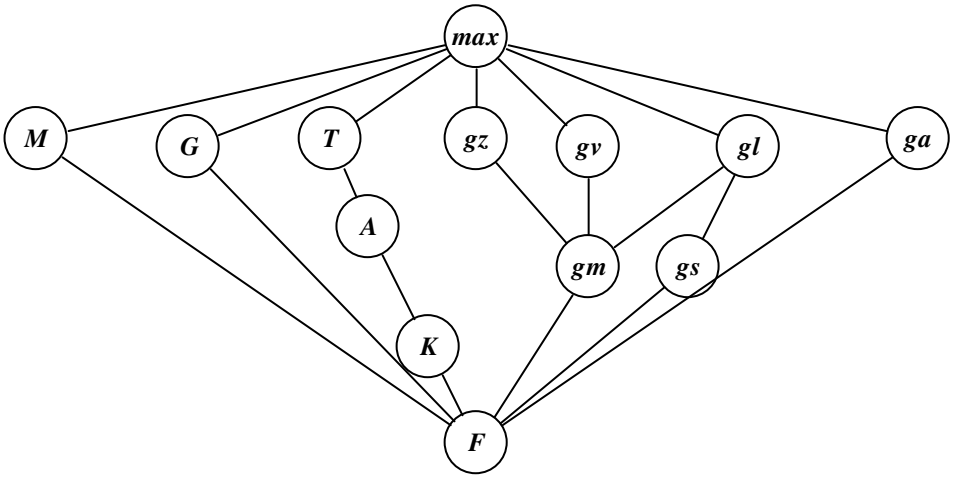


Figure 13: The Hasse diagram of the fish communities of the wetland of Gosen. Note, that the object "max" was artificially included.

Once again, going back to the fish classification according to their current – preferences, then it can be shown that the Meadow creeks are the preferred sites of limnophilic fish species, i.e. the morphometry and the hydraulics of these creeks can be characterized as still water zones, without remarkable water flows. The limnophilic fish species are quite well accomodated to habitats, for which O_2-deficits are at least temporarily possible. In Meadow creeks, which are not protected by bushes and trees high temperatures might result in a loss of dissolved oxygen, due to volatilization. Because of the presence of macrophytes there will at night an additional loss of O_2 because of plant biomass production. Contrary to the Meadow creeks, the Great Creeks are populated by eurytopic fish species, i.e. fish species that do not have a preference for currents. Beyond this the Great creeks are protected against sunlight by trees and bushes along them, therefore those fish species, which cannot accept O_2 – deficits, will be found here.

A standard test is to examine the W-matrix for the system. The first row of the W-matrix (W_{01}, W_{02}, …) is usually taken as a sensitivity measure (see for example: Brüggemann and Halfon, 2000). High values indicate that the corresponding attribute has a high influence on the ranking. Instead of reporting the values of the W-matrix a histogram of counts can be shown as an overview. The counts are drawn as a bar diagram in Figure 14. For example 5 times the entry of the W-matrix was 0 and 2 times the entry was 1 etc. There are only two striking attributes, that corresponding to the abundancies of *perch* (value 3) and that of the *crucian carp* with the value 4.

The diagram (without object "max" and without attribute "*crucian carp*") is shown in Figure 15. In comparison with the Hasse diagram including the *crucian carp*, the following changes occur: A < M and consequently: K < M; ga < gl and G < M.

Indeed, the *crucian carp* as a limnophilic fish as far as its bulk velocity preference is referred to, and as phytophilic fish, if the spawn behaviour is considered, cannot compete well with other fish species. Therefore the *crucian carp* is present in the creeks that are not inhabited by other fish species. Thus, the *crucian carp* is antagonistic to al-

most all other fish species. The *crucian carp* is indeed well accomodated to survive in habitats, where other fish species cannot survive: The *crucian carp* can tolerate O_2 deficits over several weeks.

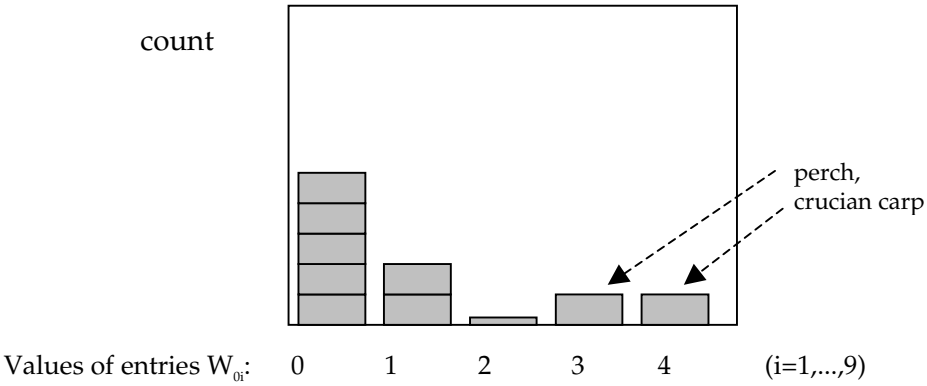


Figure 14: Distribution of values of the matrix entries W_{0i}

Now, let us go back to the main question: Which creek should be protected and which creek may be neglected in order to reduce the costs. Clearly those creeks, which are not maximal elements, are candidates to be neglected, being aware that some creeks of minor importance with respect to fish communities are needed to maintain the water exchange and the connection to the river Spree.

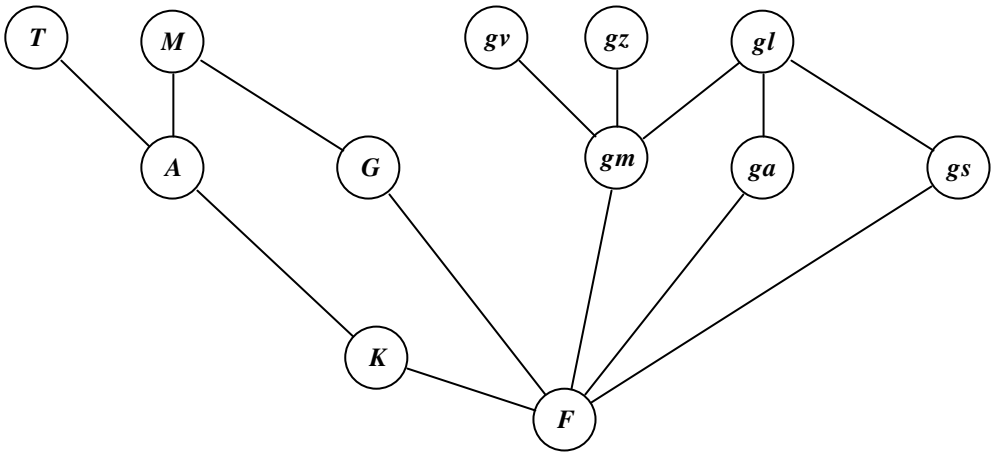


Figure 15: Hasse diagram of the fish communities in wetland of Gosen. Without the attribut: abundancy of *crucian carp*

The maximal elements should be considered more closely. In order to do this, one should try to get a fitness function expressing which creek might be the best to be protected. Such general fitness function does not exist, however, we know a little bit about the fitness. Obviously the fitness function should be a positive monotonuous function of the fish abundancies. The set of all linear extensions of a poset, includes all outcomes of fitness function, because linear extensions are order preserving, and therefore any ranking, due to the fitness function must be included in the set of linear extensions (see Brüggemann et al., 2000b, Brüggemann et al., 2001). If from these linear

extensions averaged ranks and their probability are deduced, we know the possible outcome of any fitness function, and additionally, the uncertainty due to ignoring the true relations among the fish abundancies. Therefore the averaged ranking derived from the set of linear extensions in addition to the probability distribution of getting a certain rank is called a General Ranking Model (GRM). (See also contributions of Sørensen et al., årstal mangler and Brüggemann et al., årstal mangler this workshop). The number of linear extensions is 997 920. The rank statistics are found in Table 5.

Table 5: Minimum, Averaged, Maximum Rank and its Range.

creek	Min	Averaged	Max	
	Rank	Rank	Rank	Range
K	2	4	10	8
M	2	7	12	10
A,	3	7	11	8
G	2	7	12	10
T	4	10	12	8
F	1	1	1	0
gz	3	8,3333	12	9
gm	2	3,6667	9	7
gl	4	9,4444	12	8
gv	3	8,3333	12	9
ga	2	7	12	10
gs	2	5,2222	11	9

The highest averaged ranks have gz and gl (gv). For two of the candidates of the Great creeks, the creeks gz and gl, the rank probability distribution function is shown (Figure 16).

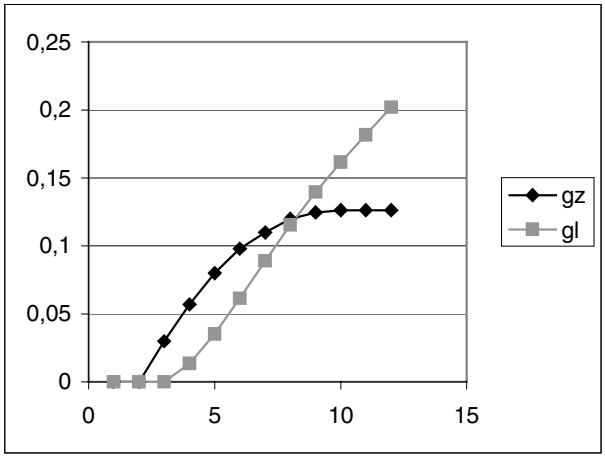


Figure 16: Rank probability distribution for two of the creeks with the highest averaged rank.

Whereas gz has some smeared out ranking, a range from 8 to 12 with nearly the same ranking probability, the creek gl gets the highest probability for the highest rank of 9.44. Thus this creek may be more closely examined for further protection. A similar study can and should be done for the Meadow creeks, because they have some singular appearance of fish species. Here the creek T is a very good candidate.

Figure 17 shows the probability of the creeks T and G, which are both maximal elements.

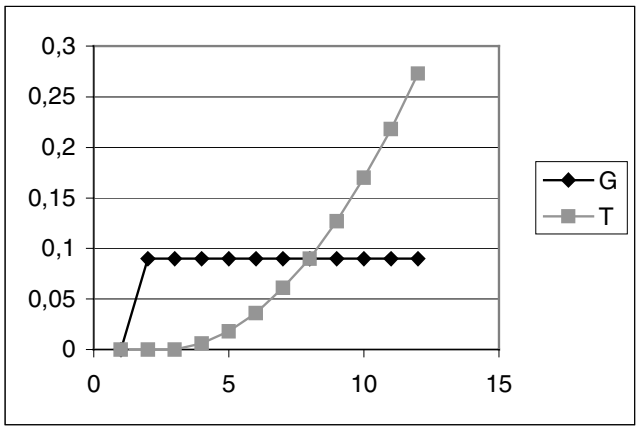


Figure 17: Rank-probability distributions of the Meadow creeks G and T

As one can see, the creek T would clearly be a more probable good habitat and therefore the aim of protection measures. Note: As this kind of representation of partial orders plays an increasing role, it is worth to state a well known fact based on the so-called Aleksandrov-Fenchel inequality: rank distribution functions derived from the set of linear extensions have to be unimodal (Daykin et al. 1984).

The question about, which of the maximal creeks should be selected for protection, can be reduced to the question of the mutual probability in the General Ranking Model GRM. It can easily be shown that the D matrix is of great use.

The D-matrix informs about common successors of any two objects. This means, the following formula is the basis of the D-matrix:

$$D_{x,y} := card\left[O(x) \bigcap O(y)\right]$$

O(x) and *O(y)* are the order ideals of the elements x and y. The operator card counts the number of elements in the corresponding set; here in the intersection of the two order ideals. So $D_{X,Y}$ denotes the number of elements, which are below both element x and y simultaniously in the partial ordered set. D_{XX} will in this way be the number of elements, which are below the element x. In order to estimate the mutual probability it is sufficient to examine the corresponding diagonal entries of the matrix D:

$prob_Q(x>y) = (D_{x,x}+1)/(D_{x,x}+D_{y,y}+2)$

x,y maximal elements

Numerically one gets:

$prob_Q(x>y) = (3+1)/(3+1+2) = 4/6 = 0.666$

Thus the probability of T > G in GRM is 0.666. There is a slight preference of the creek T in comparison to creek G.

2.2.2 Introduction into the Theory of Partially Ordered Scalogram Analysis with Coordinates (POSAC)

An overview article on the chemometrical and ecometrical research of the research group Biostatistics at the GSF - National Research Center for Environment and Health, Institute of Biomathematics and Biometry is recently given by Welzl et al., (2003). There are many statistical approaches of condensing a data-matrix by the creation of new variables. This process – called ordination – is often used to visualize relationships in two dimensions based on the first two variables. This idea can be applied when order relations are considered as the essential aspect of the data to be preserved in the analysis. This method – construction of new axes which presents correctly as many as possible of the order relations – is called Partially Ordered Scalogram Analysis with Coordinates (POSAC). POSAC is integrated in the program package SYSTAT 10 (see SPSS Science 2001) under the feature of statistics, data reduction. In POSAC, order relations are considered as the essential empirical-substantive aspect of the data to be preserved in the data analysis (see Borg & Shye, 1995). For a better interpretation of the new axes correlations between old and new variables can be calculated (F-statistics, Spearman rank, several methods). An informative book about the POSAC is Multiple Scaling by Shye (1985). For the convenience of the reader, therefore only a few remarks concerning the essential idea are cited in this article.

We have N objects, each of which is observed with respect to n variables. Every variable has a range of values, and the orientation of the values, i.e. whether a high value is good or bad is the same for all. The objects, here for example the creeks are characterized by n variables. These variables form a tuple and these tuples can be ordered as usual. Now, let us assume two creeks are comparable. For example if the two creeks A and K are considered (the small letters indicate different fish species) (see Table 6).

Table 6: Two tuples , referring to the creeks A and K

	p	r	sc	u	m	b	k	ht	f
A	0	0	30	0	0	0	21	7	4
K	0	0	26	0	0	0	6	4	3

Obviously A > K. To represent this order one new attribute would suffice and replace all 9 former variables. Now let us consider two other creeks, for example: gm, gs (Table 7).

Table 7: Two tuples, referring to the creeks gm and gs

	p	r	sc	u	m	b	k	ht	f
gm	94	0	0	1	0	7	0	1	24
gs	124	25	2	0	0	12	0	2	72

In general gs is more populated than gm and there would be no hindrance to prefer gs, if not for the rather rare fish, *bleak*, u, which is present in gm. Because we cannot balance, many fish species of one type vs one fish of another type, these two creeks are incomparable. The rare specie u is antagonistic to all other abundancies. (The concept of antagonism is explained in Simon and Brüggemann, 2000).

One can try to establish a set of pairs of attributes, which explain all comparabilities and incomparabilities. This set of all 2-score profiles may be thought of as a two dimensional Cartesian coordinate space with the one coordinate, X, indicating the first score and the other, Y, indicating the second score. Conversely, the two coordinates of each point in the XY space can be regarded as a two-score profile so that all points in the plane form a partially ordered set. The essential thing to notice is that for every given point in the coordinate space, three different regions in the space are determined:

I. A region of points that are greater than the given point

II. A region of points that are less that the given point.

III. A disjoint region of points that are incomparable to the given point.

This is illustrated in Figure 18 and is in more detail explained in the lecture of Voigt et al, this Workshop. Basically POSAC is concerned with the following question: Given a set A' of empirical tuples (as was found for the creeks in the wetland of Gosen), is it possible to assign two scores (that is, a point in the coordinate plane) to each profile in A', so that for any two observed profiles, their observed relation would be represented correctly by their corresponding two-coordinate profile?

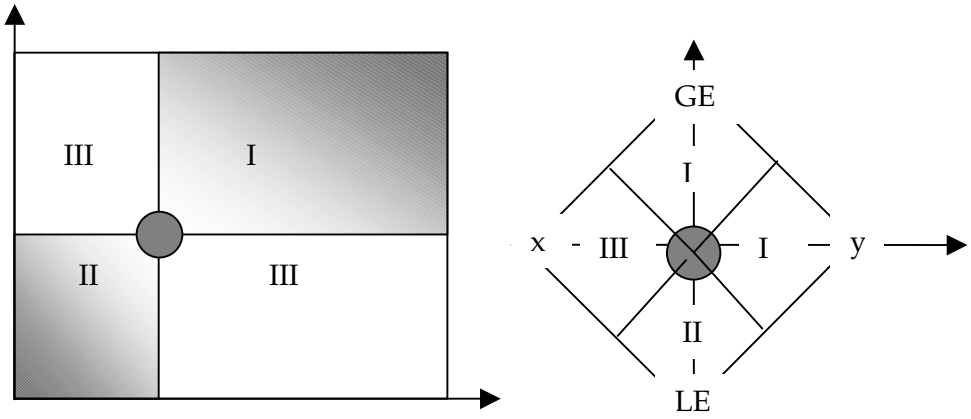


Figure 18: Regions in the Coordinate Space of POSAC (left) ; Hasse diagram - very schematically drawn as a large rectangle (with formally a greatest and a least element (GE, LE) and a large antichain from x to y and its relation to the regions of the Coordinate space (right side).

Could we as a first step find a meaning for the vertical direction? This indeed is possible. It is what in HDT is called a Level and it has to do with some averaged values of the attributes. Thus in the higher level (at least in graded posets, see Birkhoff, 1984) all the attributes have higher values than those in the lower one.

It is more difficult to assign a meaning to the horizontal direction. The difficulty will be immediately clear if one realizes that in the horizontal direction the order theory does not give any constraints with respect to the geometrical location, and even worse: If the normally arbitrary order of the attributes is changed one could underline different meanings (see Shye and Amar 1985). In POSAC theory the

vertical direction is called the joint axis or variable and the horizontal direction the lateral axis or variable. Besides these two axis one may try to find –as is done here- two latent axes / variables, which are thought of as an embedding of the poset (Figure 19). Back to the original question: Can we always find two latent variables? In general, the answer, is, of course, no. After plotting some observed profiles, it may well happen that a profile is encountered which cannot be located anywhere in the two dimensional coordinate space without misrepresenting some of its relations with the profiles already plotted. And in general there may not be a way of perfectly representing all order relations in the plane among exiting profiles of a given set. Here the connection to HDT is obvious. In HDT the dimension of the empirical poset is of importance. In Brüggemann et al (2000b), some hints on (theorems of Hiraguchi) how to calculate the dimension are given.

If the dimension is > 2, then any reduction to a graphical display in a plane will only be an approximation.

The practical POSAC problem is: Given a set A' of observed profiles, what mapping of these profiles into the two-dimensional coordinate spaces would best preserve all their relations, i.e. ≤-, ≥- and ||-relations. In order to deal with this question, a criterion must be specified by which one could determine whether a proposed mapping is better than another. When the output profile set is indeed two-dimensional or close to it, the POSAC solution is an approximation to the space of reduced dimension (Shye and Amar 1985).

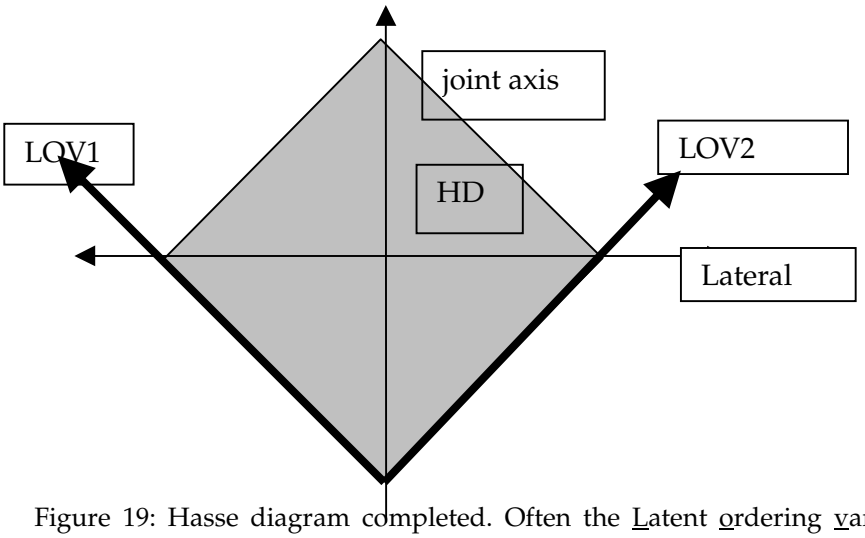


Figure 19: Hasse diagram completed. Often the Latent ordering variables (LOV) is also simply called DIM1 and DIM2, because they determine like x- and y axis in conventional geometry, the two embedding variables.

The POSAC method has already been applied on data-matrices in environmental sciences and chemistry. Welzl examined regions polluted with metals (Welzl et al., 1998). Pesticide Internet resources were analyzed with chemical and environmental evaluation criteria by Voigt et al. (2000; 2002). For further information see Voigt et al, this issue.

2.2.3 Results of POSAC analysis with respect to the wetlands of Gosen

One problem in applying POSAC in this specific case is that the Hasse diagram is not very structured. Clearly POSAC reproduces the separation of the two types of creeks. It even suggests investigating the following four classes according to the connected parts found after the POSAC procedure (exception F):

- T>A>K
- G
- ga>M
- gz>gm, gv>gm, g>gm, g>gs.

Besides the fact that POSAC finds an order relation, which was originally not present (ga>M) is due to the algorithm. There is no abiotic classification supporting this finding. The latent variables are listed in Table 8.

Table 8: Latent variables found from POSAC

creek	DIM 1	DIM 2
K	0,289	0,866
M	0,645	0,707
A	0,408	0,913
G	0,577	0,816
T	0,5	0,957
F	0	0
gz	0,816	0,645
gm	0,764	0,408
gl	0,957	0,5
gv	0,866	0,577
ga	0,707	0,764
gs	0,913	0,289
max	1	1

The two dimensions are assumed to represent the fish communities in a conventional form, namely by coordinates (Figure 20).

DIM1 may be interpreted as a measure of O_2-deficit, whereas the DIM2 may be interpreted as the water exchange and water flow bulk velocity. As one can see, ga and M are quite near to each other. Indeed the creek ga is the only one of the Great Creeks, where the *crucian carp* is living. Thus ga may belong to the group of the Meadow creeks from an abiotic point of view.

Using the Spearman correlation coefficient the relation between the new variables can be analyzed. First of all an order among the attibutes is found:

p-b-f-r-u-m-ht-sc-k.

This means that if the attributes were selected in the order given above to form a tuple, then the left side of the Hasse diagram would mainly be determined by values of p , whereas the right side of the Hasse diagram is mainly influenced by values of k. The (extremal)

attributes p (abundance of *roach*) and k (abundance of *crucian carp*) are called polar attributes, the fact that k is located on the opposite side of the ranking list, is understandable, after the WHASSE-analysis by means of the W-matrix mentioned above. The lateral variable should explain the general scale perpendicular to the vertical scale (related to more or less a sum of abundancies) of a Hasse diagram (see Figure 19). Shye is pointing out that it is sometimes very difficult to find a lateral scale (including a contextual meaning). The relation between the lateral and the one polar variable, p, is shown in Figure 21, whereas that of the lateral with k is shown in Figure 22.

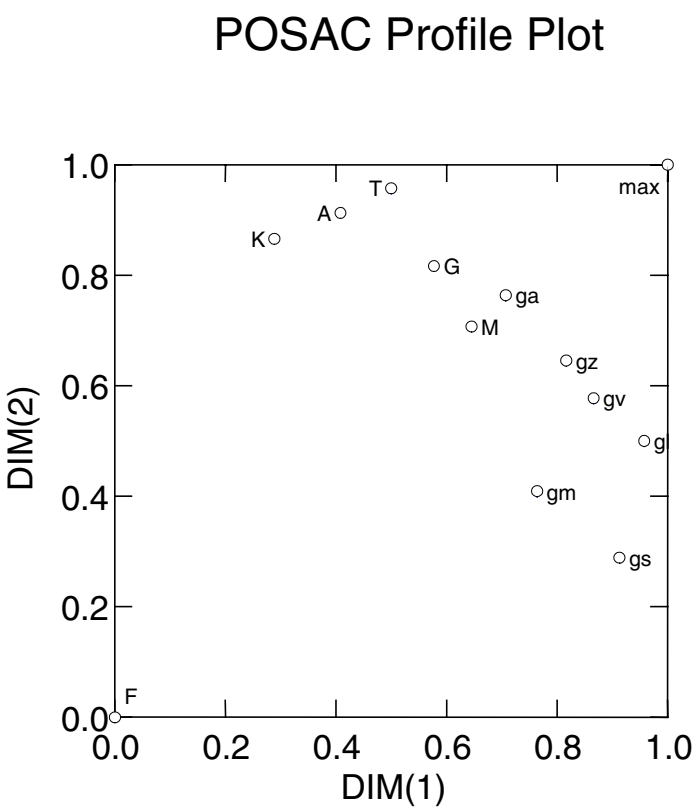


Figure 20: The presentation of the Hasse diagram in coordinates. The artificial object max is included to normalize to a 0 – 1.0 scale.

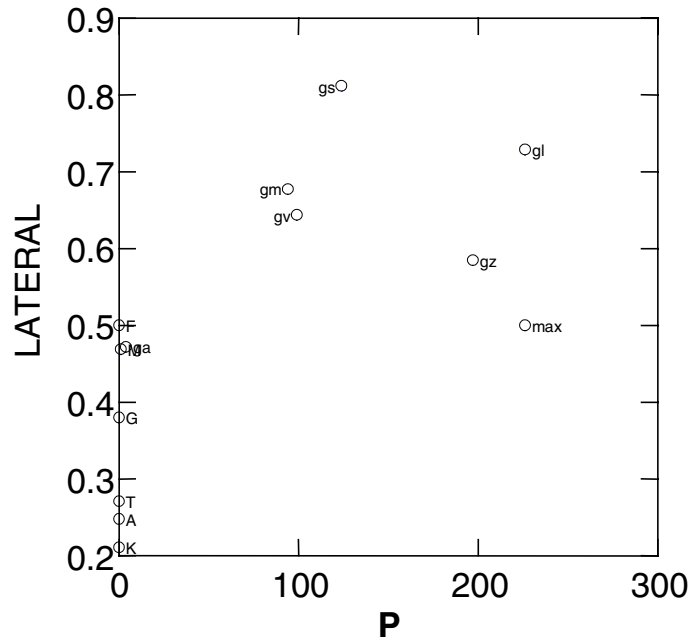


Figure 21: The lateral variable plotted against the polar variable p, the abundance of "*roach*".

The "*roach*" indicates preference of creeks with a certain bulk velocity. Therefore, the creeks are separated according to their original classification and additionally by the abundance of "*roach*", i.e. by the presence of eurytopic fish-habitats.

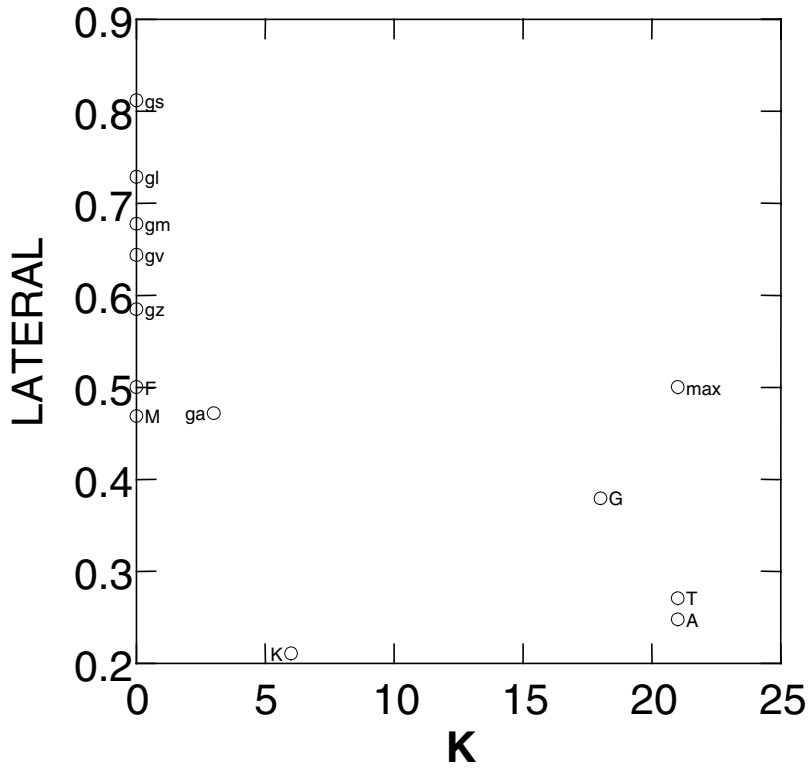


Figure 22: The lateral variable plotted against the abundance of *crucian carp*.

