

Estimation of averaged ranks by extended local partial order models

Rainer Brüggemann, Ute Simon, Silke Mey

Leibniz - Institute of Freshwater Ecology and Inland Fisheries
Müggelseedamm 310, D-12587 Berlin Friedrichshagen, Germany

Tel: +49 30 64181666 E-Mail: brg @ IGB-Berlin.de, brg_home@t-online.de

Abstract

The appearance of the White Book of the EU has refreshed the interest in ranking systems for chemicals. Many ranking schemes have a very simple mathematical structure: The chemicals are characterized by several attributes, such as fate descriptors in surface waters. These attributes are then combined by an appropriate method in order to derive a ranking index which gives a linear order. One simple method belongs to the approach of utility functions and implies by a weighting scheme the complete compensation among different attributes. An alternative approach is that of the Hasse Diagram Technique, where a dominance of one chemical over another is only established, if all attributes simultaneously support this dominance. The link between both extremes, namely that of utility functions and that of Hasse Diagram Technique, is the set of linear extensions, LE, which can be deduced from the partial order found by Hasse Diagram Technique. It is shown that any weighting scheme within the utility function approach will reproduce one linear order of the set of linear extensions. Hence it is of interest to characterize the LE set by statistical methods. One of the most promising methods is to calculate the averaged rank. The averaged rank depends solely on the structure of the Hasse Diagram, which in turn is an order theoretical representation of the data matrix.

1 Introduction

The evaluation of chemicals is of renewed interest, as the appearance of the "White book" by the European Community shows [1]. Risk assessment of chemicals based on available data is of specific interest as it allows to identify chemicals of high environmental and health hazard. However, such careful risk assessment is time and cost intensive and therefore priority setting by ranking methods is of major importance (see also [2]). Without going into details the priority setting of chemicals by ranking methods is mainly performed by a numerical combination of available chemical attributes including their weighting. The aim is to form a ranking index which provides a linear order. This procedure, however, implies a complete compensation among attribute values: A bad evaluation on one attribute can be compensated for by a good one in another attribute. As attributes describe different chemical characteristics, such as volatilization or sedimentation fluxes, a compensation can be questionable in general.

An alternative approach to sort chemicals is the Hasse Diagram Technique (HDT), which can be considered as a specific partial order. The chemicals are to be sorted with respect to their potential environmental hazard, as follows: Let be $q(i)$ appropriate attributes by which the chemicals are to be ordered. Traditionally, in HDT the set of attributes is called the information basis, IB^1 . Let further x, y be two chemicals, and $q(i,x)$ the value of the i^{th} attribute of the chemical x ($i=1, \dots, m$). Then $x \geq y$ (read x is evaluated worse or equal than y) if and only if $q(i,x) \geq q(i,y)$ for all $i=1, \dots, m$. The mere fact that x is comparable with y , i.e. $x \geq y$ or $x \leq y$ will denoted by $x \perp y$. We call the set G of N elements (here: chemicals), the ground set. By $R \subset G \times G$ an equivalence relation is introduced as follows:

$$x R y \Leftrightarrow q(i,x)=q(i,y) \quad \forall q(i) \in IB \text{ and } x,y \in G \quad (\text{Eq. 1})$$

In the following we assume that the quotient set G/R can be identified with G (i.e. there is no equivalence class with more than one element of G). Let be $x, y \in G$. If $x \perp y$ is not valid, then x is incomparable with y . Incomparability between two elements x and y is denoted as $x \parallel y$.

¹ Note that in the well known Formal Concept Analysis [3] the information base is also simply called the attribute set.

The partially ordered set (abbr.: poset) can be denoted either by (G, \leq) or by (G, IB) , because the \leq -relation is based on the simultaneous comparison of attribute values $q(i,x)$, $x \in G$, $q(i) \in IB$. For more details, see [4]. Consequently, compensation is excluded and appears as "incomparability" among chemicals.

Information which chemical should be considered as important, can be deduced from the Hasse Diagram, without a difficult weighting process. However, due to incomparabilities the identification of one worst chemical is often no more possible. Therefore it is a good strategy to add besides the results of HDT a linear order which is solely based on the graph theoretical structure of the Hasse Diagram and which also avoids the difficult weighting process. Such a linear order can be deduced by the so-called averaged height of posets [5], which is called an averaged rank in applicational studies [6]². We demonstrate the use of Hasse Diagrams by a real life example to evaluate chemicals detected in the river Main, Germany, and show that in many cases a statistical characterization by averaged ranks can easily be obtained by a simple probability scheme.

2 The link between utility function approach and partial orders

The utility function approach has a firm theoretical foundation [7], unfortunately this approach is often used in a simple manner. In the simplest form of the utility approach one proceeds as follows: (1) Select attributes $q(i) \in IB$, (2) define individual preference functions for any of these $q(i)$: $f_i(q(i))$, (3) find individual weights g_i for each $f_i(q(i))$, (4) calculate a ranking index $\Gamma(x)$ for a chemical $x \in G$ by equation 2 (see below) and (5) order the chemicals by their $\Gamma(x)$ -values.

$$\Gamma(x) := \sum_{i=1}^m g_i \cdot f_i(q(i,x)) \quad (\text{Eq. 2})$$

Between evaluation techniques like that of the utility functions and that of the HDT, described in section 1, a link can be found as follows:

1. From the partial order (G, IB) the set of linear extensions, LE can be obtained [8].
2. If there are no ties (i.e. equivalence of chemicals with respect to the function Γ) any utility function positive monotonous in $q(i)$ leads to a linear ranking, which must be

² Note that the concept of "averaged rank" does not imply that (G,IB) has a rank-function!

one of the linear extensions, obtained from the partial order. Note, however that the reverse need not be true: Simple counterexamples show that not each linear extension can be thought of as a result of an equation such as Eq. 2.

3. The statistical characterization of LE is of interest. One characteristic value which can be found from LE is the averaged rank, R_{kav} , for each element of G .
4. By R_{kav} a linear order is induced.

Taking this approach, the concept of partially ordered set is not replaced by another linear order, but a measure is derived, which is independent of weight factors g_i and which does not depend on transformations of the attributes as far as they let the partial order invariant. The benefits are that it can be used as an internal mean in order to compare results of other ranking schemes and it summarizes the complex information of a partial order. Hence, the linear order induced by the averaged ranks is thought of as an additional help for decision makers. Especially in complex diagrams this linear order might be a convenient tool. Figure 1 shows schematically the basic idea:

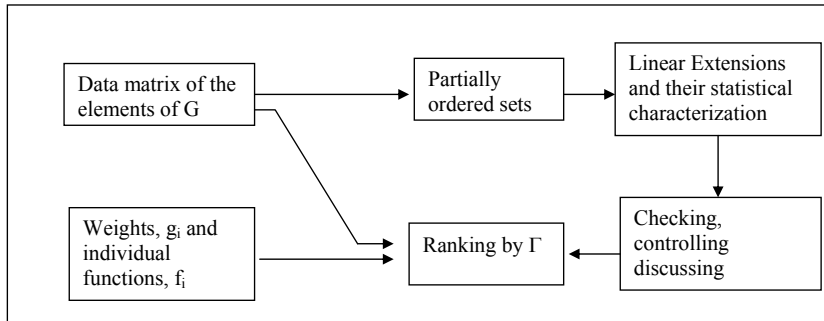


Figure 1: Relation between partially ordered sets and linear ranks induced by averaged ranks

3 The concept of averaged ranks

Taken the linear extensions of a partial order, the so called spectrum $\lambda(x)$ of an element x can be calculated (see for example [12]): The i^{th} component, $\lambda_i(x)$, of the spectrum $\lambda(x)$ is the frequency how often an element $x \in G$ has the rank i in the set of the linear extensions. (Note, as the linear extensions are linear orders, a rank function exists.) Clearly $\sum \lambda_i(x) = LT$, the number of linear extensions. The exact averaged rank of any element is then found by equation 3:

$$\text{Rkav}(x) = \sum_{i \in \{1, \dots, N\}} \frac{\lambda_i(x)}{LT} \cdot i \quad (\text{Eq. 3a})$$

Equivalently one may write:

$$\text{Rkav}(x) = \frac{\sum_j^{\text{LT}} \text{rank}(x, j)}{LT} \quad (\text{Eq. 3b}),$$

where $\text{rank}(x, j)$ is the rank of element x taken the j^{th} linear extension having a value $i \in \{1, \dots, N\}$. Note that even in this most general case, elements of the ground set G may be equivalent, as their spectrum is equivalent. One may define an appropriate equivalence relation, by the equality of Rkav -values. Especially in symmetric Hasse Diagrams one may find many nontrivial equivalence classes. The equivalence relation we define as follows:

Let be $A \subset G \times G$ then:

$$xAy; \Leftrightarrow \text{Rkav}(x) = \text{Rkav}(y) \quad (\text{Eq. 4})$$

As later different approximations for the averaged rank are discussed, equation 4 may be specified by indices, like $A_i, \text{Rkav}_{(i)}(x)$.

