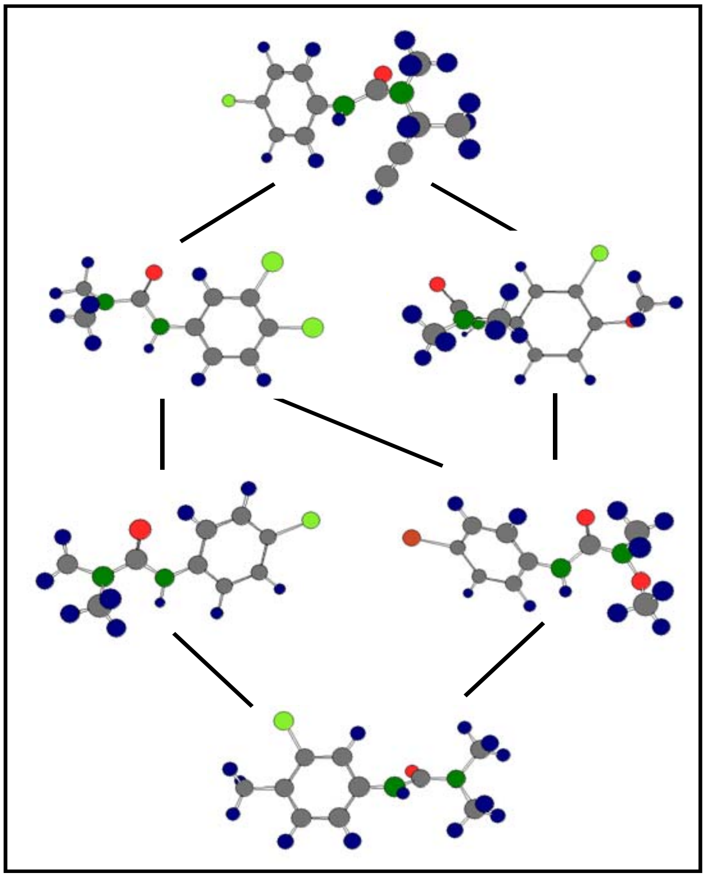
Manuela Pavan **2003**

TOTAL AND PARTIAL RANKING METHODS IN CHEMICAL SCIENCES



PhD in Chemical Sciences
Cycle XVI
University of Milano - Bicocca
Tutor : Prof. Roberto Todeschini

The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day. Never lose a holy curiosity.

Anyone who has never made a mistake has never tried anything new.

Albert Einstein, The human side, new glimpses from his archives

I would like to acknowledge Roberto for his support and contribution to the development of what is now collected in this thesis, Prof.sa Paola Gramatica, Viviana, Andrea, Davide and Morena for their cooperation, support and friendship over these years.

I also wish to thank warmly the help, the received information and the comments of Prof. Rainer Brüggemann at the Institute of Fresh Water Ecology and Inland Fisheries in Berlin and Peter Sørensen at the National Environmental Research Institute in Denmark.

This Thesis has been financially supported by European BEAM project (Bridging Effect Assessment of Mixtures to Ecosystem Situations and Regulation: R&D project “Beam”, EVK1-CT1999-00012). Prof.sa Paola Gramatica of the QSAR Research Unit, (Dep. of Structural and Functional Biology, University of Insubria), leading scientist of Partner 4 (Insubria University) in the BEAM Project, is acknowledged for giving me the opportunity to perform this study with a PhD fellowship.

CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

INTRODUCTION

Over the last century multivariate statistics have become an important tool to perform data analysis and, in recent years, its development has been mainly oriented towards mathematics and, therefore, towards the technical aspects of data analysis. With the advent of computers and the 'information age', statistical problems have grown in both size and complexity, and new fields have arisen, like data mining, chemometrics, chemoinformatics, bioinformatics. Two main aspects are faced by statistics: data exploration, which means learning from data, and data modelling.

Experiments and measurements are performed with the aim of analysing the variance of elements, measuring the distance among the elements and investigating their order relationships. Several techniques are now available for data exploration purposes. Principal Component Analysis (PCA) is one of the main methods for performing data analysis: it creates new axes to explain, to the greatest degree possible, the variance of the data matrix; furthermore it can be used to study element relationships, discovering outliers by score plots, i.e. projections of the elements in newly defined axes. These graphs allow the analysis of element relationships and element distribution in the reduced space of the principal components. Another reduced-space ordination method is multidimensional scaling: this starts with a scaling of the elements into a full-dimensional space, representing them in several dimensions and preserving their distance relationships.

Several Clustering methods are available to study the distance between elements or their similarity. Different criteria can be used to establish whether elements are close enough i.e. similar enough to be located within the same group or cluster, and different definitions of cluster are provided by different cluster measures.

Another way to perform data exploration is by rank methods which analyse the order relationship among elements. The different kinds of order methods available can be roughly classified as total (called even-scoring) and partial-order ranking methods, according to the specific order they provide. These methods are the ones needed to support and solve decision problems, setting priorities. Besides sophisticated multivariate statistics, used mostly in pre-processing and modelling data, priority setting makes use of quite simple methodologies. However the increasing of problem complexity leads to the decision processes becoming more complex, requiring the support of new tools. Thus there has been increased interest in decision making strategies and several techniques have been proposed. The intrinsic complexity of the systems analysed in chemistry research, and the multiplicity of objectives involved like economic efficiency, environmental quality and availability of resources, has led to complex multicriteria decision problems.

A decision problem is a situation where an individual has alternative courses of action available and has to select one, without any *a priori* knowledge of which is the best. A decision process can be organised into three phases: the identification phase, which consists in the recognition of the problem and in the diagnosis of the cause-effect relationships for the decision situation; the development phase, which results in a search routine to find ready-made solutions and the selection phase, which consists in a screen routine, when the search generates more solutions, in an evaluation routine and in a judgement choice. The decision process, which results in the selection of the best solution, is efficient if the procedure to reach the solution is optimal. The aims of a decision process are (a) to generate effective information on the decision problem from available data, (b) to generate effective solutions and (c) to provide a good understanding of the structure of a decision problem. MultiCriteria Decision Making (MCDM) strategies are used to rank various alternatives (scenarios, samples, objects, etc.) on the basis of multiple criteria, and are also used to make an optimal choice among these alternatives. In fact, the assessment of priorities is the typical premise before a final decision is taken. Decision support systems are computer-based systems, which assist individuals in the decision process and support judgement decision, improving the effectiveness of the decision process. Thus the focus is on the high quality of the strategy rather than on the quality of the final solution.

In recent years ranking strategies have been widely applied for different purposes: evaluation of aquatic toxicological tests [Bruggermann *et al*., 1997a; Bruggermann *et al*., 1995a], analysis of waste disposal sites [Halfon, 1989], ranking chemicals for environmental hazard [Halfon and Reggiani, 1986; Newman, 1995], comparison among ecosystems [Bruggerman *et al*., 1994; Munzer *et al*., 1994; Pudenz *et al*., 1997; Bruggermann *et al*., 1999a; Pudenz *et al*., 2000], chemicals priorization [Bruggermann *et al*., 1993a], evaluation of on-line databases [Bruggermann *et al*., 1997b; Voigt *et al*., 1999; Voigt *et al*., 2000], ranking of contaminated sites [Bruggermann *et al*.,1995b; SØrensen *et al*., 1998], evaluation of materials in car production [Pudenz et al., 1999].

In the complex systems evaluated by ranking strategies, elements (chemical substances, chemical processes, regions,...) are described by several attributes, referred to also as the criteria; thus the system must be analysed by more than one criterion, and decisions must be made by taking several criteria into account contemporaneously.

The criteria are any set of attributes which must reliably represent the system required properties and which must be orientable, i.e. for each criterion it is necessary to explicitly ascertain whether the best condition is satisfied by a minimum or maximum value of the criterion.

Let us now consider an *R*-dimensional system, with an associated (*N* x *R*) data matrix **X**. To each of the *N* elements a set of *R* attributes, criteria relevant to the decision making procedure, is associated. Each criterion can then be weighted to take account of the different importance of the criteria in the decision rule. The strategies to reach the optimal choice require the development of a ranking of the different options. Within a set E ($s, t, w, z \in$ E) a ranking (order) on E is a relation with the following properties:

$s \leq s$	reflexivity
$s \leq t$ and $t \leq s \;\Rightarrow\; t = s$	antisymmetry
$s \leq t$ and $t \leq z \;\Rightarrow\; s \leq z$	transitivity

A set E equipped with the relation ≤ is said to be an ordered set. Therefore the evaluation and even the ranking consists of two major steps: providing attributes and combining them. An evaluation method can generate:

- a complete or total ranking: s > t > w > z also called a linear order
- the best option: s > (t, w, z)
- a set of acceptable options: (s, t, w) > z
- an incomplete ranking of options s > (t, w, z) or (s, t) > (w, z).

Total and Partial order ranking (POR) strategies, which from a mathematical point of view are based on elementary methods of Discrete Mathematics, appear an attractive and simple tool to perform data analysis. A complete evaluation by the ranking technique requires a pre-processing phase to establish an adequate data matrix, and a post-processing phase to extract information and decisions on the system investigated. Obviously both pre-processing and post-processing may influence the results significantly. Pre-processing statistical techniques like Clustering, Principal Component analysis, Multidimensional Scaling and broad order statistics have been analyzed and compared with respect to their capability of supporting ranking methods. The analysis performed pointed out that broad order statistics seems to be a very suitable pre-processing tool, providing a satisfactory solution to those drawbacks related to noise and measurement error.

Total and partial order rankings can be analysed to establish the quality of the result obtained. As is usual for regression and classification strategies, the quality of a ranking procedure has to be evaluated by a deep analysis and by several indices, i.e. scalar functions which describe features of an ordered set and allow comparison among different rankings. Thus, the post-processing phase mainly consists in evaluating the quality of the ranking procedure by calculating ranking indices. For this purpose, new indices for ranking analysis are proposed here, and compared with those found in the literature and tested on both theoretical and real data. A preliminary analysis of the relationships among the ranking indices was performed by Principal Component Analysis, and a few ranking index classes were identified. This analysis revealed the new indices as being suitable to represent the main ranking properties and to encode unique information.

Moreover order ranking methods 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 statistical methods such as multi-linear regression (MLR) or classification. When data material is characterised by uncertainties, order models can be used as an alternative to statistical methods such as multi-linear regression (MLR), since they do not require a specific functional relationship between the independent variables and the dependent variables (responses). Moreover in several chemical and environmental problems the aim is to define order relationships among several chemicals, to indicate the more hazardous ones and to set priorities before final decisions are taken. For these purposes order ranking models, which allow not the finding of the quantitative response but the inter-relationships for each chemical, can be a promising approach in supporting decision making processes.

The development of an order ranking model has been investigated. To develop an order ranking model, the element ranking based on element responses (experimental ranking) is compared to the ranking based on independent variables (model ranking). If the model ranking is in agreement with the experimental ranking of the responses under investigation then the predictions of the ranking of other elements not yet investigated experimentally can be performed by the ranking model. As an exhaustive search for the best ranking models within a wide set of variables requires extensive computational resources and is time consuming due to the extremely high number of possible variable combinations, the Genetic Algorithm (GA-VSS) approach is proposed here as the variable selection method. Models based on the selected subsets of variables are compared with the experimental ranking, and evaluated in both total and partial ranking by the different parameters that measure the agreement of the two rankings.

Only the best quality models are retained in the population undergoing the evolution procedure. After a few iterations, the evolving population is usually composed of different combinations of variables that correlate well with the experimental ranking.

Total and partial ranking optimisation parameters have been investigated, and the new one proposed has been compared with those already published in the literature. Prediction calculations by ranking models have been analysed deeply and an approach is proposed here, together with a few measures for prediction precision. Moreover, the model ranking approach has been compared with traditional multilinear techniques in order to highlight the main advantages and disadvantages of these approaches.

Applications of the ranking data exploration, as well as ranking models, have been investigated and illustrated. The ranking approach has been tested on data coming from different fields: they are both real data provided by scientific collaborations and data published in literature. The case studies here described have been chosen in order to explain and verify some of the theoretical aspects introduced in the thesis.

CHAPTER 1

Total Ranking Theory

Total order ranking methods are multicriteria decision making techniques used for the ranking of various alternatives on the basis of more than one criterion. A criterion is a standard by which the elements of the system are judged. Criteria are not always in agreement, they can be conflicting, motivating the need to find an overall optimum that can deviate from the optima of one or more of the single criteria.

Total order ranking methods are based on an aggregation of the criteria y_r, where r = 1, ..R:

$$\Gamma \equiv f\,(y_1, y_2, ...y_R)$$

Thus, if an element is characterised by R criteria, then a comparison of different elements needs a scalar function, i.e. an order or ranking index, to sort them according to the numerical value of Γ. Several evaluation methods which define a ranking parameter generating a total order ranking have been proposed in the literature [Keller and Massart, 1991; Hendriks *et al*., 1992; Lewi *et al*., 1992]; those more frequently used are *Pareto Optimality*, *Desirability functions*, *Utility functions*, *Dominance functions*, *Concordance Analysis and Absolute Reference method*.

Most of these methods require the definition of the values and situations of optimum, i.e. for each criterion it is necessary to ascertain explicitly if the best condition is satisfied by a minimum or a maximum criterion value, and the trend from the minimum to the maximum must also be established.

The attribute setting is a crucial point in ranking methods since it requires the mathematization of decision criteria which are often not completely defined or explicit.
Total order ranking results are strictly dependent on the criteria setting and thus can be completely different for different settings.

1.1 Pareto Optimality

Pareto optimality is a multicriteria decision making method introduced into chemometrics by Smilde et al [Smilde et al., 1986]. The Pareto optimality technique selects the so-called Pareto-optimal points and the points that are not Pareto-optimal points are inferior to the Pareto-optimal points with respect to at least one criterion. Let us consider a two-dimensional criterion space like the one in Figure 1.1.

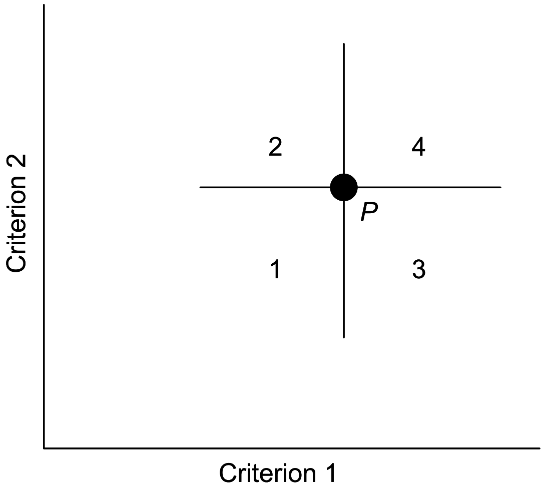


Figure 1.1 – Representation of the four quadrants in a two-dimensional criterion space around the point P.

A point corresponds to one setting of two criteria, the criterion values of which are plotted against each other. The space around the point P can be divided in four quadrants. In the case of two criteria both to be maximised, the points in the first quadrant are inferior to point P, while points in the fourth quadrant are superior to point P. The points in the second and third quadrants are incomparable with point P since they are superior to P for one criterion and inferior for the other. By definition, a Pareto optimal point is superior to all other comparable points, thus in the case of Figure 1.2 representing the space of two criteria Y_1 and Y_2, both to be maximised, a point a is superior to another point b if the following conditions are verified:

$$Y_{1a} > Y_{1b} \quad \text{and} \quad Y_{2a} > Y_{2b} \quad \text{or}$$

$$Y_{1a} > Y_{1b} \quad \text{and} \quad Y_{2a} = Y_{2b} \quad \text{or}$$

$$Y_{1a} = Y_{1b} \quad \text{and} \quad Y_{2a} > Y_{2b}$$

In other words, a point is a Pareto optimal point if no other points are found in the upper right quadrant. According to Pareto optimality, at least one point must be Pareto optimal, and all the non-inferior and incomparable points together form a set of Pareto-optimal (PO) points.

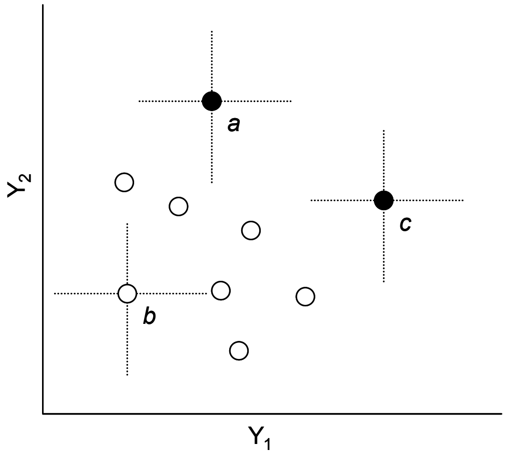


Figure 1.2 – Bivariate representation of the criteria Y1 and Y2. Points a and c are Pareto optimal points.

If the system under study is described by more than two criteria, the R-dimensional criterion space (R > 2) containing the Pareto optimal points must be projected onto a two dimensional plane. Through Principal Component Analysis (PCA) of the matrix containing the PO points, and following projection of the scores, it is possible to investigate the criterion space graphically.

1.2 Desirability and Utility functions

1.2.1 Desirability functions

Desirability functions are a well-known multicriteria decision making method. The approach is based on the definition of a desirability function for each criterion in order to transform values of the criteria to the same scale. Different kinds of functions can be used, the more common ones being linear, sigmoid, logarithmic, exponential, step, normal, parabolic, Laplace, triangular and box.

Each criterion is independently transformed into a desirability d_{ir} by an arbitrary function which transforms the actual value of each element into a value between 0 and 1:

$$d_{ir} = f_r(y_{ir}) \qquad 0 \le d_{ir} \le 1 \qquad r = 1,2,\ldots,R.$$

r being the selected criterion, *f* the function chosen and y_{ir} the actual value of the *i-th* element for the *r-th* criterion.

Once the kind of function and its trend for each criterion is defined, the global desirability *D* of each *i-th* element can be evaluated as follows:

$$D_i = \sqrt[R]{d_{i1} \cdot d_{i2} \cdot \ldots \cdot d_{iR}} \qquad 0 \le D_i \le 1$$

The overall desirability is calculated combining all the desirabilities through a geometrical mean. It must be highlighted that the desirability product is very strict: if an element is poor with respect to one criterion, its overall desirability will be poor. If any desirability d_i is equal to 0 the overall desirability D_i will be zero, whereas the D_i will be equal to one only if all the desirabilities have the maximum value of one.

In addition each criterion can be weighted in order to take into account criterion importance in the decision rule. In the case of weighted desirability functions the overall desirability of the *i-th* element is defined as follows:

$$D_i = (d_{i1}^{w_1} \cdot d_{i2}^{w_2} \cdot \ldots \cdot d_{iR}^{w_R})^{\sum_r w_r} \qquad 0 \leq D_i \leq 1$$

w_r being the weight of the *r-th* criterion and $\sum_{r=1}^{R} w_r = 1$.

Once *D* for each element has been calculated, all the elements can be ranked according to their *D* value and the element with the highest *D* can be selected as the best one, if its *D* value is acceptable. A *Desirability scale,* shown in Table 1.1, was developed by Harrington [Harrington, 1965] :

Scale of D	*Quality evaluation*
1.00	Improvement beyond this point has no preference
1.00 – 0.80	Acceptable and excellent
0.80 – 0.63	Acceptable and good
0.63 – 0.40	Acceptable but poor
0.40 – 0.30	Borderline
0.30 – 0.00	Unacceptable
0.00	Completely unacceptable

Table 1.1 – Harrington qualitative definition of the Desirability scale.

The critical feature of this approach to multicriteria decision making problems is the establishment of the relation between criteria and desirabilty values which must be performed by the decision maker.

1.2.2 Utility functions

The approach is very similar to the desirability functions; each criterion is independently transformed into a utility u_r by a function which transforms the actual value of each element into a value between 0 and 1.

$$u_{ir} = f_r(y_{ir}) \qquad 0 \le u_{ir} \le 1$$

r being the selected criterion, *f* the function selected and y_{ir} the actual value of the *i-th* element for the *r-th* criterion.
Once the kind of function and its trend for each criterion has been defined, the overall Utility *U* of each *i-th* element is defined as:

$$U_i = \frac{\sum_{r=1}^{R} u_{ir}}{R} \qquad 0 \le U_i \le 1$$

In the case of weighted utility functions the overall utility is calculated as:

$$U_i = \sum_{r=1}^{R} w_r \cdot u_{ir} \qquad 0 \le U_i \le 1$$

with $\sum_{r=1}^{R} w_r = 1$.

In this case the overall utility is calculated less severely: in fact the overall quality of an element can be high even if a single utility function is zero.
Like the desirability functions, the utility functions are affected by arbitrariness related to the *a priori* selection of the functions and corresponding upper and lower limits. Both desirability and utility functions are very easy to calculate, thus specific software is not required.

1.3 Dominance functions

The dominance function method is based on the comparison of the state of the different criteria for each pair of elements. This approach does not require the transformation of each criterion into a quantitative function, it has only to be established whether the best condition is satisfied by a minimum or maximum value of the selected criterion.

For each pair of elements (*i*, *j*) three sets of criteria are determined:
R^+(*i,j*) is the set of criteria w^+ where i dominates *j,* i.e. where *i* is better than *j*, $R^0(i,j)$ is the one where *i* and *j* are equal, and $R^-(i,j)$ is the set of criteria w^- where *i* is dominated by *j*.
The dominance function between two elements i and j is calculated considering the weights as follows:

$$C_{ij} = \frac{1+\sum_{R^+} w^+}{1+\sum_{R^-} w^-} \qquad 0.5 \le C_{ij} \le 2$$

with $\sum_{r=1}^{R} w_r = 1$.

A C_{ij} value equal to 1 means equivalence of the two elements; $C_{ij} > 1$ means that the element *i* is, on the whole, superior to the element *j*, whereas $C_{ij} < 1$ means that the element *i* is, on the whole, inferior to the element *j*. The obtained values can be normalised according to:

$$C'_{ij} = \frac{C_{ij} - 0.5}{2 - 0.5} \qquad 0 \le C'_{ij} \le 1$$

A global score of the *i-th* element is then calculated as:

$$\Phi_i = \sum_j C'_{ij} \qquad 0 \le \Phi_i \le N - 1$$

and the corresponding *i-th* scaled value is:

$$\Phi_i^{'} = \frac{\Phi_i}{N-1} \qquad 0 \le \Phi_i^{'} \le 1$$

Elements with higher values of Φ' are the optimal points.

1.4 Preference functions

The preference function ranking method was developed by Brans, Vincke and Mareschal [Brans and Vincke, 1985; Brans et al., 1986]. This approach uses subjective preference functions for each separate criterion to rank the different elements. However, differently from the desirability and utility functions, the preference function trend does not directly model the element values for each criterion, it models the difference values between each pair of elements. Thus for each r-th criterion, a preference function $P_r(i.j)$ must be defined for the difference between the function values of two elements (δ_{ij} = $f(i) - f(j)$). The preference function $P_r(i.j)$ defines the degree to which the *i-th* element is preferred to the *j-th* element and is constructed according to the following rules:

$$P_r(i,j) = 0 \qquad if\ \delta_{ij} < 0$$

$$P_r(i,j) = 1 \qquad if\ \delta_{ij} \ge \delta_r$$

$$P_r(i,j) = f_r({}^{\delta_{ij}}\!/\!_{\delta_r}) \qquad if \quad 0 < {}^{\delta_{ij}}\!/\!_{\delta r} < 1$$

where δ_r is the quantification of the outranking difference value required for the r-th criterion between two elements. If the difference between the two elements, *i* and *j*, is greater than or equal to the δ_r value, then the *i-th* element is strictly preferred to the *j-th* element; if it is less than 0, no preference exists and the two elements do not differ. In the other cases the preference value is provided by the function itself, the term being defined as $P_r(i,j) = f_r({}^{\delta_{ij}}\!/_{\delta_r})$. In a second step, a preference index *Π(i,j)* of element *i* over *j* for all the criteria, is calculated as:

$$\prod(i,j) = \sum_{r=1}^{R} w_r \cdot P_r(i,j)$$

where *R* is the total number of criteria, w_r the weight for the *r-th* criterion with $\sum_{r=1}^{R} w_r = 1$. The *Π(i,j)* values range from 0 to 1 indicating the global preference of *i* to *j*. In a third step, the positive flow and negative flow outranking for each element is calculated:

$$\Phi_i^+ = \sum\nolimits_{r^+} \prod(i,j)$$
$$\Phi_i^- = \sum\nolimits_{r^-} \prod(i,j)$$

The former measures how the *i-th* element outranks all the other elements and the sum runs over all the criteria favourable to *i*; the latter measures how the *i-th* element is outranked by all the other elements and the sum runs over all the criteria not favourable to *i*.

The global quality, called net flow outranking, of the *i-th* element is then calculated as:

$$\Phi_i = \Phi_i^+ - \Phi_i^-$$

and the calculated values are normalised according to:

$$\Phi_i^{'} = \frac{\Phi_i + (N-1)}{(N-1)+(N-1)}$$

N being the total number of elements, and (*N* - 1) and – (*N* - 1) respectively the maximum and minimum values of Φ_1

1.5 Concordance Analysis

The use of Concordance Analysis was introduced by Opperhuizen and Hutzinger as a multicriteria decision making method for the priority setting of chemicals [Opperhuizen and Hutzinger, 1982]. The main difference between Concordance Analysis and Desirability, Utility and Dominance functions is the introduction of a reference element to which each element is compared. The reference element can be a real element or a fictitious one: the centroid, i.e. the vector of the means, is frequently used as the fictitious reference element.

Because of the different dimensions of the criteria, each criterion first undergoes normalisation, and each is weighted according to its importance in the decision process. For each criterion the normalised value is compared with the normalised value of the reference element.

For each element Concordance and Discordance sets are defined. The Concordance set *ConSet*$_i$, related to the *i-th* element, is composed by those criteria for which the *i-th* element has values higher than those of the reference element i*:

$$ConSet_i = \left\{ \forall r \,\middle|\, (y_{ir} > y_{i^*r}) \cdot w_r \right\}$$

The Discordance set *DiscSet*$_i$, related to the *i-th* element, is composed by those criteria for which the *i-th* element has values lower than or equal to those of the reference element *i*:*

$$DiscSet_i = \left\{ \forall r \,\middle|\, (y_{ir} \le y_{i^*r}) \cdot w_r \right\}$$

For each element a Concordance Indicator *CI*$_i$, which measures the number of criteria for which the *i-th* element is preferred to the reference element, is calculated as the sum of the weights belonging to the criteria of the Concordance set, *ConSet*$_i$:

$$CI_i = \sum\nolimits_{r \in ConSet_i} w_r \qquad 0 \le CI_i \le 1$$

Similarly a Discordance Indicator DI_i, which quantifies not only the number of criteria with a worse *i-th* element than the reference element but also how much worse it is, is calculated as the weighted maximum difference between the criteria of the Discordance set and those of the reference element:

$$DI_i = max_{r \in DiscSet_i} \left| (y_{ir} - y_{i^* r}) \cdot w_r \right|$$

The maximum is taken over all the criteria of the Discordance set. The elements are ranked according to the global score Γ_i:

$$\Gamma_i = CI_i - DI_i$$

Since both CI_i and DI_i range from 0 to 1, the global scaled score Γ_i of the *i-th* element is calculated as:

$$\Gamma_i = \frac{CI_i - DI_i + 1}{2} \qquad 0 \le \Gamma_i \le 1$$

It must be pointed out that the Concordance Indicator, as defined in the classical Concordance Analysis proposed by Opperhuizen and Hutzinger [Opperhuizen and Hutzinger, 1982], is a measure of the number of criteria for which each element is preferred to the reference element, since the Indicator is defined as the sum of the weights belonging to the criteria of the Concordance set, however no account is taken of the real quantitative distance between the two elements.

A new and quantitative Concordance Indicator $CI_i^{'}$, which measures not only for how many criteria the *i-th* element is preferred to the reference element but also how much it is preferred, is proposed here as the sum of the weighted differences between the criteria of the Concordance set and those of the reference element:

$$CI_i^{'} = \sum\nolimits_{r \in ConSet_i} \left[1 - (y_{ir} - y_{i^*r})\right] \cdot w_r$$

If the centroid has been taken as the reference element, CI_i ranges from 0 to 0.5, and DI_i from 0 to half the maximum weight value ($Max_r \{w_r\}$), thus the global scaled score $\Gamma_i^{'}$ of the *i-th* element is calculated as:

$$\Gamma_i^{'} = \frac{CI_i^{'} - DI_i + 1}{2} \qquad 0 \le \Gamma_i^{'} \le 1$$

1.6 Absolute reference method

The absolute reference method is based measuring the distance between each element and a reference element, which is supposed to represent the overall optimum of all the considered criteria. This method requires the definition of the values and situations of optimum, i.e. for each criterion it is necessary to explicitly ascertain not only whether the best condition is satisfied with a minimum value or a maximum value of the criterion, but also the specific optimum values. To get rid of different criterion dimensions, each criterion first undergoes normalisation and weighting to account for its importance.

Once a distance measure has been selected, the Absolute reference method calculates the entire *N* distances between the elements and the reference element. If the Euclidean distance is selected, the distance of the *i-th* element from the reference element (*i**) is defined as:

$$d_{ii^*} = \sqrt{\sum_{r=1}^{R} (y_{ir} - y_{i^*r})^2 \cdot w_r}$$

For each element a measure of its similarity with the reference element is derived from the Euclidean distance according to the following expression:

$$S_i = 1 - d_{ii^*} \qquad 0 \le S_i \le 1$$

This similarity measure is used to rank the elements. It ranges from 0 (no similarity exists between the considered element and the reference one) and 1 (there is complete similarity between the considered element and the reference one).

1.7 Comparison of total order ranking methods

A comparison of the total order ranking methods described above is performed on a dataset of twelve High Production Volume Chemicals (HPVC) found in the IUCLID database [European Communities, 2000]. The environmental impact of the pesticides is described by four criteria: production volume (PV), as indicator of exposure, acute toxicity to fish (LC50), as indicator of toxicity, partitioning coefficient between n-octanol and water (LogKow), as indicator of the tendency to bioaccumulate in biota, and biodegradation as indicator of the persistence of the substance in the environment. The decision matrix analysed by the multicriteria decision making methods is shown in Table 1.2.

Substance	*Abbreviation*	*PV**	*LC50*	*LogKow*	*BD(%)*
1-Chloro-4-nitrobenzene	CNB	4	1.5	2.6	0.2
4-Nitroaniline	4NA	2	35	1.4	0
4-Nitrophenol	4NP	1	7	1.9	0.1
Atrazin	ATR	2	4.3	2.5	0.5
Chlormequat chlorid	CHL	2	80	-2.2	1
Diazinon	DIA	1	2.6	3.3	0
Dimethoate	DIM	2	7.5	0.7	0
Ethofumesate	ETH	1	11	2.7	0.4
Glyphosate	GLY	2	52	0.002	0.3
Isoproturon	ISO	2	3	2.5	30
Malathion	MAL	3	0.04	2.7	100
Thiram	THI	2	0.3	1.7	0

Table 1.2 – Data on production volume (PV), Acute Toxicity for fish (LC50), n-octanol – water coefficient (LogKow) and biodegradation for 12 pesticides. * 1 = 5.000 – 10.000 tons/year; 2 = 10.000 – 50.000 tons/year; 3 = 50.000 – 100.000 tons/year; 4 = 100.000 – 500.000 tons/year.

The environmental impact of the pesticides was studied: since a hazard substance, i.e. one with a high environmental impact, is characterised by low acute toxicity for fish (LC50) and low biodegradation a linear inverse transformation has been applied to these two criteria, whereas for production volume and n-octanol – water coefficient (logKow) a linear transformation was applied as high values determine high environmental impact. The criteria were weighted equally (w_r = 0.25) and the preference function values calculated with a delta value δ_r equal to half the standard deviation of the *r-th* criterion. The rankings obtained by Desirability, Utility, Dominance, Preference functions, Concordance analysis and Absolute reference are shown in Table 1.3.

Sub.	*Des.*	*Uti.*	*Dom.*	*Pref.*	*ConcA*	*ConcQ*	*Abs R.*
CNB	0.96	0.96	0.67	0.76	1.00	0.93	1.00
4NA	0.59	0.64	0.34	0.41	0.74	0.39	0.59
4NP	0.00	0.66	0.28	0.42	0.67	0.51	0.50
ATR	0.72	0.78	0.36	0.61	0.74	0.65	0.67
CHL	0.00	0.33	0.09	0.28	0.22	0.22	0.26
DIA	0.00	0.74	0.54	0.59	0.67	0.65	0.50
DIM	0.63	0.69	0.36	0.46	0.74	0.49	0.62
ETH	0.00	0.69	0.26	0.47	0.67	0.56	0.50
GLY	0.47	0.52	0.20	0.33	0.30	0.30	0.48
ISO	0.66	0.71	0.33	0.50	0.74	0.53	0.64
MAL	0.00	0.64	0.54	0.61	0.67	0.47	0.47
THI	0.70	0.76	0.55	0.57	0.74	0.61	0.66

Table 1.3 – Rankings obtained by Desirability, Utility, Dominance, Preference functions, classical and quantitative Concordance Analysis, (ConcA and ConcQ), and Absolute reference method.

According to the defined criterion settings, high values of the global ranking index correspond to hazard pesticides. Figure 1.3 shows the comparison of the rankings obtained by desirability and utility functions: it can be observed that they are quite different, the utility approach being more capable in discriminating elements. The most significant differences are 4-nitrophenol (4NP), diazinon (DIA) and ethofumesate (ETH) whose desirability overall index is equal to 0, meaning that they are not hazard pesticides, as they are not hazardous according to all the considered criteria, having quite a low value of production volume. Chlormequat chlorid (CHL) and malathion (MAL) are not considered hazards as they respectively have a low LogKow value and high biodegradation. Thus if a substance has a very low value of one criterion its rank may vary significantly and if the criteria have different weights another rank will occur.

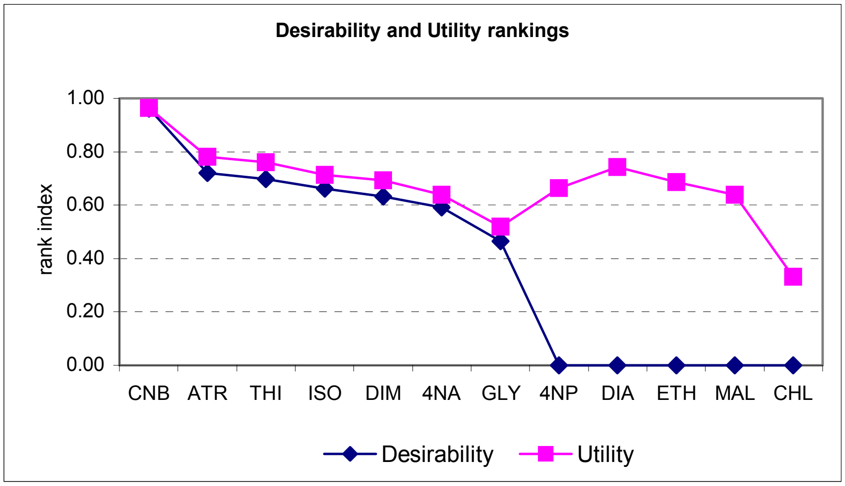


Figure 1.3 – Graphical comparison of the obtained Desirability and Utility rankings: the elements are sorted according to the Desirability index.

The rankings obtained by dominance and preference functions have been compared in Figure 1.4: it can be observed that they are quite similar, both these approaches being based on pair element comparison. The preference overall index is always greater than the dominance one, thus it estimates a higher hazard pesticide than the dominance approach. The most relevant discrepancies are detected for atrazin (ATR), isoproturon (ISO) and ethofumesate (ETH) which are considered significantly hazardous according to their preference index, whereas they are not of priority attention according to their dominance index value. In contrast to the dominance function, the overall ranking index derived from the preference function further depends on how the preference functions have been formulated. If δ_r is small the preference function becomes independent of the metric used, whereas if δ_r is large, the metric value becomes more relevant.

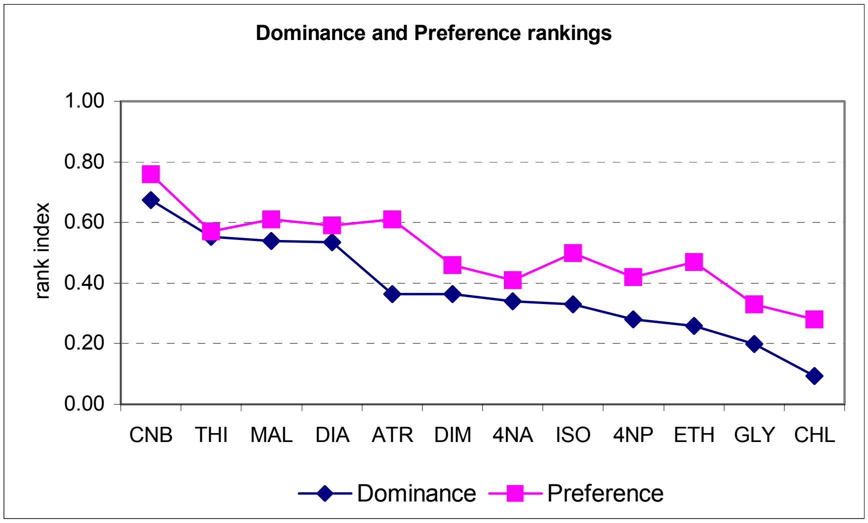


Figure 1.4 – Graphical comparison of the obtained Dominance and Preference rankings: the elements are sorted according to the Dominance index.

Both classic concordance analysis and the new quantitative one were performed using, as the fictitious reference element, the centroid i.e the vector of the means. Figure 1.5 shows the rankings by classical concordance analysis and quantitative analysis. The rankings differ and it is to be noted that classical concordance analysis is unable to differentiate the final ranking results as it only sums the weights of the elements above the reference element, and subtracts the weighted difference from the criterion, the lowest with respect to the reference element. While classical concordance analysis identifies only 5 levels (different values of the ranking index) for the twelve pesticides, the quantitative concordance approach is able to distinguish all the pesticides, showing less degeneracy than the classical approach.

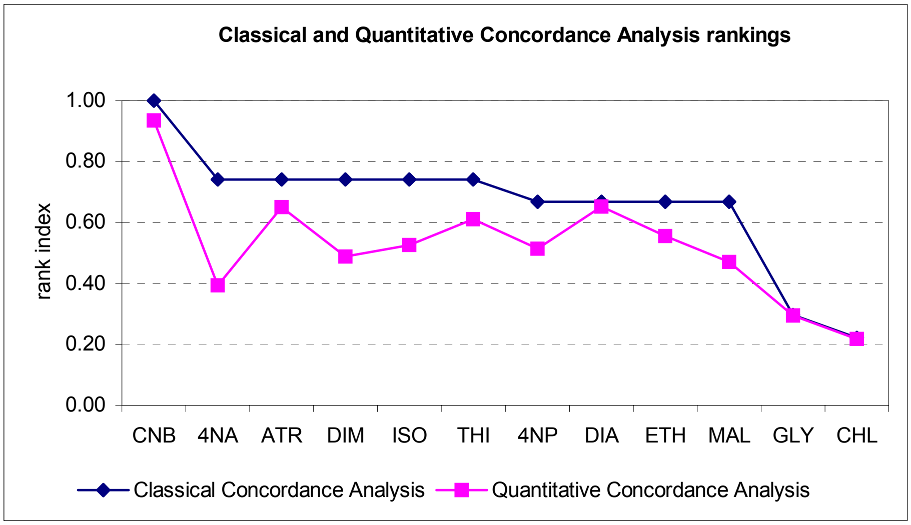


Figure 1.5 – Graphical comparison of the obtained Classical and Quantitative Concordance Analysis rankings: the elements are sorted according to the Classical Concordance Analysis.

By the absolute reference method, Figure 1.6, it can be seen that if the element with the highest environmental impact is selected as the reference element (CNB) the absolute reference method calculates the distance of all the other elements from the reference element, and all the other elements are supposed to have a lower impact than the reference one.

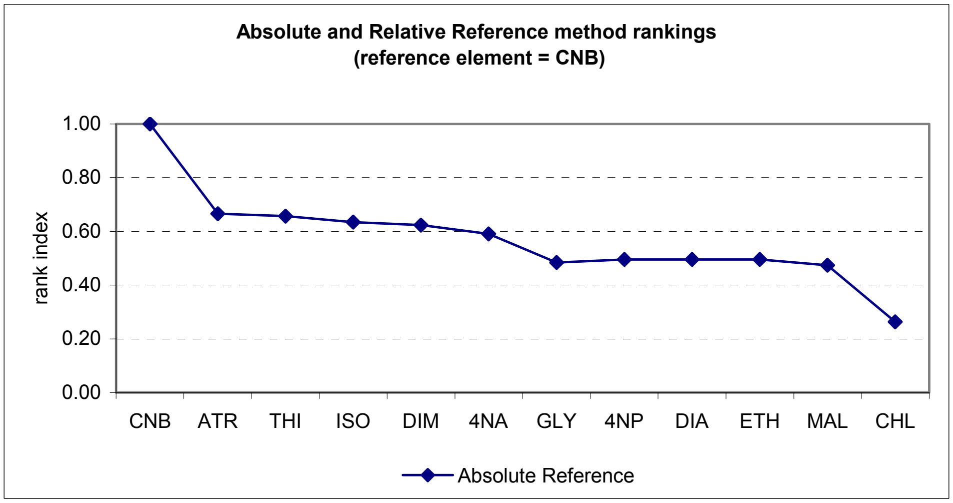


Figure 1.6 – Graphical trend of the obtained Absolute Reference ranking: the elements are sorted according to the Absolute Reference rank.

When comparing total ordered ranking methods, it is important to compare even the additional external information required. All the methods are based on a first level of subjectivity, concerning the criteria selected as representative of the system under investigation. Another level of subjectivity is added when the criteria are weighted, as this requires the identification of the more important criteria and the results are strictly influenced by the weight setting. Compared with the desirability, utility and dominance functions, the preference functions, the concordance analysis and the absolute reference approaches contain an additional level of subjectivity. The preference functions need information on a delta value δ_r, whereas concordance analysis, absolute reference approaches require the identification of a reference element. Moreover the absolute reference method is based on the assumption that no one element can be better than the reference one. Thus all the total order methods are based on a set of assumptions: the desirability and utility approaches assume a numerical and often linear relation among criteria; dominance and preference functions assume the linear comparability of the criteria; the classical concordance analysis assumes not only the existence of a reference element, as the absolute reference method, but also that the criterion weights can be linearly additive. The assumptions which the total ranking methods are based on should be kept in mind when a priority setting method is chosen as a certain method may be applicable for a given problem, but not suitable for another one.

1.8 Rank correlation

All total order ranking methods are highly influenced by the criterion selection. The criteria chosen as representative of the system under studyi are often provided by the decision-maker. To verify the real necessity of all the selected criteria, as being relevant for the multicriteria decision problem, a preliminary analysis of the criterion correlation can provide useful information. Two rank correlation coefficients are available: the Spearman *r* and the Kendall τ [Kendall, 1948]. Both coefficients quantify the correlation relationship between two criteria.

According to the Spearman coefficient *r*, two criteria are perfectly correlated if they provide the same ranks for all the elements, and the difference between two ranks (d_i) is taken as a measure of the criterion difference for the element considered. For the whole set of elements, the rank differences are squared before summing them, in order to prevent differences with opposite signs from cancelling each other out. The general formula of the Spearman r coefficient is:

$$r_{rk} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} \qquad -1 \leq r_{rk} \leq +1$$

where d_i is the rank difference for the element *i* in the two criteria *r* and *k* and *N* is the total number of elements. This coefficient ranges between +1 and –1. Criteria perfectly directly correlated, in terms of rank, assume values *r* = +1; inversely correlated values *r* = -1 and criteria not correlated values *r* = 0.

The Kendall coefficient τ is based on the sums of scores for pairs of elements in increasing and decreasing order. In rank correlation analysis

Kendall defined a score for a pair of rankings of *N* items as +1 if any two are ranked in the same order by the two rankings, as -1 if in opposite order, and zero if tied to either or both rankings. The total score S is the algebraic sum of the ½ *N*(*N*-1) contributions from pairs of items.
To better explain this concept, a small numerical example is used. The ranked data of four elements described by two criteria (r_1 and r_2) are shown in Table 1.4.

Element	*Ranks on* r_1	*Ranks on* r_2
a	3	3
b	4	1
c	2	4
d	1	2

Table 1.4 – Numerical example. Ranks of four elements on two criteria r_1 and r_2.

The elements are first rearranged in increasing ranks on one of the two criteria (here r_1), as shown in Table 1.5:

Element	*Ranks on* r_1	*Ranks on* r_2
d	1	2
c	2	4
a	3	3
b	4	1

Table 1.5 – Numerical example. The element order is rearranged in increasing order on r_1.

As the ranks are in increasing order on r_1, in order to quantify the correlation between the criteria, there must be the determination of how many pairs of ranks are in increasing order also on r_2. Considering the element on the first rank on r_1 (*d*), the first pair of ranks (2 and 4 belonging to *d* and *c*) is in increasing order and thus a score of +1 is assigned to it. The same occurs for the second pair (2 and 3, belonging to d and a). The third pair of ranks (2 and 1, belonging to *d* and *b*) is in decreasing order and thus has a negative score of –1. This operation is repeated for all the elements in successive ranks along r_1 and the sum of scores S is calculated. Kendall's rank coefficient is the sum of scores for pairs in increasing and decreasing order, divided by the total number of pairs ($N \cdot (N-1)$) defined as:

$$\tau_{rk} = \frac{2S}{N(N-1)} \qquad -1 \leq \tau_{rk} \leq +1$$

For the numerical example of Table 1.4 and 1.5:

$$\tau_{rk} = \frac{2 \cdot (1+1-1-1-1-1)}{4 \cdot 3} = \frac{2 \cdot (-2)}{12} = -0.33$$

Kendall's rank coefficient ranges from +1 in the case of complete agreement to –1 in the case of complete disagreement. If the two criteria are uncorrelated, it takes a value of 0.

Both the Spearman and Kendall rank correlation coefficients measure the correlation between two criteria, based on *N* elements. In contrast, Kendall's coefficient of concordance *W* [Kendall, 1948] measures the relationship among several rank-ordered criteria. It is calculated on a data table which contains, in each column, the ranks of *N* elements of *R* criteria according to the following expression:

$$W = \frac{12 \cdot \sum_{i=1}^{N} (Q_i - \overline{Q})^2}{R^2 \cdot (N^3 - N)} \qquad 0 \leq W \leq 1$$

Q_i being the sum of ranks of element i, $\overline{Q}$ the average rank.

This coefficient ranges from 0 if no concordance exists among the criteria to 1 in the case of maximum concordance.

The calculation of Kendall's concordance is illustrated here by the numerical example of Table 1.6. Six elements are ranked separately according to each criterion. The last column, contains for each *i-th* element the sum Q_i of its ranks on the R = 3 criteria.

Elements	*Ranks of elements on*			*Row sums*
	r_1	r_2	r_3	Q_i
a	1	1	6	8
b	6	5	3	14
c	3	6	2	11
d	2	4	5	11
e	5	2	4	11
f	4	3	1	8

Table 1.6 – Numerical example. Ranks of six elements on three criteria r_1, r_2, r_3.

Kendall's concordance for the data of Table 1.6 is computed as:

$$W = \frac{12 \cdot (6.25 + 12.25 + 0.25 + 0.25 + 0.25 + 6.25)}{9 \cdot (216 - 6)} = 0.162$$

1.9 Indices for total ranking analysis

Total order ranking can be analysed to establish the quality of the result. As is usual for regression and classification strategies, ranking procedure quality has to be evaluated by an analysis deep enough to find the main characteristics of the ranking. This requires indices, i.e. scalar functions which describe the features of an ordered set, allowing comparison of the different rankings. A few new indices for ranking analysis are proposed here.

1.9.1 Information content and degeneracy degree

One of the first aspects to be analysed after having performed a ranking procedure is to measure the amount of information made available, and when calculating the information content of a total ordered ranking consideration must be given to the number of equivalent elements, i.e. elements of the same numerical value as the scalar function Γ which is the order or ranking index used to sort the elements.

The information content of a system having *N* elements is a measure of the degree of diversity of the elements in the set [Klir and Folger, 1988]; it is defined as:

$$I_C = \sum_{c=1}^{C} n_c \log_2 n_c$$

where C is the number of different equivalence classes and n_c is the number of elements in the *c-th* class and

$$N = \sum_{c=1}^{C} n_c$$

Each *c-th* equivalence class is built by the definition of some relationships among the elements of the system. The logarithm is taken at base 2 for measuring the information content in bits.

The information content is zero if the elements differ one from the other i.e. no equivalence relationship exists; in this case there are $C = N$ different equivalence classes. On the contrary, the information content is maximal if all the elements of the set are recognized as belonging to the same class ($C = 1$). This quantity is called the maximal information content ${}^{\max}I_C$ and represents the information content needed to characterize all the *N* elements of the considered set:

$${}^{\max}I_C = N \log_2 N$$

The total information content (or negentropy) of a system having *N* elements is defined by the following:

$$I = {}^{\max}I_C - I_C = N\log_2 N - \sum_{c=1}^{C} N_c \log_2 N_c = N \cdot H$$

The term *H* is Shannon's entropy, defined below.

The total information content represents the residual information contained in the system after *C* relationships are defined among *N* elements.

The mean information content $\bar{I}$, also called Shannon's entropy *H* [Shannon and Weaver, 1949] is defined as:

$$\bar{I} \equiv H = \frac{I}{N} = -\sum_{c=1}^{C} \frac{n_c}{N} \log_2 \frac{n_c}{N} = -\sum_{c=1}^{C} \mathrm{p}_c \log_2 \mathrm{p}_c$$

where p_c is the probability of randomly selecting an element of the *c-th* class, and *I* is the total information content.

The maximum value of the entropy is $\log_2 N$, obtained when $n_c = 1$ for all *C* classes; it is called Hartley information

$$I_N = \log_2 N$$

where *N* can be interpreted as the number of alternatives regardless of whether they are realized by one selection from a set or by a sequence of selections [Hartley, 1928]. Hartley information represents the information content needed to characterize one of the *N* elements.

The standardized Shannon's entropy (or standardized information content) is the ratio between the actual mean information content and the maximum available information content (i.e. the Hartley information):

$$H^* = \frac{H}{I_N} = \frac{H}{\log_2 N} = \frac{I}{N\log_2 N} \qquad 0 \le H^* \le 1$$

The standardised Shannon's entropy is a measure of the relative efficiency of the collected information, i.e. the mean information per unit.

From the mean information content Brillouin [Brillouin, 1962] defined a complementary quantity, called the Brillouin redundancy index *R* (or redundancy index), to measure the information redundancy of the system:

$$R = 1 - \frac{H}{\log_2 N} = 1 - H^*$$

Another measure of entropy is given by the Gini index G defined as:

$$G = \sum\nolimits_{c \neq c'} p_c \cdot p_{c'} \qquad p_c = \frac{n_c}{N} \quad p_{c'} = \frac{n_{c'}}{N} \qquad 0 \leq G \leq \frac{N-1}{2 \cdot N}$$

where p_c and $p_{c'}$ are the probabilities of randomly selecting an element of the *c-th* and *c'-th* different equivalence classes, respectively; its corresponding standardised version *G** is defined by its ratio with the maximum *G* value. The Gini index increases as the diversity of the system increases. A complementary quantity to the Gini index is the informational energy content [Onicescu, 1966]: defined as:

$$I_E = \sum\nolimits_c \mathrm{p}_c^2 \qquad \frac{1}{N} \leq I_E \leq 1$$

It corresponds to a redundancy measure whose maximum and minimum values are 1 and 1/*N*, respectively.

Another degeneracy index *k*(*N*) was proposed by Brüggemann [Bruggermann and Halfon, 1999b] to measure the degeneracy of an order ranking. First proposed for partial ordered rankings, but easily applicable to total ordered rankings, it is defined as:

$$k = \sum_{c=1}^{C} n_c \cdot (n_c - 1)$$

n_c being the number of elements of the *c-th* equivalence class and *C* the total number of equivalence classes.

The corresponding standardized index is:

$$k_{std} = \frac{\sum_{c=1}^{C} n_c \cdot (n_c - 1)}{N(N-1)} \qquad 0 \le k_{std} \le 1$$

Note that in the case of two equivalence classes, one containing five elements and the other only one element, the k_{std} index takes a value equal to 0.67, whereas in the case of two equivalence classes, each containing three elements, the k_{std} index takes a value equal to 0.40. Thus the more the degeneracy is shared among the equivalence classes, the less is the numerical value of k_{std} ; thus this index depends not only on the degeneracy degree but also on the degeneracy distribution in the equivalent classes.

To avoid degeneracy distribution dependency an absolute degeneracy degree (*D*) of a ranking is proposed here and defined as:

$$D = \frac{\sum_{c=1}^{C} \left(\frac{n_c}{N} - \frac{1}{N} \right)}{\frac{(N-1)}{N}} \qquad 0 \le D \le 1$$

The numerator represents the difference between the amount of degeneracy of each equivalence class and the case of total absence of degeneracy (uniform distribution); the denominator corresponds to the maximum value reached by the numerator and is used to scale the values between 0 and 1. Degeneracy *D* allows the evaluation of the non-uniformity or diversity of the element distribution; *D* takes a value of 1 when all the elements have the same value as the ranking parameter Γ, in which case the degeneracy is maximum and the total ranking method used is not able to differentiate the elements, i.e. the elements are correlated and only one equivalence class exists. On the other hand *D* takes the value of 0 for minimum degeneracy when all the elements differ from each other, and *N* equivalence classes exist, each with only one element. The greater the degeneracy, the less the diversity of the elements. To highlight the different behaviour of the indices measuring the information content or the complementary degeneracy of totally ordered sequences, a theoretical example is illustrated in Table 1.7. It contains five total rankings obtained for six elements.

Element	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5
a	0.96	0.74	0.81	0.98	0.83
b	0.96	0.38	0.81	0.82	0.77
c	0.96	0.38	0.81	0.82	0.52
d	0.96	0.38	0.54	0.66	0.41
e	0.96	0.38	0.54	0.66	0.38
f	0.96	0.38	0.54	0.30	0.10

Table 1.7 – Numerical example. Rankings of six elements.

For each ranking the standardised Shannon and Gini entropies (*H*, G**), the Brillouin redundancy index (*R*), informational energy content (I_E), the standardised Brüggemann degeneracy (k_{std}) and the absolute degeneracy index (*D*) have been calculated. Their values are shown in Table 1.8 and can be compared in Figure 1.7.

Ranking	*H**	*G**	*R.*	I_E	k_{std}	*D*
Γ_1	0.00	0.00	1.00	1.00	1.00	1.00
Γ_2	0.25	0.33	0.75	0.72	0.67	0.80
Γ_3	0.39	0.60	0.61	0.50	0.40	0.80
Γ_4	0.74	0.86	0.26	0.28	0.13	0.40
Γ_5	1.00	1.00	0.00	0.17	0.00	0.00

Table 1.8 – Numerical example of information content and degeneracy indices.

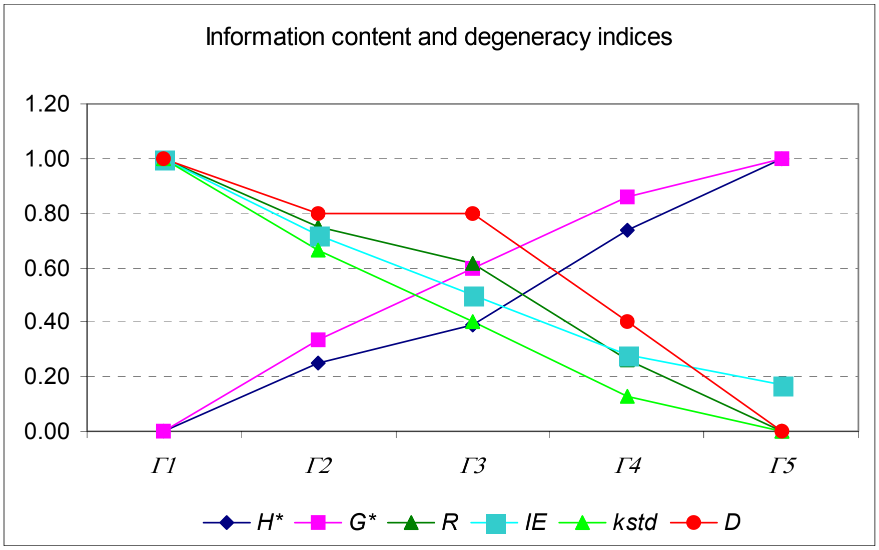


Figure 1.7 – Information content and degeneracy indices trend.

The first ranking (Γ_1) corresponds to the case of a system characterised by complete degeneracy: The two entropy indices, the standardised Shannon and Gini entropies, take values 0 and the degeneracy is maximum according to all four degeneracy indices. The fifth ranking (Γ_5) corresponds to the case of a system characterised by the absence of degeneracy: The two entropy indices take the maximum value of 1 whereas the degeneracy calculated by the four indices is equal to the minimum value (0 for R, k_{std} and D; $1/N$ for I_E). The second and third rankings (Γ_2 and Γ_3) correspond to two rankings with differently distributed degeneracy. The Shannon and Gini entropies take a greater value for ranking Γ_3 than for Γ_2. The degeneracy calculated according to the Brillouin index, the informational energy content and the standardised Brüggemann expression are all influenced by the way the degeneracy is distributed in the system, being greater in ranking Γ_2 and lower in Γ_3. In contrast, the absolute degeneracy index takes the same value for rankings Γ_2 and Γ_3, revealing the presence of two information sources in both the rankings. As far as concerns the index values for ranking Γ_4 it can be observed that the absolute degeneracy index calculates greater degeneracy than the others, the Brillouin index and informational energy content values being very similar to each other. It is to be pointed out that Bruggermann degeneracy is always the lowest, underestimating the degeneracy.

1.9.2 Discrimination power by ranking

In most of the cases, total ordered ranking methods are used in multicriteria decision-making problems with the aim of defining priorities. For this purpose, of particular relevance is the method capability of differentiating the elements with different values of the ranking parameter. The quality of an order set can be quantified by the index proposed here, called Discrimination power by Ranking (*DbyR*), which measures the capability of discriminating elements by a ranking according to the following expression:

$$DbyR = 1 - \frac{D}{L} \qquad 0 \le DbyR \le 1$$

D being the absolute degeneracy degree and *L* the number of Levels, i.e. the number of different values of the ranking parameter Γ.
This index ranges from value 0 for the case of all elements equal to each other, i.e. only one equivalent class (*D* = 1; *L* = 1), to 1 for the case of totally ordered sequence with no degeneracy (*D* = 0); and increases with the decreasing of the degeneracy index.
The Discrimination power by ranking values calculated for the theoretical example of Table 1.5 are 0 for Γ_1 (*D* = 1; *L* = 1), as the elements are not differentiated one from the other; 0.6 for Γ_2 and Γ_3 (*D* = 0.8; *L* = 2), 0.9 for Γ_4 (*D* = 0.4; *L* = 4), and 1 for Γ_5 (*D* = 0; *L* = 6), as the elements are totally separated in different ranks.

1.9.3 Stability index

A total ranking, performed by any whatever total ranking method, is strictly determined by the set of criteria used to describe the system, thus by changing the criteria different rankings arise. The set of criteria used may vary, and an additional criterion may be used. Thus it is of interest to forecast the effect on the ranking of increasing the number of considered criteria, i.e. to evaluate ranking stability.

The stability ranking index for a total ordered sequence, proposed here, is defined as:

$$StR = \frac{1-\sqrt{D}}{L} \qquad 0 \le StR \le 1/N$$

where D is the degeneracy index and L the number of levels.

This index allows the distinguishing of the case of totally ordered sequence with no degeneracy from the case of full degeneracy, in fact it ranges from 0 for full degeneracy to $1/N$, which is assumed as the stability of an ordered sequence of N elements.