Whereas the *roach* is present exclusively in the Great creeks, the *crucian carp* is only present in the Meadow creeks. Thus the *crucian carp* may be used as an indicator fish for shallow, warm creeks which are not preferred as habitat for many other fish species.

3 Conclusion and Outlook

The discussion above shows that it is possible to obtain results about fish communities by partial order theory that can be used in practice. The clear separation of the creeks found in the Hasse diagram respects the general morphometric and abiotic description of the creeks. Beside this, it can be shown that specific fish species, here the *crucian carp*, survive, because they circumvent competition and accomodate to creeks, which are avoided by most other fish species.

The tree-like form of the Hasse diagram may suggest the idea that the empirical poset of biotic communities can be the result of a

- Driving force D, which tends to add more and more individuals.
- Carrying capacities C_i, which exert an upper limit for at least some habitat i, and
- Competition behaviour, I_{ij} regulating which animals win.
- Specific interaction P_{ij}, which specifies prey-predator interactions and which –not based on the general concept of availability of mineral nutrition – will be hard to quantify in general terms.

The driving force D will even in the case of presence of I lead to an increase of at least the winning competitor. The competitors, loosing, will have to balance out, in order to maintain C_i. The fact that two creek systems are found can be additionally modelled by the assumption, that there are at least two driving forces. Each of these driving forces tends to increase its abundancies in the corresponding subsystems.

Thus one may select the *roach* as one fish, which parametrize D_1 and the c*rucian carp* as the other fish which parametrize D_2. The sum of D_i may be interpreted as a joint variable in terms of partially ordered set theory. Other fish species will increase too, but if the C_1, C_2 hitherto unknown are reached, the driving forces are still acting which allows the winner to increase at the expence of all others. This qualitative model will in its most simple form lead to two trees (connected by F) which seems to model the empirical found Hasse diagram. If more than two nonexclusively acting driving forces are acting, then the Hasse diagram will look like a network, because high abundance of the one fish does not exclude that of other fish species, if and only if one is below the C_i-limits. This procedure has to be worked out in the future.

Furthermore in the future the combination of HDT with POSAC is envisaged.

References

Ade, M. (Personal communication, 2003)

Beckenbach E. F. & Bellman R. (1971) An Inequality of Karamata. In: *Inequalities* pp. 30-31. Springer-Verlag, Berlin.

Begon M., Harper J. L., & Townsend C. R. (1996) Ecology Individuals, Populations, and Communities, 3 edn. Blackwell Science, Oxford.

Birkhoff G. (1984) Lattice theory. American Mathematical Socienty, Vol: XXV, Providence, Rhode Island.

Borg I. & Shye S. (1995). Facet Theory, Form and Content. Sage Publications: Thousand Oaks.

Brüggemann R. & Halfon E. (2000) Introduction to the General Principles of Partial Order Ranking Theory. In: *Order Theoretical Tools in Environmental Sciences - Proceedings of the Second Workshop , October 21st, 1999 in Roskilde, Denmark* (eds P. B. Sørensen, L. Carlsen, B. B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U. Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, and F. Rediske) pp. 7-43. National Environmental Research Institute, Roskilde.

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.

Brüggemann R., Fredrich F., Wolter C., Pudenz S., & Steinberg C. (2002) Partielle Ordnungen: Ein Hilfsmittel zur Beschreibung von Artengemeinschaften. In: *Theorie und Modellierung von Ökosystemen - Workshop Kölpinsee 2000* (ed A. Gnauck) pp. 173-200. Shaker-Verlag, Aachen.

Brüggemann R., Fromm O., & Steinberg C. (2000a) Biodiversitätsmaβe - kritische Überlegungen aus naturwissenschaftlicher und ökonomischer Sicht. *Wasser & Boden* 52: 31-35.

Brüggemann R., Halfon E., Luther B., & Simon U. (2000b) New Tools in Hasse diagram technique - Example: Comparative Evaluation of Near - Shore - Sediments by a battery of tests. In: *Order Theoretical Tools in Environmental Sciences - Proceedings of the Second Workshop , October 21st, 1999 in Roskilde, Denmark* (eds P. B. Sørensen, L. Carlsen, B. B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U. Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, and F. Rediske) pp. 73-94. National Environmental Research Institute, Roskilde.

Brüggemann R., Halfon E., Welzl G., Voigt K., & Steinberg C. (2001) Applying the Concept of Partially Ordered Sets on the Ranking of

Near-Shore Sediments by a Battery of Tests. *J.Chem.Inf.Comp.Sc.* 41: 918-925.

Brüggemann,R. & Drescher-Kaden, U. (2003) Einführung in die modellgestützte Bewertung von Umweltchemikalien; Abschätzung Ausbreitung , Wirkung und Bewertung. Springer-Verlag, Berlin, Heidelberg, pp. 520

Daykin D. E., Daykin J. W., & Paterson M. S. (1984) On log Concavity for Order-Preserving Maps on Partial Orders. *Discrete Mathematics* 50: 221-226.

Fromm O. & Brüggemann R. (2001) Das Konzept der Artenvielfalt. Eine sinnvolle Ergänzung ökonomischer Naturbewertung? In: *Jahrbuch Ökologische Ökonomik 2: Ökonomische Naturbewertung* (eds J. Meyerhoff, U. Hampicke, and R. Marggraf) pp. 201-219. Metropolis-Verlag, Marburg.

Ganter B. & Wille R. (1996) Formale Begriffsanalyse Mathematische Grundlagen. Springer-Verlag, Berlin.

Gutman I., Rada J., & Araujo O. (2000) The Wiener Index of Starlike Trees and a Related Partial Order. *Match* 42: 145-154.

Lande R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 76(1): 5-13.

Muirhead R. F. (1905) Proofs of an Inequality. *Proc.Edingburgh Math.Soc.* 24: 45-50.

Randic, M. (1992): In Search of Structural Invariants. *J. Math. Chem.* 9, 97-146, 1992

Salomon D. L. (1979) A Comparative Approach to Species Diversity. In: *Ecological Diversity in Theory and Practice* (eds J. F. Grassie, G. P. Patil, W. Smith, and C. Taillie) pp. 29-35. International Co-operative Publishing House, Fairland, Maryland.

Shye S. (1985). Multiple Scaling. Elsevier Publishers: Amsterdam.

Shye S. & Amar R. (1985). Partial Order Scalogram Analysis by Base Coordinates and Lattice Mapping of the Items by their Scalogram Roles. In Facet Theory, Canter D (ed), Springer-Verlag: New York; 63-71.

Simon U. & Brüggemann R. (2000) Assessment of Water Management Strategies by Hasse diagram technique. In: *Order Theoretical Tools in Environmental Sciences - Proceedings of the Second Workshop, October 21st, 1999 in Roskilde, Denmark* (eds P. B. Sørensen, L. Carlsen, B. B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U. Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, and F. Rediske) pp. 117-134. National Environmental Research Institute, Roskilde.

Trotter WT. (1991). Combinatorics and Partially Ordered Sets Dimension Theory. The Johns Hopkins University Press: Baltimore, Maryland.

Voigt K., Welzl G. & Rediske G., (2000) Environmetrical Approaches to Evaluate Internet Databases, in: Sørensen P.B., Carlsen L., Mogensen B.B., Brüggemann R., Luther B., Pudenz S., Simon U., Halfon E., Voigt K., Welzl G., Rediske G., Order Theoretical Tools in Environmental Sciences, Proceedings of the Second Workshop October 21st, 1999 held in Roskilde, Denmark, NERI – Technical Report No. 318, pp. 135-144, National Environmental Research Institute, Roskilde, Denmark, 2000.

Voigt, K. & Welzl G. (Eds.) (2002): Order Theoretical Tools in Environmental Sciences. Order Theory (Hasse diagram technique) Meets Multivariate Statistics. Shaker-Verlag, Aachen, pp. 206

Washington H. G. (1984) Diversity, Biotic and Similarity Indices - A Review with special Relevance to Aquatic Ecosystems. *Wat.Res.* 18: 653-694.

Welzl G, Faus-Kessler T, Scherb H & Voigt K. (2003) in print. Biostatistics. In EOLSS Encyclopedia of Life Support Systems, Sydow A (ed.); EOLSS Publishers Co. Ltd.: Oxford.

Welzl G, Voigt K & Rediske G. (1998). Visualisation of environmental pollution - Hasse diagram technique and explorative statistical methods. In Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences held on November 16th, 1998 in Berlin, Berichte des IGB 1998, Heft 6, Sonderheft I, Institut für Gewässerökologie und Binnenfischerei (ed.); IGB: Berlin; 101-110.

An attempt of ecological assessment in urban zones – example Berlin (Germany)

Ute Simon*, Rainer Brüggemann*,

Michael Abs+, Mathias Erfmann°

* Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin,

SIMON@IGB-Berlin.de, brg@IGB-Berlin.de

michael.abs@snafu.de, ° erfman@web.de

Abstract

In the city of Berlin (Germany) 28 churchyards have been assessed with respect to their biodiversity (richness of species) using partial order theory in its most simple version, namely the so called Hasse Diagram Technique. As a result a priority list of churchyards with respect to their bird species diversity can be derived. Additionally to the ranking of the churchyards information about the species being unique for certain yards can be given. Following the attempt of key-species the information about species composition is used to describe and to differentiate the habitats of the churchyards. Both the priority list of churchyards as well as the analysis of habitat preferences seem to be a promising attempt to prepare faithful decision making in the field of nature conservation and urban planning.

1 Introduction

Parks, green and undeveloped areas are a limited resource in large cities. They, however, cover multiple tasks as "green lungs" such as recreation and ecological functions. As economical forces are getting stronger it is to be expected that not all (undeveloped) green areas can be saved for the reasons of nature conservation. An assessment of the ecological value of each area can help to decide which areas should be conserved. From an ecological point of view, the diversity of species might be a criterion to determine the ecological value (Costanza & Perrings 1990, Weitzman 1997, Perrings et al. 1997, Fromm & Brüggemann 2001). In the city of Berlin (Germany) 28 churchyards are evaluated with respect to the diversity of their sing-

ing birds populations (avifauna) using simple elements of partial order theory, namely the Hasse Diagram Technique. Within the example of the 28 churchyards, two different types of information are obtained. (1) A topological sorting the areas according to their richness (number of species) to identify the hotspots of biodiversity and (2) an evaluation of the species composition. By this procedure the habitats can also be described qualitatively to a certain degree. Together this information can be used for a competent decision, which of these areas can achieve highest priority for nature conservation beside some valuable ecological information.

2 Methods

The churchyards are spread all over the city of Berlin, Germany from the periphery to the centre and are of different size. Table 4 in the appendix gives a brief overview. The data on the avifauna has partly been collected from literature (Dobberkau et al. 1979, Elvers 1977, Otto & Scharon 1997, Schütz 1970, Wendland 1982) and was partly obtained by the mapping of breeding birds by the co-author Abs. A 28*65 matrix was formed: 28 rows, according to the number of churchyards and 65 columns, according to the number of bird-species observed. Because of extreme differences in the sizes of the churchyards, the data are transformed into binary information. The presence or absence of each of the 65 species are signified as one or zero respectively. The richness of an area is defined as the total number of observed species. The richness was classified in four classes named levels of richness (Table 1).

Table 1: Levels of richness

Richness Level	Number of species	No. of churchyards
1	36 - 39	14, 19, 28
2	24 - 30	2, 3, 9, 11, 15, 16, 29, 31
3	13 - 21	1, 5, 6, 13, 17, 20, 21, 23, 25, 30
4	6 - 10	4, 7, 8, 10, 12, 18, 26

The evaluation of the avifauna was executed by partial order theory. There are many possibilities to introduce an order relation. Here the tuples a and b of two churchyards "a" and "b" consisting of 65 components (bird species) are ordered as follows: $a \leq b \Leftrightarrow a(k) \leq b(k)$ for all $k \in \{1, \ldots, 65\}$. Furthermore the assignment to levels of richness (see above) is done by an inverse order preserving mapping: Applying the product order on the set of churchyards a Hassediagram results. From this levels can be derived. The levels can be considered as a set of new objects, which is totally ordered. Thus the original objects are mapped on a total order, whose ground set is the set of levels $\{\tau_1, \tau_2, \ldots\}$. This mapping is order preserving (Brüggemann & Steinberg, 2000). As example consider:

a (1, 1)
b (1, 0)
c (0, 1)
d (0, 0)

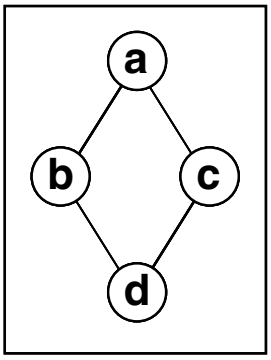


Figure 1: Example Hassediagram

The Hassediagram ordered by the pair of binary digits is shown in Figure 1. In this Hassediagram there are three levels, τ_1,..., τ_3, thus a map can be constructed. In Figure 2 this is summarised by calling φ the order-preserving map from the object set to the set of levels, and by i, the inversion.

({churchyards ...}, $(\{churchyards\},\leq) \xrightarrow{\varphi} (\{level\},\leq) \xrightarrow{i} (\{richness\},\leq)$

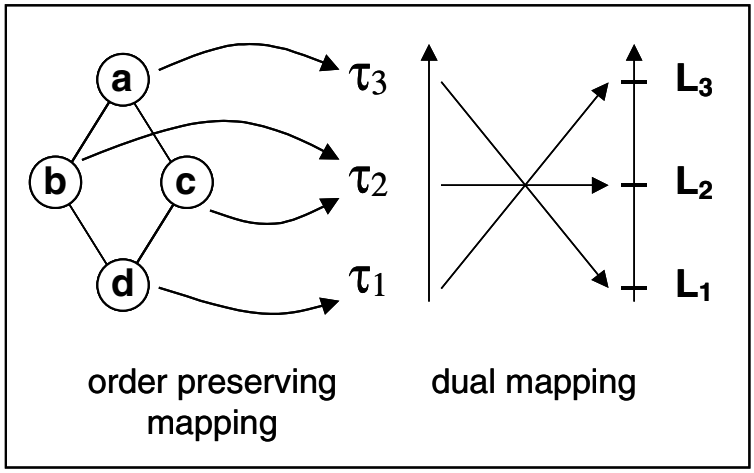


Figure 2: Order-preserving and dual mappings

Therefore the minimal elements of the partial order, i.e. churchyards, with a great richness become the level no. 1 etc. The simple construction of an order relation, mentioned above, is in environmental science known as Hasse diagram technique (HDT). This method is a specific discipline of Discrete Mathematics and combines graph-theoretical methods with basic elements of order theory. HDT may be seen as a multi-criteria assessment method, because the order relation can express a preference. Furthermore this technique avoids the comparison of different components of the species tuple; i.e. it avoids the preference among different species, which is for ecological reasons questionable. For more details see for example Brüggemann et al. (1999), Brüggemann and Drescher-Kaden (2003). Typically, one

obtains more than one option (here churchyards) as the result of the comparable assessment (partial order). The evaluation is done with respect to all indicator values, based on a ≥ comparison. Table 2 shows a small matrix with three options (O1, O2 and O3) and three indicators (I1, I2 and I3).

Table 2: Example of a data matrix. O1-O3= options, I1-I3= indicators

Options	Indicators		
	I1	**I2**	**I3**
O1	1	2	1
O2	1	1	0
O3	0	1	1

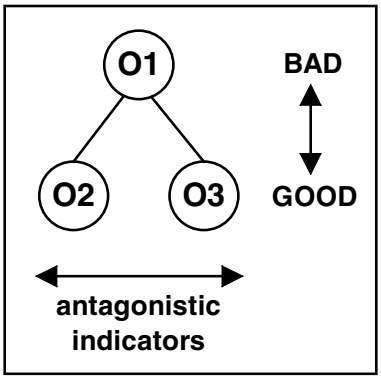


Figure 3: Hassediagram to Table 1

If high values indicate a bad evaluation, it is easy to recognize that option O1 in all indicators is evaluated worse than O2 and O3. Looking at the options O2 and O3, it is hard to say which one is better or worse because each option is evaluated better than the other one in one indicator and worse in another indicator. As a result, in HDT, these two options are incomparable with each other (Figure 3). To explain these incomparabilities, one can analyse the antagonistic indicators, which specifies the advantages and disadvantages of each assessed option (Simon & Brüggemann 2000).

3 Results

The assessment of the 28 churchyards due to the presence or absence of bird species is depicted in Figure 4. The diagram is arranged in four levels according to the species richness of the churchyards (Table 1). Only three churchyards (no. 14, 19 and 28) are at the first level with the highest number of species (36-39). They are depicted at the bottom of the diagram. At the second level of richness with 24 to 30 species there are eight churchyards (no. 2, 3, 9, 11, 15, 16, 29 and 31 as an isolated element). At the third level of richness between 13 and 21 species there are ten churchyards (no. 1, 5, 6, 13, 17, 20, 21, 23, 25 and 30). And at the fourth level of richness with only 6 to 10 species there are seven yards (no. 4, 7, 8, 10, 12, 18 and 26).

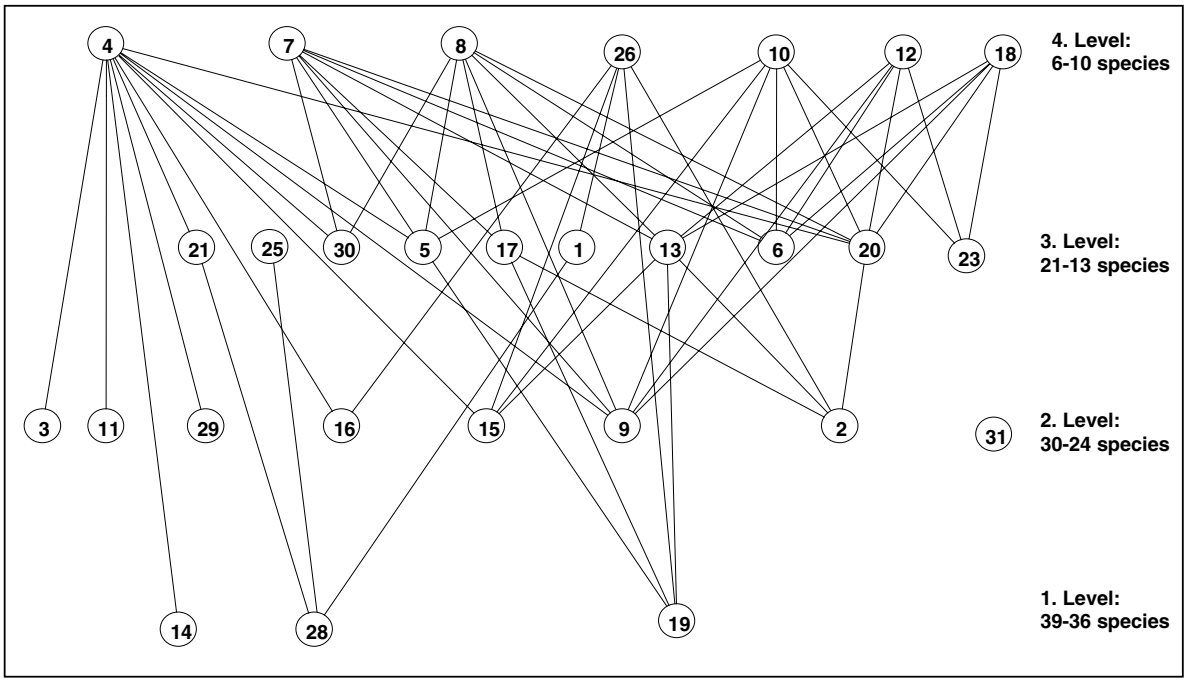


Figure 4: Hasse diagram of the evaluation of the 31 churchyards. The numbers in the circles are the short cuts of the churchyards (see Table 4).

All areas belonging to one level of richness are not comparable with each other, because each churchyard owns its specific composition of bird species. Yards at different levels, which are connected by a sequence of lines, are comparable with each other. This means that the yards got the same species composition but the yard located below in the diagram always has a higher number of species (richness) in total. Looking at the three churchyards with the highest richness it is remarkable that they are not comparable to areas at the second level of richness but only to those at the third or fourth level. They share the most basic inventory of bird species, namely the urbanised generalists and differ in composition concerning the more special birds. At the second level of richness one churchyard (no. 31) is isolated from all others. This area should be unique in its species composition as it is not comparable to any other yard.

The software program WHASSE (Brüggemann et al. 1999) includes the calculation of the stability of the diagram. The stability is a measurement of structural changes in the diagram when indicators, in our case bird species, would be added or removed. The diagram in Figure 4 has achieved a stability of 0.8. Therefore the addition of species would not influence the result of the assessment seriously whereas changes in the structure of the diagram are to be expected when species are removed from the matrix. The stability-value of 0.8 shows that the result of the assessment is rather robust against additional avifaunistical information.

4 Discussion

The richness of an area can be of general interest in the field of environmental planning and nature conservation, because a high richness stands for a high number of species living in the area under investigation. The species-richness is a good measure (indicator) of biodiversity. Additionally information about the kind of species living there is very helpful: Endangered species can give reason for special protection of their habitats for example. Furthermore the species composition itself often inform about the kind and to a certain degree about the ecological quality of a habitat (Flade 1994).

As described above the Hassediagram of Figure 4 provides information about the species-richness of the churchyards as well as the species composition. The richness is mapped at four levels. The species composition can be analysed by looking at the antagonistic indicators, which explain the incomparabilities between different yards. Looking for example at the churchyards with the highest richness, one find the bird species which has been observed only in one of these three yards (Table 3). These differences in the species compositions are with a high probability caused by environmental structures, for example the kind of habitat of the yard, the surrounding area or the size of the habitat. Comparing the Yards No. 14, 19 and 28 for example there are ten species were only observed in yard no. 14, seven species only in yard no. 19 and three only in yard no. 28. Furthermore eleven species are observed in both yards no. 19 and no. 28 but not in yard no. 14. There are only 5 species living in both yards no. 14 and no. 28 but not in no. 19 and there are no species at all living in both yards no. 14 and 19 but not in 28. Figure 5 shows schematically the distribution of bird species among these three churchyards.

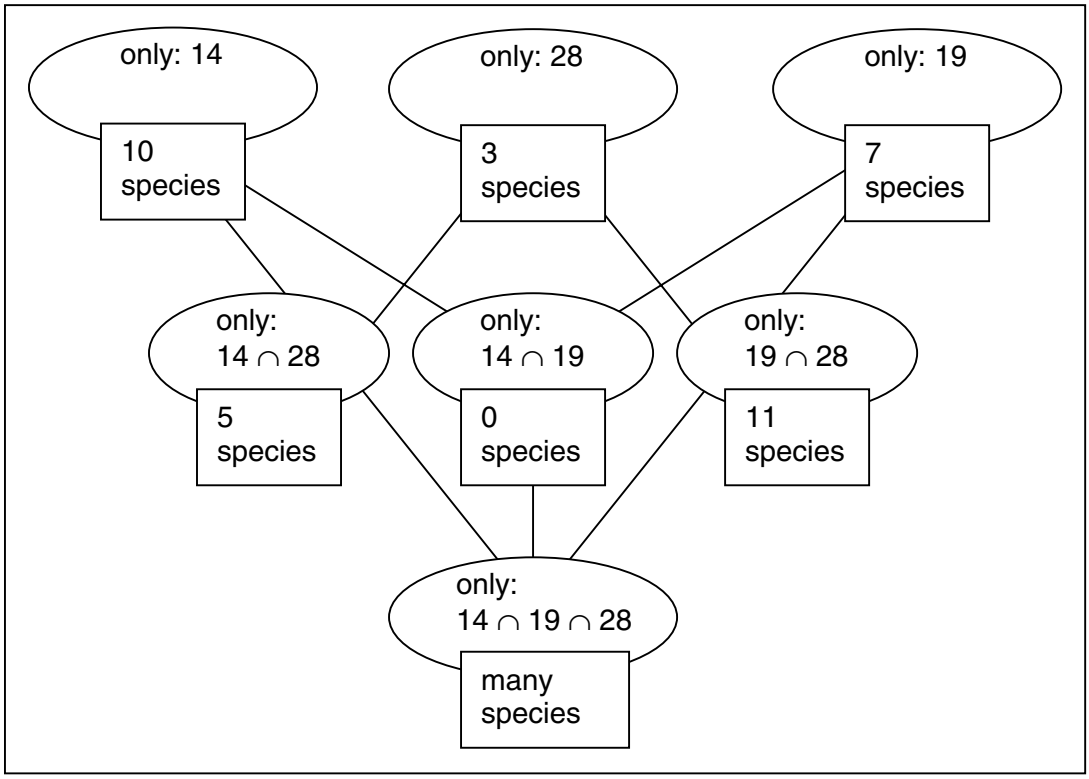


Figure 5: Schematical representation of bird species in 3 churchyards, 14, 19 and 28. In ellipses the churchyard-combinations are shown. In rectangles below the bird species which are inhabitants of that combination

From Figure 4 and 5 we deduce that 14 and 19 must be very specific with respect to the habitat conditions, whereas churchyard 28 allows a wide variety of birds.

Table 3: Example of unique species on churchyards. *: Bird of the Red List of endangered bird in Berlin.

Churchyard No.	Species only observed in this churchyard
14	Bullfinch*, Linnet, Crested Tit, Dunnock, Raven*, Black Woodpecker, Firecrest*, Coal Tit, Treecreeper, Goldcrest
19	Middle-Spotted Woodpecker*, Golden Oriole, Mallard, Marsh Warbler, Collared Dove, Tawny Owl, Wryneck*
28	Whitethroat, Common Buzzard, Long-eared Owl*

Thinking about species as indicators for certain habitats one can assume that the habitat of yard no. 14 should be somehow different especially from no. 19. Whereas the habitats of yards no. 19 and 28 seem to be more similar to each other because the number of common species is definitely higher. By looking at the species composition one finds that the habitats are of forest type. In yard no. 19 birds like the Marsh Warbler and the Mallard make it likely, that there is some water or marsh in the yard itself or in the nearby surrounding. Comparing the species, which are unique for only one yard with the key-species (Flade 1994) that yard no. 14 should be mainly a coniferous forest riddled with broad-leafed trees. Yards no. 19 and 28 should be mainly deciduous forest. Species like the Magpie, the Common Redstart or the Tree Sparrow shows that the habitat can be classified as of settlement-type in general. Thus the difference between yard no. 14 and yard no. 19 and 28 seem to be (1) the kind of forest (coniferous or

deciduous) and (2) the influences of the settlement. In yard no. 14 these influences seem to be lower. These statements are confirmed by observation and are helpful if further management decisions are necessary. Looking at the species mentioned in Table 3 one can find six species which are members of the so called "Red List" of endangered breeding birds of Berlin. In Table 3 they are marked with a *. Despite the high level of richness, this fact also confirms the special quality of these three yards. As written above the habitat of yard no. 31 seem to be very special as well, as the yard is an isolated element. The reason is the presence of the Grey Partridge, which is also a bird of the Red List of endangered species of Berlin. The Grey Partridge and the Common Pheasant in this yard may lead to the assumption that the habitat or the nearby surrounding might be field-like. Analysing the species composition of the yards of the lowest level of richness, one can find only very common species like the Greenfinch.

