

# Rule Generation by Means of Lattice Theory

R.Brüggemann<sup>1</sup>, S.Pudenz<sup>1</sup>, H.-G.Bartel<sup>2</sup>

1) Institute of Freshwater Ecology and Inland Fisheries, Dep. of Ecohydrology  
Rudower Chaussee 6a  
D-12484 Berlin  
brg@igb-berlin.de

2) Institute of Chemistry  
Department of Analytical and Environmental Chemistry  
Humboldt University Berlin  
Hessische Str. 1-2  
D-10115 Berlin  
hans-georg=bartel@rz.hu-berlin.de

## Abstract

A rather small data matrix of seven chemicals and 17 different ecotoxicological end points is examined by lattice theory. Especially the Formal Concept Analysis, developed since 1980 by the school of Wille is applied. The heart of this mathematical method is the concept, consisting of a pair of sets, which are related to each other (Galois Connection). The one set is a subset of objects, the other a subset of properties. The concepts are partially ordered due a subset relation among those sets. From the subset relation implications are derived. Here -after giving some reading examples- implications are discussed. For example the following question can be answered:

Is there a chemical that has a high ecotoxicological effect for example on both *Tanytarsus diss.* (insect) and *Rana catesbeiana* (amphibium)? Our chemical set has no one, which has these properties in common. Other examples are discussed.

## Introduction

In our paper we concentrate on the purely qualitative point of view, i.e. we would like to generate „if-.then.- rules“. Ecotoxicological test data designed to evaluate the load status of the aquatic environment are of specific interest. The premises of such rules should be structural properties of the chemicals. The conclusions should be statements about the order of amount of toxicities. A lack of numerical accuracy may be compensated for by the potency to derive many rules.

The tool to generate such rules is derived from lattice theory, which is a rather new discipline of Discrete Mathematics (since around 1930) (Birkhoff 1984, Davey and Priestley 1990). In particular the variant of Formal Concept Analysis, a sub discipline of lattice theory dating from around 1980, turns out to be useful in deriving such rules in an automatic, i.e. algorithmic way (Davey and Priestley 1990, Ganter and Wille 1996). One of the main advantages of this method is its connection to graph theory and through this the possibil-

ity of visualizing rather complex relationships in high-dimensional spaces. Some first steps to apply lattice theory in environmental sciences are already published, see e.g. Brüggemann, Voigt and Steinberg (1997).

## Data

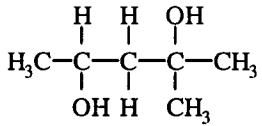
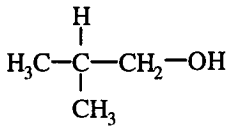
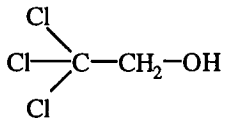
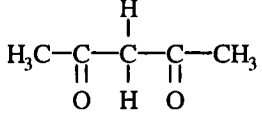
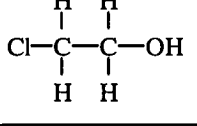
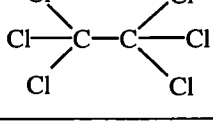
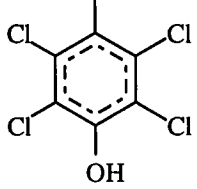
Seven chemicals and 17 test species of different trophic levels are analyzed (see table 1). The species are commonly used in the frame of environmental protection studies. From

Table 1: Test species of different trophic levels taken from Devil-lers, Thioulouse and Karcher (1993).

Species	Identifier	Type
<i>Daphnia magna</i> (LC50)	A	Crustacea
<i>Tanytarsus diss.</i>	B	Insect
<i>Orconectes immunis</i>	C	Crustacea
<i>Rana catesbeiana</i>	D	Amphibium
<i>Oncorhynchus mykiss</i>	E	Fish
<i>Lepomis macrochirus</i>	F	Fish
<i>Gambusia affinis</i>	G	Fish
<i>Ictalurus punctatus</i>	H	Fish
<i>Carassius auratus</i>	I	Fish
<i>Pimephales promelas</i>	J	Fish
<i>Arbacia punctulata</i> (early embryo)	K	Echinoderm
<i>Arbacia punctulata</i> (sperm cell)	L	Echinoderm
<i>Photobacterium phosphoreum</i>	M	Bacterium
<i>Arbacia punctulata</i> (early embryo, DANN)	N	Echinoderm
<i>Daphnia magna</i> (EC50)	O	Crustacea
<i>Daphnia pulex</i>	P	Crustacea
<i>Ceriodaphnia reticulata</i>	Q	Crustacea

a systematic point of view, more different species per group would be favored; however the data availability for such an ambitious aim is rather poor. Devillers, Thioulouse and Karcher (1993) selected mainly the chemicals by their different physiological mechanisms. There are oxy- or halogenated compounds (see table 2). The full data matrix can be examined by looking into the original paper. We call the chemicals the objects of the analysis and the set of objects  $G$ . The toxicological data are called the attributes of each object. The attribute set is called  $M$ .

Table 2: Seven chemicals and their identifier.

