***MATCH***
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **54** (2005) 583-609

ISSN 0340 - 6253

# Partial Ranking Models by Genetic Algorithm Variable Subset Selection (GAVSS) Approach for Environmental Priority Settings

*Manuela Pavan, Viviana Consonni and Roberto Todeschini*[*],

Milano Chemometrics and QSAR Research Group – Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1 – 20126 Milano (Italy)

**Abstract**

Partial order ranking (POR) strategies, which from a mathematical point of view are based on elementary methods of Discrete Mathematics, appear as an attractive and simple tool to perform data analysis. Moreover order ranking strategies seem to be a very useful tool not only to perform data exploration but also to develop order ranking models, being a possible alternative to conventional QSAR methods. In fact, when data material is characterised by uncertainties, order methods can be used as alternative to statistical methods such as multi-linear regression (MLR), since they do not require specific functional relationship between the independent variables and the dependent variables (responses).

A ranking model is a relationship between a set of dependent attributes, experimentally investigated, and a set of independent attributes, i.e. model variables. As in regression and classification models the variable selection is one of the main steps to find predictive models. In the present work, the Genetic Algorithm (GA-VSS) approach is proposed as the variable selection method to search for the best ranking models within a wide set of variables. The ranking models based on the selected subsets of variables are compared with the experimental ranking and evaluated by a set of similarity indices. A case study application is presented on a partial order ranking model developed for 23 chemicals selected as active ingredients used in agricultural practice and analysed according to their toxicity on *Scenedesmus vacuolatus*.

* Corresponding author e-mail: roberto.todeschini@unimib.it

## 1 Introduction

The increasing complexity of the systems analysed in scientific research together with the significant increase of available data require availability of suitable methodologies for multivariate statistics analysis and motivate the endless development of new methods. Moreover, the increasing of problem complexity leads to the decision processes becoming more complex, requiring the support of new tools able to set priorities and define rank order of the available options. The huge number of chemicals used and released in the environment is one of the complex problems the scientific community has to deal with. Since it is not possible to generate experimentally all necessary input for the risk assessment of these chemicals, information on the environmental fate and effects of the chemicals is usually performed by Quantitative Structure - Activity Relationships (QSAR) regression modelling. In QSAR models structural, steric and/or electronic features in series of selected chemicals are associated with modification in a given biological or physico-chemical end-point of the chemicals. QSAR modelling usually looks for unknown relations between several descriptors and the end-points; however when a relationship between a toxic activity and molecular descriptors is searched for, it should be kept in mind that toxicity data are typically multiple response endpoints, i.e. the chemical toxicity is analysed at different concentrations to detect both acute and chronic effects. Furthermore, toxicity data often include uncertainties and measurements errors. Thus, if the aim is to point out the more toxic and thus hazardous chemicals and to set priorities before final decisions are taken and data material is characterised by uncertainties, partial order models can be an attractive complement to statistical methods such as multi-linear regression (MLR). As it has been already pointed out in several studies on the use of ordering techniques for QSAR [1-5], despite conventional QSAR methods, partial order ranking by Hasse diagram technique assumes neither linearity nor any assumptions about distribution properties; thus being a parameter-free method. Moreover, it is suitable in all those environmental problems whose aim is to define order relations among several chemicals, to point out the more hazardous chemicals and to set priorities before final decisions are taken [6-8]. For these purposes order ranking models, which allow finding out not a quantitative response for each chemical but the inter-relationships, seem a promising approach in supporting environmental decision making processes.

The Genetic Algorithm (GA-VSS) approach is here used as the variable selection method to search for the best ranking models within a wide set of variables. The models based on the selected subsets of variables are compared with the experimental ranking and evaluated by a similarity index or by Tanimoto indices. Only models of the best quality, i.e. highly correlated with the experimental ranking, are retained in the population undergoing the evolution procedure. In the present study a partial ranking model for 23 chemicals selected as active ingredients used in agricultural practice is illustrated: the aim is to provide a priority list of these chemicals for the aquatic system according to their overall toxicity on *Scenedesmus vacuolatus*, contemporary accounting for their toxicity at the complete range of effect.

## 2 Theory

### 2.1 Partial ranking method: Hasse diagram technique

The Hasse diagram technique is a very useful tool to perform partial order ranking. It has been introduced in environmental sciences by Halfon [6] and refined by Brüggemann [9]. In this approach the basis for ranking is the information collected in the full set of attributes, E, which is called the "information basis" of the comparative evaluation of elements.

The typical data matrix contains $n$ elements (rows) and $R$ attributes (columns). The entry $y_{ir}$ of the matrix is the numerical value of the *r-th* attribute of the *i-th* element. Let IB be the information basis of evaluation, E the set of $n$ elements: the two elements $s$ and $t$ are comparable if *for all* $y_r \in$ IB either $y_r(s) \le y_r(t)$ or $y_r(s) \ge y_r(t)$. If $y_r(s) \le y_r(t)$ for all $y_r \in$ IB then $s \le t$, while if $y_r(s) \ge y_r(t)$ for all $y_r \in$ IB then $s \ge t$. The request "for all" is very important and is called the *generality principle*:

$$s, t \in \mathrm{E};\ s \le t \ \Leftrightarrow\ y(s) \le y(t)$$

$$y(s) \le y(t) \ \Leftrightarrow\ y_r(s) \le y_r(t) \text{ for all } y_r \in \mathrm{IB}$$

If there are some $y_r$, for which $y_r(s) < y_r(t)$ and some others for which $y_r(s) > y_r(t)$ then $s$ and $t$ are *incomparable*, and the common notation is $s \| t$. A partial order ranking is easily developed by the Hasse diagram technique comparing each pair of elements and storing this information in the Hasse matrix which is a ($n$ x $n$) antisymmetric matrix: for each pair of elements $s$ and $t$ the entry $h_{st}$ of this matrix is:

$$h_{st}\begin{cases}+1 & \text{if } y_r(s) \geq y_r(t) \text{ for all } y_r \in IB \\ -1 & \text{if } y_r(s) < y_r(t) \text{ for all } y_r \in IB \\ 0 & \text{otherwise}\end{cases}$$

The results of the partial order ranking are visualized in a diagram, named Hasse diagram, where each element is represented by a small circle, comparable elements which belong to an order relation are linked, while incomparable elements are not connected with a line and they are located as high as possible in the diagram, such that the diagram exhibits a level structure. The elements at the top of the diagram are called *maximals* (*maximal elements*) and they have none element above; the elements which have none element below are called *minimals* (*minimal elements*). In environmental field, where the Hasse diagram technique has been firstly proposed, the main assumption is that the lower the numerical value of the criteria the lower the hazard. Therefore, the maximal elements are the most hazardous and are selected to form the set of priority elements.

### 2.2 Partial ranking models

Partial ranking method has been widely used to perform data exploration, investigate the inter-relationships of objects and/or variables and set priorities. However it appears a very useful tool even for modelling purposes. Mathematical models have become an extremely useful tool in several scientific fields like environmental monitoring, risk assessment, QSAR and QSPR, i.e. in the search for quantitative relationships between the molecular structure and the biological activity/ chemical properties of chemicals.