The abundancies of the species can be visualised by a Hassediagram as well (Figure 6). The species, which has been found in many yards, are listed in the upper part of the Hassediagram, whereas rare species living only in a few areas are at the bottom of the diagram. As it is to be expected, the abundancies of the species of the 28 yards shows a good accordance with the statement of the Red List and general ornithologic knowledge.

Trying to find the reasons for the high or low richness of the yards, it is very likely to examine the size of the yards should be examined. Indeed in general there is good evidence as the three yards with greatest biodiversity are also the biggest in size and the ones of lowest richness are the smallest. Of cause there are exceptions from this rule. Yard no. 9 for example belongs to the second level of richness but is relatively small, whereas yard no. 17 is quite large, but only at richness-level 3. Here the reasons should be searched in the surrounding area of the yards. Looking at the geographical position of the yards, it seems to be of minor influence if an area is located in the centre or the periphery of the city.

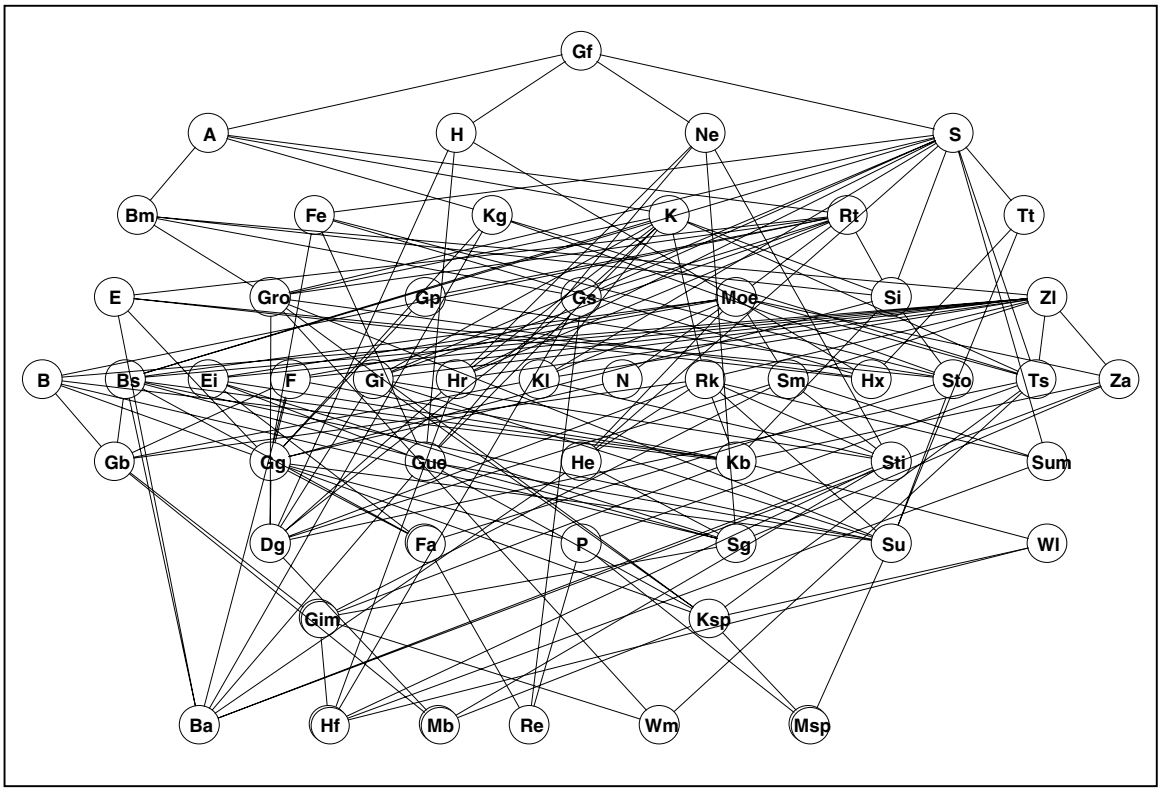


Figure 6: Extension of bird-species. Species with equivalent extensions are: (Hf, Kra, Sap, Wb), (Mb, Wo), (Msp, Wz, Wh), (Gim, Hm, Tm, Wg) and (Fa, Ku). Abbreviation of German names, for English and scientific names see Table 5.

5 Conclusions

The assessment of the 28 churchyards has in accordance with their species of breeding birds shown that even with a simple data-base containing only binary information, the Hasse Diagram Technique (HDT) leads to ornithologic reasonable results, which could be helpful not only for decision making in the field of nature protection and urban planning, but also to identify regularities in community structures. Especially a good background information by analysing the antagonistic indicators is obtained. Using for example the richness of an area as the only criterion of its ecological value, one might be wrong for purposes of endangered species. It is possible that areas with only a low number of species in total are special habitats for endangered species. In our example the HDT can be seen more like a data analysis tool than as a method for assessment. In a first step the data are sorted and the result is visualised in a diagram to give a good general overview. In a second step the information about the differences in the species composition enables the decision-makers to evaluate and differentiate the information about the species-richness looking for endangered species or for habitat conditions as outlined above. The analysis of antagonistic species may provide statements about the quality of the habitats. This technique should be developed in detail and a validation by further experimental results would be of great interest.

References

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: 170-178pp.

Brüggemann R. & Drescher-Kaden, U. 2003. Einführung in die mathematische Modellierung von Umweltchemikalien – Datenabschätzung, Ausbreitung und Verhalten, Wirkung von Chemikalien und Bewertung, Springer-Verlag, Heidelberg. (in press).

Costanza, R. & Perrings, C. 1990. A Flexible Assurance Bonding System for Improved Environmental Management. Ecological Economics, 2: 57-75 pp.

Dobberkau, T., Jander, G. & Otto, W. 1979. Untersuchungen zur Siedlungsdichte der Brutvögel Berliner Friedhöfe 1972. Beitr. Vogelkd. Leipzig 25 (3/4): 129-166.

Elvers, H. 1977. Die Brutvögel des Waldfriedhofes Heerstraße 1974. Orn. Ber. f. Berlin (West) 2 (2): 139-150.

Flade, M. 1994. Die Brutvogelgemeinschaften Mittel- und Norddeutschlands. Grundlagen für den Gebrauch vogelkundlicher Daten in der Landschaftsplanung. IHW-Verlag. Eching.

Fromm, O. & Brüggemann, R. 2001. Das Konzept der Artenvielfalt. Eine sinnvolle Ergänzung ökonomischer Naturbewertung? In: J. Meyerhoff, U. Hampicke and R. Marggraf (Editors), Jahrbuch Ökologische Ökonomik2: Ökonomische Naturbewertung. Metropolis-Verlag, Marburg, pp. 201-219.

Otto, W. & Scharon, J. 1997. Siedlungsdichte der Brutvögel einiger Berliner Friedhöfe. Berl. ornithol. Ber. 7: 38-57.

Perrings C., Mäler K.G., Folke C., Holling C.S. & Jansson B.O. 1997. Biodiversity loss – Economic and Ecological Issues. Cambridge University Press, Cambridge,1-332 pp.

Schütz, J. 1970. Die Brutvögel eines Friedhofes in Berlin-Neukölln. Berl. Naturschutzblätter 14, (41): 425-426.

Simon U. & Brüggemann, R. 2000. Assessment of Water Management Strategies by Hasse Diagram Technique. In: Order Theoretical Tools in Environmental Sciences. Proceedings of the second workshop October 21st, Roskilde, Denmark. NERI Technical report No. 318. Publisher: Ministry of Environment and Energy National Environmental Research Institute, Denmark.

Wendland, V. 1982. Die Vögel eines alten Friedhofs in Berlin (West). Orn. Ber. f. Berlin (West) 7 (2): 203-209.

Weitzman, M.L. 1997. Diversity functions. In: C. Perrings, K.-G. Mäler, C. Folke, C.S. Holling and B.O. Jansson (Editors), Biodiversity loss – Economic and Ecological Issues. Cambridge University Press, Cambridge, pp. 21-43.

Witt, K. 1991 Rote Liste der Brutvögel Berlins, 1. Fassung. Berl. ornithol. Ber. 1: 3-15.

Appendix

Table 4: List of churchyards

No.	Name	Size [ha]	Richness Level
1	Alter Friedhof der Nikolai- und Mariengemeinde 1997	3.5	3
2	Alter und Neuer Städtischer Friedhof Baumschulenweg 1972	33.8	2
3	Emmauskirchhof, St. Simeon- und St. Lukas Kirchhof 1997	12.8	2
4	Evangelischer Gemeindefriedhof Hohenschönhausen 1997	1.4	4
5	Friedhof Adlershof 1972	11.7	3
6	Friedhof am Mehringdamm 1966	6	3
7	Friedhof der Elisabeth-Gemeinde 1972	2.7	4
8	Friedhof der Friedrich-Werderschen, Dorotheenstädtischen und Französischen Gemeinde 1972	2.4	4
9	Friedhof der St. Laurentius-Gemeinde und Soldatenfriedhof 1972	8.4	2
10	Friedhof Gaillardstraße 1972	1.1	4
11	Friedhof Heiligensee 2000	13	2
12	Friedhof Pappelallee 1972	0.5	4
13	Friedhof St. Jacobigemeinde 1965	6	3
14	Friedhof Stahnsdorf 2001	50	1
15	Friedhöfe der Friedens- und Himmelfahrts, der Zion- und Gethsemanegemeinden 1972	33.3	2
16	Friedhöfe der St. Andreas-, St Markusgemeinde, der St. Hedwigs- und der St. Piusgemeinde 1997	26.3	2
17	Friedhöfe der St. Georgen II, St. Petri und St. Parochial I - Gemeinden 1972	22	3
18	Garnison Friedhof 1972	1	4
19	Jüdischer Friedhof Weißensee 1972	40	1
20	Neuer Freidhof der Nikolai- und Mariengemeinde und Freidhof I der Georgen-Parochialgemeinde 1997	7.2	3
21	Neuer St. Michael Friedhof 1997	5.1	3
23	St. Matthäus Friedhof 1999	4.7	3
25	Städtischer Friedhof Altglienicke 1997	2	3
26	Städtischer Friedhof Hohenschönhausen 1997	1	4
28	Städtischer Friedhof Marzahn 1997	20.7	1
29	Waldfriedhof Heerstarße 1974	12.5	2
30	Waldfriedhof Oberschönweide 1972	5.7	3
31	Zentralfriedof 1972	10	2

Table 5: List of bird species

Short Cut	German Name	*Scientifc Name*	English Name
A	Amsel	*Turdus merula*	Blackbird
Ba	Bachstelze	*Motacilla alba*	Pied Wagtail
Bp	Baumpieper	*Anthus trivialis*	Tree Pitpit
Bm	Blaumeise	*Parus caeruleus*	Blue Tit
B	Buchfink	*Fringilla coelebs*	Chaffinch
Bs	Buntspecht	*Picoides major*	Great Spotted Woodpecker
Dg	Dorngrasmücke	*Sylvia communis*	Whitethroat
Ei	Eichelhäher	*Garrulus glandarius*	Jay
E	Elster	*Pica pica*	Magpie
Fa	Fasan	*Phasianus colchicus*	Pheasant
Fe	Feldsperling	*Passer montanus*	Sparrow
F	Fitis	*Phylloscopus trochilus*	Willow Warbler
Gb	Gartenbaumläufer	*Certhia brachydactyla*	Short-toed Treecreeper
Gg	Gartengrasmücke	*Sylvia borin*	Garden Warbler
Gro	Gartenrotschwanz	*Phoenicurus phoenicurus*	Redstart
Gp	Gelbspötter	*Hippolais icterina*	Icterine Warbler
Gim	Gimpel	*Pyrrhula pyrrhula*	Bullfinch
Gi	Girlitz	*Serinus serinus*	Serin
G	Goldammer	*Emberiza citrinella*	Yellowhammer
Gs	Grauschnäpper	*Muscicapa striata*	Spotted Flycatcher
Gf	Grünfink	*Cardelius chloris*	Greenfinch
Gue	Grünspecht	*Picus viridis*	Green Woodpecker
Hf	Hänfling	*Carduelis cannabina*	Linnet
Hm	Haubenmeise	*Parus cristatus*	Crested Tit
Hr	Hausrotschwanz	*Phoenicurus ochruros*	Black Redstart
H	Haussperling	*Passer domesticus*	House Sparrow
He	Heckenbraunelle	*Prunella modularis*	Dunnock
Kb	Kernbeißer	*Coccothraustes coccothraustes*	Hawfinch
Kg	Klappergrasmücke	*Sylvia curruca*	Lesser Whitethroat
Kl	Kleiber	*Sitta europaea*	Nuthatch
Ksp	Kleinspecht	*Picoides minor*	Lesser Spotted Woodpecker
K	Kohlmeise	*Parus major*	Great Tit
Kra	Kolkrabe	*Corvus corax*	Raven
Ku	Kuckuck	*Cuculus canorus*	Cuckoo
Mb	Mäusebussard	*Buteo buteo*	Common Buzzard
Msp	Mittelspecht	*Picoides medius*	Middle-spotted Woodpecker
Moe	Mönchsgramücke	*Sylvia atricapilla*	Blackcap
N	Nachtigall	*Luscinia megarhynchos*	Nightingale
Ne	Nebelkrähe	*Corvus corone cornix*	Hooded Crow
P	Pirol	*Oriolus oriolus*	Golden Oriole
Re	Rebhuhn	*Perdix perdix*	Grey Partridge
Rt	Ringelatube	*Columba palumbus*	Wood Pigeon
Rk	Rotkehlchen	*Erithacus rubecula*	Robin
Sm	Schwanzmeise	*Aegithalos caudatus*	Long-tailed Tit
Sap	Schwarzspecht	*Dryocopus martius*	Black Woodpecker
Si	Singdrossel	*Turdus philomelos*	Song Thrush
Sg	Sommergoldhähnchen	*Regulus ignicapillus*	Firecrest
Hx	Stadttaube	*Columba livia*	Feral Pigeon
S	Star	*Sturnus vulgaris*	Starling
Sti	Stieglitz	*Carduelis carduelis*	Goldfinch
Sto	Stockente	*Anas platyrhynchos*	Mallard
Sum	Sumpfmeise	*Parus palustris*	Marsh Tit
Su	Sumpfrohrsänger	*Acrocephalus palustris*	Marsh Warbler
Tm	Tannenmeise	*Parus ater*	Coal Tit
Ts	Trauerschnäpper	*Ficedula hypoleuca*	Pied Flycatcher
Tt	Türkentaube	*Streptopelia decaocto*	Collared Dove
Tf	Turmfalke	*Falco tinnunculus*	Common Kestrel
Tut	Turteltaube	Streptopelia turtur	Turtle Dove
Wb	Waldbaumläufer	*Certhia familiaris*	Treecreeper
Wz	Waldkauz	*Strix aluco*	Tawny Owl
Wl	Waldlaubsänger	*Phylloscopus sibilatrix*	Wood Warbler
Wo	Waldohreule	*Asio otus*	Long-eared Owl
Wm	Weidenmeise	*Parus montanus*	Willow Tit
Wh	Wendehals	*Jynx torquilla*	Wryneck
Wg	Wintergoldhähnchen	*Regulus regulus*	Goldcrest
Za	Zaunkönig	*Troglodytes troglodytes*	Wren
Zl	Zilpzalp	*Phylloscopus collybita*	Chiffchaff

A decision support tool to prioritize chemical substances.

Partial order ranking using QSAR generated descriptors

Lars Carlsen[1*], Peter B. Sørensen[2] and Dorte B. Lerche[2]

[1] Awareness Center,
Hyldeholm 4,
Veddelev,
DK-4000 Roskilde,
Denmark
(e-mail: LC@AwarenessCenter.dk)

[2] National Environmental Research Institute,
Department of Policy Analysis,
DK-4000 Roskilde,
Denmark

Abstract

The selection and prioritization of chemical substances constitutes an important task in the possible regulation of chemicals. Partial order ranking appears as an efficient tool in this respect. Due to the shortage of experimental data, Quantitative Structure Activity Relationship (QSAR) estimates for endpoints appear as an attractive alternative. The present paper suggests a simple Decision Support Tool based on partial order ranking using QSAR generated descriptors that may help regulatory bodies as well as companies producing and/or using chemicals to disclose the environmentally more hazardous substances. The use of linear extensions in the selection process is demonstrated. The descriptors included in the present study comprise biodegradation, bioaccumulation and toxicity.

1 Introduction

The selection and prioritization of chemicals based on their potential hazard to man and environment is an important task to regulatory bodies. Thus, substances that possess so-called PBT characteristics (Persistent, Bioaccumulating and Toxic) receive in this context special

attention. However, the shortage of experimental data is obvious. Thus, according to the European Commission only in the case of approx. 14% of the HPV chemicals on the EINECS list, comprising 100.116 entries, the minimum required data for evaluating the chemicals were available. For approx. 21% of the compounds no data at all concerning the impact on the environmental and human health were found (EINECS, 1967). In a study by the Danish EPA (Niëmela, 1994) it was concluded that even in major sources of test data, such as RTECS (RTECS) and AQUIRE (AQUIRE) data on selected ecotoxicological effects could be found only for very limited number of the compounds on the EINECS list (Acute toxic effect: 10.5%, Reproductive damage: 2.2%, Genetic damage: 3.2%, Carcinogenic effect: 1.6%, Effect on the aquatic environment: 3.5%). Furthermore, intensive and experimental evaluations of chemicals are rather costly (Walker et al. 2002 and references therein). Thus, QSAR derived data for physicochemical as well as toxicological endpoints appear as an attractive alternative.

The prioritization of chemicals may take place based on selected criteria. Typically, it is advisable simultaneously to include a set of criteria, as e.g. criteria for persistence, bioaccumulation and toxicity. Typically this has been done by transforming the included criteria into one single criteria (for a discussion please see Lerche et al., 2002). The purpose of this paper is to present partial order ranking as an advantageous method to prioritize chemicals using a selection of criteria simultaneously. In the present study, we have applied this approach for prioritizing PBT substances, the single P (persistence), B (bioaccumulation) and T (Toxicity) characteristics of the chemicals being derived by QSAR.

As an illustrative example the present study includes 50 arbitrarily chosen potential PBT substances, 9 of these are being high production volume chemicals, the remaining 41 being medium production volume chemicals.

2 Methods

2.1 QSAR Modeling

QSAR modeling of the PBT characteristics of the substances studied was based on the appropriate modules in the EPI Suite (EPI, 2000). Thus, the persistence was addressed using the various biodegradation probabilities (cf. EPI, 2000) as measures; the present study applies BDP2 and BDP3 as expressed through the results of the BioWin module. In the cases of the BPP2, predicted values lower than 0.5 indicate that the substance does NOT biodegrade fast. In the cases of BPP3 (primary biodegradation) predicted values in the ranges 5.0-4.0, 4.0-3.0, 3.0-2.0, 2.0-1.0 and <1.0 indicate that biodegradation will take place within hours, days, weeks, months or longer than months, respectively. Chemicals with BDPs in the interval of 1.75 to 2 are asso-

ciated with a medium persistence potential, and BDPs smaller than 1.75 were assigned a high persistence potential.

Bioaccumulation was assessed using the BCFWin module. Chemicals with BCFs >1,000, but < 5,000 were assigned a medium bioconcentration potential. Chemicals with BCFs > 5,000 were assigned a high bioconcentration potential.

Neutral Baseline Toxicity, i.e., the calculated toxicity based on the acute neutral organics model, and chronic toxicity for algae, daphnids and fish, respectively, were derived using the ECOSAR module of the EPI Suite.

2.2 Partial Order Ranking

The theory of partial order ranking is presented elsewhere (e.g. in Davey and Priestley, 1990) and application in relation to QSAR is presented in previous papers (Carlsen et al., 2001; Brüggemann et al., 2001a; Carlsen et al., 2002). In brief, Partial Order Ranking is a simple principle, which a priori includes "≤" as the only mathematical relation. If a system is considered, which can be described by a series of descriptors p_i, a given compound A, characterized by the descriptors $p_i(A)$ can be compared to another compound B, characterized by the descriptors $p_i(B)$, through comparison of the single descriptors, respectively. Thus, compound A will be ranked higher than compound B, i.e., B ≤ A, if at least one descriptor for A is higher than the corresponding descriptor for B and no descriptor for A is lower than the corresponding descriptor for B. If, on the other hand, $p_i(A)>p_i(B)$ for descriptor i and $p_j(A)< p_j(B)$ for descriptor j, A and B will be denoted incomparable. In mathematical terms this can be expressed as

$$B \leq A \Leftrightarrow p_i(B) \leq p_i(A) \text{ for } \mathbf{all}\ i \qquad 1)$$

Obviously, if all descriptors for A are equal to the corresponding descriptors for B, i.e., $p_i(B) = p_i(A)$ for all i, the two compounds will have identical rank and will be considered as equivalent. It further follows that if A ≤ B and B ≤ C then A ≤ C. If no rank can be established between A and B these compounds are denoted as incomparable, i.e. they cannot be assigned a mutual order.

In partial order ranking – in contrast to standard multidimensional statistical analysis - neither assumptions about linearity nor any assumptions about distribution properties are made. In this way the partial order ranking can be considered as a non-parametric method. Thus, there is no preference among the descriptors. The graphical representation of the partial ordering is often given in a so-called Hasse diagram (Hasse, 1952; Halfon & Reggiani, 1986; Brüggemann et al., 2001b; Brüggemann et al., 1995).

2.3 Linear Extensions

The number of incomparable elements in the partial ordering may obviously constitute a limitation in the attempt to rank e.g. a series of chemical substances based on their potential environmental or hu-

man health hazard. To a certain extent this problem can be remedied through the application of the so-called linear extensions of the partial order ranking (Fishburn, 1974; Graham, 1983). A linear extension is a total order, where all comparabilities of the partial order are reproduced (Brüggemann et al., 2001b; Davey & Priestley, 1990). Due to the incomparisons in the partial order ranking, a number of possible linear extensions corresponds to one partial order. If all possible linear extensions are found, a ranking probability can be calculated, i.e., based on the linear extensions the probability that a certain compound have a certain absolute rank can be derived. If all possible linear extensions are found it is possible to calculate the average ranks of the single elements in a partially ordered set (Winkler, 1982; 1983). The average rank is simply the average of the ranks in all the linear extensions. On this basis the most probably rank for each element can be obtained leading to the most probably linear rank of the substances studied.

It appears virtually impossible to generate all linear extensions for Hasse diagrams containing more than 10-15 elements. Thus, in practice it has been demonstrated that fairly accurate predictions can be obtained based on a randomly selected fraction of the total set of linear extensions (Sørensen et al., 2001; Sørensen and Lerche, 2002a). In the present study the algorithm developed by Sørensen et al.(2001) and improved by Lerche et al., (2002b) has been used for generating a randomly selected fraction of linear extensions.

2.4 Results and discussion

The evaluation of chemicals for their potential environmental and/or human health effects will typically involve a series of parameters/descriptors such as (bio)degradation half life, the bioaccumulation potential and toxicity. For this purpose the partial order ranking methodology appears as an effective decision support tool. Hence, let us assume that a suite of 10 compounds has to be evaluated and that the evaluation should be based on 3 pre-selected criteria. To illustrate this we generated a Hasse diagram containing 10 elements, the individual values of 3 descriptors being chosen as random numbers between 0 and 1. The resulting Hasse diagram is depicted in Figure 1A. It is immediately seen that the 10 compounds are divided into 3 groups corresponding to the 3 levels in the diagram. Assuming that the 3 descriptors represented biodegradation, bioaccumulation and toxicity, respectively, in a way so that the more persistent, the more bioaccumulating and the more toxic the substance would be the higher in the diagram it would be found. Thus, on a cumulative basis the compounds 1, 3, 4, 7 and 8 can be classified as the environmentally more problematic of the 10 compounds studied, whereas compound 10 apparently among these 10 compounds are the less hazardous.

Studies based on actual scenarios will often include a higher number of compounds. Thus, it will typically not be possible to deal with all compounds included in the study simultaneously. The partial order ranking will obviously lead to important information as to which substances that primarily should be dealt with, e.g., through restric-

tion in the use of the compounds or substitution with other less hazardous compounds.

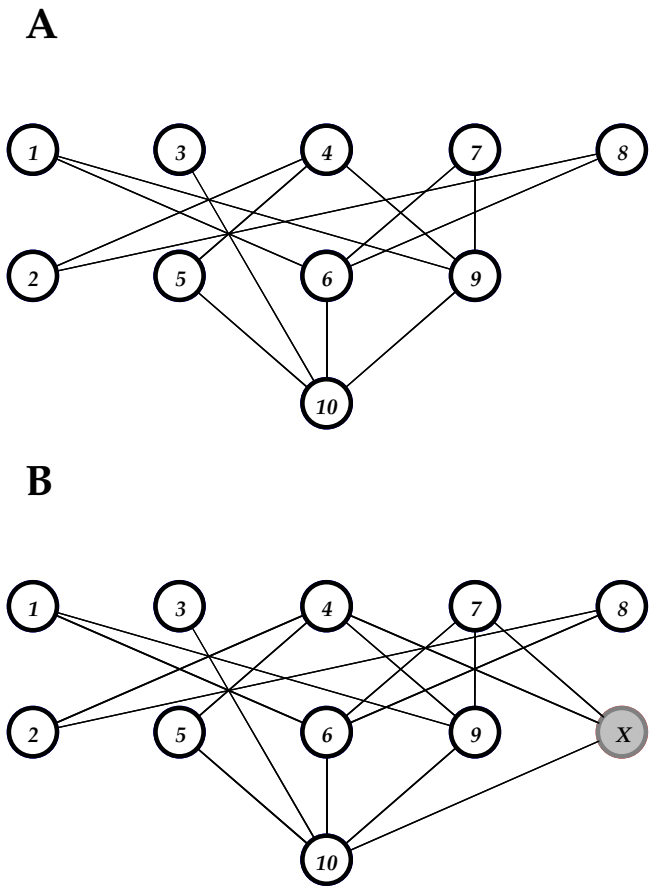


Figure 1. Illustrative Hasse diagram of A: 10 compounds using 3 descriptors and B. the same 10 compounds plus 1 new compound X.

Further the partial order ranking methodology can be use to evaluate new compounds. This may be a new compound planned to be introduced in a certain production or a compound that has been introduced in the production in order to reduce, e.g., the environmental impact. Adopting the above discussed 10 compounds and the corresponding Hasse diagram (Figure 1A) we introduced a new compound X, the corresponding Hasse diagram being visualized in Figure 1B. It is immediately noted that compound X is evaluated as less environmentally harmful than compounds 4 and 7, but more harmful than compound 10. In other words, if appears environmentally advantageous if compounds 4 or 7 could be substituted by compound X, whereas a substitution of compound 10 with compound X from an environmentally point of view should not take place. Thus, through the partial order ranking the new compound, X, has obtain an identity in the scenario with regard to its potential environmental impact.

As mentioned real scenarios will often include a higher number of compounds. As an illustrative example 50 arbitrarily chosen potential PBT substances have been studied, 9 of these being high production volume chemicals, the remaining 41 being medium production volume chemicals (Carlsen & Walker, 2003). In Figure 2the Hasse diagram corresponding to these 50 compounds based on the BioWin

descriptors BDP2 and BDP3 as well as the bioconcentration factor, BCF, as derived by BCFWin (EPI, 2000) is displayed.