The Stability index value for the ranking for Γ_1 of the theoretical example of Table 1.5 takes value 0: in the case of complete degeneracy it is strongly probable that the addition of one criterion will vary the ranking. Stability increases from Γ_2 to Γ_5 with decreasing degeneracy, with the following values: 0.05 for Γ_2 and Γ_3; 0.09 for Γ_4, and 0.17, (maximum value for a system of 6 elements) for Γ_5 .

CHAPTER 2

Partial Ranking Theory

Ordering is one of the possible ways to analyse data and to get an overview over the elements of a system. The elements are commonly characterised by more than one quantity, i.e. they are described by several variables. As a consequence of the multivariate property of the elements, their ordering requires specific techniques as "conflict" among the criteria is bound to exist. Total order ranking methods, being scalar methods, combine the different criteria values into an index, the ranking index Γ, and element comparison and ordering is performed according to the numerical value of Γ. In this way the elements are always ranked in a total or linear ordered sequence, but the information on conflict among criteria is inevitably lost. Partial order ranking is a vectorial approach that recognizes that not all elements can be directly compared with all other elements because, when many criteria are used, contradictions in the ranking can be present. An example could help to better understand what criteria conflict is. The system is made up of five, not perfectly correlated, elements (a, b, c, d, e), each described by two criteria r_1 and r_2, and the aim is to discover which element is better than the other with respect to all the criteria. The elements are sorted, arranging them according to r_1 and r_2 in the permutation diagram [Urrutia, 1987] or by parallel coordinates [Welzl *et al.*, 1998] with a vertical orientation, as shown in Figure 2.1.

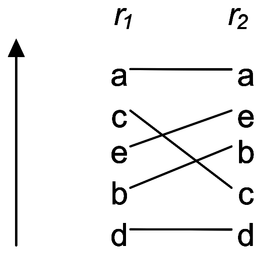


Figure 2.1 – Elements arranged in two sequences according to two different criteria.

This representation highlights the inversions between the two criteria. Elements mutually exchange their position according to the criterion used to sort them. Obviously the higher the number of criteria, the higher the probability that contradictions in the ranking exist. The partial ranking approach not only ranks elements but also identifies contradictions in the criteria used for ranking: some "residual order" remains when many criteria are considered and this motivates the term "partial order". Thus the more known concept of order is the one demanding that all elements be comparable i.e. linear or total order, while partial order is the one in which elements can be “not comparable”. If many elements are to be investigated, and especially if many criteria are to be considered, the parallel coordinates become complex and confusing. The Hasse diagram technique is a useful tool to perform partial order rankings with an easy visualisation of the obtained results.

2.1 Hasse Diagram Technique (HDT)

The Hasse diagram technique is a partial order ranking technique introduced in environmental sciences by Halfon [Halfon and Reggiani, 1986] and refined by Brüggemann [Brüggemann and Bartel, 1999c]. It is based on a specific order relation, named *product order*, and it provides a diagram, which visualises the results of the sorting.

In this approach the basis for ranking is the information collected in the full set of criteria, called even attributes, E, which is called the "*information basis*" of the comparative evaluation of elements.

The processed data matrix **Q** (*N* x *R*) contains *N* elements (rows) and *R* attributes (columns). The entry y_{ir} of **Q** is the numerical value of the *r-th* attribute of the *i-th* element. According to the product order relation, which the Hasse diagram technique is based on, IB being the information basis of evaluation and E the set of *N* elements, the two elements *s* and *t* are comparable if *for all* $y_r \in$ IB either $y_r(s) \leq y_r(t)$ or $y_r(t) \geq y_r(s)$. If $y_r(s) \leq y_r(t)$ for all $y_r \in$ IB then $s \leq t$.

The request "for all" is very important and is called the *generality principle*:

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

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

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 *not comparable*, and the common notation is $s \| t$. If only one attribute is used or all the attributes are perfectly correlated then total order is obtained, and all the elements are comparable.

Partial order is determined by the actual information base, thus by changing the information base (IB) different partial orders arise. Partial order sets can be developed easily with 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:

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

Thus according to the so-called *cover-relation*, if there is no element “*a*” of *E*, for which $s \leq a \leq t$, $a \neq s$, *t* and $s \neq t$, then *s* is covered by *t*, and *t* covers *s*.

The results of the partial order ranking is visualised in a diagram which is constructed as follows:

1. each element is represented by a small circle
2. within each circle the element name, or the equivalence class, is given. Equivalent elements are different elements that have the same numerical values with respect to a given set of attributes. The equality according to a set of attributes defines an equivalence relation
3. if an order or cover relation exists then a line between the corresponding pairs of elements is drawn, the elements belonging to an order relation are "comparable"
4. if $s \leq t$ then s is drawn below t, therefore the diagram has orientation, consequently a sequence of lines can only be read in one direction either upwards or downwards
5. if $s \leq t$ and $t \leq z$ then $s \leq z$ according to the transitivity rule; however a line between s and z is not drawn because this connection can be deduced from the lines between s and t and t and z
6. if either $s \leq t$ or $t \leq s$ then s and t are not connected by a line; thus they are called "incomparable"
7. ‘incomparable’ elements are located at the same geometrical height and as high as possible in the diagram, resulting in a structure of levels. Elements belonging to a given level are incomparable’. Note, however, that a location of elements at different levels does not imply comparability.

In the Hasse diagram, the elements at the top of the diagram are called *maximals* and there are no elements above them; instead elements which have no elements below are called *minimals* and they do not cover any further element. If there is only one minimal element, then this is called the *least element* and if there is only one maximal element, it is called the *greatest element*. In the environmental field, where the Hasse technique was first applied, the criteria describe the elements in terms of environmental hazard. The main assumption is that the lower the numerical value the lower the hazard. If a high numerical value of an attribute corresponds to low hazard the attribute values must be multiplied by -1 to invert their order.

Therefore, by this convention, the maximal elements are the most hazardous, and are selected to form the set of priority elements. Elements that are not comparable with any other element are called *isolated elements*, and can be seen as maximals and minimals at once: according to the caution principle they are located at the top of diagram within those elements that require priority attention.

A *chain* is a set of comparable elements, therefore levels can be defined as the longest chain within the diagram. An *antichain* is a set of mutually incomparable elements. The height (longest chain) and width (longest antichain) of an order set are indicators of the relative number of comparable pairs of elements compared to the total number of pairs. An example is provided to understand the Hasse diagram interpretation. Let E be the set of 10 elements, and IB the information basis of four attributes describing the elements then the data matrix processed is the one of Table 2.1.

Element	r_1	r_2	r_3	r_4
a	15	4	6	8
b	12	22	57	31
c	3	5	6	8
d	44	33	54	33
e	22	38	66	35
f	11	2	69	27
g	6	29	44	28
h	14	31	32	22
i	13	18	20	21
m	18	19	23	28

Table 2.1 – Data matrix used for the construction of the Hasse diagram.

The corresponding Hasse matrix is shown in Table 2.2.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*m*
a	-	0	0	-1	-1	0	0	0	0	-1
b	0	-	1	0	-1	0	0	0	0	0
c	0	-1	-	-1	-1	0	-1	-1	-1	-1
d	1	0	1	-	0	0	1	1	1	1
e	1	1	1	0	-	0	1	1	1	1
f	0	0	0	0	0	-	0	0	0	0
g	0	0	1	-1	-1	0	-	0	0	0
h	0	0	1	-1	-1	0	0	-	1	0
i	0	0	1	-1	-1	0	0	-1	-	-1
m	1	0	1	-1	-1	0	0	0	1	-

Table 2.2 – Hasse matrix used for the construction of the Hasse diagram.

The corresponding Hasse diagram is shown in Figure 2.2.

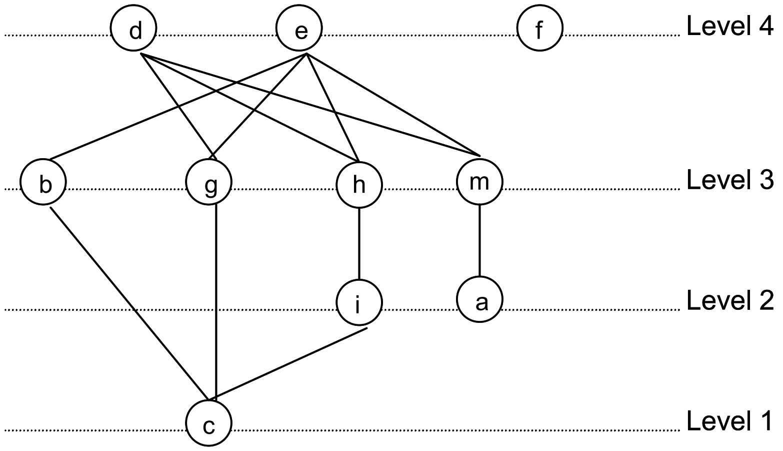


Figure 2.2. – Typical Hasse diagram.

In the Hasse diagram in Figure 2.2 there are no equivalence classes; the elements are arranged in four levels, elements *d* and *e* are maximals, and are not covered by any other element. The element *f* is an isolated element since it is not comparable with any of the other elements; *c* is a minimal and, especially, a least element. Several chains arise: for example $d \geq g \geq c$ and $e \geq h \geq i. \geq c.$ Obviously maximals are mutually incomparable. At level 3, *b* and *g* are not comparable. Incomparability is due to contradictory attributes: for each incomparable pair of elements there must be at least one pair of attributes of counteracting values. Such attributes are called antagonistic. The key diagram interpretation is provided by the meaning of chain and antichain. A chain indicates that the values of the attributes increase synchronously, whereas antichains correspond to diverse patterns. Thus if attributes describe the hazard caused by chemicals which are toxic to different species, then maximals are those elements of highest priority, the most toxic ones, whilst incomparability expresses a diverse pattern of toxicity e.g. toxicity to different species. In this case maximal elements are, in the same way, of priority attention, being toxic but in a different way.

In recent years the Hasse diagram technique (HDT) has been widely applied in several fields; a new, recently presented, application of HDT in chemistry is that of reaction diagrams of progressive substitution on a fixed molecular skeleton, forming Hasse diagrams for a partially ordered set of substituted structures [Klein and Bytautas, 2000]. An example is the case of the progressive chlorination of the hexagonal benzene skeleton illustrated in Figure 2.3. The carbon hexagon is shown and the Cl-substituted carbon vertices are shown as larger dots; an arrow points the way from structure A to structure B if B can be obtained from A by replacing a H-atom by a Cl-atom. Thus the arrow represents a single minimal step of chlorination; the ordering relation means that there is a number of Cl-atoms which can be added to structure A so as to obtain structure B.

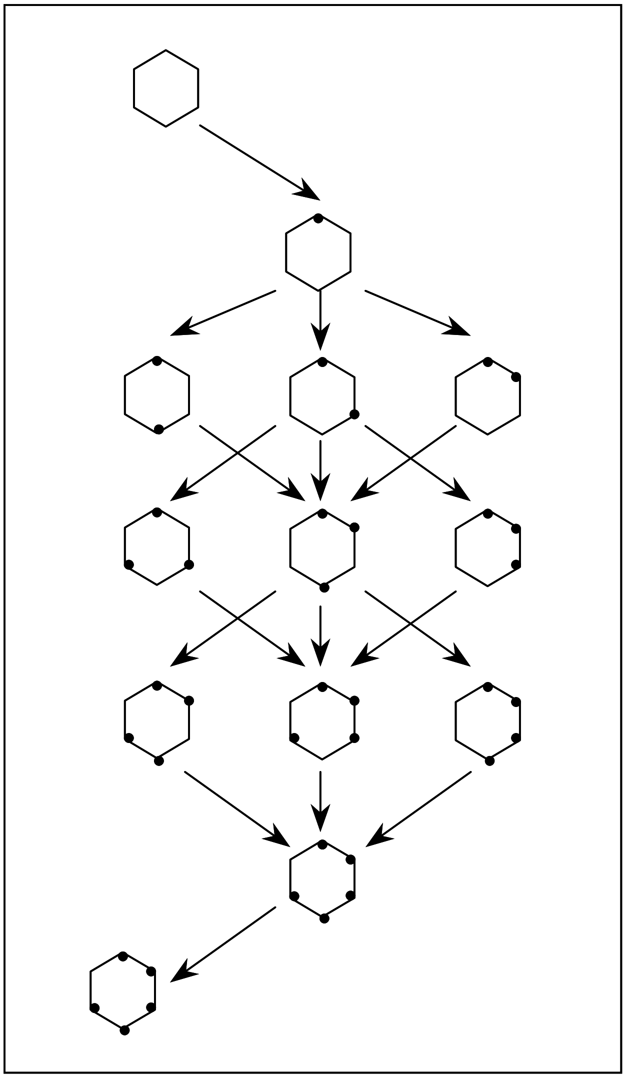


Figure 2.3 – Posetic reaction diagram for successive chlorination of the benzene skeleton; black dots identify Cl-substituted sites.

The general class of posetic reaction diagrams are those for which there is a progressive degree of reaction (substitution, addition, dissociation,…). There are many possible examples of progressive reaction graphs and many possible uses to be investigated.

In accordance with the literature [Brüggemann *et al*., 1993b] the Hasse diagram technique has some relevant advantages:

- evaluation can be represented as a graph
- the mathematics is very simple
- it can easily manage criteria of different scales (linguistic, ordinal and ratio-scaled criteria) since it does not perform any numerical aggregation of the criteria.

Nevertheless there are some severe drawbacks:

- It is strictly dependent on the clarity of the graphical diagram: diagrams that are too complex or too poorly structured, with more isolated elements than comparable elements because of conflict, are of little use.
- if there are too many contradictions criterion reduction must be performed by preliminary multivariate statistic techniques, like Principal Component Analysis (PCA) Multidimensional Scaling.
- if many elements are to be evaluated, preliminary multivariate statistic techniques, like Cluster analysis, are needed to get a readable diagram
- the generality principle is very restrictive and requires appropriate data handling. In fact it must be ensured that any two elements ordered by ">" can be considered as physically and numerically significantly different, i.e. they should have numerically significant data differences. Differences within statistical noise, numerical uncertainty and experimental error are considered physically meaningless, but the Hasse diagram technique considers such elements as different.

Data, and especially environmental data, are often associated with a significant degree of uncertainty inherent in ranking analysis. The comparison of two elements (comparable/incomparable) and thus of the ranking can obviously be affected by this uncertainty. There are two main sources of ranking uncertainty: the relationship assumed between the attributes and the phenomenon described by the ranking, and the input uncertainty. The first type of uncertainty can be minimized by increasing the number of attributes so that a large number of different aspects are taken into account; nevertheless the greater the number of attributes, the higher the probability that contradictions will occur in ranking the attributes (incomparabilities), and thus the greater the uncertainty in ranking the elements.

Uncertainty from input is the uncertainty induced from variability in the input parameters, which may be due to true variability or to errors in the procedure used to determine the values.

2.2 Pre-processing tools

Several studies on the Hasse diagram technique have highlighted that often it cannot be used alone. Hasse limits can be solved by combining the technique with statistical techniques like Clustering, Principal Component analysis or Multidimensional Scaling. Thus a complete evaluation with HDT requires a pre-processing phase to establish an adequate data matrix, and a post-processing phase to correctly extract information and decisions. Obviously both pre-processing and post-processing may significantly influence the results provided by HDT.

2.2.1 Cluster analysis

As pointed out above the Hasse diagram technique, which shows its results graphically, is not easy to read if there are many objects to be evaluated, and in such cases object reduction is required. Thus, for HDT, Cluster Analysis is a very important pre-processing tool. A partition of the object is first performed according to a given clustering method, then representatives of each cluster are defined and used to construct the order diagram. In this way the clusters are considered equivalence classes, the cluster elements being similar enough to be assumed equivalent. The main advantage is that the elements are easily replaced by a reduced number of pseudo elements, cluster centres or centrotypes. However difficulties may arise. Clustering methods are based on the concept of ‘element similarity’ and calculate element distances. Instead HDT does not combine numerically different attributes, this is done by defining distance expressions. Moreover the clustering procedure may change ordinal relations and generate new ones, thus the results obtained will depend on the clustering performed. As is well known, clustering is affected by a certain arbitrariness related to the chosen clustering method, the initial element partition, the metrics used to measure element distances, and the selection of cluster numbers. As the clustering is not aimed at assigning each element to a belonging class (previously unknown), but at reducing data, the main interest is to consider a big enough number of clusters to sample the data space as deeply and exhaustively as possible.

The clustering methods most frequently combined with HDT are single and complete linkage clustering. These are based on a sequential, agglomerative and hierarchical algorithm that starts from a distance or a similarity matrix, and proceeds in two steps: The matrix is rewritten so as to decrease the similarities or increase distances, identifying the two most similar elements followed by the second most similar pair. The clusters are then made hierarchically, starting with the two most similar elements that are combined to form a new group, and aggregating the new groups one to the other. In single linkage, an element is assigned to a cluster if it has similarity equal to the considered partition level of at least one element belonging to that cluster. In complete linkage the element must display the similarity level of all the elements already assigned to that cluster. The result of a clustering procedure is usually represented by dendrograms which clearly show the clusters generated at each partition level. The ordinate is graduated in similarities or distances, while the abscissa encodes the information on the elements or their identification codes.

Example of cluster analysis combined with Hasse diagram technique.

A partial order ranking by Hasse diagram technique was performed on the toxicity data of 83 chemicals tested experimentally for their toxicity at 01, 10, 20, 50,80 and 90 concentrations on *Scenedesmus vacuolatus* by the BEAM EU project. Table 2.3 collects the data.

SUBSTANCE	*ID*	*EC01*	*EC10*	*EC20*	*EC50*	*EC80*	*EC90*
2,4,5-Trichlorophenol	1	0.25	0.09	0.03	-0.05	-0.10	-0.13
2,4-D	2	-1.22	-2.03	-2.29	-2.67	-2.96	-3.09
4-Nitrophenol	3	-0.66	-1.14	-1.30	-1.53	-1.71	-1.78
Aldicarb	4	-0.70	-1.92	-2.31	-2.90	-3.34	-3.52
Ametryn	5	3.01	2.34	2.13	1.81	1.57	1.47
Anilazine	6	0.14	-0.34	-0.50	-0.73	-0.90	-0.98
Atraton	7	1.83	1.06	0.81	0.44	0.17	0.05
Atrazine	8	2.07	1.36	1.13	0.79	0.53	0.42
Aziprotryne	9	0.02	-0.74	-0.98	-1.34	-1.62	-1.73
Barban	10	0.18	-0.19	-0.31	-0.49	-0.63	-0.69
Biphenyl	11	0.47	0.11	0.00	-0.17	-0.30	-0.35
Bitertanol	12	0.85	0.16	-0.06	-0.39	-0.64	-0.74
Bromacil	13	2.60	1.78	1.51	1.12	0.82	0.70
Buturon	14	1.88	0.90	0.58	0.11	-0.24	-0.39
Butylate	15	-1.91	-2.50	-2.69	-2.98	-3.19	-3.28
Carbetamide	16	-1.16	-1.94	-2.19	-2.57	-2.85	-2.96
Chlorbromuron	17	2.80	1.94	1.67	1.25	0.94	0.81
Chlorbufam	18	0.18	-0.19	-0.31	-0.49	-0.62	-0.68
Chloridazon	19	0.80	-0.04	-0.32	-0.72	-1.03	-1.16
Chlorotoluron	20	2.52	1.58	1.27	0.82	0.48	0.33
Chloroxuron	21	2.66	1.99	1.77	1.45	1.21	1.10
Chlorpropham	22	-0.61	-0.75	-0.79	-0.86	-0.91	-0.93
Cyanazine	23	1.85	1.35	1.19	0.95	0.77	0.69
Cycluron	24	1.54	0.61	0.32	-0.13	-0.46	-0.60
Cyproconazole	25	1.73	1.14	0.95	0.66	0.45	0.36
Cyromazine	26	-1.34	-2.17	-2.44	-2.84	-3.14	-3.26
Desmetryn	27	2.49	1.91	1.72	1.43	1.22	1.13
Diclobutrazol	28	1.19	0.36	0.10	-0.30	-0.59	-0.72
Difenoconazole	29	1.31	0.77	0.60	0.33	0.14	0.06
Difenoxuron	30	2.13	1.42	1.19	0.85	0.59	0.49
Dimefuron	31	2.11	1.32	1.06	0.68	0.39	0.27
Dimethametryn	32	2.18	1.69	1.53	1.29	1.11	1.03
Dipropetryn	33	2.11	1.55	1.38	1.11	0.91	0.82
Diuron	34	3.07	2.22	1.95	1.54	1.24	1.11
Fenbuconazole	35	1.30	0.70	0.51	0.21	0.00	-0.09
Fenuron	36	1.30	0.22	-0.13	-0.65	-1.04	-1.21
Fluometuron	37	1.99	0.78	0.39	-0.19	-0.62	-0.81
Fluoranthene	38	1.56	1.25	1.15	1.00	0.89	0.85
Flusilazole	39	1.31	0.70	0.51	0.21	-0.01	-0.10
Flutriafol	40	0.09	-0.39	-0.54	-0.78	-0.95	-1.02
Hexaconazole	41	1.42	0.93	0.77	0.54	0.36	0.28
Hexazinone	42	1.60	1.22	1.10	0.91	0.78	0.72

Irgarol 1051	43	2.77	2.15	1.95	1.65	1.43	1.33
Isoproturon	44	2.29	1.37	1.08	0.64	0.31	0.17
Karbutylate	45	1.51	0.93	0.74	0.46	0.25	0.16
Lenacil	46	1.85	1.35	1.20	0.96	0.78	0.71
Lindane	47	0.13	-0.32	-0.46	-0.68	-0.84	-0.91
Linuron	48	3.15	1.99	1.62	1.06	0.64	0.46
Metamitron	49	1.65	0.51	0.15	-0.40	-0.81	-0.98
Methabenzthiazuron	50	1.88	1.16	0.93	0.58	0.32	0.21
Methoprotryne	51	1.97	1.53	1.39	1.18	1.02	0.95
Methoxyphenone	52	2.24	1.83	1.69	1.49	1.34	1.28
Metobromuron	53	1.53	0.67	0.39	-0.02	-0.33	-0.46
Metoxuron	54	2.32	1.21	0.85	0.32	-0.08	-0.25
Metribuzin	55	2.57	1.81	1.56	1.19	0.92	0.80
Monolinuron	56	2.06	0.92	0.56	0.01	-0.40	-0.57
Monuron	57	2.57	1.37	0.98	0.40	-0.03	-0.21
Myclobutanil	58	0.55	-0.04	-0.22	-0.50	-0.71	-0.80
Naphthalene	59	-0.71	-1.13	-1.26	-1.47	-1.62	-1.68
Neburon	60	3.41	2.32	1.97	1.45	1.06	0.89
Paclobutrazol	61	-0.50	-0.76	-0.85	-0.97	-1.07	-1.11
Paraquat	62	1.14	0.54	0.35	0.06	-0.15	-0.24
Parathion	63	-0.56	-0.95	-1.07	-1.26	-1.40	-1.46
Penconazole	64	0.71	0.27	0.13	-0.08	-0.24	-0.31
Phoxim	65	2.50	1.20	0.78	0.15	-0.31	-0.51
Prochloraz	66	1.99	1.54	1.40	1.18	1.02	0.95
Prometon	67	1.66	0.88	0.64	0.26	-0.02	-0.13
Prometryn	68	2.35	1.78	1.59	1.31	1.11	1.02
Propazine	69	1.92	1.15	0.91	0.54	0.26	0.14
Propiconazole	70	0.65	0.25	0.12	-0.08	-0.22	-0.28
Sebuthylazine	71	2.06	1.41	1.19	0.88	0.64	0.54
Secbumeton	72	2.38	1.50	1.22	0.80	0.49	0.35
Simazine	73	2.49	1.41	1.07	0.55	0.16	0.00
Simetryn	74	2.72	1.94	1.69	1.32	1.04	0.92
Tebuconazole	75	0.92	0.34	0.16	-0.12	-0.33	-0.42
Tebuthiuron	76	1.98	1.11	0.83	0.42	0.10	-0.03
Terbacil	77	2.33	1.55	1.30	0.92	0.64	0.52
Terbumeton	78	2.38	1.61	1.36	0.99	0.71	0.59
Terbuthylazine	79	2.24	1.64	1.45	1.16	0.94	0.85
Terbutryn	80	2.33	1.87	1.72	1.50	1.33	1.26
Tetraconazole	81	0.63	0.11	-0.05	-0.30	-0.49	-0.57
Triadimefon	82	0.51	-0.01	-0.18	-0.43	-0.62	-0.70
Triadimenol	83	0.37	-0.13	-0.29	-0.53	-0.71	-0.78

Table 2.3 – Toxicity (Log1/EC) data of 83 chemicals.

The chemicals are currently commonly used: antifouling agents, antioxidants, bactericides, chemotherapeutics, disinfectants, fungicides, herbicides, insecticides, tools in physiological research and industrial chemicals. Since the obtained Hasse diagram was quite complex (26 levels, 2788 comparabilities, 1230 incomparabilities) and thus difficult to interpret, a cluster analysis by Single Linkage algorithm and Euclidean distance, was performed to define a homogeneous subset of chemicals. Fifteen clusters were found with a similarity level of 90%. Figure 2.4 shows the Hasse diagram obtained on the 83 chemicals, the chemicals (circles) being coloured according to the cluster they belong to.

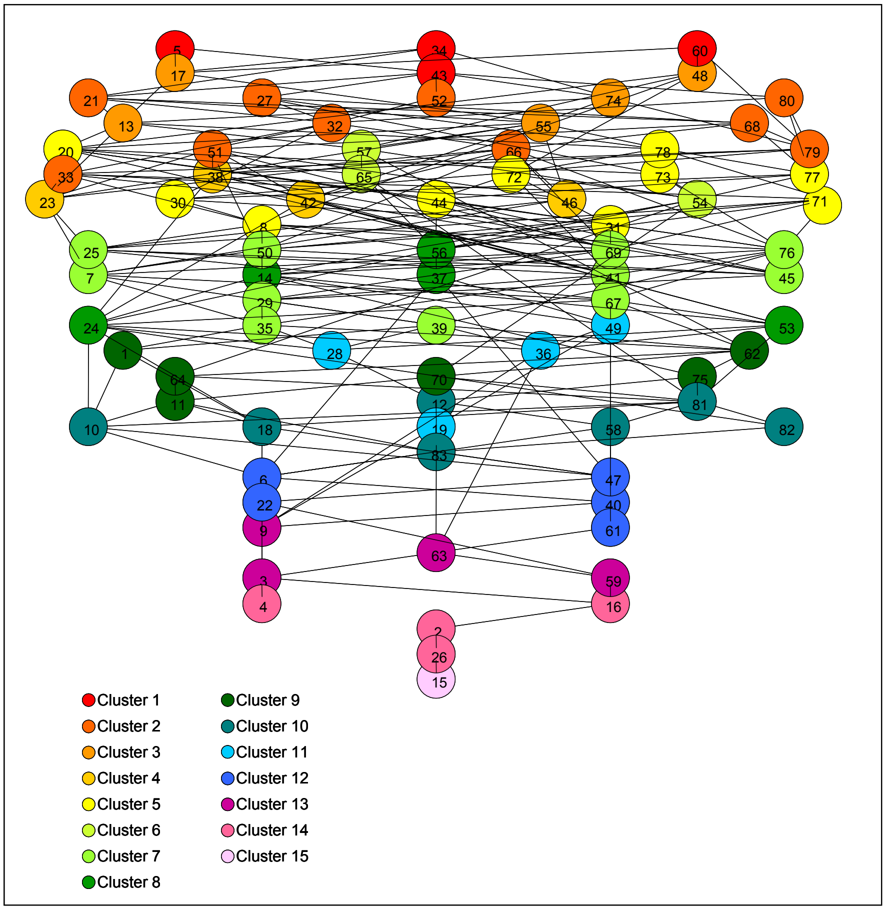
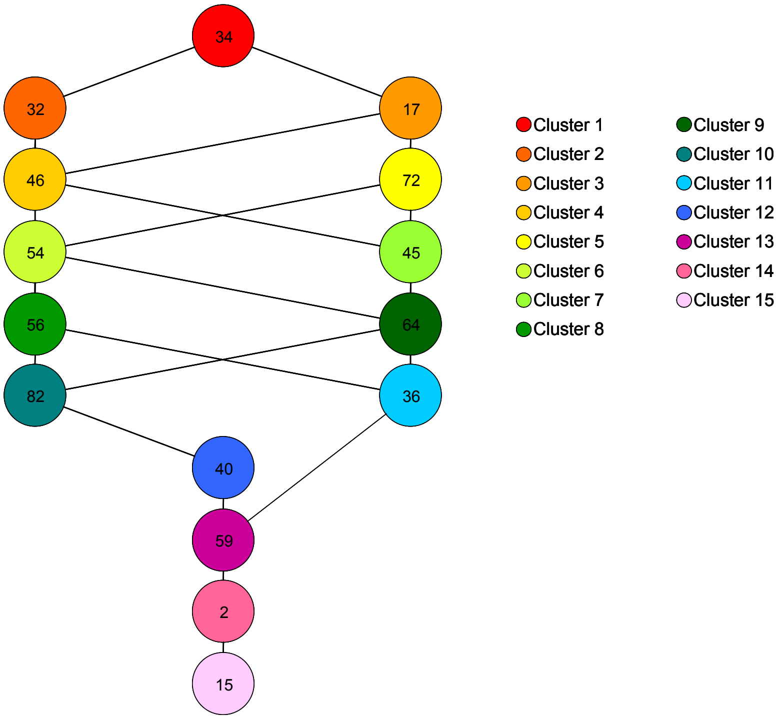


Figure 2.4 –Hasse diagram on toxicity data of 83 chemicals.

It can be highlighted that Cluster analysis and the Hasse diagram technique provide comparable information on the overall toxicity of the studied chemicals: there is quite good agreement between clusters and levels. Cluster 1 is composed of highly toxic chemicals, located at the highest level in the Hasse diagram, while clusters 14 and 15 collect chemicals of low toxicity. Also the centrotypes of each cluster were ranked in a Hasse diagram (Figure 2.5).

The obtained diagram is much clearer than the previous one. Cluster elements being characterised by a similarity level of 90% are considered similar enough to be assumed equivalent. The main order relations appear to be preserved by the clustering. Clusters 2 and 3 both consist of highly toxic chemicals showing a different kind of toxicity: chemicals of cluster 2 exhibit lower toxicity than those of cluster 3 at low concentrations (EC01, EC10, EC20), but higher toxicity at high concentrations (EC50, EC80, EC90). This behaviour results in incomparability between Clusters 1 and 2. The same occurs for Clusters 4 and 5. Thus, both methods are valuable tools to explore element relations, and their combined use can help to better understand and read the complex diagram. In any case the obtained results depended strongly on the clustering performed.



Cluster 1: { 5, 34, 43, 60 }

Cluster 2: { 21, 27, 32, 33, 51, 52, 66, 68, 79, 80 }

Cluster 3: { 13, 17, 48, 55, 74 }

Cluster 4: { 23, 38, 42, 46 }

Cluster 5: { 8, 20, 30, 31, 44, 71, 72, 73, 77, 78 }

Cluster 6: { 54, 57, 65 }

Cluster 7: { 7, 25, 29, 35, 39, 41, 45, 50, 67, 69, 76 }

Cluster 8: { 14, 24, 37, 53, 56 }

Cluster 9: { 1, 11, 62, 64, 70, 75 }

Cluster 10: { 10, 12, 18, 58, 81, 82, 83 }

Cluster 11: { 19, 28, 36, 49 }

Cluster 12: { 6, 22, 40, 47, 61 }

Cluster 13: { 3, 9, 59, 63 }

Cluster 14: { 2, 4, 16, 26 }

Cluster 15: { 15 }

Figure 2.5 –Hasse diagram on toxicity data of 83 chemicals.

2.2.2 Principal Component Analysis (PCA) and Nonmetric Multidimensional scaling

Principal Component Analysis is one of the best known procedures in multivariate statistics. Proposed by Karl Pearson in 1901 and developed by Harold Hotelling in 1933 it has been mainly used for data exploration. It allows the examination of the correlation pattern among variables and an evaluation of their relevance, the visualization of the elements by analyzing their inter-co-relationships (outliers, clusters), the synthesis of data description discarding noise, the reduction of data dimensionality by discarding unnecessary variables, and the finding of principal properties in multivariate systems.

From a mathematical point of view the aim of principal component analysis is to transform p-correlated variables into a set of orthogonal variables which reproduce the original variance/covariance structure. This means rotating a *p-th* dimensional space to achieve independence between variables. The new variables, called principal components, are linear combinations of the original variables along the direction of maximum variance in the multivariate space, and each linear combination explains a part of the total variance of the data. Being orthogonal the information contained in each PC is unique. A maximum of *p* principal axes can be derived from the original data containing *p* variables. The new variables are defined by calculating eigenvalues and eigenvectors of the correlation matrix **C** (or the covariance matrix **S**) obtained from the data matrix **X**. The principal components of a dispersion matrix **C** are found according to the following expression:

$$\mathrm{diag}(\mathbf{C}) = \mathrm{diag}\left[\frac{\mathbf{X}_C^T \cdot \mathbf{X}_C}{N-1}\right]$$

where $\mathbf{X}_C$ is the centred data matrix.

Because of their properties, principal components can often be used to summarize, in a few dimensions, most of the variability of a dispersion matrix of a large number of variables, providing a measure of the amount of variance explained by a few independent principal axes. The first two principal components define a plane, which represents the largest amount of variance. The elements are projected in this plane in such a way as to preserve, as much as possible, the relative Euclidean distances they have in the multidimensional space of the original variables.

Principal component analysis is a reduced space ordination method which starts from a scaling of the elements in full-dimensional space, representing them in a few dimensions while preserving the distance relationships among the elements. However sometimes the aim is not the exact preservation of the element distances, but their representation in a small and specified number of dimensions plotting dissimilar elements far apart in the ordination space and similar elements close to one another. For these purposes the Nonmetric Multidimensional Scaling (NMDS) method, aimed at preserving ordering relationships among elements, can be suitable. It is not limited to Euclidean distance matrices, and contrary to PCA, which is an eigenvector method, NMDS does not maximize the variability associated with individual axes; NMDS axes are arbitrary. Starting from a distance matrix, the number *m* of dimensions has to be chosen *a priori*: the output provides the coordinates of the *N* elements on the *m* axes. An iterative process is commonly used: starting from an initial configuration of the elements in *m* dimensions, the adjustment process goes on until it converges on a solution. The space of solutions can contain several local minima besides the overall minimum and strongly depends on the initial element configuration. Several solutions can be used: most objective functions are based on the sum of the squared differences between the fitted distances d_{hi} and the corresponding values forecast by the process $\hat{d}_{hi}$. Several variants have been proposed. The one commonly used in NMDS programs is the one called Stress, and defined as follows:

$$Stress = \sqrt{\frac{\sum_{hi}(d_{hi} - \hat{d}_{hi})^2}{\sum_{hi} d_{hi}^2}}$$

Another expression frequently used is defined as:

$$Stress\ (formula\ 2) = \sqrt{\frac{\sum_{hi}(d_{hi} - \hat{d}_{hi})^2}{\sum_{hi}(d_{hi} - \bar{d}_{hi})}}$$

The denominators in the above expressions are scaling terms that make the objective functions dimensionless and produce Stress values between 0 and 1. All the objective functions are measures of how far the reduced space configuration is from the original one.

When a preprocessing tool, like cluster analysis, principal component analysis or multidimensional scaling is used to reduce the number of elements or attributes to be used in the ranking analysis, metric information, which is not required by the Hasse diagram technique as it extracts only ordinal information, becomes mandatory.

An ordinal scale possesses no natural origin, and distances between points of scale are undefined. It simply preserves the ranks of the elements. Ordinal scale tends to be discrete rather than continuous. Since an ordinal scale does not have the property of distance among its values, but only reflects monotonically increasing and decreasing sequences of magnitudes, it is referred to as “nonmetric”. Metric scaling is usually seen as a “stronger” property than ordinal scaling, and this means that element ranking based on a set of attributes is seen as “basic information” which is supplemented with metric information. As a consequence, if a metric preprocessing tool is used, its effect on the data has to be analyzed, i.e. there must be the clarification of how preprocessing influences ordinal information. Thus information preservation is the minimal requirement that arises whenever several data analysis methods are combined.

Ordinal data can characterize three main different situations: first, multidimensional continuity is not observed, and there are integer ranks from which underlying continuity has to be estimated. Second, continuity can exist but cannot be used because of excessive measurement error or nonlinearity of unknown form. In this case the original variables can be replaced by their rank orders to restore linearity and delete much of the error: as a consequence of this action metric information is lost from the sampled values. Third, continuity may not exist at all: in this case the data are purely qualitative or nonmetric.

2.2.3 Principal Component Analysis on intrinsically continuous ordinal variables

A relatively straightforward approach to principal component analysis of ordinal data is the assumption of underlying continuity for the samples, not observable directly but consequently approximated by an ordinal scale. As the ordinal scales are invariant to continuous monotonic transformations, positive integers are used. A typical situation is the one in which *N* judges are asked to express a preference (agreements) on the scale 1,2,..,*k*, concerning a set of *p* products. In such a case the aim of the factor analysis is to provide an interval scale estimate of the multidimensional continua which generated the observed rankings. Another application of principal component analysis on ordinal variables is that of a measuring or scaling device of concepts which are intrinsically multidimensional. In this case the observed variables are chosen to reflect the underlying multidimensional scale. If the variables are correlated, a reduced number of dimensions can be selected to develop a scale to estimate the relative position of each sample on the scale.

2.2.4 Principal Component Analysis on ranked values obtained from a continuous scale

When data observed on a continuous scale are characterized by large measurement errors or unknown forms of nonlinearity among the variables, “quantitative” information cannot be used; in these cases the original variables can be replaced by their rank orders. This action allows a significant reduction in measurement error and introduces linear relations between the variable ranks, even if the original variables are nonlinear [Basilevsky, 1994]. Rank-order transformation can obviously result in a loss of information if applied to errorless or linear data. A high number of ties occur when only a subset of order statistics is used, for example, deciles or quartiles. The decision to use a reduced number of order statistics is related to the aim of exploring the “main features” of multivariate data, or to perform a preliminary analysis before a more complete one. Obviously the results of the analysis depend on the chosen ranking scale, however replacing the original data by quartiles or deciles may reveal qualitative features which could otherwise be submerged by quantitative information. A comparison of principal component analyses performed on original data, ranked transformed data and on quartiles were conducted by Goldstein [Goldstein,1982] on data for social and disease variables for 21 wards of Hull, England. For each ward the following quantitative information was collected: crowding, number of toilets, number of cars, unskilled, jaundice, measles, scabies. The original values, the ranked transformed values and the quartile and binary values are shown in Tables 2.3, 2.4, 2.5 and 2.6, respectively.

Ward	*Crowding*	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Quantitative data: counts							
1	28	222	627	86	139	96	20
2	53	258	584	137	479	165	31
3	31	39	553	64	88	65	22
4	87	389	759	171	589	196	84
5	29	46	506	76	198	150	86
6	96	385	812	205	400	233	123
7	46	241	560	83	80	104	30
8	83	629	783	255	286	87	18
9	112	24	729	255	108	87	26
10	113	5	699	175	389	79	29
11	65	61	591	124	252	113	45
12	99	1	644	167	128	62	19
13	79	276	699	247	263	156	40
14	88	466	836	283	469	130	53
15	60	443	703	156	339	243	65
16	25	186	511	70	189	103	28
17	89	54	678	147	198	166	80
18	94	749	822	237	401	181	94
19	62	133	549	116	317	119	32
20	78	25	612	177	201	104	42
21	97	36	673	154	419	92	29

Table 2.3 – Original quantitative data for 21 Wards of Hull, England 1968-1973.

Ward	*Crowding*	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Ranked data							
1	2	12	9	5	5	7	3
2	6	14	6	8	20	16	10
3	4	6	4	1	2	2	4
4	13	17	17	13	21	19	18
5	3	7	1	3	7	14	19
6	17	16	19	16	16	20	21
7	5	13	5	4	1	9	9
8	12	20	18	20	12	5	1
9	20	3	16	19	3	4	5
10	21	2	13	14	15	3	7
11	9	9	7	7	10	11	14
12	19	1	10	12	4	1	2
13	11	15	14	18	11	15	12
14	14	19	21	21	19	13	15
15	7	18	15	11	14	21	16
16	1	11	2	2	6	8	6
17	15	8	12	9	8	17	17
18	16	21	20	17	17	18	20
19	8	10	3	6	13	12	11
20	10	4	8	15	9	10	13
21	18	5	11	10	18	6	8

Table 2.4 – Ranked data for 21 Wards of Hull, England 1968-1973.

Ward	*Crowding*	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Quartile data							
1	1	3	2	1	1	2	1
2	2	3	2	2	4	4	2
3	1	2	1	1	1	1	1
4	3	4	4	3	4	4	4
5	1	2	1	1	2	3	4
6	4	4	4	4	4	4	4
7	1	3	1	1	1	2	2
8	3	4	4	4	3	1	1
9	4	1	4	4	1	1	1
10	4	1	3	3	3	1	2
11	2	2	2	2	2	3	3

12	4	1	2	3	1	1	1
13	3	3	3	4	3	3	3
14	3	4	4	4	4	3	3
15	2	4	3	3	3	4	4
16	1	3	1	1	2	2	2
17	3	2	2	2	2	4	4
18	4	4	4	4	4	4	4
19	2	2	2	2	3	3	3
20	2	1	3	3	2	2	3
21	4	1	2	2	4	2	2

Table 2.5 – Quartile data for 21 Wards of Hull, England 1968-1973.

Ward	*Crowding*	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Binary data							
1	1	2	1	1	1	1	1
2	1	2	1	1	2	2	1
3	1	1	1	1	1	1	1
4	2	2	2	2	2	2	2
5	1	1	1	1	1	2	2
6	2	2	2	2	2	2	2
7	1	2	1	1	1	1	1
8	2	2	2	2	2	1	1
9	2	1	2	2	1	1	1
10	2	1	2	2	2	1	1
11	1	1	1	1	1	2	2
12	2	1	1	2	1	1	1
13	2	2	2	2	2	2	2
14	2	2	2	2	2	2	2
15	1	2	2	2	2	2	2
16	1	2	1	1	1	1	1
17	2	1	2	1	1	2	2
18	2	2	2	2	2	2	2
19	1	1	1	1	2	2	2
20	1	1	1	2	1	1	2
21	2	1	2	1	2	1	1

Table 2.6 – Binary data for 21 Wards of Hull, England 1968-1973.

The aim of the analysis was to explore rank-transformation capability to reduce and delete measurement error from the data.

Correlation analysis was performed on the original data, the ranked transformed data and the quartile data. The multivariate correlation in the three sets of data was estimated by the *K* correlation index [Todeschini *et al.*, 1998] which measures the total quantity of correlation contained in the data, from the eigenvalue distribution obtained from the eigenvalue decomposition of the corresponding correlation matrix.

The *K* index is defined as the following:

$$K = \frac{\sum_{j=1}^{p} \left| \frac{\lambda_j}{\sum \lambda_j} - \frac{1}{p} \right|}{\frac{2 \cdot (p-1)}{p}} \qquad 0 \leq K \leq 1$$

λ_j being the set of *p* eigenvalues obtained by PCA applied to the correlation matrix of a data set. The denominator corresponds to the maximum value reached by the numerator and is used to scale the values between 0 and 1. The *K* correlation index, being a redundancy index, takes a value of 1 when all the variables are correlated and 0 when they are uncorrelated.

The multivariate correlation index *K* calculated on the original, rank-transformed and quartile data, takes values 0.57, 0.58, 0.59 and 0.47, respectively. Moreover the pair correlations for the four datasets were analyzed and the corresponding matrices are reported in Tables 2.7, 2.8. 2.9 and 2.10, respectively. Both the correlation matrices and the rotated loadings (Tables 2.11, 2.12, 2.13 and 2.14), reveal close similarity in the three analyses, indicating that the relationship among the variables is approximately linear.

	Crowding	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Crowding	1.000						
N. Toilet	0.084	1.000					
N. Car	0.733	0.641	1.000				
Unskilled	0.779	0.480	0.869	1.000			
Jaundice	0.380	0.487	0.534	0.403	1.000		
Measles	0.055	0.522	0.375	0.184	0.542	1.000	
Scabies	0.181	0.378	0.409	0.180	0.425	0.823	1.000

Table 2.7 – Correlation matrix of social and disease variables in the Wards.

	Crowding	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Crowding	1.000						
N. Toilet	-0.200	1.000					
N. Car	0.675	0.495	1.000				
Unskilled	0.717	0.305	0.879	1.000			
Jaundice	0.330	0.494	0.512	0.444	1.000		
Measles	-0.103	0.645	0.296	0.139	0.552	1.000	
Scabies	0.070	0.396	0.262	0.142	0.487	0.879	1.000

Table 2.8 – Correlation matrix of the ranks of social and disease variables in the Wards.

	Crowding	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Crowding	1.000						
N. Toilet	-0.105	1.000					
N. Car	0.694	0.388	1.000				
Unskilled	0.779	0.263	0.924	1.000			
Jaundice	0.447	0.484	0.541	0.484	1.000		
Measles	0.005	0.558	0.197	0.116	0.595	1.000	
Scabies	0.079	0.374	0.235	0.189	0.521	0.853	1.000

Table 2.9 – Correlation matrix of the quartile ranks of social and disease variables in the Wards.

	Crowding	*N. Toilet*	*N. Car*	*Unskilled*	*Jaundice*	*Measles*	*Scabies*
Crowding	1.000						
N. Toilet	0.045	1.000					
N. Car	0.809	0.236	1.000				
Unskilled	0.618	0.236	0.618	1.000			
Jaundice	0.427	0.427	0.618	0.427	1.000		
Measles	0.045	0.236	0.236	0.045	0.427	1.000	
Scabies	0.045	0.045	0.236	0.236	0.236	0.809	1.000

Table 2.10 – Correlation matrix of the binary data of social and disease variables in the Wards.

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*	*Factor 6*	*Factor 7*
Crowding	0.958	0.047	-0.116	0.171	0.031	0.179	0.069
N. Toilet	0.162	0.238	0.936	0.196	-0.031	-0.030	-0.015
N. Car	0.793	0.230	0.444	0.195	0.011	0.006	-0.289
Unskilled	0.887	0.029	0.325	0.112	-0.046	-0.302	0.017
Jaundice	0.251	0.268	0.210	0.906	-0.024	-0.006	-0.015
Measles	0.017	0.850	0.268	0.270	-0.362	-0.032	0.004
Scabies	0.127	0.961	0.116	0.127	0.169	0.017	-0.032
Expl. Variance	2.439	1.831	1.322	1.028	0.165	0.125	0.090

Table 2.11 – Varimax-rotated loadings of the original social and disease variables in the Wards of Hull.

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*	*Factor 6*	*Factor 7*
Crowding	-0.008	0.856	-0.348	0.175	-0.341	-0.017	0.012
N. Toilet	0.288	0.120	0.928	0.198	0.036	0.010	0.007
N. Car	0.139	0.887	0.340	0.160	-0.016	0.006	0.229
Unskilled	0.033	0.944	0.158	0.150	0.177	0.003	-0.167
Jaundice	0.315	0.288	0.209	0.880	-0.014	0.007	0.005
Measles	0.870	0.012	0.373	0.222	0.036	0.232	0.006
Scabies	0.970	0.081	0.083	0.161	-0.019	-0.137	0.008
Expl. Variance	1.900	2.514	1.313	0.967	0.151	0.073	0.081

Table 2.12 – Varimax-rotated loadings of the ranks of social and disease variables in the Wards of Hull.

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*	*Factor 6*	*Factor 7*
Crowding	0.772	0.001	-0.226	0.232	-0.547	0.008	-0.006
N. Toilet	0.132	0.254	0.942	0.166	0.061	0.015	0.006
N. Car	0.934	0.096	0.222	0.172	0.027	0.005	0.199
Unskilled	0.969	0.062	0.108	0.130	-0.044	-0.014	-0.160
Jaundice	0.336	0.362	0.227	0.835	-0.082	0.016	0.004
Measles	0.001	0.868	0.313	0.242	-0.030	0.298	0.005
Scabies	0.103	0.965	0.099	0.143	0.019	-0.168	0.002
Expl. Variance	2.548	1.893	1.158	0.904	0.314	0.118	0.065

Table 2.13 – Varimax-rotated loadings of the quartile ranks of social and disease variables in the Wards of Hull.

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*	*Factor 6*	*Factor 7*
Crowding	0.732	0.506	-0.280	-0.127	-0.239	0.219	0.096
N. Toilet	0.419	-0.114	0.845	0.211	-0.228	0.007	0.042
N. Car	0.875	0.285	-0.117	-0.156	-0.175	-0.280	-0.082
Unskilled	0.734	0.337	-0.102	0.523	0.234	0.056	-0.083
Jaundice	0.787	-0.051	0.301	-0.362	0.390	0.013	0.062
Measles	0.516	-0.809	-0.072	-0.129	-0.095	0.153	-0.160
Scabies	0.485	-0.744	-0.353	0.223	-0.013	-0.101	0.160
Expl. Variance	3.135	1.674	1.037	0.555	0.356	0.163	0.079

Table 2.14 – Varimax-rotated loadings of the binary data of social and disease variables in the Wards of Hull.

In the four principal components analyses performed, the first component mainly encodes information related to crowding, number of cars and the unskilled. The second component encodes information related to measles and scabies while the third and fourth components are mainly related to the number of toilets and jaundice respectively. The analyses are very similar to each other, revealing that replacing continuous data by ordered statistics of ranks or quartiles results in little loss of information.

Moreover partial order rankings by the Hasse diagram technique were performed using the original data, ranked data, quartile and binary data and the correspondent principal components, with the purpose of finding how the Hasse diagram changes, becoming simpler after a preprocessing analysis, and to compare the obtained results. The Hasse diagram developed on the seven original variables is shown in Figure 2.6: it is arranged in three levels, with three isolated elements, seven maximals and eight minimals and is characterized by 46 comparable pairs of elements counted in only one direction and 328 contradictions. The Hasse diagram developed on the four principal components calculated on the original data is shown in Figure 2.7: it is again arranged in three levels, with four isolated elements, five maximals and ten minimals and is characterized by 27 comparable pairs of elements and 366 contradictions.

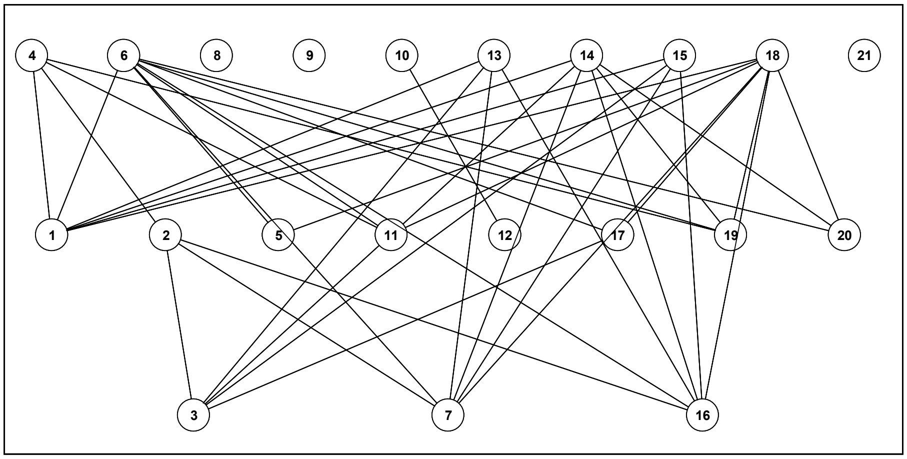


Figure 2.6 – Hasse diagram on the original quantitative data of social and disease variables in the Wards of Hull.

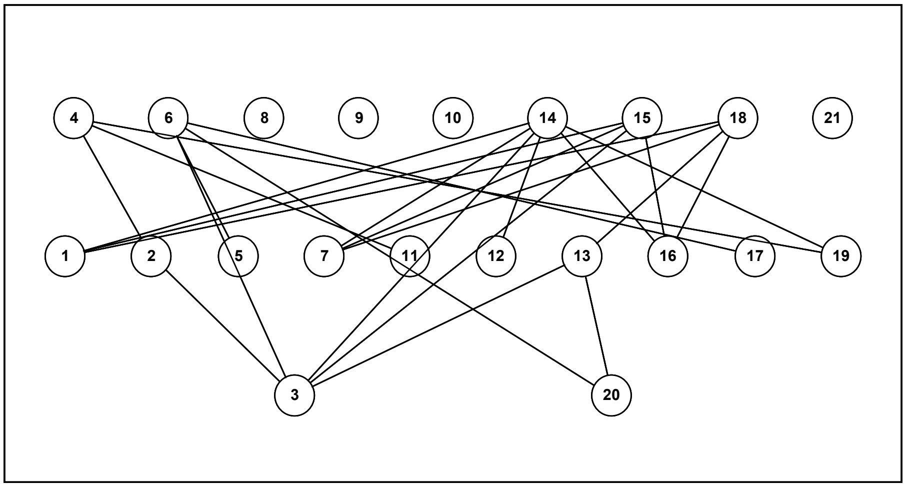


Figure 2.7 – Hasse diagram on the principal components calculated on quantitative data of social and disease variables in the Wards of Hull.

It can be observed that the number of comparable pairs of elements decreases whereas the number of contradictions increases, revealing that the principal components calculated on the original quantitative data are not able to simplify the original diagram and reduce the incomparabilities. Two main differences between the two diagrams can be highlighted: the element 13 is underestimated by the principal components with respect to the original data, being in the latter a maximal element, while the element 16 is overestimated by the components, being a minimal element in the diagram performed on original data.

The Hasse diagram developed on the seven rank transformed variables is exactly the same as the one obtained on the original data.

The Hasse diagram developed on three principal components calculated on ranked transformed data is shown in Figure 2.8: it is a four level diagram with four maximals, one isolated and nine minimals. It presents 59 comparable pairs of elements and 302 contradictions.

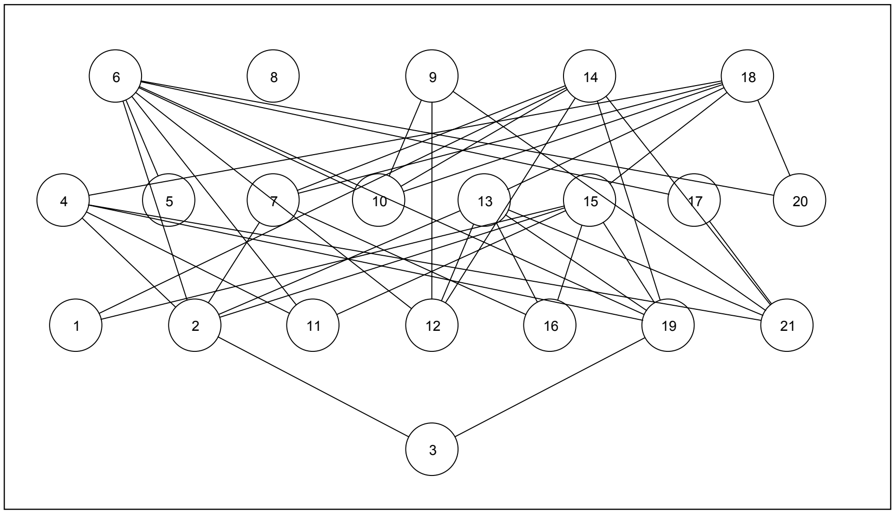


Figure 2.8 – Hasse diagram on the principal components calculated on ranks of social and disease variables in the Wards of Hull.

The Hasse diagram developed on quartile data is shown in Figure 2.9. It is organized in six levels, with two maximal equivalent elements, four minimals and no isolated elements. The number of comparabilities has increased greatly, from 46 (original data) to 93, while the incomparabilities decreased significantly, from 328 (original data) down to 236. The diagram is now clear in appearance, the relations among the elements being more evident and not one isolated element is now present.

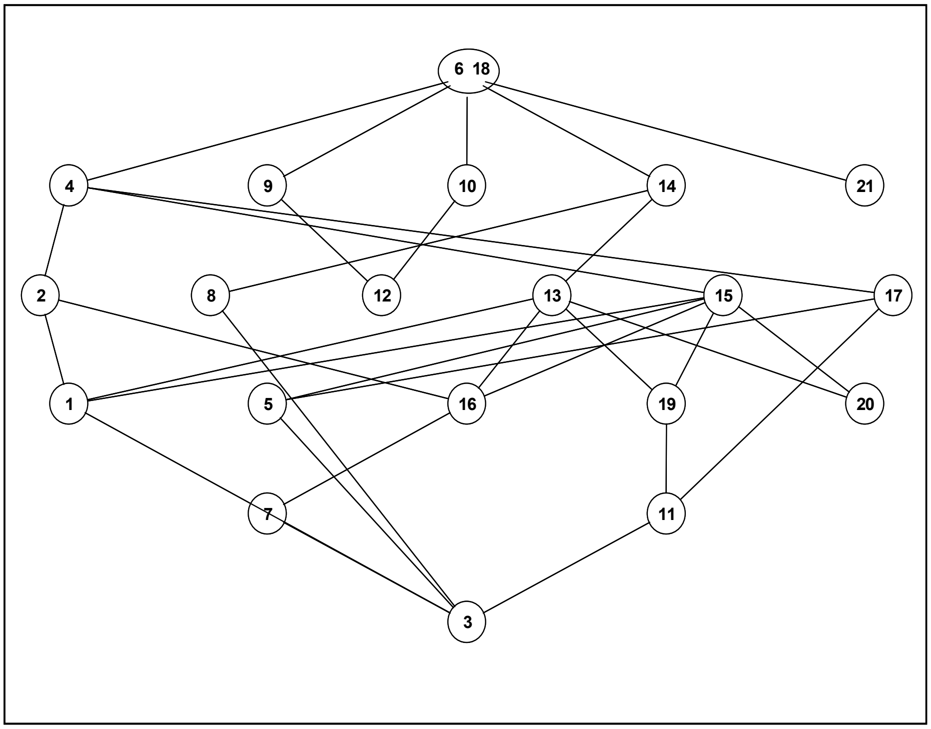


Figure 2.9 – Hasse diagram on the quartile data of social and disease variables in the Wards of Hull.

The Hasse diagram developed on three principal components calculated on quartile data is shown in Figure 2.10: it is a five level diagram with four maximals, three isolated and six minimals. It presents 60 comparable pairs of elements and 302 contradictions.

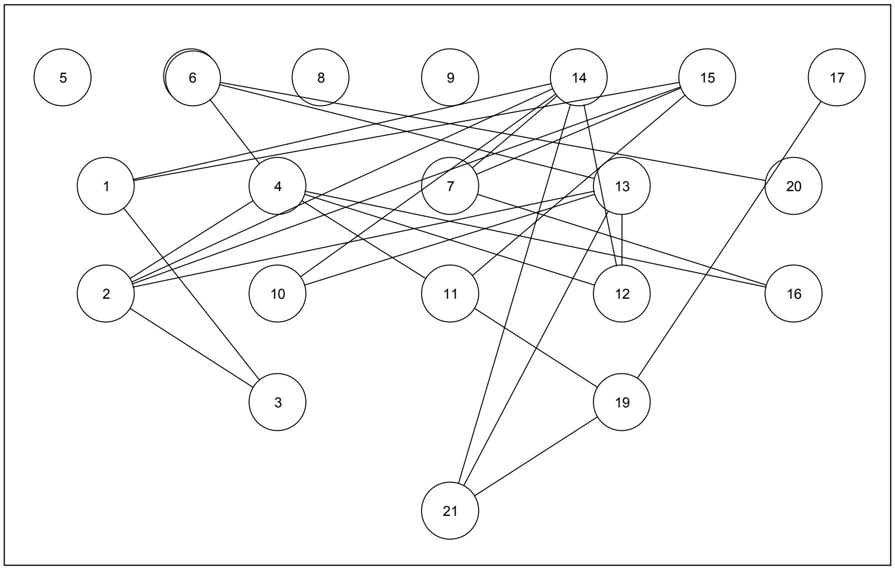


Figure 2.10 – Hasse diagram on the principal components calculated on quartile data of social and disease variables in the Wards of Hull.