4 The concept of local partial order models

4.1 Introductory remarks

In a preceding paper [9] the concept of a local partial order model (LPOM) was introduced. The main idea, outlined there, was:

- Select one element $x \in G$
- Find a partial order which represents as far as possible the order relations for x and which is simple enough to
- derive an estimation formula for the averaged rank of x .

Here we generalize the estimation procedure by introducing different degrees of approximation. That means that we intend to analyze more and more complex local partial

orders, LPOM(i), and derive the corresponding approximations for the averaged rank, $Rkav_{(i)}$ ($i=0,1,2,3,\dots$).

4.2 Basic definitions

Cardinalities

Cardinalities of finite sets are denoted by $|\dots|$. For example:

$N:=|G|$, is the number of elements of G, here the number of chemicals

Important sets

Furthermore the principal order ideals, $O(x)$, principal order filters $F(x)$, generated by $x \in G$ and the set of elements incomparable with x , $U(x)$, are of interest:

$$O(x) := \{y \leq x : y \in G\} \quad (\text{Eq. 5a})$$

$$F(x) := \{y \geq x : y \in G\}. \quad (\text{Eq. 5b})$$

$$U(x) := \{y \parallel x : y \in G\} \quad (\text{Eq. 5c})$$

We call the elements $\neq x$ of $O(x)$ the successors of x , the elements $\neq x$ of $F(x)$ the predecessors of x and the elements $y \in U(x)$ the x -incomparable elements.

Cover relation

Let x, y, z be elements of G . The element z covers x , if and only if $x < z$ and for all y we have that $x \leq y \leq z$ implies y is either x or z . We also say: z covers x or x is covered by z .

Connection

Let $x, y \in G$. If there is (in an ordinary graph theoretical sense) a path from x to y then we call x, y to be connected.

4.3 The S-x-P - construction and U(x) - transformation

4.3.1 The S-x-P - chain

A modified ground set, G' , related to x is introduced:

$$G'(x) := O(x) \cup F(x) \quad (\text{Eq. 6})$$

An order preserving map ϕ is applied, such that

$$\phi(G',IB) \text{ is a linear order.} \tag{Eq. 7a}$$

The linear order $\phi(G',IB)$ is called the **S-x-P** - chain (S ; successors of x , P predecessors of x). $z \in \mathbf{S-x-P}$ is an element of $O(x)$ or $F(x)$ which is mapped by ϕ onto the linear order $\phi(G',IB)$. If $z \neq x$, then z is an element below or above x. We write:

$$\mathbf{S-x-P} := \phi(G',IB) \tag{Eq. 7b}$$

According to incomparabilities in $O(x)$ and $F(x)$, respectively, there can be found many mappings ϕ , which do the job. Based on the experiences in deriving $Rkav_{(1)}(x)$ it is expected that $Rkav_{(1)}$ is mainly depending on $|O(x)|$, $|F(x)|$ and $|U(x)|$. Therefore any arbitrarily selected ϕ might be applied on (G',IB) . However, as Figure 3 exemplifies, the order relation $z \in G'$, $y \in U(x)$ is affected by the specific selection of ϕ .

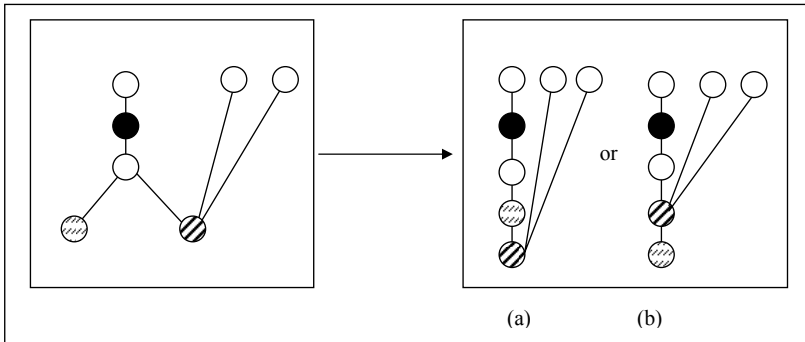


Figure 2: A simple partial order whose **S-x-P** - chain is not uniquely defined. In order to clarify the argument, the elements of G' are differently hatched. The black circle symbolizes the specifically selected element x. The right side shows two possible **S-x-P** - chains.

In the following we assume that one **S-x-P** - chain can be selected.

4.3.2 The $U(x)$ - transformation

We introduce (if necessary) new order relations such that for each single element $y \in U(x)$:

$$y \in U(x) \text{ covers exactly one } z \in \mathbf{S-x-P} \text{ or} \tag{Eq. 8a}$$

$y \in U(x)$ is covered by exactly one $z \in \mathbf{S-x-P}$ or (exclusively) (Eq. 8b)

there may be some

$y \in U(x)$: $y \parallel z$ and for all $z \in \mathbf{S-x-P}$ (these elements are isolated) (Eq. 8c)

We call the element $z \in \mathbf{S-x-P}$ of equation 8a an "lower anchor point" and write $a_y^<$.

We call the element $z \in \mathbf{S-x-P}$ of equation 8b an "upper anchor point" and we write $a_y^>$.

Note that -by assumption- the location of the anchor points within the $\mathbf{S-x-P}$ - chain is well defined.

Considering one single $y \in U(x)$, equations 8a - 8d lead to four cases:

1. $y \in U(x)$ has only one lower anchor point (Eq. 9a)

2. $y \in U(x)$ has only one upper anchor point (Eq. 9b)

3. $y \in U(x)$ is covered by an upper and covers a lower anchor point (Eq. 9c)

4. $y \in U(x)$ is not connected with the $\mathbf{S-x-P}$ - chain, i.e. is isolated at all (Eq. 9d)

Now a partitioning of $U(x)$ can be found due to the following equivalence relation:

$y_1, y_2 \in U(x)$: $y_1 \sim y_2 \Leftrightarrow$ same equation 9_i (one of the four equations)

and the same anchor point(s) (Eq. 10)

Let $U_i(x)$ be the subset of $U(x)$, obeying one of the equations 9 and with the same anchor point(s), then we write for the anchor point a_{U_i} . By equation 10 the set $U(x)$ will be partitioned:

$$U(x) = U_1(x) \oplus U_2(x) \oplus \dots \oplus U_i(x) \oplus U_{i+1}(x) \oplus \dots \oplus U_k(x) \oplus U_{k+1}(x) \oplus \dots \quad (\text{Eq. 11})$$

$i, k \in \{3, 4, \dots\}$

$U_1(x)$: all elements of $U(x)$ having the same lower anchor point a_{U_1}

$U_2(x)$: all elements of $U(x)$ having the same lower anchor points $a_{U_2} \neq a_{U_1}$

...

$U_i(x)$: all elements of $U(x)$ having the same upper anchor point a_{U_i}

$U_{i+1}(x)$: all elements of $U(x)$ having the same upper anchor point $a_{U_{i+1}} \neq a_{U_i}$

...

$U_k(x)$: all elements of $U(x)$ having the same upper and lower anchor points

$U_{k+1}(x)$: all elements of $U(x)$ having the same upper and lower anchor points, but differ in at least one anchor point from the setting of $U_k(x)$.