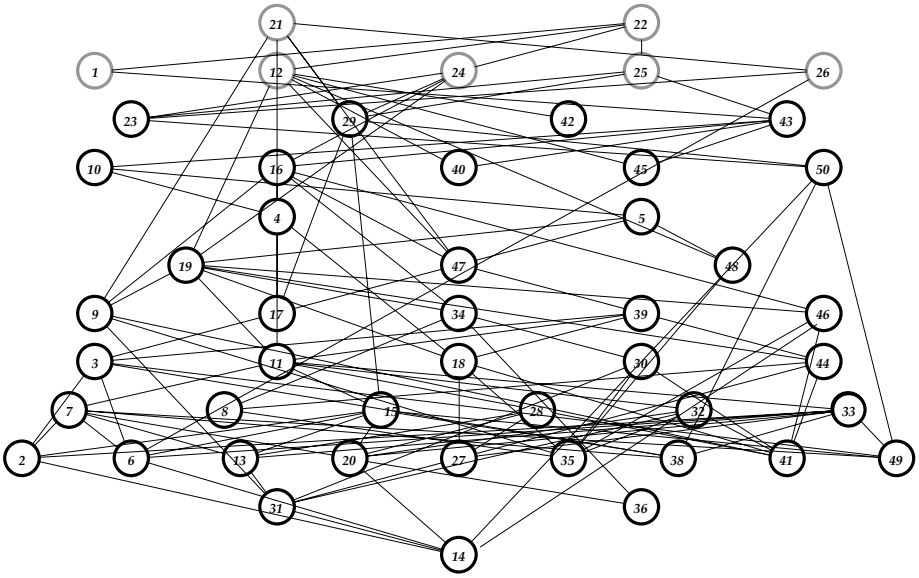


Figure 2. Hasse diagram of 50 arbitrarily chosen potential PBT substances ranked according to their biodegradation and bioaccumulation potentials

The complexity of this ranking is immediately noted. However, it should in this context be taken into account that in the present study the descriptor values applied were used as derived from the QSAR calculations. In some cases the descriptor values vary only slightly among the single compounds. However, the partial order ranking is a purely ordinal method so any differences in the descriptor values are taken as significant. If this for some studies turns out to be a problem, this may be remedied by grouping/classification of substances within certain descriptor value ranges ranking (Walker & Carlsen, 2002). Especially in cases whereas large numbers of compounds are evaluated the grouping appears appropriate. Thus, currently 2773 compound on the US EPA inventories are investigated and a preliminary ranking based on descriptor value ranges is made whereby the compounds are grouped in specific 'events' such as those exhibiting biodegradation half lives > 6 months, BCFs > 5000 and acute baseline toxicity < 1 mg/L (Carlsen et al., 200X). Subsequently, the compounds found in the single events may be ranked based on the individual specific descriptor values.

In the above example (Figure 2) the diagram nevertheless enables us to verify the environmentally most harmful compounds based on their persistence and bioaccumulation. In the present example the compounds in the two upper levels, i.e., level 1: compounds 21 and 22 and level 2: compounds 1, 12, 24, 25 and 26, represents the 10% of the compounds that must be regarded as the environmentally most hazardous.

Obviously, the use of other descriptor combinations, i.e. other measures for biodegradation as well as toxicity descriptors, will lead to different results. However, overall the same trend is observed, i.e.,

that virtually the same compounds constitute the "top-10%". Thus, compounds 21 and 22 appear at level 1 in all cases, compounds 24, 25 an 26 appear typically at level 2, inclusion of toxicity descriptors brings compound 26 to level 1 and further, by inclusion of toxicity descriptors, a few "new" compounds are brought into the "Top-10%".

As mentioned previously the Hasse diagrams are typically characterized to the presence of a number of comparisons. The actual number of incomparisons is roughly speaking a result of interplay between the number of compounds and the number of descriptors (Sørensen et al., 2000). Thus, increasing the number of descriptors will, for the same number of compounds, increase the number of incomparisons.

A priori the incomparisons may turn out as an Achilles' heel of the partial order ranking method. However, the adoption of the linear extension approach apparently remedies this, at least to a certain extent.

Turning back to the model diagram (Figure 1B) it can be noted that e.g. the compounds 4 and 7 are incomparable, i.e. looking just for these two compounds it cannot from the Hasse diagram be concluded which of them are the more hazardous. However, bringing the linear extensions into play gives us the probability for these two compounds to have a certain absolute rank. In Figure 3A the probability distribution for the compounds 4 and 7 for the possible absolute ranks is visualized. It is easily seen that the probability for finding compound 4 at rank 1 or 2 are higher than for compound 7 (Rank 1is equal to top rank). On the other hand, compound 7 are more probable to be found at rank 4-7 than compound 4. On this basis we can conclude that comparing compounds 4 and 7, the most probable absolute ranking will place compound 4 above compound 7. In Figure 3B the probability distribution for compound 10 is shown. The probabilities of finding compound 10 at rank 11 are approx. 70% and at rank 10 approx. 30%. The incomparability between compounds 10 and 2 accounts for this since compound 2 has an approx. 30% probability to be occupy rank 11.

The 'new' compound, X, introduced in the diagram displayed in Figure 1B apparently is comparable only with compound 4, 7 and 10 and thus incomparable with the remaining 7 compounds in the scenario. The high number of incomparisons immediately indicates the presence of a relative broad probability distribution for compound X. This is nicely demonstrated in Figure 4 displaying the probability distribution of compound X for being found at specific absolute ranks.

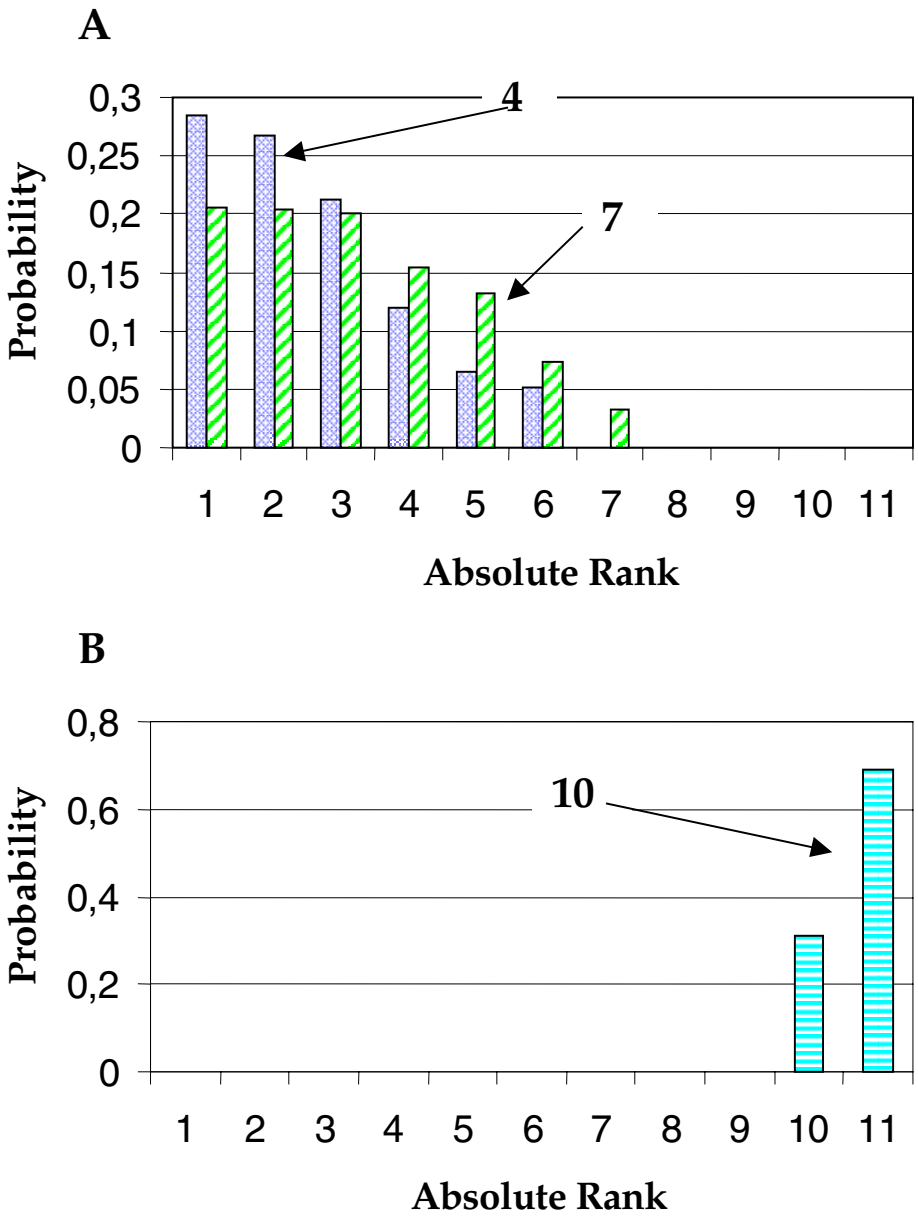


Figure 3. Probability distribution of A: compounds 4 and 7 and B: compound 10 to occupy specific absolute ranks (rank 1 and 11 is top and bottom rank respectively).

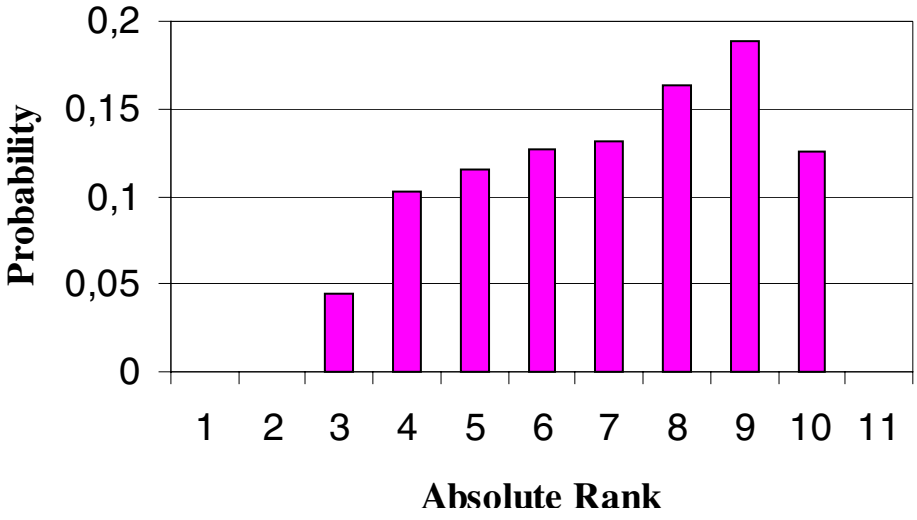


Figure 4. Probability distribution of compound X to occupy specific absolute ranks.

The probability distribution of compound X in relation to compounds 4, 7, 10 and X is visualized in Figure 5. It must in this connection be remembered that although the probability distribution of compound X overlaps those of compounds 4, 7 and 10, compound X must be located between compounds 4 and 7 and compound 10 (cf. Figure 1B).

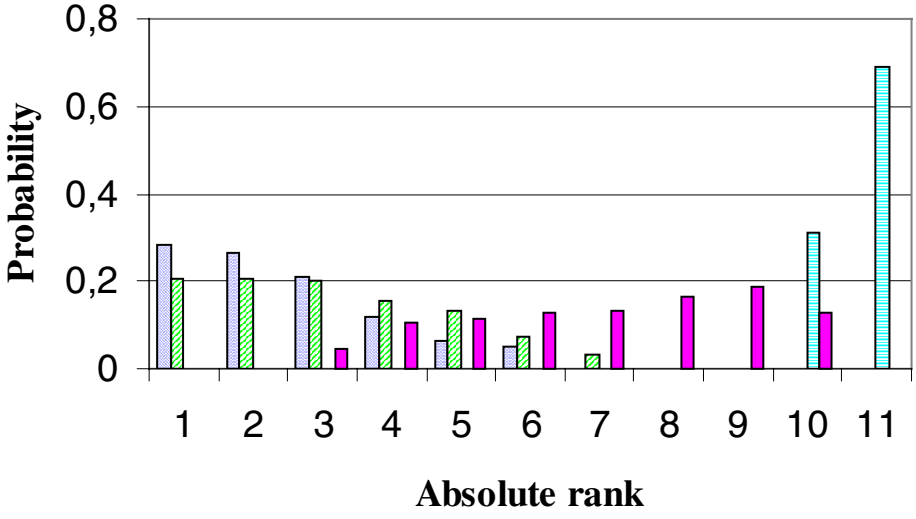


Figure 5. Probability distribution of compound X in relation to compounds 4, 7 and 10 to occupy specific absolute ranks.

Now turning back to the 50 arbitrarily chosen PBT compounds (Figure 2). In order further to illustrate the use of linear extensions, we looked at the three incomparable compounds 24, 25 and 26, which are all located in level 2. In Figure 6 the probability distributions for these three compounds to occupy specific absolute ranks are displayed.

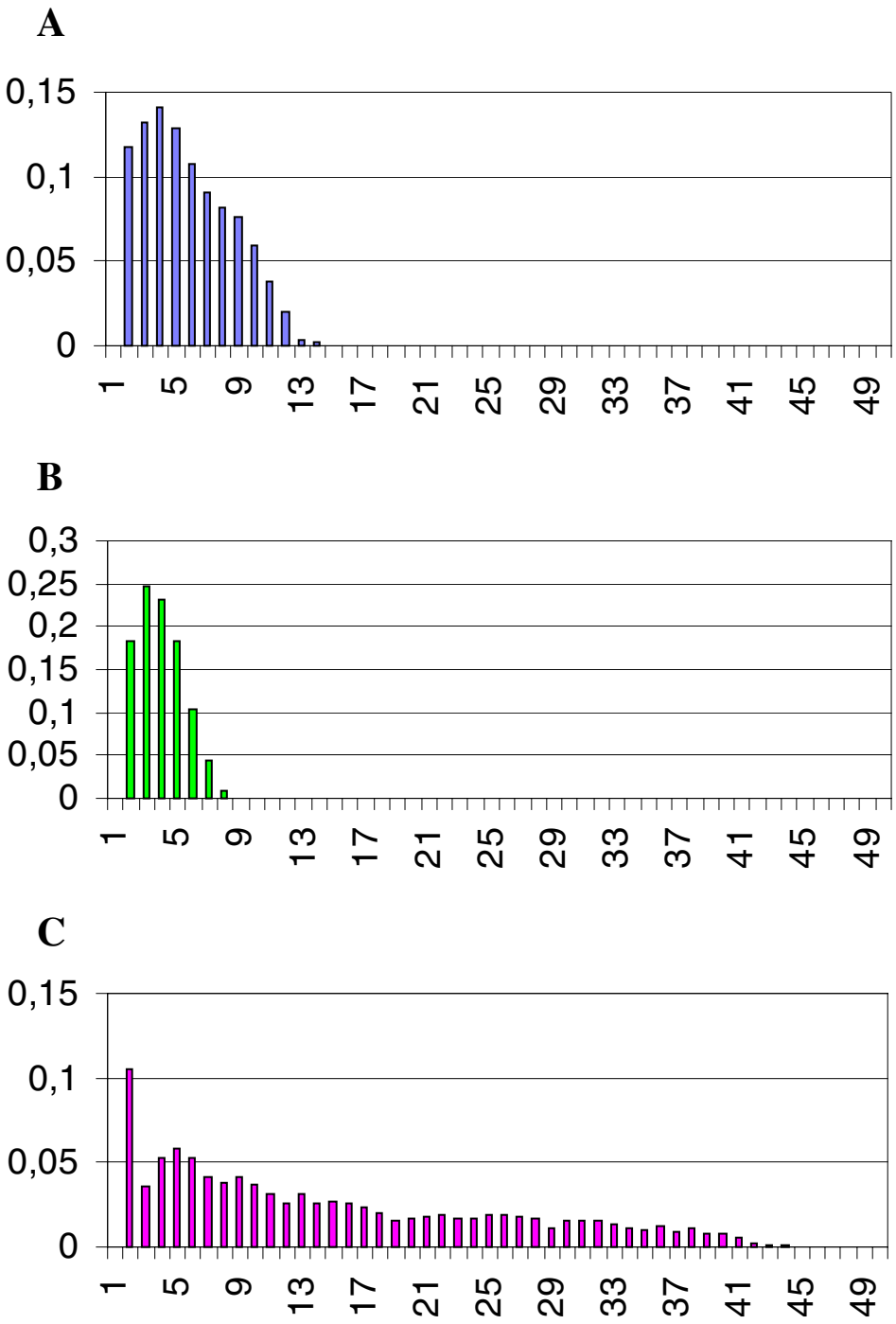


Figure 6. Probability distributions of A: compound 24, B: compound 25 and C: compound 26 to occupy specific absolute ranks.

From the figures it can easily be calculated that their is a probability of 95% to find compound 24 at ranks 2 to 13, compound 25 at ranks 2-8 and compound 26 at ranks 2-38, respectively. Thus, the mutual ranking of these three compounds most probably will be 25 > 24 > 26. A closer look at the diagram depicted in Figure 2 unambiguously discloses that the broad probability distribution found in the case of compound 26 can be ascribed to the fact that this compound are comparable only to relative few other compounds.

3 Conclusions

The present study has demonstrated that chemical substances, such as PBT substances, can be prioritized or ranked using a partial order ranking technique. Obviously a simple "yes/no" classification would be achievable based on QSARs alone with reference to selected PBT criteria. However, partial order ranking in combination with the use of linear extensions apparently provide addition valuable information with regard to which substances on a cumulative basis are the environmentally more hazardous taking into account the influence of several descriptors such as e.g. persistence, bioaccumulation and toxicity.

Assuming that we are dealing with meaningful data material the partial order ranking methodology offers several advantages as a decision support tool for prioritizing chemical substances. The method is robust and transparent, it takes into account several decision criteria simultaneously, and it offers an immediate selection of environmentally more problematic compounds. Further, the use of linear extensions leads to probabilities for absolute ranks.

The limitations of the method comprise the fact that the ranking typically is incomplete and that the method is sensitive to inversely correlated data.

To summarize we conclude that the combination of QSAR modeling and partial order ranking constitutes an effective decision support tool that could be used to facilitate regulatory actions.

References

AQUIRE, http://epa.gov/med/databases.html#aquire

Brüggemann, R. Halfon, E. & Bücherl, C., 1995. Theoretical base of the program "Hasse", GSF-Bericht 20/95, Neuherberg. The software may be obtained by contacting Dr. R. Brüggemann, Institute of Freshwater Ecology and Inland Fisheries, Berlin.

Brüggemann, R., Pudenz, S., Carlsen, L., Sørensen, P.B., Thomsen, M. & Mishra R.K., 2001a. The use of Hasse diagrams as a potential approach for inverse QSAR, *SAR QSAR Environ. Res.*, 11, 473-487 (2001).

Brüggemann, R., Halfon, E., Welzl, G., Voigt, K. & Steinberg, C.E.W., 2001b. Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests, *J. Chem. Inf. Comput. Sci*. 41, 918-925.

Carlsen, L., Sørensen, P.B. & Thomsen, M., 2001. Partial order ranking based QSAR's: Estimation of solubilities and octanol-water partitioning, *Chemosphere*, 43, 295-302.

Carlsen, L., Sørensen, P.B., Thomsen, M. & Brüggemann, R., 2002. QSAR´s Based on Partial Order Ranking, *SAR and QSAR Environ. Res*, 13, 153-165.

Carlsen, L. & Walker, J.D. 2003. QSARs for Prioritizing PBT Substances to Promote Pollution Prevention, *Quant.Struct.-Activ.Relat.*, in press

Carlsen, L., Walker, J.D., Mekenyan, O.G. & Russom, C.L. 2003. Using the Hasse Diagram Technique to Prioritize Potential PBTs. QSAR Comb. Sci. 22, 49-57.

Davey, B.A. & Priestley, H.A. 1990. Introduction to lattices and Order. Cambridge University Press, Cambridge, 1990.

EINECS, 1967. EINECS (European Inventory of Existing Commercial Chemical Substances). cf. European Commission 1967: Directive 67/548/EEC on the application of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances and the 6th amendment: Directive 79/831/EEC; art. 13.

EPI, 2000. Pollution Prevention (P2) Framework, EPA-758-B-00-001; may be obtained through the link 'P2 Manual 6-00.pdf' found at http://www.epa.gov/pbt/framwork.htm, US EPA.

Fishburn, P.C. 1974. On the family of linear extensions of a partial order, *J.Combinat.Theory*, **17**, 240-243.

Graham, R.L. 1982. Linear Extensions of Partial Orders and the FKG Inequality. in: Ordered Sets, I Rival (ed.), pp. 213-236

Halfon, E. & Reggiani, M.G. 1986. On the ranking of chemicals for environmental hazard, *Environ. Sci. Technol.*, 20, 1173-1179.

Hasse, H. 1952. Über die Klassenzahl abelscher Zahlkörper, Akademie Verlag, Berlin.

Niëmela, J. 1994. Working document on the availability of data for classification and labelling of chemical substances at the European market.

Lerche, D., Brüggemann, R., Sørensen, P., Carlsen, L. & Nielsen, O.J. 2002a. A Comparison of Partial Order Technique with Three Methods of Multi-Criteria Analysis Techniques for Ranking of Chemical Substances, J.Chem.Inf.Comput.Sci., 42, 1086-1098.

Lerche, D Sørensen P., Brüggemann R. 2002b. An attempt to derive a general model by partial order theory, Part I. Improved Estimation of the Ranking Probabilities in Partial Orders using Random Linear Extensions, Submitted to J.Chem.Inf.Comput.Sci. in press

RTECS, http://www.nisc.com/factsheets/rtecs.htm

Sørensen, P.B., Mogensen, B.B., Carlsen, L. & Thomsen, M. 2000. The influence of partial order ranking from input parameter uncertainty. Definition of a robustness parameter *Chemosphere*, 41, 595-601.

Sørensen, P.B., Lerche, D.B., Carlsen, L. & Brüggemann, R. 2001. Statistically approach for estimating the total set of linear orders. A possible way for analysing larger partial order sets. in: Order Theoretical Tools in Environmental Science and Decision Systems, R. Brüggemann, S. Pudenz and H.-P. Lühr, eds, Berichte des IGB, Leibniz-Institut of Freshwater Ecology and Inland Fisheries, Berlin, Heft 14, Sonderheft IV, pp. 87-97

Sørensen, P.B. & Lerche, D.B. 2002. Quantification of the uncertainty related to the use of a limited number of random linear extensions, in: K. Voigt and G. Welzl, eds., 'Order Theoretical Tools in Environmental Sciences. Oreder Theory (Hasse Diagram Technique) Meets Multivariate Statisrics, Shaker Verlag, Aachen, pp. 65-72

Walker, J.D., Carlsen, L., Hulzebos, E & Simon-Hettich, B. 2002. Government Applications of Analogues, SARs and QSARs to Predict Aquatic Toxicity, Chemical or Physical Properties, Environmental Fate Parameters and Health Effects of Organic Chemicals. *SAR QSAR Environ.Res*., 13, 607-619

Walker, J.D. & Carlsen, L. 2002. QSARs for identifying and prioritizing substances with persistence and bioaccumulation potential, *SAR QSAR Environ.Res*. 13, 713-726

Winkler, P.M. 1982. Average height in a partially ordered set. *Discrete Mathematics.* 39, 337-341.

Winkler, P.M. 1983. Correlation among partial orders. *Siam J Alg Disc Meth.* 4, 1-7.

Probability approach applied for prioritisation using multiple criteria

Cases: Pesticides and GIS

Peter B. Sørensen, Steen Gyldenkærne, Dorte Lerche, Rainer Brüggemann, Marianne Thomsen, Patrik Fauser and Betty B. Mogensen

National Environmental Research Institute (NERI),
DK-4000 Roskilde,
Denmark
E-mail: pbs@dmu.dk

Abstract

A paradigm of using a probability concept of ranking is applied and discussed using the concept of maximal entropy. The principle of the raking is based on the set of linear extensions formed from a partial order set. An existing method using a linear extension analysis is applied and combined with a measure for the entropy level as defined in the information science. It is shown how the entropy level reflects the degree of uncertainty in the probability space of possible ranking values. Finally the concept of ranking probability is applied in GIS to predict the ranking of environmental impact from the use of pesticides in relation to the surface waters. In this way it is possible to identify the positions in the landscape where the environmental impact is highest due to the use of pesticides.

1 Introduction

Prioritisation among alternatives (objects) is a general problem in environmental management. Examples of specific prioritisation problems could be: (1) Prioritisation of soil waste sites based on environmental and health data in order to identify the best soil remediation strategy. (2) Ranking of chemicals in relation to the potential environmental risk. (3) Identification of the position in a geographical information system where the possible environmental impact due to pesticide usage will be at the highest level. Two problems arise in such a ranking analysis: (1) How to make a manageable ranking procedure which can perform a sufficient precise and valid description

of the phenomenon of concern and (2) how to deal with the uncertainty associated with the necessary input parameters (descriptors) for the ranking procedure. The topic of this paper will be the design and application of a ranking procedure, leaving the uncertainty aspect with respect to the input data for other investigations. In the area of environmental management the problems will typically be rather complex where multiple factors are going to be taken into account simultaneously. So, dealing with higher complexity more than one criterion is often needed to provide a sufficient valid description and this paper deals with this type of multiple criteria ranking. The conventional ways for ranking of objects is to use a model, which can yield a single ranking number for each object. The outcome of the model is then an exact rank of each object in relation to all the other objects. Such a model for exact ranking can be more or less complicated. If the model is judged to be sufficiently valid then this method will solve the ranking problem. However, in some cases dealing with high complexity it can be difficult to suggest a model of sufficiently known validity. One could be tempted to use a more or less doubtful scoring system and add the scoring of each descriptor to form a single number. The actual validity of such a system can easily be too unclear and difficult to quantify on a scientific basis.

The problem of multi criteria ranking is illustrated in a simple example in Figure 1 where two criteria are applied in terms of two descriptor values for each object to be ranked. The four objects are denoted $X_{1\text{-}4}$ and the descriptor values are shown in the table in Fig 1. In the same table two different models for exact ranking are shown each in one column and based on respectively addition and multiplication of the two descriptors and the resulting values are shown in the table. The resulting ranking from these two models are different because $X_1>X_2$ using addition while $X_1<X_2$ using multiplication. As an alternative, a partial order can be applied where no model is applied to form an exact ranking. In the partial order a pair of objects is only ranked when no contradiction exists among the ranking of the single descriptors (Davey and Priestley, 1990). So, the ranking between X_3 and X_1 is included in the partial order ranking because both 7>2 and 7>5 is true yielding the rank $X_3>X_1$. Contrary, the ranking between X_1 and X_2 is not included in the set of rankings because 2<3 and 7>5, which makes the ordering of the objects X_1 and X_2 indefinite. The partial order is graphically mapped in the Hasse diagram, where connecting lines are drawn between objects between which there exist determined orders (Hasse, 1952).

The indefiniteness in the ranking using partial order forms the basis of ranking probability. This is illustrated in Figure 1, where three possible exact rankings are seen to exist all being in agreement with the partial order (Hasse diagram). Each of these exact rankings is denoted a linear extension and any exact ranking model using all the descriptors will make a rank similar to one of the linear extensions. In this way the exact ranking model using addition is seen to reproduce one of the linear extensions while the model based on multiplication identifies another linear extension. A third linear extension is seen to exist, which is identified neither by the simple addition nor by the simple multiplication of the descriptors but other exact ranking models can reproduce this third linear extension.