A *ranking model* is defined as a relationship between one or more dependent attributes, investigated experimentally, and a set of theoretically defined independent attributes, also called model attributes, such as molecular descriptors (for example graph theoretical invariants or quantum chemical properties):

$$rank_i(y_{i1}, y_{i2}, \ldots, y_{iR}) = f(x_{i1}, x_{i2}, \ldots, x_{ip})$$

where $f$ is a ranking function applied on the training set elements (TS), $R$ the number of dependent attributes and $p$ the number of independent attributes. A model ranking development is based on the following steps:

1. *Experimental ranking:* the partial ranking method is applied to experimental attributes (dependent attributes).

2. *Model ranking*: the partial ranking method is applied to a subset of selected model attributes (independent attributes).
3. *Experimental and model ranking comparison*: evaluation of the degree of agreement between two rankings, i.e. analysis of model ranking reliability.
4. *Model ranking evaluation*: for each element the interval of each experimental attribute is compared with the interval derived from the model ranking.

Thus, the ranking model is given by the chosen ranking function and the ordered training set.

In the first phase, elements are ranked according to the experimental attributes describing them. Thus, the Hasse diagram technique is applied to the experimental attributes providing a diagram of partially ordered elements. In the second phase the Hasse diagram technique is applied to a selected subset of model attributes, and the elements are ranked according to the selected model attributes.

Then, the two Hasse diagrams are compared to evaluate if the model ranking is able to reproduce the element ranking based on the experimental attributes. In this way the similarity between two diagrams of partially ordered elements, is measured. Finally, if the agreement between the model ranking and the experimental ranking is considered satisfactory, predictions of the ranking of other elements, not being investigated experimentally, can be performed by the model ranking.

As in multilinear regression (MLR) methods, the selection of variables (attributes) is crucial to developing an acceptable ranking model. The aim of variable subset selection is to reach optimal model complexity in predicting response variables by a reduced set of independent variables [10, 11]. Ranking models based on the optimal subsets of a few predictor attributes have the great advantage of being more statistically stable, interpretable and showing higher predictive power. One of the simplest techniques for variable selection, - "sentimental selection" -, is based on the *a priori* selection of a few variables, by experience, tradition, availability, opportunity or previous knowledge. Another more mathematically based, but common, method of performing variable selection is the one based on an exhaustive examination of all the possible $k$ variables models (the model size) obtained by a set of $p$ variables. However, when many variables are available, an exhaustive examination of all possible models is not feasible as, given the extremely high number of possible variable combinations, it requires

extensive computational resources and is time consuming. In such cases a variable selection technique is needed. The Genetic Algorithm Variable Subset Selection (GA-VSS) approach is used here as the variable selection method to search for the best ranking models within a wide set of variables.

### 2.3 GA-VSS applied to partial ranking models

Genetic algorithms (GA) are an evolutionary method widely used for complex optimisation problems in several fields such as robotics, chemistry and QSAR [12, 13]. Since complex systems are described by several variables, a major goal in system analysis is the extraction of relevant information, together with the exclusion of redundant and noisy information. A special application of Genetic algorithms is variable selection for modelling purposes [14-18]. Variable selection is performed by GAs by considering populations of models generated through a reproduction process and optimised according to a defined *objective function* related to model quality. The procedure is illustrated in Figure 1.

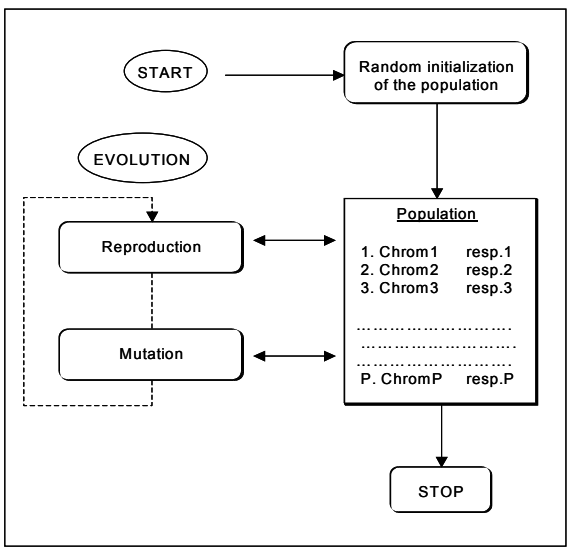


Figure 1: – Genetic algorithm procedure.

It consists in the evolution of a *population* of models, i.e. a set of ranked models according to some objective function, based on the crossover and mutation processes, which are alternatively repeated until a stop condition is encountered (e.g., a user-defined maximum number of iterations) or the process is ended arbitrarily.

It is to be highlighted that the GA-VSS method provides not a single model but a population of acceptable models; this characteristic allows the evaluation of variable relationships with response from different points of view. Moreover, when variable subset selection is applied to a huge number of variables, the genetic strategy can be extended to more than one population, each based on different variable subsets, evolving from each other independently. In this case, after a number of iterations, these populations can be combined according to different criteria, obtaining a new population with different evolutionary capabilities [18].

### 2.4 Partial ranking optimisation parameters

Variable subset selection is performed by GAs optimising populations of models according to a defined *objective function* related to model quality. In partial ranking models *objective function* is an expression of the degree of agreement between the element ranking resulting from experimental attributes and that provided by the selected subset of model attributes.

For the same $n$ elements the correlation between the experimental partial ranking and the model ranking (denoted as E and M, respectively) can be evaluated by a set of similarity measures, called Tanimoto indices [19 - 24] $T(I_E, I_M)$. Each Tanimoto index can be used as the measure of "goodness of fit" (degree of agreement) as it is the ratio of the number of agreements over the number of disagreements, i.e. contradictions in the ranking of two elements in the model and experimental ranking.

Another similarity index is here proposed as a measure of the agreement between two partial rankings. It is calculated comparing the experimental and model Hasse matrices, denoted **E** and **M** respectively, according to the following expression:

$$S(\mathbf{E},\mathbf{M}) = 1 - \frac{\sum_{st} \left| h_{st}^{\mathbf{E}} - h_{st}^{\mathbf{M}} \right|}{2n \cdot (n-1)} \qquad 0 \le S(\mathbf{E},\mathbf{M}) \le 1$$

where:

$h_{st}$ is the entry of the Hasse matrix for each pair of elements $s$ and $t$ and

$$s,t \in E \quad \text{and} \quad s \neq t \quad \text{and} \quad h_{st} \begin{cases} +1 & if \quad y_r(s) \geq y_r(t) \quad for\ all\ y_r \in IB \\ -1 & if \quad y_r(s) < y_r(t) \quad for\ all\ y_r \in IB \\ 0 & otherwise \end{cases}$$

S(**E**,**M**), being a similarity index, ranges from 0 (no similarity) to 1 (complete similarity) and expresses the differences between the two compared matrices; if two elements ($s$ and $t$) have the same mutual rank in both rankings, their contribution is 0. Thus it can be forecast that if two elements ($s$ and $t$) have different ranks, but not opposite ones, in the two rankings ($h_{st}^E = \pm 1$ and $h_{st}^M = 0$, or $h_{st}^E = 0$ and $h_{st}^M = \pm 1$), then their contribution is 1, while if the mutual ranks are opposite ($h_{st}^E = 1$ and $h_{st}^M = -1$, or $h_{st}^E = -1$ and $h_{st}^M = +1$), their contribution is 2. In this way the discrepancies due to opposite mutual rankings are evaluated more deeply than those due to comparable element pairs that have become incomparable, and *vice versa*.