The Hasse diagram developed on binary data is shown in Figure 2.11. It is organized in six levels, with five maximal equivalent elements, one minimal and no isolated elements. The number of comparabilities has increased greatly, from 46 (original data) to 149, while the incomparabilities decreased significantly, from 328 (original data) down to 150. The diagram is now very clear in appearance and the relations among the elements are even more evident than in the quartile Hasse diagram.

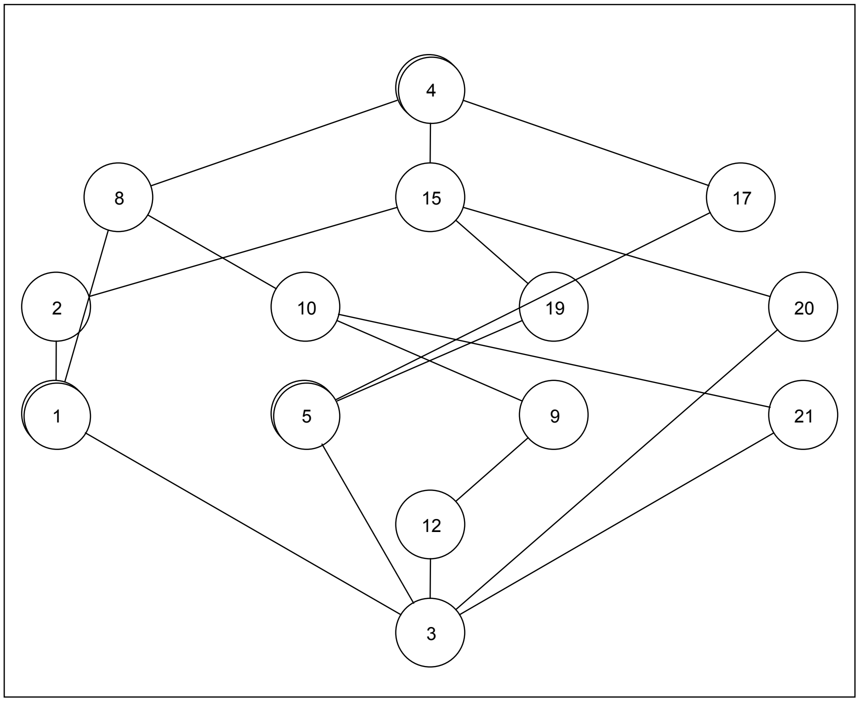


Figure 2.11 – Hasse diagram on binary data of social and disease variables in the Wards of Hull.

Finally, the Hasse diagram developed on three principal components calculated on binary data is shown in Figure 2.12: it is a three level diagram with 69 comparable pairs of elements and 310 contradictions.

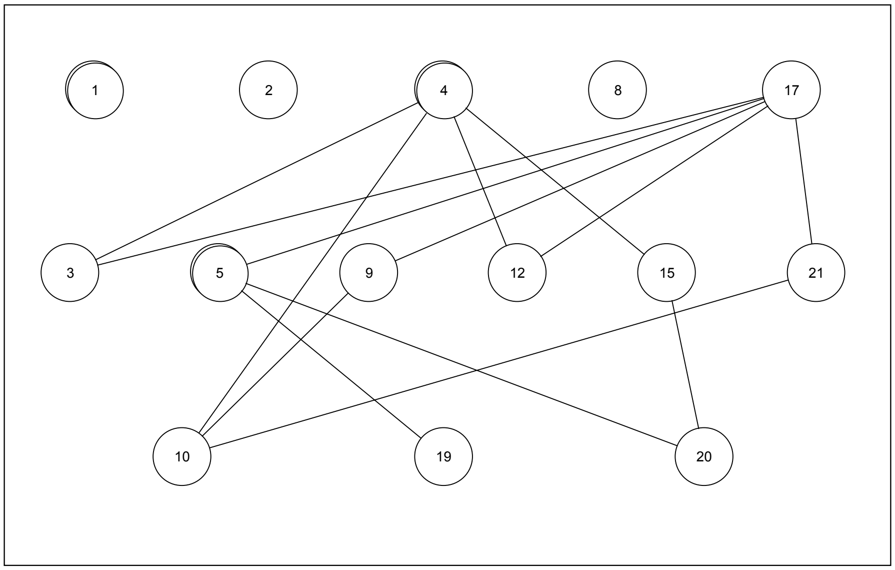


Figure 2.12 – Hasse diagram on principal components calculated on binary data of social and disease variables in the Wards of Hull.

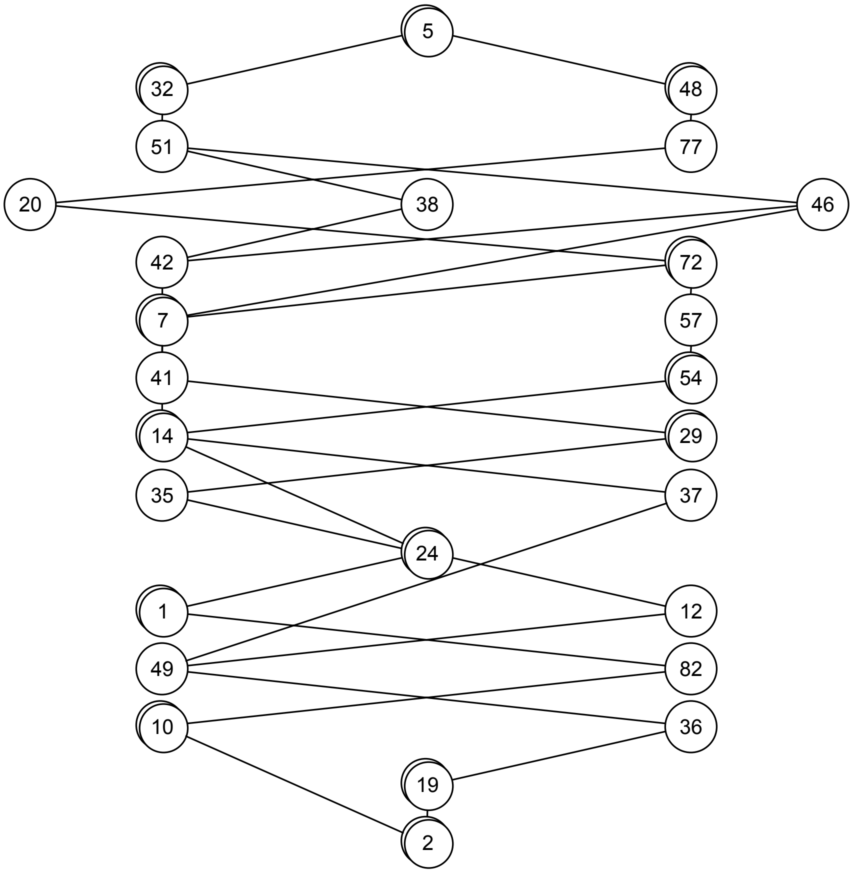
Table 2.15 summarises the differences among the diagrams:

Dataset	*Variables*	*Levels*	*Comparability*	*Incomparability*
Original data	7	3	46	328
PCA on original data	4	3	27	366
Ranked data	7	3	46	328
PCA on ranked data	3	4	59	302
Quartile data	7	6	93	236
PCA on quartile data	3	5	60	302
Binary data	7	6	149	150
PCA on binary data	3	3	69	310

Table 2.15 – Characteristics of the Hasse diagrams.

According to the results obtained, it seems that if the Hasse diagram performed on the original data is characterized by too many contradictions, induced by data variability that was probably due to measurement error, rank transformation like quartile and binary transformations appears a more suitable approach than Principal Component data reduction. Therefore, broad order statistics seems to be a very useful tool to support the partial order ranking technique as it allows the significant reduction, or even the deletion, of measurement error, thus solving the incomparabilities of Hasse diagrams. Instead Principal Component Analysis is a common pre-processing tool to be used when the purpose is to synthesize data description, discarding noise, to reduce data dimensionality, discarding unnecessary variables, and to find principal properties of the multivariate systems.

Moreover, as a high number of ties occurs when only a subset of order statistics is used, for example, deciles or quartiles, broad order statistics provides a significant object reduction. Thus a comparison with cluster analysis was performed, applying quartiles and binary transformation to the toxicity data used above for the cluster example (Table 2.3). The Hasse diagrams obtained on quartile and binary data for 83 chemicals are shown in Figures 2.13 and 2.14, respectively. The diagram resulting from quartile transformation is much more readable than that developed on original data; it is arranged on fifteen levels, and the original 83 chemicals are reduced to 27 equivalence classes. Moreover the number of comparabilities increases from 2788 to 3484, while the number of incomparabilities decreases from 1230 to 416.



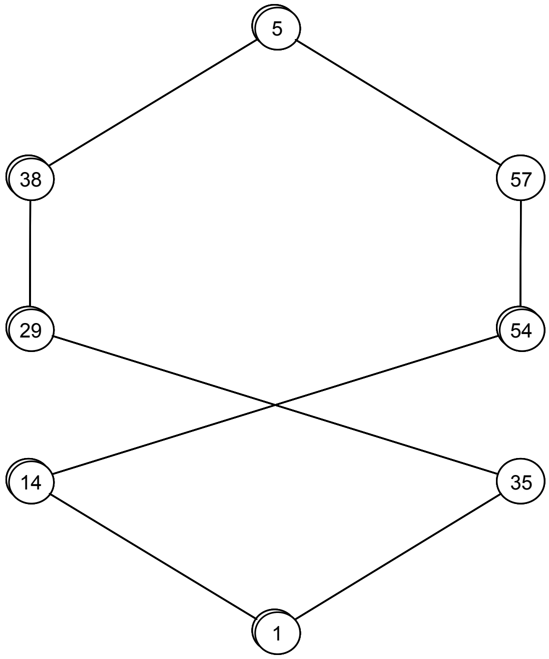
Equivalent chemicals

{1;11}
{2;3;4;6;9;15;16;22;26;40;47;59;61;63;83}
{5;13;17;21;27;34;43;55;60;68;74;80}
{7;8;23;25;30;31;44;50;69;71;76}
{10;18}
{14;56}
{19;58}

{24;28;39;53;62;64;67;70;75;81}
{29;45}
{32;33;52;66;79}
{48;78}
{54;65}
{72;73}

Figure 2.13 – Hasse diagram on quartile data of 83 chemicals

The diagram that resulted from binary transformation is even simpler than the one developed on quartile data; it is arranged on five levels and the original 83 chemicals are reduced to 8 equivalence classes. Moreover the number of comparabilities increases from 2788 to 4600, while the number of incomparabilities decreases from 1230 to 72.



Equivalent chemicals

{1;2;3;4;6;9;10;11;12;15;16;18;19;22;24;26;28;36;39;40;47;49;53;58;59;61;62;63;
64;67;70;75;81;82;83}

{5;7;8;13;17;20;21;23;25;27;30;31;32;33;34;43;44;46;48;50;51;52;55;60;66;68;69;
71;72;73;74;76;77;78;79;80}

{14;37;56}

{29;45}

{38;41;42}

{54;65}

Figure 2.14 – Hasse diagram on binary data of 83 chemicals

The analysis performed confirms that broad order statistics seems to be a suitable pre-processing tool to support the Hasse diagram technique, also providing a satisfactory solution to the two main drawbacks:

- noise and measurement error reduction (or elimination)
- element reduction

Moreover, compared with Principal Components analysis, broad order statistics has some main advantages:

- it preserves ordinal relations among the elements
- it allows an easy result interpretation as it does not create new variables

Compared with traditional Clustering methods, broad order statistics shows the following advantages:

- it does not measure element distance, thus it does not require metric choice or a variable scaling procedure
- it preserves ordinal relations among the elements
- it is a robust method, as it does not depend on subjective choices (metric, number of clusters, element similarity within each cluster..).

2.2.5 Broad order distance (similarity) matrices

Broad order statistics logic seems natural, so that it can be used as a pre-processing tool to support multivariate data analysis techniques when “quantitative” information cannot be used. If a ranked transformation has been performed, comparisons among objects require a distance measure suitable for rank transformed data. Once quartile rank transformation has been performed on the original data matrix, the distance between the two elements *i* and *k* can be calculated as follows:

$$d_{ik} = \sum_{j=1}^{p} \left| q_{ij} - q_{kj} \right|$$

q_{ij} being the quartile value of the *i-th* element for the *j-th* variable. The normalized distance is then derived by dividing the distance d_{ik} to its maximum value, according to the following expression:

$$d^{*}_{ik} = \frac{d_{ik}}{(Q-1) \cdot p}$$

Q being the quartile used and *p* the total number of variables. This distance takes its maximum value of 1 when two objects are entirely different, and its minimum value of 0 for objects that are identical over all the quartile transformed variables. This distance can be used to measure the association between objects. It is an Euclidean metric distance with the following properties:

1. minimum 0; if a = b, then $d^{*}_{ab} = 0$
2. positiveness; if a ≠ b, then $d^{*}_{ab} > 0$
3. symmetry: $d^{*}_{ab} = d^{*}_{ba}$
4. triangle inequality: $d^{*}_{ab} + d^{*}_{bc} \geq d^{*}_{ac}$

Moreover being a normalized distance it allows the evaluation of the absolute distance among objects.

A similarity coefficient can be derived from the above distance as:

$$S_{ik} = (1 - d^{*}_{ik})$$

This distance can be used to develop association matrices, like similarity or distance matrices, to be successively analyzed by any clustering method or multidimensional scaling.

2.2.6 Principal Component Analysis on ranks representing qualitative random variables

When the ranks represent qualitative categories related by hierarchical monotonic ordering, the only relationship among the rank values is “less than” or “greater than”. Three approaches can be used to perform principal component analysis on such data. First, the nonmetric nature of the data can be ignored, and Spearman’s correlation used to measure the correlation among the multivariate rankings. Secondly, a nonmetric correlation coefficient, like Kendall’s rank coefficient τ, can be used, and the correlation matrix derived from such coefficient decomposition. Third, a nonmetric algorithm can be developed where the component structure is invariant under monotone transformations of the data. The use of Kendall’s rank correlation in place of Spearman’s rank correlation corrects for the nonmetric nature of continuous rank order variables. However it can be argued that the correction is only partial, as it does not involve the algorithm used to compute the loading and the score coefficients. An alternative to principal component analysis performed on ranked data is simply to recover the minimum dimensional factor representation which is invariant under monotone transformations of the data. Such an analysis is called “nonmetric”, in contrast to the usual “metric” analysis. Kruskal and Shepard [Kruskal and Shepard, 1974] developed an algorithm to perform nonmetric analysis by a least squares monotonic regression. The minimization is carried out by iterative numerical methods. Although it has theoretical appeal, this approach is affected by two practical drawbacks: First, it demands long computational times, even for not so big data sets. Second, the nonmetric approach can hardly compete with principal components when variables are nonlinear, and especially when, in addition, they have errors in measurement.

Rank data lack much of the quantitative information present in continuous random variables. If continuity can be reasonably assumed, as a working hypothesis or by *a priori* theoretical reasoning, then a factor approach can be used as if the variables were continuous. If the continuity assumption is weak, a nonparametric coefficient like Kendall’s τ can be used, and in this case the principal component analysis provides a fictitious but useful summary of the data.

2.3 Indices for partial ranking analysis

Once a data analysis is performed by a partial order ranking method, as for total ranking, the quality of the obtained results has to be established. As is usual for regression and classification strategies, the quality of a ranking procedure must be evaluated by a deep analysis and by several indices, i.e. scalar functions that describe features of a partial ordered set, and must allow the comparison of different rankings. A Hasse diagram can be roughly described by a set of numbers of relevance, counting the number of equivalence classes with more than one element (NECA), the maximum number of elements in an antichain (W(E)), the number of lines in the longest chain (L(E)), the number of levels (NL), the number of elements in the level that contains the most elements (NEL), the number of maximals and minimals (N.Max and N.Min), the number of equivalence classes (Z), the number of comparabilities (V) and the number of incomparabilities (U). Based on these numbers, several indices for partial ranking analysis have been proposed and new ones are proposed here and compared with those already defined in the literature.

2.3.1 Information content and degeneracy index

As for total order ranking, one of the main aspects to be analysed once a ranking procedure is performed is the amount of information available from the ranking. The indices used to analyse total ranking information content and degeneracy can be used analogously for partially ordered sets. Thus the information content (I_c), which quantifies the degree of diversity of the elements in the partial set, the maximal information content (${}^{\max}I_C$), which represents the information content needed to characterize all of the N elements, the total information content (I), the standardized Shannon's entropy (H^*), the Hartley information index (I_N), the Brillouin redundancy index (R), the standardised Gini index (G^*), the informational energy content (I_E), the Bruggermann degeneracy index $k(N)$ and the absolute degeneracy degree (D) of a ranking can be calculated according to the expressions defined in chapter 1.

2.3.2 Comparability degree

Peculiar information encoded in a partial ordered set is that related to the comparability degree which can be quantified by a simple comparability index (χ),proposed here. Taking into account the number of comparabilities in the ranking, it is defined as:

$$\chi = \frac{V(N,R)}{N(N-1)/2} \qquad 0 \le \chi \le 1$$

where the numerator *V*(*N,R*) represents the number of comparable pairs of elements counted in only one direction and the denominator corresponds to the maximum theoretical value and is used to scale the values between 0 and 1.

This index assumes value 1 for the chain case, i.e. total order, which represents the maximum comparability, whereas, value 0 is assumed for the antichain case where no comparabilities exist. It must be observed that this index assumes value 1 for both the one chain case and the theoretical case of all elements equal each other, as in both these cases the comparability is maximum.

2.3.3. Discrimination power by ranking

When partial order ranking is performed for priority settings it is of great relevance to evaluate the ranking procedure capability of discriminating elements according to different ranks; the discrimination power by ranking (*DbyR*) of an order set, proposed in a similar formula even for totally ordered rankings, can be calculated as:

$$DbyR = \chi - \frac{D}{L} \qquad 0 \le DbyR \le 1$$

where χ is the comparability degree, *D* the degeneracy degree and *L* the number of levels. It can be observed that in the case of a chain, total order ranking, the comparability degree takes value 1, thus this expression is equal to the one defined for total ranking. The discrimination power index ranges from value 0 for the case of one antichain to 1 for the case of one chain, and increases with the increasing of the comparability degree and the decreasing of the degeneracy index: it can be observed that *DbyR* , by taking into account both the number of comparabilities in the ranking and the amount of degeneracy of each equivalence class, permits the distinguishing of the case of one chain with no degeneracy, for which χ = 1, *D* = 0 and thus *DbyR* = 1, from the case of all elements equal to each other for which χ = 1, *D* = 1 and thus *DbyR* = 0.

2.3.4 Stability indices

Partial order ranking is determined by the criteria considered in the ranking procedure, the actual information base, thus by changing the information base (IB), different orders arise. The set of criteria used may vary, and an additional criterion may be used in the information basis. Thus it is of interest to forecast the effect on the ranking of increasing the number of considered criteria, i.e. evaluate the ranking stability. The stability index proposed in the literature [Brüggemann and Voigt, 1996] is defined as follows:

$$P(N,R) = U(N,R) \,/\, S(N)$$

with

$$S(N) = U(N,R) + 2 \cdot V(N,R) - k(N,R)$$

where *V*(*N,R*) is the number of comparabilities, *U*(*N,R*) the number of incomparabilities (counted in both the directions) and *k*(*N,R*) the Bruggermann degeneracy index. This stability index ranges from 0 to 1: when *P*(*N,R*) is near zero, then *U*(*N,R*) must be near zero and in such a case adding an attribute may have quite a big influence on the ranking, in fact the higher the number of criteria, the greater the probability that contradictions (incomparabilities) in ranking exist among criteria.
Conversely, when *P*(*N,R*) is near 1, then *U*(*N,R*) must be near *S*(*N*), and adding an attribute may have a little influence on the ranking.

The quantity *P*(*N,R*) does not differentiate between full degeneracy and the one chain case because, in both cases, no incomparability appears (*U*(*N,R*)=0).

Nevertheless the stability of the case of full degeneracy is different from the one chain case stability. An example may be useful to better understand this concept: let *E* be the set constituted of two elements (*s,t*) and *IB* the actual information base of *R* attributes

In the case of full degeneracy:

$$(E, IB) = \{(s,t)\}$$

on adding an attribute, full degeneracy may still exist or a chain may arise.

In the case of a chain:

$$(E, IB) = \{s,t\}$$

on adding an attribute, the chain may still exist or an antichain may arise. Thus, assuming the antichain case to be the case of maximum stability, because adding an attribute changes nothing as the number of incomparabilities is already maximum, from the moment that the one-chain is nearer the antichain case than the full degeneracy, the one-chain case should be more stable than the full degeneracy case.

To take account of the differing stability of the one chain case and the case of full degeneracy, a new ranking stability index is proposed:

$$StR = \left(\frac{1-\sqrt{D}}{L}\right)^{\chi} \qquad 0 \le StR \le 1$$

where *D* is the absolute degeneracy index, χ the comparability degree defined above and *L* the number of levels. This index ranges from 0 for full degeneracy to 1 for the case of one antichain, and increases with decreasing degeneracy, with the comparability decreasing and with the decreasing of the number of levels. It can be observed that for a chain with no degeneracy, where $L = N$ and $D = 0$, *StR* takes the value of $1/N$, which is assumed as the stability of an ordered chain with N elements.

2.3.5 Complexity index

The Hasse diagram technique providing a graphical representation of the results needs to be clear and not too complex. For this reason, the appearance of the diagram can be analysed, as far as concerns complexity, by a complexity index [Bruggeman *et al.*, 2001a] defined as follows:

$$C_x = \frac{U(N,R)}{S} \qquad 0 \le C_x \le 1$$

where $U(N,R)$ is the number of incomparabilities, counted for each pair of elements twice, and S is the number of all connections. Thus Cx takes the value 1 in the case of a total antichain and the diagram is not complex; it takes a value equal to 0 in both the case of a total chain and the case of all the elements belonging to one equivalence class, and both these cases are considered not complex. It takes a value between 0 and 1 in all the other cases where some complexity exists. As the three cases, chain, antichain and all elements equal, all correspond to a not complex diagram, a modified index of complexity is proposed:

$$C'_x = 1 - \left|C_x - \chi\right| \qquad 0 \le C'_x \le 1$$

This modified index takes the value 0 for one chain and one equivalent class ($C_x = 0$, $\chi = 1$) as well as for a total antichain ($C_x = 1$, $\chi = 0$) and it takes the maximum value 1 when the two contributions of incomparability and comparability are equal.

2.3.6 Diversity index

Other useful information encoded in partial ordered ranking is the diversity existing among the elements [Pudenz *et al.*, 1999]. A ranking characterised by many incomparabilities between elements, indicates that the elements analysed are of high diversity as far as concerns the criteria they are described with. Therefore antichain corresponds to maximum diversity which can be measured as:

$$div = \frac{NEL(N,R) - 1}{N - 1} \qquad 0 \leq div \leq 1$$

where *NEL*(*N*,*R*) is the number of elements in the level, which contains the most elements, and *N* is the total number of elements. In an antichain *NEL*(*N*,*R*) = *N* and *div* = 1; whereas for a chain *NEL*(*N*,*R*) = 1 and *div* = 0.

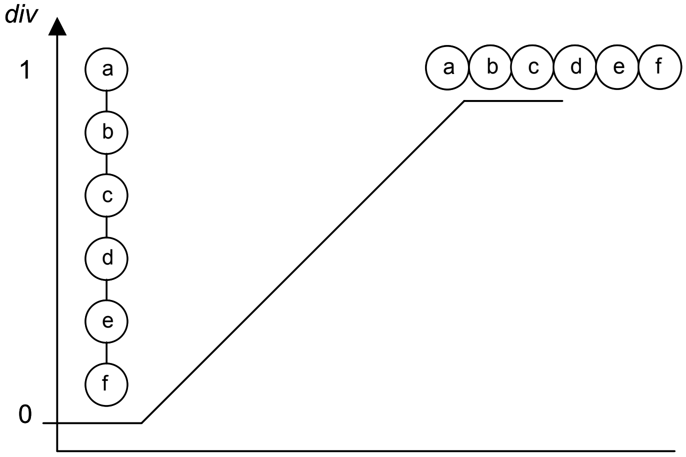


Figure 2.15 – Minimum and maximum diversity of ranking.

If equivalent classes with more than one element exist, the diversity is calculated as:

$$div = \frac{NEL(N,R)-1}{Z-1} \qquad 0 \le div \le 1$$

where Z is the number of equivalent classes.

2.3.7 Selectivity index

The selectivity of a partial ordered ranking is a measure of its capability to providing a unique orientation from “good” to “bad“ and therefore it is assumed maximum in a total chain and minimum in a total antichain, as in this case all the elements are incomparable with each other and no orientation is founded. A selectivity index has been proposed [Pudenz *et al*., 1999] and is defined as:

$$T = \frac{L-1}{N-1} \qquad 0 \le T \le 1$$

and for equivalent classes with more than one element, it is computed as:

$$T = \frac{L-1}{Z-1} \qquad 0 \le T \le 1$$

where Z is the number of equivalent classes.

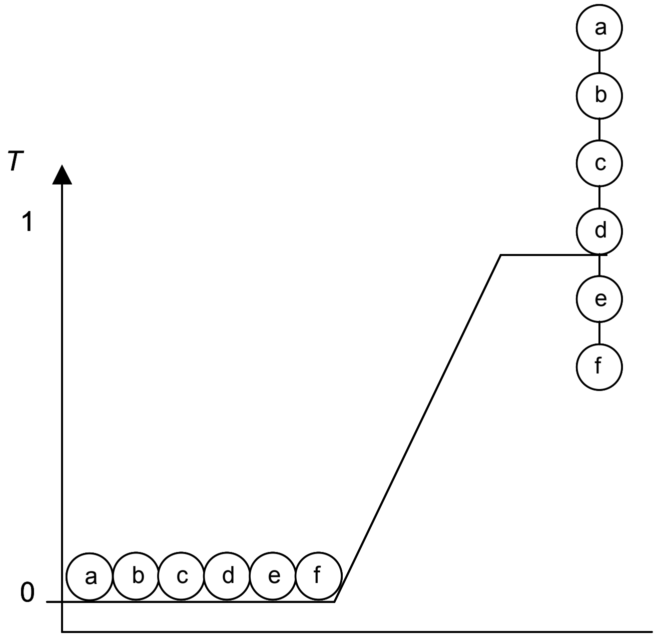


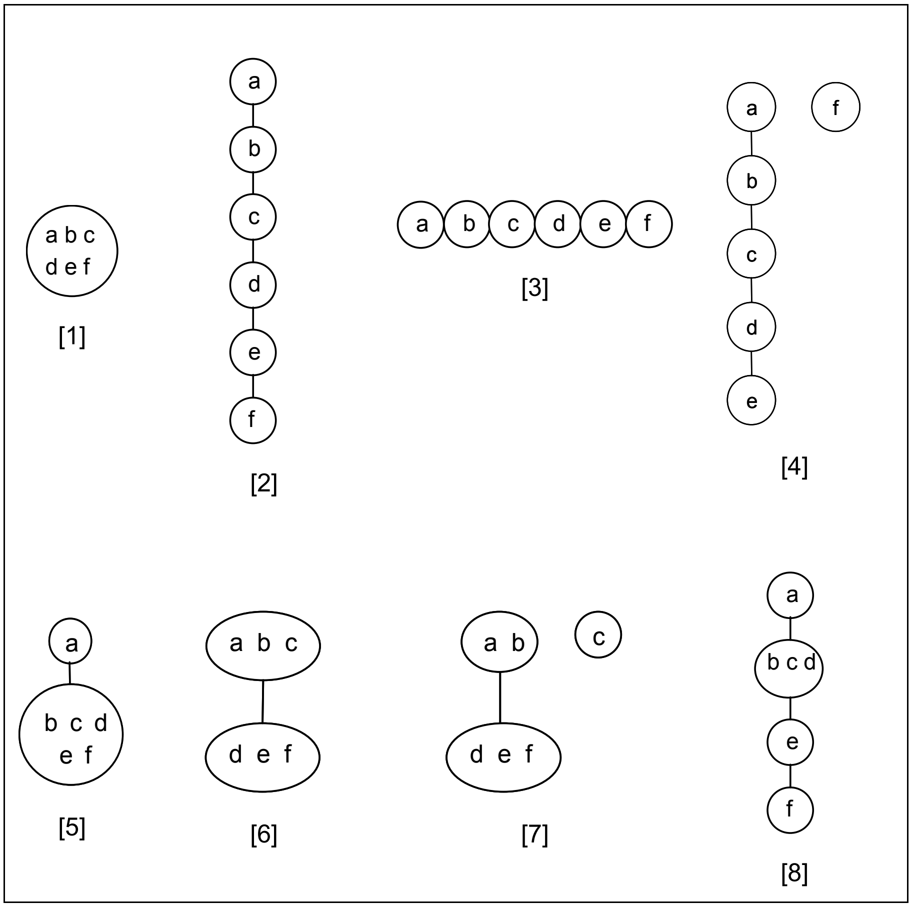
Figure 2.16 – Minimum and maximum selectivity of ranking.

2.4 Ranking indices comparison

To highlight the different information encoded by the indices described above, the indices have been calculated and compared on theoretical examples and a real data set.

Theoretical examples

Figure 2.17 shows the theoretical examples, each of six elements, analysed and compared by the ranking indices.



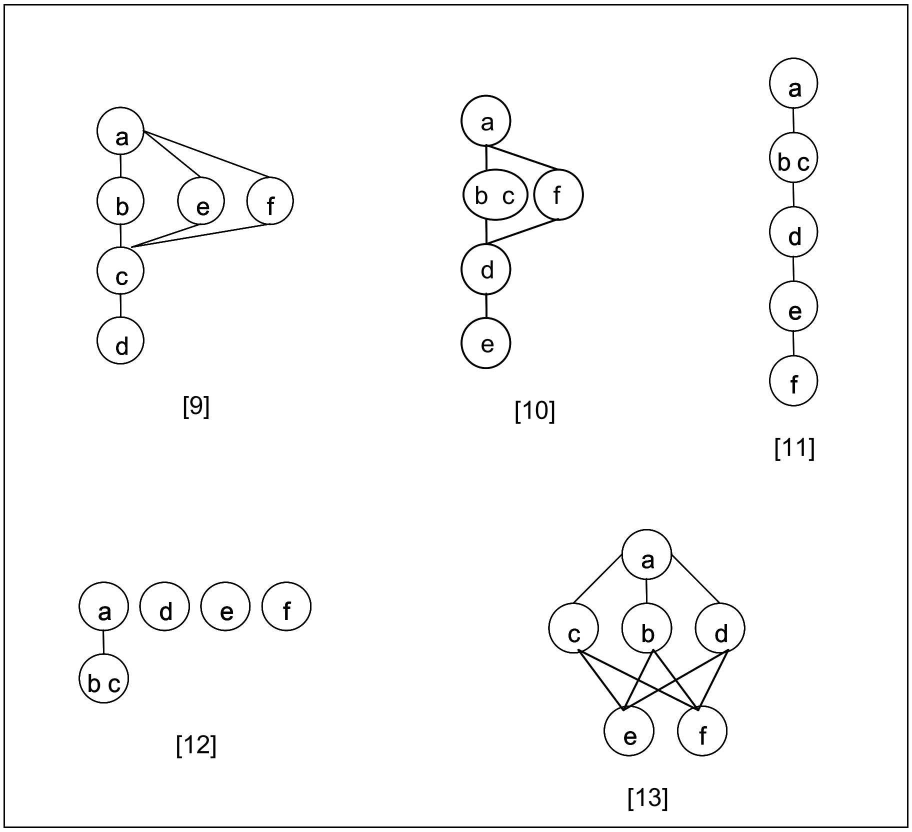


Figure 2.17 – Theoretical examples of partially ordered sets

For each theoretical example we calculated: the standardized Shannon's entropy index (*H**), the standardized Gini entropy index (*G**), the informational energy content (I_E), the Brüggemann standardized degeneracy index (k_{std}), the absolute degeneracy degree (*D*), the comparability degree (χ), the discrimination power by ranking (*DbyR*), the two stability indices (*StR* and *P*), the complexity indices (*Cx* and *Cx'*), the diversity (*div*) and selectivity indices (*T*). Table 2.16 lists the values.

The first case represents a system characterised by complete degeneracy with all the elements being equal: the two entropy indices, the standardised Shannon and the Gini entropies, take their minimum value of 0, whereas degeneracy is maximum according to the three degeneracy indices (I_E, k_{std}, D). In this case all the elements are comparable, thus comparability (χ) is maximum and takes a value of 1. Elements are not discriminated by ranking ($DbyR = 0$): according to the criteria used they look completely similar, thus no ranking can be established. The stability (P and StR) is minimum as adding a criterion is likely to change the diagram. The diagram is not complex ($Cx = 0$, $Cx' = 0$); no diversity ($div = 0$) exists among the elements and no orientation ($T = 0$) is provided by the ranking procedure.

Example	*H**	*G**	I_E	k_{std}	*D*	χ	*DbyR*	*StR*	*P*	*Cx*	*Cx'*	*div*	*T*
Case 1	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Case 2	1.00	1.00	0.17	0.00	0.00	1.00	1.00	0.17	0.00	0.00	0.00	0.00	1.00
Case 3	1.00	1.00	0.17	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00
Case 4	1.00	1.00	0.17	0.00	0.00	0.67	0.67	0.34	0.33	0.33	0.67	0.20	0.80
Case 5	0.25	0.33	0.72	0.67	0.80	1.00	0.60	0.05	0.00	0.00	0.00	0.00	1.00
Case 6	0.38	0.60	0.50	0.40	0.80	1.00	0.60	0.05	0.00	0.00	0.00	0.00	1.00
Case7	0.56	0.74	0.39	0.27	0.60	0.67	0.37	0.23	0.45	0.33	0.67	0.50	0.50
Case 8	0.69	0.79	0.33	0.20	0.40	1.00	0.90	0.09	0.00	0.00	0.00	0.00	1.00
Case 9	1.00	1.00	0.17	0.00	0.00	0.80	0.80	0.33	0.20	0.20	0.40	0.40	0.60
Case 10	0.87	0.93	0.22	0.07	0.20	0.87	0.82	0.18	0.14	0.13	0.27	0.25	0.75
Case 11	0.87	0.93	0.22	0.07	0.20	1.00	0.96	0.11	0.00	0.00	0.00	0.00	1.00
Case 12	0.87	0.93	0.22	0.07	0.20	0.20	0.10	0.77	0.86	0.80	0.40	0.75	0.25
Case 13	1.00	1.00	0.17	0.00	0.00	0.73	0.73	0.45	0.27	0.27	0.53	0.40	0.40

Table 2.16 – Numerical examples of ranking indices.

The second case is a chain with maximum entropy and minimum degeneracy. Comparability (χ) is maximum as all the elements are comparable and, despite the previous case, discrimination power by ranking is maximum (*DbyR* = 1) providing a totally ordered diagram; the ranking procedure is able to discriminate all the elements according to their different ranks. The stability *P* index does not distinguish between this case and the previous one, while the proposed *StR* stability index calculates greater stability (equal to 1/*N*) for the chain than for the one equivalent class case. The diagram has no complexity, no element diversity and the selectivity is maximum (*T* = 1) providing a unique orientation from “good” to “bad“. The third case corresponds to an antichain, with no degeneracy: thus the information content indices and the degeneracy indices take the same values as in the previous case. The comparability is minimum as elements are not comparable, and the ranking procedure provides no discrimination, being unable to assign ranks to the elements. Stability is maximum since adding a criterion leaves the diagram totally unchanged, as the number of incomparabilities does not decrease on increasing the number of criteria. The complexity calculated by *Cx* is equal to 1, while the new *Cx’* index evaluates the diagram as not complex. Diversity among the elements is maximum, no relationships can be found among them, whereas selectivity is minimum as no vertical orientation is established. Cases 1 to 3 are the theoretical extreme cases, all the others are located within these extremes.

Case 4 corresponds to a chain with an isolated element and no degeneracy. This case seems similar to the one of a total chain, however the ranking index values are quite different. Comparability is much lower than the chain case as one element of the six cannot be compared with the others, and also discrimination power is lower for the same reason: only five elements of the six are discriminated according to their ranks, and there is no idea of the isolated element’s relation with the others. Stability increases with increasing incomparability: this is the reason for the greater stability of case 4 with respect to case 2. The complexity estimated by *Cx* is lower than *Cx’,* and the diagram has only a slight degree of diversity. Selectivity is quite high as vertical orientation is provided for five of the six elements. The fifth and sixth cases correspond to chains with differently distributed degeneracy: both consist of two equivalence classes of different density. The standardised Shannon entropy and the standardised Gini entropy take a greater value for case 6 than for case 5. The informational energy content and the standardised Brüggemann expression are all influenced by the way the degeneracy is distributed in the system, being greater in case 5 than in case 6. In contrast, the absolute degeneracy degree takes the same value for cases 5 and 6, revealing the presence of two information sources in both the posets. All the other ranking indices take the same values for case 5 and case 6: the comparability index, the *P* stability index, the complexity, the diversity and the selectivity indices take the same values as for the one chain case (case 2), whereas the discrimination power by ranking and the stability *StR* take lower values with respect to the one chain case because of their degeneracy. Case 7 is a chain with two equivalence classes and an isolated element. The absolute degeneracy degree provides a higher degeneracy value than the informational energy content and the Bruggermann degeneracy, which seems to underestimate the degeneracy. Besides high comparability, the discrimination power by ranking is low because of the high degeneracy; stability calculated by the *StR* index is lower than that by the *P* index, confirming that the former is more sensitive to system degeneracy than the latter. The diversity and selectivity indices take the same values as the comparabilities and incomparabilities are perfectly balanced. Case 8 represents a chain, with a degeneracy greater than case 2 and lower than cases 5 and 6, as confirmed by the information content and degeneracy indices. Comparability is maximum; discrimination power is high even if the maximum is not reached because of the degeneracy. Being a chain, the comparability and selectivity indices take the maximum value, whereas the complexity and diversity take the minimum and stability is very low. Case 9 corresponds to a diagram with three different chains and no degeneracy: comparability and discrimination power are high, while stability is low. The complexity is lower than that of Cases 4 and 7; moreover the *Cx* index seems to underestimate the complexity with respect to *Cx'*. Diversity exists among the elements, as shown by the incomparabilities of elements *b*, *e* and *f*. Selectivity is quite high as three different orientations are provided. Case 10 is quite similar to case 9, but degeneracy replaces incomparability. This is the reason for the decreased element diversity and increased selectivity. Case 11 is a chain again, similar to case 8 but with lower degeneracy, this is the reason for the slight increase in discrimination capability. Case 12 is characterised by the presence of a short chain, a little degeneracy and several incomparabilities. The comparability degree, the discrimination power by ranking and the selectivity values are low, while the stability and element diversity are high. Case 13 is not too far from case 9, but it is characterised by more incomparabilities as indicated by the decreased values of comparability degree, discrimination power and selectivity and by the increased values of stability.

It is now evident that the visualisation of the partially ordered set is just the first step in the examination of the results of a partial ranking procedure; a lot of information can be extracted just by looking at the Hasse diagram, but the need to compare different diagrams while avoiding any arbitrary interpretation prompts the need for several ranking indices.

A case study application

A partial ranking procedure by the Hasse diagram technique has been performed on data of fruit composition [Sicheri and Borsarelli, 1989]. The ranking indices proposed in the work were calculated and compared with those already defined in the literature. The dataset is reported in Table 2.17: it consists of 31 fruits described according to their composition. The following variables were used in the ranking procedure: eatable content (%), water content, protein, lipid, available glucides, amide glucides, soluble glucides, fibred glucides, energy, Iron, Calcium, Phosphorus, Thiamine, Riboflavin, Niacin, Vitamin A and Vitamin C.

Fruit	ID	*Eat.*	H_2O	*Prot*	*Lip*	*Glu ava*	*Glu ami*	*Glu sol*	*Glu fib*
Apricot	1	94	86.3	0.4	0.1	6.8	0	6.8	0.6
Egriot cherry	2	85	84.2	0.8	0	10.2	0	10.2	1
Pineapple	3	57	86.4	0.5	0	10	0	10	0.4
Peanut	4	79	7.1	26	47.2	11.2	6.7	4.5	2.3
Orange	5	80	87.2	0.7	0.2	7.8	0	7.8	0.6

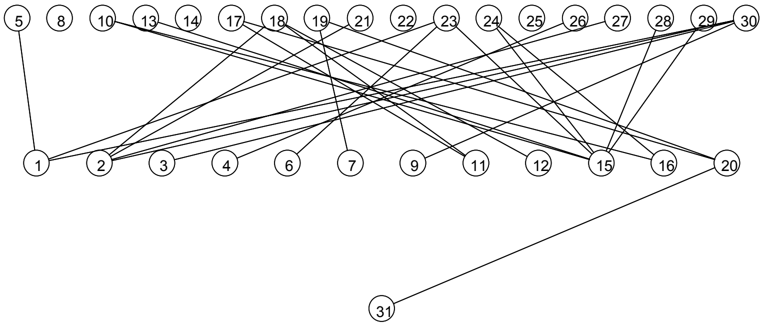
Orange juice	6	100	89.3	0.5	0	8.2	0	8.2	0
Banana	7	65	76.8	1.2	0.3	15.5	2.4	12.8	0.5
Chestnut	8	69	41	3.5	1.8	42.4	34.3	8.1	1
Cherry	9	86	86.2	0.8	0.1	9	0	9	1
Watermelon	10	52	95.3	0.4	0	3.7	0	3.7	0
Figs	11	75	81.9	0.9	0.2	11.2	0	11.2	0.7
Prickly pear	12	64	83.2	0.8	0	13	1.6	13	0.3
Strawberry	13	94	90.5	0.9	0.4	5.3	0	5.3	0.6
Raspberry	14	100	84.6	1	0.6	6.5	0	6.5	3
Lemon	15	64	89.5	0.6	0	2.3	0	2.3	0.6
Lemon juice	16	100	92.1	0.2	0	1.4	0	1.4	0
Persimmon	17	97	82	0.6	0.3	16	1.6	16	0.5
Tangerine	18	87	85.3	0.8	0.2	12.8	0	12.8	0.3
Manderin	19	80	81.4	0.9	0.3	17.6	0	17.6	0.8
Pomegranate	20	59	80.5	0.5	0.2	15.9	0	15.9	0.2
Apple	21	94	85.6	0.2	0.3	11	0	11	1
Apple Cot	22	79	84.3	0.3	1	6.3	0	6.3	1.7
Sum.Melon	23	47	90.1	0.8	0.2	7.4	0	7.4	0.3
Wint. Melon	24	51	94.1	0.5	0.2	4.9	0	4.9	0.3
Medlar	25	66	85.3	0.4	0.4	6.1	0	6.1	0.5
Walnut	26	58	19.2	10.5	57.7	5.5	2.1	3.4	2
Pear	27	91	85.2	0.3	0.4	9.5	0	9.5	0.6
Peach	28	91	90.7	0.8	0.1	6.1	0	6.1	0.6
Grapefruit	29	70	91.2	0.6	0	6.2	0	6.2	0.6
Plum	30	89	87.5	0.5	0.1	10.5	0	10.5	0.3
Grapes	31	94	80.3	0.5	0.1	15.6	0	15.6	0.2

Fruit	*ID*	*Energy*	*Fe*	*Ca*	*P*	*Thi*	*Rib*	*Niam*	*Vit. A*	*Vit. C*
Apricot	1	28	0.5	16	16	0	0	0.5	360	13
Egriot cherry	2	41	0.4	15	17	0	0.1	0.4	24	7
Pineapple	3	40	0.5	17	8	0.1	0	0.2	7	17
Peanut	4	571	3.2	60	239	1.5	0.1	0.4	0	2
Orange	5	34	0.2	49	22	0.1	0.1	0.2	71	50
Orange juice	6	33	0.2	15	17	0.1	0	0.4	38	44
Banana	7	66	0.8	7	28	0.1	0.1	0.7	45	16
Chestnut	8	189	1.2	38	89	0.2	0.4	1.4	0	18.2
Cherry	9	38	0.6	30	18	0	0	0.5	19	11
Watermelon	10	15	0.2	7	2	0	0	0	37	8
Figs	11	47	0.5	43	25	0	0	0.4	15	7
Prickly pear	12	53	0.4	30	25	0	0	0.4	10	2
Strawberry	13	27	0.8	35	28	0	0	0.5	0	54
Raspberry	14	34	1	49	52	0.1	0	0.5	13	25
Lemon	15	11	0.1	14	11	0	0	0.3	0	50
Lemon juice	16	6	0.2	14	10	0	0	0.2	0	43
Persimmon	17	65	0.3	8	16	0	0	0.3	237	23
Tangerine	18	53	0.3	30	19	0.1	0.1	0.3	25	37
Manderin	19	72	0.3	32	19	0.1	0.1	0.3	18	42
Pomegranate	20	63	0.3	0	10	0.1	0.1	0.2	0	8
Apple	21	45	0.3	6	12	0	0	0.3	8	5
Apple Cot	22	34	0.1	4	14	0	0	0.7	0	14
Sum.Melon	23	33	0.3	19	13	0.1	0	0.6	189	32
Wint. Melon	24	22	0.3	21	16	0	0	0.5	5	12
Medlar	25	28	0.3	16	11	0	0.1	0.4	27	1
Walnut	26	582	2.6	131	238	0.6	0.2	0.8	6	0

Pear	27	41	0.3	6	11	0	0	0.1	0	4
Peach	28	27	0.4	4	20	0	0	0.5	27	4
Grapefruit	29	26	0.3	17	16	0.1	0	0.2	0	4
Plum	30	42	0.2	13	14	0.1	0.1	0.5	16	5
Grapes	31	61	0.4	27	4	0	0	0.1	4	6

Table 2.17 – Experimental data of fruit composition used for the ranking analysis.

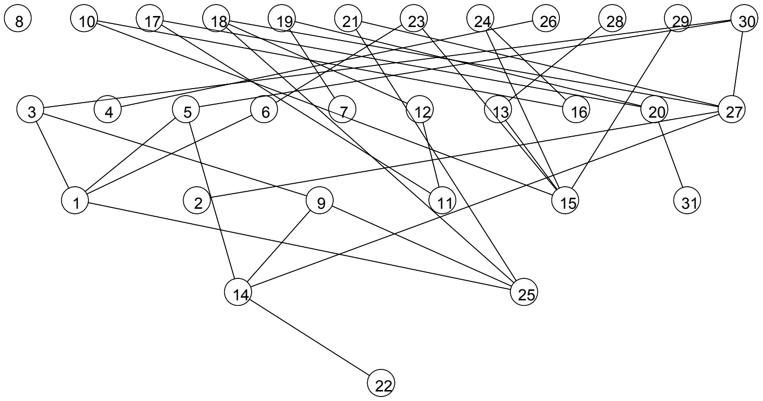
By using different subsets of variables, different partial ordered rankings were obtained. The variable selection was performed so as to obtain significantly different Hasse diagrams and no alimentary meaning is necessarily associated with them. The obtained Hasse diagrams were analysed and compared by the ranking indices described above. Table 2.18 shows the ten attribute combinations considered; together with the corresponding Hasse diagrams.

Attributes used	*Hasse diagram*
H_2O Lipid Energy Case 1	 5 8 10 13 14 17 18 19 21 22 23 24 25 26 27 28 29 30 1 2 3 4 6 7 9 11 12 15 16 20 31

H_2O

Energy

Case 2

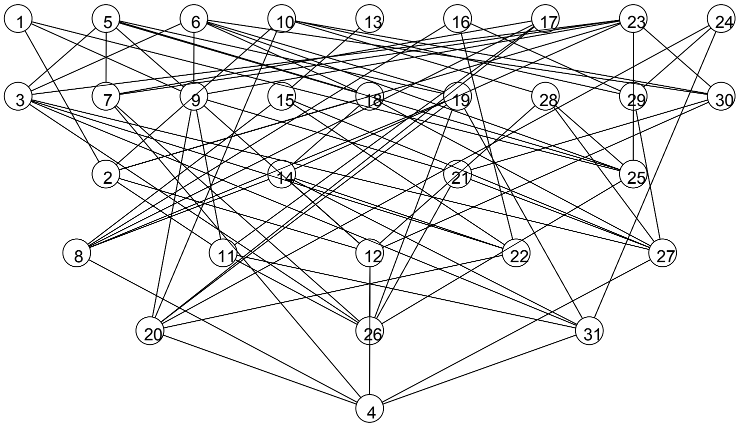


H_2O

VitaminA

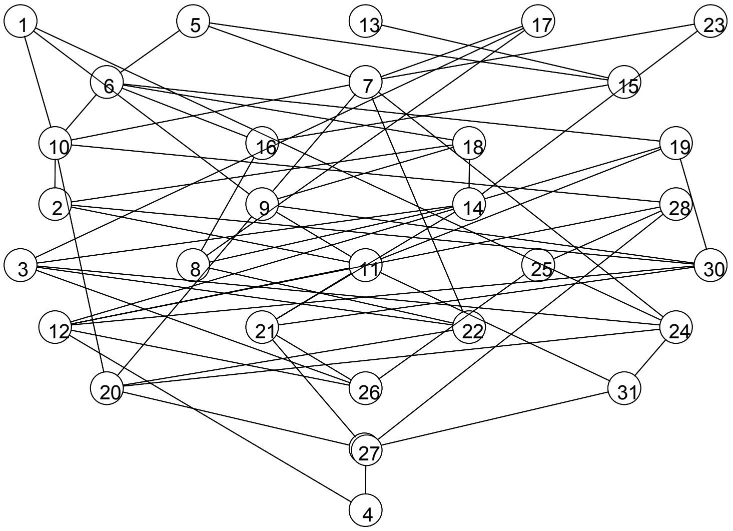
VitaminC

Case 3



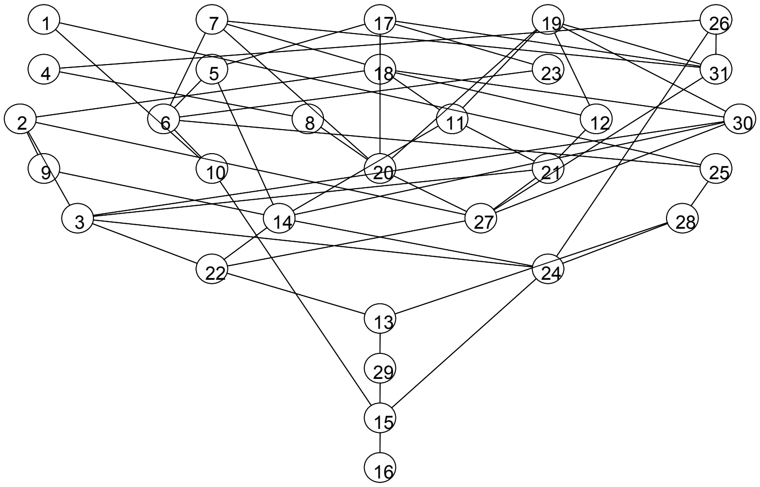
VitaminA
VitaminC

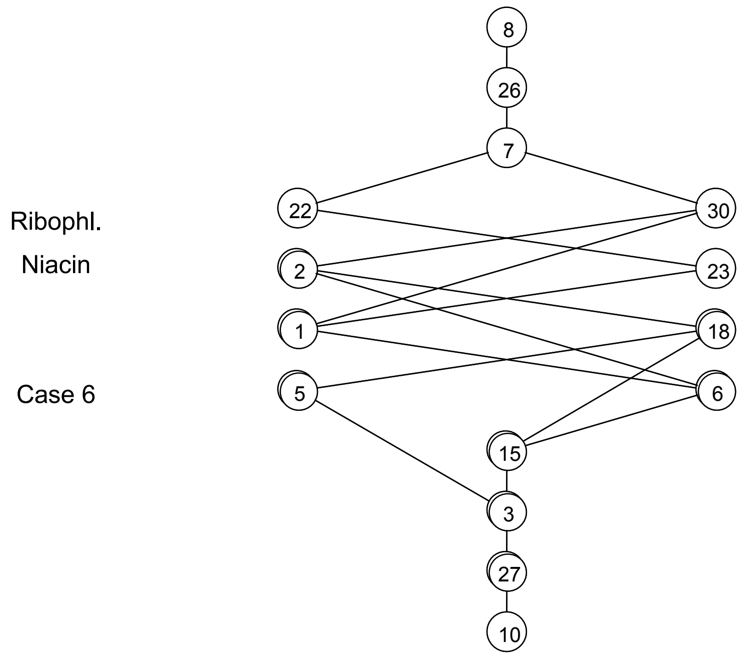
Case 4

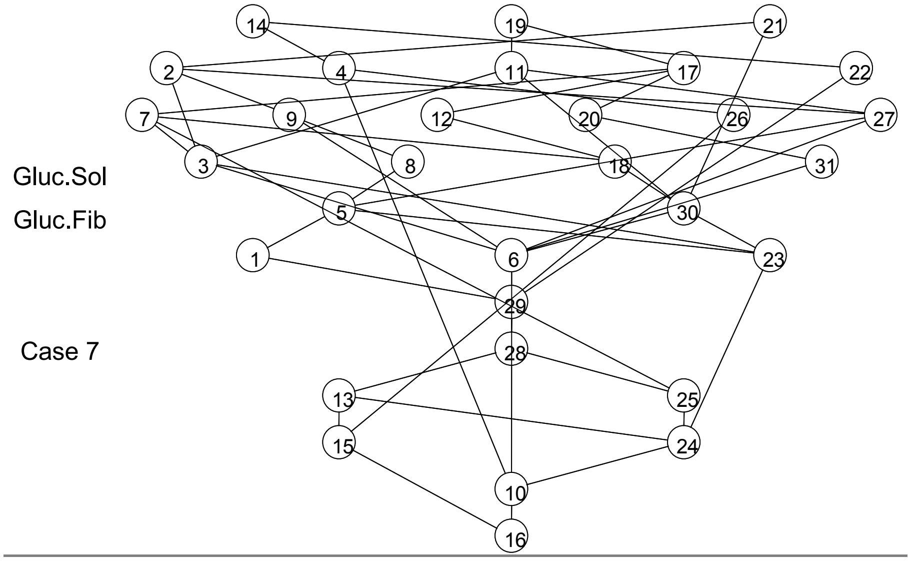


Energy
VitaminA

Case 5



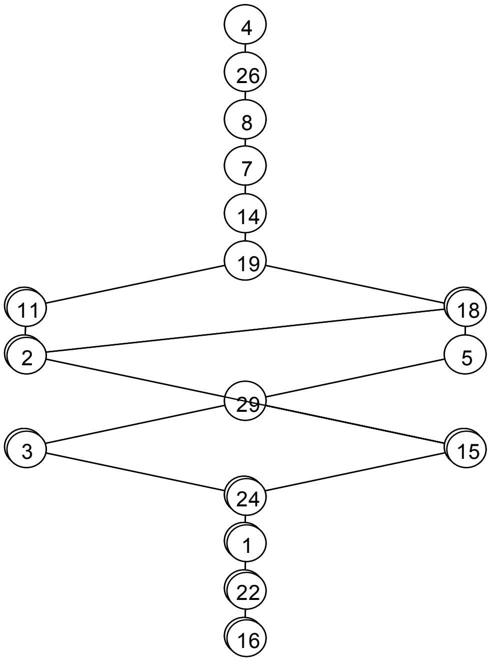

Ribophl.
Niacin
Case 6
8
26
7
22
30
2
23
1
18
5
6
15
3
27
10


Gluc.Sol
Gluc.Fib
Case 7
14
19
21
2
4
11
17
22
7
9
12
20
26
27
3
8
18
31
5
30
1
6
23
29
28
13
25
15
24
10
16

Protein

Thiamin

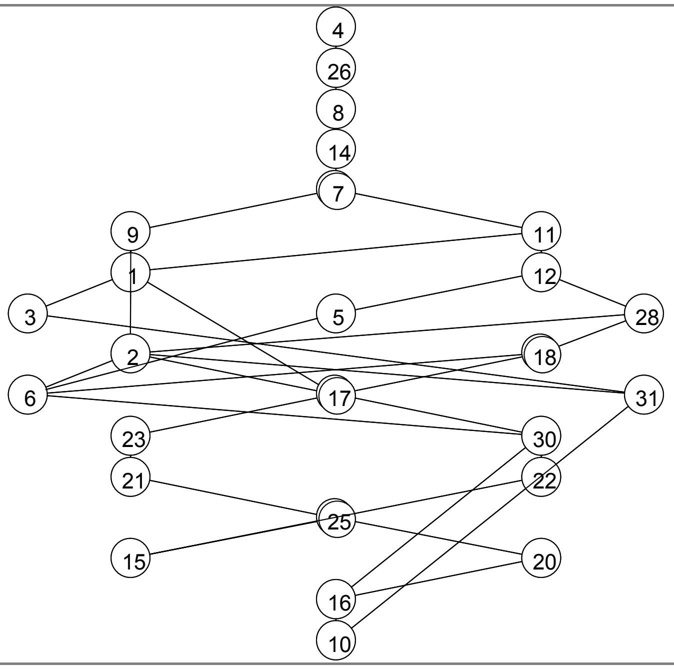
Case 8



Fe

P

Case 9



Gluc. Ava
Gluc. Amid

Case 10

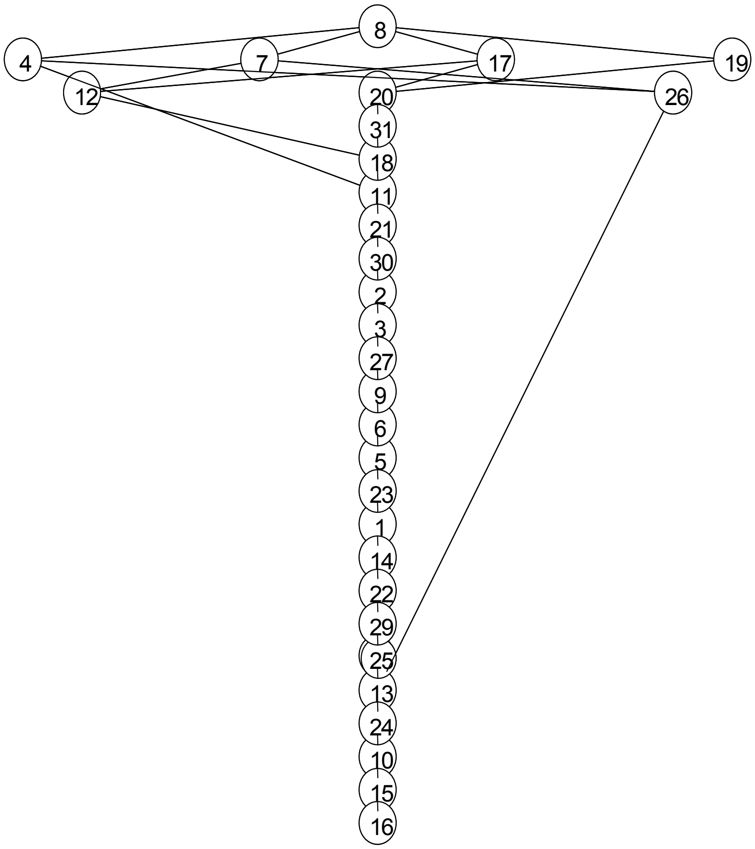


Table 2.18 – Comparison of Hasse diagrams obtained from different combinations of attributes.

As can be easily observed the results of partial ranking analysis depend strictly on the attributes used to perform the analysis. The diagrams corresponding to the different attribute combinations are fairly different from each other. To compare them in a more objective way than the one based only on their appearance, the ranking indices were calculated and their numerical values are shown in Table 2.19. The informational energy content (I_E) is not reported in the table as, for all the cases, it assumes a value of 0.03.

Case	*H**	*G**	k_{std}	*D*	χ	*DbyR*	*StR*	*P*	*Cx*	Cx'	*div*	*T*
1	1.00	1.00	0.00	0.00	0.06	0.06	0.94	0.94	0.94	0.12	0.57	0.07
2	1.00	1.00	0.00	0.00	0.13	0.13	0.81	0.87	0.87	0.26	0.37	0.13
3	1.00	1.00	0.00	0.00	0.38	0.38	0.51	0.62	0.62	0.76	0.27	0.17
4	0.99	1.00	0.00	0.03	0.60	0.59	0.24	0.41	0.40	0.80	0.14	0.28
5	1.00	1.00	0.00	0.00	0.53	0.53	0.30	0.47	0.47	0.94	0.30	0.17
6	0.74	0.94	0.06	0.53	0.84	0.79	0.04	0.17	0.16	0.32	0.21	0.71
7	1.00	1.00	0.00	0.00	0.50	0.50	0.29	0.50	0.50	1.00	0.13	0.37
8	0.79	0.96	0.05	0.47	0.90	0.86	0.03	0.11	0.10	0.20	0.25	0.81
9	0.93	1.00	0.01	0.17	0.82	0.81	0.07	0.19	0.18	0.36	0.12	0.60
10	0.99	1.00	0.00	0.03	0.92	0.92	0.04	0.08	0.08	0.16	0.10	0.83

Table 2.19 – Numerical values of ranking indices calculated for different attribute combinations.