Objects	Descriptors		Addition	Multiplication
X_1	2	7	9	14
X_2	3	5	8	15
X_3	7	7	14	49
X_4	3	4	7	12

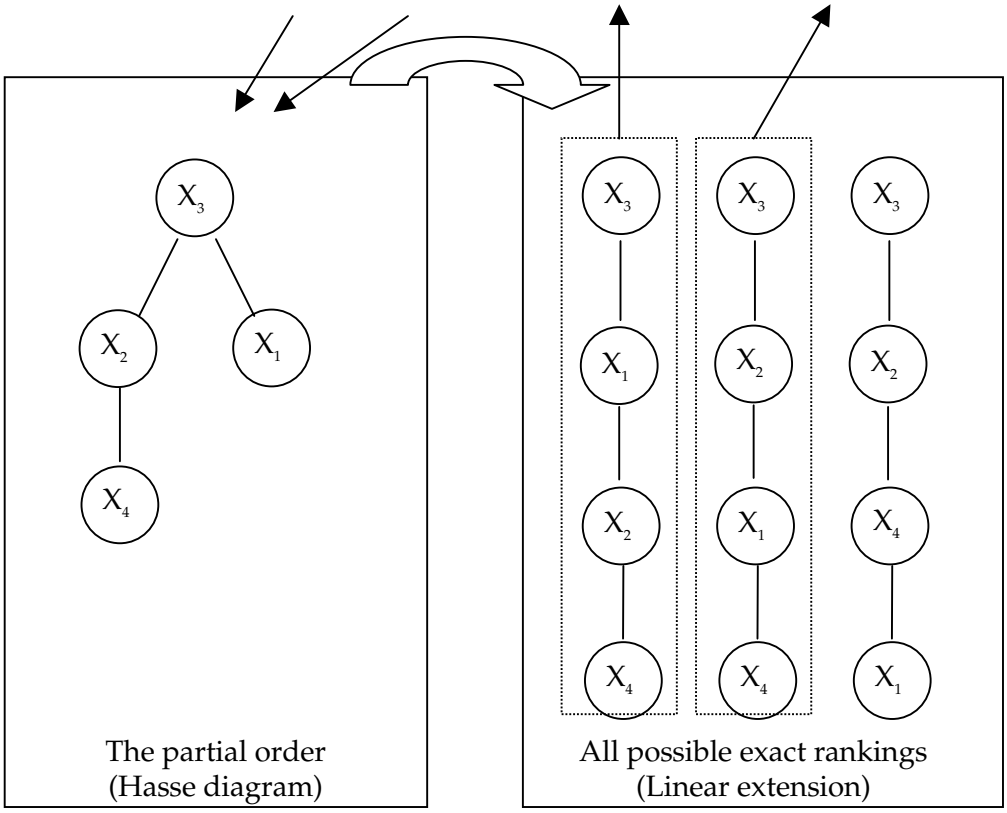


Figure 1. A simple example for illustration of the relationship between the partial order and models for exact ranking, here in terms of addition and multiplication of the descriptors.

In this paper a situation is considered where no model for exact ranking can be identified and where no ranking between two objects a priori can be claimed to be more relevant (important) than another ranking. So, in this way in Figure 1 none of the rankings $X_3>X_2$, $X_3>X_1$ nor $X_3>X_4$ will be considered a priori as more relevant than the others. But, the fact that: $X_3>X_2$ and $X_2>X_4$ in the partial order yields a higher resulting weight to the ranking of $X_3>X_4$ compared to the ranking $X_3>X_2$. Under these circumstances, none of the linear extensions can be said to be most likely and random selections of linear extensions can thus create a probability space for a specific object to be ranked at a specific level. In this way it can be seen for object X_2 that the most likely rank is at level 3 (next highest) as this is the rank in two out of 3 linear extensions. So, more precisely the object X_2 is ranked at level 3 by a probability of 2/3 and ranked at level 2 by a probability of 1/3. This principle is described by e.g. Winkler, 1982, and Trotter, 1992.

A more general illustration is seen in Figure 2 for at series of N objects where each has I different descriptors. In such cases the partial order theory can be a better alternative for solving the ranking problem as illustrated in Figure 1 (Lerche et al., 2002). But, there is a price

applying the partial order where the total ranking level for an object comes out in form of a probability distribution for a series of possible ranking levels as illustrated in Figure 2. The ranking probability makes the result more fuzzy and thus less useful compared to an exact rank. However, the probability distribution yields a direct measure for the uncertainty related to the missing quantified interrelation between the different descriptors.

The ranking probability formed by the partial ordering yields a probability distribution for all possible ranking values of each object. The probability distribution thus represents in Bayesian terms the a priori knowledge related to the ordering using the descriptors alone without knowing a valid model for exact ranking. The use of the ranking probability may also be regarded as a maximum entropy (maximal variability) estimate of the ranking variability purely as a result of the selected descriptors where no exact ranking model is assumed. When it is possible to make a useful conclusion based on this ranking probability the validity is relatively strong because the uncertainty related to any assumed inter-relationship between the different descriptors is avoided.

The concept of entropy in partial order theory has been discussed by several references (Dhar, 1978, Dhar, 1980 and Brightwell et al., 1996). The findings in these references lead to the consideration of partial orders as a real gas, where the phase transition is considered as a transition from an unordered state having a higher degree of indefiniteness in the ranking towards partial orders having a higher degree of determined rankings. The phase transition of partial order is now an item of intensive mathematical research (Proemel, personal communication).

For discrete systems an equation for entropy (E) for each object can be defined (Berger, 1985) as

$$E = -\sum_{i=1}^{n} p_i \cdot \log(p_i) \qquad (1)$$

where n is the number of possible alternatives and p_i is the probability for alternative i to be true. The alternatives in our context are the different rankings so n is equal to the number of objects. If no ordering is realised in a partial order then the probability for a given object to be placed at a given position in the linear extension is $\frac{1}{n}$ and Eq. 1 becomes:

$$E_{no\ \inf} = -\sum_{i=1}^{n} \frac{1}{n} \cdot \log\left(\frac{1}{n}\right) = \log(n) \qquad (2)$$

This relation sets up the upper limit of entropy equivalent to "no information about ranking". The lowest possible entropy value is zero and comes out when an exact rank exists for an object as the p value in this case will be unity for the true rank and all other p values will be zero. So the interval of E is closed between zero and $\log(n)$.

Object	Descriptor							
	$x_{.,1}$	$x_{.,2}$	-	-	$x_{.,i}$	-	-	$x_{.,I}$
X_1	$x_{1,1}$	$x_{1,2}$	-	-	$x_{1,i}$	-	-	$x_{1,I}$
X_2	$x_{2,1}$	$x_{2,2}$	-	-	$x_{2,i}$	-	-	$x_{2,I}$
X_3	$x_{3,1}$	$x_{3,2}$	-	-	$x_{3,i}$	-	-	$x_{3,I}$
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
X_n	$x_{n,1}$	$x_{n,2}$	-	-	$x_{n,i}$	-	-	$x_{n,I}$
-			-	-		-	-	
-			-	-		-	-	
X_N	$x_{N,1}$	$x_{N,2}$	-	-	$x_{N,i}$	-	-	$x_{N,I}$

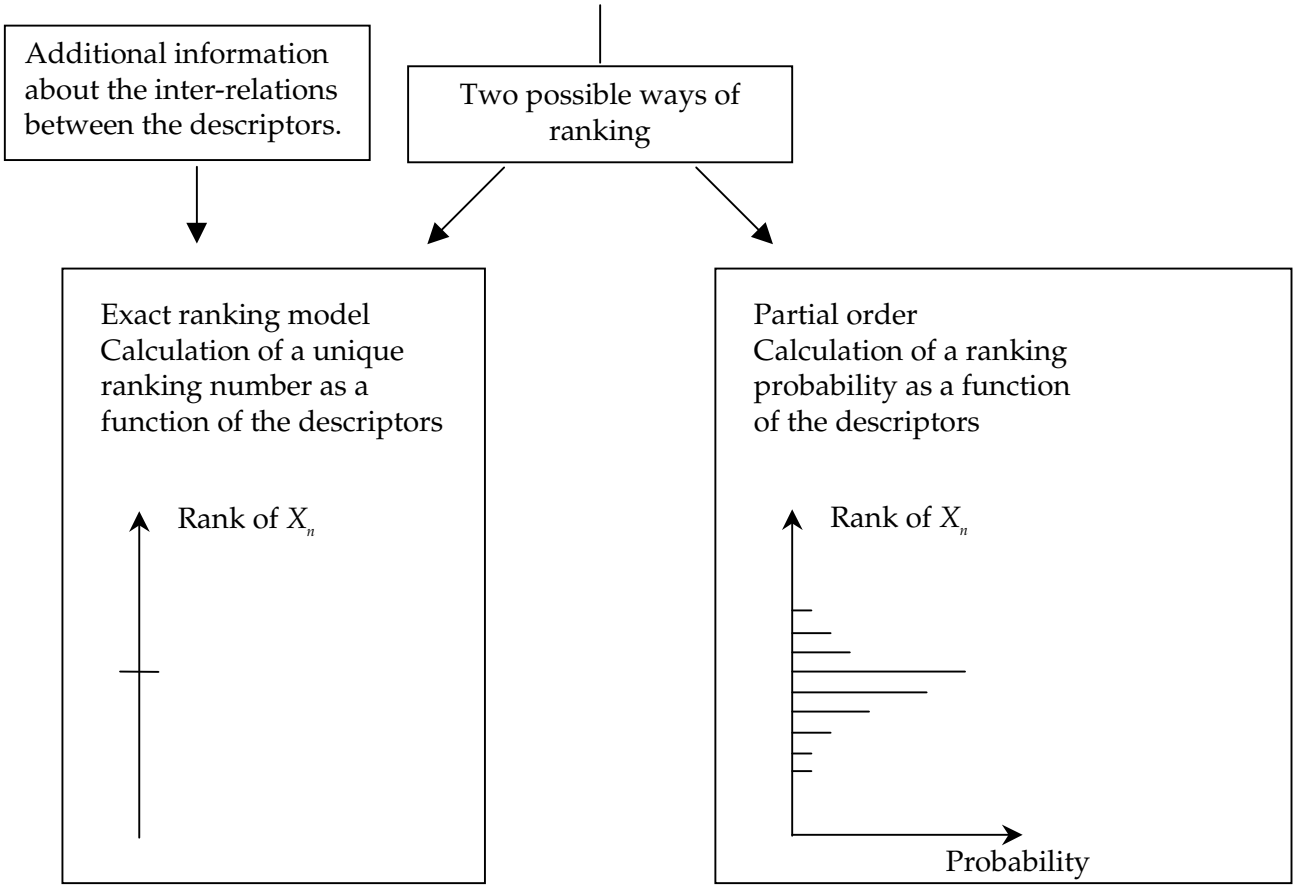


Figure 2. The principle of ranking using either a ranking model or a partial order.

In recent years new methods have been developed to generate ranking probabilities for prioritisation of up to a few thousand objects depending on the computer power available and this opens up for a wide use of the method of partial ordering (Sørensen et al., 2001 and Lerche and Sørensen, 2003). The concept of ranking probability is the fundament in the paradigm in this paper and we will introduce the use of this concept for practical purposes in cases where the ordering can be formulated as an event space and used for a high number of objects in GIS.

2 The event space

In this investigation the approach of an event space is applied. In the event space a continuos value scale of each descriptor is transformed to a finite interval system. The finite interval scale used in this investigation is: low, medium, high leaving three possible values for every descriptor. The event space is made using three descriptors. This makes 27 possible ways of combining the three descriptors and the three different levels for each object. The ordering relations are shown in Figure 3 using the Whasse software (Brüggemann et al., 1999).

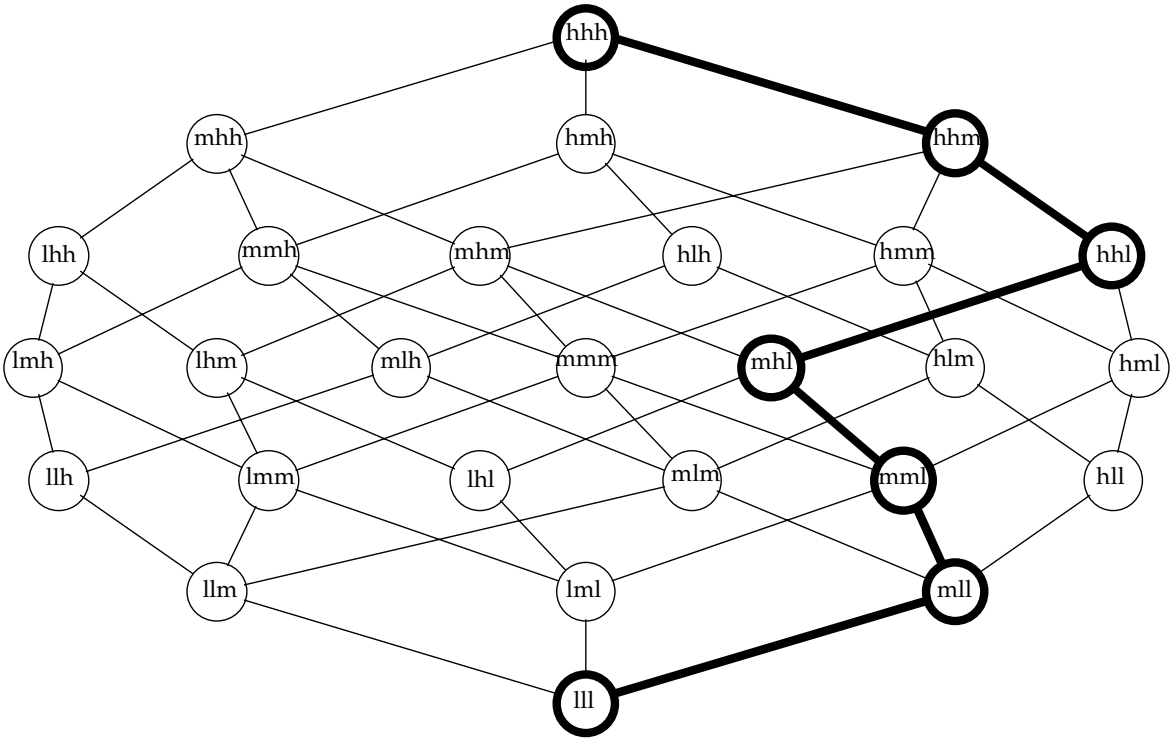


Figure 3. The event space using three descriptors, where every descriptor can take the values High (h), Medium (m) or Low (l).

The descriptor values are shown in Figure 3 in a way so e.g. the name hlm means high level for the first, low level for the second and medium level for the last descriptor etc. For illustration of the principle in the Hasse diagram a series of ranked elements is shown by thick lines as hhh>hhm>hhl>mhl>mml>mll>lll. However, some indefiniteness is also identified in relation to ranking due to incomparable objects. An example of a incomparable pair of objects is mmh and mhm, where the rank due to the single descriptors are in conflict and thus no ranking can be made between these two objects. This forms a "window" of possible rankings for each object and can be considered as a window of uncertainty in relation to ranking the object in a total order where all object are compared to each other. In this way the object mmh is compared with 11 objects below and 3 objects above thus 14 objects are compared to mmh and the remaining 12 objects are incomparable with mmh. The window of uncertainty for mmh

allows a rank of mmh starting from level 12 (11 objects ranked below) and stopping at level 25 always having 3 objects ranked above. However, it seems not so likely for mmh to be placed at level 12 even it is possible, because all the incomparable objects then need to be placed above mmh simultaneously. A position in the middle of the ranking window around rank 18 seems more likely to be true. This likeliness is expressed by the ranking probability for being situated at a given rank.

3 Estimation of ranking probability

A large number of linear extensions exist where every of them coincide with all the ordering done in Figure 3., e.g. in Figure 3 the relation mhh>lhh exists and thus mhh>lhh will also exist in every possible linear extension. The whole set of linear extensions generates a probability space (i.e. fulfils the axioms of probability, according to Kolmogoroff) and can be used to estimate the ranking probability (Winkler, P., 1982). The principle of using the linear extensions to find the probability is shown in Figure 4. In all the linear extensions the object hhh will be in the top and lll will be in the bottom because all objects are compared to these two objects as respectively being below and above.

A finite number of different linear extensions related to the partial order in Figure 3 exist but the number is huge. In Figure 1 only three linear extensions exist but simple combinatorial considerations tell that the maximal number of linear extensions is $n!$, where n is the number of objects, so in case of Figure 3: n=27 and $n!\approx10^{28}$. Therefore, it will be impossible in this case for even an extremely fast computer to find all the linear extensions for the partial order in Figure 3 within a sufficient time scale for calculating. In general it is possible to find all linear extensions if the partial order consists of less than 18-20 objects. The solution for this problem is to form random sampling among the possible linear extension (Sørensen et al., 2001) and in this way 50000 linear extensions are identified from Figure 3 and used in the following analysis. For a more detailed description of the method for finding linear extensions randomly see Lerche and Sørensen (2003). The principle of finding the ranking probability is to count the number of linear extensions where a given object is placed at a given position and then divide the number by the total number of identified linear extensions used (in this case 50000). The principle is illustrated in Figure 4.

Two ranking probability distributions are shown for the elements hhl and llh in Figure 5. They are incomparable to each other but the ranking probability separates them in two distinct different ways. The object hhl tends to be ranked higher than the object llh, but the condition hhl<llh seems to be possible in seldom cases. However, it is very important to make clear that the mutual probability for two objects to be ranked above/below each other can not be calculated directly based of the probability distributions shown in Figure 4. The

reason is that two distributions are not necessary independent of each other. The mutual probability between every pair of objects can easily be deducted from the set of linear extensions, but this is not the topic of this paper.

In the analysis using linear extensions no functional relationship is assumed between the single descriptors as discussed above. Thus the probability distribution shown in Figure 5 can be considered as a maximal entropy estimate of the rank and a more detailed discussion of this concept will be given in the following paragraph.

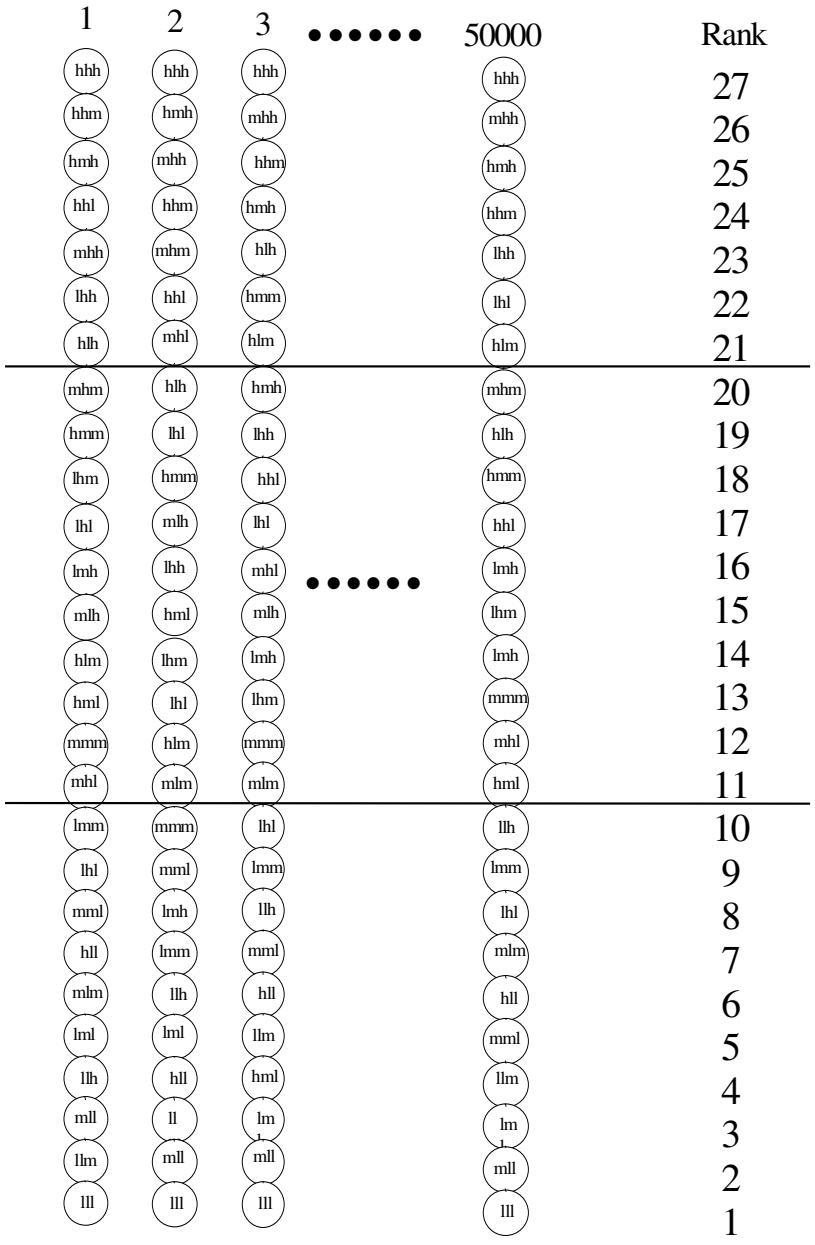


Figure 4 The principle of using the linear extensions to find the ranking probability. The probability for a given object to be at a given rank is determined as the ratio between the number of linear extensions of the object at the given rank and the total number of linear extensions.

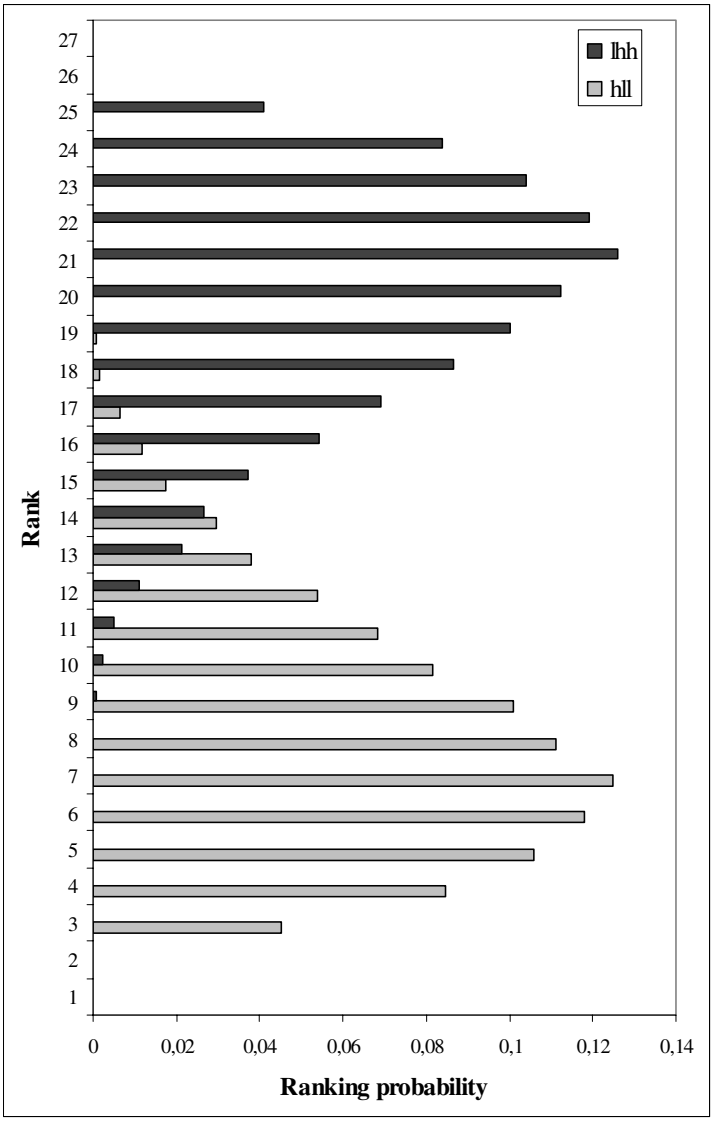


Figure 5. The ranking probability distribution for the two selected objects hhl and llh respectively.

The concept of Eq. 1 can be applied on the ranking probability distribution and the result from such an analysis is shown in Figure 6. The $E_{no\ inf}$ value (Eq. 2) for the partial order in Figure 3 having 27 objects is log (27)=1.43. The entropy level is zero for respectively the objects hhh and lll as the rank for these two objects is exact, because they both are compared to all objects. It seems reasonable that the entropy level is highest for those objects, which are incomparable with many other objects because the ranking uncertainty in this case is high. But also other factors affect the entropy, which will be illustrated in the following lines.

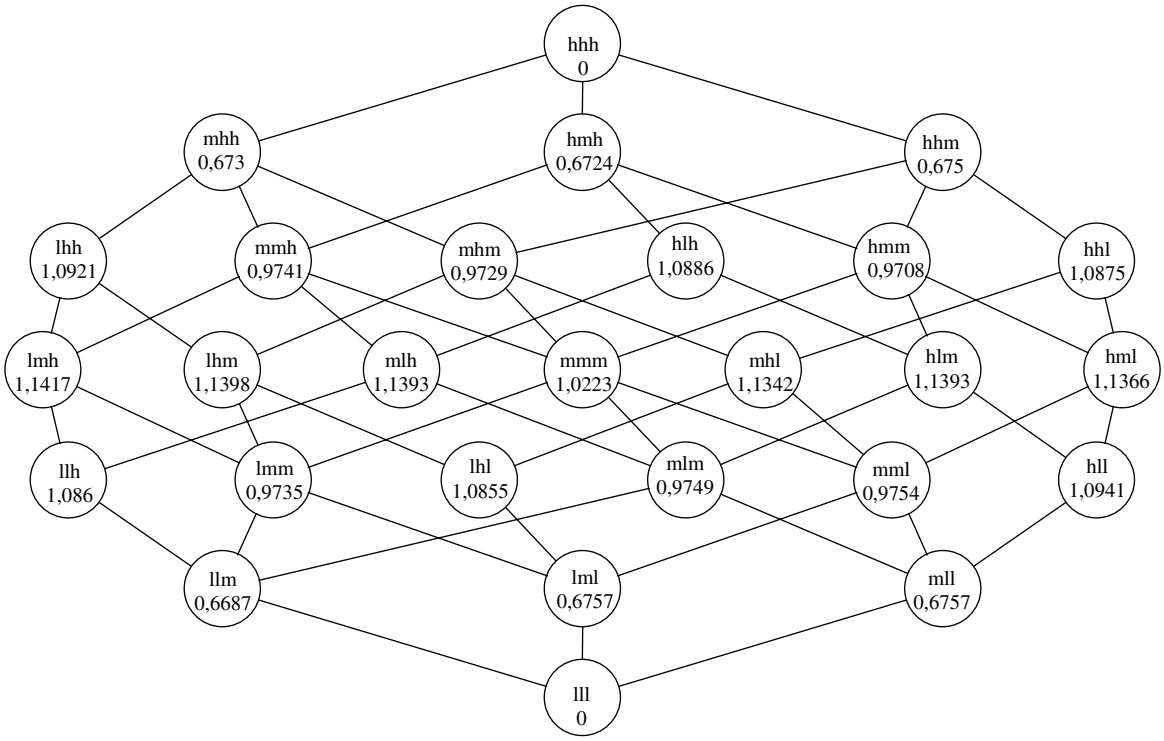


Figure 6. The partial order including the entropy levels for each object, where a higher value is equivalent to higher uncertainty.

The effect on the entropy level is illustrated by comparing the objects mmh and mmm. Both objects are incomparable with 12 other objects, so with respect to this they are equal but in Figure 6 it is seen that the object mmm has higher entropy than object mmh. This means that the ranking of object mmh is associated with smaller uncertainty than the ranking of object mmm. This is actually reasonable when the ranking probability distribution is displayed, see Figure 7. The distribution for mmm is nearly symmetric while the distribution for mmh is tailing to one side and this makes the distribution for mmh more focussed and lesser dispersed than the distribution for mmm. This effect is reflected in the E value.