Figure 2 shows the procedure used to compare the partial experimental ranking and the partial model ranking.

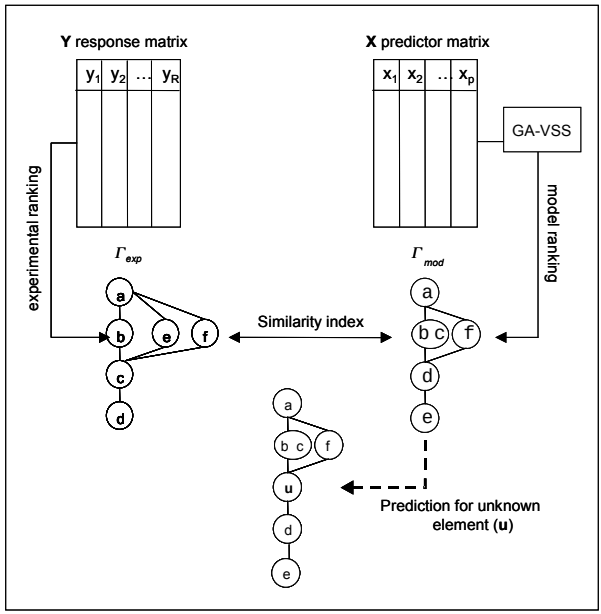


Figure 2: – Scheme of the procedure used to compare the experimental and model ranking.

### 2.5 Ranking predictions

Once the "goodness of fit" of the model ranking has been verified, predictions can be performed for new elements. The experimental ranking of new compounds that have not yet been investigated experimentally can be estimated by the ranking model; from the set of model attributes $\{x_{u1},\ldots, x_{up}\}$ describing any unknown element *u*, prediction of the experimental ranking of element *u* can be performed on the basis of the training set elements:

$$f\{x_{u1},.....,x_{up}\} \xrightarrow{training\ set} rank_u$$

To explain ranking predictions, a directed connectivity operator *C* is introduced. Being *s* and *t* two diverse elements in a partial ranking (PR), and N the set of natural numbers, then the connectivity operator $C(s,t)$ is defined as follows:

$$\text{if } s,t \in TS \text{ and } s \neq t \rightarrow C(s,t) \in N$$

$$C(s,t) = 0 \quad \text{iff} \quad s \text{ is incomparable with } t \quad (s \| t)$$
$$C(s,t) \in N^+ \quad \text{iff} \quad s \text{ is above } t$$
$$C(s,t) \in N^- \quad \text{iff} \quad s \text{ is below } t$$

The operator $C(s,t)$ has the following properties:

- $C(s,t) = k \qquad 0 \le |k| \le L\text{-}1$
- $C(s,t) = -\,C(t,s) \quad \rightarrow \quad$ antisymmetry
- $C(s,t) = p$ and $C(t,z) = q \quad \Rightarrow \quad C(s,z) = p + q \quad$ if $p,q > 0$ □ transitivity

where the absolute value $|k|$ of $k$ is the topological distance between the two elements $s$ and $t$ in the Hasse diagram, i.e. the shortest path length in the diagram and $L$ the number of levels in the ranking. According to the first property, the operator is an integer, taking a value equal to the path length between $s$ to $t$. If $s$ is above $t$, and is located in the level immediately above $t$ then $C(s,t)$ takes a value equal to 1. The maximum length of a Hasse diagram, i.e. the maximum number of lines in the longest chain, is equal to $L$-1, $L$ being the number of HD levels. If no path exists between $s$ and $t$, meaning that $s$ and $t$ are incomparable ($s \| t$), then $C(s,t)$ equals 0. Reflecting the ranking order relation properties, the connectivity operator has antisymmetry and transitivity properties. Thus, through the connectivity operator, predictions of the experimental ranking of any unknown element $u$ can be performed looking for the two elements $s$ and $t$ which satisfy the following conditions:

$$\min_s C(s,u) > 0 \quad \wedge \quad \min_t C(u,t) > 0 \quad \wedge \quad \min[(y_s - y_t)] > 0$$

where $s$ and $t$ are the two elements connected (comparable) to $u$, i.e. $C(s,u) > 0$ (with $s$ above $u$) and $C(u,t) > 0$ (with $u$ above $t$), located on the shortest path, and whose experimental difference value constitutes the smallest positive interval. Moreover, $C(s,u)$ represents the *u-above rank radius* and $C(u,t)$ the *u-below rank radius*, whereas $C(s,t)$ is the *u rank diameter.*

### 2.5 Prediction uncertainty

According to the proposed prediction calculation procedure, it is clear that the actual distance between the two elements $s$ and $t$, which satisfies the prediction conditions for any unknown element $u$, is crucial, and the larger the distance the larger the potential uncertainty in the prediction. Thus a first topological measure of the prediction

precision is provided by the connectivity operator $C(s,t)$ previously defined: the precision decreases for increased $C(s,t)$.

$$1 \le C(s,u) \le L-1 \quad \text{and} \quad 1 \le C(u,t) \le L-1$$

Moreover a normalised distance measure for each prediction from the upper and lower limits of the interval can be evaluated according to the expression:

$$D_u^{\sup} = \frac{C(s,u)-1}{L-2} \qquad 0 \le D_u^{\sup} \le 1$$

$$D_u^{\inf} = \frac{C(u,t)-1}{L-2} \qquad 0 \le D_u^{\inf} \le 1$$

$s$ and $t$ being the two elements which, satisfying the prediction conditions, are selected to predict the experimental interval of the unknown element $u$. $D_u^{\sup}$ and $D_u^{\inf}$ give a measure of the normalised rank uncertainty, above and below respectively. Note that if $u$ is a priority element (maximal) $C(s,u)$ is not defined, as no element exists above $u$, thus $D_u^{\sup}$ is not defined and only $D_u^{\inf}$ can be evaluated. Analogously, if $u$ is a minimal element $C(u,t)$ is not defined as no element exists below $u$, thus $D_u^{\inf}$ is not defined and only $D_u^{\sup}$ can be evaluated.

Another way to measure prediction uncertainty is to evaluate the experimental interval width of the prediction on the *r-th* experimental attribute:

$$Ry_{ur} = \frac{y_{sr} - y_{tr}}{\max_{y_r} - \min_{y_r}} \qquad 0 \le Ry_{ur} \le 1$$

where $y_{sr}$ and $y_{tr}$ are the experimental values of $s$ and $t$ for the *r-th* attribute respectively, and $max_{yr}$ and $min_{yr}$ the maximum and minimum experimental values of the *r-th* attribute. The greater the width, the greater the uncertainty. For maximal and minimal elements $Ry_{ur}$ is not defined, as their estimated interval is an open interval. Therefore, $D_u^{\sup}$ and $D_u^{\inf}$ measure the normalised *rank uncertainty* of the estimated interval, above and below respectively, whereas $Ry_{ur}$ measures the *experimental uncertainty*.