From cases 1 to 10 the number of levels increases as a consequence of the decreasing number of incomparabilities. As far as concerns information and the degeneracy indices, it must be pointed out that all the diagrams show high entropy and low degeneracy. For cases 1, 2, 3, 5 and 7, the standardised Shannon entropy and Gini entropy take the maximum value of 1, while the degeneracy calculated by the standardised Brüggemann degeneracy (k_{std}) and the absolute degeneracy degree (*D*) are equal to the minimum value of 0. The diagrams corresponding to the cases (1, 2, 3, 5, 7) are all characterized by the absence of degeneracy, each equivalent class being composed by only one element. From cases 4, 10, 9 to cases 8 and 6 the entropy decreases and degeneracy increases, these cases being based respectively on 30, 30, 26, 17 and 15 equivalence classes. This different entropy degree is well encoded in the standardised Shannon entropy, whereas the Gini entropy index seems less susceptible to reflecting low entropy differences. Analogously, the standardised Brüggemann degeneracy (k_{std}) with respect to the absolute degeneracy index (D) seems not able to catch low degeneracy differences. With the increasing number of levels from cases 1 to 10, the comparability among the elements and the ranking procedure capability of discriminating elements according to different ranks (*DbyR*) increases, even if they are not perfectly correlated directly. Case 4 corresponds to a nine level diagram, with 278 comparabilities, Case 5 to a ten level diagram with 245 comparabilities and Case 7 has a twelve level diagram with 231 comparabilities. Except for small differences, the higher the number of levels, the higher the comparability and the higher the discrimination power by ranking. As far as concerns the two stability indices, they show an opposite trend with respect to the comparability and discrimination power indices as they decrease with increasing number of comparabilities, tending to 0 for a totally ordered sequence (chain). For all the cases analysed the new stability index (*StR*) takes values smaller than the one proposed in the literature (*P*). For the diagrams investigated it is evident that the higher the number of incomparabilities, the higher the complexity index C_x value: it takes its highest value for Case 1, and smallest for Case 10, the former being the case with the higher incomparabilities (874) and the latter that with the lower (74). The diagram complexity is evaluated in a different way according to the new index *Cx'*, which takes the maximum value of 1 for the diagram of Case 7 since the two contributions of incomparability and comparability are balanced. The diversity index, being a measure of the diversity degree of elements as far as concerns the criteria used to describe them, is correlated to the number of incomparabilities and thus takes the highest value for Case 1, which has the highest number of incomparabilities (874), and the lowest value for Case 10, which has the lowest number of incomparabilities (74). An opposite trend is the one of the selectivity index which, being a measure of the unique orientation of the diagram, tends to 1 when the number of levels tend to the number of elements.

As pointed out above, one of the main drawbacks of the Hasse diagram technique is its strict dependence on the clear appearance of the graphical diagram, since very poorly structured diagrams with more incomparabilities than comparabilities, because of the roughness of the conflict, are useless. When there are too many contradictions or when data observed on a continuous scale are suspected of being affected by large measurement error or unknown forms of nonlinearity among the variables, “quantitative” information cannot be used. In such cases the original variables can be replaced by their rank orders. A high number of ties occur when only a subset of order statistics is used, for example, deciles or quartiles. The decision to use a reduced number of order statistics can be related to the aim of exploring the “main features” of multivariate data, or to perform a preliminary analysis before a more complete one. Obviously the results of the analysis depend on the chosen ranking scale, however replacing the original data by quartiles or deciles could reveal qualitative features which would otherwise be submerged by quantitative information. As an example of the significant reduction of incomparabilities induced by order statistics, the original variables used in Case 1, which is the one with the highest number of incomparabilities, were replaced by their quartile rank orders. The Hasse diagram obtained is shown in Figure 2.18. Circles with a double line indicate equivalence classes with more than one element. The diagram is made of the following 20 equivalent classes: (1,15), (2,12), (3,6,9), (4,8,26), (5), (7,19), (10,16,28,29), (11), (13), (14,22), (17), (18), (20), (21), (23), (24), (25), (27), (30), (31).

The number of incomparabilities decreases from 874 (original data) to 670 (quartile transformed data), and the number of levels increases from 3 to 5.

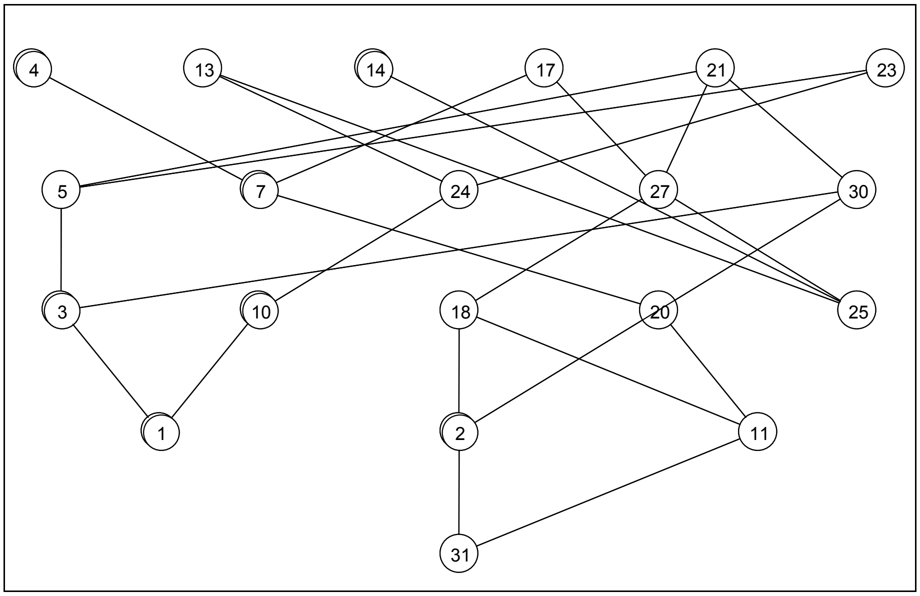


Figure 2.18 – Hasse diagram developed on quartile data of H_20, lipid and energy.

A deeper comparison of the Hasse diagrams developed on the original variables (A) and one on the quartile data (B) is provided by the ranking indices collected in Table 2.20.

	*H**	*G**	I_E	k_{std}	D	χ	*DbyR*	*StR*	*P*	*Cx*	Cx'	*div*	*T*
A	1.00	1.00	0.03	0.00	0.00	0.06	0.06	0.94	0.94	0.94	0.12	0.57	0.07
B	0.83	0.98	0.07	0.03	0.37	0.28	0.21	0.49	0.75	0.72	0.56	0.42	0.21

Table 2.20 – Numerical values of ranking indices calculated on original and quartile data of H_20, lipid and energy.

As expected, the rank transformation results in a decrease of the information content measured by the standardised Shannon and Gini entropies, and an increase of the degeneracy by the standardised Brüggemann degeneracy (k_{std}) and the absolute degeneracy degree (*D*).

The comparability (χ) and the ranking capability of discriminating elements according to different ranks (*DbyR)* increases significantly, whereas the stability decreases. The diagram complexity measured according to the *Cx* index decreases, as a consequence of the decreasing number of incomparabilities, whereas according to the new index *Cx'* it increases as a consequence of the decreased discrepancy between comparabilities and incomparabilities.

The diversity index, reflecting the incomparability decrease, itself decreases in the diagram developed on quartile data, whereas the selectivity index reflecting the level increase, increases.

2.5 Correlation analysis

A first rough analysis of the relationships among the above-described indices was performed by Principal Component Analysis on a data matrix made up of twenty-four cases described by the defined eleven ranking indices. The first thirteen cases are the theoretical examples of Figure 2.17, whereas the other eleven cases are the ones obtained on the fruit data. The first six principal components provide 99.7% of the total information (note that in the absence of experimental error also less informative PCs can be considered).

Figure 2.19 shows the loading plot relative to the first and second components.

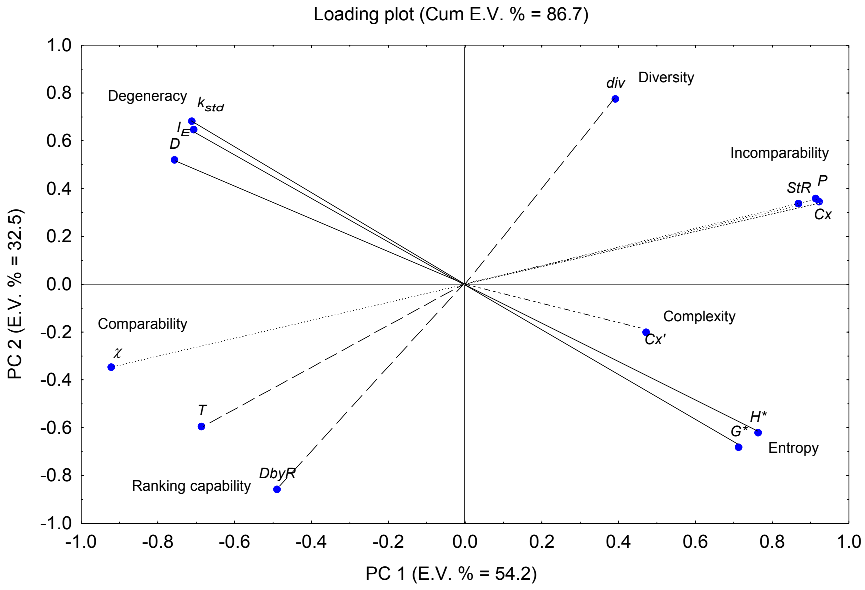


Figure 2.19 – Loading plot of the PC1 versus PC2.

The first two principal components explain 86.7% of the total variance and reveal four main classes of ranking indices, collected in Table 2.21:

Class	*Meaning*	*Indices*	*Dual meaning*	*Indices*
1	Entropy	*H*, G**	Degeneracy	*D*, k_{std}, I_E
2	Ranking capability	*DbyR, T*	Diversity	*div*
3	Comparability	χ	Incomparability	*StR, P, Cx*
4	Complexity	*Cx'*	-	-

Table 2.21 – Three main classes of ranking indices

A first class of indices is constituted by the entropy indices and their complementary degeneracy indices (indicated by continuous lines in all the PC graphs); as expected, both the standardized Shannon (*H**) and the Gini entropy indices (*G**) are closely correlated and inversely correlated to the information energy content (I_E), the Brüggemann standardized degeneracy index (k_{std}) and the absolute degeneracy degree (*D*).

A second class of indices, highlighted by the PC1 – PC2 loading plot, represents the ranking capability, i.e. the Hasse diagram vertical/horizontal orientation, where *DbyR* and *T* measure the vertical orientation (total ranking) and *div* the horizontal orientation, i.e. low total ranking ability (signed by hatched lines in all the PC graphs).

A third class of indices represents comparability versus its dual meaning of incomparability, where *StR* and *P* are both closely correlated and inversely to the comparability χ (signed by dotted lines in all the PC graphs). The complexity index *Cx* also belongs to this class and seems closely correlated to the stability *P* index.

The new complexity index *Cx'* constitutes the fourth class, as it shows different behaviour from the indices of the third class.

The remaining four components (13.0 % of explained variance) explain the internal differences among the indices of each class.

In particular, the loading plot of the third and fourth PCs (10.2% of explained variance, Figure 2.20) highlights how the behaviour of the complexity index Cx' differs from all the other ranking indices; moreover also the entropy/degeneracy indices show different behaviour (class 1), decorrelating *D* from k_{std} and *G** from *H**. For the second class, the same graph highlights the difference between *T* and *DbyR* and, for the third class, the difference of *StR* from *P* and χ.

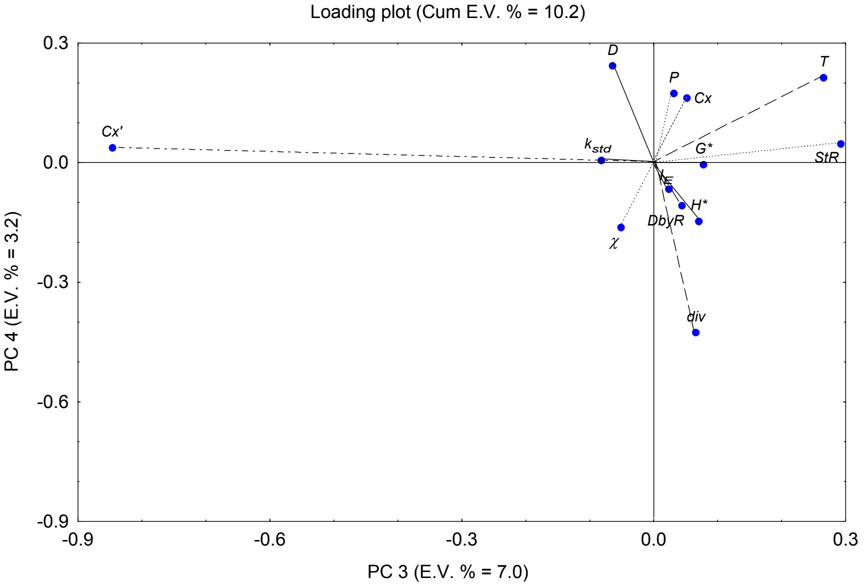


Figure 2.20 – Loading plot of the PC3 versus PC4.

Finally, further differences are shown in the third loading plot (2.1% of explained variance, Figure 2.21) for I_{E} and k_{std} (class 1), *DbyR* and *div* (class 2).

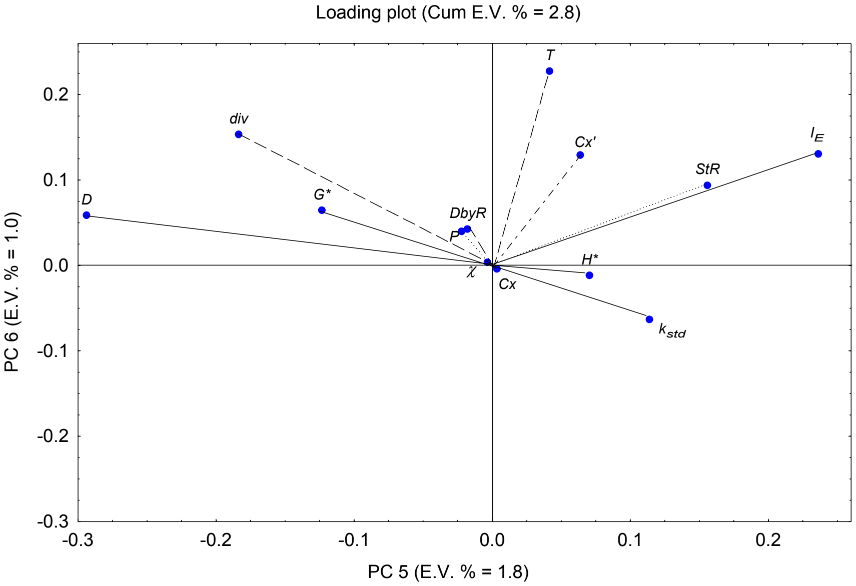


Figure 2.21 – Loading plot of the PC5 versus PC6.

Thus, according to this analysis, only *G** and k_{std} appear quite similar in their behaviour. The indices proposed (*D*, χ, *DbyR*, *StRand Cx’*) encompass all three classes of indices, with a meaning different from existing indices.

2.6 Criteria similarity analysis

Once partial ordered ranking has been developed, it is of interest to establish the degree of similarity among the criteria used to develop the ranking: for this purpose the criteria similarity index *CS* is proposed. For each pair of criteria *j* and *k*, the similarity between the two criteria is calculated according to the following expression:

$$CS_{jk} = 1 - \frac{U_{jk}}{N(N-1)} \qquad 0 \le CS_{jk} \le 1$$

U_{jk} being the number of incomparabilities in the diagram developed by using only the two criteria *j* and *k*, and *N* the total number of elements. In the case of an antichain, the number of incomparabilities being maximum and equal to *N*(*N*-1), the similarity between the two considered criteria is zero.

Once the similarity among each pair of criteria is calculated, the overall similarity among all the criteria can be defined:

$$CS = \frac{2 \cdot \sum_{jk} CS_{jk}}{R(R-1)} \qquad j > k \qquad 0 \le CS \le 1$$

R being the total number of criteria used for the order ranking.

For a total ordered ranking, the number of incomparabilities is zero, thus each contribution from a criteria pair being equal to 1, the overall similarity among the criteria reaches the maximum value of 1.

2.7 Sensitivity analysis

The analysis of the structure of an ordered set includes an analysis of the influence of each attribute on the ranking, i.e. the sensitivity analysis.

2.7.1 Senstivity by **W** matrix

The approach proposed by Halfon and Bruggermann [Halfon and Brüggemann, 1998; Brüggemann *et al*. ,2001b] to assess the importance of an attribute requires a comparison of the results from ranking performed with different attribute subsets; this implies Hasse diagram comparisons. To compare Hasse diagrams an appropriate metric must be found, by which the distance between any two partial order rankings can be calculated. The results are stored in a matrix **W**, which is a way of defining the distances among different posets. Matrix **W** stores the mutual Hasse diagram comparison. Selecting an element of interest, preferably a maximal element, is the starting point for the analysis. The selected element is called the "*key element*". The analysis of the key element requires a search for all the elements located lower than the key element that can be reached by a path, a sequence of connecting edges. These elements together with the elements equivalent but not identical to the key element are called *successors*. The set of successors of a key element "*k*" is denoted as *G*(*k*) and, by definition, *G*(*k*) does not include the key element itself. A successor set depends on the key element and on the attributes used. The sensitivity analysis is based on the analysis of different successor sets arising from different attribute subsets. To measure each attribute's influence on ranking, Hasse diagrams from each attribute subset are compared. This is done by selecting a key element and quantifying the effect of each attribute set on its successor set. For this purpose the residual set *R(k,B,C)* is introduced:

$$R(k,B,C)=\frac{G(k,B)}{G(k,C)}$$

G(k,B) being the successor set of the key element *k* on the B attribute subset , and *G(k,C*) being the successor set on the C attribute subset. The symmetric difference set *W(k,B.C)* is defined as:

$$W(k,B,C)=card\,R(k,B,C)+card\,R(k,C,B)$$

The square matrix, denoted by **W**(*k*) has *L* = 2^R-1 columns and rows, respectively. The sensitivity analysis is performed starting from the **W** matrix. However this matrix does not always need to be analysed in its entirely because, in the most of the cases, the interest is only in a few attribute sets. Thus, the sensitivity analysis of the criteria can be performed with the following steps:

- To find each criterion's relevance, requiires only the comparison of the full attribute set A with the A_i subsets, thus only one row of the **W** matrix is of interest:

$$W(k,A,A)\quad W(k,A,A_1)\quad W(k,A,A_2).....\quad W(k,A,A_R)$$

- To find each attribute's influence it is enough to compare the diagram induced by the full set of attributes (A) with those induced by the attributes sets with only (*R*-1) attributes. The effect of dropping one attribute is given by the remaining *R* entries of the first row. The remaining *R* matrix element of the first row generates a "sensitivity rule" of the key element *k*:

  $$\sigma(k) = \left[W(k,A,A_1)\ldots\ldots W(k,A,A_R)\right]$$

  where:

  $A = \{r_1, r_2, \ldots\ldots\ldots, r_R\}$ full set of attributes
  $A_i = \{r_1, \ldots, r_{i-1}, r_{i+1}, \ldots\ldots, r_R\}$ i-*th* attribute skipped

  σ(k) can also be written as [σ(1), σ(2),…, σ(*R*)]. The larger σ_i, the larger is the symmetrized difference between *G(k,A)* and *G(k,A$_i$)* and thus the larger the influence of the attribute r_i on the position of the key element *k* within the Hasse diagram on *A* compared with that on A_i.

- The matrix W(k) depends on the selection of the key element k. When more elements are to be analysed, the generalised expression *W(K,A$_i$,A$_j$)* is introduced:

  $$W(K,A_i,A_j) = \sum_{k=1}^{N} W(k,A_i,A_j) \qquad k \in K$$

  where *K* is any set of key elements and $W(K) = \sum_{k=1}^{N} W(k)$.

- All elements are selected as the key element; therefore instead of the **W***(k)* matrix we have a **W**(*E*), i.e. total matrix of set *E*, is analysed as measure of sensitivity which is quantified by:

  $$\sigma_i = W(E,A,A_i) \qquad 1 \le i \le R$$

$$\mathbf{W}(K)=\begin{vmatrix} W(k_1,A,A_1) & W(k_1,A,A_2) & \ldots\ldots\ldots & W(k_1,A,A_R) \\ W(k_2,A,A_1) & W(k_2,A,A_2) & \ldots\ldots\ldots & W(k_2,A,A_R) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ W(k_N,A,A_1) & \ldots\ldots\ldots & \ldots\ldots\ldots & W(k_N,A,A_R) \end{vmatrix}$$

$$\downarrow$$

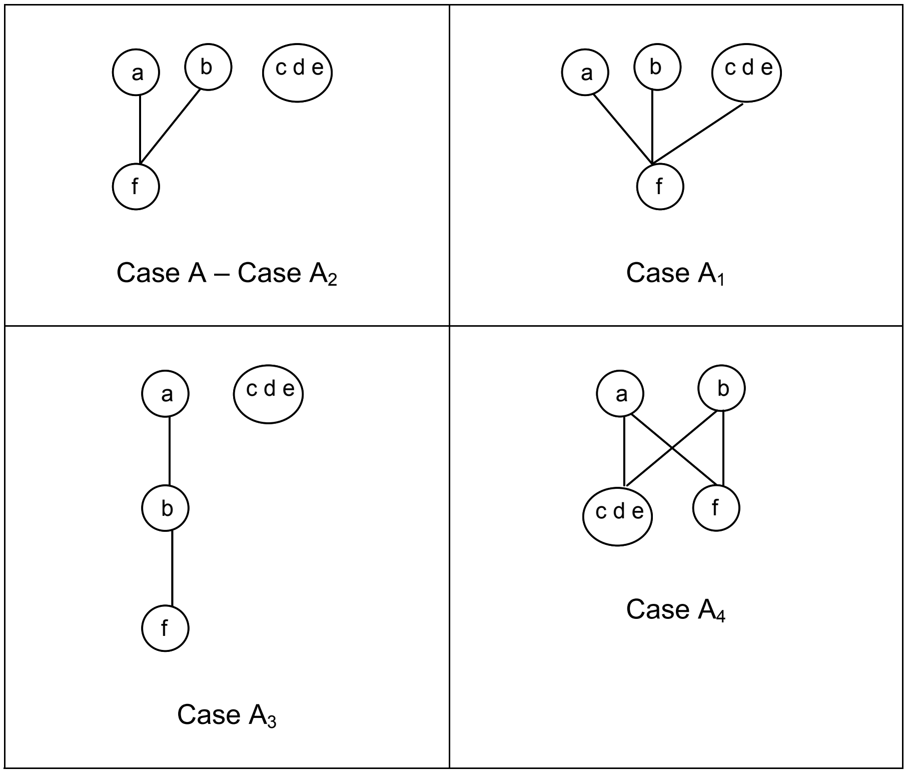
$$\sigma_1 = \sum_{j=1}^{N} W(k_j,A,A_1) \text{ = influence of the attribute 1.}$$

An example of sensitivity analysis is provided here on six elements described by four criteria. Table 2.22 shows the data matrix.

Element	r_1	r_2	r_3	r_4
a	5	9	11	12
b	4	6	22	10
c	1	2	4	16
d	1	2	4	16
e	1	2	4	16
f	3	1	2	3

Table 2.22 - Data matrix

The diagrams for comparison are illustrated in Figure 2.22.



Case A – Case A_2

Case A_1

Case A_3

Case A_4

Figure 2.22 – Hasse diagrams for different attribute sets on example data.

The corresponding total matrix for different combinations of attributes is:

$$\mathbf{W} = \begin{vmatrix} W(A,A) & W(A,A_1) & W(A,A_2) & W(A,A_3) & W(A,A_4) \\ W(A_1,A) & W(A_1,A_1) & W(A_1,A_2) & W(A_1,A_3) & W(A_1,A_4) \\ W(A_2,A) & W(A_2,A_1) & W(A_2,A_2) & W(A_2,A_3) & W(A_2,A_4) \\ W(A_3,A) & W(A_3,A_1) & W(A_3,A_2) & W(A_3,A_3) & W(A_3,A_4) \\ W(A_4,A) & W(A_4,A_1) & W(A_4,A_2) & W(A_4,A_3) & W(A_4,A_4) \end{vmatrix} =$$

$$= \begin{vmatrix} 0 & 3 & 0 & 1 & 6 \\ 3 & 0 & 3 & 4 & 9 \\ 0 & 3 & 0 & 1 & 6 \\ 1 & 4 & 1 & 0 & 7 \\ 1 & 0 & 0 & 0 & 0 \\ 6 & 9 & 6 & 7 & 0 \end{vmatrix}$$

As an example the *W*(A, A$_4$) value is calculated as follows:

W (a, A, A$_4$) = card *R*(a, A, A$_4$) + card *R*(a, A$_4$, A) = 0 + 3 = 3
W (b, A, A$_4$) = card *R*(b, A, A$_4$) + card *R*(b, A$_4$, A) = 0 + 3 = 3
W (c, A, A$_4$) = card *R*(c, A, A$_4$) + card *R*(c, A$_4$, A) = 0 + 0 = 0
W (d, A, A$_4$) = card *R*(d, A, A$_4$) + card *R*(d, A$_4$, A) = 0 + 0 = 0
W (e, A, A$_4$) = card *R*(e, A, A$_4$) + card *R*(e, A$_4$, A) = 0 + 0 = 0
W (f, A, A$_4$) = card *R*(f, A, A$_4$) + card *R*(f, A$_4$, A) = 0 + 0 = 0

$$\text{W(A, A4)} = \sum_{j=1}^{N} W(k_j, A, A_4) = 3 + 3 + 0 + 0 + 0 + 0 = 6$$

From the W matrix, the sensitivities are:

$$\sigma_1 = \sum_{j=1}^{N} W(k_j, A, A_1) = 3$$

$$\sigma_2 = \sum_{j=1}^{N} W(k_j, A, A_2) = 0$$

$$\sigma_3 = \sum_{j=1}^{N} W(k_j, A, A_3) = 1$$

$$\sigma_4 = \sum_{j=1}^{N} W(k_j, A, A_4) = 6$$

Therefore, attribute r_4 is the most important, whereas attribute r_2 does not have any influence on the order.

2.7.2 Backward sensitivity analysis

Another way to evaluate the importance of the attributes used to perform partial ranking is proposed here. The method consists in a stepwise technique starting with all the attributes and then selecting one attribute at a time, based on Hasse diagram comparison evaluated by the similarity index *S* (see Chapter 3). According to this index, the similarity between two partial order rankings is calculated comparing their Hasse matrices (A and B) as follows:

$$S(A,B) = 1 - \frac{\sum_{st} \left| h_{st}^{A} - h_{st}^{B} \right|}{2N \cdot (N-1)} \qquad\qquad 0 \le S(A,B) \le 1$$

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

$$s,t \in 1,2,\ldots,N \quad \text{and} \quad s \neq t \quad \text{and} \quad h_{st} \begin{cases} +1 & \mathit{if} \quad y_r(s) \ge y_r(t) \quad \text{for all } y_r \in IB \\ -1 & \mathit{if} \quad y_r(s) < y_r(t) \quad \text{for all } y_r \in IB \\ 0 & \text{otherwise} \end{cases}$$

All the attributes are removed one at a time and a calculation is made of the similarity between the diagram developed by the full attribute set A and those induced by the attribute sets with only (*R*-1) attributes. The *i*-

th attribute associated with the minimum S is the most influential one (i.e. the attribute that, if removed, provides the more diverse diagram).

$$Min\,S(A,A_i)\rightarrow i-th\ \text{attribute more influent ial}$$

being:

$A=\{r_1,r_2,.........,r_R\}$ full set of attributes

$A_i=\{r_1,....,r_{i-1},r_{i+1},......,r_R\}$ i-*th* attribute skipped

The less influential attribute is then removed and the procedure repeated on the remaining attributes; the new full set A' will be composed by *R*-1 attributes: The diagram developed by the new full set of attributes (A') is compared with those induced by the attribute sets with only (*R*-2) attributes. Thus, at any step the less influential attribute is identified and deleted. The procedure stops when the full set consists of only two attributes; thus a sequence of attribute influence is provided. An example will explain the procedure better: Table 2.23 shows the data of five elements described by four criteria.

Element	r_1	r_2	r_3	r_4
a	2	4	5	34
b	22	4	3	30
c	34	4	33	61
d	68	15	28	23
e	12	65	11	22

Table 2.23 the data matrix for sensitivity example.

The similarities between the diagram from all the attribute sets and those obtained deleting one attribute at a time are the following:

$S(\mathrm{A}, \mathrm{A}_1) = 0.95$

$S(\mathrm{A}, \mathrm{A}_2) = 0.90$

$S(\mathrm{A}, \mathrm{A}_3) = 1.00$

S(A, A$_4$) = 0.85

The third attribute, showing the highest similarity value is the least influential, thus it is deleted and the procedure repeated. The new full set of attributes A' is composed of attributes r_1, r_2 and r_4. The similarities between the diagram developed with the new full set A' and those induced by the attribute sets with only (R-2) attributes are calculated.

S(A', A'$_1$) = 0.95

S(A', A'$_2$) = 0.85

S(A', A'$_4$) = 0.75

being A'$_1$ = {r_2, r_4}

being A'$_2$ = {r_1, r_4}

being A'$_4$ = {r_1, r_2}

The first attribute showing the highest similarity value is the least influential, thus it is deleted. The remaining two attributes constitute the most influential attribute pair. Thus the following influence order arises:

$$r_3 \leq r_1 \leq r_2 \leq r_4$$

Therefore, the fourth attribute together with the second, is the most influential on the ranking; the first follows it. The third attribute is the least influential. The sensitivity analysis performed on the **W** matrix confirms the obtained results ($\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 0; \sigma_4 = 3$).

2.8 Total – Partial ranking comparison

The Partial ranking approach by the Hasse diagram technique has been applied to the data of twelve High Production Volume Chemicals (HPVC) described by production volume (PV), acute toxicity to fish (LC50), the partitioning coefficient between *n*-octanol and water (LogKow), and biodegradation [European Communities, 2000], previously used in Chapter 1 to compare total order ranking methods. The total order results are shown in Table 2.23.

Sub.	Des.	*Uti.*	*Dom.*	*Pref.*	*ConcA*	*ConcQ*	*Abs R.*
CNB	0.96	0.96	0.67	0.76	1.00	0.93	1.00
4NA	0.59	0.64	0.34	0.41	0.74	0.39	0.59
4NP	0.00	0.66	0.28	0.42	0.67	0.51	0.50
ATR	0.72	0.78	0.36	0.61	0.74	0.65	0.67
CHL	0.00	0.33	0.09	0.28	0.22	0.22	0.26
DIA	0.00	0.74	0.54	0.59	0.67	0.65	0.50
DIM	0.63	0.69	0.36	0.46	0.74	0.49	0.62
ETH	0.00	0.69	0.26	0.47	0.67	0.56	0.50
GLY	0.47	0.52	0.20	0.33	0.30	0.30	0.48
ISO	0.66	0.71	0.33	0.50	0.74	0.53	0.64
MAL	0.00	0.64	0.54	0.61	0.67	0.47	0.47
THI	0.70	0.76	0.55	0.57	0.74	0.61	0.66

Table 2.23 – Rankings obtained by Desirability, Utility, Dominance, Preference functions, classical and quantitative Concordance Analysis, (ConcA and ConcQ), and Absolute reference method.

The obtained Hasse diagram is shown in Figure 2.23.

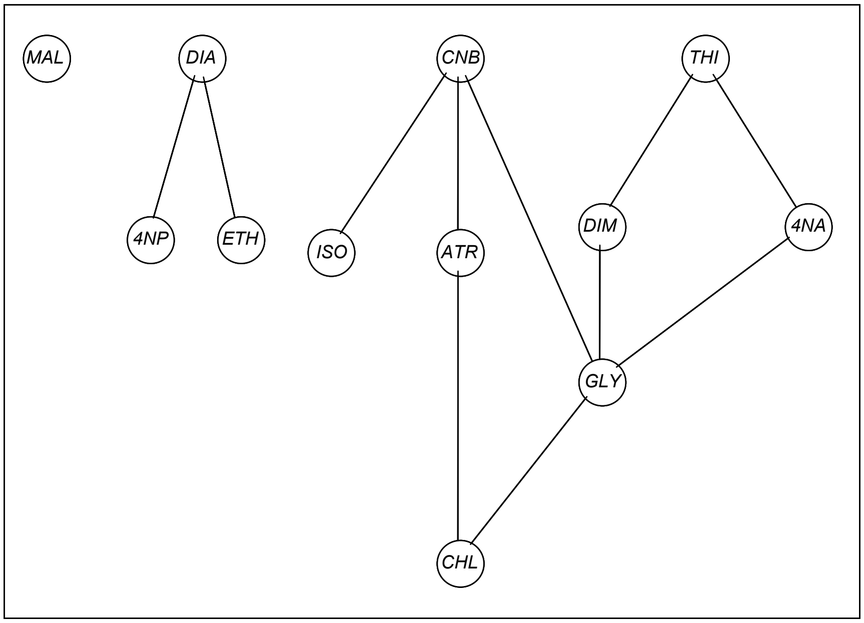


Figure 2.23 – Hasse diagram of twelve High Production Volume Chemicals.

The Hasse diagram is arranged in four levels; four elements are identified as maximals: Malathion, Linuron, 1-Chloro-4-Nitrobenzene and Thiram. The 12 HPVC are separated into two groups, plus an isolated element. The larger group is composed by eight elements, the smaller by three. The elements in the large group are characterised by high production volume and low LogKow values, whereas the elements in the small group have a reverse trend. The production volume and the partitioning coefficient between *n*-octanol and water are "antagonistic" attributes. Malathion is an isolated element because of its singular behaviour, which is the lowest value of LC50 and the fastest biodegradation.

Comparing the results from HDT with those from total ranking methods, it can be highlighted that, from among the total ranking methods, the Dominance approach is the one closest to HDT, being based on element pair comparisons. Moreover in the Hasse diagram technique the elements are divided, on the basis of levels, into groups. Unlike total ranking methods, which are based on a subjective definition of attribute weights, HDT allows the estimation of the quantitative importance of the attributes used to perform the ranking by sensitivity analysis. The example above is used for sensitivity analysis performed by the **W** matrix approach:

	A	*A₁*	*A₂*	*A₃*	*A₄*
A	0	9	2	9	22
A₁	9	0	11	18	31
A₂	2	11	0	11	24
A₃	9	18	11	0	31
A₄	22	31	24	31	0

Where A is the full set of attributes; A_1, A_2, A_3 and A_4 are the sets with PV, LC50, LogKow and BD skipped; respectively. The numbers indicate the marked importance of biodegradation in the ranking.

The same results were provided by backward sensitivity which, at the first step, selects LC50 as the least influential attribute, followed by PV selected at the second step. Thus, if the attribute weighting is assumed to be a level of subjectivity, the Hasse diagram technique can be considered a more scientifically based method than those of total ordering.

CHAPTER 3

Ranking Models

The previous chapters explained how to use total and partial ranking methods to perform data exploration, investigate the inter-relationships of objects and/or variables and set priorities. Moreover order ranking methods appear a very useful tool not only to perform data exploration but also to develop order ranking models, being a possible alternative to conventional statistical methods such as multi-linear regression (MLR) or classification. 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; moreover they are used for process control purposes as well as in chemical research.

A mathematical model is an important tool that allows the synthesis of knowledge on an investigated system and that can be used to perform predictions of future events based on the model itself. A *mathematical model* can be defined as a mathematical formulation of the relationship between a set of *dependent* (or *response*) variables ($\mathbf{y}_1$, …, $\mathbf{y}_R$) and a set of *independent* (or *explanatory*) variables ($\mathbf{x}_1$, …, $\mathbf{x}_p$), which allows the prediction of dependent variable values from the given values of independent variables:

$$(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_R) = f(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p)$$

Mathematical models can be divided into three main groups:

1. regression models: the dependent variable, i.e. the response, is a *quantitative* variable.
2. classification models: the dependent variable is a *qualitative* variable.
3. ranking models: the dependent variable is an *ordinal* variable.

Data modelling is usually performed with different purposes:

- Description: the aim is to find the best functional relationship among variables in the model. In mathematics, a function $y = f(x)$ is a rule of correspondence that associates a value of y to each x value.
- Inference: the aim is to generalise the results of a set of objects to the whole target system.
- Prediction: the aim is to predict response variables values from the model for new samples not yet experimentally investigated.

Thus, searching for a mathematical model is a complex procedure articulated in four main phases: The first consists in identifying the type of model, i.e. in the choice of the type of model that is supposed to be more appropriate for the system under study and for the objectives of the analysis. The second phase consists in developing the model by removing noise from the useful information and by the real model parameter estimation. The third phase is validation, which consists in checking stability and predictive model capability. Once reliability has been verified the fourth phase comes into effect, the model can be used to predict unknown events.

When data material is characterized by uncertainties, order models can be used as an alternative to statistical methods like multi-linear regression (MLR), since order models do not require a specific functional relationship between the independent and the dependent variables (responses). Moreover in several chemical and environmental problems the aim is to define order relations among several chemicals, to point out the more hazardous ones and to set priorities before taking final decisions. For these purposes, order ranking models, which allow the finding of inter-relationships for each chemical even though no quantitative values are provided, can be a promising approach to support decision making processes.

A ranking model is defined as a relationship between a set of dependent attributes, investigated experimentally, and a set of theoretically defined independent attributes, also called model attributes, which are calculated attributes:

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

where *f* is a ranking function.A model ranking development is based on the following steps:

1. *Experimental ranking:* a total or partial ranking method is applied to experimental attributes (dependent attributes).
2. *Model ranking*: the total or 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. *Interval estimations*: experimental ranking of an element is evaluated by the ranking model obtained.

In the first phase, elements are ranked according to the experimental attributes describing them. If the aim is to develop a total ranking model, a total ranking method is selected and applied to the experimental attributes. In this case the result will be a totally ordered element sequence. On the other hand, if the aim is to develop a partial ranking model, the Hasse diagram technique will be applied and the result will be a Hasse diagram of partially ordered elements. In the second phase the same ranking method (total or partial), previously applied to the experimental attributes, is now applied to a selected subset of model attributes, and the elements are ranked according to the selected model attributes. Then, the two rankings 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 totally ordered sequences, or two diagrams, 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, are performed by the ranking model.

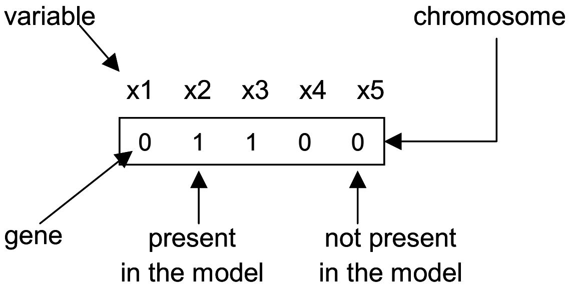
As in multilinear regression (MLR) methods, the selection of variables (attributes) is crucial to developing a reasonable ranking model. The aim of variable subset selection is to reach optimal model complexity in predicting response variables by a reduced set of independent variables [Hocking, 1976; Miller, 1990]. Ranking models based on the optimal subsets of a few predictor attributes have the great advantage of being more stable and showing higher predictive power. One of the simplest techniques for variable selection, also called "sentimental selection", is based on the *a priori* selection of a few variables, by experience, tradition, availability, opportunity or knowledge. Ranking models, still proposed in today's literature, are mostly based on a subjective selection of attributes which are supposed to reflect the chemical and physical features to be modelled [Carlsen *et al*., 1999; Carlsen, 2001; Walker and Carlsen, 2002; Carlsen *et al*., 2002*a*; Sørensen *et al*., 2003]. 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. As the procedure consists in evaluating the quality of all the models with one variable, all the models with two variables to all possible models with *k* variables, the required computer time increases extraordinary when *p* and *k* are quite large. In fact, the total number of models *t* is given by the following expression:

$$t = \sum_{k=1}^{L} \frac{p!}{k! \cdot (p-k)!} \leq 2^p - 1$$

where *L* is the maximum user-allowed model size. The main advantage of this method is the exhaustive search for the best ranking model in the model space. However in many multivariate applications elements are described by too many variables. This happens either because the researcher does not know *a priori* which variables are relevant to the study on hand (and as many as possible are measured to make sure all the important ones are included), or because the experiment is very costly and difficult to organize (and the researcher measures as many variables as possible in case the opportunity to repeat the experiments does not present itself again), or, finally, because the variables are calculated variables. When many variables are available, relevant information should be separated from redundant and noisy information. Moreover, when many variables are available, an exhaustive examination of all possible models is not feasible as, given the extremely high number of possible attribute combinations, it requires extensive computational resources and is time consuming. In such cases a variable selection technique is needed. The Genetic Algorithm (GA-VSS) approach is proposed here as the variable selection method to search for the best ranking models within a wide set of variables.

3.1 GA-VSS applied to ranking models

Genetic algorithms (GA) are an evolutionary method used widely for complex optimisation problems in several fields such as robotics, chemistry and QSAR [Goldberg, 1989; Wehrens and Buydens, 1998]. A specific application of GA is variable subset selection (GA-VSS) [Leardi, *et al.*, 1992; Leardi, 1994; Luke, 1994; Leardi, 1996; Todeschini *et al.*, 2003]. 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. 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 based on the evolution of a *population* of models, i.e. a set of ranked models according to some objective function. In genetic algorithm terminology each individual population is called *chromosome* and is a *p*-dimensional binary vector **I**, where each position (a *gene*) corresponds to a variable (1 if included in the model, 0 otherwise).



Each chromosome represents a model given by a subset of variables. The objective function to be optimised must be defined along with the model population size P and the maximum number L of allowed variables in a model; the minimum number of allowed variables is usually assumed equal to one. Moreover, *crossover probability* and *mutation probability* are to be defined. The genetic algorithm procedure is illustrated in Figure 3.1.

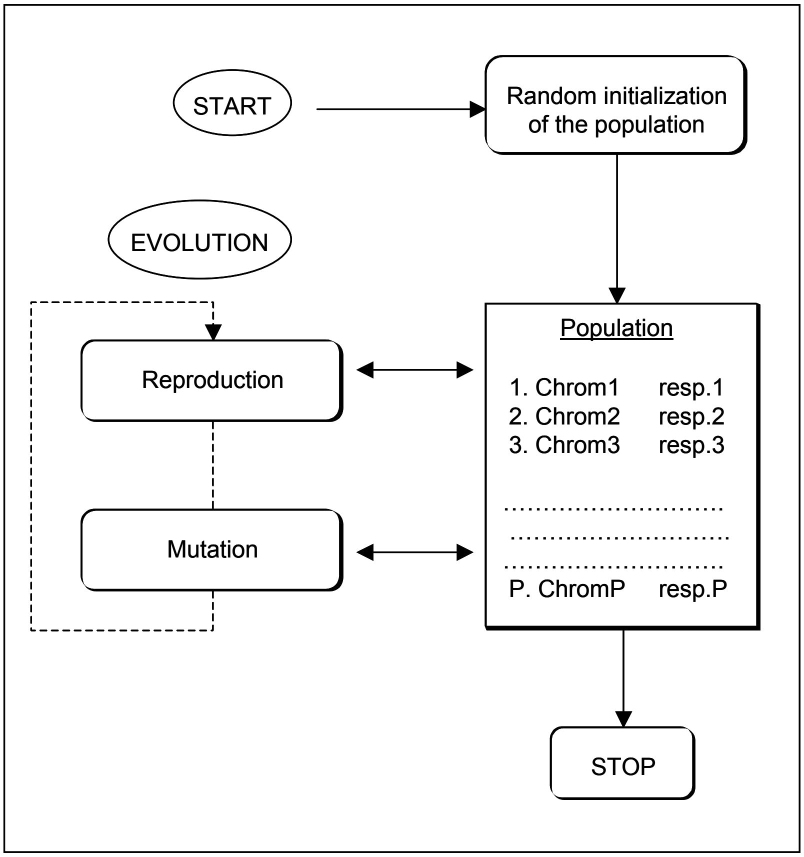


Figure 3.1 – Genetic algorithm procedure.

Once the leading parameters are defined the genetic algorithm evolution starts, based on three main steps:

1. Random initialisation of the population

The model population is built initially by random models with a number of variables between 1 and L. The value of the selected objective function of each model is calculated in a process called *evaluation*. The models are then ordered with respect to the selected objective function – model quality - (the best model is placed first in the population, the worst at position P).

2. Crossover

From the actual population, pairs of models are selected to be used as parents. Parent selection can be performed randomly, if no account is taken of quality, or by the so-called *roulette wheel* (RW) which is biased towards the best individuals, the chance of an individual being selected being a function of its quality (or rank). In this case the concept of quality survival comes into play by applying *selection pressure*. Additional pressure can be introduced by using the roulette wheel operator several times to produce a *tournament selection* of a subset of individuals: the best individual is then chosen as the selected parent. Then, from each pair of selected models (*parents*), new individuals (*offspring*) are generated, preserving the common characteristics of the parents (i.e. variables excluded in both models remain excluded, variables included in both models remain included) and mixing the opposite characteristics according to the crossover probability. Let Parent 1 and Parent 2 be the selected parents:

Parent 1: 0 1 **0** 0 **1** 1 0 0
Parent 2: 0 1 **1** 0 **0** 1 0 0

Each offspring derived from these two parents will preserve their common genetic part, being a chromosome like 0 1 **?** 0 **?** 1 0 0.

Offspring generation is performed using one parent at a time and analysing its changeable genes by comparing a random number with the crossover probability (*unbiased uniform crossover*). For each variable included in one parent but not in the other, a number is extracted randomly and compared with the crossover probability: if this randomly selected number is lower than the crossover probability then the variable is included if not present in the parent (0 → 1), or excluded if present in the parent (1 → 0), otherwise it remains unchanged. If the generated son coincides with one of the individuals already present in the actual population, it is rejected; otherwise, it is evaluated. If the objective function value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank;

otherwise, it is no longer considered. This iterative procedure is repeated for several pairs.

3. Mutation

After a number of crossover iterations, the population proceeds through the mutation process. Mutation is a mechanism that produces, by a completely random process, new genetic material during population evolution. For each individual present in the population, *p* random numbers are tried, *p* being the number of individual genes, and each is compared, one at a time, with the defined mutation probability: each gene remains unchanged if the corresponding random number exceeds the mutation probability, otherwise, it is changed from zero to one or *viceversa*. Low values of mutation probability allow only a few mutations, resulting in new chromosomes not too different from the generating chromosomes.

4. Stop conditions

The second and third steps are repeated until some stop condition is encountered (e.g., a user-defined maximum number of iterations) or the process is arbitrarily ended. The models based on the selected subset of attributes are tested and evaluated by optimisation parameters, i.e. indices that quantify the agreement of the model ranking with the experimental ranking.

It is to be highlighted that the GA-VSS method provides not a single model but a population of acceptable models; this characteristic, sometimes considered a disadvantage, makes the evaluation of variable relationships with response from different points of view possible. 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 [Todeschini *et al.*, 2003].

3.2 Optimisation parameters

Variable subset selection is performed by GAs optimising populations of models according to a defined *objective function* related to model quality. In 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. To measure the agreement of two rankings they have to be compared: in the case of a total ranking model two totally ordered element sequences are brought into comparison, while in a partial ranking model the comparison is performed between two Hasse diagrams. The total and partial ranking models are evaluated by different objective functions.

3.2.1 Total ranking optimisation parameters

To develop a total ranking model, a total order ranking method is first applied to the experimental attributes y_1, …, y_R, defining an experimental ranking parameter, Γ_{exp}. According to the experimental parameter, a specific experimental rank is associated to each *i-th* element:

$$\Gamma_i^{\text{exp}} \equiv f(y_{i1}, y_{i2}, \ldots y_{iR}) \quad \rightarrow \quad rank_i^{\text{exp}}$$

Then the total order ranking method is applied to the model attributes x_1, …, x_p, defining a model ranking parameter, Γ_{mod} and *a*ccording to that, a model rank is associated to each *i-th* element:

$$\Gamma_i^{\text{mod}} \equiv f(x_{i1}, x_{i2}, \ldots x_{ip}) \quad \rightarrow \quad rank_i^{\text{mod}}$$

The correlation between the two ranking parameters (Γ_{exp}, Γ_{mod}) can then be evaluated by Spearman's rank correlation coefficient, according to the following expression:

$$r_{exp-mod} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} \qquad -1 \le r_{exp-mod} \le +1$$

where d_i is the rank difference for the element *i* in the two rankings and *N* is the total number of elements.

The correlation between experimental and model rankings can also be evaluated by Kendall's rank correlation coefficient τ. These indices can be used as optimisation parameters in a genetic evolution algorithm to quantify the correlation, i.e. the similarity, between the total experimental ranking, for example a *desirability function* D_{exp} and the total model ranking, a *desirability function* D_{mod} .

Figure 3.2 shows the procedure used to compare the total experimental ranking and the total model ranking.

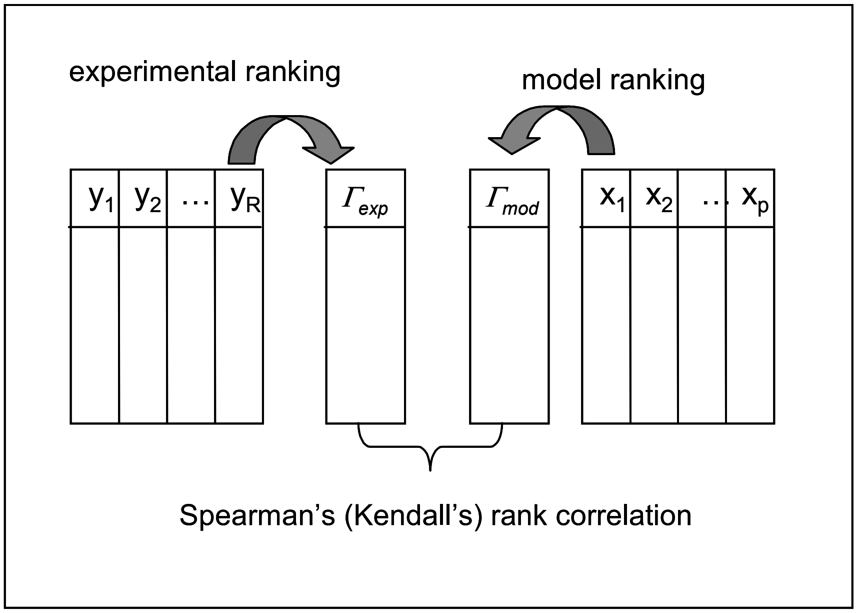


Figure 3.2 – Scheme of the procedure used for the comparison of the total experimental ranking with the total model ranking.

3.2.2 Partial ranking optimisation parameters

According to the procedure described above, once the experimental ranking has been developed, for each subset of selected model attributes the agreement degree between the two corresponding diagrams is evaluated. 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 [Rogers and Tanimoto 1960; Brüggemann *et al*., 1995b; Bath *et al*., 1993; Moock *et al*., 1998; Sørensen *et al*., 2003] $T(I_E, I_M)$ defined as:

$$T(I_E, I_M) = \frac{E \cap M}{E \cup M} = \frac{\sum sr}{\sum sr + \sum rr + I_E \sum ir + I_M \sum ir}$$

where I_E and I_M are weights which can take the value either 0 or 1, and the denominator terms are defined as follows:

$\sum sr$: sum of element pairs with $y_{sj} < y_{tj}\ \forall j \in E$ and $x_{sj} < x_{tj}\ \forall j \in M$ "same ranking"

$\sum rr$: sum of element pairs with $y_{sj} < y_{tj}\ \forall j \in E$ and $x_{sj} > x_{tj}\ \forall j \in M$ "reverse ranking"

$\sum_E ir$: sum of elements pairs with $y_{sj} \le y_{tj}\ \forall j \in E$ and $x_{sj} \| x_{tj}\ \forall j \in M$ "incomplete ranking in *E*"

$\sum_M ir$: sum of pairs with $y_{sj} \| y_{tj}\ \forall j \in E$ and $x_{sj} \le x_{tj}\ \forall j \in M$ "incomplete ranking in *M*"

where $s, t \in 1, 2, ..., N$ and $s \ne t$.

Depending on the specific correlation problem it is not always relevant to use $\sum_E ir$ and/or $\sum_E ir$ in Tanimoto index calculation.

Thus, 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, i.e. the number of the same mutual rank of two elements identified in the model and in the experimental ranking, over the number of disagreements, i.e. contradictions in the ranking of two elements in the model and experimental ranking. The most severe similarity measure is I_E = 1 and I_M =1.

The Tanimoto indices, being similarity indices, range from 0, when no similarity exists between the two rankings and, thus, no element has the same mutual rank, to 1, when the two rankings are totally similar, i.e. all the elements have the same mutual rank and no contradiction exists.

A numerical example of the Tanimoto indices calculation is illustrated here, see the two diagrams shown in Figure 3.3.

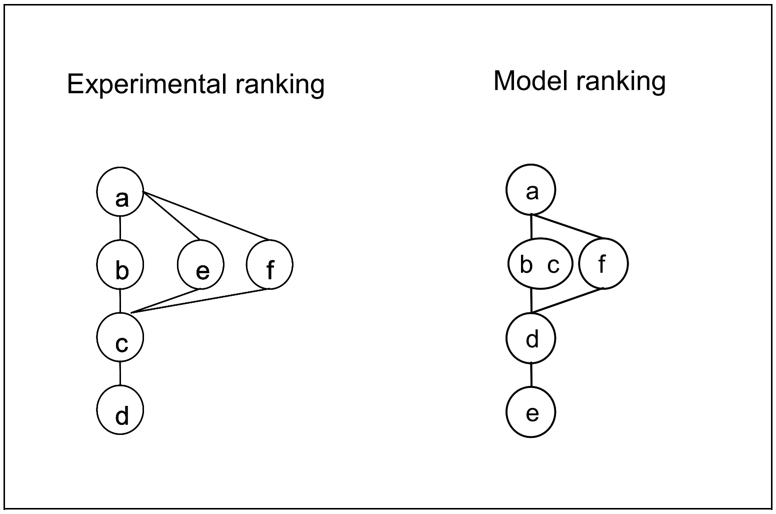


Figure 3.3 – Experimental and model ranking comparison.

Number of reverse rankings ($\sum rr$) = 2 {CE, DE}

Number of same rankings ($\sum sr$) = 9 {AB, AC, AD, AE, AF, BC, BD, CD, DF}

Number of incomplete rankings in E ($\sum_E ir$) = 1 {CF}

Number of incomplete rankings in M ($\sum_M ir$) = 2 {BE, EF}

Thus, the Tanimoto indices are:

$$T(0,0) = \frac{E \cap M}{E \cup M} = \frac{\sum sr}{\sum sr + \sum rr} = \frac{9}{9+2} = 0.82$$

$$T(0,1) = \frac{E \cap M}{E \cup M} = \frac{\sum sr}{\sum sr + \sum rr + \sum ir} = \frac{9}{9+2+2} = 0.69$$

$$T(1,1) = \frac{E \cap M}{E \cup M} = \frac{\sum sr}{\sum sr + \sum rr + \sum ir + \sum ir} = \frac{9}{9+2+2+1} = 0.64$$

In searching for ranking models, optimising the Tanimoto index T(0,0) has been demonstrated to be overoptimistic and not able to give optimal models. In fact, the selected models often turned out to be quite unlike the experimental model as this index in the model discloses only the percentage of rankings, which can be again found in the experimental ranking. The Tanimoto indices T(0,1) are more severe than T(0,0), as the denominator even takes into account the dissimilarity due to element pairs that, being incomparable in the experimental ranking, become comparable in the model ranking. As far as concerns the Tanimoto indices T(1,1), it is to be highlighted that this is the most severe case as the denominator accounts for both dissimilarities; this is due to element pairs that, being incomparable in the experimental ranking, become comparable in the model ranking and *viceversa*. However, as the numerator factor only accounts for rankings, i.e. comparabilities, but ignores similarity due to incomparabilities, it turns out to be over pessimistic.

To have a more realistic measure of the agreement between two partial rankings a similarity index is proposed. 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

$$s,t \in 1,2,\ldots,N \quad \text{and} \quad s \neq t \quad \text{and} \quad h_{st}\begin{cases} +1 & \text{if} \quad y_r(s) \ge y_r(t) \quad \text{for all } y_r \in IB \\ -1 & \text{if} \quad y_r(s) < y_r(t) \quad \text{for all } y_r \in IB \\ 0 & \text{otherwise} \end{cases}$$

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 ($h_{st}^{\mathbf{E}} = \pm 1$ and $h_{st}^{\mathbf{M}} = 0$, or $h_{st}^{\mathbf{E}} = 0$ and $h_{st}^{\mathbf{M}} = \pm 1$), then their contribution is 1, while if the mutual ranks are opposite ($h_{st}^{\mathbf{E}} = +1$ and $h_{st}^{\mathbf{M}} = -1$, or $h_{st}^{\mathbf{E}} = -1$ and $h_{st}^{\mathbf{M}} = +1$), 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 *viceversa*.

A numerical example of the similarity index calculation is provided for the two diagrams shown in Figure 3.4.

To allow easy calculation of the similarity index, the experimental and model Hasse matrices are displayed here: the discrepancies are highlighted in bold. The order relations between element b and f is opposite in the two rankings, as *b* covers *f* in the experimental ranking, while it is covered in the model ranking. Moreover the model is not able to reproduce the order relations between *c* and *e* and *d* and *e*; they are incomparable elements, as their order relations have not been solved.

Experimental ranking

	a	*b*	*c*	*d*	*e*	*f*
a	-	1	1	1	1	1
b	-1	-	1	1	0	**1**
c	-1	-1	-	1	**-1**	-1
d	-1	-1	-1	-	**-1**	-1
e	-1	0	**1**	**1**	-	0
f	-1	**-1**	1	1	0	-

Model ranking

	a	*b*	*c*	*d*	*e*	*f*
a	-	1	1	1	1	1
b	-1	-	1	1	0	**-1**
c	-1	-1	-	1	**0**	-1
d	-1	-1	-1	-	**0**	-1
e	-1	0	**0**	**0**	-	0
f	-1	**1**	1	1	0	-

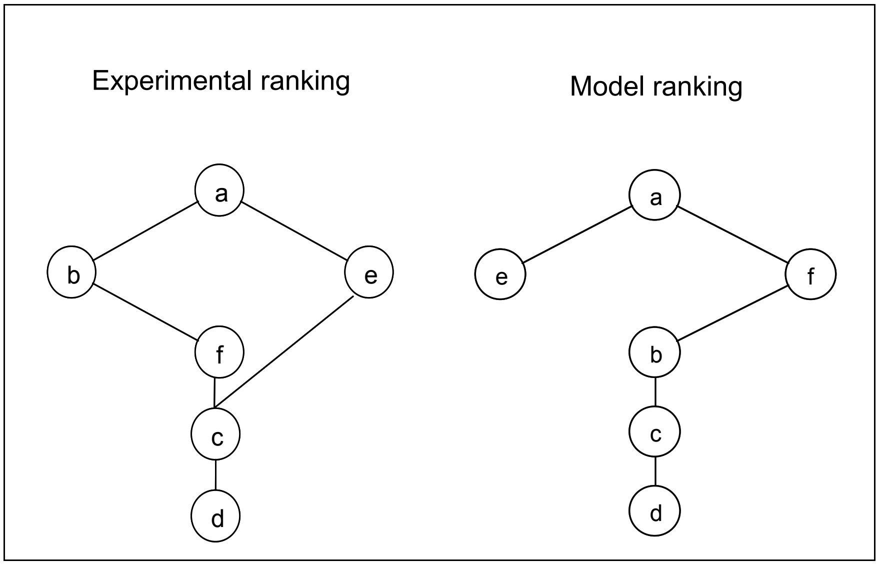


Figure 3.4 – Experimental and model ranking comparison.