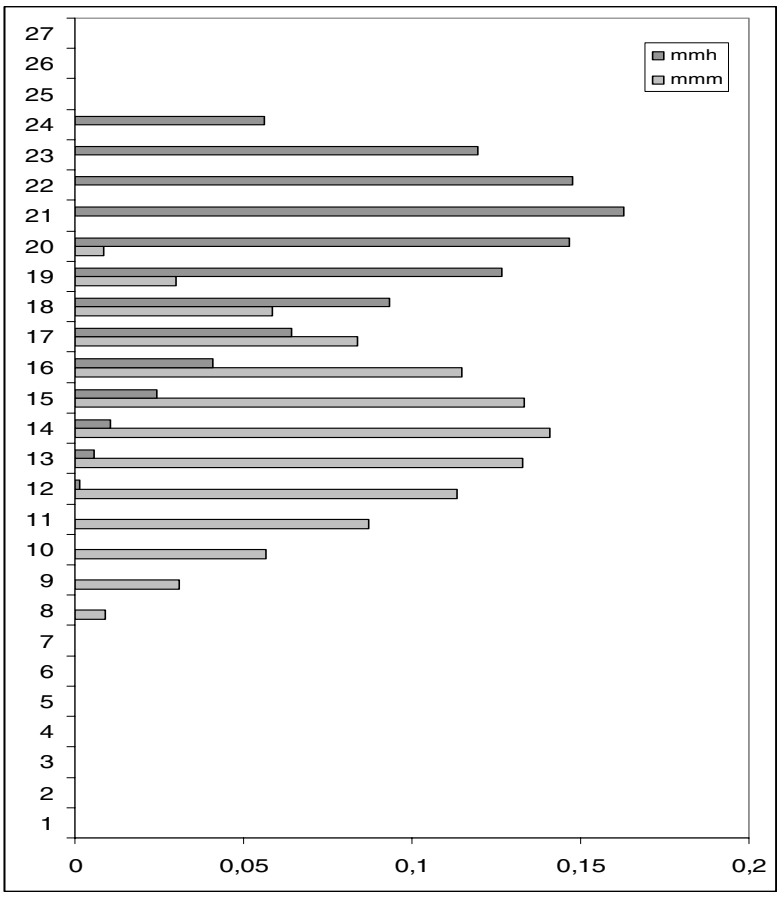


Figure 7. The rank probability distribution for the objects mmm and mmh.

4 Application of the ranking probability in GIS

Spatial prioritisation is a central issue in environmental management. E.g. strategies for environmental monitoring often need a way to predict the most important location to be monitored in order to be sure that the worst possible state is under control In other words a large amount of resources can easily be wasted in monitoring programs yielding unsatisfactory results. This paper applies the principle of ranking probability using maximum entropy estimates on GIS data in a spatial prioritisation having focus on the eco-toxicological impact on surface waters from the usage of pesticides.

Several investigators have used GIS to describe the environmental impact from pesticides (see e.g. Verro et al., 2002 and Tiktak 1999). Typically more or less complicated fate and toxicity models are used in such calculations. In our study the target is the eco-toxicological effect on the surface water living organisms.

4.1 The usage data

The pesticide dosage is calculated in GIS (for Denmark) for area blocks of 5-20 ha each. The procedure is described in more detail by Sørensen et al. 2002. The reference can be ordered from the first author (pbs@dmu.dk). Three data sources are used: (1) A database containing information about agricultural praxis in terms of the given farm size and type, and crop type and area for each blocks in a central record. (2) A database having information about dose levels, which are gathered from actual reported farming praxis. (3) A geographical reference database containing the positions of every area block for use in GIS. The central record about agricultural praxis is obtained from "The General Agricultural Field Register" which is a digital field mapping register. "The General Agricultural Field Register" is based on the "Integrated Administration and Control System (IACS)" by the European Union. The register is available for research purposes in Denmark. The relationship between the dose level and the usage descriptors in form of farming size/type and crop type is established for every active ingredient. The dosage for each active ingredient is calculated as the total use divided by the block area. Often there are several different agricultural fields inside a single block and the amount used of a specific active ingredient is the accumulated mass.

The use of a given pesticide cannot be completely determined as a unique function of the selected usage descriptors. This is because the farmers will perform treatment in relation to the actual local need for pest control and select the active ingredient from a series of alternative commercial products. Further complexity is added from the fact that often more than one active ingredient is sold in the same commercial product. An averaging procedure could be applied but this will induce relatively low dose levels compared to actual dose levels when the pesticides are applied. Therefore, such a procedure is very problematic for an eco-toxicological viewpoint, which is illustrated in Figure 8.

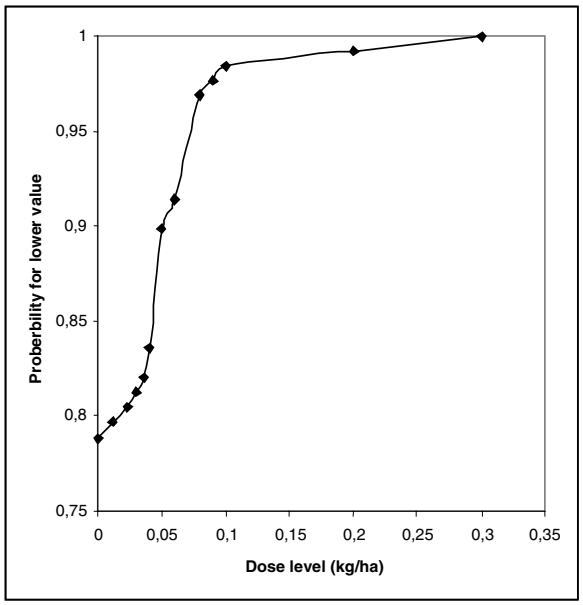


Figure 8. The probability for dose level values of Bromoxynil for same usage descriptor values (same type and size of farming and same crop type).

In Figure 8 it is demonstrated that the eco-toxicological impact can hardly be judged based on mean values of dose as the mean will be much lower compared to realistic applied dose levels because many application is "not used". E.g. for Bromoxynil in Figure 8 the probability for no usage (zero dose) is 0.79. One solution of this problem could be to apply the mean under the condition that the substance is used for all possible fields. This will, however, result in a general dramatic over estimation of the pesticide use leading to unrealistic usage scenarios. Introducing a stochastic approach, where every field application is considered as a whole, solves this problem. Thus, a field application is selected to represent a crop type in the specific block including all the active ingredients involved.

The selection of field application is done randomly among alternatives in the database having information about dose levels. In this way the statistics of the pesticide usage covered by this database will be transfered into the GIS system.

4.2 Eco-toxicological descriptors

The calculated spatial distributed use of pesticides is utilised to form a multi-criteria measure for the potential eco-toxicological impact on organisms in surface water. Three organisms are selected: fish, daphnia and algae. For these organisms there exist rather complete data from legislation of pesticides and data from the work of Møhlenberg et al., (2001). It seems reasonable to claim that the selected organisms cover a rather wide spectrum of endpoints as both vertebrates, invertebrate and plants are included. The potential environmental impact of each area block is calculated in relation to surface water ecotoxicology. An indicator in form of a load index is calculated for each end point in every block using the relationship:

$$Load_{block}^{end\ point} = \sum_{i=1}^{m} \frac{Dose_i}{Toxicity_i} \qquad (3)$$

where m is the number of active ingredients used in agriculture. For a discussion of the indicator see Møhlenberg et al., (2001). Thus, three numbers are calculated in every block one for respectively fish, daphnia and algae. In order to take into account the variability in toxicity testing data values a random selection procedure is applied for selecting toxicity values among different reported values. This procedure is described in Møhlenberg et al., (2001).

4.3 Application of ranking probability in GIS

Every block is assigned to an object of the event space shown in Figure 3. This is done by a ranking of each of the load indexes separately. In this way it was possible to identify the upper one third as “high” the middle one third as “middle” and the lowest one third as “low”. So every block gets an identity as e.g. mlm as shown for the objects displayed in Figure 3.

Linear extensions are then subsequently used to form the ranking probability distributions for every object in the event space. It is now simple to apply a Monte Carlo type algorithm for assigning specific total rank values to a block by random sampling from the ranking probability. The principle can be illustrated by considering object lhh in Figure 5. In most cases when a block is assigned to lhh a ranking level of 18-24 will be assigned to that block, however, in some seldom cases a rank of 11 can be given.

This stochastic procedure yields a pattern, which reflects variability due both to the heterogeneity of the usage and toxicity data and the uncertainty in the multi-criteria ranking when no exact ranking model is applied. The resulting GIS picture is shown in Figure 9.

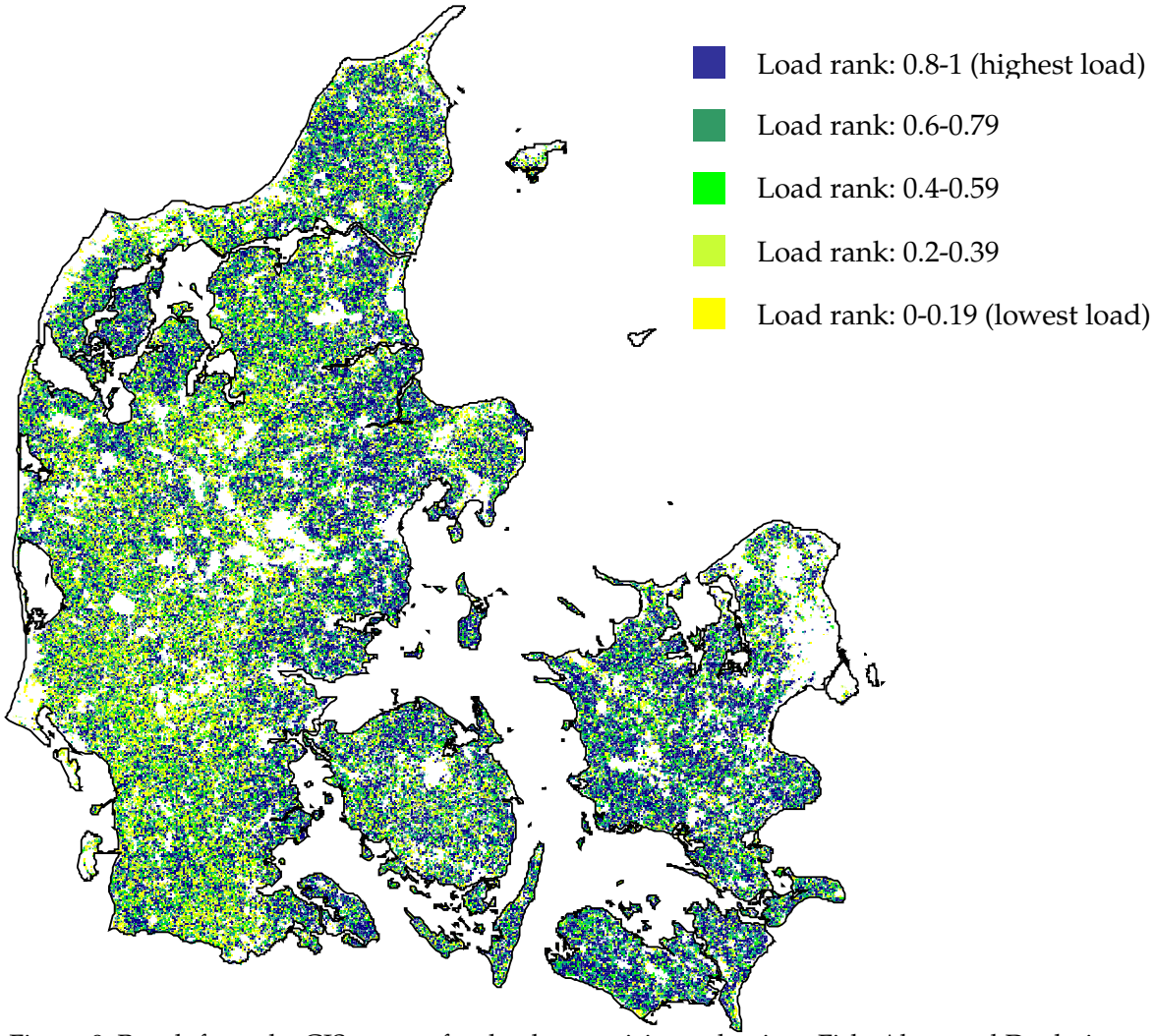


Figure 9. Result from the GIS system for the three toxicity end points: Fish, Algae and Daphnia

5 Conclusion

A new paradigm of applied ranking probability based on partial order technique is presented and described. It is shown how the concept of entropy can be used as a measure for missing information and thus for the uncertainty due to incomplete ranking in the partial order system. A direct application of ranking probability is shown for a GIS problem where the eco-toxicological load is described using a three-parameter characterisation of the load.

References

Berger, O.B. 1985. Statistical Decision Theory and Bayesian Analysis, Springer-Verlang, New York, 1985.

Brightwell, G.R., Prömel, H.J., & Steger, A. 1996. The average number of linear extensions of a partial order. *J.Combin.Theory (A)* 73: 193-206.

Brüggemann, R., Bücherl, C., Pudenz, S. & Steinberg, C.E.W. 1999. Application of the concept of partial order on comparative evaluation of environmental chemicals. *Acta Hydrochim Hydrobiol* 23: 170-178.

Davey, B.A. & Priestley, H.A. 1990. Introduction to lattices and Order. Cambridge University Press, Cambridge, 1990.

Dhar, D. 1978. Entropy and phase transitions in partially ordered sets. *J.Math.Phys.* 19: 1711-1713.

Dhar, D. 1980. Asymptotic enumeration of partially ordered sets. *Pacific Journ.of Mathem.* 90: 299-305.

Hasse, H. 1952. Über die Klassenzahl abelscher Zahlkörper, Akademie Verlag, Berlin.

Lerche, D.B. and Sørensen, P.B. 2003. Evaluation of the ranking probabilities for partial orders based on random linear extensions, accepted by Chemosphere march 2003.

Lerche, D., Brüggemann, R., Sørensen, P.B., Carlsen, L. & Nielsen, O.J. 2002. A comparison of Partial Order Technique with Three Modelling-based Priority Setting Scheme with Partial order Theory for Ranking Chemicals Substances, J. Chem. Inf. Comput. Sci., Vol. 42, pp. 1086-1098.

Møhlenberg, F., Gustavson, K. & Sørensen, P.B. 2001. Pesticide Aquatic Risk Indicators - an examination of the OECD indicators REXTOX, ADSCOR and the Danish indicators FA and LI based on Danish sales data from 1992-2000, OECD report available at: http://www.oecd.org/pdf/M00030000/M00030795.pdf.

Sørensen, P.B., Gyldenkærne, S., Iversen, H.L. & Elmegaard, N. 2002.

Mogensen, B.B., Hansen, H.S. & Schou, J.S. 2002. Spatial prioritisation of environmental impact from pesticides on surface water ecosystems using GIS, SETAC Poster Vienna 2002.

Sørensen, P.B., Lerche, D., Carlsen, L. & Brüggemann, R. 2001. Statistically Approach for Estimating the Total Set of Linear Orders – A possible way for analysing partial ordered sets, (Eds) S. Pudenz, R. Brügemann and H. P. Lühr, Order Theoretical Tools in Environmental Science and Decision Systems, Proceedings of the Third Workshop November 6-7, 2000 Berlin, Germany, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, pp. 222 Heft 14.

Tiktak, A. 1999. Modelling Non-Point Sources Pollutants in Soils. Application to the Leaching and Accumulation of Pesticides and cadmium, Ph.D. Thesis, University of Amsterdam.

Trotter, W.T. 1992. Combinatorics and Partially Ordered Sets, Dimension Theory, The John Hopkins University Press, USA.

Verro, R., Calliera, M., Maffioli, G., Auteri, D., Sala, S., Finizio, A. & Vighi, M. 2002. GIS-Based System for Surface Water Risk Assessment of Agricultural Chemicals. 1. Methodological Approach., Environmental Science and Technology, Vol. 36, 1532-1538.

Winkler, P. 1982. Average height in a partially ordered set, Discrete Mathematics, 39, 1982, 337-341.

A Java-based software for data evaluation and decision support

Stefan Pudenz

Criterion – Evaluation and Information Management,
Mariannenstr. 33,
D-10999 Berlin,
Germany
Email: stefan.pudenz@criteri-on.de

Abstract

The well-known program WHASSE, which is based on visualisation of partial orders, has grown to a considerably scientific tool for data evaluation (Brüggemann and Halfon, 1995; Brüggemann et al., 1999). However, users, which are not as much involved in the theory of partial ordering as many researchers are, may have problems of operation. Moreover, users from e.g. industry or public authorities have specific requirements concerning data evaluation. This means that often only a part of the palette of options that partial ordering yields is required.

Here, a software development with special emphasis on customising is presented. In order to include both independence of operating system of computers and the extensive options of internet such as e-business the computer language JAVA is used for software programming. The modular design of the program supports extensions or substitution of program modules (e.g. user-specific evaluation and decision tools, interfaces to databases). The dynamic download of the modules enables shifting of functional range without compiling the main program. Beside these functional features by means of examples of use the present state of application will be presented.

1 Introduction

Multi-criteria data analysis and decision aid are complex processes and many tools are available based on different approaches. Depending on the problem to be solved, e.g. the decision about the best location of a store in a town, which car to buy or which river section is the most polluted, a specific approach may be preferred. However independent of the approach, the corresponding software has to sat-

isfy some standard requirements. These requirements can roughly be defined as

- transparency and comprehensibility of the evaluation/decision process on the one hand and
- demand for specific interfaces on the other hand.

Transparency and comprehensibility can be further specified: Decisions and in particular multi-criteria data-analysis are often time-consuming because for instance the basic data have to be modified by pre-processing (e.g. classification), then evaluated, again classified and so on. So it is easily imaginable that at the end of the analysis a complex system consisting of different results, tables, diagrams etc. has arisen. All the more it is important that the sequence of steps can be reconstructed.

The software presented here is mainly based on partial ordering. However it is not only seen as a commercial version of WHASSE (Brüggemann and Halfon, 1995; Brüggemann et al., 1999). Several applications have shown that partial ordering often has to be combined with statistics or even other approaches, particularly when the resulting partial order and Hasse diagram respectively, presents a messy system of lines and several optimal alternatives arise. For this reason, further approaches like e.g. METEOR (Pudenz and Brüggemann, 2002) and clustering methods will be considered in a later release too.

In the following paragraph basic features and an example of software application within a business routine is presented.

2 Basic structure

2.1 Flexibility – Operating system

The program is written in Java-language and therefore it can run independent from the operating system of the computer (MS-Windows, Mac-Os, Linux). Furthermore in the basic program structure the application as applet is considered too, which enables the use via internet in future.

2.2 Flexibility – Customizing

In order to fulfil the demand for (user-) specific interfaces and modules, specific flexibility within the program structure was taken into account. Following some known model-view-controller-concepts this was obtained by a strict separation of the three levels:

2.2.1 Model

This level and package respectively, contains all classes (instead of objects the Java-language uses classes) which presents the interior structure of the governed data independent from the GUI (graphical user interface). For instance the classes responsible for the data matrix and the partial ordered set belong to the model. All classes whose content is saved when saving a file belong to the model can be classified as rule of thumb.

2.2.2 MetaGUI

This level contains all classes based on the model and providing the functions for modifications of the model, e.g. calculating a sensitivity analysis or specific properties like the number of comparabilities in a Hasse diagram. These classes do not contain any data for saving. What exactly can be defined as a Meta-component is controlled by a specific interface. Further examples for Meta-components are e.g. the table that allows a view on the data matrix, or the Diagram that allows a view on the Hasse diagram.

2.2.3 GUI

This level is responsible for real presentation and its layout on the screen. This level contains the so-called component manager, which accepts meta-components and translates their content into Java-components.

As mentioned above, this partitioning enables high flexibility with regard to modifications of specific features. For instance, the visual design of the program (user interface) can be changed by replacing or modifying GUI-classes in the way that no programming at the model or the MetaGUI is necessary. Furthermore it enables changes in data management (e.g. caching) without the need for changes at the MetaGUI and GUI. Finally other or more functionalities can be added by writing a new Meta-component without touching the MetaGUI or GUI.

Figure 1 shows an example of a program sequence from the view of the table (in the MetaGUI), that means the user opens the view on the Hasse diagram (which will be operated by the model) by a button of the table view.

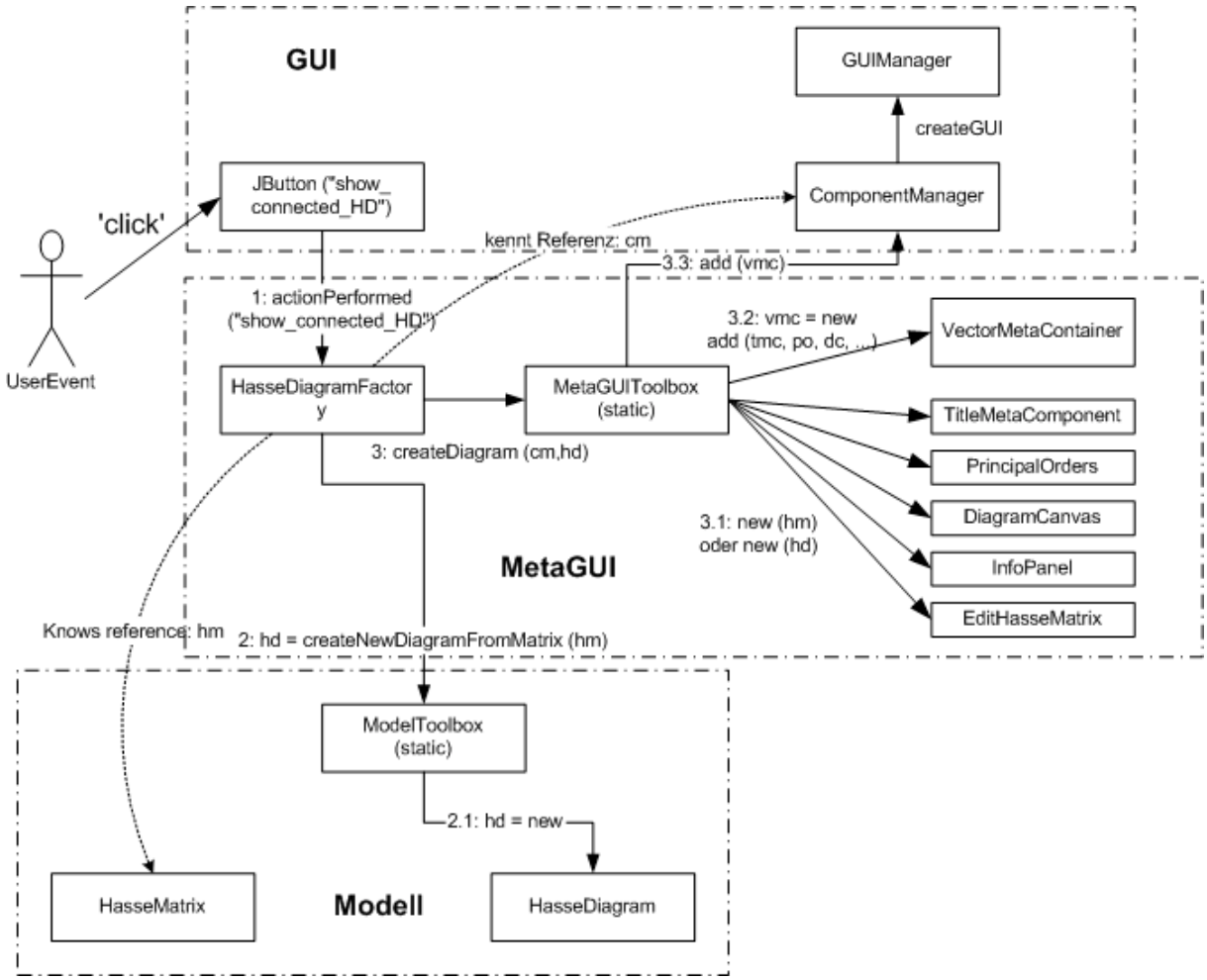


Figure 1: Program sequence from the view of the table: the user opens the view on the Hasse diagram from table view as part of the MetaGUI..

2.3 Transparency and comprehensibility

Here, transparency and comprehensibility are defined as ability to reproduce all steps within a data analysis and decision procedure respectively (see also above). For these purposes the so-called component manager was developed. In figure 1 it is seen that the GUI-functionality was subdivided into the component manager and the GUI-manager. The latter one is responsible for the final translation of a Meta component in Java components. The GUI-manager sets the "look-and-feel" of the program; for instance he defines that all Meta components (e.g. tables, diagrams, dialogs etc.) are shown in the main program window, alternatively each component could be arranged as register, as it may be known from MS-Excel sheets. The component manager knows the actual GUI-manager and passes the components to be shown to the GUI-manager. In doing so the component manager is responsible for the management of the Meta components shown. Figure 2 shows an example for this management.

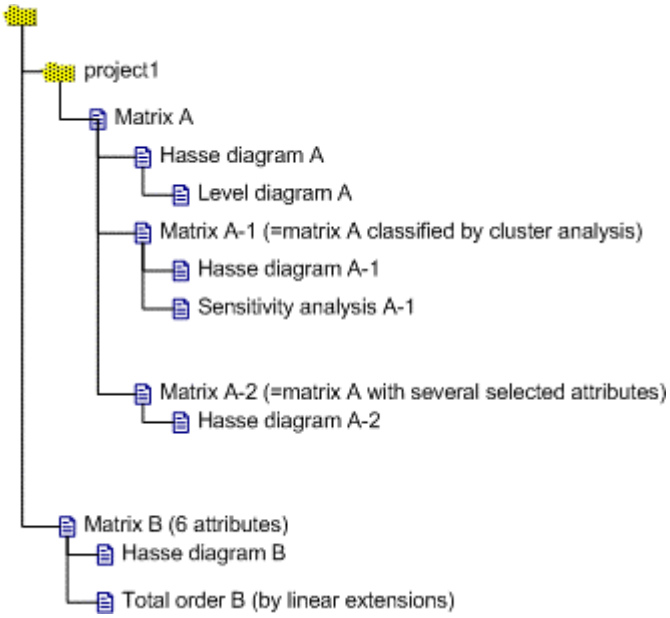


Figure 2: Component manager responsible for management of Meta components.

For example project 1 was generated in a previous session and has been started from Matrix A. As it is seen several operations were executed where each step can be reconstructed. When project 1 will be opened and the user selects one item, e.g. Hasse diagram A-2, then the Hasse diagram will be calculated; it is not saved as diagram but only the information that a Hasse diagram belongs to matrix A. However, in case of a cluster analysis the result is saved in a matrix and therefore by selecting the Matrix A-1 a table containing clustered data will be opened. During a session of course it is also permitted to open further tables or projects. When saving the file, the components (diagram, matrix, etc.) will be selected by the user from the component manger.

3 Software demonstration

This example is about toxicity estimation of products and recipes respectively, for detergent or familiar industries. The software called Criterion ToxEstimator and is based on the approach of partial ordered sets.

The problem can be described as follows: Before new detergent products can be placed on the market the so-called german detergent law demands a proof about no harmful effects in use. A clear definition of these effects is given by legislation and is expressed by hazard symbols and basic rules for handling hazardous materials, i.e. so-called r-p phrases (see Figure 3).

Xn
harmful to health
e.g. R20/21: harmful by inhalation and in contact with skin

Xi
Irritant
e.g. R38: Irritating to skin

F
R10: flamable

Figure 3: Hazard symbols and r-p phrases.