### 2.6 Model analysis

Further verification of model ranking applicability can be obtained by applying the described ranking prediction procedure to the training set elements initially used to develop the model. This results in the creation of a number of modified data sets from

which the elements will be deleted from the data one by one. For each element of the training set the *experimentally derived intervals* are calculated from the experimental ranking; the other training set elements are then used to calculate the experimental intervals of that element from the experimental ranking. In the same way, the *model calculated intervals* are obtained by deleting one element at a time from the model ranking, and using the remaining training set elements to calculate the model intervals of the deleted element from the model ranking. Once having obtained the experimentally derived intervals and the calculated intervals, they are compared to establish the model ranking quality.

On comparing two intervals, six different cases, illustrated in Figure 3, can be identified.

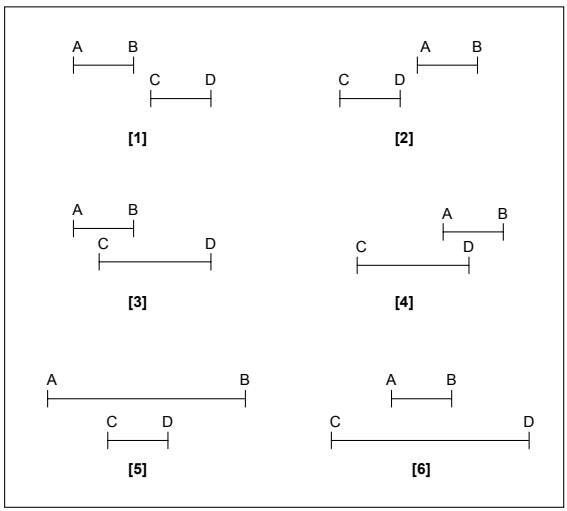


Figure 3: – Interval comparison.

As A and B are respectively the lower and upper values of the experimental interval, and C and D those of the model interval, Cases 1 and 2 represent disjoint intervals; Cases 5 and 6 intervals contained one in the other, and Cases 3 and 4 partially overlapped intervals.

Analysing one experimental attribute at a time, for each *i-th* element the disagreement $\delta_{ir}$ between its experimentally derived interval (A-B) and its model calculated interval (C-D) on the *r-th* attribute is calculated, assuming the worst case, according to the following expressions:

- Case 1: $\delta_{ir} = |D - A|$
- Case 2: $\delta_{ir} = |B - C$
- Case 3, 4, 5, 6: $\delta_{ir} = |C - A| + |D - B|$

A standardised interval disagreement for the *i-th* element on the *r-th* attribute is then derived as:

$$\delta^*_{ir} = \frac{\delta_{ir}}{\max_{y_r} - \min_{y_r}}$$

$max_{yr}$ and $min_{yr}$ being the maximum and minimum values of the *r-th* attribute respectively.

The average disagreement between the experimental and the model calculated intervals is then calculated:

$$\overline{\delta}_r = \frac{\sum_{i=1}^{N} \delta^*_{ir}}{n}$$

and a measure of the *ranking model quality*, as far as concerns the *r-th* attribute is calculated as:

$$Q_r = 1 - \overline{\delta}_r$$

The overall ranking model quality, i.e. taking into account all the *R* responses, can be evaluated by the following expressions:

$$Q_T = \frac{\sum_{r=1}^{R} Q_r}{R} \qquad Q_G = \sqrt[R]{Q_1 \cdot \ldots \cdot Q_R} \qquad Q_M = \min_r \{Q_r\}$$

$Q_T$ being the arithmetic mean of all the *R* attributes of the ranking model represents the least demanding parameter for evaluating overall model ranking quality. Instead the geometric mean $Q_G$ is a more severe parameter, able to enhance models not able to reproduce a correct experimental ranking for only a few attributes. The most demanding

evaluation parameter of model quality is $Q_M$, which assumes minimum quality among the $R$, calculated as the representing overall model quality.

This procedure for evaluating model ranking quality is based on ranking interval comparison. Moreover, as the metric scale is usually seen as a "stronger" property than the ordinal scale, it is of interest to measure the loss of information due the replacement of the original "quantitative" information with rank orders.

Thus, being the quantitative experimental values intervals with equal lower and upper values, they are compared with the experimentally derived intervals (A-B), and for each *r-th* attribute the standardised interval disagreement ${}^{0}\delta_{ir}^{*}$ is calculated the same way, as described above. The arithmetic mean of the average disagreement between the quantitative experimental values and their derived intervals on the *r-th* attribute provides a measure of the uncertainty increase due to the replacement of a metric scale with an ordinal scale and is calculated as:

$$\widetilde{\delta}_r = \frac{\sum_{i=1}^{N} {}^{0}\delta_{ir}^{*}}{n}$$

## 3 Partial order ranking QSAR model for agriculture chemicals.

Today, more than 100.000 chemical are in use and constitute a potential risk to the environment. Human activities introduce a large amount of different chemicals into the aquatic environment, either by accident, in wastewaters (surfactants and pharmaceuticals from household use, heavy metals from industry) or in run-off waters from agriculture (herbicides and fungicides used in plant protection products). Even so, it is the professed aim of the European Communities to ensure the sustainable use of water and to protect the structure and function of the aquatic ecosystem (EU parliament 2000). Thus, a methodology is needed for risk assessment of chemicals. In the case of "new" chemicals that entered the European market after 18 September 1981, the hazard assessment is based on a minimum set of toxicity data from bacteria, algae, and daphnia. However, it is not practically possible experimentally to generate all the necessary input information for the risk assessment of these chemicals. For this reason, it appears necessary to obtain part of the information concerning the chemicals fate and effect in the environment by models. The development of efficient and inexpensive

technologies for effective risk assessment and to predict physical, chemical and biological properties of new compounds is thus of major interest.
An application of a Quantitative Structure - Activity Relationships (QSARs) by partial ranking technique for agriculture chemicals is illustrated here.

### 3.1 Toxicity experimental data

The toxicity data have been provided by the EU project: BEAM [25]. The dataset consists of 23 chemicals selected as active ingredients used in agricultural practice: they are included among the 10 major European crops in quantitative terms and they are representative of agriculture of various European areas (North, Central, South). The chemicals have been tested for toxicity on freshwater algae *Scenedesmus vacuolatus* by the research group of Bremen University, coordinator of the EU project BEAM. The dependent variables selected for describing their toxicity were the reproduction inhibition responses with 3 concentrations (μmol/L) provoking 10% (EC10) 50% (EC50), 90% (EC90) effect, respectively. Table 1 shows the EC toxicity values of the 23 chemicals.

### 3.2 Molecular descriptors

The chemical structures of the agriculture chemicals have been described with more than 1500 molecular descriptors, in order to catch all the structural information.

The molecular descriptors have been calculated by the *Dragon* program [26] on the basis of the minimum energy molecular geometries optimized by *HyperChem* package [27] (PM3 semiempirical method). In this study the following sets of molecular descriptors have been calculated: constitutional descriptors, topological descriptors [28-29], walk and path counts, connectivity indices [30], information indices, Moreau-Broto 2D-autocorrelations [31-33], edge adjacency indices [34], Burden eigenvalue descriptors [35-36], topological charge indices [37-38], eigenvalue based indices [39], Randic molecular profiles [40-41], geometrical descriptors, radial distribution function descriptors [42], 3D-MoRSE descriptors [43-44], WHIM descriptors [45-46], GETAWAY descriptors [47], functional group counts and atom centred fragments. Definitions and further information regarding all these molecular descriptors can be found in the *Handbook of Molecular Descriptors* [48].