An example might be helpful (Figure 4):

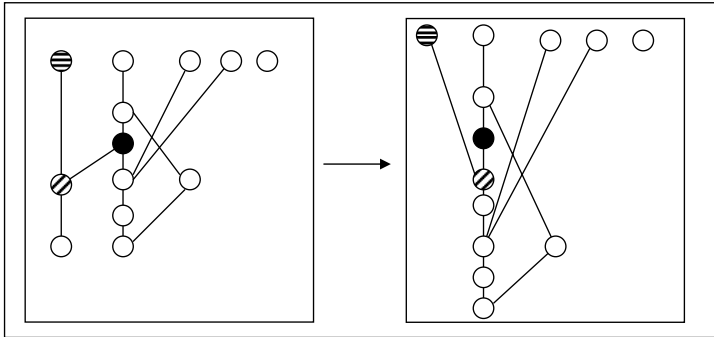


Figure 3: For element x in the empirical Hasse Diagram (left side, x : black circle) a LPOM(i) is to be found. The lower anchor points are arbitrarily located in the chain below x . The upper anchor point gets a unique position, because for all elements $z \in F(x)$ $z \perp a_y^>$. The set $U(x)$ (five elements) can be partitioned into 4 subsets.

4.4 Different Local partial order models (LPOM(i))

In order to estimate the averaged rank by local partial order models the main idea is:

1. to select the element x ,
2. find $O(x)$, $F(x)$ and $U(x)$
3. to apply mapping ϕ (equation 7)
4. to select one **S-x-P** - chain and
5. to perform the $U(x)$ - transformation.

4.4.1 LPOM(0)³

This model assumes

1. the **S-x-P** - construction and
2. $U(x)$ is an empty set.

³ Note that the enumeration scheme of the local partial order models used here deviates slightly from that used in [9].

If there are no elements $y \in U(x)$, then the averaged rank of x in (G, IB) is easily determined: As any linear extension is obtained from an order preserving map, all elements $y \in (O(x) - \{x\})$ will be located below x . Similarly, all elements $y \in (F(x) - \{x\})$ will be located above x . Therefore, the number of elements $\leq x$ is independent of the selected linear extension (i.e. independent of the selected **S-x-P** - chain) and is always $|O(x)|$. Hence we arrive (applying equation 3b) at

$$Rkav_{(0)}(x) = |O(x)| \quad (\text{Eq. 12.})$$

4.4.2 LPOM(1):

This model assumes:

1. the **S-x-P** - construction
2. $U(x)$ is not empty
3. for all $y \in U(x)$, $z \in \mathbf{S-x-P}$ is valid: $y \parallel z$
4. the $U(x)$ transformation, i.e. all elements of $U(x)$ are mutually incomparable

As there are no cover relations between elements of **S-x-P** and elements of $U(x)$, the arbitrary choice of one **S-x-P** - chain does not affect the estimation of the averaged rank. In a previous publication [9] it was shown that then two relations can be derived: The first one is based on the fact that the averaged rank $Rkav_{(1)}(x)$ takes values in the closed interval $[|O(x)|, |O(x)|+|U(x)|]$. $Rkav_{(1)}(x)$ is just the mean of the two limiting cases (in [9] called $Rkav_{(0)}$).

$$Rkav_{(1)}(x) = |O(x)|+|U(x)|/2 \quad x \in G \quad (\text{Eq. 13})$$

However, $Rkav_{(1)}(x)$, calculated by equation 13 is a poor approximation, see for details [9]. The second one takes into account that the averaged rank will not only depend on $|O(x)|$ and $|U(x)|$ but also on the number of predecessors, i.e. on $|F(x)|$. The calculation of $Rkav_{(1)}(x)$ was fully discussed in [9]. However, in order to show the similarity in the arguments, beginning with LPOM(0), stopping with LPOM(3), we sketch the derivation:

The minimum value of $Rkav$ is, as in equation 12:

$$Rkav_{(1)\min}(x) = |O(x)|$$

The maximum value would be

$$Rkav_{(1)max}(x) = |O(x)| + |U(x)|$$

Both equations may be combined by:

$$Rkav_{(1)}(x) = |O(x)| + w \cdot |U(x)| \tag{Eq. 14}$$

The quantity w varies between 0 and 1 and can be interpreted as probability that $|U(x)|$ sees only the positions below x within the **S-x-P** - chain. Hence the probability w can be estimated by the number of positions for $U(x)$ below x divided by the number of all available positions. As a chain, the **S-x-P** - chain is supposed, the probability is given as follows:

$$w = |O(x)| / (|O(x)| + |F(x)|)$$

Within LPOM(1) the final result is therefore:

$$Rkav_{(1)}(x) = |O(x)| + \frac{|O(x)|}{(|O(x)| + |F(x)|)} \cdot |U(x)| \tag{Eq. 15}$$

Equation 15 shows that the estimation of the averaged rank by equation 13 will only lead to reliable results, if $|O(x)| \approx |F(x)|$. The equation 15 implies that the averaged rank $Rkav_{(1)}(x)$ - and by this a linear order - can be deduced from very simple quantities, namely $|O(x)|$, $|F(x)|$ and $|U(x)|$. As in the theory of topological indices, a motive for deriving more sophisticated estimates of the averaged rank is just to reduce the degeneracy, i.e. to get as small equivalence classes due to A (or A_i) due to equation 4. E.g., the reduction of the degeneracy is documented by equation 15 (three variables) in comparison to equation 13 (only one variable).

The LPOM(1) is satisfying in that sense that all further models depend more or less crucially on the assumption that exactly one **S-x-P** - chain is selected. As there are no cover relations between $z \in \mathbf{S-x-P}$ and $y \in U(x)$ the result (equation 15) does not depend on the selection of **S-x-P**.

If such cover relations are to be taken into account, we will at least reduce the problem of degeneracy, however, if the assumptions made are justified is a question which we solved by an empirical study.

4.4.3 LPOM(2)

The derivation of equation 15 is based on the assumption that x -incomparable elements are isolated. This assumption is seldomly justified. For example in the poset, whose Hasse Diagram is shown in Figure 4 all elements of $U(x)$ are connected with one element $z \in O(x)$.

We assume:

1. validity of the **S-x-P** - construction
2. $U(x)$ is not empty
3. the elements of $U(x)$ cover either exactly one lower anchor point or (exclusively) are covered by exactly one upper anchor point.
4. the elements of $U(x)$ are mutually incomparable

The procedure is sketched for the case, where a lower anchor point is assumed.

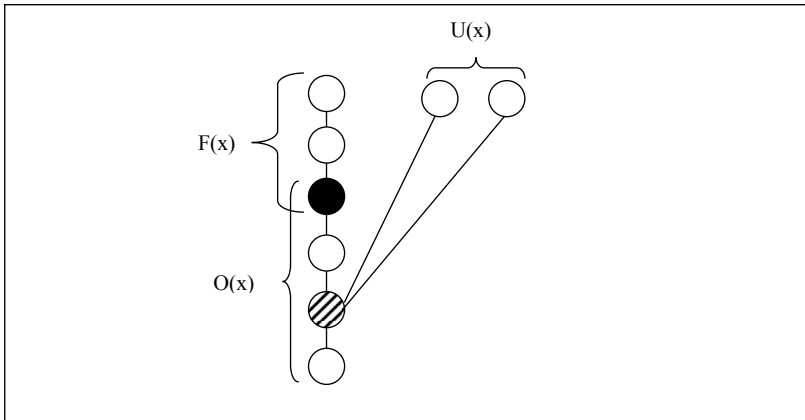


Figure 4: The type of local partial order which is considered in LPOM(2). The element x is symbolized as grey circle. $|F(x)|=3$, $|O(x)|=4$, $|U(x)|=2$. The element y covered by the elements of $U(x)$ is called the anchor point of $U(x)$: $a \prec y$.