The similarity index is:

$$S(\mathbf{E},\mathbf{M}) = 1 - \frac{2+1+1+1+1+2}{2\cdot 6\cdot 5} = 1 - \frac{8}{60} = 0.87$$

The Tanimoto indices for the example above take the following values:

$$T(0,0)=\frac{\sum sr}{\sum sr+\sum rr}=\frac{10}{10+1}=0.91$$

$$T(0,1)=\frac{\sum sr}{\sum sr+\sum rr+\sum ir}=\frac{10}{10+1+0}=0.91$$

$$T(1,1)=\frac{\sum sr}{\sum sr+\sum rr+\sum ir+\sum ir}=\frac{10}{10+1+0+2}=0.77$$

Thus, the “goodness of fitting” of the partial ranking model of Figure 3.4 calculated by the similarity index is lower than that calculated by both T(0,0) but higher than the one by T(1,1), suggesting that the similarity index could be a more reasonable compromise between the over optimistic over pessimistic evaluation provided by T(0,0) and T(1,1) respectively, then the T(0,1) index, which not necessary provides a diverse result with respect to T(0,0).

3.3 Ranking predictions

Once the “goodness of fitting” of the model ranking has been verified by either Spearman’s rank index for total ranking or the similarity index (or Tanimoto indices) for partial ranking, predictions can be performed for new elements. The experimental ranking of new compounds that have not yet been investigated experimentally can be predicted by the ranking model. Since total ranking can be assumed as a particular case of partial ranking, the proposed prediction procedure is here described for a more general partial ranking.

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:

$$\left\{x_{u1},\ldots,x_{up}\right\}\xrightarrow{\text{training set}} rank_{ur}$$

To explain ranking predictions, a directed connectivity operator *C* is introduced. This operator is defined for element pairs of digraphs (directed graphs) to which Hasse belongs. Interpreting the Hasse diagram as a digraph, it consists of a set of *N* vertices (circles representing elements) and a set of oriented edges each connecting two vertices. No path exists between incomparable elements.
Being *s* and *t* two diverse elements in a Hasse diagram (HD), and N the set of integer numbers, then the connectivity operator *C*(*s*,*t*) is defined as follows:

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

$$C(s,t) \in N^{+} \quad \text{iff} \quad \text{s covers t}$$

$$C(s,t) \in N^{-} \quad \text{iff} \quad \text{t covers s}$$

$$C(s,t) = 0 \quad \text{iff} \quad \text{s} \parallel \text{t}$$

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

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

k being the absolute value of the path length between the two elements *s* and *t* , and *L* the number of levels in the Hasse diagram.
According to the first property, the operator is an integer number, taking a value equal to the path length between *s* to *t*. If *s* covers *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 ($s \parallel t$), then *C*(*s*,*t*) equals 0. Reflecting the ranking order relation properties, the connectivity operator has antisymmetry and transitivity properties.
As the connectivity operator is a key element to performing ranking prediction, an example is illustrated in Figure 3.5.

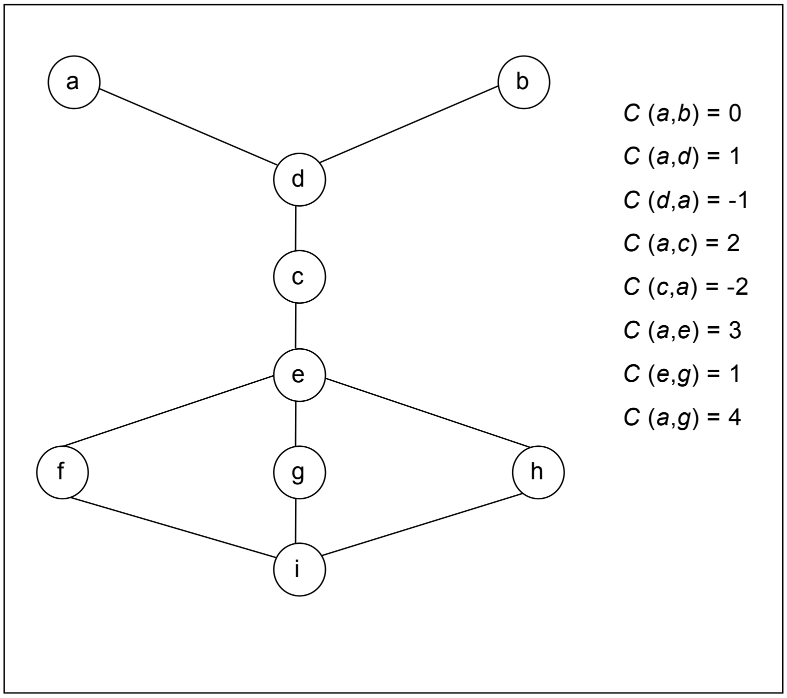


Figure 3.5 – Connectivity operator values for Hasse diagram.

The connectivity operator $C(a,b)$ on elements a and b of the Hasse diagram in Figure 3.5 takes a value equal to 0, as the elements are not connected: no line exists between them as they are incomparable. For the same reason $C(f,g) = 0$, $C(f,h) = 0$, $C(g,h) = 0$. Elements a and d are comparable, and thus connected by a path with an absolute value of length equal to 1. Moreover, according to the antisymmetry property $C(a,d) = 1$, as a covers d, whereas $C(d,a) = -1$, as d is covered by a. Elements a and e are connected by a path length equal to 3. According to the transitivity property, as $C(a,e) = 3$ and $C(e,g) = 1$, $C(a,g) = 4$. It can be observed that for the Hasse diagram of Figure 3.5 the maximum value of the path length is 5 (= L - 1).

Moreover, for totally ordered ranking the connectivity operator does not take a value equal to 0, as all the objects are comparable and no incomparability exists among them.

Thus, through the connectivity operator, the 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:

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

Thus, for any unknown element *u*, a search is performed for the two elements *s* and *t* which are 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 which 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.*

Figure 3.6 summarizes the procedure used to perform partial ranking predictions. In a first step, a partial order ranking method like the Hasse diagram is applied on the experimental attributes y_1, …, y_R of the *N* training set elements; in the second step the Hasse diagram is developed on the selected model attributes x_1, …, x_p. Once agreement between the two rankings has been verified, the model is used to perform predictions of the experimental ranking of a new element *u*, not yet tested experimentally.

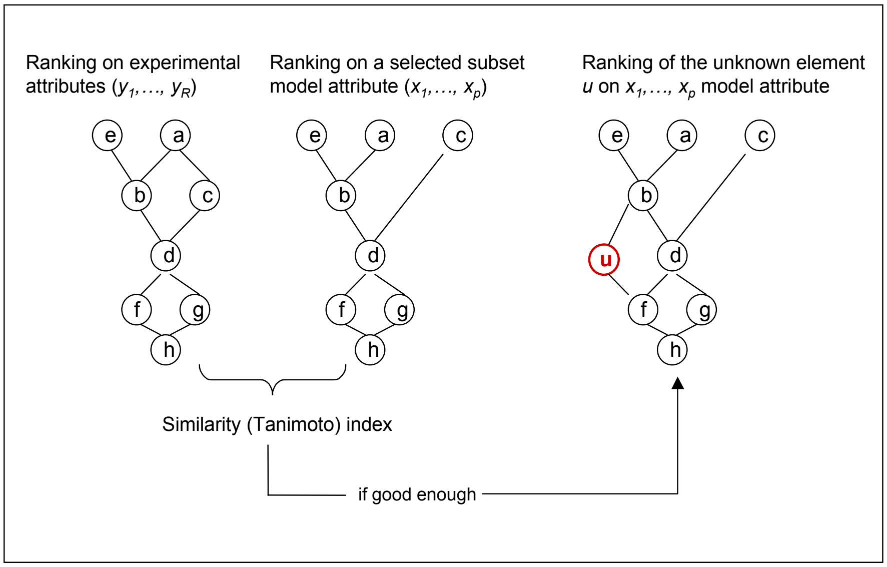

Ranking on experimental attributes (y_1,…, y_R)
Ranking on a selected subset model attribute (x_1,…, x_p)
Ranking of the unknown element u on x_1,…, x_p model attribute
e
a
b
c
d
f
g
h
u
Similarity (Tanimoto) index
if good enough

Figure 3.6 – Scheme of the procedure used for partial ranking predictions.

A numerical example is provided here to better explain the prediction calculation. For the sake of simplicity, let us consider an experimental ranking developed on two experimental attributes y_1 and y_2; Table 3.1 shows their numerical values. The corresponding experimental Hasse diagram is shown in Figure 3.7

Element	y_1	y_2
a	180	400
b	150	420
c	130	240
d	140	270
e	90	190
f	100	230
g	120	200
h	90	235
i	82	88

Table3.1 – Numerical values of the experimental attributes.

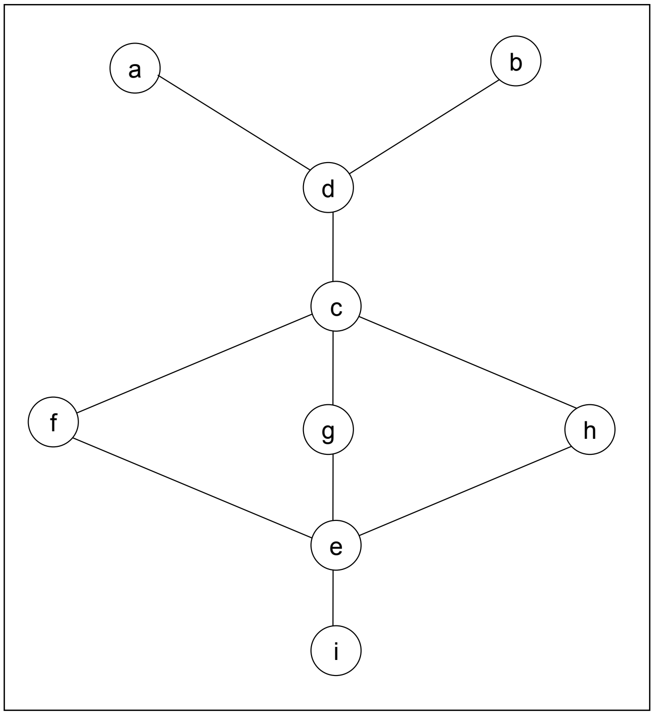


Figure 3.7 –Experimental Hasse diagram.

Let us suppose the Hasse diagram in Figure 3.7 to be the ranking model developed on the training set composed by 9 elements {*a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*}; having verified the model's agreement with the experimental ranking (S(**E**,**M**) = 0.92; T(0,0) = T(0,1) = T(1,1) = 0.91) the set of model attributes $\{x_{u1},\ldots, x_{up}\}$ describing the unknown element *u* can, on the basis of the training set elements, perform the prediction of the experimental ranking of *u*:

$$\left\{x_{u1},\ldots,x_{up}\right\} \xrightarrow{TS} rank_u$$

Figure 3.8 shows the model ranking projection of the new unknown element *u* in the model ranking diagram.

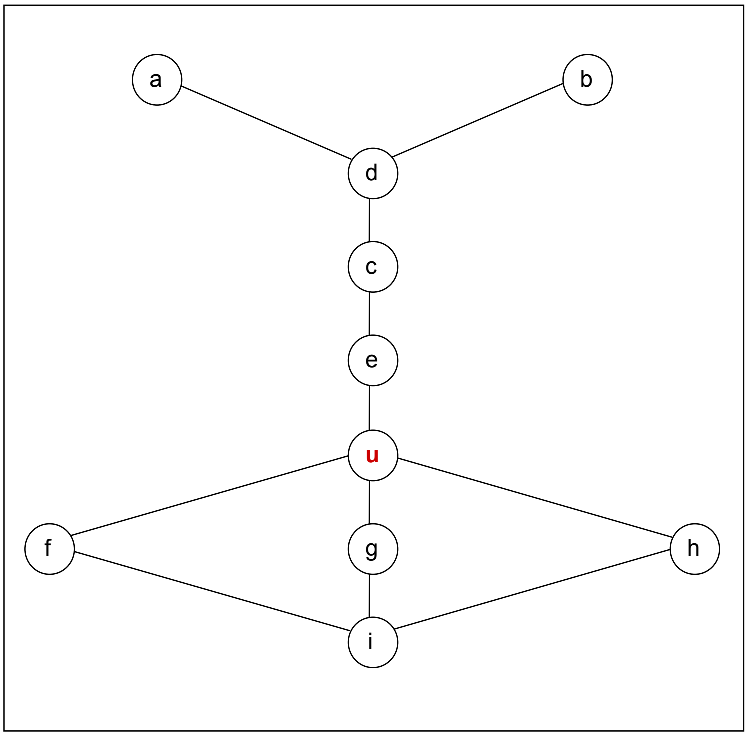


Figure 3.8 –Projection of the unknown element *u* in the model ranking diagram.

To predict the experimental values of the unknown element *u*, a search is made for the element pair located on the shortest path from *u* and with an experimental value difference that constitutes the smallest positive interval.

Firstly, an examination is made of the elements comparable to *u* and located on a path length equal to 1. The experimental values y_1 and y_2 of elements *e*, *f*, *g* and *h* are taken into account and the differences between *e* (located above *u*) and *f*, *g* and *h* (located below *u*) are investigated.

As far as concerns the experimental attribute y_1, the model provides the following intervals:

$$y_{e1} - y_{f1} = 90 - 100 = -10$$

$y_{e1} - y_{g1} = 90 - 120 = -30$
$y_{e1} - y_{h1} = 90 - 90 = 0$

All three intervals are rejected as they are not positive interval. Thus the elements located on a path length equal to 2 are examined. The elements *c* and *i* are considered and the following intervals examined:

$y_{c1} - y_{f1} = 130 - 100 = 30$

$y_{c1} - y_{g1} = 130 - 120 = 10$

$y_{c1} - y_{h1} = 130 - 90 = 0$

$y_{e1} - y_{i1} = 90 - 82 = 8$

$y_{c1} - y_{i1} = 130 - 82 = 48$

The smallest positive interval for y_1 is the one provided by elements *e* and *i*, thus the experimental value y_1 of the unknown element *u* is predicted as:

$$y_{i1} \leq y_{u1} \leq y_{e1} \quad \Rightarrow \quad 82 \leq y_{u1} \leq 90$$

It can be observed that element *e* and *i* satisfy all the conditions required to perform a ranking prediction, i.e.:

$$C(e,u) = 1\ (> 0) \quad \wedge \quad C(u,i) = 2\ (> 0) \quad \wedge \quad min[(y_e - y_i)] = 8\ (> 0)$$

In the same way the following intervals are provided for y_2:

$y_{e2} - y_{f2} = 190 - 230 = -40$
$y_{e2} - y_{g2} = 190 - 200 = -10$
$y_{e2} - y_{h2} = 190 - 135 = \text{-}45$

All the three intervals are rejected, and again the intervals provided by elements *c* and *i* located at a length path equal to 2 are examined:

$y_{c2} - y_{f2} = 240 - 130 = 10$
$y_{c2} - y_{g2} = 240 - 200 = 40$
$y_{c2} - y_{h2} = 240 - 235 = 5$

$y_{e2} - y_{i2} = 190 - 88 = 102$

$y_{c2} - y_{2i} = 240 - 88 = 152$

The smallest positive interval for y_2 is the one provided by elements *c* and *h*, thus the experimental value y_2 of the unknown element *u* is predicted as:

$$y_{h2} \le y_{u2} \le y_{c2} \quad \Rightarrow \quad 235 \le y_{u2} \le 240$$

Elements *c* and *h* satisfy all the conditions required to perform a ranking prediction, i.e.:

$$C(c,u) = 2\ (> 0) \quad \wedge \quad C(u,h) = 1\ (> 0) \quad \wedge \quad min[(y_c - y_h)] = 5\ (> 0)$$

According to the position of the unknown element *u* in the model ranking, four different cases can be identified, each characterized by specific prediction:

1. *u* is located in a chain → $y_{tr} \le y_{ur} \le y_{sr}$
2. *u* is a minimal → $y_{ur} \le y_{sr}$
3. *u* is a maximal → $y_{tr} \le y_{ur}$
4. *u* is isolated → $y_{ur} = ?$

s and *t* being two elements in HD respectively located above and below *u*. In particular, for case 2, being *u* a minimal, its rank is predicted to be smaller than the lowest value of the comparable elements ranked above; thus, the rule is the following:

$$C(s,u) = 1 \quad \wedge \quad min_s(y_s)$$

which means that the estimated interval of *u* is open on the left and only the first shell of neighbourhoods above is taken into account. If, for the example above, the unknown element *u* is a minimal (Figure 3.9), then its predicted ranks will be:

$$y_{u1} \leq y_{i1} \quad \Rightarrow \quad y_{u1} \leq 82$$

$$y_{u2} \leq y_{i2} \quad \Rightarrow \quad y_{u2} \leq 88$$

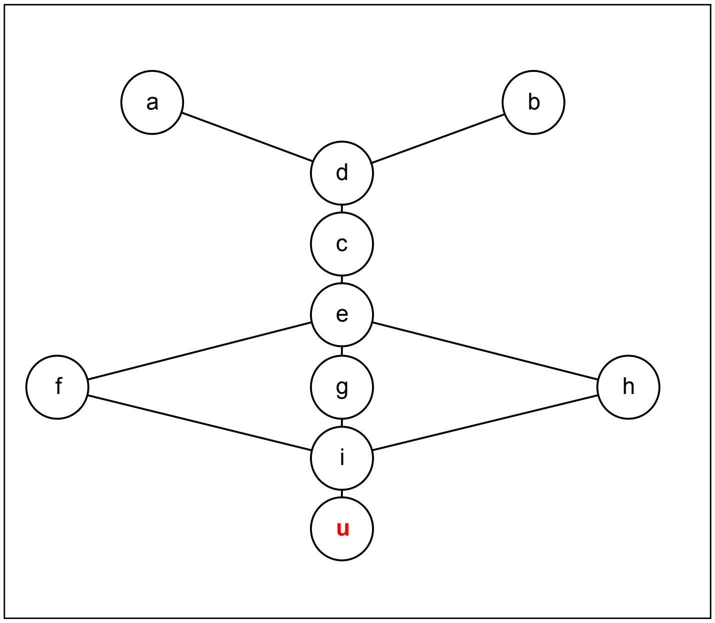


Figure 3.9 – Projection of the unknown element *u* in the model ranking diagram.

Moreover, for case 3 where *u* is a maximal, there is no comparable element above, and its rank is predicted to be larger than the highest value of the comparable elements ranked below; thus, the rule is:

$$C(u,t) = 1 \quad \wedge \quad max_t(y_t)$$

which means that the estimated interval of *u* is open on the right and only the first shell of neighbourhoods below is taken into account.

For the example described above, if the unknown element *u* is a maximal (Figure 3.10) then its predicted ranks will be:

$$y_{a1} \leq y_{u1} \quad \Rightarrow \quad 180 \leq y_{u1}$$

$$y_{b2} \leq y_{u2} \quad \Rightarrow \quad 420 \leq y_{u2}$$

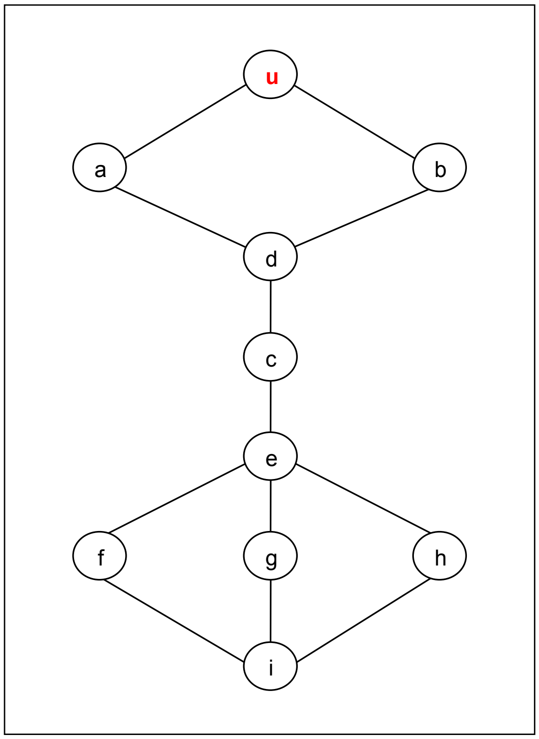


Figure 3.10 – Projection of the unknown element *u* in the model ranking diagram.

In the last case *u* is an isolated element, i.e. it is not comparable with any of the elements of the training set, thus its rank cannot be predicted by the model ranking developed.

As mentioned above, total ranking can be assumed to be a particular case of partial ranking. Thus the procedure described for partial ranking predictions can be used in the same way for total ranking predictions. A numerical example is provided here. Let us consider the experimental ranking developed on the two experimental attributes y_1 and y_2, of Table 3.1. Assuming the two attributes of equal importance, the experimental total ranking provided by the desirability function is illustrated in Table 3.2. Elements are sorted according to decreasing desirability.

Element	D_{exp}
a	0.969
b	0.833
d	0.570
c	0.474
g	0.362
f	0.280
h	0.190
e	0.158
i	0.000

Table 3.2 – Numerical values of the total experimental ranking.

Suppose that the Desirability method applied on a selected subset of model attributes provides the ranking shown in Table3.3. Having verified the good agreement between the experimental and model rankings ($r_{exp-mod} = 0.90$), the unknown element *u* is projected in the model ranking and, according to its model attributes, is ranked below *e* and above *g*.

Element	D_{mod}
a	0.969
b	0.833
d	0.570
c	0.474
e	0.441
u	0.436
g	0.362
f	0.280
h	0.190
i	0.000

Table 3.3 – Numerical values of the model experimental ranking.

Firstly elements comparable to *u* and located on a path length equal to 1 are examined. The experimental desirability values of elements *e* and *g* are taken into account and the differences between *e* (located above *u*) and *g* (located below *u*) are investigated.
The interval provided by the model is:

$$D_e^{exp} - D_g^{exp} = 0.158 - 0.362 = -0.204$$

This interval is rejected, as it is not a positive interval. Thus the elements located on a path length equal to 2 are examined. Elements *c* and *f* are considered and the following intervals examined:

$$D_c^{exp} - D_g^{exp} = 0.474 - 0.362 = 0.112$$
$$D_e^{exp} - D_f^{exp} = 0.158 - 0.280 = -0.122$$
$$D_c^{exp} - D_f^{exp} = 0.474 - 0.280 = 0.194$$

The second interval is rejected again as not being positive. The smallest positive interval is the one provided by elements *c* and *g*, thus the experimental desirability value of the unknown element *u* is predicted as:

$$D_g^{exp} \leq D_u^{exp} \leq D_c^{exp} \quad \Rightarrow \quad 0.362 \leq D_u^{exp} \leq 0.474$$

It can be seen that elements *c* and *g* satisfy all the conditions required to perform a ranking prediction, i.e.:

$$C(c,u) = 2 \ (>0) \quad \wedge \quad C(u,g) = 1 \ (>0) \quad \wedge \quad min[(D_c^{exp} - D_g^{exp})] = 0.112 \ (>0)$$

3.3.1 Ranking predictions by arithmetic mean values

According to the approach proposed in the literature [Carlsen *et. al*, 1999, 2001, 2002a; 2002b; Walker and Carlsen 2002], the prediction of experimental values of new elements, not yet investigated experimentally, can be derived as simple arithmetic means between the elements ranked above and below. This ranking prediction procedure is summarised briefly in the following. Partial order models predict only for elements within the “ordering net”, those elements respectively located in the top (maximals) and bottom (minimals) layers cannot be predicted. The predicted values for a given element *u* (Value *u*) are obtained by simple arithmetic means between the lowest value of the comparable elements ordered above *u* (minAbove) and the highest value of the comparable elements ordered below *u* (maxBelow).

$$Value\ u = \frac{(\min Above + \max Below)}{2}$$

The uncertainty of the value u is equally distributed in the interval
$\pm 0.5|\min Above - \max Below|$, thus the predicted value u is calculated as:

$$Value\ u = \frac{(\min Above + \max Below)}{2} \pm 0.5|\min Above - \max Below|$$

The approach has two main drawbacks: the first is that it is based on strong extrapolation, as the ranking model is used to derive not a ranking but a quantitative value. The second is that no check is performed on the interval consistency of the two elements selected as minAbove and maxBelow. To better explain this latter point, the prediction procedure is applied to the numerical partial ranking example discussed above (Table 3.1, Figures 3.7 and 3.8). Thus the predicted experimental value y_1 for the unknown element *u* would be:

$$y_{u\,1} = \frac{(y_{e1} + y_{g1})}{2} \pm 0.5\left|y_{e1} - y_{g1}\right| = \frac{(90 + 120)}{2} \pm 0.5\left|90 - 120\right| = 105 \pm 15$$

whereas the experimental range values of y_1, predicted by the approach here proposed was:

$$y_{i1} \le y_{u1} \le y_{e1} \quad \Rightarrow \quad 82 \le y_{u1} \le 90$$

In this case the elements *e* and *g* were selected to predict the experimental value of the unknown element *u*, but experimental values do not provide a real interval and the procedure does not take this into account, meaning that the *e* and *g* ranking established by the model was inverted with respect to the experimental ranking. In fact, according to the experimental ranking, *g* covers *e*, while for model ranking *e* covers *g*. For this reason they should not be used to perform prediction of the experimental values of *u*. In the same way, the predicted experimental value y_2 would be:

$$y_{u\,2} = \frac{(y_{e\,2} + y_{h\,2})}{2} \pm 0.5\left|y_{e\,2} - y_{h\,2}\right| = \frac{(190 + 235)}{2} \pm 0.5\left|190 - 235\right| = 212.5 \pm 22.5$$

whereas the experimental range of the y_2 values, predicted by the proposed approach, was:

$$y_{h\,2} \le y_{u\,2} \le y_{c\,2} \quad \Rightarrow \quad 235 \le y_{u\,2} \le 240$$

3.3.2 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 length path between *s* and *t*, that is *C*(*s,t*): the precision decreases for increased *C*(*s,t*). Moreover,

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

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 maximal element *C*(*s*,*u*) is not defined, as no comparable 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*(*t*,*u*) is not defined as no comparable 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 $y_{s\,r}$ and $y_{t\,r}$ are the experimental values of *s* and *t* for the *r-th* attribute respectively, and max_{y_r} and min_{y_r} the maximum and minimum 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*.

3.3.3 Hasse diagram evaluation

Further verification of model ranking applicability can be obtained by applying the above described ranking prediction procedure to the training set elements used initially 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 reduced data set the model is calculated, and from this model the interval values for the deleted elements are calculated. The calculated intervals are compared to the corresponding, experimentally derived, intervals.

For each element of the training set the *experimentally derived intervals* are calculated by deleting it from the experimental ranking diagram; the remaining training set elements are then used to calculate the experimental intervals of the deleted element from the experimental ranking diagram.

In the same way, the *model calculated intervals* are obtained by deleting one element at a time from the model ranking diagram, and using the remaining training set elements to calculate the model intervals of the deleted element from the model ranking diagram. 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 3.11, can be identified. 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.

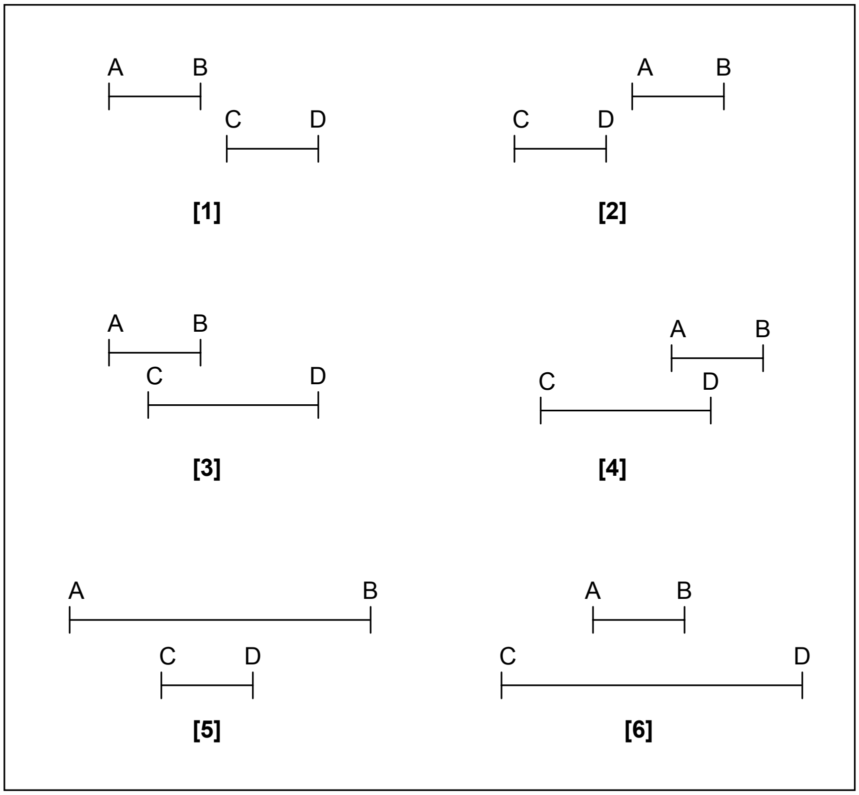


Figure 3.11 – Interval comparison.

Analysing one experimental attribute at a time, for each *i-th* element the disagreement δ_{ir} between its experimentally derived interval (A-B) and its model calculated interval (C-D) on the *r-th* attribute is calculated:

- 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_{y_r} and min_{y_r} 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:

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

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

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

A rougher evaluation of ranking quality can be derived from the non-error rate of the model (*NER*), defined as the ratio of the number of intervals correctly calculated by the model out of the total number of intervals, according to the following expression:

$$NER_r = \frac{C_r}{N}$$

C_r being the total number of intervals for the *r-th* attribute, where $\mathrm{C} \le y_{ir} \le \mathrm{D}$, and *N* the total number of intervals.

Note that isolated elements for which the interval is not calculable are considered errors.

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

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

$$NER_T = \frac{\sum_{r=1}^{R} NER_r}{R} \qquad NER_G = \sqrt[R]{NER_1 \cdot \ldots \cdot NER_R} \qquad NER_M = min_r\{NER_r\}$$

.

Q_T and NER_T being, respectively, the arithmetic mean of all the *R* attributes of the ranking model qualities and no-error rates, they are represent the least demanding parameters for evaluating overall model ranking quality. Instead the geometric means Q_G and NER_G are more severe parameters, able to enhance models not able to reproduce a correct experimental ranking for only a few attributes. The most demanding evaluation parameters of model quality are Q_M and $NER_{M,}$ these assuming 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, provided by the Hasse diagram developed on the experimental attributes and calculated intervals of the diagram developed on model attributes. 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, assuming the quantitative experimental values as intervals with equal lower and upper values, i.e. E = F, they are compared with the experimentally derived intervals (A-B), and for each *r-th* attribute the standardised interval disagreement ${}^{0}\delta^{*}_{i\,r}$ is calculated the same way, as described above.

The average disagreement between the quantitative experimental values and their derived intervals on the *r-th* attribute is calculated as:

$$\tilde{\delta}_r = \frac{\sum_{i=1}^{N} {}^{0}\delta^{*}_{i\,r}}{N}$$

The arithmetic mean calculated on all the *R* experimental attributes provides a measure of the uncertainty increase due to the replacement of a metric scale with an ordinal scale.

The numerical example already used to explain prediction calculation is used here to clarify the model ranking evaluation procedure.
Figure 3.12 shows both the experimental ranking developed on two experimental attributes y_1 and y_2 of Table 3.1, and the model ranking for the 9 training set elements.

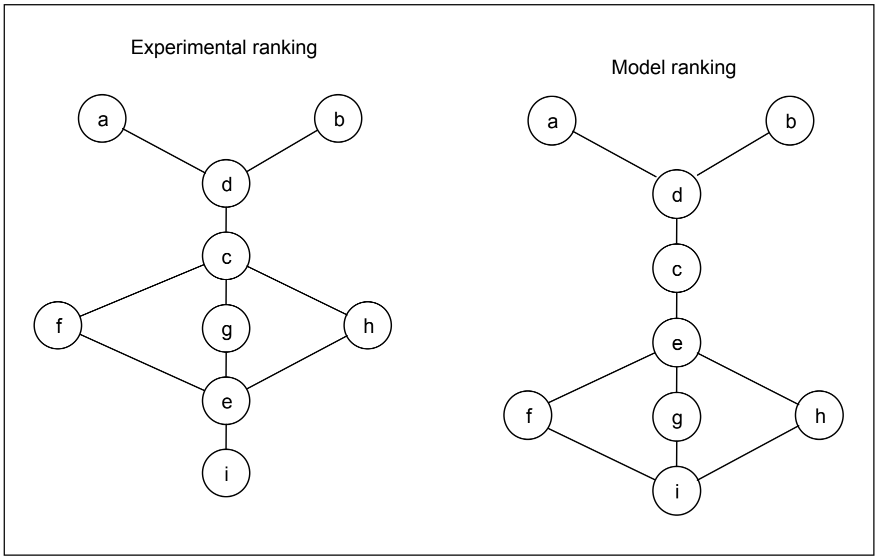


Figure 3.12 – Experimental *versus* model ranking Hasse diagrams.

Table 3.4 collects the experimentally derived intervals, and those calculated, for y_1, together with their disagreement δ^*_{i1}. The last three columns contain the rank uncertainty of the calculated intervals above and below (D_i^{sup} and D_i^{inf}) plus the experimental interval uncertainty *Ri*, from the model.

Element	y_1	Δy_1 *Exp*	Δy_1 *Mod*	δ^*_{i1}	$y_1\ D_i^{sup}$	$y_1\ D_i^{inf}$	$y_1\ R_{i1}$
a	180	> 140	> 140	0.000	-	0	-
b	150	> 140	> 140	0.000	-	0	-
c	130	120-140	90-140	0.306	0	0	0.510
d	140	130-150	130-150	0.000	0	0	0.204
e	90	82-90	**120-130**	0.490	0	0	0.102
f	100	90-130	**82-90**	0.490	0	0	0.082
g	120	90-130	**82-90**	0.490	0	0	0.082
h	90	90-130	82-90	0.490	0	0	0.082
i	82	<90	<90	0.000	0	-	-

Table 3.4 – Experimentally and calculated intervals comparison for y_1.

For the y_1 attribute the results are:

$\bar{\delta}_1 = 0.252$

$Q_1 = 0.748$

$NER_1 = 0.667$

In the same way the experimental intervals of y_2 were derived and the model intervals calculated. They are collected in Table 3.5 together with their disagreement δ_i^*, the rank and experimental interval uncertainties.

Element	y_2	*Δy_2 Exp*	*Δy_2 Mod*	$\delta^*_{i\,2}$	$y_2\ D_i^{sup}$	$y_2\ D_i^{inf}$	$y_2\ R_{i2}$
a	400	> 270	> 270	0.000	-	0	-
b	420	> 270	> 270	0.000	-	0	-
c	240	235-270	190-270	0.136	0	0	0.241
d	270	240-400	240-400	0.000	0	0	0.482
e	190	88-200	**235-240**	0.458	0	0	0.015
f	230	190-240	**88-190**	0.458	0	0	0.307
g	200	190-240	**88-190**	0.458	0	0	0.307
h	235	190-240	**88-190**	0.458	0	0	0.307
i	88	<190	<200	0.030	0	-	-

Table 3.5 – Experimentally and calculated intervals comparison for y_2.

For the y_2 attribute the following results were obtained:

$\bar{\delta}_2$ = 0.222

$Q_2 = 0.778$

$NER_2 = 0.556$

The overall model quality values calculated according to the parameters described above are shown in Table 3.6.

Q_T	NER_T	Q_G	NER_G	Q_M	NER_M
0.763	0.611	0.763	0.609	0.748	0.556

Table 3.6 – Overall quality parameter values.

Table 3.7 allows the comparison of the quantitative values of y_1 and y_2 and their corresponding experimentally derived intervals.

Element	y_1	Δy_1 *Exp*	$^{0}\delta^{*}_{i1}$	y_2	Δy_2 *Exp*	$^{0}\delta^{*}_{i2}$
a	180	> 140	0.408	400	> 270	0.392
b	150	> 140	0.102	420	> 270	0.392
c	130	120-140	0.204	240	235-270	0.105
d	140	130-150	0.204	270	240-400	0.482
e	90	82-90	0.082	190	88-200	0.337
f	100	90-130	0.408	230	190-240	0.151
g	120	90-130	0.408	200	190-240	0.151
h	90	90-130	0.408	235	190-240	0.151
i	82	<90	0.082	88	<190	0.307

Table 3.7 – Experimental values and experimentally derived intervals of y_1 *and* y_2.

As $\tilde{\delta}_1 = 0.256$ and $\tilde{\delta}_2 =$ = 0.274, the overall uncertainty increase is $\tilde{\delta}_T = 0.265$.

3.4 *W* Kendall rule

In searching for ranking models by evolutionary methods, optimising only Spearman's rank correlation or the Tanimoto or similarity indices could be overoptimistic and not sufficient to find optimal predictive models. In fact, these models could be affected by unwanted properties like chance correlation or the presence of noisy variables in the models. To avoid unlike model properties like chance correlation, a fitness function similar to the *QUIK rule* used for regression models [Todeschini *et. al*., 1998] is proposed here. The fitness function is based on Kendall's coefficient of concordance *W* [Kendall, 1948]. By using the *W* Kendall rule in an evolutionary algorithm for optimal model population searching, there should bed the maximising of Spearman's (or the Kendall) rank coefficient for total ranking models, or the similarity index (or Tanimoto indices) for partial ranking models; the models are accepted only if the following test is satisfied:

$$W_{XY} - W_X > \delta W \qquad\qquad (W\ \text{Kendall rule})$$

This is a simple test that allows the rejection of models with high predictor collinearity, which can lead to chance correlation. The *W* Kendall rule is based on Kendall's coefficient of concordance *W* that measures the total correlation of a set of rank-ordered variables. This rule is derived from the assumption that the total correlation in the set given by the model attributes (x_1, …, x_p), plus the experimental attributes (y_1, …, y_R) should always be greater than that measured only in the set of model attributes. Therefore, the *W* Kendall rule accepts only ranking models with the W_{XY} correlation among the [X+Y]–variables greater than the W_X correlation among the [X]–variables or

$$W_{XY} - W_X < \delta W \qquad \rightarrow \qquad \text{reject the model}$$

where δW is a user-defined threshold, greater than zero.

The W Kendall rule has been demonstrated to be very effective in avoiding models with multi-collinearity without prediction power.

3.5 Comparison of ranking model with traditional statistical techniques

As pointed out above, searching for a mathematical model is a complex procedure which requires first the identification of the type of model that is supposed to be more appropriate for the system investigated and for the objectives of the analysis. For this reason, ranking model properties can be examined by comparing them with statistical methods such as multi-linear regression (MLR).

Being based on elementary methods of Discrete Mathematics, total and partial ranking methods look very simple compared to multilinear regression (MLR) or PLS, thus they can be a very useful and simple tool for QSAR modelling. While multilinear regression assumes linearity over the whole training set with respect to predictors or functions of predictors and, at the same time, requires the normal distribution of the residuals, partial order ranking does not assume linearity nor does it call for distribution qualities. Compared to MLR the main disadvantage of partial ranking could be the need for the pre-processing of data to avoid the effects of stochastic noise. However, it has been demonstrated that the influence of uncertainty on the ranking can be significantly reduced by broad order statistics. One of the main advantages of PLS is the finding of orthogonal latent variables together with a potential dimension reduction; however, as in MLR, the problem of finding relationships among latent variables is connected to the supposition of a specific functional relationship.

As in traditional statistical QSAR approaches, ranking methods require a selection of variables to find the subset of variables which better reproduces the experimental ranking, excluding highly correlated variables, i.e. attributes which rank all the elements of the dataset in the same way.

One of the main disadvantages of partial ranking models is that the results are difficult to visualise when the number of elements is high since each element, or equivalence class, is represented by a small circle in the diagram. In such a case a cluster pre-processing analysis could be sufficient to solve the problem. An advantage MLR has over the ranking method is that all the numerical information is retained, and the predicted properties are sharply quantified. However, this higher information content leads to lower robustness. Moreover, 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 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. Thus, for when carrying out exposure analyses and risk assessment the use of ranking models is recommended, not to substitute conventional statistics but to supplement them.

CHAPTER 4

Ranking applications

The previous chapters explained how to use total and partial ranking methods to perform both data exploration and data modelling.

In the following some case study applications are illustrated and discussed. The data analysed come from diverse fields; they are both real data provided by scientific collaborations and data published in literature, mainly used to check the approaches proposed on already well studied data. Each application has been chosen in order to explain and illustrate some of the theoretical aspects introduced in the fist chapters. The first one is a case of study performed in collaboration with the organic research group of the Department of Environmental Sciences of Milano-Bicocca: total ranking strategies have been applied on paper industry data with the aim of selecting the best bleaching process among the ones proposed in the last years. The second application here presented refers to the ranking analysis performed within the European BEAM project on toxicity data, the aim being to compare EC curves and provide a method able to detect easily the similarity degree of the mechanism of action of the chemicals under study. The third case of study concerns the use of partial order ranking strategies as usefull tool to support waste management decision strategies: the waste production and discharching data of Italian regions have been compared in order to explorate the efficiency of their waste policy and to set a priority list for the governement. In the fourth study a dataset of 158 chemicals already analysed in literature for their environmental effects and exposure potential, has been chosen to compare pre-processing tools and to give evidence of the usefulness of partial ranking strategies in environmental decision problems. The fifth study has been performed in collaboration with the Italian Society of Chemistry and Cosmetology: both total and partial ranking analysis have been performed in order to evaluate the sensory panel and ranking the shampoo prototypes. The sixth analysis illustrates a total order ranking model developed for polychlorinated biphenyl compounds, which have been analysed according to some of their physico-chemical properties which play an important role on their environmental impact. The last study is an example of partial ranking model developed for 23 chemicals selected within the EU project: BEAM as active ingredients used in agricultural practice and tested for toxicity on *Scenedesum vacuolatus*. This analysis is here illustrated to point out that for exposure analysis and risk assessment ranking model can be a very useful tool in supporting decision making processes.

4.1 Optimisation of the “pulp and bleaching” process in the paper industry: a green chemistry problem

The paper industry is a high energy consuming industry, which makes a large use of feedstock and chemical additives. Much of the global pulp and paper industry has been making significant progress in pollution reduction in recent years. Many affords have been made to adopt new technology that holds promise to reduce energy consuming, environmental impact and use of chemical products. Green chemistry technologies have been researched and developed in paper industry for both their environmental benefits as well as economic benefits. Moreover new techniques based on diverse products are now under studies, the aim being selecting those that involve safer chemicals, i.e. not toxic and biodegradable chemicals, according to the green chemistry principles. The paper industry processes require a sequence of oxidative reactions commonly known as highly pollutant:

1. Pulp process, which removes lignin from cellulose: it is a multi-stages process, as lignin can not be selectively and completely oxidised in an unique step by any reagent.

2. Bleaching process to obtain bright white cellulosic fibre for quality papermaking
3. Waste products treatment

The dominant chemical technology for obtaining pulp suitable for bleaching is called the "*kraft* process": it is able to reduce lignin content from 18-35 % to 2-6%. However the following oxidative bleaching is necessary to obtain white pulp that does not yellow on ageing.

For many years, dominant oxidation technologies for papermaking have been provided by chlorine-based oxidants. Chlorine dioxide has expanded the most in the last decades because, even if it is a relatively expensive agent, it is highly selective for attack at the lignin over the cellulose. Nevertheless, chlorine-dioxide bleaching would appear to be less than ideal for the industry. It is one of the most expensive bleaching chemicals; it is produced in the mill by reduction of chlorate salts. Moreover, it is not totally chlorine free (TCF) and chlorine-containing effluents cannot be burned in the recovery boiler because they lead to corrosion problems. In addition, there is the potential for chlorinated dioxin production if the combustion proceeds in the presence of chlorine. For these reasons, environmental regulations coupled to market-pressures have forced the pulp and paper industry to explore alternatives to chlorine based bleaching practises.

Various technologies and bleaching chemicals have been suggested as candidates for chlorine replacement. The main bleaching processes investigated are the following:

- O_2 bleaching: the delignification is effective but its effectiveness is limited to about 50% after which a more severe treatment is required. The principal advantages concern the environment and the relative low chemical costs. However, the principal disadvantages are the high capital costs of O system and the low selectivity.
- O_2 + H_2O_2 bleaching: the hydrogen peroxide reinforced oxygen delignification is aimed to enhancing the efficiency of lignin removal. It allows the use of lower temperatures and the production of pulps with better strengths properties.

- O_2 + H_2SO_5 bleaching: the peroxymonosulfuric acid was found to enhance the oxygen delignification effectiveness as well.
- H_2O_2 bleaching: the hydrogen peroxide has been a technologically attractive oxidant to pulp and paper industry. The process involves a series of oxygen containing compounds that are formed and consumed dependent on pH, temperature and organic/inorganic contaminants.
- H_2O_2 activated by slow iron catalysts: it was observed that the hydrogen peroxide process can be employed by activators able to provide higher selectivity
- H_2O_2 activated by fast iron catalysts: modified activators have been analysed in order to obtain high selectivity and efficiency with smaller catalysts charge
- POMs $Na_5SiVW_{11}O_{40}$ bleaching: it is a new environmental friendly technology. They selectively oxidise lignin under anaerobic conditions and they can be reoxidise by oxygen. $Na_5SiVW_{11}O_{40}$ shows a high chemical selectivity; however, it is not stable at pH levels above 4.
- POMs $Na_5SiVW_{11}O_{40}$ bleaching in two successive stages: this process explores the possibility of further optimisation of the POM-based delignification.
- POMs $Na_6SiV_2W_{10}O_{40}$ bleaching: this process is based on a new kind of POM, which is stable at pH levels above neutral and is re-oxidised by oxygen on the basis of a new synthetic approach
- O_3 bleaching: it is highly capital intensive and the process is comparatively complex. Nevertheless, it is fast and the coupling of the advantage to the TCF environmental benefits has induced several companies to work on this technology.

It is now evident that the choice of the best bleaching process by the pulp and paper industry is not easy, as each of the mentioned process shows advantages and disadvantages at the same time. Moreover, the process selection is influenced by several aspects; in addition to the cited process aspects regarding the selectivity as well as the effectiveness, environmental properties together with economic aspects have to be taken into account. As a consequence of that, the selection of the best bleaching process is a complex problem: since the final objective consists in the contemporarily optimisation of several sub-objectives, decisions have to be taken contemporarily accounting several criteria. The evaluation of the overall quality of each process has to be based on several criteria.

4.1.1 Bleaching processes data

Multicriteria decision making methods have been applied on 12 bleaching processes performed on pulp delignificated by the *kraft* method. Each process has been evaluated according to process, efficiency, environmental and economic criteria.

The process criteria analysed are the following:

1. number of reaction steps
2. type of condition: aerobic / anaerobic
3. number of oxidant agents
4. reaction time
5. temperature required for the process
6. temperature range
7. kind of reaction: stechiometric / catalytic / catalytic with regeneration of the catalyst

The efficiency or selectivity criteria are:

1. pulp viscosity: the higher the selectivity, the longer the cellulose polymer chains in the fibres, the stronger the final paper product. Pulp strength is proportional to the length of the cellulose chains and the viscosity of pulp solutions is an indicator of pulp strength
2. kappa number: it is a measure of the amount of lignin present on the pulp

The environmental aspects are:

1. number of reagents involved in the process
2. pH
3. totally chlorine free TCF

The economic criteria are:

1. catalyst cost
2. oxidant agent cost

Table 4.1 shows the identification code used for twelve bleaching processes compared in the multicriteria analysis.

ID	*Activating agent*
D	Chlorine dioxide
O	Oxygen
OP	Oxygen and hydrogen peroxide
Opx	Oxygen and peroxymonosulphate
*QP**	Uncatalyzed hydrogen peroxide bleaching process
*QP*Fe slow*	Hydrogen peroxide with iron slow catalysts
*QP*Fe fast*	Hydrogen peroxide with iron fast catalysts
POMs (1)	Oxygen and $Na_5SiVW_{11}O_{40}$
POMs (2)	Oxygen and $Na_5SiVW_{11}O_{40}$ in two successive steps
POMs (3)	Oxygen and $SiV_2W_{10}O_{40}$
OZE (1)	Oxygen and Ozone (inlet ozone concentration = 2.0 wt)
OZE (2)	Oxygen and Ozone (inlet ozone concentration = 0.6 wt)

Table 4.1 – Bleaching processes analysed.

The seven process criteria values of the bleaching processes under study are collected in Table 4.2 .

	Process criteria						
Process	*N.step*	*Cond* [a]	*N. oxid*	*Time (min)*	*T (C°)*	*ΔT*	*React.*[b]
D	1	1	1	60	50	0	1
O	1	1	1	240	100	0	1
OP	2	1	2	85	90	26	1
Opx	2	1	2	85	90	24	1
*QP**	2	1	1	360	90	0	1
*QP*Fe slow*	2	1	1	60	90	0	3
*QP*Fe fast*	2	1	1	37.5	90	0	3
POMs (1)	2	2	2	30	125	0	4
POMs (2)	4	2	2	60	90	0	4
POMs (3)	2	2	2	180	150	0	4
OZE (1)	3	1	2	93	115	45	1
OZE (2)	3	1	2	93	115	45	1

Table 4.2 – Process criteria.

[a]: Reaction condition: 1 = aerobic; 2 = aero-anaerobic.
[b]: Reaction: 1 = stechiometric; 3 = catalytic; 4 = catalytic with catalyst regeneration

Table 4.3 shows the environmental, efficiency and economic criterion values of the bleaching processes.

	Environmental criteria			*Efficiency criteria*		*Economic criteria*	
Process	*N. reag.*	*pH*	*TCF*[a]	*ΔK*	*Δ visc*[b]	*$ catalyst*[c]	*$ oxid*[d].
D	1	3	0	18.5	6.4	0	1
O	2	10.5	1	14.5	14.4	0	3
OP	1	11.5	1	11.4	15.1	0	3
Opx	1	9.7	1	17.2	21.2	0	2
*QP**	4	11.6	1	15.2	13.4	0	2
*QP*Fe slow*	4	11.6	1	14.7	12.1	1	2
*QP*Fe fast*	4	11.6	1	13.5	10.3	1	2
POMs (1)	0	7	1	11.9	6.5	3	5
POMs (2)	0	5.5	1	12.6	1.4	3	5
POMs (3)	0	9.3	1	25	11	3	5
OZE (1)	2	2	1	7.9	8.4	0	4
OZE (2)	2	2	1	6.4	5	0	4

Table 4.3 – Environmental, efficiency and economic criteria.
[a]: Totally chlorine free process: 0 = no; 1 = yes
[b]: Viscosity variation (cP)
[c]: Catalyst cost: 0 = null; 1 = low; 2 = medium; 3 = high
[d]: Oxidant cost: 1 = very low; 2 = low; 3 = medium; 4 = high; 5 = very high

4.1.2 Bleaching ranking analysis

Since to perform ranking analysis of the bleaching processes, it was necessary to explicit whether the best condition was satisfied with a minimum value or a maximum value of the criterion and the trend from the minimum to the maximum, scientifically expert considerations have been taken into account.

The criterion setting is illustrated in Table 4.4, the criteria were weighted equally. The overall quality of the processes have been calculated by Desirability, Utility and Dominance functions; the obtained results are shown in Table 4.5 and the corresponding histograms in Figure 4.1.

Criterion	*Function*
Number of reaction steps	Inverse linear
Reaction condition	Logarithmic
Number of oxidant agents	Inverse exponential
Reaction time	Inverse exponential
Temperature required	Inverse sigmoid
Temperature range	Inverse linear
Kind of reaction	Linear
Number of reagents involved	Inverse linear
pH	Triangular
TCF	Linear
Kappa number variation	Sigmoid
Pulp viscosity variation	Inverse sigmoid
Catalyst cost	Inverse logarithmic
Oxidant agent cost	Inverse sigmoid

Table 4.4 – Criterion setting.

Process	*Desirability*	*Utility*	*Dominance*
O	0.703	0.751	0.346
*QP*Fe slow*	0.663	0.729	0.378
Opx	0.660	0.745	0.363
*QP*Fe fast*	0.658	0.730	0.391
OP	0.640	0.705	0.284
*QP**	0.632	0.706	0.324
POMs (1)	0.526	0.731	0.415
POMs (2)	0.516	0.719	0.418
OZE (1)	0.512	0.617	0.177
POMs (3)	0.509	0.728	0.363
OZE (2)	0.498	0.620	0.186
D	0.000	0.787	0.541

Table 4.5 –Desirability, Utility and Dominance values calculated on fourteen criteria (sorted according to desirability values).

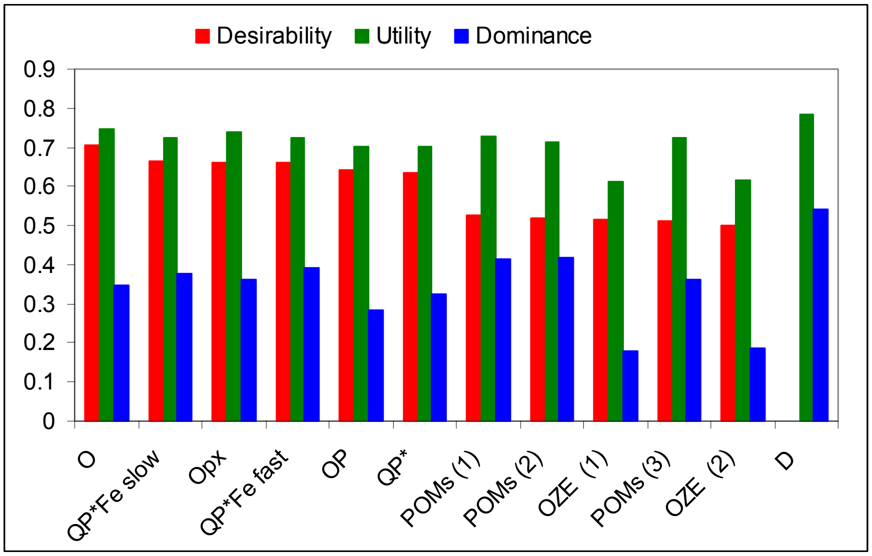


Figure 4.1 –Desirability, Utility and Dominance histograms.

It can be pointed out that the desirability quality evaluation is much more demanding than the utility function. The chlorine dioxide bleaching process low desirability (equal to 0) is due to the fact that it is not a TCF process and thus it is a high environmental impact process. The highest quality according to desirability index is provided by the oxygen bleaching, followed by hydrogen peroxide with iron slow catalysts, oxygen and peroxymonosulphate bleaching and hydrogen peroxide with iron fast catalysts process. These are the bleaching processes that show the overall highest quality; contemporary accounting all the criteria they are described with. They result as the ones that mostly satisfy the required properties of an acceptable bleaching process, as far as concerns the process, efficiency, environmental and economic criteria.

The evaluation provided by the utility method is much less severe than the desirability one: in fact, the overall quality of a process can be high even if a single utility function is low. The results provided by the dominance methods are based on pair comparison but it is less susceptible to the criterion setting.

Moreover, as the aim of the study was to find out the best bleaching process able to provide a satisfactory alternative to the highly environmental impact chlorine dioxide process, a further multicriteria analysis has been performed accounting separately for the process, efficiency, environmental and economic criteria in order to have a deeper understanding of the advantages and disadvantages of the processes. The multicriteria ranking on seven process criteria, was performed with the aim of finding out the bleaching process conducted in the best process conditions. The derived ranking illustrated in Table 4.6 confirms the good performances of the bleaching process based on hydrogen peroxide with iron slow and fast catalysts. In fact, these processes are realised in only two steps, in aerobic condition, with only one oxidant, a very low reaction time (37-60 min), at a 90°C temperature and by a catalytic reaction.

Process	*Desirability*	*Utility*	*Dominance*
*QP*Fe fast*	0.871	0.879	0.555
*QP*Fe slow*	0.871	0.879	0.526
POMs (1)	0.801	0.849	0.485
D	0.800	0.870	0.578
POMs (2)	0.775	0.842	0.448
O	0.746	0.812	0.351
Opx	0.702	0.755	0.290
OP	0.699	0.752	0.274
*QP**	0.691	0.750	0.324
POMs (3)	0.676	0.811	0.337
OZE (1)	0.587	0.644	0.123
OZE (2)	0.587	0.644	0.123

Table 4.6 –Desirability, Utility and Dominance values calculated on seven process criteria (sorted according to desirability values).

A satisfactory result is provided by oxygen and $Na_5SiVW_{11}O_{40}$ bleaching process (*POMs (1)*) as well. In fact, it is carried out in two steps, with two oxidants ($Na_5SiVW_{11}O_{40}$ and O_2), with a reaction time pretty low (30 min), at a 125°C temperature and it does not require changing temperature during the process. Moreover the reaction is catalytic with catalyst auto regeneration.

When the oxygen and $Na_5SiVW_{11}O_{40}$ bleaching process is realised in two successive steps (*POMs (2)*), its overall process quality is a little bit lower as it requires four reaction steps.

The oxygen and $SiV_2W_{10}O_{40}$ bleaching process (*POMs (3)*) is slightly penalised with respects to the other bleaching process based on polyoxomethalate oxidants, since it requires longer time (180 min) and higher temperature (150°C). Being the oxygen and oxygen plus hydrogen peroxide processes based on stechiometric reactions, their overall process quality is not in the first positions.

The ozone processes seem to be the worse ones as far as concerns process criteria; in fact, they are quite complex processes provided by three reaction steps and requiring temperature changing during the process.

The ranking evaluation of the twelve processes based on their environmental impact (Table 4.7) identified the polyoxomethalate oxidant bleaching processes as the best ones, according to the low reagents involved and their moderate pH.

Obviously, the ranking resulted from the efficiency criteria which measure the real process capability to remove lignin from cellulose in such a way to provide bright white cellulosic fibre for quality papermaking was of great interest (Table 4.8).

Process	*Desirability*	*Utility*	*Dominance*
POMs (1)	1.000	1.000	0.758
POMs (2)	0.923	0.929	0.722
POMs (3)	0.876	0.890	0.687
Opx	0.789	0.805	0.564
O	0.669	0.700	0.413
OP	0.659	0.719	0.447
OZE (1)	0.556	0.629	0.185
OZE (2)	0.556	0.629	0.185
*QP*Fe fast*	0.409	0.514	0.176
*QP**	0.409	0.514	0.176
*QP*Fe slow*	0.409	0.514	0.176
D	0.000	0.410	0.282

Table 4.7 –Desirability, Utility and Dominance values calculated on three environmental criteria (sorted according to desirability values).

Process	*Desirability*	*Utility*	*Dominance*
POMs (3)	0.874	0.878	0.636
D	0.849	0.854	0.818
*QP*Fe slow*	0.587	0.600	0.394
*QP**	0.571	0.573	0.394
*QP*Fe fast*	0.559	0.602	0.364
POMs (2)	0.554	0.650	0.576
O	0.502	0.504	0.273
POMs (1)	0.498	0.603	0.394
OP	0.337	0.362	0.091
OZE (1)	0.278	0.493	0.242
Opx	0.276	0.394	0.273
OZE (2)	0.228	0.510	0.303

Table 4.8 –Desirability, Utility and Dominance values calculated on two efficiency criteria (sorted according to desirability values).

From the obtained results it is to be highlighted that the oxygen and $SiV_2W_{10}O_{40}$ bleaching process (POMs (3)) is the only one which being even more efficient than the chlorine dioxide process, can realistically compete with it. In fact it is the process that provides the major lignin removal from cellulose, whereas all the other processes can not compete with chlorine process, according to the process efficiency.

Since the economic aspect cannot be ignored being one of the most important for the paper industry, the economic properties have been analysed as well. The results (Table 4.9) motivate the spread use of chlorine dioxide bleaching process, being the one more convenient; at the same time it is well explained the reason for the low overall quality of bleaching processes based on polyoxomethalate oxidants.