Table 1: – Experimental toxicity data (Log1/EC) and values of the two descriptors selected by GA for 23 chemicals.

| *ID* | *Substance* | *Log(1/EC10)* | *Log(1/EC50)* | *Log(1/EC90)* | *nN* | *CIC2* |
|---|---|---|---|---|---|---|
| 1 | Aclonifen | 2.024 | 1.527 | 1.067 | 2 | 1.228 |
| 2 | Atrazin | 1.574 | 0.745 | 0.415 | 5 | 1.376 |
| 3 | Lenacil | 1.916 | 1.306 | 1.027 | 2 | 2.114 |
| 4 | Chloridazon | -0.045 | -0.723 | -1.155 | 3 | 0.885 |
| 5 | Alachlor | 1.215 | 0.853 | 0.621 | 1 | 1.366 |
| 6 | Metolachlor | 0.434 | 0.087 | -0.078 | 1 | 1.241 |
| 7 | Tribenuron-methyl | 1.683 | 0.597 | -0.095 | 5 | 1.134 |
| 8 | Thifensulfuron-methyl | 0.057 | -1.139 | -2.335 | 5 | 0.744 |
| 9 | Bromoxynil | -1.878 | -2.115 | -2.352 | 1 | 0.571 |
| 10 | Carbofuran | -1.169 | -2.121 | -2.728 | 1 | 1.194 |
| 11 | Cycloxydim | -1.498 | -2.445 | -3.048 | 1 | 1.603 |
| 12 | Ethofumesate | 0.112 | -1.588 | -2.671 | 0 | 1.244 |
| 13 | Isofenphos | 0.952 | -0.890 | -2.119 | 1 | 1.622 |
| 14 | Isoxaflutol | -1.211 | -1.956 | -2.431 | 1 | 0.885 |
| 15 | MCPA | -2.076 | -2.902 | -3.729 | 0 | 0.523 |
| 16 | Terbuthylazin | 1.642 | 1.159 | 0.852 | 5 | 1.799 |
| 17 | Metamitron | 0.657 | -0.329 | -0.957 | 4 | 1.005 |
| 18 | Ioxynil | -0.689 | -1.534 | -2.072 | 1 | 0.571 |
| 19 | Triasulfuron | 1.391 | 0.273 | -0.440 | 5 | 0.655 |
| 20 | Isoproturon | 1.363 | 0.641 | 0.166 | 2 | 1.719 |
| 21 | Linuron | 1.990 | 1.057 | 0.463 | 2 | 0.971 |
| 22 | Pendimethalin | 2.706 | 2.069 | 1.663 | 3 | 1.718 |
| 23 | 2,4-Dichlorophenoxyacetic acid | -1.891 | -2.932 | -3.369 | 0 | 0.461 |

### 3.3 Experimental ranking

The Hasse diagram technique has been applied on the three toxicity responses of algae reproduction inhibition with 3 concentrations of 10, 50, 90 □mol/l. Figure 4 shows the experimental Hasse diagram: it is arranged on twelve levels and characterized by 223 comparable pairs of elements and 60 contradictions. The diagram is of simple interpretation: the more toxic chemicals are located on the top while the less toxic are on the bottom. The diagram points out pendimethalin as a maximal element, since it is

characterized by the highest toxicity values at all the three concentration levels. It is the most toxic chemical among the 23 investigated, followed by aclonifen. Linuron and lenacil can be considered at the same toxicity level but with diverse behavior: the former explicates high toxicity at low concentration (acute effect), the latter at high concentrations (chronic effect). MCPA (2-methyl-4-chlorophenoxyacetic acid) and 2,4-dichlorophenoxyacetic acid are minimals, showing the low toxicity values at all the three concentration levels.

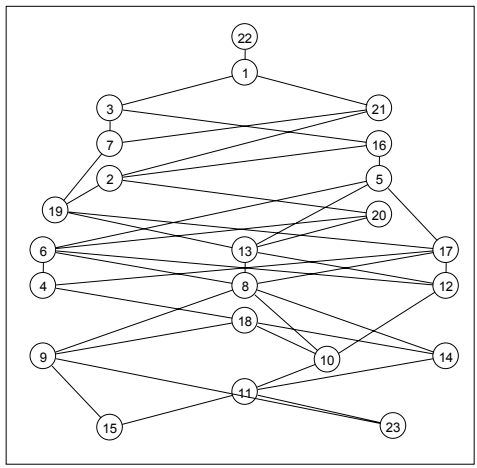


Figure 4: – Experimental Hasse diagram.

### 3.4 Model ranking

The correlations between the toxicity of the considered chemicals and the molecular descriptors have been estimated by the partial ranking Hasse diagram technique (HDT). However as an exhaustive search for the best ranking models within a wide set of descriptors requires extensive computational resources and is time consuming, given the extremely high number of possible descriptor combinations, the Genetic Algorithm (GA-VSS) approach has been used as the variable selection method. Starting from a population of 100 random models with a number of variables equal to or less than 3, the algorithm has explored new combinations of variables, selecting them by a mechanism

of reproduction/mutation similar to that of biological population evolution. The models based on the selected subsets of variables have been tested and evaluated by similarity index (S(**E**,**M**)). All of the calculations have been performed by the in-house software *RANA* for variable selection for WINDOWS/PC [49].

The optimal model obtained is a very simple model, made of two variables: the number of nitrogen atoms (nN) and the complementary information content (neighbourhood symmetry of order 2) CIC2. The maximal elements of the experimental Hasse diagram are the more toxic element (priority elements), whereas the minimal elements are the less toxic. According to the model Hasse diagram, the more toxic elements are those with a greater number of nitrogen atoms and with a greater value of CIC2. The model Hasse diagram is shown in Figure 5: it is arranged on eleven levels and characterized by 171 comparable pairs of elements and 164 contradictions. The two model descriptor values are illustrated in Table 1. The diagram points out lenacil and terbuthylazin as maximal elements, the former is characterized by the highest CIC2 value (CIC2 = 2.114), the latter by both high number of nitrogen atoms (nN = 5) and quite high CIC2 value (CIC2 = 1.799). 2,4-dichlorophenoxyacetic acid is the least element, followed by MCPA (2-methyl-4-chlorophenoxyacetic acid): they are both characterised by absence of nitrogen atoms and low CIC2 value.

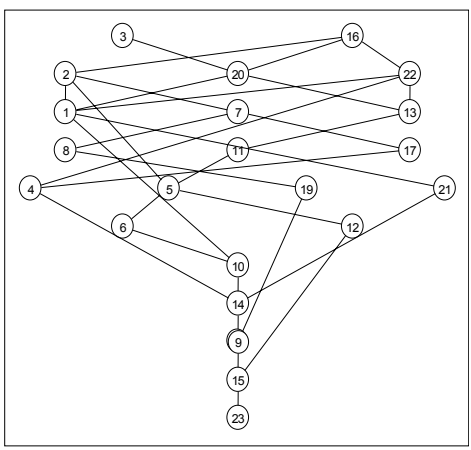


Figure 5: – Model Hasse diagram developed with nN and CIC2 descriptors.