As a **S-x-P** - chain is supposed, it is useful to define the interval

$$I(a_y^<, x) = \{z \in \mathbf{S-x-P}: a_y^< < z \leq x\} = (O(x) - O(a_y^<)) \quad (\text{Eq. 16})$$

As there is no partitioning of $U(x)$ we write for the lower anchor point simply a_U .
Similar to the approach in LPOM(1) we set:

$$Rkav_{(2)}(x) = |O(x)| + w \cdot |U(x)|$$

w is the probability to find a location below x and is estimated by:

$$w = \frac{|I(a_U, x)|}{|I(a_U, x)| + |F(x)|}$$

Hence we arrive at:

$$Rkav_{(2)}(x) = |O(x)| + \frac{|I(a_U, x)|}{|I(a_U, x)| + |F(x)|} \cdot |U(x)| \quad (\text{Eq. 17})$$

Now, the averaged rank is not only depending on $|O(x)|$, $|F(x)|$ and $|U(x)|$ but also on the location of the anchor point, here of the lower anchor point. If, several $\mathbf{S-x-P}$ - chains can be found, where the location of the anchor point varies, an uncertainty arises. Future studies have to solve this ambiguity. Beyond this, applying the $\mathbf{S-x-P}$ - and $U(x)$ - transformation on empirical posets such that only one anchor point is allowed, is seldom a good approximation. In the next model the $\mathbf{S-x-P}$ - construction is still maintained and also the $U(x)$ - transformation, however within the $\mathbf{S-x-P}$ - chain there are several anchor points allowed which are related to subsets of $U(x)$.

4.4.4 LPOM(3)

Assumption and steps

1. A $\mathbf{S-x-P}$ - chain will be formed
2. Anchor points can be defined after forming the $\mathbf{S-x-P}$ - chain.
3. $U(x)$ will be partitioned according to equation 11.
4. Elements of $U_i(x)$ are mutually incomparable and obey the four cases, described by equations 9a to 9d.

By LPOM(3) the graph - theoretical structure, regarding the elements of $U(x)$ is considered in more detail. Figure 5 shows two examples. Figure 5 (a): Step 1: The **S-x-P** - construction: The **S-x-P** - chain of this partial order can be uniquely defined. A unique lower anchor point (covered by one element of $U(x)$, hatched diagonally) and a unique upper anchor point (covering two elements of $U(x)$, hatched diagonally) can be found. Here an arbitrary selection was done. Step 2: The upper anchor point is considered as covering two elements of $U(x)$.

Figure 5 (b): Step 1: The **S-x-P** - chain of this partial order **cannot** be uniquely defined. The vertically hatched element of $O(x)$ may be located in several ways. However, only one possibility is shown. Step 2: $U(x)$ - transformation: Partitioning of $U(x)$. Two elements (horizontally hatched) are connected with a lower anchor point, for the other four elements of $U(x)$ (diagonally hatched) there is no comparable element of the **S-x-P** - chain.

Definition of several quantities:

As now by the principle of LPOM(3) only the **S-x-P** - chain and the x -incomparable elements in cover relations are considered we simplify -as before- the notation and introduce the following quantities:

$$I(r,s) := |\{z : z, r, s \in S - x - P \text{ and } s < z < r\}| \quad (\text{Eq. 18a})$$

$$I(r) := |\{z : z, r \in S - x - P \text{ and } z < r\}| \quad (\text{Eq. 18b})$$

$$IT := I(a_2) + I(a_1, a_2) + I(x, a_1) \quad (\text{Eq. 18c})$$

$$n_{\text{tot}} := IT + |U_1(x)| + |F(x)| + 2 \quad (\text{Eq. 18d})$$

Note that the intervals do not contain the interval limits.

In Figure 6b the defined quantities are exemplified: Partitioning is performed due to equation 11. All elements of $U_i(x)$ have the same lower anchor point ($i=1,2$). The black circle symbolizes $x \in G$. Thus, we consider a model system with two sets : $U_1(x) = \{u_{1,1}, u_{2,1}, \dots, u_{k,1}\}$ and $U_2(x) = \{u_{1,2}, u_{2,2}, \dots, u_{k,2}\}$ (only two are shown). All elements of any of these two U_i - sets are considered as mutually incomparable according to the demand of covering in equation 8 ($U(x)$ - transformation).

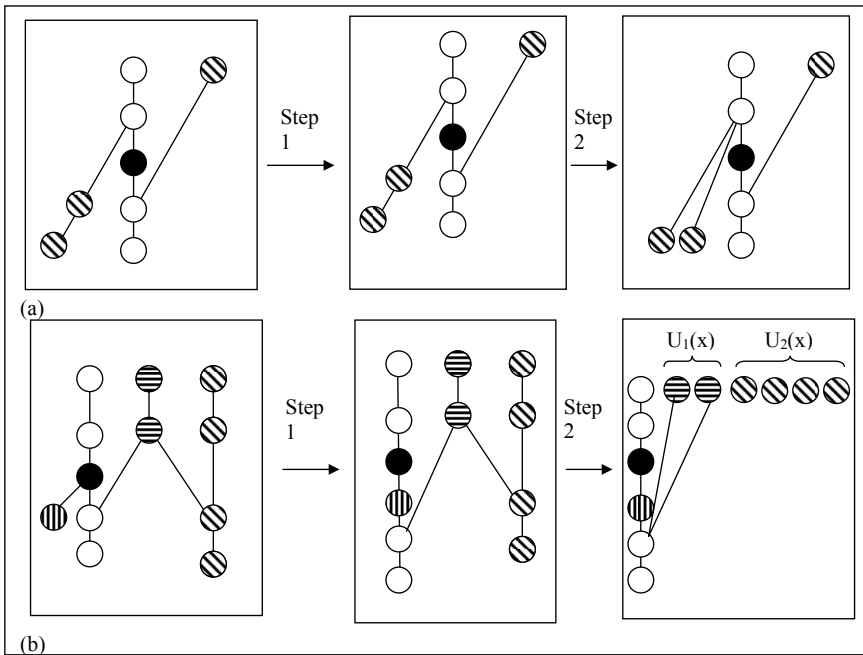


Figure 5 a: Examples of partial orders, where x -incomparable elements are not isolated. The element x is black coloured, the incomparable elements are hatched. The **S-x-P** - construction and the $U(x)$ - transformation are performed in corresponding two steps (see text).

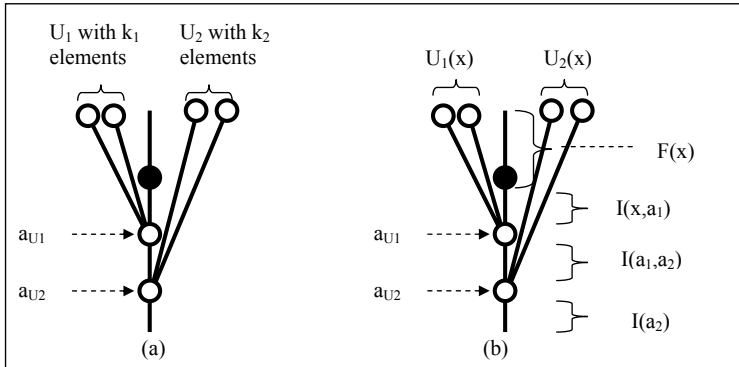


Figure 6 (a): The S-x-P - chain and two subsets of U(x). (b): Explanation of the notation used in equations 18a - d.

Derivation of an estimate of $Rkav_{(3)}$ according to LPOM(3)

Instead of deriving a closed formula of the LPOM(3) we outline the strategy of deriving equations with the example of the Hasse Diagram, shown in Figure 6.

For deriving an equation for $Rkav_{(3)}$ one may start as for the other models:

$$Rkav_{(3)}(x) = |O(x)| + w'_1 \cdot |U_1(x)| + w'_2 \cdot |U_2(x)| \tag{Eq. 19}$$

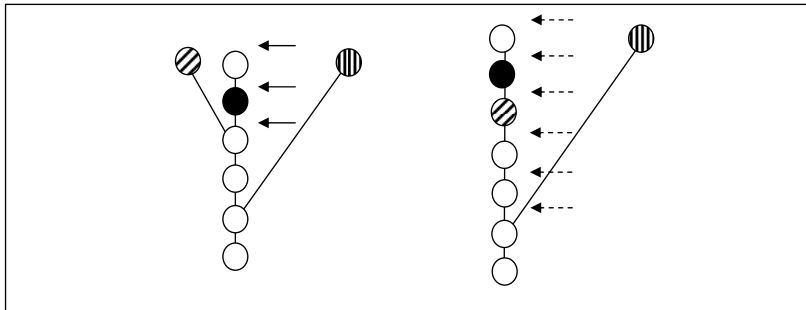


Figure 7: The diagonal hatched element sees three position in the S-x-P - chain. These positions are marked with arrows. If a position of an element below x (the black circle) is taken, then the element vertically hatched sees 4 positions below and two above x (dotted arrows).