These investigations are time consuming and as consequence (and more important for the manufacturer) expensive too. Therefore the company is anxious to proof already against harmful effects in the development of new products. So let us now start the procedure with ToxEstimator:

In Figure 4 the main program window on a windows operating system is shown. The table view represents the matrix consisting of products P1, P2, .., P21 characterized by concentration of ingredients I1, I2, …, I27. The first column is a Boolean column where several products are classified with an obligation to mark. In the next (string) column the corresponding hazard symbol is listed. Furthermore, the ingredients itself are classified too however on the basis as a single chemical. That means, if a product only consists of e.g. ingredient I1 it has to be declared by the symbol Xi and the corresponding r-phrase. Assuming now, that on the basis of its ingredient concentration, all products have been already classified/declared in the past (Indeed the existence of such a matrix and declarations done in the past is the condition for applying the ToxEstimator).

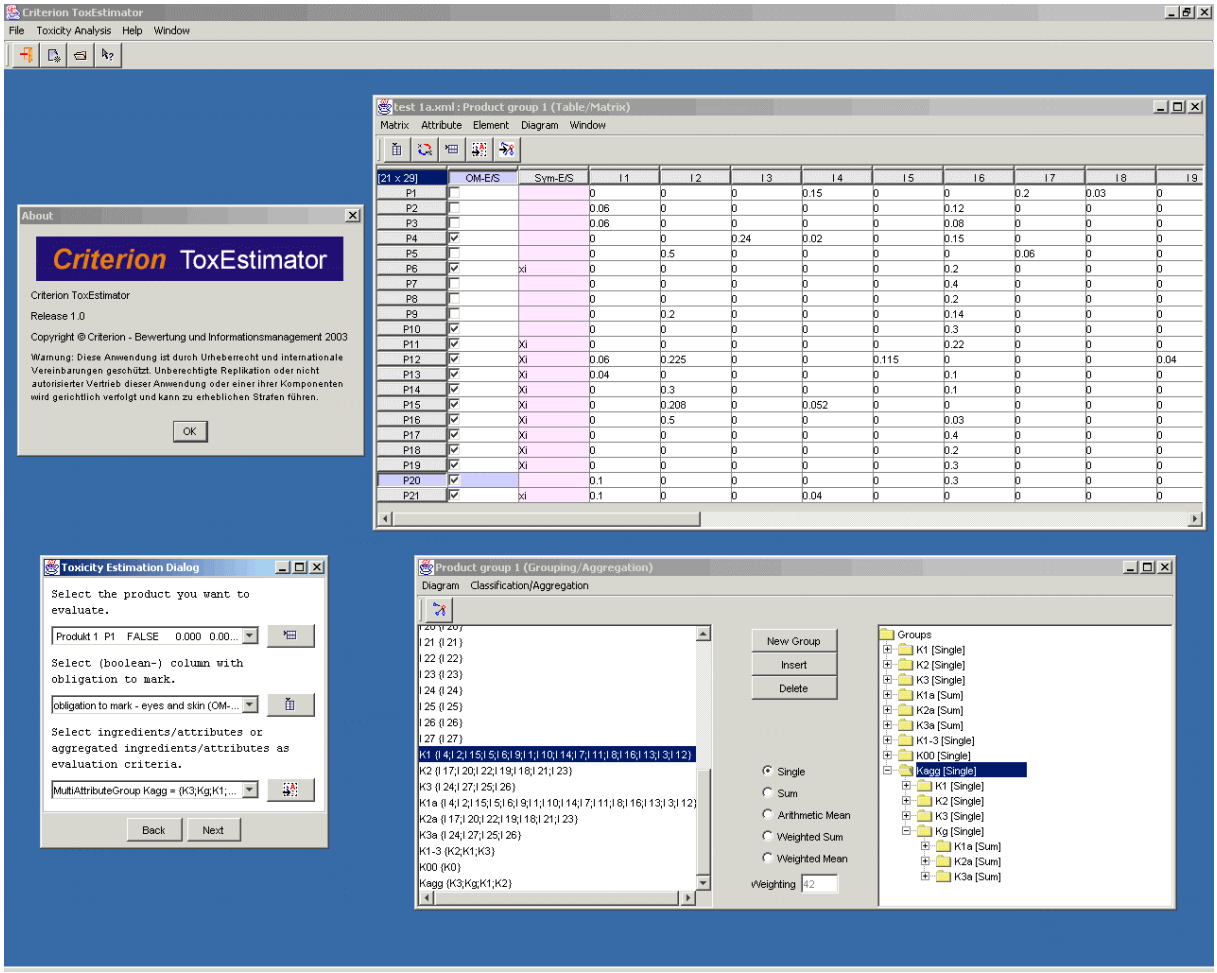


Figure 4: Window of the main program.

Following the interactive dialog box *Toxicity Estimation Dialog*, a product for evaluation (here P1), the Boolean column containing obligation to mark and finally the ingredients/attributes (the information base) as basis for evaluation have to be selected. The latter one is assisted by a *grouping/aggregation dialog* that enables aggregation of ingredients to so-called ingredient groups by different algorithm (sum, arithmetic mean,…). This has to be carefully done by the expert. After defining ingredient groups, here e.g. K1={I1,…,I16}

K2={I17,…,I23)

K3=(I24,…,I27)

Kg={K1a, K2a, K3a}; where $K1a = \sum_{x=1}^{x=16} I_x$, $K2a = \sum_{x=17}^{x=23} I_x$; $K3a = \sum_{x=24}^{x=27} I_x$,

the *Next* button activates the estimation procedure, where the results are successively presented. The first result is the evaluation by means of K1 (figure 5) and as it can be seen there is no product comparable to P1 which has an obligation to mark. If there would be products with an obligation to mark and which have lower quantities in all ingredients as P1, then P1 could be also subject of obligation to mark. This is the (relatively simple) idea of the toxicity estimation procedure. However as mentioned above, the expert has to define groups

of ingredients which give sense for comparison with regard to toxicity classification.

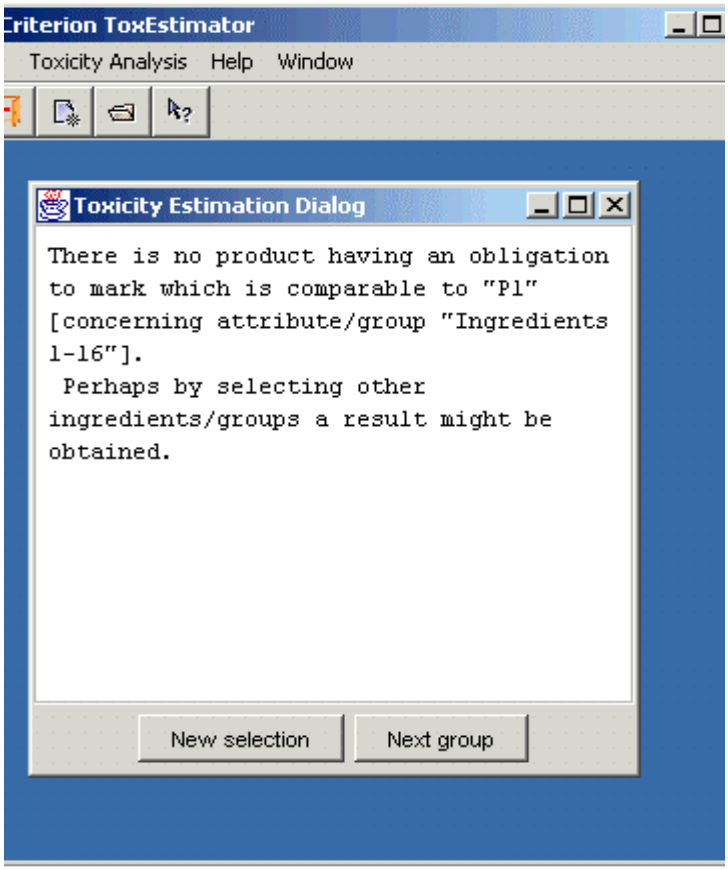


Figure 5: Result of evaluation using ingredients 1-16.

Figure 6 shows the result for e.g. Kg. For P1 the dialog box gives an estimation "obligation to mark is possible", because P1 has higher values in K1a, K2a and K3a as the products below which all have an obligation to mark. Additionally a table can be opened to show the values of P1 compared to the other products and as in the diagram view only the products comparable to P1 are shown.

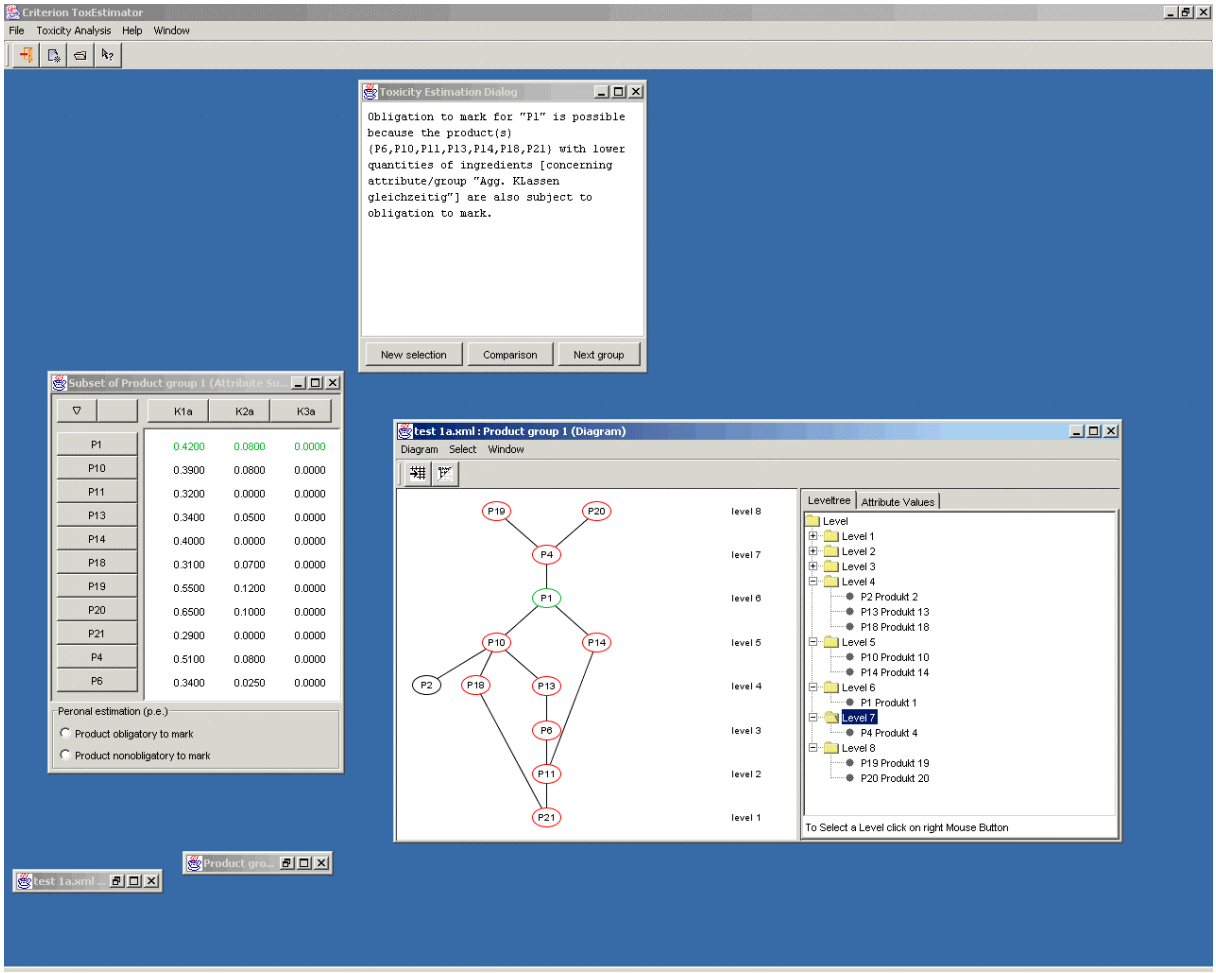


Figure 6: Evaluation step within the toxicity estimation procedure.

After several cycles depending of the number groups defined in the aggregation dialog (here four: Kagg=K1,K2,K3,Kg; see figure 4) a result of all evaluations is shown (figure 7).

4 Summary

The demonstration shown above is an example of the customized software. It includes user-specific features and leads interactively through an evaluation procedure. Aggregated ingredient groups, diagrams and results can be saved in a project file, written in XML. The software enables import of MS-Excel sheets as wells as text-files and import from clipboard. In principle there is no restriction about the size of data files; the handling is controlled by the computer performance alone. The program offers several formats for diagram export including the new vector format SVG.

The concept of partitioning the software into GUI, model, MetaGUI (and component manager) enables high flexibility with regard to (user-specific) modifications. Transparency and comprehensibility

for evaluation and decision procedures respectively are realized by the component manager.

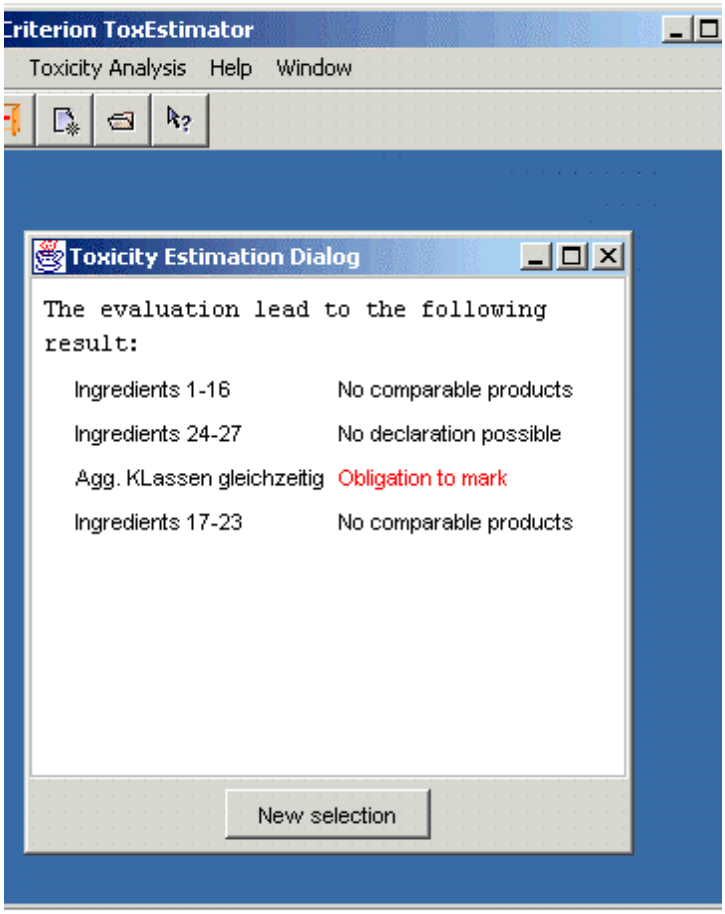


Figure 7: Result window of a toxicity estimation procedure.

A general standard software includes several more features about pre-processing, data analysis, graphical presentations like e.g.

- Boolean arithmetic to select objects and/or attributes in order to define subsets of the data matrix
- Explorative statistics
- Classifying by cluster analysis [4,5]
- Similarity of partially ordered sets [6]
- Sensitivity analysis with regard to the attributes of a partial ordered set (poset) [6]
- Linear extensions, mutual probabilities and total order (see also contributions of Brüggemann et al., Carlsen et al., Soerensen et al., this issue)
- Method of partial evaluation (METEOR) [3]

The consideration of further decision support tools is planned.

References

Brüggenman, R. & Halfon, E. 1995. Theoretical base of the program "Hasse". GSF-Bericht 10/95, München.

Brüggemann, R., Bücherl, C., Pudenz, S. & Steinberg, C. 1999. Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim. Hydrobiolog 27 (3), 170-178.

Pudenz, S. & Brüggemann, R. 2002. A New Decision Support System - METEOR. In: Voigt, K,. Welzl, G. (Eds.), Order Theoretical Tools in Environmental Sciences - Order Theory Meets Multivariate Statistics, Shaker, Aachen, 103 - 112.

Pudenz, S., Brüggemann, R., Luther, B., Kaune, A. & Kreimes, K. 2000. An Algebraic/Graphical Tool to Compare Ecosytems with respect to their Pollution V: Cluster Analysis and Hasse Diagrams, Chemosphere 40 (12), 1373-1382.

Luther, B., Brüggemann, R. & Pudenz, S. 2000. An Approach to Combine Cluster Analysis with Order Theoretical Tools in Problems of Environmental Pollution, MATCH 42, 119-143.

Brüggeman, R. & Halfon, E. 2000. Introduction into the general principles of the partial order ranking theory. In: P.B. Soerensen, L. Carlsen, B.B. Mogensen, R. Brüggemann, B. Luther, S. Pudenz, U Simon, E. Halfon, T. Bittner, K. Voigt, G. Welzl, F. Rediske (Eds.), Order Theoretical Tools in Environmental Science - Proceedings of the Second Workshop, October 21st, 1999 in Roskilde, Denmark, Nationale Environmental Research Institute (NERI), Roskilde, 7-44.

Partial order as tool in monitoring data interpretation

Marianne Thomsen[1)], **R. Brüggemann**[2)], **P.B. Sørensen**[1)], **B. Kronvang**[3)], **S. Gyldenkærne**[1)], **P. Fauser**[1)]

1) National Environmental Research Institute,
Department of Political Analysis,
DK-4000 Roskilde

2) Leibniz-Institute of Freshwater Ecology and Inland Fisheries,
Department Ecohydrology,
D-12587 Berlin

3) National Environmental Research Institute,
Department of Freshwater Ecology,
DK-4000 Roskilde

Abstract

A key factor in the design and planning of monitoring activities is to be able to deliver data, which support the DPSIR-principle as good as possible. The number of compounds on national and international concern lists is far beyond the number of compounds, which can be included within the National monitoring programme under the available economic resources. Therefore, in this paper we focus on the identification of redundant stations and substances in order to cost optimise the monitoring activities in a way that allows for new compounds to be implemented. In this way it can be possible to perform the best possible coverage of the societies environmental impact. Preliminary data analysis has identified the potential of using partial order theory in order to remove redundant stations and substances. Basically, if a substances is registered in a large number of stations in similar concentration levels then resources seems lost if the only purpose of the monitoring is to control the possible occurrence of that substance. Because in that case the substance only has to be identified once. Similarly if one substance is always found in lower concentration than another substance then the most important task is to register the highest concentration level substance. As long as the highest concentration substance is registered in low concentration values there are no reason to seek for the other lower concentration substance. For the majority of chemicals within the Danish monitoring programme the measured concentration levels is below any of the individual chemicals effect concentration. However, the approach

described in this paper is only valid provided that the in-situ mixture toxicity is non-specific and additive if present.

1 Introduction

It is important to perform systematic monitoring studies. The Ministry of Environment in Denmark has so far used the following definition of monitoring activities: A systematic and repeatedly collection of data, analysis and assessment over time of a given set of information based on a pre-designed survey. The objective is to be able to establish time trends, areas of concern and possible cause-relations. According to the terminology used within the Water Framework Directive, this definition corresponds to operational monitoring. Such monitoring activities are rather costly and thus limited compared with the large number of chemicals. It is therefore very important to combine monitoring activities with data interpretation in order to gain maximum knowledge from the existing data and to set up plans for future activities.

Two goals needs to be approached, i.e., 1) the monitoring data needs to be informative and 2) the resources spent on the monitoring activity need to be minimised.

1.1 Informative monitoring

The occurrence of chemicals within the natural environment is a function of

1) emission patterns and sources,
2) the inherent environmental parameters characteristic for different catchment areas and
3) the physico-chemical properties of the individual chemicals.

However, such interpretation will typically be associated with a rather large degree of uncertainty as the transport and process parameters are difficult to quantify under full-scale and realistic conditions. Thus, robust methods are needed for data handling, which can focus on the most general information about the fate and occurrence of the chemicals. Methods based on process-oriented compartment/transport models or on metric statistics will often be difficult to apply due to the inherent assumptions. Therefore, more robust ordinal (ranking) methods may be a good alternative or merely methods to support cause-relations and optimisation of future activities. Consequently this investigation will introduce aspects of the partial order theory for data interpretation of monitoring data. The first attempt of this study was to follow the DPSIR-concept, i.e. try to relate different parameters/indicators to

- Driving forces (D)
- Pressure (P)

- States (S)
- Impacts (I)
- Reactions (R).

Figure 1 shows the elements of the DPSIR-concept of which the Pressure-State relation is central for environmental management. Focus of the present study is on the Pressure and State. The State of the environment may be expressed in many ways, however, this investigation focus on State in relation to pesticide contamination of surface waters.

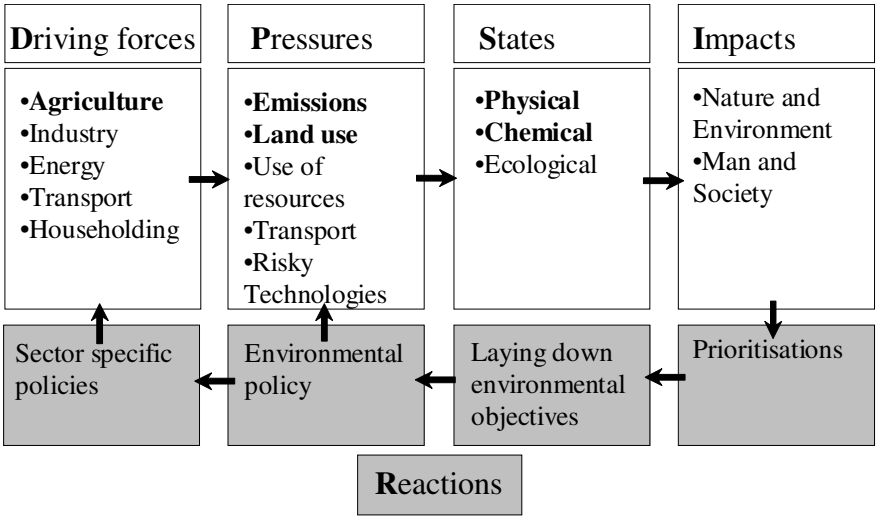


Figure 1. The Driving forces, Pressures, States and Impacts relationship which form the basis for environmental management, i.e. Reactions, having focus on the usage of pesticide in the agriculture.

The main Driving Force in relation to surface water contamination by pesticides is agriculture. The Pressure caused by the use of pesticides is quantified by data on 1) emission patterns and sources. In relation to chemical contamination, the State may be directly related to the Pressure data (Thomsen et al., 2003). In addition, State is expected to be a function of 2) the inherent environmental parameters characteristic for different catchment areas and 3) the physico-chemical properties of the individual chemicals.

In the preprocessing of monitoring data it turned out that there was a significant difference in the variance within and between sampling stations. However, for many of the sampling stations there was no significant difference in the mean concentration within and between catchment areas. Detection limits vary between the different sampling locations. When the maximum reported detection limits for each specific pesticide are used as a requirement for further data analysis, 204 out of 392 mean values at national scale are rejected. For more detailed information concerning the monitoring data please refer to Thomsen et al. 2003.

Here we present a methodological concept, which is able to identify redundant sampling stations and at the same time is able to cover all pesticides at minimum cost, i.e. sampling locations. We may assume that the objective is to fulfil the generation goal by the year 2020. Therefore, the highest risk catchment areas, still covering all priority pesticides, should be monitored covering short-term state and the

development of these areas should be controlled every six year within each six-year programme period.

The task is now to select representative sampling stations covering the highest risk areas and all priority pesticides, which is to form the basis for operative monitoring within the next programme period.

1.2 Minimising the demand of resources

One way to optimise the monitoring activities is to identify the redundant stations and/or substances in order to be able to implement new compounds within the monitoring programme under the given economic resources. In this way it is possible to perform the best coverage of the compounds and data that serve management decision needs regarding control monitoring. In this paper preliminary data analysis will identify the potential of using partial order theory in order to remove redundant stations and substances. Basically, if a substance is registered in a large number of stations in similar concentration levels resources seem to be lost when the only purpose with the monitoring is to control the possible occurrence of that substance. Because in that case the substance only have to be identified once. Similarly, when one substance is always found in lower concentration than another substance the most important task is to register the highest concentration level substance. As long as the highest concentration substance is registered in low concentration values, i.e. control monitoring level, there is no reason to seek for the other lower concentration substance. Given that a fairly simple relation between the usage, e.g. dose, of a given pesticide in a given catchment area, and the occurrence in the recipient surface water exists (Sørensen et al. 2003), a tool which takes into account the above considerations may form the basis for a dynamic and progressive monitoring program.

A more detailed derivation of guidelines for designing monitoring activities is outside the scope of this proceeding. The focus of this paper is to disclose redundancy in existing monitoring data using partial order theory.

2 Data

The topic of this investigation is to perform a preliminary investigation of measured concentration levels for the pesticide active ingredients in smaller Danish streams. The starting point is a data matrix where concentration levels of different active ingredients are recorded as a series of single measurements during a year. This data matrix is recorded for sampling stations located in different streams. A single number describing the concentration level for each active ingredient in each sampling station is going to be applied in the data analysis. So first such a characteristic number has to be formed based on the measurements. Three possibilities seem to be obvious: (1)

Mean value, (2) Median value, (3) Maximal monitored concentration. Which of these three alternatives are to be selected depends on the topic of the investigation, the data value variability and the number of the single measured concentration levels. The median value will in general be more robust against high variability combined with relatively few data points compared to estimated mean values. However, the mean value may better reflect the concentration level taking into account the existence of high values. The maximum values will be highly influenced by the number of data points and thus not be robust compared with the two other measures, but the maximum levels may be most relevant as an acute eco-toxicological characterisation for the conditions in the streams. The maximum concentration level together with the frequency of detection was suggested by Sorensen et al. (2003). A single number should be applied as a descriptor for the occurrence in order not to end up having a three-dimensional data matrix (different substances * different streams * different descriptors) but only a two-dimensional data matrix (different substances * different streams). In this investigation the median for each stream is therefore chosen as the most robust descriptor.

The number of sampling stations in the investigation is 27 and the location and catchment size is graphically displayed in Figure 2. The investigated substances are listed in Table 2, where the total number of measurements is reported too. It is seen in Table 2 that the typical total number of measurements is 189 equally distributed among the 27 sampling stations yielding 7 single samples at each station.

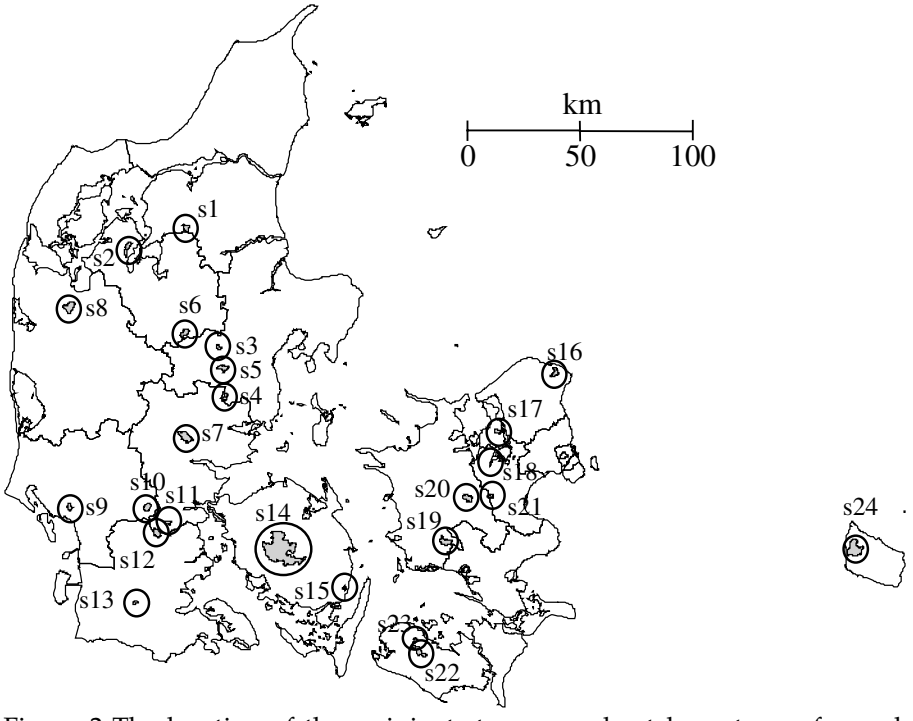


Figure 2 The location of the recipient streams and catchment area for each sampling station included in the Danish monitoring program. The catchment areas are coloured in grey with a black circle around.