The agreement degree between experimental and model diagrams is quite satisfactory (**S**(**E**,**M**) = 76.3). The Tanimoto indices have been also calculated:

T(0,0) = 87.9 T(0,1) = 80.7 T(1,1) = 58.2

The "goodness of fit" of the partial ranking model calculated by the similarity index is lower than that calculated by both T(0,0) and T(0,1) but higher than the one by T(1,1), confirming that the similarity index S(**E**,**M**) is a reasonable compromise between the over optimistic and the over pessimistic evaluation provided by T(0,0), T(0,1) and T(1,1) respectively.

### 3.5 Interval estimation

The experimental ranking of each chemical has been estimated according to the procedure described above. The calculated intervals have been compared to the corresponding experimentally derived intervals, obtained by deleting each chemical from the experimental ranking diagram; and using the remaining training set elements to calculate the experimental intervals of the deleted element from the experimental ranking diagram. Analysing one experimental response at a time, for each chemical the standardised disagreement $\delta_{ir}$ between its experimentally derived interval and model-calculated interval has been calculated. The experimentally derived intervals and the calculated intervals for Log(1/EC10), Log(1/EC50), Log(1/EC90), together with the corresponding standardised disagreements are illustrated in Table 2, 3 and 4, respectively.

### 3.6 Overall model quality

By comparing the experimentally derived intervals with the calculated ones, an average disagreement has been calculated on each response:

$$\overline{\delta}_{Log(1/EC10)} = 0.314 \qquad \overline{\delta}_{Log(1/EC50)} = 0.276 \quad \overline{\delta}_{Log(1/EC90)} = 0.293$$

The average disagreement between the quantitative experimental values and their derived intervals has been calculated:

$$\widetilde{\delta}_{Log(1/EC10)} = 0.171 \qquad \widetilde{\delta}_{Log(1/EC50)} = 0.190 \qquad \widetilde{\delta}_{Log(1/EC90)} = 0.150$$

Table 2: – Experimental Log(1/EC10) interval estimation. (bold fonts indicate disjoint intervals). [a]2,4-Dichlorophenoxyacetic acid.

| *Response: Log(1/EC10)* | | *Experimental* | | *Calculated* | | |
|---|---|---|---|---|---|---|
| *ID* | *Substance* | *Min* | *Max* | *Min* | *Max* | $\delta_{EC10}$ |
| 1 | Aclonifen | 1.990 | 2.706 | **-1.169** | **1.363** | 0.810 |
| 2 | Atrazin | 1.391 | 1.642 | 1.215 | 1.642 | 0.037 |
| 3 | Lenacil | 1.683 | 2.024 | > 1.363 | - | 0.067 |
| 4 | Chloridazon | -0.689 | 0.434 | -1.211 | 0.657 | 0.156 |
| 5 | Alachlor | 0.952 | 1.642 | 0.434 | 1.574 | 0.123 |
| 6 | Metolachlor | 0.112 | 1.363 | -1.169 | 1.215 | 0.299 |
| 7 | Tribenuron-methyl | 1.391 | 1.916 | 0.657 | 1.574 | 0.225 |
| 8 | Thifensulfuron-methyl | -1.169 | 0.434 | **1.391** | **1.683** | 0.596 |
| 9 | Bromoxynil | -1.891 | -0.689 | -2.706 | -1.211 | 0.280 |
| 10 | Carbofuran | -1.498 | -0.689 | -1.211 | 0.434 | 0.295 |
| 11 | Cycloxydim | -1.891 | -1.169 | **0.434** | **0.952** | 0.595 |
| 12 | Ethofumesate | -1.169 | 0.434 | -2.706 | 1.215 | 0.485 |
| 13 | Isofenphos | 0.112 | 1.363 | -1.498 | 1.363 | 0.337 |
| 14 | Isoxaflutol | -1.498 | -0.689 | **-0.689** | **-0.045** | 0.304 |
| 15 | MCPA | - | < -1.498 | -1.891 | -1.878 | 0.079 |
| 16 | Terbuthylazin | 1.574 | 1.916 | > 2.706 | - | 0.237 |
| 17 | Metamitron | 0.112 | 1.363 | -0.045 | 1.683 | 0.100 |
| 18 | Ioxynil | -1.169 | -0.045 | **-2.706** | **-1.211** | 0.556 |
| 19 | Triasulfuron | 0.952 | 1.574 | **-0.689** | **0.057** | 0.473 |
| 20 | Isoproturon | 0.952 | 1.574 | 0.952 | 1.642 | 0.014 |
| 21 | Linuron | 1.683 | 2.024 | **-1.211** | **0.657** | 0.676 |
| 22 | Pendimethalin | > 2.024 | - | **0.952** | **1.642** | 0.224 |
| 23 | [a]2,4- D | - | < -1.498 | - | < -2.706 | 0.253 |

Table 3: – Experimental Log(1/EC50) interval estimation. (bold fonts indicate disjoint intervals). [a]2,4-Dichlorophenoxyacetic acid.

| *Response: Log(1/EC50)* | | *Experimental* | | *Calculated* | | |
|---|---|---|---|---|---|---|
| *ID* | *Substance* | *Min* | *Max* | *Min* | *Max* | $\delta_{EC50}$ |
| 1 | Aclonifen | 1.306 | 2.069 | 1.057 | 2.069 | 0.050 |
| 2 | Atrazin | 0.273 | 1.057 | 0.853 | 1.159 | 0.136 |
| 3 | Lenacil | 1.159 | 1.527 | > 0.641 | - | 0.104 |
| 4 | Chloridazon | -1.534 | 0.087 | -1.956 | -0.329 | 0.168 |
| 5 | Alachlor | 0.087 | 1.159 | 0.087 | 0.745 | 0.083 |
| 6 | Metolachlor | -0.723 | 0.641 | -2.121 | 0.853 | 0.322 |
| 7 | Tribenuron-methyl | 0.273 | 1.057 | -0.329 | 0.745 | 0.183 |
| 8 | Thifensulfuron-methyl | -1.956 | -0.329 | **0.273** | **0.597** | 0.510 |
| 9 | Bromoxynil | -2.902 | -1.534 | -2.902 | -1.956 | 0.084 |
| 10 | Carbofuran | -2.445 | -1.588 | -1.956 | 0.087 | 0.433 |
| 11 | Cycloxydim | -2.902 | -2.121 | **-1.588** | **-0.890** | 0.402 |
| 12 | Ethofumesate | -2.121 | -0.890 | -2.902 | 0.853 | 0.505 |
| 13 | Isofenphos | -1.139 | 0.273 | -2.445 | 0.641 | 0.335 |
| 14 | Isoxaflutol | -2.445 | -1.534 | **-1.534** | **-0.723** | 0.344 |
| 15 | MCPA | - | < -2.445 | -2.932 | -2.115 | 0.066 |
| 16 | Terbuthylazin | 0.853 | 1.306 | > 2.069 | - | 0.243 |
| 17 | Metamitron | -0.723 | 0.641 | -0.723 | 0.597 | 0.009 |
| 18 | Ioxynil | -1.956 | -0.723 | **-2.902** | **-1.956** | 0.436 |
| 19 | Triasulfuron | -0.329 | 0.597 | **-1.534** | **-1.139** | 0.426 |
| 20 | Isoproturon | 0.087 | 0.745 | -0.890 | 1.159 | 0.278 |
| 21 | Linuron | 0.745 | 1.527 | **-1.956** | **-0.329** | 0.696 |
| 22 | Pendimethalin | > 1.527 | - | **-0.723** | **1.159** | 0.450 |
| 23 | [a]2,4- D | - | < -2.445 | - | < -2.902 | 0.091 |