Even though, we finally arrive at an equation such as equation 19, this equation hides a trap: The terms w'_1 and w'_2 are not independent from each other: Merging $U_1(x)$ into the **S-x-P** - chain, will change the probability for $U_2(x)$ to be located below or above x , and the other way round. By Figure 7 the problem of mutual influence of the $U_i(x)$ -subsets is demonstrated. In seeking for positions of $U_1(x)$ within the **S-x-P** - chain there are two possibilities:

- (1) $U_1(x)$ sees **successors** of x and locating $U_1(x)$ below x may have the probability w_1 ; and
- (2) $U_1(x)$ sees **predecessors** of x . The probability to locate the $U_1(x)$ above x is w_2 .

After locating the $U_1(x)$ - set in one of the accessible parts above or below x in the **S-x-P** - chain, the $U_2(x)$ - set sees an extended **S-x-P** - chain. The location of the $U_2(x)$ -set depends now on how many positions are available below or above x . For example, if the extended **S-x-P** - chain is found by locating the $U_1(x)$ set below x (with probability w_1) then the $U_2(x)$ -set has the probability w_{11} also to be located below x . The probability w_{11} must take into account the number of positions below x , in comparison to all -now available- positions, by counting the additional positions of $U_1(x)$ too. The probability that both events, $U_1(x)$ below x and $U_2(x)$ below x in the extended **S-x-P** - chain appear, is the product $w_1 * w_{11}$. Therefore an event-tree is the appropriate graphical scheme to describe the process of locating subsets of $U(x)$ within the **S-x-P** - chain, see Figure 8. The number of positions for merging $U_2(x)$ depend on the step done before, i.e. by which $U_1(x)$ is positioned within the **S-x-P** - chain. According to the below- x , above- x - consideration and assuming only two U_i -subsets there are four possible ranks and correspondingly four probabilities to get these ranks (equations 20a-d):

$$\text{prob} [\text{rk}(x) = \text{IT}+3+|U_1(x)|+|U_2(x)|] = w_1 * w_{11} \quad (\text{Eq. 20a})$$

$$\text{prob} [\text{rk}(x) = \text{IT}+3+|U_1(x)|] = w_1 * w_{12} \quad (\text{Eq. 20b})$$

$$\text{prob} [\text{rk}(x) = \text{IT}+3+|U_2(x)|] = w_2 * w_{21} \quad (\text{Eq. 20c})$$

$$\text{prob} [\text{rk}(x) = \text{IT}+3+0] = w_2 * w_{22} \quad (\text{Eq. 20d})$$

The probabilities w_i are easily calculated as only the available positions for the $U_1(x)$ -set above or below x in the **S-x-P** - chain are to be counted and divided by the count of all accessible positions in the original obtained **S-x-P** - chain. Similarly the probabilities w_{ij} for the location of the $U_2(x)$ are to be calculated in counting the available positions above and

below x in the extended **S-x-P** - chain. Compare the event tree, shown in Figure 8 and Table 1 for the relevant equations.

Table 1: Calculation of w_i, w_{ij}

First step	Second step
$w_1 = (I(x, a_{k1}) + 1) / (I(x, a_{k1}) + F(x) + 1)$	$w_{11} = (I(a_1, a_2) + I(x, a_1) + U_1(x) + 2) / n_{tot}$
	$w_{12} = F(x) / n_{tot}$
$w_2 = F(x) / (I(x, a_{k1}) + F(x) + 1)$	$w_{21} = (I(a_1, a_2) + I(x, a_1) + 2) / n_{tot}$
	$w_{22} = (F(x) + U_1(x)) / n_{tot}$
$n_{tot} = I(a_1, a_2) + I(x, a_1) + U_1(x) + F(x) + 2$	

Combining the information, given in Table 1 and that of equations 20 a-d, one arrives at the final equation to calculate $Rkav_{(3)}$ and the system, shown in Figure 6

$$\begin{aligned}
 Rkav_{(3)}(x) = & w_1 \cdot w_{11} \cdot (|O(x)| + |U_1(x)| + |U_2(x)|) \\
 & + w_1 \cdot w_{12} \cdot (|O(x)| + |U_1(x)|) + w_2 \cdot w_{21} \cdot (|O(x)| + |U_2(x)|) \quad \text{Eq. (21)} \\
 & + w_2 \cdot w_{22} \cdot |O(x)|
 \end{aligned}$$

Rearranging the equation 21:

$$Rkav_{(3)}(x) = (w_1 \cdot w_{11} + w_1 \cdot w_{12} + w_2 \cdot w_{21} + w_2 \cdot w_{22}) \cdot |O(x)| + \dots$$

The sums $w_{11} + w_{12}$, $w_{21} + w_{22}$ and $w_1 + w_2$ equal 1, therefore we arrive at:

$$Rkav_{(3)}(x) = |O(x)| + w_1 \cdot |U_1(x)| + (w_1 \cdot w_{11} + w_2 \cdot w_{21}) \cdot |U_2(x)| \quad \text{(Eq. 22)}$$

Hence, comparing with equation 19 $w_1' = w_1$, however $w_2' = (w_1 w_{11} + w_2 w_{21})$ is a little bit more complicated expression, taking into account that merging of $U_2(x)$ depends on the manipulation of $U_1(x)$. We observe that obviously the order of merging of the two subsets $U_1(x)$ and $U_2(x)$ respectively is important. The reason is that the merging of $U_1(x)$ and $U_2(x)$ into the **S-x-P** - chain should be done simultaneously or at least by merging the elements of $U(x)$ one by one into the **S-x-P** - chain. However, by deriving equation 21 the merging was

done stepwise, "en bloc" for any subsets of $U(x)$. Empirically we found that better results are found if first that $U(x)$ -subset should be merged, which "sees" more positions below x . More complicated event trees can be analyzed, if other model systems under the premises of LPOM(3) are of interest. For example a LPOM(3) like that shown in Figure 9 would need an event tree with four steps, because four $U(x)$ - subsets can be found, whereas local partial orders like those shown in Figure 6 can be represented by an event tree with two steps (Figure 8).

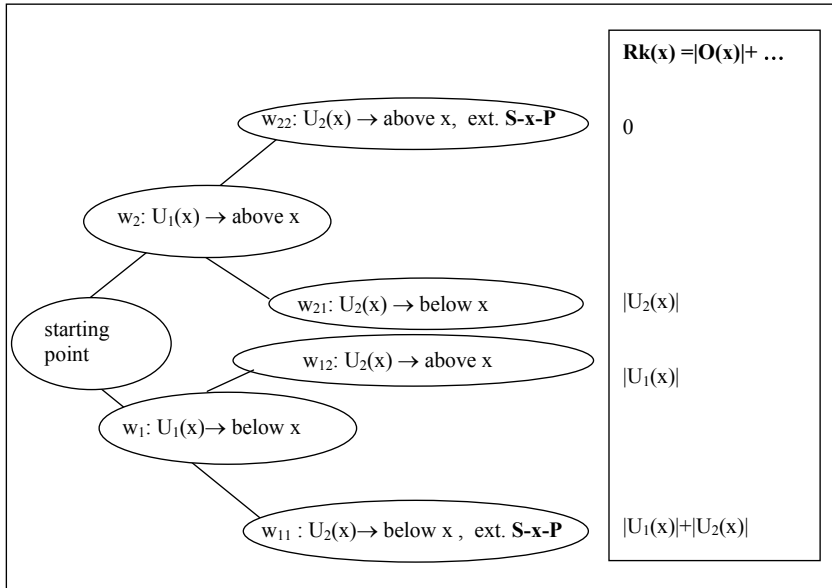


Figure 8: Event tree. Beginning with the "starting point": Lines downwards are associated with merging an $U_i(x)$ subset into the part below x and hence enhancing the averaged rank, lines upwards are associated with merging into the part above x . "ext. **S-x-P**": by $U_1(x)$ extended **S-x-P** - chain.