Process	*Desirability*	*Utility*	*Dominance*
D	1.000	1.000	0.818
*QP**	0.894	0.900	0.677
Opx	0.894	0.900	0.677
OP	0.775	0.800	0.515
O	0.775	0.800	0.515
*QP*Fe slow*	0.775	0.775	0.444
*QP*Fe fast*	0.775	0.775	0.444
OZE (1)	0.632	0.700	0.414
OZE (2)	0.632	0.700	0.414
POMs (3)	0.224	0.225	0.061
POMs (1)	0.224	0.225	0.061
POMs (2)	0.224	0.225	0.061

Table 4.9 –Desirability, Utility and Dominance values calculated on two economic criteria (sorted according to desirability values).

The hypothesis of both the relevant POM bleaching quality and their penalty according to economic criteria has been confirmed by a further ranking analysis performed on all the criteria excepted the economic ones. In this case POM processes resulted the best ones.

According to all the analyses performed, at the time, the best bleaching processes, which can realistically be valid alternatives to the chlorine dioxide process, are:

Oxygen bleaching

Oxygen and peroxymonosulphate bleaching

Hydrogen peroxide with iron slow catalysts

Hydrogen peroxide with iron fast catalysts

Moreover, it is to be highlighted that POM bleaching processes seem to be the more promising process according to environmental, process and efficiency criteria; however, their wide use is not permitted because of the high costs of their synthesis.

The development of new cheaper technologies for POM synthesis would make the POM bleaching process economically accessible and competitive with the traditional chlorine dioxide process.

4.2 Comparison of toxicity profiles by Hasse diagram technique

Natural environments and ecosystems are not exposed to individual chemicals but to complex multi-component mixtures of chemicals of various origin (industry, agriculture, urban sewage). Nevertheless, most ecotoxicological research and chemical regulation focus on hazard and exposure assessment of individual chemicals only and the chemical mixtures in the environment is ignored to a large extent. Therefore, there is the need for developing risk assessment procedures no longer restricted to single toxicants and instead considering combined effects resulting from multiple chemical exposure. The BEAM project (Bridging Effect Assessment of Mixtures to Ecosystem Situations and Regulation) (see also: http://www.aquatox.uni-bremen.de/beam) is an European research project started in spring 2000 with the aim of developing procedures capable to derive Environmental Quality Standards for mixtures of chemicals likely to occur in the environment. The predictive mixture toxicities approaches imply that the chemical composition of the mixture of interest in known. Two different concepts, termed Concentration Addition and Independent Action, are thought of being more generally applicable and allow to calculate an expected mixture toxicity on the basis of known toxicities of the mixture components. Both concepts consider cases in which all substances in a mixture affect the same experimental endpoint, and both require precise knowledge about qualitative and quantitative composition of the mixture of study. However they are based on opposite assumption with respect to the similarity of the mechanism of action of the individual components. The Concentration Addition concept is based on the idea of a similar mechanism of action of all the substances in the examined mixture and assumes that the total effect of the mixture is the sum of the concentrations of the single compounds in the mixture scaled for their potency. The alternative concept of Independent Action assumes that the mixture components act dissimilarly and estimates the mixture effect from the product of the chemical effects applied singly in a concentration that corresponds to the concentration of the chemicals in the mixture.

As these two concepts are based on opposite assumption with respect to the similarity of the mechanism of action of the individual components, the first step in mixture risk assessment procedure is to evaluate the similarity of the mechanism of action of the individual components.

The mode of action of a toxicant is frequently described by its Effect-Concentration curve, where the concentrations of the toxicant that are estimated to cause a predefined effect are plotted. Typically the entirely curve is drawn from the EC01, EC10, EC50 and EC90 values, which are the concentration at which 1%, 10%, 50% and 90% of the test population are affected, respectively. If the common assumption that similarly acting substances show similar EC curves, whereas dissimilarly acting substances show dissimilar EC curves, characterised by many crossing, is accepted then, the EC curves have to be compared to establish whether the mixture components act with the same mechanism of action or not, and thus whether using the Concentration Addition or Independent Action mixture toxicity prediction model. The Hasse diagram technique has been proposed as an useful tool to compare and rank toxicities of chemicals studied in the BEAM project. Being Hasse diagram technique a multivariate explorative method it is not limited to one level of biological response and thus it seems suitable for Concentraction Effect curves comparison, highlighting different EC curve shapes.

4.2.1 Toxicity data

A ranking explorative analysis by Hasse diagram technique has been performed to compare EC curves both of similarly and dissimilarly acting substances, the aim being to evaluate the Hasse diagram capability of comparing toxicity profiles and provide diverse results for EC curves of substances with the same mode of action from the ones of substances with different mode of action. The Hasse diagram technique has been applied on two datasets:

- 12 phenylureas → similar acting chemicals
- 21 diverse chemical → dissimilar acting chemicals

4.2.2 Comparison of concentration effect curves of similar acting substances

The EC01, EC10, EC50 and EC90 values of 12 phenylureas analysed by Hasse diagram technique are collected in Table 4.10.

		Log(1/EC)			
ID	*Substance*	EC01	EC10	EC50	EC90
1	Buturon	1.877	0.897	0.111	-0.390
2	Chlobromuron	3.244	2.059	1.222	0.824
3	Chlortoluron	2.523	1.576	0.815	0.332
4	Diuron	3.071	2.223	1.538	1.109
5	Fenuron	0.927	0.137	-0.633	-1.214
6	Fluometuron	2.030	0.772	-0.173	-0.849
7	Isoproturon	2.226	1.363	0.642	0.166
8	Linuron	3.155	1.990	1.056	0.463
9	Metobromuron	1.430	0.630	-0.019	-0.490
10	Metoxuron	2.320	1.209	0.319	-0.249
11	Monolinuron	2.058	0.920	0.007	-0.575
12	Monuron	2.569	1.367	0.402	-0.212

Table 4.10 –Toxicity values of 12 phenylureas.

The Hasse diagram, shown in Figure 4.2, is a quite simple diagram arranged in seven levels. It identifies two maximals: chlobromuron and diuron, selected as the most toxic. They are incomparable since some contradictions exist among their EC values: chlobromuron is more toxic than diuron on 01 and 50 concentration levels, but it is less toxic on 10 and 90 concentration levels. Fenuron is the less toxic substance, as it is characterised by the lowest effect concentration values. Since the 12 phenylureas are similar acting substances, not too many contradictions in their EC values exist. Thus, as expected, the Hasse diagram is characterised by a few number of incomparabilities: 18 over 132 comparisons. The 9 incomparabilities (counted in only one direction) are listed below:

- (2) Chlobromuron || (4) Diuron
- (4) Diuron || (8) Linuron
- (3) Chlortoluron || (12) Monuron
- (7) Isoproturon || (12) Monuron
- (7) Isoproturon || (10) Metoxuron
- (1) Buturon || (11) Monolinuron
- (1) Buturon || (6) Flometuron
- (9) Metobromuron || (11) Monolinuron
- (6) Flometuron || (9) Metobromuron

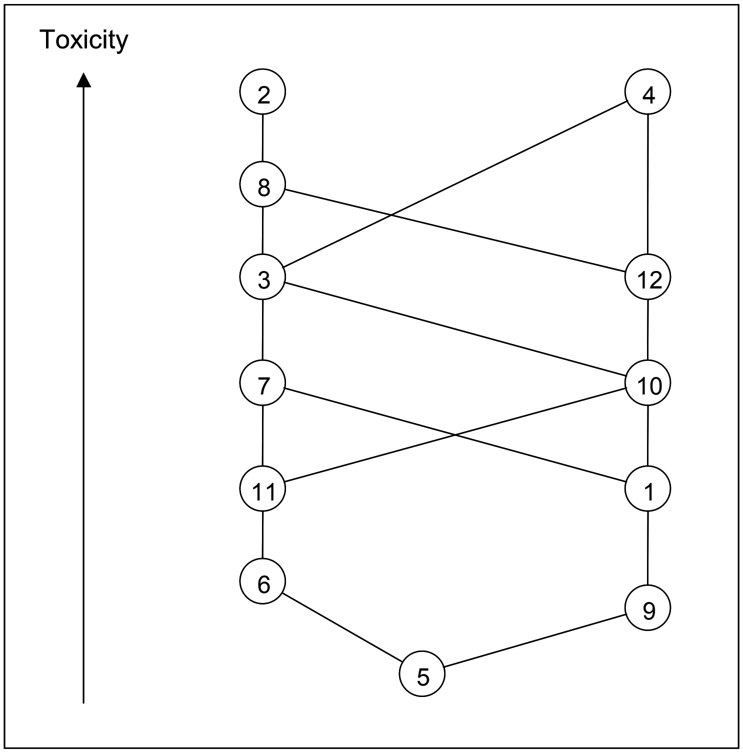


Figure 4.2 – Hasse diagram of 12 phenylureas.

In Figure 4.3, the phenylureas EC curves are illustrated: the EC curves are quite similar and only a few crossing, corresponding to the above incomparabilities can be identified.

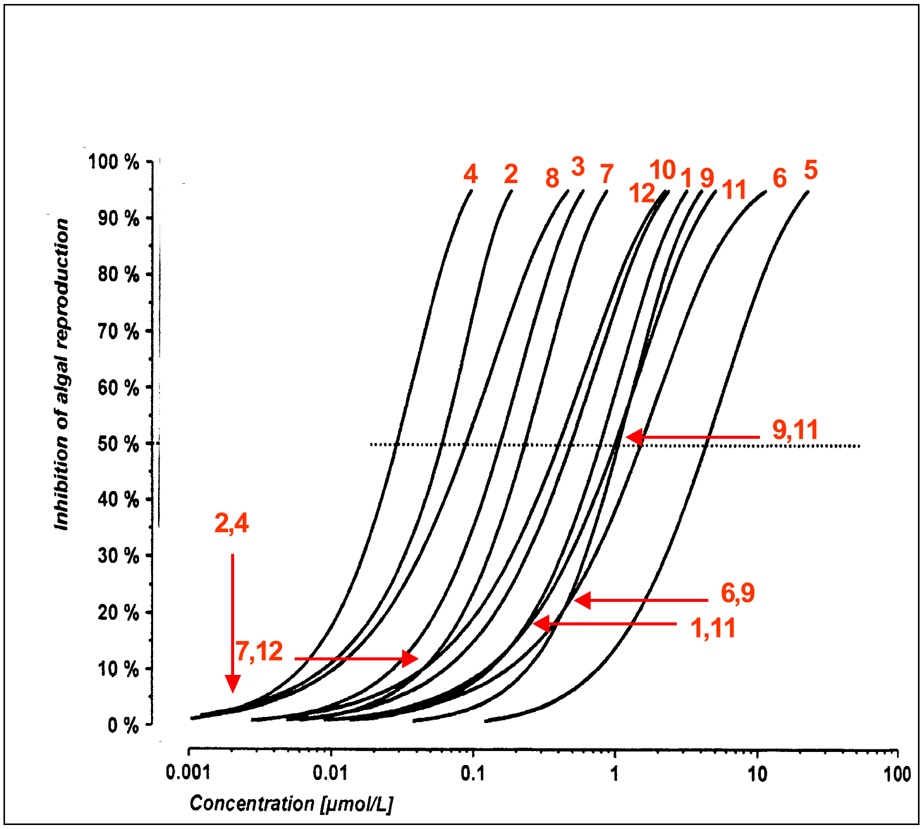


Figure 4.3 – Effect concentration curves of 12 phenylureas.

Therefore, the analysis performed not only allows to rank the substances according to their toxicity values but also provide a clear and simple way to detect EC curves crossing.

4.2.3 Comparison of concentration effect curves of dissimilar acting substances

The 21 substances with diverse mechanism of action analysed by Hasse diagram technique together with their EC01, EC10, EC50 and EC90 values of are collected in Table 4.11.

		Log(1/EC)			
ID	*Substance*	EC01	EC10	EC50	EC90
1	Aclonifen	2.455	2.022	1.523	1.067
2	8-Azaguanine	2.347	1.815	1.150	0.134
3	Azaserin	2.592	0.927	-0.250	-0.873
4	CCCP	0.979	0.488	-0.092	-0.653
5	Chloramphenicol	-0.483	-0.721	-1.100	-1.955
6	Fenfuram	1.106	0.139	-0.635	-1.129
7	5-Fluoruracil	1.983	1.582	1.013	-0.059
8	Fusidic acid	0.588	0.247	-0.256	-0.906
9	Kresoxim-methyl	0.793	0.296	-0.310	-0.925
10	Metalaxyl	-0.784	-1.633	-2.312	-2.745
11	Metazachlor	2.268	1.182	0.775	0.587
12	Metsulfuron-methyl	1.366	0.273	-0.604	-0.162
13	Nalidixic acid	-0.700	-1.419	-1.994	-2.362
14	Norflurazon	2.249	1.959	1.699	1.428
15	Paraquat	0.668	0.417	0.107	-0.292
16	n-Propyl-gallate	-1.441	-1.845	-2.214	-2.583
17	Pyrimethamin	-0.218	-1.157	-1.711	-2.008
18	Seanine	1.155	0.924	0.635	0.187
19	Steptomycin-sulphate	-0.213	-0.676	-1.099	-1.523
20	Terbuthylazine	2.243	1.642	1.161	0.852
21	Triadimenol	0.496	-0.109	-0.539	-0.76

Table 4.11 –Toxicity values of 21 dissimilar acting substances.

The Hasse diagram, shown in Figure 4.4, is a quite complex diagram arranged in nine levels. It identifies three maximals, aclonifen, azaserin and norflurazon, and two minimals metalaxyl and n-propyl-gallate. Being the 21 substances characterised by a diverse mechanism of action several contradictions exist among their EC values. The diagram identifies 76 incomparabilities over 420 comparisons.

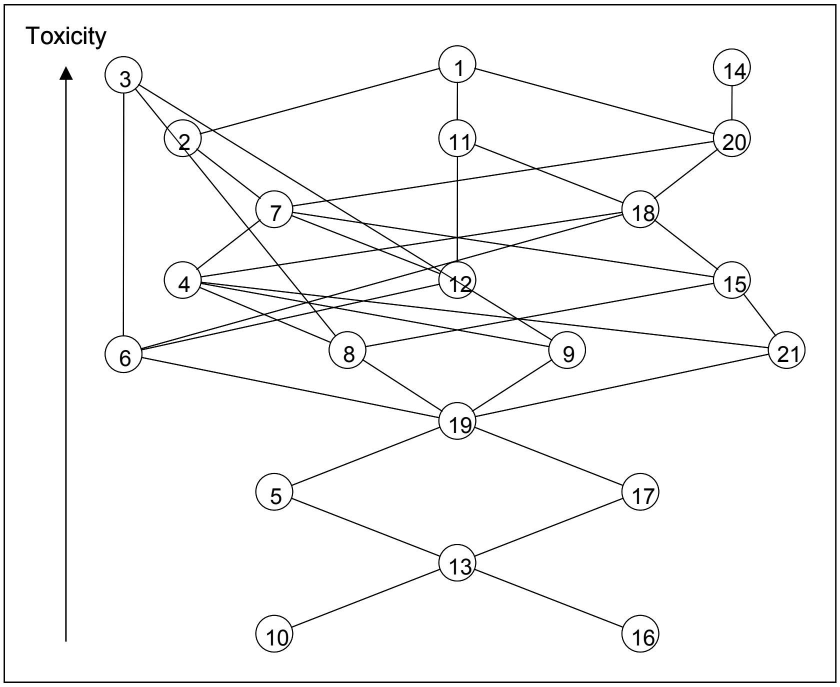


Figure 4.4 – Hasse diagram of 21 dissimilar acting substances.

In Figure 4.5, the 21 substances EC curves are illustrated: the EC curves are pretty diverse and many crossing, corresponding to the Hasse diagram incomparabilities can be identified.

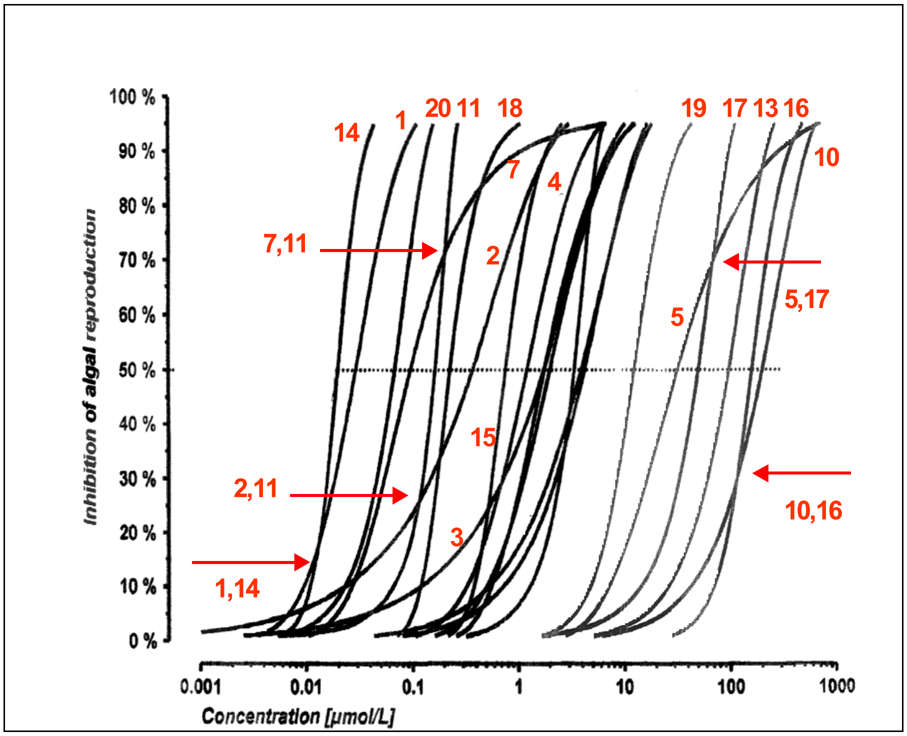


Figure 4.5 – Effect concentration curves of 21 dissimilar acting substances.

The obtained results on the two datasets investigated highlight that the Hasse diagram technique can be an efficient and simple tool not only to rank the substances according to their toxicity values but also to detect the similarity of toxicants mechanism of action. In fact, the higher the number of incomparabilities in the diagram, the higher the number of crossing in the EC curves, and thus the higher the dissimilarity in the mechanism of action of the substances under study. Thus, the complexity and the number of incomparabilities provided by partial ranking analysis on the EC values of substances with an unknown mode of action can be used as a measure of the dissimilarity degree in their mechanism, suggesting which mixture toxicity prediction model to be used (Concentration Addition or Independent Action).

4.3 Waste management analysis

Over 1.8 billion tonnes of waste are generated each year in Europe. This is mainly made up of waste coming from households, commercial activities (e.g., shops, restaurants, hospitals etc.), industry (e.g, pharmaceutical companies, clothes manufacturers etc.), agriculture (e.g., slurry), construction and demolition projects, mining and quarrying activities and from the generation of energy. With such vast quantities of waste being produced, it is of vital importance that it is managed in such a way that it does not cause any harm to either human health or to the environment. Between 1990 and 1995, the amount of waste generated in Europe increased by 10%, according to the Organisation for Economic Cooperation and Development (OECD). Most of what we throw away is either burnt in incinerators, or dumped into landfill sites (67%). But both these methods create environmental damage. Landfilling not only takes up more and more valuable land space, it also causes air, water and soil pollution, discharging carbon dioxide (CO_2) and methane (CH_4) into the atmosphere and chemicals and pesticides into the earth and groundwater. This, in turn, is harmful to human health, as well as to plants and animals. Under EU policy, landfilling is seen as the last resort and should only be used when all the other options have been exhausted , i.e., only material that cannot be prevented, re-used, recycled or otherwise treated should be landfilled. The EU's Sixth Environment Action Programme identifies waste prevention and management as one of four top priorities. The EU is aiming for a significant cut in the amount of rubbish generated, through new waste prevention initiatives, better use of resources, and encouraging a shift to more sustainable consumption patterns. It wants to reduce the quantity of waste going to 'final disposal' by 20% from 2000 to 2010, and by 50% by 2050, with special emphasis on cutting hazardous waste. In the present study, partial ranking analysis has been performed on waste data provided by National Waste Cadaster, with the aim of comparing italian regions as far as concerns the amount of waste they produce and discharge in landfills. Moreover, ranking analysis providing a list of priority italian regions is a suitable tool to support waste management decisions problems.

4.3.1 Waste data

A ranking explorative analysis by Hasse diagram technique has been performed to compare italian regions according to the amount of waste they produce and discharge in landfills. The data have been provided by National Waste Cadaster and are collected in Table 4.12 and 4.13: each region is decribed in terms of urban waste, non-hazardous and hazardous waste production and amount of their discharging in landfills.

ID	*Region*	*UW*	*NHW*	*HW*
1	Piemonte	2006853	3841809	402117
2	Valle d'Aosta	62614	39189	2918
3	Lombardia	4279974	8494457	1172724
4	Trentino Alto Adige	508272	711490	42048
5	Veneto	2112601	5335021	440440
6	Friuli Venezia Giulia	572480	1326474	128783
7	Liguria	898758	922149	89203
8	Emilia Romagna	2413949	5998763	419496
9	Toscana	2105665	5012818	230292
10	Umbria	422108	1373125	21173
11	Marche	761011	1037527	43161
12	Lazio	2779686	1884997	121110
13	Abruzzo	608995	676999	48058
14	Molise	113930	300191	13203
15	Campania	2561546	1736932	84664
16	Puglia	1802608	2805891	98445
17	Basilicata	218822	474929	6947
18	Calabria	821129	375414	43988
19	Sicilia	2552727	970225	89318
20	Sardegna	760186	1526171	313231

Table 4.12 – Waste production in italian region in 1999. *UW*: urban waste in t/year; *NHW*: non-hazardous waste in t/year; *HW*: hazardous waste in t/year.

ID	*Region*	*L-UW*	*L-NHW*	*L-HW*
1	Piemonte	1526554	864865	17543
2	Valle d'Aosta	54923	67070	0
3	Lombardia	1504586	3979258	100729
4	Trentino Alto Adige	308143	506370	27834
5	Veneto	1489658	2274287	115059
6	Friuli Venezia Giulia	334832	461971	1766
7	Liguria	833126	1219190	44287
8	Emilia Romagna	1879281	411610	6625
9	Toscana	1275113	1641492	88308
10	Umbria	324790	718974	3630
11	Marche	684174	249353	737
12	Lazio	2619169	1170487	10660
13	Abruzzo	477690	202509	5754
14	Molise	111560	26834	1561
15	Campania	2635617	59228	4925
16	Puglia	1776093	1061765	1521
17	Basilicata	198057	143951	4131
18	Calabria	724757	136005	17987
19	Sicilia	2412985	530890	7755
20	Sardegna	573584	1443890	278340

Table 4.13 – Waste discharging in lanfills in italian region in 1999. *L-UW*: urban waste discharged in landfills in t/year; *L-NHW*: non-hazardous waste discharged in landfills in t/year; *L-HW*: hazardous waste discharged in landfills in t/year.

4.3.2 Waste ranking analysis

Partial ranking analysis has been performed on the waste production data, the aim being to analyse the efficiency of italian region waste management strategy to reduce the amount of waste generated. Figure 4.6 shows the priority ranking in form of a Hasse diagram according to waste production. The twenty regions have been compared have been ranked in eight levels.

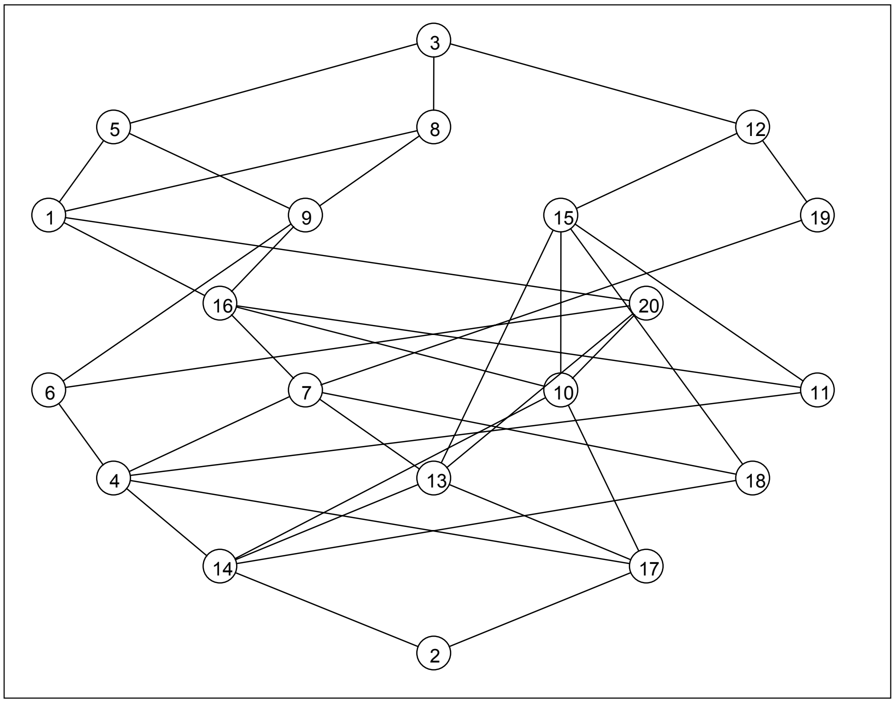


Figure 4.6 – Hasse diagram of 20 italian regions ranked according to their amount of waste production. (*UW*, *NHW, HW*).

The diagram allows an easy comparison of the regions: it is not surprising that Lomabrdia is the maximal region for waste production, followed by Veneto, Emilia Romagna and Lazio, which are all characterside by a high waste production even if they are not comparable. These three regions must be considered of equally priority even if for different reasons. In fact, the urban and non-hazardous waste production of Emilia Romagna is greater than the Veneto one, but the hazardous waste production of Veneto is higher than the Emilia Romagna one, due to the great number of industries located in Veneto region. Moreover, Lazio is the region with the second highest urban waste production and is characterised by a lower non-hazardous and hazardous waste production than Veneto and Emilia Romagna. As it was expected Valle d'Aosta is minimal, being characterised by the lowest waste production; it is follwed by Molise and Basilicata: the former shows a higher production of hazardous waste and a lower production of urban and non-hazardous waste than the latter.

As European Member States should ensure that existing landfill sites may not continue to operate unless they comply with the provisions of the Council Directive 99/31/EC of 26 April 1999 on the landfill of waste, it was of relevance for the Italian governement to evaluate the amount of waste discharged in landfills. Thus, a partial ranking analysis has been performed on waste discharging data, and the corresponding Hasse diagram is shown in Figure 4.7.

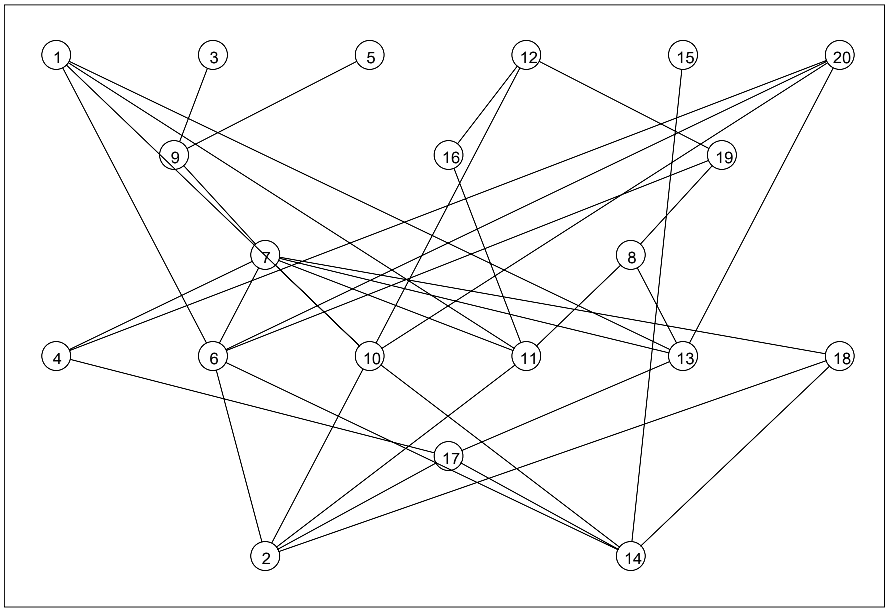


Figure 4.7 – Hasse diagram of 20 italian regions ranked according to their amount of waste discharged in landfills. (*L-UW, L-NHW, L-HW*).

The diagram points out six regions that are the most environmentally hazardous because of the huge amount of waste they discharge in landfills: Piemonte, Lombardia, Veneto, Lazio, Campania and Sargegna as maximals: even if differences exist related to the kind of waste discharged in landfills, all these regions being at the first priority level must be considered of major attention. Moreover, attention should be paid even to the second priority regions: Toscana, Puglia and Sicilia wich are then followed by Friuli Venezia Giulia, Liguria, Umbria and Marche, on the third level. When all the variables are contemporarly taken into account, the priority ranking obtained in form of Hasse diagram is the one shown in Figure 4.8.

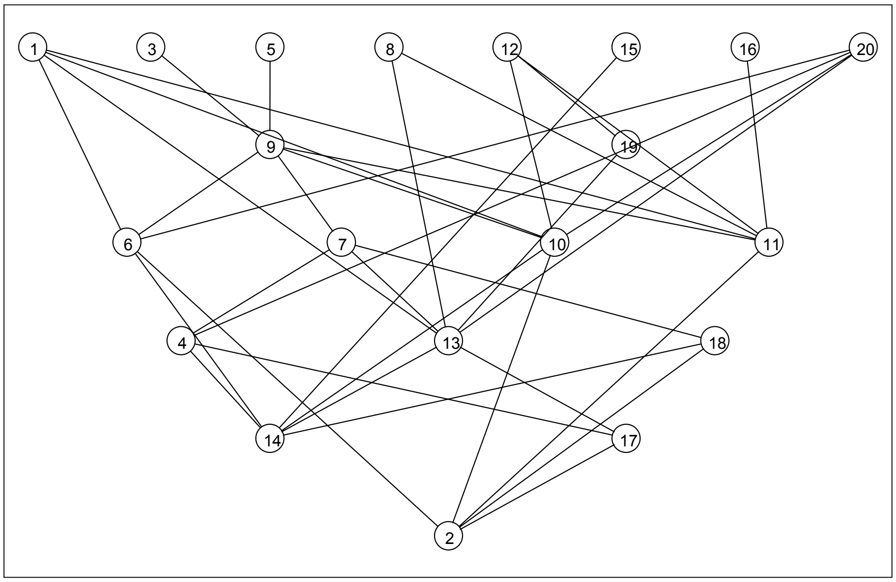


Figure 4.8 – Hasse diagram of 20 italian regions ranked according to their amount of waste produced and discharged in landfills. (*UW*, *NHW, HW, L-UW, L-NHW, L-HW*).

It can be pointed out that the twenty regions have been ranked in six priority levels: Puglia is now at the first attention level together with Piemonte, Lombardia, Veneto, Emilia Romagna, Lazio, Campania and Sargegna. Toscana and Sicilia are located at the second level.

The present study shows that Hasse diagram technique is a solution to the understanding of multivariate data without reducing the information in a trivial chain. Moreover, partial ranking methods are suitable techniques for analysing environmental problems, which are based on multiple criteria to estimate hazard. Hasse diagrams allow visual comparison of regions based on multiple variables, which might otherwise be very confusing when displayed in table form. The ranking procedure provides a ranking of regions into distinct hazard groups and a visual identification of contradictions. The approach proposed can be included into an expert system that help decision makers interested in objective data ranking analysis.

4.4 Ranking chemicals for environmental hazard: comparison of pre-processing methodologies.

In March 1999 the European Council required the Commission to establish a list of substances priorised on the basis of their risk to the aquatic environment and to human health via the aquatic environment. In order to establish a list of priority substances in accordance with the given provisions, a combined monitoring-based and modelling based priority scheme have been elaborated. Several approaches have been proposed: often total ranking methods have been used. These methods rank chemicals according to an index function which combines chemical properties describing the toxicity, exposure and persistence in the environment of the chemicals investigated. The choice of the index function and the weights assigned to each chemical property is subjective and depend on the developer of the analysis. However, by these approaches much information is lost as the same global index can be provided to chemicals which effects on the environment may be totally different. Thus, in all that cases which the aims are both to provide a ranking of chemicals for environmental hazard and to detect their different effect on the environment, a partial ranking analysis is more suitable than a total ranking approach. In fact, Hasse diagrams make contradictions in data evident and allow to easily relate the ranking position of chemicals of interest to any contradictory data available. However, as pointed out in chapter 2, one of the main drawback of the Hasse diagram technique is its strongly dependence on the clear appearance of the diagram, which is hardly achievable when many chemicals are analysed and compared, or when many criteria are considered and when data are affected by uncertainty. In these cases, pre-processing techniques are required. In the present study a priority setting scheme based on partial ranking method is proposed. The environmental hazard of 140 chemicals including inorganic chemicals plus 18 high volume pesticides has been analysed by comparing their relative risk, computed by Hasse diagram technique. Chemicals have been described by many criteria accounting to their effect on human health and on environment with the aim of setting protocols and select priority chemicals to be submitted to revision or to supplementary testing.

Pre-processing statistical techniques have been analysed and compared with respect to their capability to support ranking analysis.

4.4.1 The dataset

The dataset analysed is made of 158 chemicals collected by Swanson [Swanson *et al*., 1997] and described by their human health effects, environmental effects and exposure potential. One hundred and forty are from the 1989 EPA Toxic Release Inventory (TRI) and 21 are high volume pesticides. Table 4.14 shows their toxicological and exposure endpoints and Table 4.15a and 4.15b all the data. The 1989 TRI data are pretty old however in this study the interest is in the methodology rather than in how the data are since these results are not relevant by themselves. Moreover, this dataset being already analysed by partial ranking approach by Halfon and Bruggemman [Halfon and Bruggemman, 1998] has been selected in order to allow a comparison of the proposed pre-processing technique with the one adopted by Brüggemann and Halfon. The dataset include toxicity data, persistence data and loadings (10 criteria for each substance).

	Toxicological endpoint	*Meaning*
LD50	Rodent Oral LD50	The concentration of a substance, expressed in mass of the substance per mass of the animal, that will kill half of a group of rodents within 14 days when administered orally as a single dose.
LC50	Rodent Inhalation LC50	The concentration of a substance in air that will kill half of a group of rodents when inhaled continuously for 8 h or less.
NCAR	Evidence of carcinogenicity	Based on EPA and International Agency for Research on Cancer (IARC) classification
NTOX	Other specific effects	Includes positive evidence of mutagenicity, developmental effects, reproductive effects, other chronic effects, and neurotoxicity.
F-LC50	Fish LC50	The concentration of a chemical in water that causes death in %0% of the fish tested in a 96-h test.
NOEL	Fish NOEL	The highest dosage administered that does not produce observable toxic effects, estimated from LC50 data.
BOD	Biological oxygen demand	The time required to biodegradate a chemical such that its BOD in water is reduced by half.
Hydro	Hydrolysis half life	The time required for the amount of a chemical to be reduced by half through hydrolysis reaction in water, at pH 7.
BCF	Aquatic bioconcetration factor	The ratio of the concentration of a chemical in an aquatic organism to that in water at steady-state.
RWF	Amount released into the environment, air and water.	The amount of annual releases or transfer of chemicals into the environment. Modified by logarithmic transformation.

Table 4.14 – Toxicological and exposure endpoints.

ID	*Substance*	*LD50*	*LC50*	*NCAR*	*NTOX*	*F-LC50*
1	1,1,1-Trichloroethane	11240	2000	0	3	48
2	1,1,2-Trichloroethane	150	2000	2	1	7
3	1,2,4-Trichlorobenzene	300	1100	0	2	3
4	1,2,4-Trimethylbenzene	5000	3655	0	0	8
5	1,2-Dichlorobenzene	1400	1700	0	1	0.55
6	1,2-Dichloroethane	780	2063	4	4	136
7	1,2-Dichloropropane	3000	5554	0	0	127
8	1,3-Butadiene	3210	128850	4	4	4
9	1,4-Dichlorobenzene	3790	1100	4	1	34
10	1,4-Dioxane	3150	6368	4	0	10352
11	2,4-D	275		4	3	71
12	2,4-Dinitrophenol	30		0	3	11
13	2,4-Dinitrotoluene	268		4	3	24
14	2-Ethoxyethanol	1400	3185	0	3	16305
15	2-Methoxyethanol	950	2590	0	3	22655
16	2-Nitropropane	725	600	4	3	5
17	4,4'-Isopropyldenediphenol	2500	200	0	0	5
18	4,4'-Methylenedianiline	185	163	4	1	45
19	4-Nitrophenol	620	50	0	2	41
20	Acetaldehyde	1930	1500	4	0	34
21	Acetone	3000	42000	0	1	7200
22	Acetonitiile	3800	15000	0	3	1640
23	Acrylamide	107	1000	4	4	109
24	Acrylic acis	193	1200	0	1	186
25	Acrylonitrile	78	576	4	3	10
26	Allyl chloride	425	926	2	3	72
27	Alluminium (fume or dust)	9999	500	0	0	
28	Ammonia	350	2377	0	1	2
29	Ammonium nitrate	4500		0	1	800
30	Ammonium sulfate	3000		0	0	4000
31	Aniline	250	306	0	1	108
32	Anthracene	17000	250	1	1	0.01
33	Antimony compounds	20000		0	2	833
34	Arsenic compounds	8		5	3	32
35	Asbestos (friable)	9999	9999	5	1	
36	Barium compounds	132		0	2	200
37	Benzene	4700	17500	5	3	19
38	Benzoyl chloride	2460	163	0	2	35
39	Biphenyl	3280	25	0	2	2
40	Bis(2-ethylhexyl) adipate	9110		0	1	0.35
41	Bromomethane	214	780	0	2	11
42	Butyl acrylate	3730	2730	0	0	2
43	Butyl benzyl phthalate	2330		0	1	43
44	Butyraldehyde	2490	7547	0	0	32
45	Cadmium compounds	88	306	4	3	0.1
46	Carbon disulfide	2780	1604	0	4	694
47	Carbon tetrachloride	2800	19052	4	3	41
48	Carbonyl sulfide		10000	0	1	2685
49	Catechol	260		0	0	9
50	Chlorine	8910	34	0	1	0.34
51	Chlorine dioxide	292	130	0	2	0.17
52	Chlorobenzene	1440	1100	0	2	17
53	Chloroethane	7500	29	0	0	16
54	Chloroform	908	5720	4	3	71
55	Cloromethane	1800	3063	2	3	550
56	Chlorophenols [o]	261	100	0	0	19
57	Cloroprene	260	3253	0	5	2

ID	*Substance*	*LD50*	*LC50*	*NCAR*	*NTOX*	*F-LC50*
58	Chlorothalonil	6000	7	0	2	0.05
59	Chrominum compounds	97		5	1	33
60	Cobalt compounds	55		0	1	0.38
61	Copper compounds	300		0	2	0.33
62	Cresol (mixed isomers)	760	50	0	1	13
63	Cumene	2910	8000	0	2	6
64	Cumene hydroperoxide	382	200	1	0	62
65	Cyclohexane	29820	500	0	0	5
66	Decabromdiphenly oxide	2570		0	2	0.06
67	Di(2-ethylhexyl) phthalate	30000		4	4	1
68	Diaminotoluene (mixed isomers)	13000		0	1	0.93
69	Dibutyl phthalate	260	100	4	2	37
70	Dichlorobenzene (mixed isomers)	9000	500	0	3	1
71	Dichloromethane	2600	1100	0	1	0.54
72	Diethanolamine	1600	17400	4	1	330
73	Diethyl phthalate	710	484	0	0	4710
74	Dimethyl phthalate	9000	537	0	0	32
75	Epichlorohydrin	2400	500	0	0	121
76	Ethylbenzene	40	500	4	4	35
77	Ethylene	5460	5000	0	3	11
78	Ethylene glycol	9999	950000	0	1	14
79	Ethylene oxide	6610	1000	0	1	227634
80	Formaldehyde	270	835	4	5	474
81	Freon 113	260	480	4	3	24
82	Glycol ethers	43000	10000	0	1	290
83	Hexachloro-1,3-butadiene	1200	850	0	0	1490
84	Hexachlorobenzene	102	35	2	3	0.09
85	Hexachloroethane	4000	308	4	4	22
86	Hydrochloric acid	4970	10000	2	3	1
87	Hydrogen cyanide	900	277	0	1	19
88	Hidrogen flouride	4	18	0	1	1385
89	Hydroquinone	50	86	0	4	265
90	Isobutyraldehyde	320		0	1	141
91	Isopropyl alcohol	2810	6681	0	1	41
92	Lead compounds	3600	32000	5	2	8623
93	m-Xylene	1500		4	4	5
94	Maleic anhydride	5000	4550	0	2	16
95	Manganese compounds	465	1000	3	1	2963
96	Methanol	615		0	3	150
97	Methyl ethyl ketone	5628	64000	0	1	29400
98	Methyl isobutyl ketone	2737	6766	0	4	3220
99	Methyl methacrylate	2080	5672	0	2	505
100	Methyl tert-butyl ether	8000	7500	0	3	259
101	Methylenebis (phenylisocyanate)	4000	23568	0	0	786
102	Molybdenum trioxide	2200	5	0	0	66
103	N,N-Dimethylaniline	125		0	2	370
104	n-Butyl alcohol	1410	1225	0	2	65
105	Di-n-OctylPhthalate	790	8000	0	1	1860
106	N-nitrosodiphenylamine	1650		4	1	1
107	Naphtalene	2200	30	1	2	6
108	Nikel compuonds	350		5	3	27
109	Nitric acid	500	65	0	0	26
110	Nitrobenzene	640		0	2	119
111	o-Xylene	5000	4550	0	3	16
112	p-Cresol	207	50	0	0	25
113	p-Xylene	5000	4550	0	2	2
114	Phenol	530	46	0	2	34

ID	*Substance*	*LD50*	*LC50*	*NCAR*	*NTOX*	*F-LC50*
115	Phosphoric acid	1530	14	0	0	70
116	Phosphorus (yellow or white)	3		0	2	0.02
117	Phthalic anhydride	2000	1000	0	2	364
118	Picric acid	30		0	0	170
119	Polychlorinated biphenyls	1300		4	3	3
120	Propionaldehyde	1200	4581	3	1	44
121	Propylene	9999	10500	0	0	5
122	Propylene oxide	690	1740	4	5	306
123	Pyridine	1580	1000	0	1	100
124	Sec-butylalcohol	6480	8000	0	0	3670
125	Styrene	1000	2528	4	3	4
126	Sulfuric acid	2140	14	0	1	31
127	Terephthalic acid	18800		0	1	29
128	Tert-butyl alcohol	3500	8000	0	0	1954
129	Tetrachloroethylene	8100	5200	4	4	17
130	Thorium dioxide	1140		0	0	
131	Titanium tetrachloride	1000	7	0	0	25
132	Toluene	5050	6675	0	2	34
133	Toluene-2,4-diisocyanate	5800	10	4	1	53
134	Trichloroethylene	2402	8450	4	4	44
135	Vinyl acetate	1613	3680	0	1	100
136	Vinyl chloride	500	100	5	4	143
137	Vinylidene chloride	200	6350	2	3	108
138	Xylene (mixed isomers)	4300	6350	0	3	13
139	Zinc (fume or dust)	9999	1000	0	3	
140	Zinc compuonds	7950		0	0	17
141	Alachlor	1065		0	0	5
142	Atrazine	1750	540	0	0	16
143	Butylate	4659		0	0	7
144	Captan	7500	168	0	4	0.2
145	Carbaryl	500	25000	0	4	8
146	Chlorpyrifos	151		0	0	2
147	Cyanazine	261	230	0	0	18
148	1,3-Dichloropropene	140	996	4	1	0.24
149	EPTC (ethyl dipropylthiocarbante)	916	4062	0	0	27
150	Glyphosate	4873		0	0	600
151	Malathion	570	6	0	0	0.1
152	Maneb	4400		0	4	2
153	Metam Sodium (MeNHCS2Na)	285	888	0	0	0.39
154	Methyl Parathion	14	3	0	0	9
155	Metolachlor	2780		0	0	15
156	Metribuzin	7500		0	0	80
157	Terbufos (tBuSCH2SP(=S)(Oet)2	3	1	0	0	0.01
158	Trifluralin	500	47	0	3	0.11

Table 4.15a – Original data of 140 chemicals.

ID	*Substance*	*NOEL*	*BOD*	*Hydro*	*BCF*	*RWF*
1	1,1,1-Trichloroethane	7	503	30	1.5	8.94
2	1,1,2-Trichloroethane	1	503	30	1.2	3.59
3	1,2,4-Trichlorobenzene	0.2	550	1000	2.9	3.97
4	1,2,4-Trimethylbenzene	0.68	502	1000	2.3	5.4
5	1,2-Dichlorobenzene	0.05	6	1000	2.3	3.1
6	1,2-Dichloroethane	34	508	30	0.6	5.5
7	1,2-Dichloropropane	23	503	30	1.3	4.03
8	1,3-Butadiene	1	502	1000	1	5.56
9	1,4-Dichlorobenzene	3	6	1000	2.3	3.87
10	1,4-Dioxane	2588	520	1000	-1	1
11	2,4-D	6	503	1000	2.4	3.08
12	2,4-Dinitrophenol	3	550	1000	0.6	2.04
13	2,4-Dinitrotoluene	6	550	1000	1	4.37
14	2-Ethoxyethanol	4076	9	1000	-1.3	4.82
15	2-Methoxyethanol	5664	9	1000	-1.5	3.05
16	2-Nitropropane	1	550	1000	0.2	2.35
17	4,4'-Isopropyldenediphenol	0.42	8	1000	2.2	1.88
18	4,4'-Methylenedianiline	11	8	1000	0.7	1
19	4-Nitrophenol	10	550	1000	-0.1	6.07
20	Acetaldehyde	9	7	1000	-1	9.14
21	Acetone	1800	7	1000	-1	6.79
22	Acetonitiile	410	5	1	-1.1	5.31
23	Acrylamide	27	9	360	-1.4	6.76
24	Acrylic acis	47	8	1000	-0.5	6.14
25	Acrylonitrile	3	5	1	-1.6	2.09
26	Allyl chloride	18	6	2	-1	9.65
27	Alluminium (fume or dust)		500	9999		7.98
28	Ammonia	0.09	9	1000	-1.2	10.12
29	Ammonium nitrate	40	9999	9999	-2	5.15
30	Ammonium sulfate	200	9999	9999	-2	1.65
31	Aniline	27	9	1000	0	3.41
32	Anthracene	0	502	1000	3.2	2.24
33	Antimony compounds	42	9999	9999	1.6	3.92
34	Arsenic compounds	2	9999	9999	2.5	4.83
35	Asbestos (friable)		9999	9999	1	7.06
36	Barium compounds	10	9999	9999	1	4.04
37	Benzene	4	10	1000	1.2	1.65
38	Benzoyl chloride	7	9	1000	1.2	4.79
39	Biphenyl	0.12	9	1000	2.8	2.59
40	Bis(2-ethylhexyl) adipate	0.02	9	1000	4.2	2.55
41	Bromomethane	3	6	30	0.3	4.26
42	Butyl acrylate	0.31	9	50	1.4	1.85
43	Butyl benzyl phthalate	2	8	400	3.5	8.42
44	Butyraldehyde	8	7	1000	0	6.71
45	Cadmium compounds	0	9999	9999	3.5	2.9
46	Carbon disulfide	174	9999	1000	0	5.25
47	Carbon tetrachloride	5	503	30	1.8	5.4
48	Carbonyl sulfide	671	9999	1000	-0.7	7.06
49	Catechol	2	9	1000	0	6.03
50	Chlorine	0.02	9999	1	1	1.87
51	Chlorine dioxide	0.01	5	1		3.97
52	Chlorobenzene	2	6	1000	1.8	1
53	Chloroethane	4	6	30	0.5	5.28
54	Chloroform	18	503	30	1	2.61
55	Cloromethane	138	6	30	0	6.33
56	Chlorophenols [o]	3	6	1000	1.3	4.91
57	Cloroprene	0.56	503	2	0.5	5.31

ID	*Substance*	*NOEL*	*BOD*	*Hydro*	*BCF*	*RWF*
58	Chlorothalonil	0	550	1	3.7	2.44
59	Chrominum compounds	2	9999	9999	2.3	1.05
60	Cobalt compounds	0.02	9999	9999	1.7	3.92
61	Copper compounds	0.02	9999	9999	-1	1.66
62	Cresol (mixed isomers)	3	9	1000	1	3.23
63	Cumene	0.49	502	1000	2.5	1.84
64	Cumene hydroperoxide	11	9	1000	1.3	8.52
65	Cyclohexane	0.39	502	1000	2.3	1.53
66	Decabromdiphenly oxide	0	550	1000	3.9	2.81
67	Di(2-ethylhexyl) phthalate	0.08	9	400	3.6	3.42
68	Diaminotoluene (mixed isomers)	0.05	9	400	3.9	5.98
69	Dibutyl phthalate	9	550	1000	0.5	7.54
70	Dichlorobenzene (mixed isomers	0.05	9	400	3.6	7.08
71	Dichloromethane	0.05	6	1000	2.3	4.94
72	Diethanolamine	83	508	30	0.4	6.9
73	Diethyl phthalate	1178	9	1000	-2.1	7.96
74	Dimethyl phthalate	5	9	400	1.5	7.69
75	Epichlorohydrin	30	9	400	0.4	1
76	Ethylbenzene	9	510	30	-0.5	1
77	Ethylene	1	10	1000	2.1	9.73
78	Ethylene glycol	3	10	1000	0.2	5.24
79	Ethylene oxide	56909	9	1000	-2.5	6.17
80	Formaldehyde	118	510	12	-1.1	2.82
81	Freon 113	6	502	1000	0.2	3.24
82	Glycol ethers	73	503	30	0.7	5.54
83	Hexachloro-1,3-butadiene	373	9	1000	0	5.72
84	Hexachlorobenzene	0	503	2	3.6	3.03
85	Hexachloroethane	1	550	1000	3	6.16
86	Hydrochloric acid	0.13	503	30	2.2	9.33
87	Hydrogen cyanide	0.95	10	1000	0.5	8.67
88	Hidrogen flouride	346	9999	1000	-0.7	7.26
89	Hydroquinone	13	9999	1000	-0.4	5.03
90	Isobutyraldehyde	35	9	1000	-0.2	4.93
91	Isopropyl alcohol	10	7	1000	-0.2	3.56
92	Lead compounds	2156	9	1000	-0.5	3.14
93	m-Xylene	0.26	9999	9999	1.8	1.57
94	Maleic anhydride	2	10	1000	2.1	7.48
95	Manganese compounds	741	8	1000	-0.7	0.6
96	Methanol	8	9999	9999	1	4.66
97	Methyl ethyl ketone	7350	9	1000	-1.4	3.3
98	Methyl isobutyl ketone	805	7	1000	-0.5	4.41
99	Methyl methacrylate	126	7	1000	0.3	2.51
100	Methyl tert-butyl ether	65	9	1000	0.5	5.37
101	Methylenebis (phenylisocyanate)	197	508	1000	0	8.19
102	Molybdenum trioxide	6	550	1000	2.3	5.01
103	N,N-Dimethylaniline	19	9999	9999		1
104	n-Butyl alcohol	12	513	1000	1.3	3.72
105	Di-n-OctylPhthalate	465	9	1000	0	6.97
106	N-nitrosodiphenylamine	0.13	8	1000	2.1	4.55
107	Naphtalene	0.59	9	1000	2.1	3.6
108	Nikel compuonds	1	9999	9999	1.6	3.76
109	Nitric acid	1	9999	1000	-0.3	7.33
110	Nitrobenzene	30	9	1000	0.9	9.11
111	o-Xylene	2	10	1000	1.7	3.45
112	p-Cresol	6	9	1000	1	4.69
113	p-Xylene	0.2	10	1000	2.1	7.06
114	Phenol	8	9	1000	0.6	1

ID	*Substance*	*NOEL*	*BOD*	*Hydro*	*BCF*	*RWF*
115	Phosphoric acid	4	9999	1000	-0.9	0.96
116	Phosphorus (yellow or white)	0		0	1	9.36
117	Phthalic anhydride	91	550	1000	0.4	1.74
118	Picric acid	41	550	1000	1.1	7.61
119	Polychlorinated biphenyls	0.14	550	1000	4.2	5.72
120	Propionaldehyde	11	7	1000	-0.5	4.06
121	Propylene	1	10	1000	0.8	2.32
122	Propylene oxide	77	9	1000	-0.8	8.81
123	Pyridine	25	9	1000	-0.2	6.97
124	Sec-butylalcohol	918	9	1000	-0.2	8.46
125	Styrene	0.44	10	1000	1.9	1
126	Sulfuric acid	2	9999	1000	-1.3	1
127	Terephthalic acid	7	550	1000	0.7	1
128	Tert-butyl alcohol	488	508	1000	-0.5	1
129	Tetrachloroethylene	2	503	1000	1.6	1
130	Thorium dioxide		9999	9999		1
131	Titanium tetrachloride	1	9999	9999		1
132	Toluene	4	10	1000	1.7	1
133	Toluene-2,4-diisocyanate	13	550	1000	0.8	1
134	Trichloroethylene	8	503	1000	1.3	1
135	Vinyl acetate	25	9	1000	-0.1	1
136	Vinyl chloride	36	6	1000	0.6	1
137	Vinylidene chloride	27	503	1000	0.9	1
138	Xylene (mixed isomers)	1	10	1000	1.9	1
139	Zinc (fume or dust)		500	9999	-2	1
140	Zinc compuonds	0.86	9999	9999	3	1
141	Alachlor	0.51	503	2	2	5.67
142	Atrazine	3	503	1000	1.3	2.23
143	Butylate	0.54	518	1000	2.5	5.07
144	Captan	0.05	503	30	0.9	8.72
145	Carbaryl	1	8	1000	1.4	5.76
146	Chlorpyrifos	0.12	503	1000	3.7	6.67
147	Cyanazine	5	503	1	0.9	4.12
148	1,3-Dichloropropene	0.06	508	2	1	4.28
149	EPTC (ethyl dipropylthiocarbante)	3	9	1000	2.1	1
150	Glyphosate	150	9	1000	-3.7	3.98
151	Malathion	0.01	9	1000	1.8	3.18
152	Maneb	0.09				5.1
153	Metam Sodium (MeNHCS2Na)	0.1	10	1	0.1	7.44
154	Methyl Parathion	0.88	0	1000	2.1	6.55
155	Metolachlor	1	503	2	2.4	3.39
156	Metribuzin	20	508	1000	0.8	4.05
157	Terbufos (tBuSCH2SP(=S)(Oet)2	0	508	15	3.3	1
158	Trifluralin	0.01	503	30	2	1

Table 4.15b – Original data of 140 chemicals.

4.4.2 Pre-processing method comparison

Since Hasse method is very sensitive to non-discrete values, a complete evaluation by HDT requires an adequate pre-processing to establish a suitable data matrix, as well as a post-processing to correctly extract information and decisions. To reduce its sensitivity to uncertainty the data have been firstly transformed using decimal logarithms. Being the Hasse diagram based on the assumption that the higher the numerical value of a criterion, the higher is the hazard associated to that criterion, an inverse transformation has been applied to those criteria whose low values correspond to high hazard (rodent oral LD50, rodent inhalation LC50, fish LC50 and fish NOEL). Moreover, the missing data (empty cells in the matrix) have replaced with the highest in their relative column. In spite of the logarithmic transformation performed, the ranking analysis performed on the all 10 criteria provided a very complex Hasse diagram: it is a four levels and is not here shown as, being a not readable diagram is useless. To reduce much more its sensitivity to non-discrete values, Halfon and Brüggemann removed all the decimal significant digits. Notwithstanding this data simplification, the obtained Hasse diagram, shown in Figure 4.9, is still quite complex.

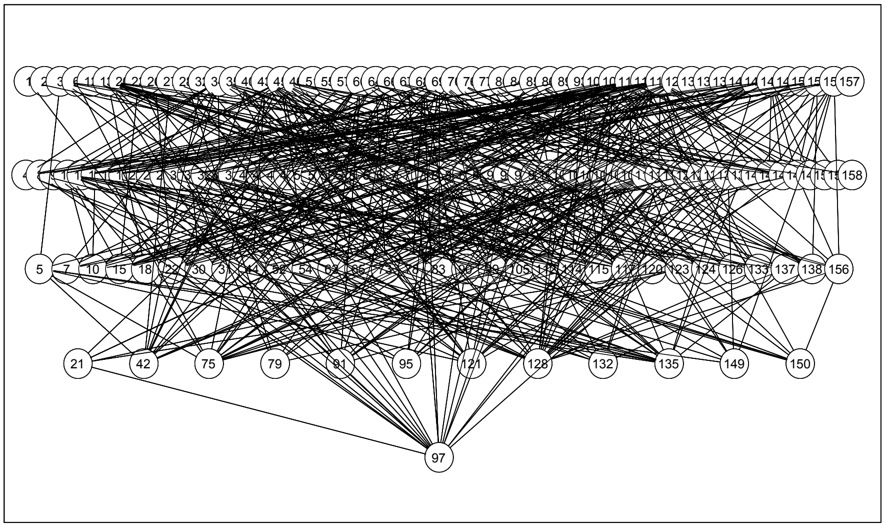


Figure 4.9 - Hasse diagram developed on log-transformed data with zero digits.

The diagram has five levels and it is not readable, since its complexity is still not resolved. In fact, it is characterized by a high number of incomparabilities (23338 over 24806 comparisons). To resolve the complexity that characterize Hasse diagrams developed on data described by many criteria, Halfon and Brüggemann proposed a classification of the data, i.e. dividing the range of each property in three equal classes, roughly corresponding to good, bad and average. When this classification is performed several chemicals take the same score and thus the obtained diagram, shown in Figure 4.10 is more sprayed.

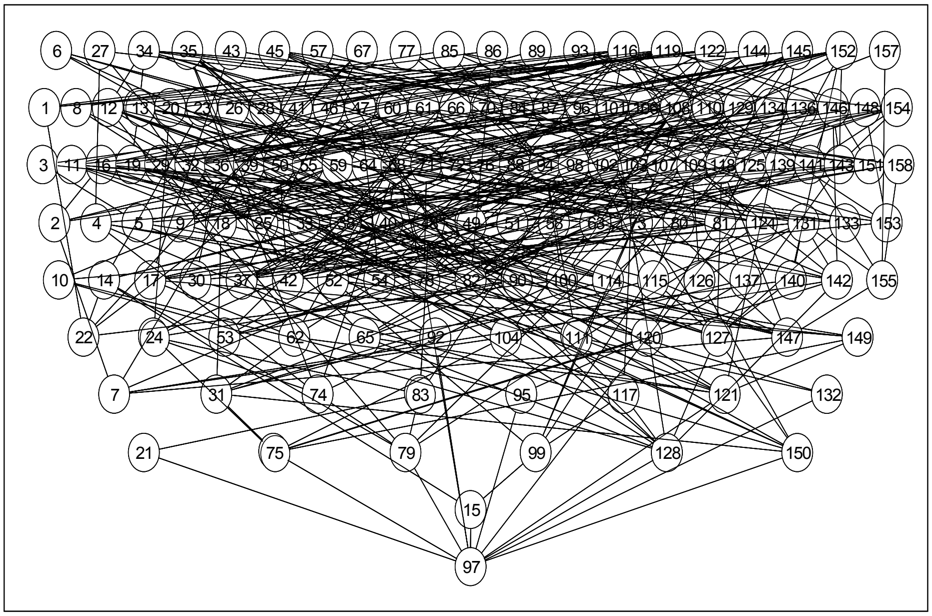


Figure 4.10 - Hasse diagram developed on classified data.

The Hasse diagram is now organised in 10 levels and as expected, the degeneracy has significantly increased, being the number of equivalence classes equal to 146. Thus, each variable has been transformed from quantitative values into hazard classes by an arbitrary classification.