Table 4: – Experimental Log(1/EC90) interval estimation. (bold fonts indicate disjoint intervals). [a]2,4-Dichlorophenoxyacetic acid.

| *Response: Log(1/EC90)* | | *Experimental* | | *Calculated* | | |
|---|---|---|---|---|---|---|
| *ID* | *Substance* | *Min* | *Max* | *Min* | *Max* | $\delta_{EC90}$ |
| 1 | Aclonifen | 1.027 | 1.663 | **0.463** | **0.852** | 0.810 |
| 2 | Atrazin | 0.166 | 0.463 | **0.621** | **0.852** | 0.037 |
| 3 | Lenacil | 0.852 | 1.067 | > 0.166 | - | 0.067 |
| 4 | Chloridazon | -2.072 | -0.078 | -2.431 | -0.957 | 0.156 |
| 5 | Alachlor | -0.078 | 0.852 | -0.078 | 0.415 | 0.123 |
| 6 | Metolachlor | -1.155 | 0.166 | -2.728 | 0.621 | 0.299 |
| 7 | Tribenuron-methyl | -0.440 | 0.463 | -0.957 | 0.415 | 0.225 |
| 8 | Thifensulfuron-methyl | -2.352 | -2.119 | **-0.440** | **-0.095** | 0.596 |
| 9 | Bromoxynil | -3.369 | -2.072 | -3.729 | -2.431 | 0.280 |
| 10 | Carbofuran | -3.048 | -2.671 | **-2.431** | **-0.078** | 0.295 |
| 11 | Cycloxydim | -3.369 | -2.728 | **-2.671** | **-2.119** | 0.595 |
| 12 | Ethofumesate | -2.728 | -2.119 | -3.729 | 0.621 | 0.485 |
| 13 | Isofenphos | -2.335 | -0.440 | -3.048 | 0.166 | 0.337 |
| 14 | Isoxaflutol | -3.048 | -2.335 | **-2.072** | **-1.155** | 0.304 |
| 15 | MCPA | - | < -3.048 | -3.369 | -2.671 | 0.079 |
| 16 | Terbuthylazin | 0.621 | 1.027 | > 1.663 | - | 0.237 |
| 17 | Metamitron | -1.155 | 0.166 | -1.155 | -0.095 | 0.100 |
| 18 | Ioxynil | -2.352 | -1.155 | **-3.729** | **-2.431** | 0.556 |
| 19 | Triasulfuron | -0.957 | -0.095 | **-2.352** | **-2.335** | 0.473 |
| 20 | Isoproturon | -0.078 | 0.415 | -2.119 | 0.852 | 0.014 |
| 21 | Linuron | 0.415 | 1.067 | **-2.431** | **-0.957** | 0.676 |
| 22 | Pendimethalin | > 1.067 | - | **-1.155** | **0.852** | 0.224 |
| 23 | [a]2,4- D | - | < -3.048 | - | < -3.729 | 0.253 |

The uncertainty increase due to the replacement of a metric scale with an ordinal scale, calculated as arithmetic mean on all the three experimental attributes, is equal to 0.170.

For each response, the model quality has been evaluated by complement of the average disagreement between experimental and calculated intervals ($Q_r$):

$$Q_{Log(1/EC10)} = 0.686 \qquad Q_{Log(1/EC50)} = 0.724 \qquad Q_{Log(1/EC90)} = 0.707$$

The overall ranking model quality, i.e. taking into account all the three responses, has been evaluated from the above parameters by arithmetic means ($Q_T$), geometric mean ($Q_G$) and by the minimum value obtained on the three responses ($Q_M$):

$$Q_T = 0.705 \qquad Q_G = 0.705 \qquad Q_M = 0.686$$

The present case study reveals that partial order ranking provides an attractive alternative to conventional QSAR modelling tools. The method appears, from a mathematical point of view, robust and transparent. It is thus possible using partial ranking techniques to develop ranking models and it is suggested that ranking models have a general potential in the area of risk assessment of environmentally hazardous chemicals. However, further analyses of the proposed method appear appropriate to investigate validation techniques suitable for ranking models and to evaluate the potential of ranking models for QSAR modelling.

## 4 Conclusions

A complete procedure to perform a partial ranking model has been here proposed, based on the following main steps: experimental and model ranking development, comparison of the experimental and model rankings to evaluate model reliability, and finally interval estimations to provide experimental ranking from the ranking model obtained. In order to allow processing of data described by a wide set of variables the Genetic Algorithm (GA-VSS) approach has been proposed as the variable selection method. Even if the information obtained by a ranking model is not quantitative information, but simply information regarding element inter-relations, in most environmental and chemical problems where the aim of the statistical methods used in QSAR strategies is to find priorities, i.e. identify which chemicals are more toxic or hazardous and which sites require quick intervention, partial ranking models appear as a useful and promising approach. Thus, for exposure analyses and risk assessment the use of ranking models is recommended, not to substitute conventional statistics but to supplement them. It is worthwhile to highlight that the procedure proposed can be located between fitting and predictive approaches, since the interval estimation and the model validation appear combined in one step. In fact, the model calculated intervals are obtained by deleting one element at a time from the model ranking, and using the remaining training set

elements to calculate the model intervals of the deleted element from the model ranking. Thus, it seems quite similar to a leave – one – out cross validation procedure (LOO technique), where each element is taken away, one at a time and the response for the deleted element is calculated from the model. In ranking model searching, the validation is not performed during the evolutionary optimisation procedure, but the model predictive ability is simulated once the model has been defined. The approach proposed seems, from a mathematical point of view well grounded. However, further analyses of the interval estimation procedure as well as of the uncertainty evaluation are required. Moreover, one of the main theoretical aspect not yet fully investigated concerns the search for validation techniques suitable for ranking models.

**Acknowledgements**

Financial support from the Commission of the European Union (R&D project "Beam", EVK1-CT1999-00012) is acknowledged.

**References**

1. Brüggemann, R., Pudenz, S., Carlsen, L., Sørensen, P.B., Thomsen, M., Mishra, R.K. (2001). The use of Hasse Diagrams as a Potential Approach for Inverse QSAR. *SAR and QSAR in Environmental Research*, *11*, 473-487.

2. 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.

3. Carlsen, L., Sørensen, P.B., Thomsen, M., Brüggemann, R. (2002a). QSAR's Based on Partial Order Ranking. *SAR and QSAR in Environmental Research*, *13*, 153-165.

4. Carlsen, L., Lerche, D.B., Sørensen, P.B. (2002b). Improving the Predicting Power of Partial Order Based QSARs through Linear Extensions. *J.Chem.Inf.Comput.Sci*., *42*, 806-811.

5. 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. *Environmental Toxicology and Chemistry*, *22*, 661-670.

6. Halfon, E., Reggiani, M.G. (1986). On Ranking Chemicals for Environmental Hazard. *Environ. Sci. Technol.*, *20*, 1173-1179.

7. Halfon, E. (1989). Comparison of an Index Function and a Vectorial Approach Method for Ranking of Waste Disposal Sites. *Environ. Sci. Technol.*, *23*, 600-609.