4.5 Validation study

A validation study was performed, taken from a series of Hasse Diagrams of the type shown in Figure 6 with varying $|U_1(x)|$, $|U_2(x)|$, $S(x, a_1)$, $S(a_1, a_2)$, $S(a_2)$. As for any element of the Hasse Diagrams the averaged rank can be calculated there are in the total 93 test cases.

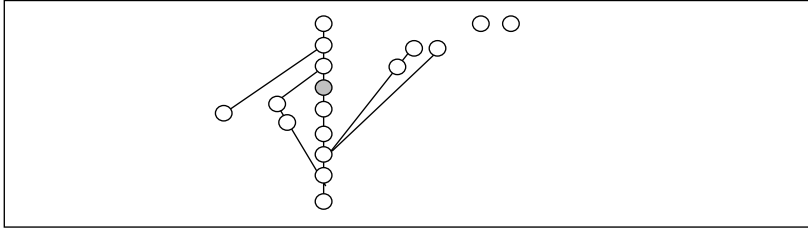


Figure 9: The local partial order found for x (grey circle) contains one U-subset (two elements) having an upper and a lower anchor point, one U-subset (only one element) having exclusively one upper anchor point, one U-subset (three elements) having one lower anchor point and finally one U-subset of isolated elements at all.

No one of these test Hasse Diagrams had more than 17 elements in order to be able to calculate the exact values $Rkav_{\text{exact}}$ by applying the software WHASSE [11]. The results are shown in Table 2. Although the correlation coefficients r^2_{DF} are very good in both cases, the dramatically improved F-value and especially the small value for t in case of $Rkav_{(3)}$ show the improvements. As even for partial orders of the type shown in Figure 6 the equation 19 has to be considered as an approximation, the validation study shows promising results. The "en bloc" merging as a crucial point of approximation is at least for partial orders of moderate size acceptable.

Table 2: Comparison of the equation for $Rkav_{(3)}$ (equation 21) with the $Rkav_{(1)}$ (equation 15). $Rkav_{(3)} = s \cdot Rkav_{\text{exact}} + t$, estimation of s and t by SPSS^(R)

	cases	N	$ U_1(x) + U_2(x) $	r^2_{DF}	t	s	F-statistics
$Rkav_{(3)}$	93	≤ 17	≤ 7	1.00	-1.2E-5	1.00	1,8E10
$Rkav_{(1)}$	93	≤ 17	≤ 7	0.959	-1.99	1.102	2172.5

5 The "real life" example.

For river management purposes, the chemical pollution of the river Main, Germany (Bavarian part) was investigated. Typically in water samples polycyclic aromatics (PAHs), polychlorinated biphenyls (PCBs) and some volatile chemicals were detected (Table 3). The steady state model EXWAT [13], as part of the evaluative model package E4CHEM [13, 14]

was applied to derive fate descriptors. These were the degree (scores) of (i) sedimentation, (ii) volatilization, (iii) persistence and (iv) transport down streams (see appendix). The diagram, discussed here, is a very simple one and averaged ranks are considered here just for demonstration of the method (nevertheless convenient for decision makers). Note that in [14] a Hasse Diagram, of the river Main pollution is shown, which includes more chemicals.

After selecting an element $x \in G$ of interest, instead the empirical Hasse Diagram like that shown in Fig. 10 a local diagram around x is considered. If $|G|$ elements are in the ground set G , then -at the maximum- $|G|$ different local model diagrams are to be analyzed. In the most simple approximation, namely the $Rkav_{(1)}$, each local Hasse Diagram consists of a chain, containing the actually selected x object and some objects incomparable to x (x -incomparable elements or objects) which however are considered as isolated. In Figure 11 the local Hasse Diagrams (LPOM(1)) of Fluoroanthene, fl and Chloroforme, ch, are shown.

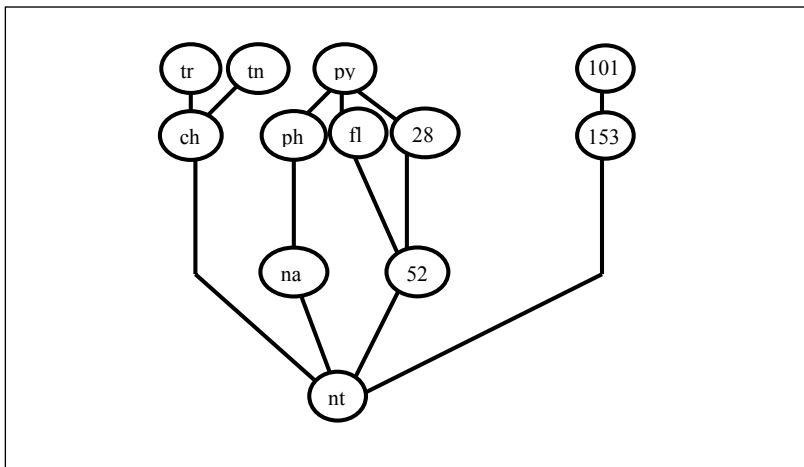


Figure 10: Hasse Diagram of 12 chemicals, found in river Main, Germany. The abbreviations are explained in Table 3.

Table 3: Examples of chemicals found in the river Main. Classification: CCl: small chlorinated hydrocarbons, PAH: polyaromatic hydrocarbons, PCB: polychlorinated biphenyls, OTH: molecules else (see appendix for the data matrix)

abbreviation	classification	name	abbreviation	classification	name
tr	CCl	Trichloroethene	nt	OTH	Nitriloacetic acid (NTA)
tn	CCl	Trichloroethane	28	PCB	PCB 28
ch	CCl	Chloroforme	52	PCB	PCB 52
py	PAH	Pyrene	101	PCB	PCB 101
ph	PAH	Phenanthrene	153	PCB	PCB 153
na	PAH	Naphthalene	fl	PAH	Fluoroanthene

From Figure 10 we exemplify for the chemicals fl and ch the derivation of the characteristics, which are needed in the subsequent calculation of $Rkav_{(1)}$ according to LPOM(1), see Figure 11.

$$\begin{aligned}
 U(\text{fl}) &= \{\text{na, ch, tr, tn, 28, 101, 153, ph}\} & U(\text{ch}) &= \{\text{na, ph, fl, 52, 28, 101, 153, py}\} \\
 |U(\text{fl})| &= 8 & |U(\text{ch})| &= 8 \\
 |O(\text{fl})| &= 3 & |O(\text{ch})| &= 2 \\
 |F(\text{fl})| &= 2 & |F(\text{ch})| &= 3
 \end{aligned}$$

Note that in reality (i.e. in the empirical Hasse Diagram, Figure 10) all elements of $U(\text{fl})$ and $U(\text{ch})$ respectively are connected (in the sense of ordinary graphs) with fl and ch, respectively. Therefore the procedure explained in section 4, especially the approach of LPOM(3) (section 4.4.4) is to be applied. Instead of LPOM(1) the local partial order is as shown in Figure 12. Three different approximations, i.e. local partial order models were selected as the best matching for the actual element x : LPOM(1), LPOM(2) and LPOM(3). Additionally by the software program WHASSE the exact values for the averaged ranks for all chemicals were calculated. We begin with showing the results of LPOM(1). The results of LPOM(1) in comparison with exact averaged ranks, $Rkav_{\text{exact}}$ are summarized in Table 4.