Chemical	Identifier
 -Methyl-2,4-pentadiol	$\alpha$
 2-Methyl-propanol	$\beta$
 2,2,2-Trichloroethanol	$\gamma$
 4-Pentadione	$\delta$
 2-Chloroethanol	$\epsilon$
 Hexachloroethane	$\zeta$
 Pentachlorophenol	$\eta$

In order of robustness an attribute-wise classification is performed. Each attribute is divided into five intervals as follows (table 3):

Table 3: Classification of the attribute values.

Interval N°	attribute-interval
1	0.68 - 1.83
2	1.89 - 3.00
3	3.09 - 4.02
4	4.24 - 5.06
5	5.22 - 6.52

Instead of the raw measured data  $p_i$ , each object  $x$  is now ecotoxicologically characterized by a 17-tuple

$$n(x) = (n_1, n_2, \dots, n_{17})$$

corresponding to the 17 different species/end points. The quantities  $n_i$  are integer numbers, as defined by table 3. A simplifying notation can be introduced: <value in terms of  $n_i$ ><identifier for the species>.

### Lattice theory

For introductory literature the reader is referred to Birkhoff (1984), Davey and Priestley (1990). Consider a set of objects  $G$ , then binary relations among the elements of this set may be considered. For example the order relation:

An order relation has to be

- reflexive
- antisymmetric and
- transitive.

For example the comparison of two numbers follows these three axioms. Sets equipped with an order relation are called partially ordered sets (posets).

For any two objects one may look for upper and lower bounds. Simplified spoken if such bounds exist and there is an unique upper and lower bound resp. then the mathematical structure of a lattice arises.

Here we are interested in lattices of specific constructions, namely lattices of Formal Concept Analysis (FCA) (Ganter and Wille 1996).

The basic notion of FCA is a triple  $(G, M, I)$  called formal concept.  $G$  is the set of objects under consideration and  $M$  the set of their (original or suitable transformed) attributes. An element  $gIm$  ( $g \in G, m \in M$ ) of the binary relation  $I \subseteq G \times M$  ( $gIm \equiv (g, m) \in I$ ) is read: "Object  $g$  has attribute  $m$ ". Another form for representing a context  $(G, M, I)$  is a cross table. Its rows are labeled by the objects  $g \in G$  and its columns by the attributes  $m \in M$ . A "X" at the position of cross-



The third rule however has not a firm basis because its premise is only fulfilled by one compound, which is the only one in the whole set of substances.

The problem becomes more evident by another rule:

aromatics  $\Rightarrow$  -Cl

This is true within the actual chosen set but can clearly not be generalized. The rules are extracted from a finite set of objects and attributes, therefore the generation of rules should better be seen as an automatic hypotheses-generation.

## Discussion

A data matrix is converted into a HASSE diagram of concepts by several steps (Classification, extending by structural codes, forming simple valued context). From the HASSE diagram and the subset-superset relation implications can be derived. Among the many implications (the total of 63 rules is the outcome of that simple data matrix), which relate the attributes, those are of primary interest, which relate structural information with ecotoxicological end points.

There are some distinct drawbacks:

Clearly the substance basis is still by far too weak to establish real rules; however the method allows in principle to do this, as was shown with a few examples.

The real measured data are converted into discrete values without any statistical guidance. Thus there is no safety against slight changes in the data transformation. This however could be done and is not on the focus of this paper.

The HASSE diagram may consist of a complex system of lines and therefore difficult to read. However, there are techniques to circumvent such unreadable diagrams.

One may argue that the type of ecotoxicological end point was not discussed so far. However here some results are given in Bartel (1999), which are more or less methodological and which are therefore suppressed here.

What are the advantages?

Clearly if a graphical presentation is possible, it is favoured against boring tables. As pointed out there can be performed many interesting studies, which are as more non trivial as more not only single relations like  $gIm$  are examined but relations like  $G'I'$  or  $?IM'$  with  $G' \subseteq G$  or  $M' \subseteq M$ .

Finally the subset-superset relation allows to deduce implications. Examples are discussed. A crucial point is, whether a premise is empty or not. If the premise is not empty then one gets some kind of probability. As more often the premise is fulfilled as more weight it has as a real rule. However it should be clarified that even the machinery of lattice theory cannot „generate“ new insights. They are all already included in the data set,- the problem is only, to distillate the

information from them.

## References

- Birkhoff, G. 1984. *Lattice theory*. Providence, Rhode Island: American Mathematical Society, Vol 25.
- Davey, B. A. and Priestley, H. A. 1990. *Introduction to Lattices and Order*. Cambridge: Cambridge University Press.
- Ganter, B. and Wille, R. eds. 1996. *Formale Begriffsanalyse Mathematische Grundlagen*. Berlin: Springer-Verlag.
- Brüggemann, R., Voigt, K. and Steinberg, C. 1997b. Application of Formal Concept Analysis to Evaluate Environmental Databases. *Chemosphere* 35(3): 479-486.
- Devillers, J., Thioulouse, J. and Karcher, W. 1993. Chemometrical Evaluation of Multispecies-Multichemical Data by Means of Graphical Techniques Combined with Multivariate Analyses. *Ecotoxicol. Environ. Saf.* 26: 333-345.
- Bartel, H.-G. (1999). Formal Concept Analysis in Ecotoxicology. In *Proceedings of I. Workshop on Order Theoretical Tools in Environmental Sciences*. Germany, Berlin. To appear as IGB Report.