Table 2. The pesticides are assigned by an id and name representing the 44 active ingredients, which are involved in this investigation. In addition the total number of measurements for each pesticide is given.

Id	Name	Number of single measurements	Id	Name	Number of single measurements	Id	Name	Number of single measurements
a1	dichlorprop	189	a16	dichlorbenzamid (BAM)	189	a31	metribuzin	a183
a2	MCPA	189	a17	desethylatrazin	189	a32	metsulfuron_methyl	183
a3	mecoprop	189	a18	desethylterbuthylazin	183	a33	nitrophenol	184
a4	DNOC	189	a19	desisopropylatrazin	189	a34	pendimethalin	189
a5	dinoseb	189	a20	dichlobenil	188	a35	pirimicarb	187
a6	atrazin	189	a21	dimethoat	189	a36	propiconazol	187
a7	simazin	189	a22	ethofumesat	186	a37	terbuthylazin	189
a8	24- D	189	a23	fenpropimorph	187	a38	aminomethylphosphon-acid (AMPA)	182
a9	bentazon	189	a24	hexazinon	189	a39	diuron	188
a10	bromoxynil	187	a25	hydroxycarbofuran	181	a40	glyphosat	182
a11	carbofuran	187	a26	ioxynil	187	a41	hydroxyatrazin	186
a12	chloridazon	183	a27	isoproturon	189	a42	trifluralin	174
a13	chlorsulforon	181	a28	lenacil	183	a43	desethylisopropylatrazin	168
a14	cyanazin	189	a29	maleinhydrazid	162	a44	ethylenthiourea (ETU)	183
a15	dalapon	170	a30	metamitron	189			

3 Results

3.1 Ranking of stream sites

The sampling stations are ranked using the median values of the concentration level for each active ingredient as a descriptor. Thus every station is associated with 44 descriptors, which are taken into account simultaneously in the ranking. The result is shown in the Hasse diagram in Figure 3 using the software WHASSE of Brüggemann & Halfon, (1995), Welzl et al. (1998) and Pudenz et al (2000). A line is only drawn between two stations when they can be ranked certainly so the station placed above has at least one descriptor value higher than the station placed below, and no descriptor value, which is lower. For 16 stations it is not possible to make any certain rankings and they are all listed in the box to the left in Figure 3. This relatively low number of rankings leaves only a spared knowledge about ranking relations in general using the medium concentration data.

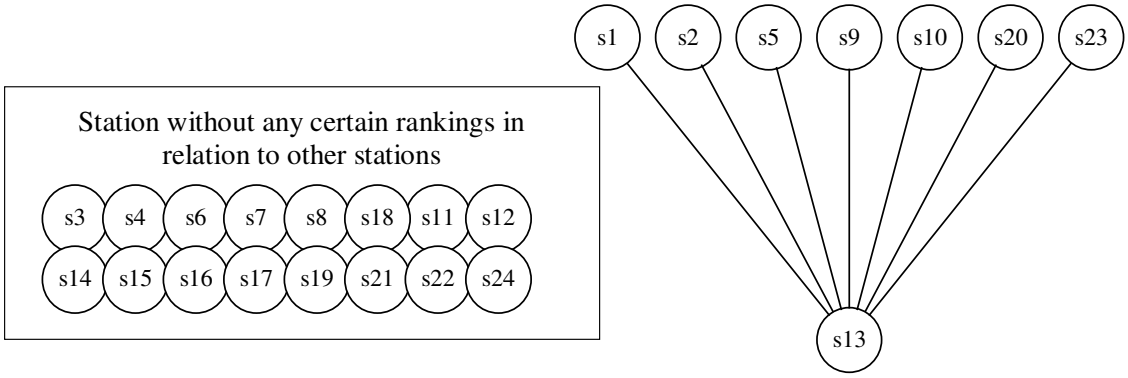


Figure 3. Hasse diagram: Sampling stations ranked due to median of concentration level for all 44 substances. The stations without connecting lines listed in the box can not be certain ranked with any station when all 44 descriptors are taken into account.

It is a rather restrictive analysis to use the concentration data directly as descriptor, because small and thus not important differences between concentration values may create many ambiguous rankings. In order to filter out such non-significant ambiguity a p-filter can be applied as suggested by Brüggemann et al., (1999). The principle of the p-filter is to do the following for each descriptor, i.e. median concentration level for each pesticide: (1) The objects, i.e. sampling stations, are ranked only in relation to the specific descriptor; (2) The descriptor values, which is ranked below a pre-selected ranking level (in this investigation below top 5) is assigned the value of zero. In this way it is possible to remove ambiguous rankings between small and thus not so important descriptor values. In the following analysis the p-filter is used to select the top 5 surface water concentration values for all descriptors and the lower ranked descriptors are sat equal to zero; (3) An additional reduction of ambiguity is obtained by setting all remaining non-zero values equal to unity. The result is a data matrix where all values are either unity or zero and the resulting Boolean Hasse diagram is shown in Figure 4. Let us consider two stations that are ranked in this diagram, e.g. s1 and s13. This case describes a situation in which the same substances among the top 5 has been found in station s1 and s13 but at least one more substance in station s1 than in station s13. From the p-filter approach, however, still only three stations, namely s13, s14 and s17 can be selected as redundant from the top-priority sample stations.

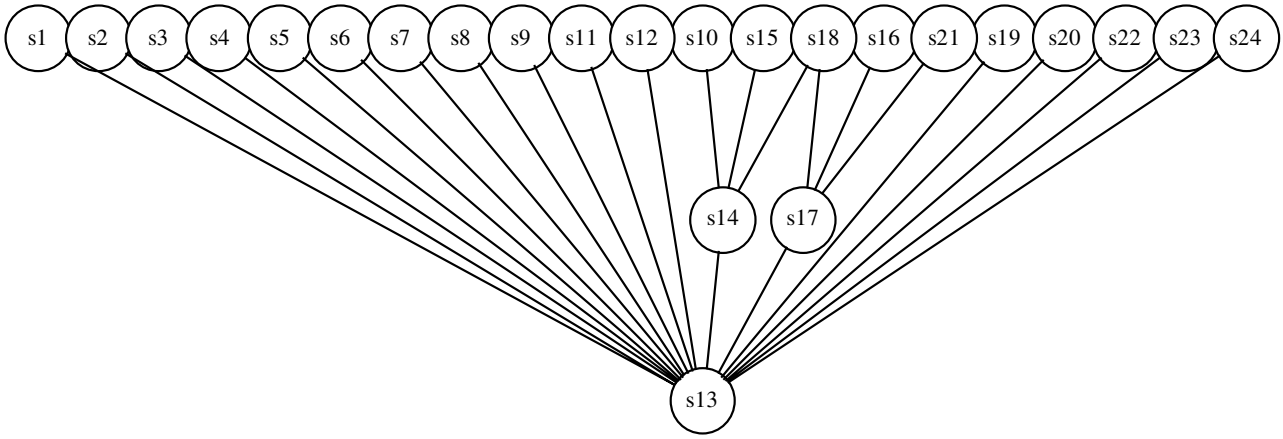


Figure 4: Hasse diagram: Sampling stations ranked due to median of concentration level for all 44 substances. The principle of a p-filter is applied as described in the text. There are 21 top priority sample stations, which have the highest load of pesticide occurrence. Compared to the 21 sample stations, three stations have a lower load of top 5 surface water concentration values, i.e. upon occurrence pesticides with highest concentration.

However, identification of redundant stations is not necessarily done by selecting the non-maximal objects in Figure 4. The problem we are going to solve is to be able to identify a substance in the stream water. We do not care in what station sample the substance is measured. The data interpretation is therefore pointing at: "What is the lowest possible number of stations we can include in a subset of stations if we still want to be able to measure all the substances we have registered in the complete set of stations". As such, there may be a combination of maximal elements in Figure 4, which together cover all pesticides. Therefore a complete computer algorithms is under development for a randomised first step combinatory analysis of those maximal objects consisting of a minimum number of sample stations and maximum number of redundant stations with a lower rank (Thomsen et al., 2003). The results of this stepwise selection of a reduced set of sample stations covering all pesticides results in a combined set of 12 stations, which is identified as a possible combined maximal object for the remaining set of stations. In this preliminary approach it seems possible to reduce the set of stations to about half of the original set and still be able to identify, and monitor, the same number of pesticides. These combined sets of sampling stations are s1, s3, s4, s5, s7, s10, s11, s12, s19, s21, s22 and s23.

3.2 Results: Ranking of substances

The substances can be ranked based on their concentrations in different sampling sites. The list of substances is found in Table 2.

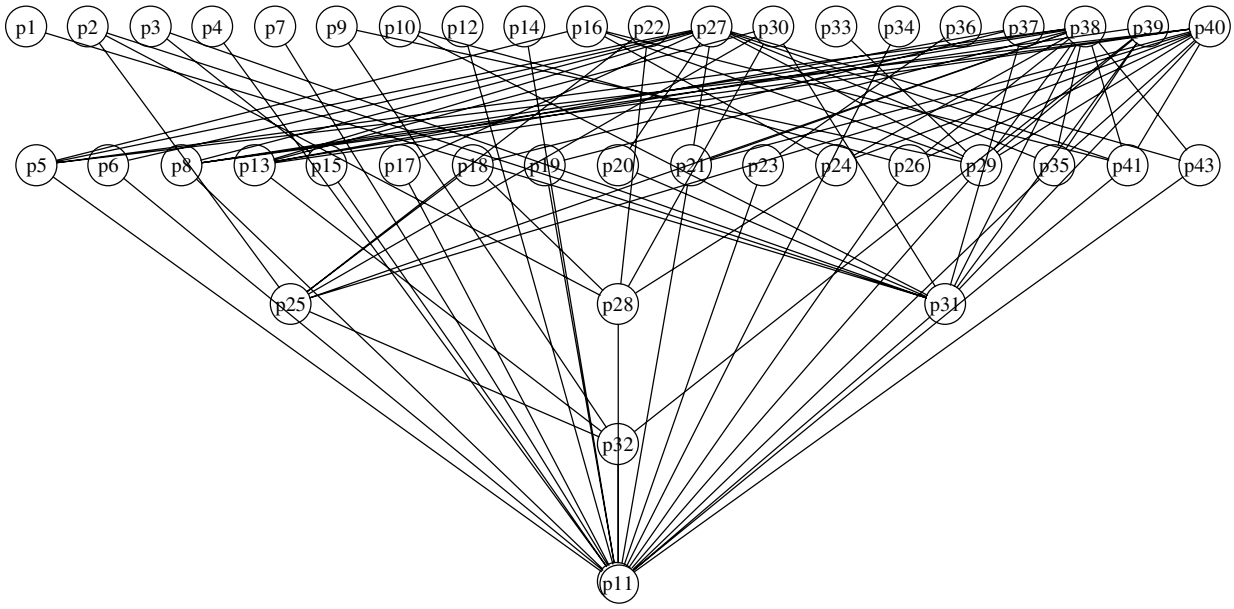


Figure 5. Hasse diagram showing 44 substances ranked according to their median concentration levels in 24 stations. Equivalent rank was assigned to the following set: {p11, p42, p44}, which represents the sub set of substances not found in the data set.

Relationships between single descriptors can be investigated using a descriptor sensitivity analysis as described by Brüggemann et al., (2001). A key parameter in such an analysis is denoted the W matrix. The sensitivity analysis is performed by subsequently neglecting one descriptor and observing the resulting change in the Hasse diagram. The changes upon this "leave-one-out" approach are mapped in the W matrix. The only possible change in a Hasse diagram when one descriptor is neglected is an increase in the number of certain rankings. If a single descriptor is the only reason for the ambiguity of a ranking between two objects then this ranking will become certain if that descriptor is neglected. E.g. consider two substances denoted p1 and p2 respectively, where p1 and p2 may have equal values (p1 = p2) with respect to 10 descriptors and where p1 > p2 for 23 descriptors and p2>p1 for one single descriptor. If that is the case then p1 and p2 will not be ranked in the Hasse diagram in Figure 7, because both p1 > p2 and p2 > p1 are possible according to the total set of descriptors. This creates an ambiguous ranking, but if the single descriptor, where p2>p1, is neglected then the ranking between p1 and p2 will become certain and equal to p1>p2 and this descriptor thus have an influence on the structure of the Hasse diagram. The W matrix values calculated by the WHASSE program can determinate the number of additional certain rankings due to the removal a single descriptor. The results are shown in Table 3 below.

Table 3: W matrix result calculated by WHASSE program (Brüggemann et al., 1999).

Station	W result	cumulative W result
s11	42	42
s23	40	82
s22	37	119
s1	36	155
s20	29	184
s4	17	201
s9	10	211
s8	8	219
s5	6	225
s12	6	231
s6	3	234
s19	3	237
s15	2	239
s24	2	241
s3	1	242
s10	1	243
s14	1	244
s2	0	244
s7, s21	0	244
s2,s13,s16, s17, s18	0	244

It can be seen that the station s11 is the most important single descriptor being responsible for 41 ambiguous rankings. The grey coloured cells in Table 3 are the stations, which are the reduced set of maximal objects from Figure 5. In the right column in Table 3 is listed the accumulated number of additional rankings from the top. The upper four stations in Table 3 are responsible for total 155 additional countings for added comparisons. From Table 3 it is possible to see that 244 ambiguity rankings in total are a result of subsequently leaving out a single descriptor at a time. This is a rather high number compared to the total number of ≤- relations realized in Figure 5, which are 214. This indicates that many of the stations do have individual properties in opposition to the other stations, which makes the Hasse diagram structure sensitive in relation to many of the single stations. However, this is in some way just a verification of the properties already identified in Figure 3, where only a few ≤- relations are seen to exist. This shows that every station do have unique individual properties in opposition to the other stations. What the result in Table 3 shows, which cannot be deduced from Figure 3, is that some of the stations are more in opposition to the other stations than others.

The reduced set of maximal objects in Figure 5 should be a sufficient set of stations for control monitoring the occurrence of the pesticide substances. So it may be of interest to see how the Hasse diagram for the substances looks like when only the subset of stations imbedded in the reduced set of maximal objects is used as descriptor. This is shown in Figure 6.

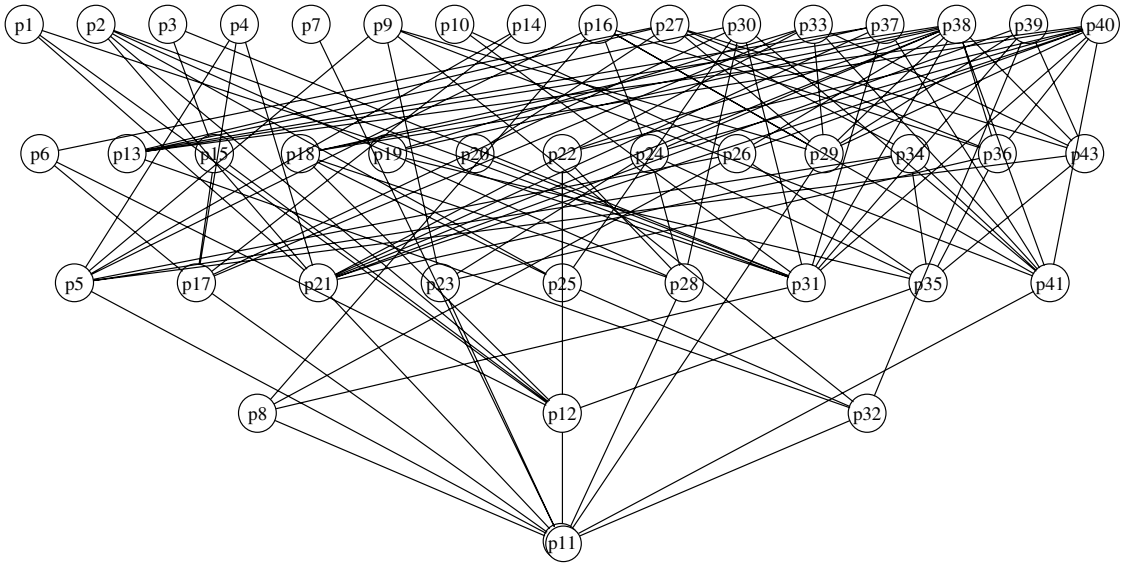


Figure 6. The Hasse diagram for the substance median concentration, when only the stations included in the set of reduced maximal objects in Figure 5 are included.

Comparing Figure 5 and 6 it can be seen that the ordering in Figure 6 is more complete and the W matrix shows that there are 87 additional rankings in Figure 6 compared to Figure 5.

A combination of the results from the Figures 4 and 6 yields a procedure for removing redundant stations and substances in control monitoring programs of the occurrence of pesticide active ingredients. Simply taking the maximal objects from both figures does this and the result is selection of the following 12 stations: s 1, 3, 4, 5, 7, 10, 11, 12, 19, 21, 22, 23 and the following 16 substances: p: 1, 2, 3, 4, 7, 9, 10, 14, 16, 27, 30, 33, 37, 38, 39, 40. Originally the data set included 24 stations and 44 substances, which yields 1056 determination of median concentration levels. These amounts of samples and compounds are now reduced to 12 stations and 16 substances. In total a reduction from 1056 to 192 determinations of median concentration levels.

4 Discussion and Conclusion

Based on an assumed generic proportional relationship between use pattern and occurrence, and that the toxicity potentials of the individual pesticides are similar, a procedure based on a search of maximal elements of the binary data matrix and its transposed form for minimisation of the control monitoring activity has been briefly outlined. The criteria for identifying the top-priority chemicals and sample location may of cause be modified according to the purpose of the prioritisation, i.e. control, operational or investigative monitoring. Furthermore, additional criteria and boundaries may be brought into play when developing similar monitoring design tools. As such, in

the case of operational or investigative monitoring a whole other set of data analysis and design tools needs to be taken into use (e.g. Sorensen et al., 2003). The biggest issue here is which environmental management questions needs to be answered and what are the restrictions of the monitoring programme, e.g. does it only cover national scale level control monitoring, or does it include operational monitoring as well. Operational monitoring as defined within the water framework directive i.e. monitoring at landscape level requires design according to P-S and/or P-S-I, the goal being to answer if the existing pressures are low enough to keep contamination below certain criteria given in certain directives, environmental objectives etc. Within this paper we have presented a conceptual tool for optimisation of the existing pesticide monitoring programme in Danish surface waters. The method may be adjusted and modified to meet individual requirements at different organisational and management levels.

References

Brüggemann, R., Bücherl C., Pudenz, S. & Steinberg C.E.W. 1999. Application of the concept of partial order on comparative evaluation of environmental chemicals Acta Hydrochimica et Hydrobiologica, Vol. 23, No. 3, pp. 170-178.

Brüggemann, R. & Halfon, E. 1995. Theoretical Base of the Program "Hasse". GSF-Bericht 20/95, München-Neuherberg.

Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., & Steinberg, C. 2001. Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests. J.Chem.Inf.Comp.Sc., 41:918-925.

Pudenz, S., Brüggemann, R., Luther, B., Kaune, A., & Kreimes, K. 2000. An algebraic/graphical tool to compare ecosystems with respect to their pollution V: cluster analysis and Hasse diagrams. Chemosphere, 40:1373-1382.

Sørensen, P.B., Brüggemann, R., Carlsen, L., Mogensen, B.B., Kreuger J. & Pudenz, S. 2003. "Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory - Data interpretation and model development", Environmental Toxicology and Chemistry, Vol 22, No. 3, pp. 661-670.

Thomsen, M., Sørensen, P.B., Worrall, F., Gyldenkærne, S., Kronvang, B. 2003. "Landscape level design and prioritisation of chemicals within the national monitoring programme. In prep.

Welzl, G., Voigt, K., & Rediske, G. 1998. Visualisation of Environmental Pollution - Hasse diagram technique and Explorative Statistical Methods. In: Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences Berichte des IGB Heft 6, Sonderheft I, 1998, edited by Group Pragmatic Theoretical Ecology, pp. 101-110. IGB, Berlin.

[Blank page]

National Environmental Research Institute

The National Environmental Research Institute, NERI, is a research institute of the Ministry of the Environment. In Danish, NERI is called *Danmarks Miljøundersøgelser (DMU)*.
NERI's tasks are primarily to conduct research, collect data, and give advice on problems related to the environment and nature.

Addresses:

URL: http://www.dmu.dk

National Environmental Research Institute
Frederiksborgvej 399
PO Box 358
DK-4000 Roskilde
Denmark
Tel: +45 46 30 12 00
Fax: +45 46 30 11 14

Management
Personnel and Economy Secretariat
Monitoring, Research and Advice Secretariat
Department of Policy Analysis
Department of Atmospheric Environment
Department of Marine Ecology
Department of Environmental Chemistry and Microbiology
Department of Arctic Environment

National Environmental Research Institute
Vejlsøvej 25
PO Box 314
DK-8600 Silkeborg
Denmark
Tel: +45 89 20 14 00
Fax: +45 89 20 14 14

Monitoring, Research and Advice Secretariat
Department of Marine Ecology
Department of Terrestrial Ecology
Department of Freshwater Ecology

National Environmental Research Institute
Grenåvej 12-14, Kalø
DK-8410 Rønde
Denmark
Tel: +45 89 20 17 00
Fax: +45 89 20 15 15

Department of Wildlife Biology and Biodiversity

Publications:
NERI publishes professional reports, technical instructions, and the annual report. A R&D projects' catalogue is available in an electronic version on the World Wide Web.
Included in the annual report is a list of the publications from the current year.

Faglige rapporter fra DMU/NERI Technical Reports

2003

Nr. 455: Kvantificering af næringsstoffers transport fra kilde til recipient samt effekt i vandmiljøet. Modeltyper og deres anvendelse illustreret ved eksempler. Nielsen, K. et al. 114 s. (elektronisk)

Nr. 456: Opgørelse af skadevirkninger på bundfaunaen efter iltsvindet i 2002 i de indre danske farvande. Af Hansen, J.L.S. & Josefson, A.B. 32 s. (elektronisk)

Nr. 457: Kriterier for gunstig bevaringsstatus. Naturtyper og arter omfattet af EF-habitatdirektivet & fugle omfattet af EF-fuglebeskyttelsesdirektivet. Af Søgaard, B. et al. 2. udg. 460 s. (elektronisk)

Nr. 458: Udviklingen i Vest Stadil Fjord 2001-2002. Af Søndergaard, M. et al. 25 s. (elektronisk)

Nr. 459: Miljøøkonomiske beregningspriser. Forprojekt. Af Andersen, M.S. & Strange, N. 88 s. (elektronisk)

Nr. 460: Aerosols in Danish Air (AIDA). Mid-term report 2000-2002. By Palmgren, F. et al. 92 pp. (electronic)

Nr. 461: Control of Pesticides 2002. Chemical Substances and Chemical Preparations. By Krongaard, T., Petersen, K. & Christoffersen, C. 30 pp. (electronic)

Nr. 462: Bevaringsstatus for fuglearter omfattet af EF-fuglebeskyttelsesdirektivet. Af Pihl, S. et al. 130 s. (elektronisk)

Nr. 463: Screening for effekter af miljøfarlige stoffer på algesamfund omkring havneanlæg. Af Dahl, K. & Dahllöff, I. 37 s. (elektronisk)

Nr. 465: Miljøundersøgelser ved Maarmorilik 2002. Af Johansen, P., Riget, F. & Asmund, G. 62 s. (elektronisk)

Nr. 466: Atmosfærisk deposition 2002. NOVA 2003. Af Ellermann, T. et al. 88 s. (elektronisk)

Nr. 467: Marine områder 2002 - Miljøtilstand og udvikling. NOVA 2003. Af Rasmussen, M.B. et al. 103 s. (elektronisk)

Nr. 468: Landovervågningsoplande 2002. NOVA 2003. Af Grant, R. et al. 131 s. (elektronisk)

Nr. 469: Søer 2002. NOVA 2003. Af Jensen, J.P. et al. 63 s. (elektronisk)

Nr. 470: Vandløb 2002. NOVA 2003. Af Bøgestrand, J. (red.) 76 s. (elektronisk)

Nr. 471: Vandmiljø 2003. Tilstand og udvikling - faglig sammenfatning. Af Andersen, J.M. et al. 157 s., 100,00 kr.

Nr. 472: Overvågning af Vandmiljøplan II - Vådområder 2003. Af Hoffmann, C.C. et al. 83 s. (elektronisk)

Nr. 473: Korrektion for manglende indberetninger til vildtudbyttestatistikken. Af Asferg, T. & Lindhard, B.J. 28 s. (elektronisk)

Nr. 474: Miljøundersøgelser ved Mestervig 2001. Af Aastrup, P., Tamsfort, M. & Asmund, G. 47 s. (elektronisk)

Nr. 475: Vandrammedirektivet og danske søer. Del 1: Søtyper, referencetilstand og økologiske kvalitetsklasser. Af Søndergaard, M. (red.) et al. 140 s. (elektronisk)

Nr. 476: Vandrammedirektivet og danske søer. Del 2: Palæoøkologiske undersøgelser. Af Amsinck, S.L. et al. 118 s. (elektronisk)

Nr. 477: Emissions of Greenhouse Gasses and Long-Range Transboundary Air Pollution in the Faroe Islands 1990-2001. By Lastein, L. & Winther, M. 59 pp. (electronic)

Nr. 480: Danske søer – fosfortilførsel og opfyldelse af målsætninger. VMP III, Fase II. Af Søndergaard, M. et al. 37 s. (elektronisk)

Nr. 481: Polybrominated Diphenyl Ethers (PBDEs) in Sewage Sludge and Wastewater. Method Development and validation. By Christensen, J.H. et al. 28 pp. (electronic)

2004

Nr. 482: Background Studies in Nuussuaq and Disko, West Greenland. By Boertmann, D. (ed.) 57 pp. (electronic)

Nr. 483: A Model Set-Up for an Oxygen and Nutrient Flux Model for Århus Bay (Denmark). By Fossing, H. et al. 65 pp., 100,00 DDK.

Nr. 484: Satellitsporing af marsvin i danske og tilstødende farvande. Af Teilmann, J. et al. 86 s. (elektronisk)

Nr. 485: Odense Fjord. Scenarier for reduktion af næringsstoffer. Af Nielsen, K. et al. 274 s. (elektronisk)

Nr. 487: Effekt på akvatiske miljøer af randzoner langs målsatte vandløb. Pesticidhandlingsplan II. Af Ravn, H.W. & Friberg, N. 43 s. (elektronisk)

Nr. 489: Overvågning af bæver Castor fi ber i Flynder å, 1999-2003. Af Elmeros, M., Berthelsen, J.P. & Madsen, A.B. 92 s. (elektronisk).

This is a collection of proceedings from the fifth workshop in Order Theory in Envi-ronmental Science. This workshop series concern the development of the concept of Partial Order Theory is development in relation to practical application and the use is tested based on specific problems. The Partial Order Theory will have a potential use in cases where more than one criterion is included in a prioritisation problem both in relation to decision support and in relation to data-mining and interpretation. Espe-cially the problems where a high degree of complexity results in considerable uncer-tainty are good candidates for application of Partial Order Theory.

National Environmental Research Institute
Ministry of the Environment

ISBN 87-7772-783-5
ISSN 1600-0048