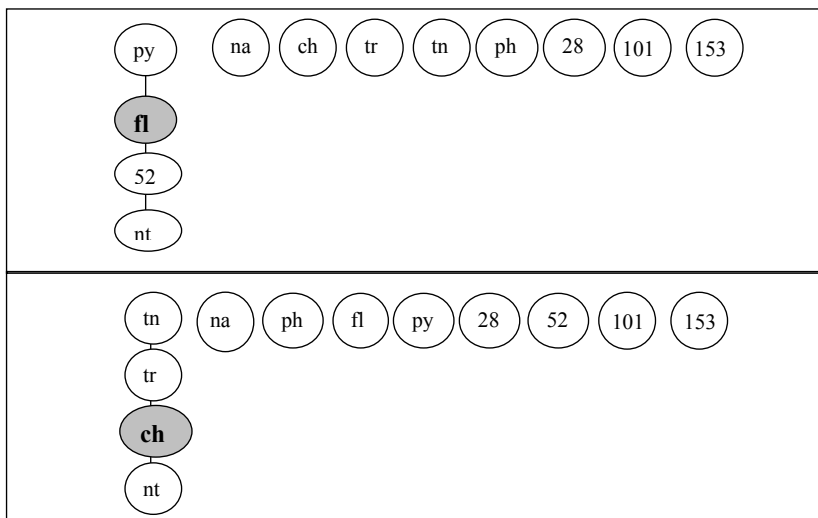


Figure 11: Local Partial Order , LPOM(1) with respect to the chemical Fluoroanthene, fl (top), LPOM(1) with respect to chemical chloroforme, ch (bottom). In both cases there are 8 incomparable elements. Note that in the original HD (Fig. 10) $tn \parallel_{IB} tr$. The index IB refers to the original poset, induced by IB. These two elements are now considered as comparable, here we arbitrarily selected $tn > tr$ (as here no anchor point is affected, this ambiguity is not relevant).

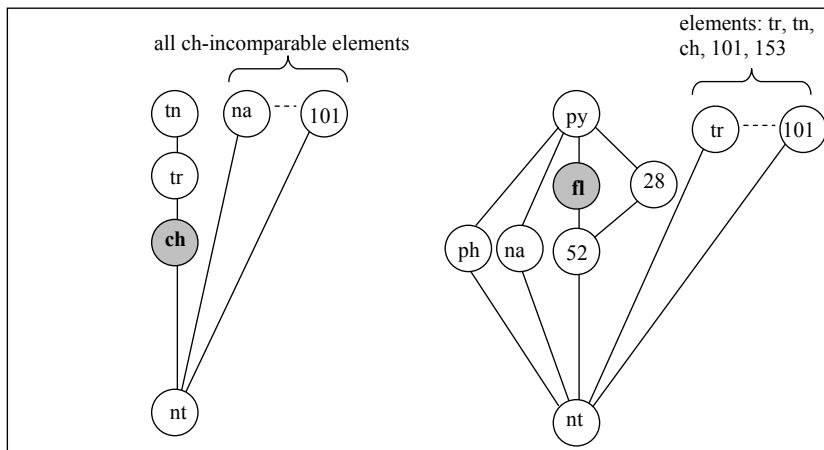


Figure 12: Other local partial order models for chemicals ch and fl (see text).

For calculating approximately the averaged rank of the chemical ch, the LPOM(2) is sufficient, whereas for the estimation of the averaged rank $Rkav(fl)$ the approximation due to LPOM(3) may be more suitable. Here the event tree will contain three steps, because $U(fl)$ is partitioned into three subsets, namely $U1(fl) = \{ph, na\}$, $U2(fl) = \{28\}$ and $U3(fl) = \{tr, tn, ch, 101, 153\}$ according to the $U(x)$ - transformation (section 4.3.2).

The individual deviations from the exact values can be up to 1.5 as the example of chemical 153 shows. Figure 13 displays the results of a comparison of $Rkav_{exact}$ with $Rkav_{(1)}$:

Table 4: Real life example, $Rkav_{exact}$ (calculated directly from WHASSE)

abbreviation	$Rkav_{exact}$	$Rkav_{(1)}$	abbreviation	$Rkav_{exact}$	$Rkav_{(1)}$
nt	1	1.0	ph	7.86	7.8
na	4.43	5.2	fl	7.43	7.8
52	3.57	4.33	28	7.43	7.8
ch	4	5.2	py	11.29	11.375
tr	8.5	9.75	153	5	6.5
tn	8.5	9.75	101	9	9.75

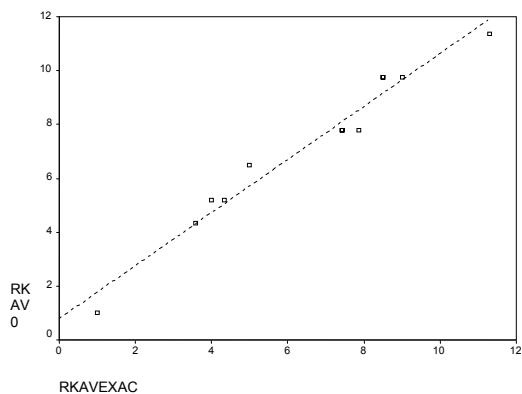


Figure 13: Scatter diagram of $Rkav_{(1)}$ (Ordinate) and $Rkav_{exact}$. The dotted line is the regression line, with $r^2_{DF} = 0.963$. See below.

The statistical results of the regression analysis are quite well. However, all $U(x)$ -elements are connected with any $x \in G$, hence the estimation by LPOM(1) is questionable. In Table 5 the

results can be seen if the methodology, outlined in section 4 is applied for LPOM(1), and for LPOM(3) with and without the restriction of a presentation by a partial order system like that shown in Figure 6. As a measure of quality of the different approximations, the regression

$$Rkav_{\text{exact}} = s * Rkav_{(3)} + t \quad (\text{Eq. 23})$$

is analyzed. Ideally s should be 1 and $t=0$. The regression coefficient r^2_{DF} should be 1. The F-statistics should have as large values for F as possible.

The estimation of the coefficients and statistical data are taken from SPSS^(R). As one can see the best results are obtained, if the full methodology of LPOM(3) is selected in order to find that partial order which matches best the structure around the actual interesting chemical.

6 Discussion and Conclusions


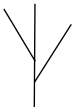
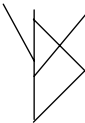
A methodological scheme was presented, how to derive equations for local partial order models. Clearly the most crucial approximations were

1. the **S-x-P** - construction which leads to a linear order for (G', IB) (see section 4.3), and
2. the U(x) - transformation which simplifies drastically the order relations among the elements of U(x).

Both approximations were found to be imperative to find a procedure to estimate the average rank. The sequence of local partial order models can be extended to LPOM(4), LPOM(5) etc. to take into account more and more details of the empirical partial order. This task, however should be solved in cooperation with other scientific groups.

Another problem is the following (see also section 4.4.4): If U(x) is partitioned into several subsets, then the problem arises in which order the merging process has to be performed. This problem might be solved, if the approach of merging the elements of the U(x)-subsets "en bloc" is given up. The alternative might be to merge the elements one after another of the whole U(x)-set. As this aspect was and will be of much concern it will be explained in more detail here:

Table 5: Statistical data for the real life example. According to 12 chemicals, twelve averaged ranks are estimated by LPOM(3), compared with the exact values (using the software WHASSE) and with the results of LPOM(1). The comparison was done by applying equation 23.