8. Halfon, E., Brüggemann, R. (1998), On Ranking Chemicals for Environmental Hazard. Comparison of methodologies. *Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences*, 11-48.

9. Brüggemann, R., Bartel, H-G. (1999c). A Theoretical Concept to Rank Environmentally Significant Chemicals. *J.Chem.Inf.Comput.Sci*., *39*, 211-217.

10. Hocking, R.R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, *32*, 1-49.

11. Miller, A.J. (1990). *Subset Selection in Regression*. Chapman & Hall, London (UK), 230 pp.

12. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Massachusetts, MA.

13. Wehrens, R. and Buydens, L M C. (1998). Evolutionary optimization: a tutorial. TrAC, *Trends in Analytical Chemistry*, *17*(4), 193-203.

14. Leardi, R., Boggia, R., and Terrile, M. (1992). Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.*, *6*, 267-281.

15. Leardi, R. (1994). Application of Genetic Algorithms to Feature Selection Under Full Validation Conditions and to Outlier Detection. *J.Chemom.*, *8*, 65-79.

16. Luke, B.T. (1994). Evolutionary Programming Applied to the Development of Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J.Chem.Inf.Comput.Sci.*, *34*, 1279-1287.

17. Leardi, R. (1996). Genetic Algorithms in Feature Selection. In *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design. Vol. 1* (Devillers, J., ed.), Academic Press, London (UK), pp. 67-86.

18. Todeschini R., Consonni V., Mauri A., Pavan M. (2004). MobyDigs: software for regression and classification models by genetic algorithms in *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks* (R.Leardi Ed.), Chapter 5, Elsevier, p.141-167.

19. Rogers, D.J. and Tanimoto, T.T. (1960). A computer Program for Classifying Plants. *Science*, *132*, 1115-1118.

20. Brüggemann, R., Zelles, L., Bai, Q.Y., Hartmann, A.(1995b). Use of Hasse Diagram Technique for Evaluation of Phospholipid Fatty Acids Distribution as Biomarkers in Selected Soils. *Chemosphere*, *30*, 1209-1228.

21. Bath, P.A., Morris, C.A., Willet, P. (1993). Effects of Standardization on Fragment-Based Measures of Structural Similarity. *J. Chemom*, *7*, 543-550.

22. Moock, T.E., Grier, D.L., Hounshell, W.D., Grethe, G., Cronin, K., Nourse, J.G., Theodosious, J. (1998). Similarity Searching in the Organic Reaction Domain. *Tetrahedron Computer Methodology*, *1*, 117-128.

23. 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. *Environmental Toxicology and Chemistry*, *22*, 661-670.

24. Pudenz, S., Brüggemann, R., Komoßa, D., Kreimes, K. (1997). An Algebaic/Graphical Tool to Compare Ecosystems with Respect to their Pollution by Pb/Cd III: Comparative Regional Analysis by Applying a Similarity Index. *Chemosphere*, *36*, 441-450.

25. Backhaus, T., Altenburger, R., Arrhenius, A., Blanck, H., Faust, M., Finizio, A., Gramatica, P., Grote, M., Junghans, M., Meyer, W., Pavan, M., Porsbring, T., Scholze, M., Todeschini, R., Vighi, M., Walter, H., Grimme, L.H. (2003). The BEAM-project: prediction and assessment of mixture toxicities in the aquatic environment. *Continental Shelf Research 23*, 1757-1769.

26. Todeschini, R., Consonni, V., Mauri, A., Pavan, M. (2004). DRAGON, rel. 5 for Windows; Talete srl: Milano, Italy.

27. HYPERCHEM. Rel 4 for Windows. 1995. Autodesk. Inc. Sausalito. CA. USA

28. Bonchev, D. (1983). Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press: Chichester.

29. Devillers, J. and Balaban, A.T. (2000). Topological Indices and Related Descriptors in QSAR and QSPR. Gordon & Breach: Amsterdam.

30. Kier, L.B. and Hall, L.H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press - Wiley, Chichester, 262 pp.

31. Moreau, G.and Broto, P. (1980a). The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv.J.Chim*., *4*, 359-360.

32. Moreau, G. and Broto, P. (1980b). Autocorrelation of Molecular Structures, Application to SAR Studies. *Nouv.J.Chim*., *4*, 757-764.

33. Broto, P., Moreau, G., Vandycke, C. (1984). Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor. *Eur.J.Med.Chem*., *19*, 66-70.

34. Estrada, E. (1995). Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *J.Chem.Inf.Comput.Sci*., 35, 31-33.

35. Pearlman, R.S.and Smith, K.M. (1998). Novel Software Tools for Chemical Diversity. In 3D QSAR in Drug Design - Vol. 2; Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer/ESCOM: Dordrecht; pp. 339-353.

36. Pearlman, R. S. (1999). Novel Software Tools for Addressing Chemical Diversity. *Internet Communication*, http://www.netsci.org/Science/Combichem/feature08.html.

37. Gálvez, J., Garcìa, R., Salabert, M.T., and Soler, R. (1994). Charge Indexes. New Topological Descriptors. *J.Chem.Inf.Comput.Sci*. 34, 520-525.

38. Gálvez, J., Garcìa-Domenech, R., De Julián-Ortiz, V., and Soler, R. (1995). Topological Approach to Drug Design. *J.Chem.Inf.Comput.Sci*. 35, 272-284.

39. Balaban, A.T., Ciubotariu, D. and Medeleanu, M. (1991). Topological Indices and Real Vertex Invariants Based on Graph Eigenvalues or Eigenvectors. *J.Chem.Inf.Comput.Sci*., 31, 517-523.

40. Randic, M. (1995). Molecular Shape Profiles. *J.Chem.Inf.Comput.Sci*. 35, 373-382.

41. Randic, M. (1996). Quantitative Structure-Property Relationship - Boiling Points of Planar Benzenoids. *New J.Chem.* 20, 1001-1009.

42. Hemmer, M.C., Steinhauer, V. and Gasteiger, J. (1999). Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vibrat.Spect*., 19, 151-164.

43. Schuur, J. ,and Gasteiger, J. (1996). 3D-MoRSE Code - A New Method for Coding the 3D Structure of Molecules. In, Software Development in Chemistry - Vol. 10 (J. Gasteiger, Ed.). Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main.

44. Schuur, J., and Gasteiger, J. (1997). Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation. *Anal.Chem*. 69, 2398-2405.

45. Todeschini, R., Lasagni, M., Marengo, E. (1994). New Molecular Descriptors for 2D- and 3D-Structures. Theory. *J.Chemom*., *8*, 263-273.

46. Todeschini, R., Gramatica, P. (1997). 3D-Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant.Struct.-Act.Relat*., *16*, 113-119.

47. Consonni, V., Todeschini, R. and Pavan, M. (2002). Structure / Response Correlation and Similarity / Diversity Analysis by GETAWAY Descriptors. Part 1. Theory of the Novel 3D Molecular Descriptors. *J.Chem. Comput. Sci*. *42*, 693-705.

48. Todeschini, R., and Consonni, V. (2000). *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, p. 667.

49. Todeschini, R., Consonni, V., Mauri, A., Pavan, M. (2003). RANA for Windows; Talete srl: Milano, Italy.