As highlighted in chapter 2 a very useful pre-processing tool is the one provided by rank-order transformation, i.e. by replacing the original data with a reduced number of order statistics for example, deciles or quartiles. In such a way a high number of ties occur. Obviously the original quantitative information is lost, however when the aim is to explore the “main features” of multivariate data, or to perform a preliminary analysis before a more complete one, replacing the original data by quantiles may be useful to reveal qualitative features, which could be otherwise submerged by quantitative information. Thus, the logarithmic transformed data have been replaced by deciles, quartiles and binary data and analyzed by Hasse diagram technique. The corresponding diagrams are shown in Figure 4.11, 4.12 and 4.13 respectively.

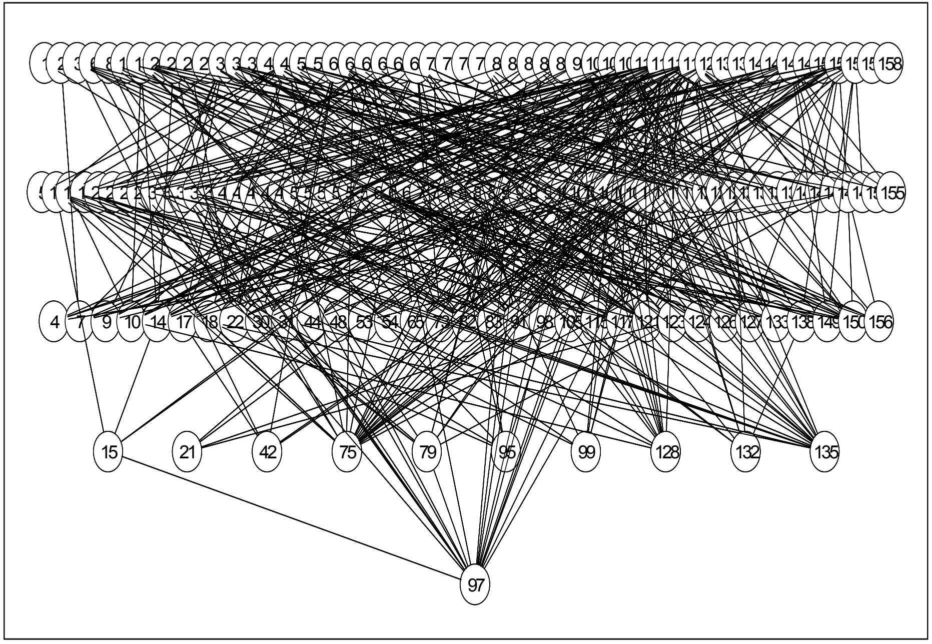


Figure 4.11 - Hasse diagram developed on deciles data.

Both deciles and quartile rank transformations provide pretty complex diagram, with five and six levels, respectively, similar to that obtained by log-transformed data with zero digits. The number of comparabilities increases from 639 for decile data up to 1201 for quartile data, however many ranking relations are not solved, being the number of incomparabilities equal to 23528 and 22406, for decile and quartile, respectively.

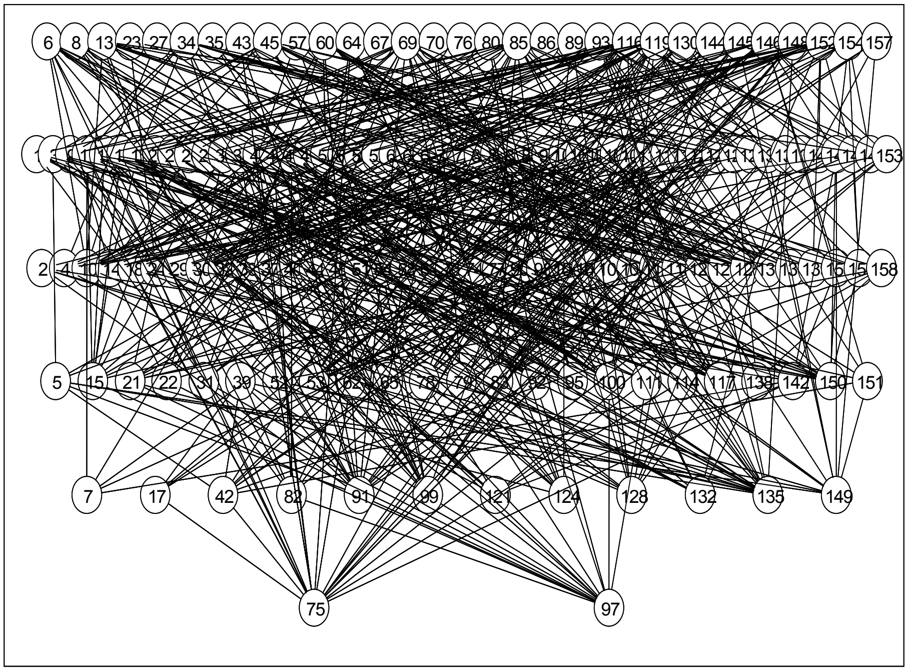


Figure 4.12 - Hasse diagram developed on quartile data.

Being the number of chemicals compared huge and the number of criteria accounted pretty high, to better understand the data and to obtain a further simplification of the Hasse diagram the binary transformation is required. The Hasse diagram on binary data (Figure 4.13) is now more readable, being organised in nine levels. It is characterised by a higher degeneracy (118 equivalence classes) with respect to the ones developed on deciles and quartiles; the number of comparabilities increases up to 3278 and the incomparabilities decrease to 18378.

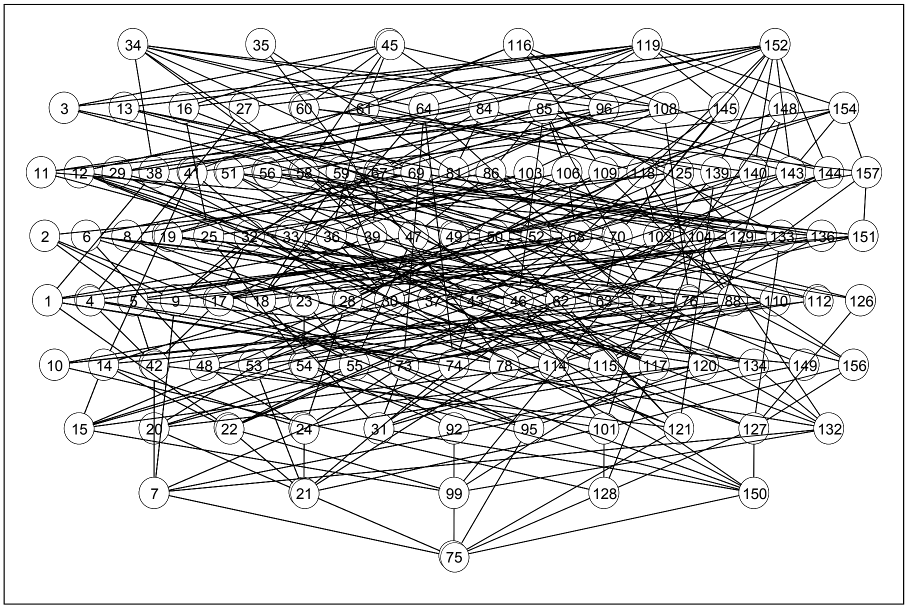


Figure 4.13 - Hasse diagram developed on binary data.

The main differences among the diagrams obtained with different kinds of rank-order transformations, in terms of number of levels, number of equivalence classes, maximals, minimals, isolated elements, comparabilities and incomparabilities are summarize in Table 4.16.

	Levels	*N.Equiv.*	*Max*	*Min*	*Iso*	*Comp.*	*Incomp.*
Logarithmic (0 digits)	5	158	45	40	8	734	23338
Hazard classes	10	146	20	8	0	2081	20672
Deciles	5	158	44	53	9	639	23528
Quartiles	6	157	31	20	0	1201	22406
Binary	9	118	7	4	0	3278	18378

Table 4.16 – Difference among Hasse diagrams.

To evaluate the quality of the ranking analysis and to better compare the diagrams obtained, ranking indices have been calculated and their values are collected in Table 4.17.

	D	*DbyR*	*T*	*div*	χ	*StR*	*P*	*Cx*	*Cx'*
Log.(0 digits)	0.00	0.06	0.03	0.33	0.06	0.91	0.94	0.94	0.12
Haz. classes	0.08	0.16	0.06	0.13	0.17	0.65	0.83	0.83	0.33
Deciles	0.00	0.05	0.03	0.33	0.05	0.92	0.95	0.95	0.10
Quartiles	0.01	0.10	0.03	0.19	0.10	0.83	0.90	0.90	0.20
Binary	0.25	0.23	0.07	0.12	0.26	0.47	0.74	0.74	0.52

Table 4.17 – Numerical values of ranking indices calculated for different rank-order transformations.

It can be highlighted that no degeneracy is provided in diagrams developed on logarithmic and decile values, whereas a little degeneracy occurs when quartile data are used and increases from the hazard class values to binary data. The discrimination power by ranking (*DbyR*) and the selectivity index (*T*) reveal that the capability of discriminating elements and to providing a unique orientation from “good” to “bad“ of the binary data diagram is pretty greater than the other diagrams, whereas the diversity (div) is lower than the other diagrams, since the number of incomparabilities is lower. Moreover, the binary data diagram provides a lower value of the stability indices (*StR* and *P*) and complexity index (*Cx*) and a higher value of the comparability degree (χ) than the diagrams resulted from the other transformations, since it is the one which better resolves ranking relations among the elements. Finally, the complexity *Cx’* providing information related to the balance between comparabilities and incomparabilities takes a higher value for the binary diagram than for the others.

Being the aim of the study to define a priority list of chemical, it is of great interest to analyse the elements selected as maximal, since these are the ones of highest hazard to be deeply investigated.

The binary diagram selects as maximals the following 7 chemicals, out of 158, which are equally hazardous, even if for different reasons:
arsenic compound (34), asbestos (35), cadmium compound (45), m-xylene (93), phosphorus (116), polychlorinated biphenyl (119) and maneb (152). Thus, even if the chemicals on the top level can not be compared with each other, these 7 chemicals can be compared with the

chemicals in the lower levels. It must be pointed out that phoshorus is identified as a priority element because it has high LD50 and its missing values for inhalation LC50 and BCF have been replaced with the maximum values. A similar reasoning should be applied on maneb, which is missing for inhalation LC50, BOD, hydrolisis and BCF. Comparing the binary results with the ones obtained by a hazard classes definition reveals that the 7 chemicals selected by binary diagram as priority chemicals are within the 20 selected as the more dangerous by the hazard classes transformation; the 13 maximal elements selected in addition to the 7 above identified are the following: 1,2-dichloroethane (6), aluminium (27), butyl benzyl phthalate (43), chloroprene (57), di(2-ethylhexyl) phthalate (67), ethylene (77), hexachloroethane (85), hydrochloric acid (86), hydroquinone (89), propylene oxide (122), captan (144), catbaryl (145), terbufos (157).

The analysis performed confirms that broad order statistics seems to be a suitable pre-processing tool to support Hasse diagram technique, since it provides a satisfactory solution both for noise and measurements errors reduction and element reduction. The ordinal relations among the elements are preserved and it does no subjective choices, like classes definitions, are required.

4.4.3 Sensitivity analysis

Since the knowledge of which criteria are important to rank the analysed chemicals is equally important as the knowledge of the ranking, a sensitivity analysis has been performed in order to find out the influence of each attribute on the ranking. The sensitivity analysis performed on the binary rank-transformed data by the backward sensitivity method proposed in chapter 2 provides the following criteria importance order:

RWF > BCF > Oral Rodent LD50 > NTOX > BOD > Fish LC50 > NCAR > Inhalation Rodent LC50 > Hydrolysis > NOEL

Thus, RWF and BCF are the more important and influent criteria, whereas NOEL does not seem very informative and exhibit a low influence in the chemical ranking. To confirm the obtained results, they have been compared with the ones obtained by Halfon and Brüggemann. The sensitivity analysis performed by Halfon and Brüggemann on the hazard classes by the W matrix method provided the following criteria importance order:

RWF > NCAR > BCF > Oral Rodent LD50 > BOD > Hydrolysis >
Inhalation Rodent LC50 > NTOX > Fish LC50 > NOEL

Obviously, some discrepancies among the sensitivity results obtained by the backward stepwise technique on binary data and the W matrix method on hazard classes data are expected. The importance of RWF obtained from the backward analysis is confirmed by the W matrix method. While RWF is still the most importance criterion, NCAR is now more important than all the other criteria. Other differences concern the importance of NTOX and Fish LC50 which are much more relevant according to the backward analysis.

This study has been presented to give evidence of the important role played by partial ranking methods in environmental decision strategies. Moreover, it has been pointed out the opportunity to support Hasse diagram analysis by adequate pre-processing statistical techniques, like broad order statistics in order to make partial ranking analysis able to process huge datasets described several criteria.

4.5 Sensory data analysis

The success and profit of a company often depends on its capability of launching a new product able to satisfy customer sensorial requirements. This is of particularly relevance in the cosmetic field, being the cosmetic product unequivocally characterised by the gratification the customer received from the product application or use. Sensory evaluation is a dynamic field concentrating on the utilization of humans for the measurement of sensory perceptions and/or their effect on product characteristics. It brings information on products regarding their perception through the 5 senses: sight, hearing, touch, taste and smell. The challenge is to understand customer perception and behaviour in order to adjust the product to meet consumer expectations. New market approaches are now available to input consumer perception in marketing decisions. Many companies have now realized the enormous creative potential for product development and scale-up, marketing insight and quality assurance their in-house descriptive sensory teams. On the other hand, companies without the time and/or resources to develop and use in-house research look to an outside tester to help facilitate internal production processes. Sensory testing companies often are called in if development time is short, quality issues have surfaced or a supplier opportunity is imminent. They then can review the data, understand and report the outcomes of individual scenarios, and offer educated suggestions on the next step to take. Descriptive testing can assess an existing category, investigate across-the-board applications, explore trend applications and/or measure the attributes behind consumer liking. Using trained or consumer panels gives researchers access to consumer reaction, both informal and structured at the very early stages of product development. Such input can be obtained easily, quickly, and at little cost. By sensory evaluation, scientists are no longer compelled to develop early product prototypes with attributes that rely heavily on their own perceptions. Early consumer input, even though it is largely qualitative and non-projectable, offers a prediction of the prototype's acceptability. Sensory research is thus an evaluative tool, the use of which enables laboratory personnel to determine product attributes that are fundamental to the development of the successful cosmetic product in a particular category. Sensory analysis make use of trained or customer panels to test new products; developing a professional descriptive panellist is a lengthy process requiring education as well as experience. Potential panellists are usually selected from approximately 100 applicants: they are trained for six to seven months and met three to four days a week to learn how to evaluate products. A trained panel should be composed by 12 – 15 persons who evaluate each product three times. Each individual evaluation is judged according to reproducibility and uniformity with respect to the other panellists. It is a common practice that the panel has to evaluate defined sensorial properties, related to the product use, appearance, fragrance…. Once performed the sensory analysis, the panel uniformity should be evaluated by objective methods able to detect the agreement/disagreement degree among the panellist, and identify panellist providing singular sensory evaluations. If the panel training can be considered satisfactory and the panel sensory evaluation reliable, the tested prototypes can be ranked and their differences analysed.

Total and partial ranking analysis has been performed on shampoo sensory data provided by a cosmetic industry with the following purposes:

1. single property perception and panel uniformity assessment
2. global judge panel analysis
3. ranking six prototypes and detecting the main differences among them

4.5.1 Shampoo sensory data

The dataset is composed by 6 shampoos, identified as A, B, C, D, E and F evaluated by 20 trained panellists. The shampoos have been tested by the panellists according to 10 properties, some of which specifically related to the shampoos, to the hair and to the shampoo general quality. The properties accounted in the analysis performed are shown in Table 4.18:

Each property has been evaluated by the panellists with a score, taking values in the range 1 (if the property is totally absent) to 10 (if the property is totally expressed).

Shampoo	*Hair*	*General*
Consistency	Comb facility before drying	Cleanliness maintenance
Rising power	Comb facility after drying	Overall agreeableness
Fragrance	Electricity	
	Volume	
	Brightness	

Table 4.18 – Shampoo properties.

Cosmetic expert considerations have been taken into account and each property has been weighted according to its importance. The properties setting and the weights are collected Table 4.19.

Property	*Function*	*Weight*
Consistency	Triangular	0.082
Rising power	Linear	0.082
Fragrance	Triangular	0.055
Comb facility before drying	Linear	0.110
Comb facility after drying	Linear	0.110
Electricity	Inverse linear	0.096
Brightness	Linear	0.082
Volume	Linear	0.109
Cleanliness maintenance	Linear	0.137
Overall agreeableness	Linear	0.137

Table 4.19 – Criterion setting.

The data matrix analysed is a three-way matrix, shown in Figure 4.14. Two axes represent the shampoos and sensory properties, the third axis the panellist votes. Thus, the resulted matrix is a 6x10x20 matrix. Three unfolding have been performed and ranking analysis has been applied on the three matrices obtained. A first unfolding has been performed in order to analyse the panel uniformity on each single property and to detect property not well perceived (Step A). Then, the global panel uniformity has been evaluated taking into account panellists overall agreement by using desirability scores (Step B). Finally, the averaged judge votes have been account in order to compare the shampoos according to the ten properties evaluated by the sensory analysis and detect their main differences (Step C).

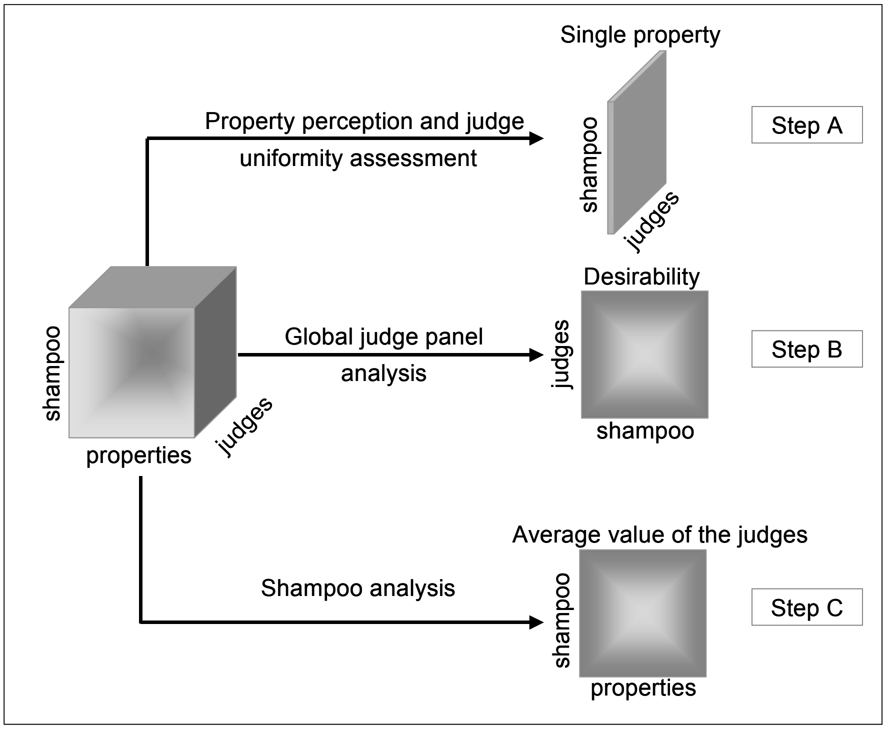


Figure 4.14 – Three-way matrix unfolding.

4.5.2 Property perception and judge uniformity assessment

Since the sensory analysis was performed by trained panellists, it was of interest to evaluate their similarity in perceiving shampoo properties. Each individual evaluation has been analysed according to its uniformity with respect to the other panellists. To find out the agreement degree among the panellists on each property accounted in the shampoo evaluation, a partial ranking analysis has been performed comparing the 20 panellists according to the way they perceived the six shampoos.

An example is here illustrated as far as concerns the Hasse diagram resulted by the analysis of the panellists consistency perception.

The panellists votes for the consistency quality of the six shampoos are collected in Table 4.20.

	Shampoo					
Panellist	*A*	*B*	*C*	*D*	*E*	*F*
1	8	7.5	5	5	6	6
2	6.5	6	3.5	3.5	4.5	4.5
3	6.5	6	3.5	3.5	4.5	4.5
4	8	7	5	5	6	6
5	8	7	5	5	6	6
6	8	7	5	5	6	6
7	8	7	5	5	6	6
8	8	7	5	5	6	6
9	8	7	5	5	6	6
10	8	7	5	5	6	6
11	8	7	5	5	6	6
12	6.5	6	3.5	3.5	4.5	4.5
13	6.5	6	3.5	3.5	4.5	4.5
14	6	6.5	3	3	4	4
15	6	6.5	3	3	4	4
16	6	6.5	3	3	4	4
17	6	6.5	3	3	4	4
18	6	6.5	3	3	4	4
19	8	7	5	5	6	6
20	4	3	1	1	2	2
Average	7	6.5	4	4	5	5

Table 4.20 – Panellist votes on shampoo consistency.

The Hasse diagram obtained is illustrated in Figure 4.15. The ranking analysis performed on the judge panel allows to highlight the degree of agreement/disagreement among the judges.

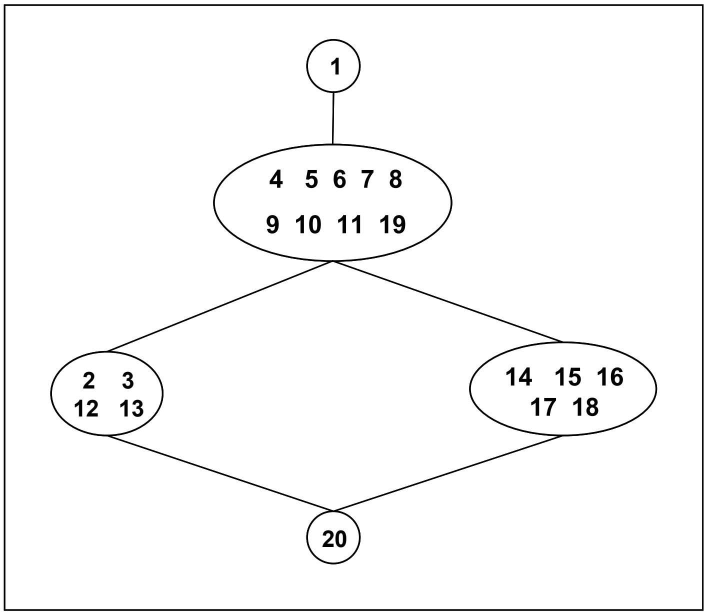
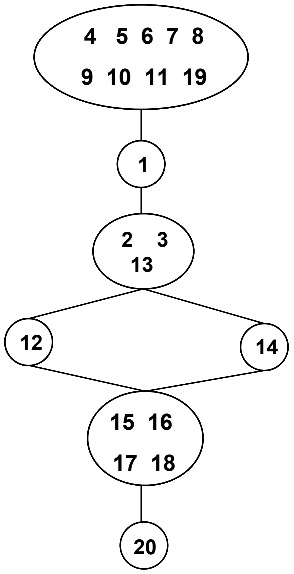
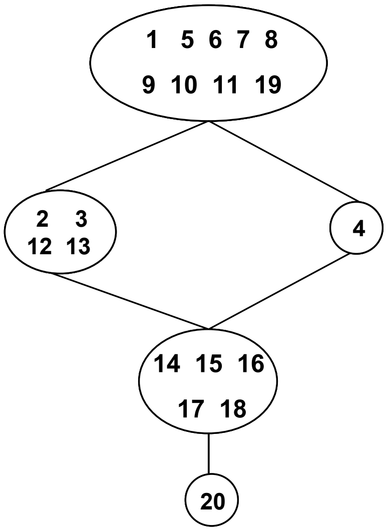


Figure 4.15 – Consistency sensory panel evaluation by Hasse diagram.

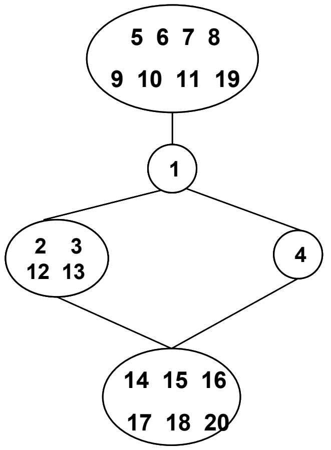
The panellists are sprayed in three main groups: panellists 4, 5, 6, 7, 8, 9, 10, 11, 19 felt the shampoo consistency tested on the six samples in a higher way than the other panellists, excepted for panellist 1 who overestimates the shampoo consistency marking more all the shampoos. A total agreement exists among panellists 2, 3, 12 and 13 as well as among panellists 14, 15, 16, 17 and 18. However the two groups are in disagreement since the panellists 2, 3, 12 and 13 felt shampoo consistency in higher way than panellists 14, 15, 16, 17 and 18 excepted for shampoo B, thus a contradictions exists among the two group votes. Panellist 20 is a minimal and he felt shampoo consistency in a fewer way than all the others. The analysis performed allows to detect the presence of two panellists who perceived the consistency sensory property in a singular way with respect to all the others: this result may suggests not to account for their votes in ranking the shampoos or to train them on the consistency property detection. Before final decision are taken, the panellists behaviour has been analysed on the other sensory properties. The Hasse diagrams developed for the other properties are collected in Table 4.21.



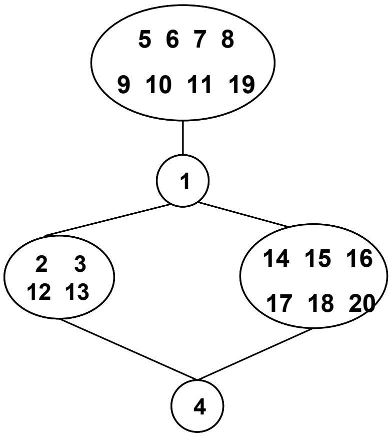
Rising power



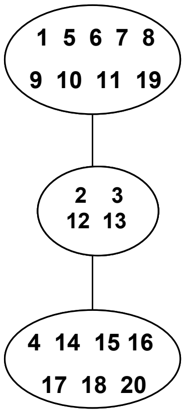
Fragrance



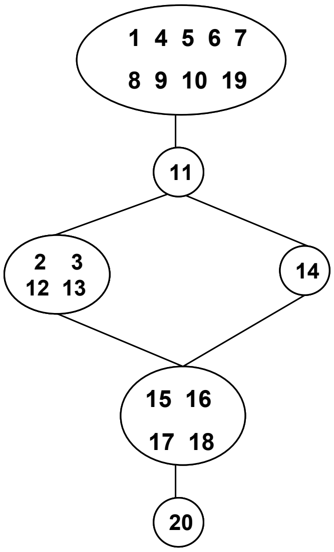
Comb facility before drying



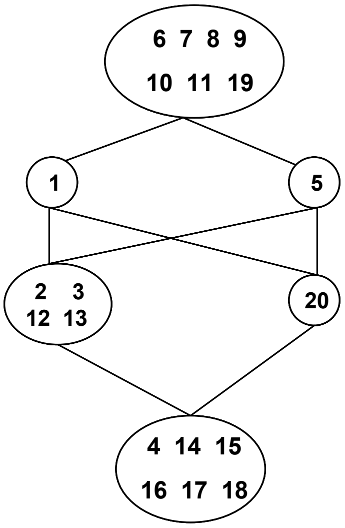
Comb facility after drying



Electricity – Cleanliness maintenance



Brightness - Volume



Overall agreeableness

Table 4.21 – Hasse diagrams of 20 panellists.

The analysis performed confirms the existence of three main groups of panellists. Panellists 6, 7, 8, 9, 10, 19 being always in agreement constitute a well defined group; they felt the sensory properties in a higher way than the panellists 15, 16, 17 and 18 who belong to another unanimous group. The group composed by panellists 2, 3, 12 and 13 is mostly located between the two previous ones: the panellists belonging to this group frequently mark lower than the first group and higher than the second group. Moreover, the three groups identified are mostly comparable, excepted for the consistency and comb facility after drying, meaning that their disagreement is mainly a *quantitative* disagreement, since the first group provided higher votes than the second and the second higher than the third. Being the aim of the analysis to evaluate the panel uniformity, a major importance should be the attached to those disagreement which are not only quantitative but also *qualitative*. This is the case of panellists 1, 4 and 20. The first one is often a singleton and shows a singular behaviour: he is pretty close to the first group (high votes) but he felt higher the consistency and lower the rising power, comb facility before and after drying and the overall agreeableness.

Panellist 4 seems quite confused and his behaviour is sometimes close to the one of the first group (consistency, rising power, brightness and volume), sometimes to the one of the second group (electricity, cleanliness maintenance and overall agreeableness) and sometimes he behaved as singleton (fragrance, comb facility before and after drying). Panellist 20 is frequently a minimal since he felt shampoo properties in a fewer way than all the others.

4.5.3 Global judge panel analysis

To have a further confirm of the obtained results, the panel has been analysed taking into account all the properties at the same time. Each panellist has been characterised by the overall quality (desirability score) he assigned to each shampoo. The matrix processed is the one of Table 4.22, where each value is the desirability score calculated on all the weighted properties. The corresponding Hasse diagram is shown in Figure 4.16.

Even if the ranking has been now provided by desirability scores, which aggregated the panellists behaviour on all the sensory properties evaluated, the previous results are pretty confirmed.

	Shampoo					
Panellist	*A*	*B*	*C*	*D*	*E*	*F*
1	0.53	0.61	0.79	0.78	0.83	0.85
2	0.50	0.57	0.65	0.66	0.76	0.78
3	0.50	0.57	0.65	0.66	0.76	0.78
4	0.43	0.51	0.67	0.66	0.76	0.79
5	0.53	0.62	0.79	0.79	0.87	0.84
6	0.53	0.62	0.79	0.79	0.87	0.84
7	0.53	0.62	0.79	0.79	0.87	0.84
8	0.53	0.62	0.79	0.79	0.87	0.84
9	0.53	0.62	0.79	0.79	0.87	0.84
10	0.53	0.62	0.79	0.79	0.87	0.84
11	0.53	0.62	0.79	0.79	0.86	0.84
12	0.50	0.57	0.65	0.66	0.76	0.78
13	0.50	0.57	0.65	0.66	0.76	0.78
14	0.48	0.53	0.60	0.61	0.74	0.74
15	0.48	0.53	0.60	0.61	0.73	0.74
16	0.48	0.53	0.60	0.61	0.73	0.74
17	0.48	0.53	0.60	0.61	0.73	0.74
18	0.48	0.53	0.60	0.61	0.73	0.74
19	0.53	0.62	0.79	0.79	0.87	0.84
20	0.42	0.47	0.49	0.51	0.65	0.66

Table 4.22 – Panellist desirability scores on shampoo.

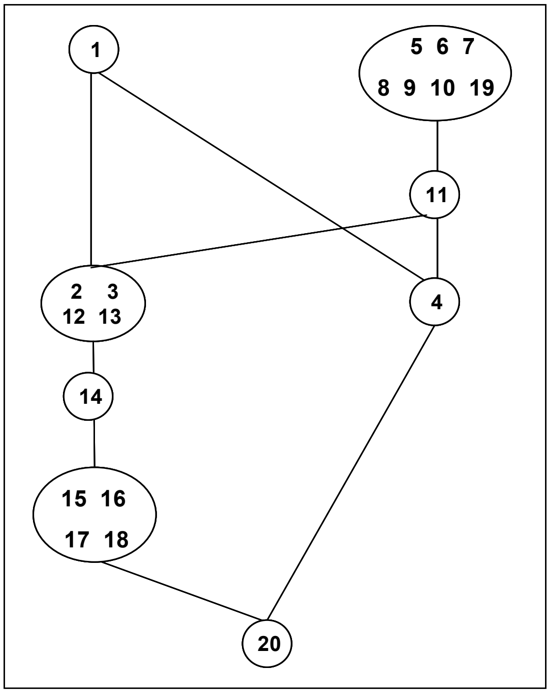


Figure 4.16 – Hasse diagram of 20 panellists based on desirability values of 6 shampoos.

The partial ranking analysis performed highlighted a quite good sensory panel characterised by a satisfactory agreement degree excepted for panellists 1, 4 and 20. Being the one analysed a trained panel, the obtained results suggested a further training of panellists 1, 4 and 20.

4.5.4 Shampoo analysis

Sensory analysis is often used by industries in the first phases of the product developing process to reduce the number of prototypes and to improve the optimisation process. Thus, it is of major importance ranking the samples according to their overall quality.

A total and partial ranking analysis has been performed on the 6 shampoos in order to compare them and detect their differences on 10 properties. The processed matrix, obtained by averaging the panel votes is shown in Table 4.23.

ID	*Cons*	*RP*	*Fr*	*Comb b*	*Comb a*	*Elect.*	*Bright.*	*V*	*Maint.*	*Agreeb.*
A	7	6	8.5	5	6	5	6	5	5	5
B	6.5	7	8	6	6	4.5	6	5.5	5	6
C	4	8	3	7	8	4	7	6	7	7
D	4	7	3	8	7	3.5	8	7	6	7
E	5	8	4	9	8	2	8	8	7	8
F	5	9	6	8	8	3	7	8	8	9

Table 4.23 - Shampoo average values on 10 properties: consistency (Cons), rising power (RP), fragrance (Fr), comb facility before drying (Comb b), comb facility before drying (Comb a), electricity (Elect.), brightness (Bright), volume (V), cleanliness maintenance (Maint.) and overall agreeableness (Agreeb.)

To perform ranking analysis of the six shampoos, it should be explicated whether the best condition was satisfied with a minimum value or a maximum value of the property and the trend from the minimum to the maximum.

The overall quality of the shampoos have been calculated by Desirability, Utility and Dominance functions; the obtained results are shown in Table 4.24 and the corresponding histograms in Figure 4.17.

Shampoo	*Desirability*	*Utility*	*Dominance*
F	0.813	0.819	0.784
E	0.810	0.815	0.810
D	0.702	0.707	0.483
C	0.700	0.705	0.482
B	0.578	0.583	0.155
A	0.521	0.528	0.014

Table 4.24 –Desirability, Utility and Dominance values calculated on 10 sensory properties (sorted according to desirability values).

The analysis performed points out that shampoo F and E show the highest overall quality: they are the two prototypes which mostly satisfy the panellist sensory requirements. They are followed by shampoo D and C, which provide pretty good results as well, while shampoo B and A are the less satisfactory. Thus if the required overall quality score was equal to 0.7, shampoo E, F, D and C can be considered acceptable, while shampoo A and B has to be significantly modify.

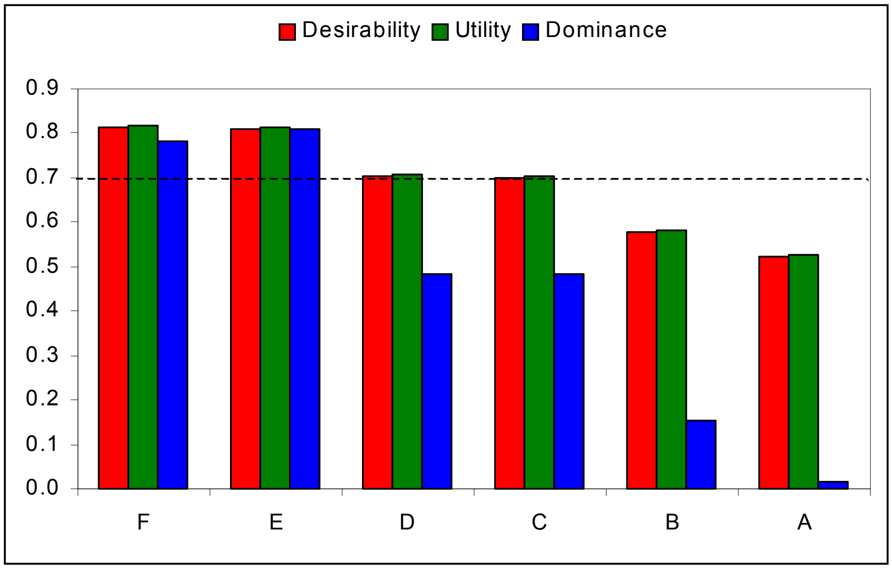


Figure 4.17 –Desirability, Utility and Dominance histograms.

Since a total order ranking approach provide a totally ordered sequence of the sample, a partial ranking analysis has been performed in order to highlight differences among the shampoos and highlight comparabilities and incomparabilities of them, if they occur.

The Hasse diagram technique has been applied on the 6 shampoos and the corresponding diagram is illustrated in Figure 4.18.

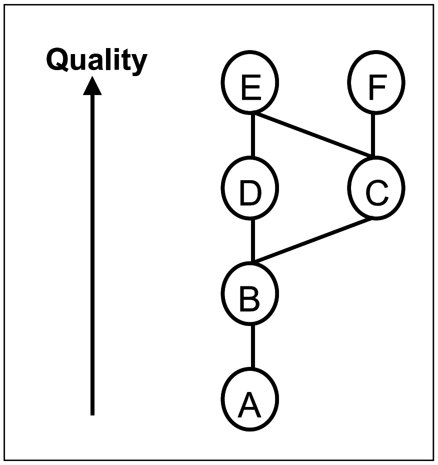


Figure 4.18 – Hasse diagram of 6 shampoo.

The Hasse diagram is arranged on four levels. Shampoo E and F are of equal overall quality and their incomparability is due to small discrepancy on the sensory properties. Shampoo F is better than E as far as concerns its rising power, its capability of cleanliness maintenance and its overall agreeableness; however it is worse than E as far as concerns the comb facility before drying, the hair electricity and brightness.

4.6 Total order ranking QSPR model for physico-chemical properties of polychlorinated biphenyls (PCBs).

PCBs are widespread contaminants in the environment primarily in soil and freshwater systems. Their persistence in the environment has been a growing concern due to their low degradability, toxicity, mutagenicity and because of their tendency to bioaccumulate. Polychlorinated biphenyls (PCBs) are a group of industrial chemicals that share a common structure. PCBs were commercially produced as complex mixtures beginning in 1929 and are not known to occur naturally in the environment. Principle uses of PCBs include uses in transformers, capacitors, printing inks, paints, dedusting agents, pesticides, plasticisers, lubricant inks, paint additives etc., and they were marketed for these uses. PCBs are an environmental hazard due to their inability to degrade in the environment. PCBs are non-polar compounds. Their non-polar nature makes them only slightly soluble in water. The solubility of PCBs is also influenced by the environment as these compounds or preparations show a strong affinity for sediment and organic fractions. Owing to their low solubility's in water, PCBs are often associated with the solid fraction of the aquatic and terrestrial environments. They are highly lipophilic leading to their profound persistence and ability to bioaccumulate. The sorption reactions of PCBs in aquatic and terrestrial systems play an important role in determining their fate and transport in the environment. Due to the need to know the PCBs behaviour in the environment it is very useful to have a good knowledge of their physico-chemical properties, like solubility, aqueous activity coefficient and octanol –water partitioning coefficient. Owing to the relevant importance of these chemicals in th environment many studies have been perfomed searching for quantitative structure – property relationships (QSPR) [Gramatica *et al*., 1998]. In the present study a total ranking model for three physico-chemical properties of PCBs is illustrated, the aim to provide a list of priority PCBs according to their environmental impact, contemporary accounting for their solubility, aqueous activity coefficient and lipophilicity. Despite classical multilinear regressions models, the total order ranking model does not require any assumptions about distribution properties and it allows a multiresponses (multiproperties) modelling.

4.6.1 PCB experimental data

The solubilities and the logK$_{ow}$ data have been taken from Patil [Patil, 1991], while aqueous activity coefficient values have been taken from Myrdal [Myrdal *et al.*, 1992]. The experimental values of the three properties analysed were available only for 64 PCBs over the 209. The 64 PCBs analysed together with their experimental values are collected in Table 4.25.

ID	*Name*	*- Log Sw (mol/l)*	*- Log Yw*	*Log Kow*
1	3-	5.39	4.94	4.66
2	4-	5.33	4.78	4.63
3	2,2'-	5.72	5.06	4.72
4	2,3'-	5.26	5.59	4.84
5	2,4-	5.56	5.36	5.15
6	2,4'-	5.46	5.40	5.09
7	3,3'-	6.45	5.76	5.27
8	4,4'-	6.37	5.65	5.23
9	2,2',3-	6.10	5.91	5.12
10	2,2',5-	6.17	5.84	5.33
11	2,2',6-	5.90	5.15	5.04
12	2,3,4'-	5.80	5.82	5.29
13	2,3,6-	6.49	6.05	5.44
14	2,3',5-	6.14	5.98	5.65
15	2,4,4'-	6.22	6.00	5.71
16	2,4',5-	6.18	5.98	5.68
17	2,3',4'-	6.21	5.94	5.71
18	2,2',3,3'-	6.83	6.08	5.67
19	2,2',3,4'-	6.96	6.25	5.72
20	2,2',3,5'-	6.91	6.28	5.73
21	2,2',3,6'-	6.30	5.44	4.84
22	2,2',4,4'-	7.23	6.43	5.94
23	2,2',4,5-	6.86	6.68	5.69
24	2,2',4,6-	6.94	5.46	5.75
25	2,2',5,5'-	7.00	6.20	5.79
26	2,2',5,6'-	6.65	5.85	5.55
27	2,2',6,6'-	6.20	6.33	5.24
28	2,3,4,4'-	6.86	5.72	6.24
29	2,3,4',5-	6.77	6.26	6.10
30	2,3',4,4'-	6.63	5.90	5.98
31	2,3',4,6-	7.26	6.94	6.03

ID	*Name*	*- Log Sw (mol/l)*	*- Log Yw*	*Log Kow*
32	2,3',4',5-	6.69	6.21	6.22
33	2,4,4',5-	6.77	6.02	6.10
34	2,4,4',6-	7.26	6.15	6.03
35	2,3',4',5'-	6.71	6.06	5.98
36	2,2',3,4,5-	7.87	6.32	6.38
37	2,2',3,4,5'-	7.66	6.53	6.23
38	2,2',3,4,6-	7.92	7.06	6.50
39	2,2',3,5',6-	7.19	6.10	5.92
40	2,2',3,4',5'-	7.76	6.51	6.30
41	2,2',4,4',6-	7.66	6.94	6.23
42	2,3,3',4',6-	7.65	7.05	6.20
43	2,3,4,4',5-	7.50	6.59	6.71
44	2,3',4,4',5-	7.33	6.55	6.57
45	2,2',3,3',4,5-	8.42	7.15	6.76
46	2,2',3,3',5,6-	7.65	7.08	6.20
47	2,2',3,3',5,6'-	7.82	6.70	6.32
48	2,2',3,4,4',5-	8.52	7.12	6.82
49	2,2',3,4,4',5'-	8.38	7.14	6.73
50	2,2',3,4,5,5'-	8.42	7.09	6.75
51	2,2',3,5,5',6-	7.93	6.68	6.42
52	2,2',4,4',5,5'-	8.49	7.50	6.80
53	2,2',4,4',6,6'-	8.12	7.88	6.54
54	2,3,3',4,4',5-	8.31	6.81	7.44
55	2,3,3',4,4',6-	8.48	6.84	6.78
56	2,3,3',4',5,6-	8.48	7.21	6.78
57	2,2',3,3',4,4',5-	8.90	6.97	7.08
58	2,2',3,3',4,5,6'-	8.59	6.84	6.85
59	2,2',3,4,4',5,5'-	9.10	7.15	7.21
60	2,2',3,4,4',5',6-	8.85	7.33	7.04
61	2,2',3,4,5,5',6-	8.75	7.10	6.99
62	2,2',3,3',4,4',5,5'-	9.70	8.23	7.62
63	2,2',3,3',4,4',5,5',6-	10.18	8.21	7.94
64	2,2',3,3',4,4',5,5',6,6'-	10.89	8.75	8.20

Table 4.25 – Physico-chemical properties of 64 polychlorinated biphenyls.

4.6.2 Molecular descriptors

The chemical structures of the PCBs 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 [Todeschini *et.al*., 2003] on the basis of the minimum energy molecular geometries optimized by *HyperChem* package [HYPERCHEM, 1995] (PM3 semiempirical method). In this study the following sets of molecular descriptors have been calculated: constitutional descriptors, topological descriptors [Bonchev, 1983; Devillers and Balaban, 2000], walk and path counts, connectivity indices [Kier and Hall, 1986], information indices, Moreau-Broto 2D-autocorrelations [Moreau and Broto, 1980a; 1980b; Broto *et al*., 1984], edge adjacency indices [Estrada, 1995], BCUT descriptors [Pearlman and Smith, 1998; Pearlman, 1999], topological charge indices [Galvez *et al*., 1994; 1995], eigenvalue based indices [Balaban *et al*., 1991], Randic molecular profiles [Randic 1995, 1996], geometrical descriptors, radial distribution function descriptors [Hemmer *et al*., 1999], 3D-MoRSE descriptors [Schuur, and Gasteiger, 1996, 1997], WHIM descriptors [Todeschini *et al*., 1994; Todeschini and Gramatica, 1997], GETAWAY descriptors [Consonni *et al.*, 2002], 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* of Todeschini and Consonni [Todeschini and Consonni, 2000].

4.6.3 Experimental ranking

The experimental ranking of the 64 PCBs was obtained by the desirability functions method. A linear transformation has been applied on the three properties, equally weighted, with the aim of providing a total rank of the chemicals according to their impact on the environment. Thus, the overall environmental impact has been calculated combining all the desirabilities through a geometrical mean. Once calculated the *D* for each chemical, all the PCBs have been ranked according to their *D* value and the PCBs with a higher *D* are identified as the ones with a highest impact on the environment. The PCBs experimental ranking, is illustrated in Table 4.26 with the corresponding environmental impact score.

ID	*Name*	*Envirn. Impact Score*
64	2,2',3,3',4,4',5,5',6,6'-	1.000
63	2,2',3,3',4,4',5,5',6-	0.888
62	2,2',3,3',4,4',5,5'-	0.831
59	2,2',3,4,4',5,5'-	0.665
60	2,2',3,4,4',5',6-	0.652
57	2,2',3,3',4,4',5-	0.626
52	2,2',4,4',5,5'-	0.621
61	2,2',3,4,5,5',6-	0.621
54	2,3,3',4,4',5-	0.602
53	2,2',4,4',6,6'-	0.597
56	2,3,3',4',5,6-	0.595
48	2,2',3,4,4',5-	0.594
45	2,2',3,3',4,5-	0.585
49	2,2',3,4,4',5'-	0.579
50	2,2',3,4,5,5'-	0.579
58	2,2',3,3',4,5,6'-	0.576
55	2,3,3',4,4',6-	0.564
38	2,2',3,4,6-	0.522
51	2,2',3,5,5',6-	0.485
46	2,2',3,3',5,6-	0.477
42	2,3,3',4',6-	0.475
43	2,3,4,4',5-	0.473
41	2,2',4,4',6-	0.471
47	2,2',3,3',5,6'-	0.471
40	2,2',3,4',5'-	0.449
44	2,3',4,4',5-	0.447
36	2,2',3,4,5-	0.445
37	2,2',3,4,5'-	0.439
31	2,3',4,6-	0.424
22	2,2',4,4'-	0.377
34	2,4,4',6-	0.364
29	2,3,4',5-	0.346
39	2,2',3,5',6-	0.346
32	2,3',4',5-	0.344
23	2,2',4,5-	0.343

ID	*Name*	*Envirn. Impact Score*
25	2,2',5,5'-	0.330
33	2,4,4',5-	0.326
19	2,2',3,4'-	0.325
20	2,2',3,5'-	0.325
35	2,3',4',5'-	0.316
28	2,3,4,4'-	0.312
18	2,2',3,3'-	0.299
30	2,3',4,4'-	0.296
26	2,2',5,6'-	0.258
24	2,2',4,6-	0.253
13	2,3,6-	0.252
15	2,4,4'-	0.252
17	2,3',4'-	0.247
16	2,4',5-	0.244
14	2,3',5-	0.238
27	2,2',6,6'-	0.224
7	3,3'-	0.211
10	2,2',5-	0.204
8	4,4'-	0.194
9	2,2',3-	0.180
12	2,3,4'-	0.167
21	2,2',3,6'-	0.122
11	2,2',6-	0.107
5	2,4-	0.105
6	2,4'-	0.090
3	2,2'-	0.053
1	3-	0.020
2	4-	0.000
4	2,3'-	0.000

Table 4.26 – Environmental impact by desirability function of 64 polychlorinated biphenyls.

The experimental ranking highliths that the environmetal impact of PCBs is strictly correlated to their degree of chlorination, since their sorption increases with the degree of chlorination, and their solubility and water activity decreases with the increasing number of chlorine atoms.

4.6.4 Model ranking

The correlations between the environmetal impact of the PCBs and the molecular descriptors have been estimated by the desirability and utility methods. However because of the extremely high number variables, 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 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 Spearman's rank correlation coefficient ($r_{exp-mod}$) has been used as optimization parameters in the genetic evolution algorithm to quantify the correlation between the total experimental ranking and the total model ranking. All of the calculations have been performed by the in-house software *RANA* for variable selection for WINDOWS/PC [Todeschini, *et al*. 2003].

The best models are collected in Table 4.27.

Size	*Method*	*Variables*	$r_{exp-mod}$
3	Utility	S3K piPC06 BELm2*	98.13
2	Utility	X1A* BELm2*	97.93

Table 4.27 – Best total ranking models: star indicates an inverse transformation.

A very good result is provided by simple utility model, made of two variables: the average connectivity index (X1A) and a Burden descriptor (BELm2). The first one is a topological descriptor calculated from the vertex degree of the atoms in the H-depleted molecular graph while the second one is the second lowest eigenvalue of the Burden matrix weighted by atomic masses. Both the descriptors are inversely correlated to the environmental impact of the PCBs, as their values decrease with the increasing size of the PCBs. The correlation between experimental and model ranking is pretty high (97.93). The model ranking is illustrated in Table 4.28.

ID	*Name*
64	2,2',3,3',4,4',5,5',6,6'-
63	2,2',3,3',4,4',5,5',6-
62	2,2',3,3',4,4',5,5'-
59	2,2',3,4,4',5,5'-
61	2,2',3,4,5,5',6-
60	2,2',3,4,4',5',6-
57	2,2',3,3',4,4',5-
58	2,2',3,3',4,5,6'-
52	2,2',4,4',5,5'-
54	2,3,3',4,4',5-
55	2,3,3',4,4',6-
48	2,2',3,4,4',5-
56	2,3,3',4',5,6-
50	2,2',3,4,5,5'-
53	2,2',4,4',6,6'-
49	2,2',3,4,4',5'-
51	2,2',3,5,5',6-
45	2,2',3,3',4,5-
47	2,2',3,3',5,6'-
46	2,2',3,3',5,6-
44	2,3',4,4',5-
43	2,3,4,4',5-
41	2,2',4,4',6-
40	2,2',3,4',5'-
36	2,2',3,4,5-
42	2,3,3',4',6-
37	2,2',3,4,5'-
38	2,2',3,4,6-
39	2,2',3,5',6-
33	2,4,4',5-
29	2,3,4',5-
34	2,4,4',6-
31	2,3',4,6-
30	2,3',4,4'-
32	2,3',4',5-
22	2,2',4,4'-
28	2,3,4,4'-
25	2,2',5,5'-
35	2,3',4',5'-
23	2,2',4,5-

ID	*Name*
24	2,2',4,6-
19	2,2',3,4'-
20	2,2',3,5'-
26	2,2',5,6'-
18	2,2',3,3'-
15	2,4,4'-
16	2,4',5-
21	2,2',3,6'-
14	2,3',5-
27	2,2',6,6'-
12	2,3,4'-
17	2,3',4'-
10	2,2',5-
13	2,3,6-
8	4,4'-
9	2,2',3-
11	2,2',6-
7	3,3'-
5	2,4-
6	2,4'-
4	2,3'-
3	2,2'-
2	4-
1	3-

Table 4.28 – Model experimental ranking calculated by X1A and BELm2 utility function.

4.6.5 Interval estimation

The experimental ranking of each PCB has been estimated by the obtained ranking model, according to the procedure described in Chapter3: the experimental ranking of any chemical has been estimated looking for those two elements located at the shortest path and which experimental value difference constitutes the smallest positive interval. Then, 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 property at a time, for each chemical the standardised disagreement $\delta_{i\,r}$ between its experimentally derived interval and model-calculated interval has been calculated. The experimentally derived intervals and the calculated intervals for solubility (-*Log*S_w), aqueous coefficient (-*LogYw*) and hydrophobicity (LogK_{ow}), together with the corresponding standardise disagreements and *experimental uncertainties Ry* are illustrated in Table 4.29, 4.30 and 4.31, respectively.

Response: -LogS_w		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	$\delta_{\text{-LogSw}}$	*Ry*
1	3-	5.26	5.72	-	< 5.33	0.07	-
2	4-	-	< 5.39	5.39	5.72	0.06	0.06
3	2,2'-	5.39	5.46	5.33	5.46	0.01	0.02
4	2,3'-	-	< 5.39	5.33	5.46	0.01	0.02
5	2,4-	5.46	5.90	5.46	6.45	0.10	0.18
6	2,4'-	5.39	5.56	5.26	5.56	0.02	0.05
7	3,3'-	6.17	6.20	5.56	5.90	0.11	0.06
8	4,4'-	6.10	6.17	6.10	6.49	0.06	0.07
9	2,2',3-	5.80	6.37	5.90	6.37	0.02	0.08
10	2,2',5-	6.37	6.45	6.10	6.21	0.06	0.02
11	2,2',6-	5.56	6.30	5.56	6.10	0.04	0.10
12	2,3,4'-	6.30	6.37	6.17	6.20	0.04	0.01
13	2,3,6-	6.21	6.94	6.10	6.17	0.15	0.01
14	2,3',5-	6.20	6.21	6.20	6.30	0.02	0.02
15	2,4,4'-	6.21	6.94	6.18	6.83	0.02	0.12
16	2,4',5-	6.14	6.21	6.14	6.22	0.00	0.01
17	2,3',4'-	6.18	6.22	6.17	6.20	0.01	0.01
18	2,2',3,3'-	6.63	6.86	6.22	6.65	0.11	0.08
19	2,2',3,4'-	6.71	6.77	6.91	6.94	0.04	0.01
20	2,2',3,5'-	6.71	6.77	6.65	6.96	0.04	0.06
21	2,2',3,6'-	5.90	6.10	6.14	6.18	0.05	0.01
22	2,2',4,4'-	7.26	7.26	6.86	7.26	0.07	0.07
23	2,2',4,5-	7.00	7.19	6.94	7.00	0.04	0.01
24	2,2',4,6-	6.22	6.65	6.65	6.86	0.11	0.04
25	2,2',5,5'-	6.77	6.86	6.71	6.86	0.01	0.03
26	2,2',5,6'-	6.22	6.63	6.83	6.91	0.12	0.01
27	2,2',6,6'-	6.17	6.18	5.80	6.14	0.07	0.06

Response: -LogS$_w$		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	$\delta_{\text{-LogSw}}$	*Ry*
28	2,3,4,4'-	6.63	6.71	7.00	7.23	0.11	0.04
29	2,3,4',5-	6.69	7.26	6.63	6.77	0.10	0.02
30	2,3',4,4'-	6.65	6.83	6.69	7.26	0.08	0.10
31	2,3',4,6-	7.23	7.66	6.63	7.26	0.18	0.11
32	2,3',4',5-	6.86	7.19	7.23	7.26	0.07	0.01
33	2,4,4',5-	6.91	7.00	6.77	7.19	0.06	0.07
34	2,4,4',6-	6.77	7.23	6.63	6.77	0.11	0.02
35	2,3',4',5'-	6.86	6.91	6.86	7.00	0.02	0.02
36	2,2',3,4,5-	7.26	7.33	7.65	7.76	0.09	0.02
37	2,2',3,4,5'-	7.26	7.87	7.19	7.65	0.05	0.08
38	2,2',3,4,6-	7.93	8.48	7.19	7.65	0.23	0.08
39	2,2',3,5',6-	6.69	7.26	6.77	7.92	0.13	0.20
40	2,2',3,4',5'-	7.33	7.66	7.65	7.66	0.06	0.00
41	2,2',4,4',6-	7.76	7.82	7.65	7.65	0.03	0.00
42	2,3,3',4',6-	7.50	7.65	7.66	7.87	0.07	0.04
43	2,3,4,4',5-	7.66	7.93	7.66	7.82	0.02	0.03
44	2,3',4,4',5-	7.66	7.76	7.50	7.65	0.05	0.03
45	2,2',3,3',4,5-	8.38	8.52	7.82	7.93	0.12	0.02
46	2,2',3,3',5,6-	7.65	7.93	7.33	7.82	0.08	0.09
47	2,2',3,3',5,6'-	7.76	7.82	7.65	8.42	0.13	0.14
48	2,2',3,4,4',5-	8.42	8.48	8.48	8.48	0.01	0.00
49	2,2',3,4,4',5'-	8.48	8.52	7.93	8.12	0.10	0.03
50	2,2',3,4,5,5'-	8.48	8.52	8.12	8.48	0.07	0.06
51	2,2',3,5,5',6-	7.65	7.92	7.82	8.38	0.11	0.10
52	2,2',4,4',5,5'-	8.31	8.90	8.31	8.59	0.06	0.05
53	2,2',4,4',6,6'-	8.48	8.49	8.38	8.42	0.02	0.01
54	2,3,3',4,4',5-	8.12	8.49	8.48	8.49	0.06	0.00
55	2,3,3',4,4',6-	7.92	8.59	8.48	8.49	0.12	0.00
56	2,3,3',4',5,6-	8.42	8.49	8.42	8.52	0.01	0.02
57	2,2',3,3',4,4',5-	8.49	8.85	8.59	8.85	0.02	0.05
58	2,2',3,3',4,5,6'-	7.92	8.38	8.49	8.90	0.17	0.07
59	2,2',3,4,4',5,5'-	8.85	9.70	8.75	9.70	0.02	0.17
60	2,2',3,4,4',5',6-	8.90	9.10	8.90	9.10	0.00	0.04
61	2,2',3,4,5,5',6-	8.49	8.90	8.85	9.10	0.10	0.04
62	2,2',3,3',4,4',5,5'-	9.10	10.18	9.10	10.18	0.00	0.19
63	2,2',3,3',4,4',5,5',6-	9.70	10.89	9.70	10.89	0.00	0.21
64	2,2',3,3',4,4',5,5',6,6'-	>10.18	-	> 10.18	-	0.00	-

Table 4.29 – Experimental solubility interval estimation.

By comparing the experimentally derived intervals with the calculated ones, the average disagreement turned out to be pretty low ($\bar{\delta}_{-\mathrm{LogS_w}} = 0.06$). Moreover, the average disagreement between the quantitative experimental values and their derived intervals calculated is equal to $\tilde{\delta}_{-\mathrm{LogS_w}} = 0.06$. The model quality for the solubility evaluated by complement of the average disagreement between experimental and calculated intervals is satisfactory ($Q_{-\mathrm{LogS_w}} = 0.94$). In addition, the standard deviation error has been calculated: SDE = 0.263.