Model system	Scheme	r_{DF}^2	s	t	F
LPOM(1)		0.963	0.981	-0.558	284
LPOM(3) Only LPOM selected like those shown in Figure 6		0.982	0.963	0.290	603
Full methodology of LPOM(3)		0.995	1.002	-0.127	1995

Considering the merging of subsets of $U(x)$ into the **S-x-P** - chain, we implicitly assumed that the elements of a $U(x)$ -subset can be located "en bloc" below or above x in the **S-x-P** - chain. Figure 14 and the subsequent text shows the problem:

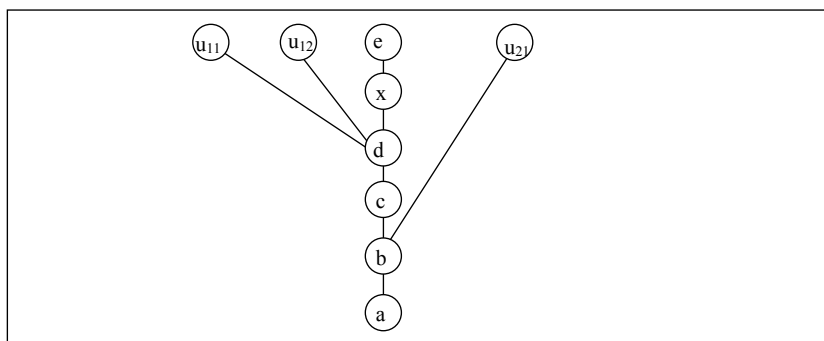


Figure 14: Example, to explain the "en bloc" distribution (see text)

In the Hasse Diagram of Figure 14, the anchor points of $U_1(x) = \{u_{11}, u_{12}\}$ is d and that of $U_2(x) = \{u_{21}\}$ is b . As only the part $d < x < e$ of the **S-x-P** - chain in Fig. 14 is of interest the distribution "en bloc" of $U_1(x)$ is demonstrated only for this part:

"en bloc":

$$d < \{u_{11}, u_{12}\} < x < e$$

$$d < x < \{u_{11}, u_{12}\} < e \text{ and}$$

$$d < x < e < \{u_{11}, u_{12}\}$$

The notation "{...}" within the sequences above means that in the subsequent estimation of the rank of x , the elements u_{11}, u_{12} will be counted independently, whether $u_{11} < u_{12}$ or $u_{12} < u_{11}$. In contrast to this "en bloc" merging of $U(x)$ -subsets into the **S-x-P** - chain one may merge the single elements of $U(x)$ until $U(x)$ is exhausted. In contrast to "en bloc" - merging we call this a "one by one" merging. Once again Figure 14 serves as example:

"one by one":

$$d < u_{11} < u_{12} < x < e$$

$$d < u_{11} < x < u_{12} < e$$

$$d < u_{11} < x < e < u_{12}$$

$$d < x < u_{11} < u_{12} < e$$

$$d < x < u_{12} < u_{11} < e$$

$$d < x < u_{11} < e < u_{12}$$

$$d < x < e < u_{11} < u_{12}$$

etc.

At least in the case of LPOM(1) it could be shown that both methods, the "en bloc" - and the "one by one" merging lead to the same results [9, 15]. However, that simple local partial order model does not give rise to partition the $U(x)$ -set. This combinatorial exercise we hope to solve in the near future.

Taking into account such an extent of approximations and assumptions one may question the general value of the local partial order model. The answer to the question whether local partial order models are helpful or too overloaded with assumptions is difficult: Empirically there is an improvement in a statistical sense. However, on the one side this finding depends on the

sample of empirical posets which was analyzed, on the other side a statistical improvement does not necessarily guarantee that the $Rkav_{(i)}$ -values induce the same linear order as the exact averaged ranks, obtained by the WHASSE software. The case that $Rkav_{(i)}(x) = Rkav_{exact}(x)$ would only be the case, when the regression equation 23 has ideal statistical characterizations. It is planned to develop a more stringent graph-theoretical setting. Which empirical posets can now be safely handled? Empirical posets can be safely handled if x -incomparable elements within the empirical poset are isolated. In that case the structure of the order ideal $O(x)$ or that of $F(x)$ can vary without affecting the numerical result of $Rkav_{(1)}$.

If the upper anchor points are comparable with all other elements of $F(x)$ or the lower anchor points are comparable with all other elements of $O(x)$ then the **S-x-P** - construction may not be unique, albeit the anchor point(s) can uniquely located within the **S-x-P** - chains. Hence the number of elements within the intervals $I(x, a_1)$, $I(a_1, a_2)$, $I(a_2)$ is uniquely defined. Therefore only the approximation due to the $U(x)$ - transformation remains.

The derivation of approximate equations to calculate the averaged rank enables us to analyze several types of posets without arbitrariness in the **S-x-P** - construction and $U(x)$ - transformation. Hence we get an impression what the leading factors for the averaged rank are. Indeed it is evident that:

- chain length of **S-x-P** and
- distance between anchor points

play a main role. The derivation of approximate equations for averaged ranks helps to get a linear order from a Hasse Diagram. Thus the crucial determination of weights, as needed in most other evaluation procedures can be avoided. However the question remains, which local partial order should be selected to calculate the averaged rank? A program written in the interpreter language PYTHON [16] is in preparation, by which several local partial order models can be selected to perform further studies in order to find an answer.

References

[1] Friege, H. (2002) Das EU-Weißbuch zum Umgang mit Stoffen: Chancen und offene Fragen bei der Umsetzung. UWSF - Z.Umweltchem.Ökotox. 14, 254

- [2] Larsen, H.F., M. Birkved, M. Hauschild, D.W. Pennington, and J.B. Guinée (2004). Evaluation of Selection Methods for Toxicological Impacts in LCA - Recommendations for OMNIITOX. *Int J. LCA* 9, 307-319
- [3] Ganter, B. and R. Wille (1996) *Formale Begriffsanalyse Mathematische Grundlagen*. Springer-Verlag, Berlin
- [4] Brüggemann, R., E. Halfon, G. Welzl, K. Voigt and C. Steinberg (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
- [5] Winkler, P. (1982) Average height in a partially ordered set. *Discr. Math.* 39, 337-341
- [6] Lerche, D. and P. Sørensen (2003) Evaluation of the ranking probabilities for partial orders based on random linear extensions. *Chemosphere* 53, 981-992.
- [7] Schneeweiss, C. (1991) *Planung 1 - Systemanalytische und entscheidungstheoretische Grundlagen*. Springer-Verlag, Berlin
- [8] Trotter, W.T. (1992) *Combinatorics and Partially Ordered Sets Dimension Theory*. The Johns Hopkins University Press, Baltimore, Maryland
- [9] Brüggemann, R., P. Sørensen, D. Lerche and L. Carlsen (2004) Estimation of Averaged Ranks by a Local Partial Order Model. *J.Chem.Inf.Comp.Sc.* 44, 618-625
- [10] Brüggemann, R. and H.-G. Bartel (1999) A Theoretical Concept to Rank Environmentally Significant Chemicals. *J.Chem.Inf.Comp.Sc.* 39, 211-217
- [11] Brüggemann, R., C. Bücherl, S. Pudenz, and C. Steinberg (1999) Application of the concept of Partial Order on Comparative Evaluation of Environmental Chemicals. *Acta hydrochim. hydrobiol.* 27, 170-178
- [12] Atkinson, M.D. (1990) On the computing the Number of Linear Extensions of a Tree. *Order* 7, 23-25

[13] Brüggemann, R., S. Trapp and M. Matthies (1991) Behavior Assessment for a Volatile Chemical in the Middle and Lower German Part of the Rhine River. *Envir.Tox.Chem.* 10, 1097-1103

[14] Brüggemann, R. and U. Drescher-Kaden (2003) Einführung in die modellgestützte Bewertung von Umweltchemikalien - Datenabschätzung, Ausbreitung, Verhalten, Wirkung und Bewertung, 1 edn. Springer-Verlag, Berlin

[15] Brüggemann R. (2004) Lokale partielle Ordnungen, ein Hilfsmittel für die Bewertung? In: Wittmann, J. and Wieland, R. (eds) *Simulation in Umwelt- und Geowissenschaften - Workshop Müncheberg*. Shaker Verlag, Aachen

[16] Von Löwis, M. and N. Fischbeck (2001) *Python 2 - Einführung und Referenz der objektorientierten Skriptsprache*, Addison-Wesley, München, 1-605 pp

Appendix: Data matrix of 12 chemicals.

The indicators are: Vol: score of volatilization flux, Sed: score of sedimentation flux, Pers: score of persistence, Adv: score of advective flux down streams

Chemicals	Vol	Sed	Pers	Adv	Chemicals	Vol	Sed	Pers	Adv
na	3	2	2	3	tn	4	1	2	3
ph	3	2	2	4	tr	4	2	2	2
py	3	3	2	4	28	3	3	2	2
fl	2	3	2	4	52	2	3	2	2
nt	1	1	0	1	101	2	4	2	1
ch	4	1	2	2	153	1	4	2	1