Response: -LogY$_w$		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	$\delta_{\text{-LogYw}}$	*Ry*
1	3-	4.78	5.06	-	< 4.78	0.07	-
2	4-	-	< 4.94	4.94	5.06	0.03	0.03
3	2,2'-	4.94	5.40	4.78	5.59	0.09	0.20
4	2,3'-	-	< 4.94	5.06	5.40	0.12	0.09
5	2,4-	5.40	5.44	5.40	5.76	0.08	0.09
6	2,4'-	5.06	5.36	5.59	5.76	0.18	0.04
7	3,3'-	5.84	6.33	5.36	5.91	0.23	0.14
8	4,4'-	5.82	5.84	5.91	6.05	0.06	0.04
9	2,2',3-	5.82	5.84	5.15	5.65	0.17	0.13
10	2,2',5-	5.65	5.76	5.65	5.94	0.05	0.07
11	2,2',6-	5.36	5.44	5.76	5.91	0.14	0.04
12	2,3,4'-	5.44	5.91	5.94	6.33	0.22	0.10
13	2,3,6-	5.94	6.00	5.65	5.84	0.09	0.05
14	2,3',5-	5.76	5.98	5.82	5.98	0.02	0.04
15	2,4,4'-	5.94	6.00	5.98	6.08	0.03	0.03
16	2,4',5-	5.98	6.05	5.44	6.00	0.15	0.14
17	2,3',4'-	5.98	6.00	5.84	6.33	0.12	0.12
18	2,2',3,3'-	5.90	6.06	6.00	6.28	0.08	0.07
19	2,2',3,4'-	6.06	6.28	6.28	6.68	0.16	0.10
20	2,2',3,5'-	6.06	6.25	5.85	6.25	0.05	0.10
21	2,2',3,6'-	5.15	5.82	5.98	5.98	0.21	0.00
22	2,2',4,4'-	6.15	6.94	5.72	6.21	0.29	0.12
23	2,2',4,5-	6.20	6.21	5.46	6.06	0.19	0.15
24	2,2',4,6-	6.00	6.08	6.25	6.68	0.17	0.11
25	2,2',5,5'-	6.02	6.68	6.06	6.43	0.07	0.09
26	2,2',5,6'-	5.46	5.90	6.08	6.28	0.21	0.05
27	2,2',6,6'-	5.76	5.98	5.82	5.98	0.02	0.04
28	2,3,4,4'-	5.90	6.06	6.20	6.43	0.13	0.06

Response: -LogY$_w$		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	δ_{-LogYw}	*Ry*
29	2,3,4',5-	6.21	6.43	5.90	6.02	0.13	0.03
30	2,3',4,4'-	5.85	6.08	6.21	6.94	0.27	0.18
31	2,3',4,6-	6.43	6.53	5.90	6.15	0.16	0.06
32	2,3',4',5-	6.20	6.26	6.43	6.94	0.19	0.13
33	2,4,4',5-	6.06	6.20	6.26	7.06	0.25	0.20
34	2,4,4',6-	6.10	6.43	5.90	6.26	0.09	0.09
35	2,3',4',5'-	5.72	6.25	5.46	6.20	0.08	0.19
36	2,2',3,4,5-	6.53	6.55	6.53	6.94	0.10	0.10
37	2,2',3,4,5'-	6.43	6.55	6.10	7.05	0.21	0.24
38	2,2',3,4,6-	6.68	6.84	6.10	6.53	0.19	0.11
39	2,2',3,5',6-	6.21	6.43	6.02	7.06	0.21	0.26
40	2,2',3,4',5'-	6.55	6.70	6.32	6.94	0.12	0.16
41	2,2',4,4',6-	6.51	6.59	6.51	6.59	0.00	0.02
42	2,3,3',4',6-	6.59	7.08	6.10	6.53	0.25	0.11
43	2,3,4,4',5-	6.51	7.05	6.51	6.55	0.13	0.01
44	2,3',4,4',5-	6.32	6.51	6.59	7.08	0.19	0.12
45	2,2',3,3',4,5-	7.09	7.12	6.70	7.14	0.10	0.11
46	2,2',3,3',5,6-	7.05	7.06	6.55	6.70	0.13	0.04
47	2,2',3,3',5,6'-	6.51	6.59	7.08	7.15	0.16	0.02
48	2,2',3,4,4',5-	7.15	7.21	7.21	7.50	0.09	0.07
49	2,2',3,4,4',5'-	6.84	7.15	6.68	7.88	0.22	0.30
50	2,2',3,4,5,5'-	6.84	7.15	7.14	7.21	0.09	0.02
51	2,2',3,5,5',6-	7.05	7.06	6.70	7.14	0.11	0.11
52	2,2',4,4',5,5'-	6.81	6.97	6.81	6.84	0.03	0.01
53	2,2',4,4',6,6'-	7.21	7.50	7.14	7.21	0.09	0.02
54	2,3,3',4,4',5-	7.21	7.50	6.84	7.50	0.09	0.17
55	2,3,3',4,4',6-	7.06	7.14	7.12	7.50	0.11	0.10
56	2,3,3',4',5,6-	7.12	7.88	7.09	7.12	0.20	0.01
57	2,2',3,3',4,4',5-	7.10	7.33	6.84	7.33	0.07	0.12
58	2,2',3,3',4,5,6'-	6.84	7.09	6.81	6.97	0.04	0.04
59	2,2',3,4,4',5,5'-	7.33	8.23	7.10	8.23	0.06	0.28
60	2,2',3,4,4',5',6-	6.97	7.15	6.97	7.10	0.01	0.03
61	2,2',3,4,5,5',6-	6.81	6.97	6.97	7.15	0.09	0.05
62	2,2',3,3',4,4',5,5'-	7.15	8.21	7.15	8.21	0.00	0.27
63	2,2',3,3',4,4',5,5',6-	8.23	8.75	8.23	8.75	0.00	0.13
64	2,2',3,3',4,4',5,5',6,6'-	> 8.21	-	> 8.21	-	0.00	-

Table 4.30 – Aqueous coefficient activity interval estimation.

The average disagreement obtained by comparing the experimentally derived intervals with the calculated ones, is quite low ($\bar{\delta}_{-LogY_w} = 0.12$). The average disagreement between the quantitative experimental values and their derived intervals calculated is equal to $\tilde{\delta}_{-LogY_w} = 0.09$. The model quality for the aqueous activity coefficient evaluated by complement of the average disagreement between experimental and calculated intervals is quite good ($Q_{-LogY_w} = 0.88$), and the standard deviation error is equal to = 0.421.

Response: LogK$_{ow}$		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	$\delta_{-Logkow}$	*Ry*
1	3-	4.63	4.72	-	4.63	0.03	-
2	4-	-	< 4.66	4.66	4.72	0.02	0.02
3	2,2'-	4.66	5.09	4.63	4.84	0.08	0.06
4	2,3'-	-	< 4.66	4.72	5.09	0.12	0.10
5	2,4-	4.72	5.04	5.09	5.27	0.15	0.05
6	2,4'-	4.72	5.15	4.84	5.15	0.03	0.09
7	3,3'-	5.23	5.24	5.09	5.12	0.04	0.01
8	4,4'-	5.12	5.33	5.12	5.44	0.03	0.09
9	2,2',3-	5.29	5.33	5.04	5.23	0.08	0.05
10	2,2',5-	5.23	5.27	5.44	5.71	0.13	0.08
11	2,2',6-	5.15	5.29	5.15	5.23	0.02	0.02
12	2,3,4'-	4.84	5.12	5.33	5.65	0.23	0.09
13	2,3,6-	5.71	5.75	5.23	5.33	0.15	0.03
14	2,3',5-	5.24	5.68	5.24	5.68	0.00	0.12
15	2,4,4'-	5.71	5.75	4.84	5.67	0.25	0.23
16	2,4',5-	5.65	5.71	4.84	5.71	0.23	0.24
17	2,3',4'-	5.68	5.71	5.23	5.29	0.13	0.02
18	2,2',3,3'-	5.98	6.24	5.71	5.73	0.15	0.01
19	2,2',3,4'-	5.98	6.10	5.73	5.75	0.10	0.01
20	2,2',3,5'-	5.98	6.10	5.55	5.72	0.15	0.05
21	2,2',3,6'-	5.04	5.29	5.65	5.68	0.18	0.01
22	2,2',4,4'-	6.03	6.03	5.79	6.22	0.12	0.12
23	2,2',4,5-	5.79	6.22	5.75	5.98	0.08	0.06
24	2,2',4,6-	5.44	5.55	5.72	5.98	0.15	0.07
25	2,2',5,5'-	6.10	6.22	5.98	6.24	0.04	0.07
26	2,2',5,6'-	5.75	5.98	5.67	5.73	0.09	0.02
27	2,2',6,6'-	5.27	5.65	5.29	5.65	0.01	0.10
28	2,3,4,4'-	5.67	5.98	5.79	5.94	0.04	0.04

Response: $LogK_{ow}$		*Experimental*		*Calculated*			
ID	*Name*	*Min*	*Max*	*Min*	*Max*	δ_{Logkow}	*Ry*
29	2,3,4',5-	5.69	6.03	6.03	6.10	0.11	0.02
30	2,3',4,4'-	5.55	5.67	5.94	6.03	0.13	0.03
31	2,3',4,6-	5.94	6.23	5.98	6.03	0.07	0.01
32	2,3',4',5-	5.69	5.92	5.94	5.98	0.08	0.01
33	2,4,4',5-	5.72	5.79	6.10	6.50	0.22	0.11
34	2,4,4',6-	5.92	5.94	6.03	6.10	0.05	0.02
35	2,3',4',5'-	5.67	5.72	5.69	5.79	0.03	0.03
36	2,2',3,4,5-	6.23	6.57	6.20	6.30	0.08	0.03
37	2,2',3,4,5'-	6.03	6.38	6.20	6.38	0.05	0.05
38	2,2',3,4,6-	6.42	6.78	5.92	6.23	0.24	0.09
39	2,2',3,5',6-	5.69	6.03	6.10	6.50	0.23	0.11
40	2,2',3,4',5'-	6.57	6.71	6.20	6.23	0.14	0.01
41	2,2',4,4',6-	6.30	6.71	6.30	6.71	0.00	0.11
42	2,3,3',4',6-	6.32	6.42	6.23	6.38	0.04	0.04
43	2,3,4,4',5-	6.23	6.42	6.23	6.57	0.04	0.10
44	2,3',4,4',5-	6.23	6.30	6.23	6.32	0.01	0.03
45	2,2',3,3',4,5-	6.73	6.82	6.32	6.42	0.14	0.03
46	2,2',3,3',5,6-	6.20	6.42	6.57	6.76	0.16	0.05
47	2,2',3,3',5,6'-	6.30	6.71	6.20	6.76	0.04	0.16
48	2,2',3,4,4',5-	6.76	6.78	6.75	6.78	0.00	0.01
49	2,2',3,4,4',5'-	6.50	6.76	6.42	6.54	0.08	0.03
50	2,2',3,4,5,5'-	6.50	6.76	6.54	6.78	0.02	0.07
51	2,2',3,5,5',6-	6.20	6.50	6.32	6.73	0.10	0.11
52	2,2',4,4',5,5'-	6.54	7.08	6.78	6.85	0.13	0.02
53	2,2',4,4',6,6'-	6.78	7.44	6.73	6.75	0.20	0.01
54	2,3,3',4,4',5-	6.54	6.80	6.78	6.80	0.07	0.01
55	2,3,3',4,4',6-	6.50	6.85	6.82	7.44	0.25	0.17
56	2,3,3',4',5,6-	6.82	7.44	6.75	6.82	0.19	0.02
57	2,2',3,3',4,4',5-	6.80	7.04	6.85	7.04	0.01	0.05
58	2,2',3,3',4,5,6'-	6.50	6.73	6.80	7.08	0.16	0.08
59	2,2',3,4,4',5,5'-	7.04	7.62	6.99	7.62	0.01	0.18
60	2,2',3,4,4',5',6-	7.08	7.21	6.85	6.99	0.10	0.04
61	2,2',3,4,5,5',6-	6.54	7.08	7.04	7.21	0.18	0.05
62	2,2',3,3',4,4',5,5'-	7.21	7.94	7.21	7.94	0.00	0.20
63	2,2',3,3',4,4',5,5',6-	7.62	8.20	7.62	8.20	0.00	0.16
64	2,2',3,3',4,4',5,5',6,6'-	> 7.94		> 7.94		0.00	-

Table 4.31 – Hydrophobicity interval estimation.

By comparing the experimentally intervals with the calculated ones for the hydrophobicity a 0.10 average disagreement has been obtained ($\bar{\delta}_{-\mathrm{LogK}_{\mathbf{ow}}} = 0.10$), and the same value has been obtained for the average disagreement between the quantitative experimental values and their derived intervals ($\tilde{\delta}_{-\mathrm{LogK}_{\mathbf{ow}}} = 0.10$). The hydrophobicity model is of good quality ($\mathrm{Q}_{-\mathrm{LogK}_{\mathrm{ow}}} = 0.90$). In addition, the standard deviation error has been calculated: SDE = 0.245.

4.6.6 Overall model quality

Once calculated the model quality for the three single properties, the overall ranking model quality has been evaluated from the single quality parameters by arithmetic means (Q_{T}), geometric mean (Q_{G}) and by the minimum value obtained on the three responses (Q_{M}):

$$Q_{\mathrm{T}} = 0.91 \qquad Q_{\mathrm{G}} = 0.91 \qquad Q_{\mathrm{M}} = 0.88$$

The multilinear regression has been performed developed on the environmental impact score of the 64 PCBs starting from the same molecular descriptors used for the ranking model

The total ranking model obtained for PCBs physico-chemical properties allows the defining of a priority list for PCBs according to their environmental impact, contemporary accounting for their solubility, aqueous activity coefficient and hydrophobicity. The model obtained is a very simple model, based on only two descriptors able to provide a multiresponses (multiproperties) modelling. Moreover, by multilinear regression modelling performed on the environmental impact score of the 64 PCBs starting from the same molecular descriptors used for the ranking model it has been obtained a two dimensional model with a leave-one-out explained variance equal to 0.95. The two selected descriptors are the second lowest eigenvalue of the Burden matrix weighted by atomic masses (BELm2) and a 3D-MoRSE descriptor (Mor29u). It can be poited out that BELm2 descriptor has been selected by the ranking model too.

4.7 Toxicity partial ranking model

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 Europe, the ecotoxicity of a chemical for the aquatic environment is typically estimated on the basis of a set of data from simple, standardised bioassays on surrogate organism taken as representative of the major trophic levels. 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 daphnids. 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.

Quantitative Structure - Activity Relationships (QSARs) are estimation methods developed and used to predict certain effects or properties of chemical substances, which are primarily based on the structure of the chemicals. QSAR models describe variation in a given end point of chemicals, from the variation in their structural and electronic features. Based on the developed QSAR model, end-point of new, structurally related chemicals, not yet experimentally investigated, may be predicted. QSAR models can be used with many purposes: screening chemical databases and virtual libraries before the synthesis of chemicals; reducing reliance on animal testing. Moreover, they contribute to the decision making process on whether further testing is needed to clarify an end-point of interest, and, if further testing is necessary, to optimise testing strategies, when appropriate.

The development of quantitative-structure activity relationships (QSARs) often relies on the application of statistical methods such as multilinear regression (MLR) or partial least squares regression (PLS).

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, order models can be used as alternative to statistical methods such as multi-linear regression (MLR). Order ranking models assume, contrary to standard multidimensional statistical analysis, neither linearity nor any assumptions about distribution properties; it can be consider as a parameter-free method. Thus, even if the information provided by order ranking models is not a quantitative information but a simpler information regarding the ordinal relation among chemicals, for exposure analysis and risk assessment ranking models can be a very useful tool in supporting decision making processes.

4.7.1 Toxicity experimental data

The toxicity data have been provided by the EU project: BEAM EVK1-1999-00012. 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 *Scenedesum vacuolatus* by the research group of Prof. Grimme, Bremen University, EU project: BEAM EVK1-1999-00012. The dependent variables selected for describing their toxicity were the algae inhibition with 3 concentrations of 10, 50, 90 mmol/l. Table 4.32 shows the toxicity values of the 23 chemicals.

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

Table 4.32 – Experimental Toxicity (Log1/EC) data of 23 chemicals.

4.7.2 Molecular descriptors

The chemical structures of the chemicals have been described with more than 1500 molecular descriptors, in order to catch all the structural information. The molecular descriptors used to search for the best partial ranking model of the toxicity activity of the selected chemicals have been calculated by the *Dragon* program [Todeschini *et.al*., 2003] on the basis of the minimum energy molecular geometries optimized by *HyperChem* package [HYPERCHEM, 1995] (PM3 semiempirical method). In this study the following sets of molecular descriptors have been calculated: constitutional descriptors, topological descriptors [Bonchev, 1983; Devillers and Balaban, 2000], walk and path counts, connectivity indices [Kier and Hall, 1986], information indices, Moreau-Broto 2D-autocorrelations [Moreau and Broto, 1980a; 1980b; Broto *et al*., 1984], edge adjacency indices [Estrada, 1995], BCUT descriptors [Pearlman and Smith, 1998; Pearlman, 1999], topological charge indices [Galvez *et al*., 1994; 1995], eigenvalue based indices [Balaban *et al*., 1991], Randic molecular profiles [Randic 1995, 1996], geometrical descriptors, radial distribution function descriptors [Hemmer *et al*., 1999], 3D-MoRSE descriptors [Schuur, and Gasteiger, 1996, 1997], WHIM descriptors [Todeschini *et al*., 1994; Todeschini and Gramatica, 1997], GETAWAY descriptors [Consonni *et al.*, 2002], 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* of Todeschini and Consonni [Todeschini and Consonni, 2000].

4.7.3 Experimental ranking

The Hasse diagram technique has been applied on the three toxicity responses of algae inhibition with 3 concentrations of 10, 50, 90 mmol/l. Figure 4.19 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.

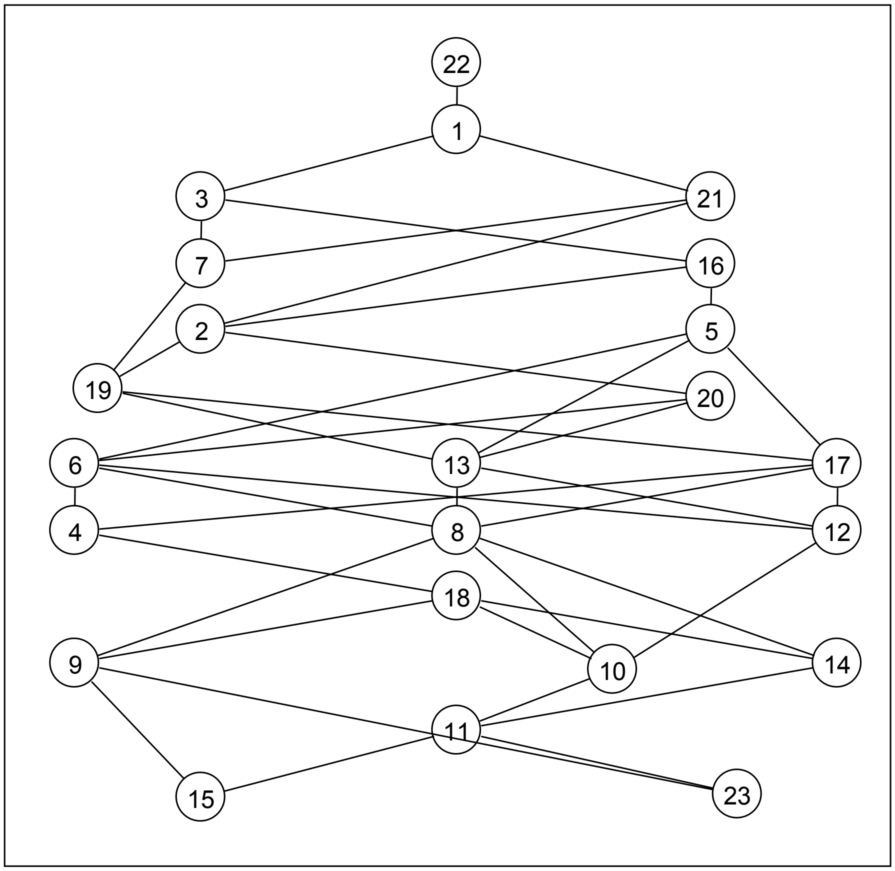


Figure 4.19 – Experimental Hasse diagram

No equivalence classes with more than one element exists, thus no degeneracy occurs (D = 0; k_{std} = 0) and the entropy is maximum (H^* = 1; G^* = 1). The diagram shows a high capability of discriminating elements according to different ranks (*DbyR* = 0.88), good selectivity (T = 0.50) and low element diversity (*div* = 0.09). The number of comparability is quite high (X = 0.88) and thus the stability quite low (*P* = 0.12; *StR* = 0.11). The diagram is not complex (C_x' = 0.24) and element relationships can be easily investigated.

4.7.4 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 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 [Todeschini, et al. 2003].

The best 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 4.20: it is arranged on eleven levels and characterized by 171 comparable pairs of elements and 164 contradictions. The two model descriptors value are illustrated in Table 4.33. 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.

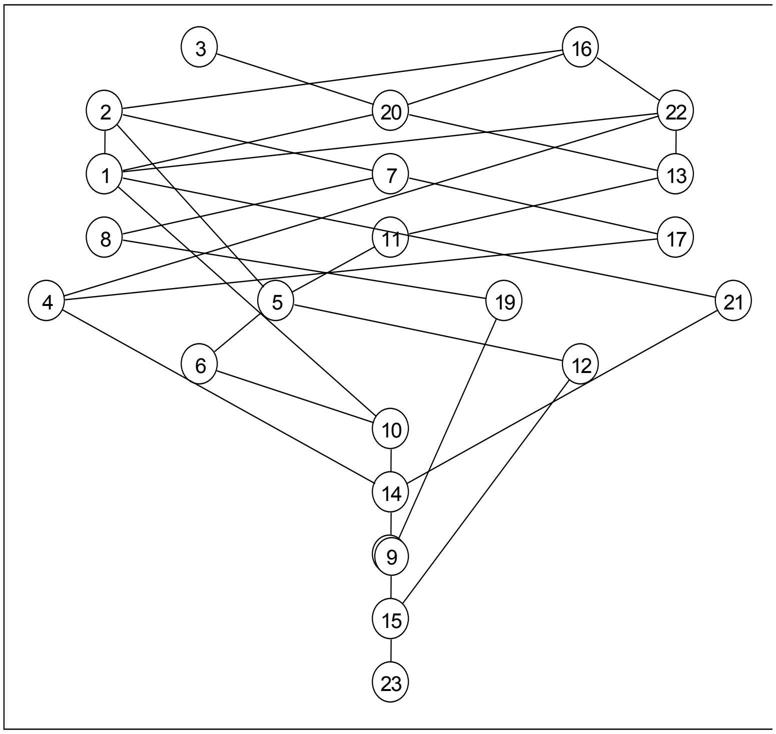


Figure 4.20 – Model Hasse diagram developed with nN and CIC2 descriptors.

The model diagram provides 22 equivalent classes over 23 chemicals: bromoxynil and ioxynil belong to the same equivalence class. A low degeneracy (*D* = 0.05; k_{std} = 0) and a high entropy (*H** = 0.98; *G** = 1) are detected. The diagram shows a medium capability of discriminating elements according to different ranks (*DbyR* = 0.67) and selectivity (*T* = 0.48) and low element diversity (*div* = 0.10). The number of comparability is not too high (X = 0.68) and thus the stability quite low (*P* = 0.33; *StR* = 0.17). The diagram is pretty complex (C_x' = 0.65), since the two contributions of comparability and incomparability are similar.

ID	*Substance*	*nN*	*CIC2*
1	Aclonifen	2	1.228
2	Atrazin	5	1.376
3	Lenacil	2	2.114
4	Chloridazon	3	0.885
5	Alachlor	1	1.366
6	Metolachlor	1	1.241
7	Tribenuron-methyl	5	1.134
8	Thifensulfuron-methyl	5	0.744
9	Bromoxynil	1	0.571
10	Carbofuran	1	1.194
11	Cycloxydim	1	1.603
12	Ethofumesate	0	1.244
13	Isofenphos	1	1.622
14	Isoxaflutol	1	0.885
15	MCPA	0	0.523
16	Terbuthylazin	5	1.799
17	Metamitron	4	1.005
18	Ioxynil	1	0.571
19	Triasulfuron	5	0.655
20	Isoproturon	2	1.719
21	Linuron	2	0.971
22	Pendimethalin	3	1.718
23	2,4-Dichlorophenoxyacetic acid	0	0.461

Table 4.33 – Model descriptors value for 23 chemicals.

4.7.5 Experimental and model ranking comparison

Variable subset selection has been performed by GAs optimising the similarity index S(E,M) defined in chapter 3. The agreement degree between experimental and model diagrams is quite satisfactory (S(E,M) = 76.3). The Tanimoto indices have been calculated:

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

The "goodness of fitting" 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 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, than the T(0,1) index. According to the ranking indices values differences between experimental and model diagram can be highlighted. Being the model diagram characterised by a higher number of incomparabilities than the experimental diagram (164 *vs* 60) it shows lower values of comparability, discrimination power by ranking, selectivity and higher values of diversity and stability as well.

4.7.6 Interval estimation

The experimental ranking of each chemical has been estimated by the obtained ranking model, according to the procedure described in Chapter3. Thus, by the connectivity operator, the experimental ranking of any chemical *u* has been estimated looking for those two elements *s* and *t*, which are connected (comparable) to *u*, i.e. *C*(*s*,*u*) > 0 (with *s* above *u*) and *C* (*u*,*t*) > 0 (with *u* above *t*), located at the shortest path and which experimental value difference constitutes the smallest positive interval. Then, 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 δ_{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 standardise disagreements are illustrated in Table 4.34, 4.35 and 4.36, respectively.

Response: LOG(1/EC10)		*Experimental*		*Calculated*		δ_{EC10}
ID	*Substance*	*Min*	*Max*	*Min*	*Max*	
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 4.34 – Experimental LOG(1/EC10)interval estimation. (Bold fonts indicate wrong interval estimations). [a]2,4-Dichlorophenoxyacetic acid.

Response: LOG(1/EC50)		*Experimental*		*Calculated*		
ID	*Substance*	*Min*	*Max*	*Min*	*Max*	δ_{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 4.35 – Experimental LOG(1/EC50)interval estimation. (Bold fonts indicate wrong interval estimations). a2,4-Dichlorophenoxyacetic acid.

Response: LOG(1/EC90)		*Experimental*		*Calculated*		
ID	*Substance*	*Min*	*Max*	*Min*	*Max*	δ_{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

Table 4.36 – Experimental LOG(1/EC90)interval estimation. (Bold fonts indicate wrong interval estimations). a2,4-Dichlorophenoxyacetic acid.

4.7.7 Prediction uncertainty

For each response the interval estimation uncertainty has been calculated. The *topological uncertainty* measures (TU), the normalised rank uncertainties above and below ($D_u^{\sup}$ and $D_u^{\inf}$), and the *experimental uncertainty* Ry_r for Log(1/EC10), Log(1/EC50) and Log(1/EC90) responses are collected in Table 4.37, 4.38 and 4.39, respectively.

It can be pointed out that the entire set of intervals has been estimated by the first shell of neighbourhoods, excepted for Cycloxydim and Aclonifen, the latter only as far as concerns on Log(1/EC90) response.
This result means that the interval calculated are mostly in agreement with the experimental ones and only in a few cases the interval provided by exploring the first shell of neighbourhoods was not a positive interval requiring the second shell of neighbourhoods exploration.

Response: Log(1/EC10)					
ID	*Substance*	*TU*	D_u^{sup}	D_u^{inf}	Ry_u
1	Aclonifen	2	0.00	0.00	0.529
2	Atrazin	2	0.00	0.00	0.089
3	Lenacil	-	-	0.00	-
4	Chloridazon	2	0.00	0.00	0.391
5	Alachlor	2	0.00	0.00	0.238
6	Metolachlor	2	0.00	0.00	0.499
7	Tribenuron-methyl	2	0.00	0.00	0.192
8	Thifensulfuron-methyl	2	0.00	0.00	0.061
9	Bromoxynil	2	0.00	0.00	0.313
10	Carbofuran	2	0.00	0.00	0.344
11	Cycloxydim	3	0.00	0.11	0.108
12	Ethofumesate	2	0.00	0.00	0.820
13	Isofenphos	2	0.00	0.00	0.598
14	Isoxaflutol	2	0.00	0.00	0.135
15	MCPA	2	0.00	0.00	0.004
16	Terbuthylazin	-	-	0.00	-
17	Metamitron	2	0.00	0.00	0.361
18	Ioxynil	2	0.00	0.00	0.313
19	Triasulfuron	2	0.00	0.00	0.156
20	Isoproturon	2	0.00	0.00	0.144
21	Linuron	2	0.00	0.00	0.391
22	Pendimethalin	2	0.00	0.00	0.144
23	[a]2,4- D	-	0.00	-	-

Table 4.37 – Uncertainty measures on Log(1/EC10) interval estimation. [a]2,4-Dichlorophenoxyacetic acid.

Response: Log(1/EC50)					
ID	*Substance*	*TU*	D_u^{sup}	D_u^{inf}	Ry_u
1	Aclonifen	2	0.00	0.00	0.202
2	Atrazin	2	0.00	0.00	0.061
3	Lenacil	-	-	0.00	-
4	Chloridazon	2	0.00	0.00	0.325
5	Alachlor	2	0.00	0.00	0.132
6	Metolachlor	2	0.00	0.00	0.595
7	Tribenuron-methyl	2	0.00	0.00	0.215
8	Thifensulfuron-methyl	2	0.00	0.00	0.065
9	Bromoxynil	2	0.00	0.00	0.189
10	Carbofuran	2	0.00	0.00	0.408
11	Cycloxydim	3	0.00	0.11	0.140
12	Ethofumesate	2	0.00	0.00	0.751
13	Isofenphos	2	0.00	0.00	0.617
14	Isoxaflutol	2	0.00	0.00	0.162
15	MCPA	2	0.00	0.00	0.163
16	Terbuthylazin	-	-	0.00	-
17	Metamitron	2	0.00	0.00	0.264
18	Ioxynil	2	0.00	0.00	0.189
19	Triasulfuron	2	0.00	0.00	0.079
20	Isoproturon	2	0.00	0.00	0.410
21	Linuron	2	0.00	0.00	0.325
22	Pendimethalin	2	0.00	0.00	0.376
23	[a]2,4- D	-	0.00	-	-

Table 4.38 – Uncertainty measures on Log(1/EC50) interval estimation. a2,4-Dichlorophenoxyacetic acid.

Response: Log(1/EC90)					
ID	*Substance*	*TU*	D_u^{sup}	D_u^{inf}	Ry_u
1	Aclonifen	3	0.11	0.00	0.072
2	Atrazin	2	0.00	0.00	0.043
3	Lenacil	-	-	0.00	-
4	Chloridazon	2	0.00	0.00	0.273
5	Alachlor	2	0.00	0.00	0.091
6	Metolachlor	2	0.00	0.00	0.621
7	Tribenuron-methyl	2	0.00	0.00	0.254
8	Thifensulfuron-methyl	2	0.00	0.00	0.064
9	Bromoxynil	2	0.00	0.00	0.241
10	Carbofuran	2	0.00	0.00	0.436
11	Cycloxydim	3	0.00	0.11	0.102
12	Ethofumesate	2	0.00	0.00	0.807
13	Isofenphos	2	0.00	0.00	0.596
14	Isoxaflutol	2	0.00	0.00	0.170
15	MCPA	2	0.00	0.00	0.129
16	Terbuthylazin	-	-	0.00	-
17	Metamitron	2	0.00	0.00	0.197
18	Ioxynil	2	0.00	0.00	0.241
19	Triasulfuron	2	0.00	0.00	0.003
20	Isoproturon	2	0.00	0.00	0.551
21	Linuron	2	0.00	0.00	0.273
22	Pendimethalin	2	0.00	0.00	0.372
23	[a]2,4- D	-	0.00	-	-

Table 4.39 – Uncertainty measures on Log(1/EC90) interval estimation. a2,4-Dichlorophenoxyacetic acid.

4.7.8 Model quality

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

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

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

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

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, both by complement of the average disagreement between experimental and calculated intervals (Q_r) and by the ratio of the number of interval correctly calculated by the model on the total number of intervals (NER_r):

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

$$\text{NER}_{\text{Log(1/EC10)}} = 0.565 \quad \text{NER}_{\text{Log(1/EC50)}} = 0.565 \quad \text{NER}_{\text{Log(1/EC90)}} = 0.478$$

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}; NER_{T}), geometric mean (Q_{G};NER_{G}) and by the minimum value obtained on the three responses (Q_{M}; NER_{M}):

$$Q_{\text{T}} = 0.705 \qquad Q_{\text{G}} = 0.705 \qquad Q_{\text{M}} = 0.686$$

$$NER_{\text{T}} = 0.536 \qquad NER_{\text{G}} = 0.535 \qquad NER_{\text{M}} = 0.478$$

The present 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 to 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.

CHAPTER 5

RANA: software for Ranking ANalysis Alghoritms

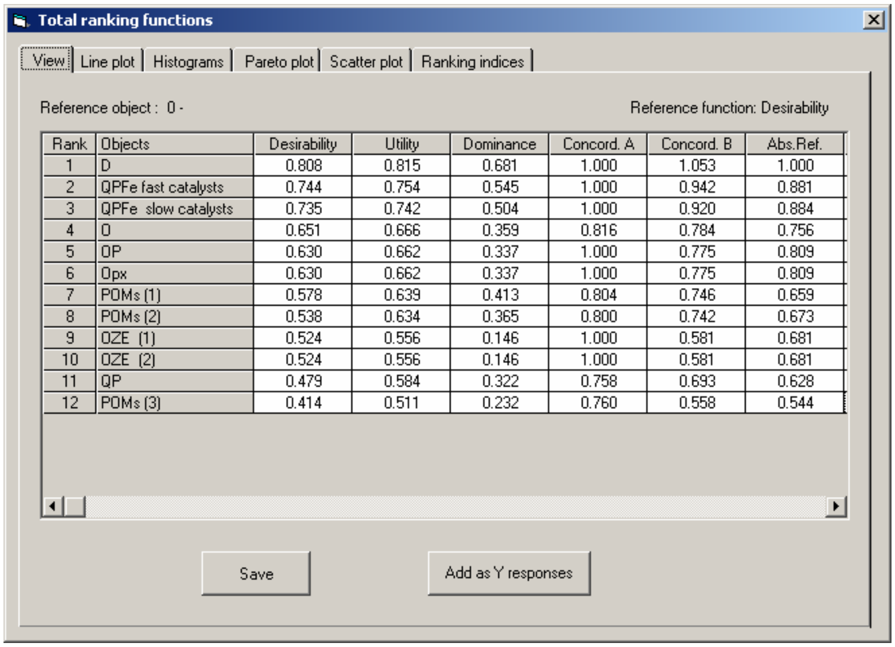
RANA software has been developed for ranking analysis data exploration and modelling. It allows performing ranking evaluation by both total and partial ranking procedures. The implemented total order ranking methods are: Desirability functions, Utility functions, Dominance functions, classical and quantitative Concordance analysis, Absolute reference method. Ranking analysis by Hasse diagram technique is also provided. Both total and partial analysis can be analysed by ranking indices. Moreover both total and partial ranking models by genetic algorithms variable subset selection can be performed. RANA software is a 32 bit application and can be run on Windows platforms. The programming language is Microsoft Visual Basic 6.0. Figure 5.1 shows the main form of the software.

5.1 Data setup

Data in the standard ASCII text format can be loaded by RANA. Once the data has been loaded, the data setup menu allows the user to select the independent variables X and the response variable Y. By default, all the objects are assigned to the training set except for those lacking values in at least one independent variable and the response. The user can modify, by hand, the object allocation by forcing an object's exclusion from the analysis. Both X and Y variables can be transformed before analysis. The available transformations are the logarithmic, inverse and square root. Moreover, to perform ranking analysis the user has to define the values and situations of optimum, i.e. for each Y variable it is necessary to ascertain explicitly if the best condition is satisfied by a minimum or a maximum criterion value, and the trend from the minimum to the maximum must also be established. The Y rank transformation available are: linear, sigmoid, logarithmic, exponential, step, normal, parabolic, Laplace, triangular and box and their correspondent inverse transformations. Each variable can be weighted in order to take into account its importance in the ranking analysis.

5.2 Ranking explorative analysis

Once the data have been loaded and the data and criteria setup performed the *Ranking methods* menu is activated. Two options are available: total order ranking or partial order ranking. The total ranking option provides a table form organised in six tables (Figure 5.1). The first one provides the ranking list of the element according their ranking scores calculated by Desirability functions, Utility functions, Dominance functions, classical and quantitative Concordance analysis, Absolute reference method. The total ranking results can be store in ACII text format. The second and third tables provides line and histograms plots respectively. These plots allow an easy comparison of the ranking scores calculated by the seven ranking methods, thus they are useful to visually detect method differences. The third table provides the so-called Pareto plot, which is a histogram plot where the elements are ranked according to their ranking score. The user can easily select the ranking method, whose score is then used in the graph. In the fifth table scatter plots and 3D plots can be develop, in order to analyse and compare elements according to their ranking scores. The last Table provides ranking indices values in order to allow an immediate analysis of the ranking quality and an easy comparison of the obtained rankings.



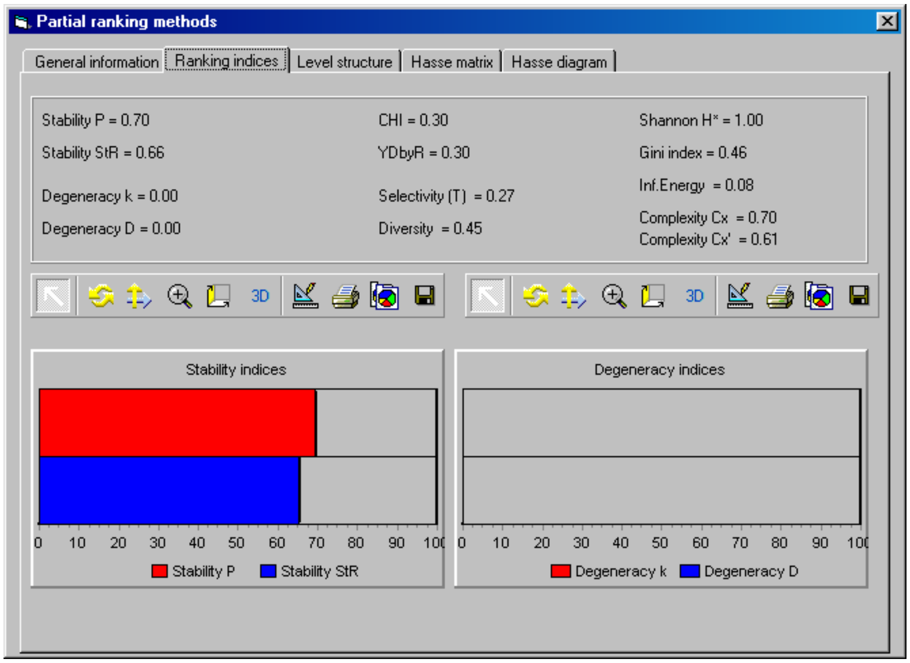
Reference object : 0 -

Reference function: Desirability

Rank	Objects	Desirability	Utility	Dominance	Concord. A	Concord. B	Abs.Ref.
1	D	0.808	0.815	0.681	1.000	1.053	1.000
2	QPFe fast catalysts	0.744	0.754	0.545	1.000	0.942	0.881
3	QPFe slow catalysts	0.735	0.742	0.504	1.000	0.920	0.884
4	O	0.651	0.666	0.359	0.816	0.784	0.756
5	OP	0.630	0.662	0.337	1.000	0.775	0.809
6	Opx	0.630	0.662	0.337	1.000	0.775	0.809
7	POMs (1)	0.578	0.639	0.413	0.804	0.746	0.659
8	POMs (2)	0.538	0.634	0.365	0.800	0.742	0.673
9	OZE (1)	0.524	0.556	0.146	1.000	0.581	0.681
10	OZE (2)	0.524	0.556	0.146	1.000	0.581	0.681
11	QP	0.479	0.584	0.322	0.758	0.693	0.628
12	POMs (3)	0.414	0.511	0.232	0.760	0.558	0.544

Figure 5.1 – Total ranking table form.

The partial ranking option allows performing partial ranking analysis by Hasse diagram technique. Its table form is made of five tables (Figure 5.2). The first one provides general information related to the Hasse diagram obtained: the number of levels (NL), the number of equivalence classes with more than one element (NECA), the number of elements in the level that contains the most elements (NEL), the number of maximals and minimals (N.Max and N.Min), the number of equivalence classes (Z), the number of comparabilities (V) and the number of incomparabilities (U). In the second table several ranking indices for partial rankings are collected: the *StR* and *P* stability indices, the Brüggemann standardized degeneracy index (k_{std}) and the absolute degeneracy degree (*D*), the comparability index (χ), the discrimination power by ranking index (*DbyR*), the selectivity index (*T*), the diversity index (*div*), the standardized Shannon (*H**), the standardized Gini entropy index (*G**), the information energy content (I_E) and the two complexity indices *Cx* and *Cx"*.



Stability P = 0.70
Stability StR = 0.66
Degeneracy k = 0.00
Degeneracy D = 0.00
CHI = 0.30
YDbyR = 0.30
Selectivity (T) = 0.27
Diversity = 0.45
Shannon H* = 1.00
Gini index = 0.46
Inf.Energy = 0.08
Complexity Cx = 0.70
Complexity Cx' = 0.61

Figure 5.2 – Partial ranking table form.

The level structure table provide information on the level structure, providing the number and list of elements in each level. Then, the fourth table contains the Hasse matrix, which can be stored by the user. Finally the Hasse diagram is provided in the last table.

5.3 Ranking modelling

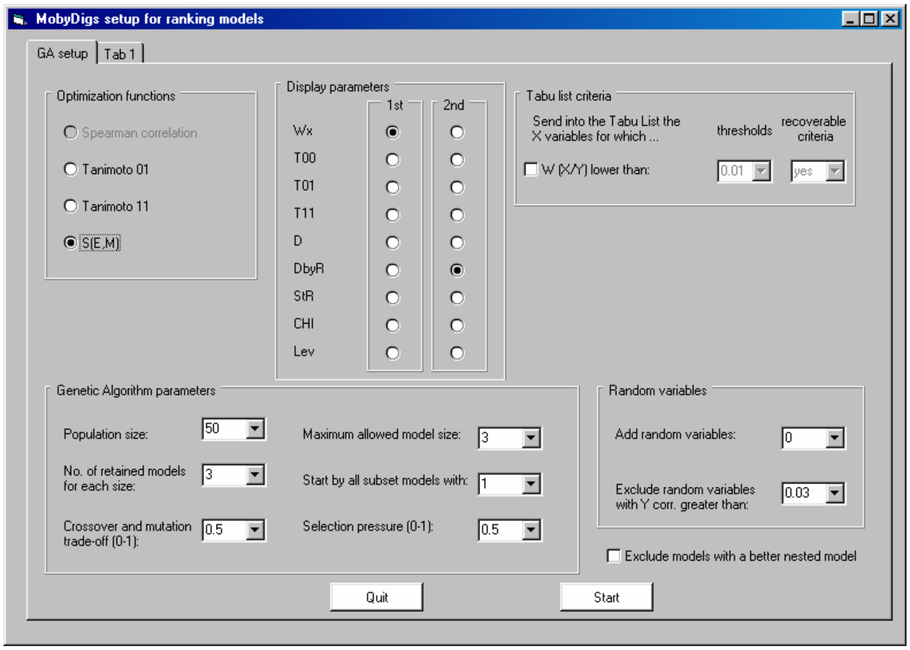
Both total and partial ranking models can be developed by RANA software. The Genetic Algorithm (GA-VSS) approach is here used as the variable selection method to search for the best ranking models within a wide set of variables. Once the data have been loaded and the criteria setup performed the *GA setup* menu is activated. The options available in this menu are concerned with the Genetic Algorithm parameters, the variable management and the choice of objective function.
The five main parameters for the Genetic Algorithm are:

1. Population size: maximum number of models in a population (default: 50).
2. Maximum allowed model size: maximum number of variables in a model (default: 3).
3. Crossover and mutation trade-off: user-defined value of the T parameter which sets the values of the crossover and mutation probabilities (default: 0.5).
4. Number of retained models for each size: number of the best models for each size surviving in the population regardless their quality (default: 3). This option is important to save, in the final population, also the best models of lower complexity e.g., the first three models with one variable, the first three models with two variables, etc.
5. Selection pressure: user-defined value of the B parameter which sets the parent selection operator (default: 0.5).

The user-defined genetic algorithm parameters can be differentiated in the evolution procedure. Several objective functions are available in RANA for evaluating the quality of population individuals. For total ranking model the optimised parameter is the Spearman rank correlation coefficient, while for partial ranking models the user can select the following objective functions:

1. T(0,1) Tanimoto index
2. T(1,1) Tanimoto index
3. S(E,M) Similarity index

Together with the chosen objective function, the user can select another two parameters to be displayed during the evolution process. The displayed parameters are the Kendall's coefficient of concordance W_X within the X variables, the T(0,0), T(0,1), T(1,1) Tanimoto indices (when they are not the one optimised), the absolute degeneracy index (*D*), the discrimination power by ranking (*DbyR*), the *StR* stability index, the number of levels and (see Figure 5.3).



Optimization functions

○ Spearman correlation

○ Tanimoto 01

○ Tanimoto 11

● S(E,M)

Display parameters

	1st	2nd
Wx	●	○
T00	○	○
T01	○	○
T11	○	○
D	○	○
DbyR	○	●
StR	○	○
CHI	○	○
Lev	○	○

Tabu list criteria

Send into the Tabu List the X variables for which ...	thresholds	recoverable criteria
□ W (X/Y) lower than:	0.01	yes

Genetic Algorithm parameters

Population size: 50

No. of retained models for each size: 3

Crossover and mutation trade-off (0-1): 0.5

Maximum allowed model size: 3

Start by all subset models with: 1

Selection pressure (0-1): 0.5

Random variables

Add random variables: 0

Exclude random variables with Y corr. greater than: 0.03

□ Exclude models with a better nested model

Figure 5.3 – RANA GA setup.

5.3.1 Population evolution view

Once population optimisation is started, the user can follow the populations on two screens. In the first screen the best model is shown, described by the objective function value, two other selected fitness parameters, its size and variables, together with the maximum allowed number of model variables and the number of population individuals (models). The second screen shows the evolution of the population; all the population models are displayed, described by their objective function values, two other selected fitness parameters, model size, model variables.

5.3.2 Modify population evolution

At any time the user can choose to modify the population evolution genetic parameters to run through new evolution directions. The parameters are usually changed when the population has become so stable that no new individuals enter the population. The maximum number of allowed variables in a model could be increased if the optimal model complexity has not yet been reached. If model searching has been initialised by small sized models, it is common practice to increase the model variable number for a satisfactory exploration of the model space, until the population becomes stable or the desired objective function has been obtained. Since the coverage of the search space is influenced by crossover/mutation trade off and selection pressure, these can be changed to augment the diversity of the population when it falls into a local optima and tends to be constituted by very similar individuals. In this case the crossover/mutation trade off should be increased, while the selection pressure should be decreased to enhance the random process of GA and thus produce new genetic material in the population. On the other hand, when the population has been initialised favouring high diversity among individuals, the crossover/mutation trade off should be decreased and the selection pressure increased to force the evolution towards local, or hopefully absolute, optima.

5.3.3 Variable frequency analysis

This menu allows the analysis of variable frequency in the final population. In particular, the total number and percentage of variables present in the final models is given, together with a table showing the frequency of each variable in the population calculated on the basis of only the final models. Variable frequency analysis can be useful to detect those variables showing high importance in ranking modelling.

5.3.4 Saving results

The final population models can be saved in a tabulated ASCII file where the models are listed according to the decreasing value of their quality. Each model is described by:

- model size;
- model variables;

- optimised parameter
- multivariate Kendall's coefficient of concordance among the independent variables X, W_X;
- multivariate Kendall's coefficient of concordance among the independent variables X plus the response, W_{XY};

Other results that can be saved in *RANA* are:

Experimentally derived and model intervals calculated or predicted by all the selected models and their standardised disagreement δ_{ir}

Rank uncertainty measures ($D_u^{\sup}$ and $D_u^{\inf}$)

Experimental uncertainty *Ry*

Overall ranking model quality parameters (Q_T , Q_G , Q_M)

5.3.5 *RANA* constraints

Maximum number of objects: 3000
Maximum number of variables: 2000
Maximum number of variables in a model: 20
Maximum number of individuals (models): 100

CONCLUSIONS

The intrinsic 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. Ordering is one of the possible ways to analyse data and to get an overview over the elements of a system. In the present thesis order ranking strategies have been investigated. Ordinal scaling is usually seen as a "weaker" property than metric scaling, and this means that element ranking based on a set of attributes is seen as "basic information" which is supplemented with metric information. Even thought often considered as less informative techniques, total and partial order ranking (POR) methods gave evidence to be efficient tool to perform data analysis, evaluating order relationship among the elements of the system investigated. The well known total order ranking methods, have been investigated and some new contributions have been here proposed.

The less known partial ranking analysis by Hasse diagram technique has been deeply examined and compared with the total order approach. Since a complete evaluation by ranking technique needs to be supported by a pre-processing phase to define an adequate data matrix, as well as a post-processing phase to extract information and decisions on the system investigated, investigations have been carried out on both these phases. Comparison of the main pre-processing statistical techniques, clustering, principal component analysis and broad order statistics pointed out that broad order statistics seems to be a very suitable pre-processing tool, providing a satisfactory solution to those ranking drawbacks related to noise and measurement error.

As far as concerns the post-processing phase, which deals with the ranking quality establishment, new ranking indices have been here proposed, and compared with those found in literature. Tested on both theoretical and real data, they result suitable to represent the main ranking properties and to encode unique information. However, further analysis on diverse datasets appears appropriate to fully elucidate their meaning and utility.

Order ranking methods have been analysed even for modelling purposes and they have been demonstrated to be a possible alternative to conventional statistical modelling such as multilinear regression (MLR) or classification.

A complete procedure to perform a 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. Total and partial ranking optimisation parameters have been investigated, and the new one proposed has been compared with those already published in the literature. Interval estimation by ranking models have been analyzed deeply and a new approach has been proposed here, together with a few measures of prediction uncertainty. It is further worthwhile to highlight that the procedure proposed can be located between fiitting 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 diagram, and using the remaining training set elements to calculate the model intervals of the deleted element from the model ranking diagram. 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. However, 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 recommended. Moreover, one of the main theoretical aspect not yet fully investigated concerns the search for validation techniques suitable for ranking models.

Finally, several different new fields have been here explored by ranking methodologies, these revealing their effective capability to catch new useful kind of information from multivariate data.

Among them, it seems particularly interesting the ranking method applicability on three-way data.

BIBLIOGRAPHY

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.

Basilevsky, A. (1994). *Statistica Factor Analysis and Related Methods. Theory and Applications*, Wiley-VCH, NY (USA).

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

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

Brans, J.P., Vincke, Ph. (1985). A Preference Ranking Organitation Method (the PROMETHEE Method for Multiple Criteria Decision Making). *Management Science.*, *31*, 647-656.

Brans, J.P., Vincke, Ph., Mareschal, B. (1986). How to Select and How to Rank Projects: the PROMETHEE Method. *European Journal of Operation Research.*, *24*, 228-238.

Brillouin, L. (1962). Science and Information Theory. Academic Press (2nd ed.), New York (NY).

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

Brüggemann, R., Münzer, B. (1993a). A Graph - Theoretical Tool for Priority Setting of Chemicals. *Chemosphere*, *27*, 1729-1736.

Brüggemann, R., Bucherl, C., Pudenz, S., Steinberg, C.E.W. (1993b). Application of the Concept of Partial Order on Comparative Evaluation of Environmental Chemicals. *Acta Hydrochim. Hydrobiol.*, 27, 170-178.

Brüggemann, R., Munzer, B., Halfon, E. (1994). An Algebaic/Graphical Tool to Compare Ecosystems with Respect to their Pollution - The German River "Elbe" as an Example - I: Hasse - Diagrams. *Chemosphere*, *28*, 863-872.

Brüggemann, R., Schwaiger, J., Negele, R.D. (1995a). Applying Hasse Diagrams Technique for the Evaluation of Toxicological Fish Tests. *Chemosphere*, *39*, 1767-1780.

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

Brüggemann, R. and Voigt, K. (1996). Stability of Comparative Evaluation, Example: Environmental Databases. *Chemosphere*, *33*, 1997-2006.

Brüggemann, R., Oberemm, A., Steinberg, C. (1997a). Ranking of Aquatic Effect Tests Using Hasse Diagrams. *Toxicological and Environmental Chemistry*, *63*, 125-139.

Brüggemann, R., Voigt, K., Steinberg, C.E.W. (1997b). Application of Formal Concept Analysis to Evaluate Environmental Databases. *Chemosphere*, *35*, 479-486.

Brüggemann, R., Pudenz, S., Voigt, K., Kaune, A., Kreimes, K.(1999a). An Algebaic/Graphical Tool to Compare Ecosystems with Respect to their Pollution IV: Comparative Regional Analysis by Boolean Arithmetics. *Chemosphere*, *38*, 2263-2279.

Brüggemann, R. and Halfon, E. (1999b). Introduction to the General Principles of the Partial Order Ranking Theory. In *Order Theoretical Tools in Environmental Sciences*; Proceedings of the Second Workshop on Order Theoretical Tools in Environmental Sciences, 7-43.

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

Brüggemann, R. and Welz, G. (2001a). Order Theory Meets Statistics – Hasse Diagram Technique. In *Order Theoretical Tools in Environmental*

Sciences; Proceedings of the Fourth Workshop on Order Theoretical Tools in Environmental Sciences, 9-39

Brüggemann, R., Halfon, E., Welz, G, Voigt, K., Steinberg, C.E.W. (2001b). Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests. *J.Chem.Inf.Comput.Sci.*, *41*, 918-925.

Carlsen, L., Sørensen, P.B., Thomsen, M. (1999). Estimation of Octanol-Water Distribution Coefficients. In *Order Theoretical Tools in Environmental Sciences*; Proceedings of the Second Workshop on Order Theoretical Tools in Environmental Sciences, 105-115.

Carlsen, L. (2001). Predicting Power of Partial Order Technique. Comparison to Partial Least Squares Regression. In *Order Theoretical Tools in Environmental Sciences*; Proceedings of the Fourth Workshop on Order Theoretical Tools in Environmental Sciences, 159-168.

Carlsen, L., Sørensen, P.B., Thomsen, M. (2001). Partial Order Ranking-based QSAR's: estimation of solubilities and octanol-water partitioning. *Chemosphere*, *43*, 295-302.

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

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.

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.

Derringer GC and Suich R. (1980). Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology*, *12*, 214-219.

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

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

European Communities. IUCLID CD-ROM Year 2000 Edition, Public Data on High Volume Chemicals, EUR 19559EN. European Communities: Luxembourg. 2000.

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

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.

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

Goldstein, M. (1982). Preliminary Inspections of Multivariate Data. *Journal of American. Statistical association*, *30*, 823-831.

Grammatica, P., Navas, N., Todeschini, R. (1998). 3D-Modelling and prediction by WHIM descriptors. Part9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). *Chemom.Intell.Lab.Syst.*, *40*, 53-63.

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

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.

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.

Harrington, E.C. (1965). The Desirability Function. *Industrial Quality Control*., *21*, 494-498.

Hartley, R.V.L. (1928). Transmission of Information. *Bell. Syst. Tech. J.*, *7*, 535-563.

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.

Hendriks, M.M.W.B., Boer, J.H. , Smilde, A.K., Doorbos, D.A.(1992). Multicriteria Decision Making. *Chemom.Intell.Lab.Syst.*, *16*, 175-191.

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

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

Keller, H.R., Massart, D.L. (1991). Multicriteria Decision Making: a case study. *Chemom.Intell.Lab.Syst.*, *11*, 175-189.

Kendall, M.G. (1948). Rank Correlation Methods.Charles Griffin & Co., London., 195, 202-204.

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

Klein, D.J. and Bytautas, L. (2000). Directed Reaction Graphs as Poset. *Communications in Mathematical and Computer Chemistry*. 42, 261-290.

Klir, G.J. and Folger, T.A. (1988). *Fuzzy Sets, Uncertainity, and Information*. Prentice-Hall, Englewood Cliffs (NJ), 356 pp.

Kruskal, J.B. and R.N. Shepard. (1974). A Nonmetric Variety of Linear Factor Analysis. *Psycometrika*, *39*, 123- 157.

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

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

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.

Lewi, P.J., Van Hoof, J., Boey, P. (1992). Multicriteria Decision Making Using Pareto Optimality and PROMETHEE Prefernce Ranking. *Chemom.Intell.Lab.Syst.*, *16*, 139-144.

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.

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

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.

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

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

Munzer, B., Brüggemann, R., Halfon, E. (1994). An Algebaic/Graphical Tool to Compare Ecosystems with Respect to their Pollution II: Comparative Regional Analysis. *Chemosphere*, *28*, 873-879.

Myrdal, P., Ward, G. H., Dannenfelser, R-M., Mishra, D., Yalkowsky., S. H. (1992).AQUAFAC 1: Acqueous functional group activity coefficients: Application to hydrocarbons. *Chemosphere*, *24*:1047-1061.

Newman, A. (1995). Ranking Pesticides by Environmental Impact. *Environ. Sci. Technol., 29*, 324-326.

Onicescu, O. (1966). Energie informationelle. *C.R.Acad.Sci.*, Paris, *263 – Ser.A*, 841-842.

Oppernhuizen, A., Hutzinger, O. (1982). Multicriteria Analysis and Risk Assesment. *Chemosphere.*, *11*, 675-678.

Patil, G. S. (1991). Correlation of aqueous solubility and octanol-water partition coefficient based on molecular structure. *Chemosphere*, *22*(8), 723-38.

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, The Netherlands; pp. 339-353.

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

Pudenz, S., Brüggemann, R., Komoβa, 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.

Pudenz, S., Bittner, T., Brüggemann, R. (1999). Comparative Evaluation of Materials in Car Production. In *Order Theoretical Tools in Environmental Sciences*; Proceedings of the Second Workshop, 95-104.

Pudenz, S., Brüggemann, R., Luther, B., Kaune, A., Kreimes, K. (2000). An Algebaic/Graphical Tool to Compare Ecosystems with Respect to their Pollution V: Cluster Analysis and Hasse Diagrams. *Chemosphere*, *40*, 1373-1382.

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

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

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

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, Germany.

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.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Univesrity of Illinois Press, Urbana (ILL).

Sicheri, G., Borsarelli, S.M. (1989). *Scienza dell'Alimentazione*, Hoepli, Milano (Italy).

Smilde, A.,K. A., Knevelmann, P.M.J. (1986). Introduction of Multicriteria Decision Making in Optimisation Procedures for High-performance Liquid Chromatographic Separations. *Journal of Chromatography*., *369*, 1-10.

Sørensen, P.B., Mogensen, B.B., Gyldenkærne, S., Rasmussen, A.G. (1998). Pesticides Leaching Assessment Mathod for Ranking Both Single Substances and Scenarios of Multiple Substance Use. *Chemosphere*, *36*, 2251-2276.

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.

Swanson, M.B., Davis, G.A., Kincaid, L.E., Schultz, T.W., Bartmess, J.E., Jones, S.L., George, E.L. (1997). A screening method for ranking and scoring chemicals by potential human helath and environmental impacts. *Environmental Toxicology and Chemistry*, *16*, 372-383.

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

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.

Todeschini, R., Consonni, V., Maiocchi, A. (1998). The K Correlation Index: Theory Development and its Applications in Chemometrics. *Chemom.Intell.Lab.Syst.*, *46*, 13-29.

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

Todeschini, R., Consonni, V., Mauri, A., Pavan, M. (2003). DRAGON, rel. 4.0 for Windows; Talete srl: Milano, Italy.

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

Todeschini, R., Consonni, V., Mauri, A., Pavan, M. (2003). MobyDigs: Software for Regression and Classification Models by Genetic Algorithms. Chapter 5 in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier.

Urrutia, J. (1987). Partial Orders and Euclidian Geometry, in Rival I. (ed.) *Algorithms and Order,* NATO ASI Series, Series C: Mathematical and Physical Science Vol. 255. Kluwer Academic Publishers, Dordrecht, S., p. 387-434.

Voigt, K., Gastaiger, J., Brüggemann, R. (2000). Comparative Evaluation of Chemicals and Environmental Online and CD-ROM Databases. *J. Chem. Inf. Comput. Sci.*, *40*, 44-49.

Voigt, K., Welz, G., Rediske, G. (1999). Environmental Approaches to Evaluate Internet Databases In *Order Theoretical Tools in Environmental Sciences*; Proceedings of the Second Workshop on Order Theoretical Tools in Environmental Sciences, 135-144.

Walker, J.D. and Carlsen, L. (2002). QSARs for Identifying and Prioritizing Substances with Persistence and Bioconcentration Potential. *SAR and QSAR in Environmental Research*, *13*, 713-725.

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

Welzl, G., Voigt, K., Rediske, G. (1998). Visualisation of Environmental Pollution – Hasse Diagram Technique and Explorative Statistical Methods. In Group Pragmatic Theoretical Ecology (Ed). Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences, 101-110.