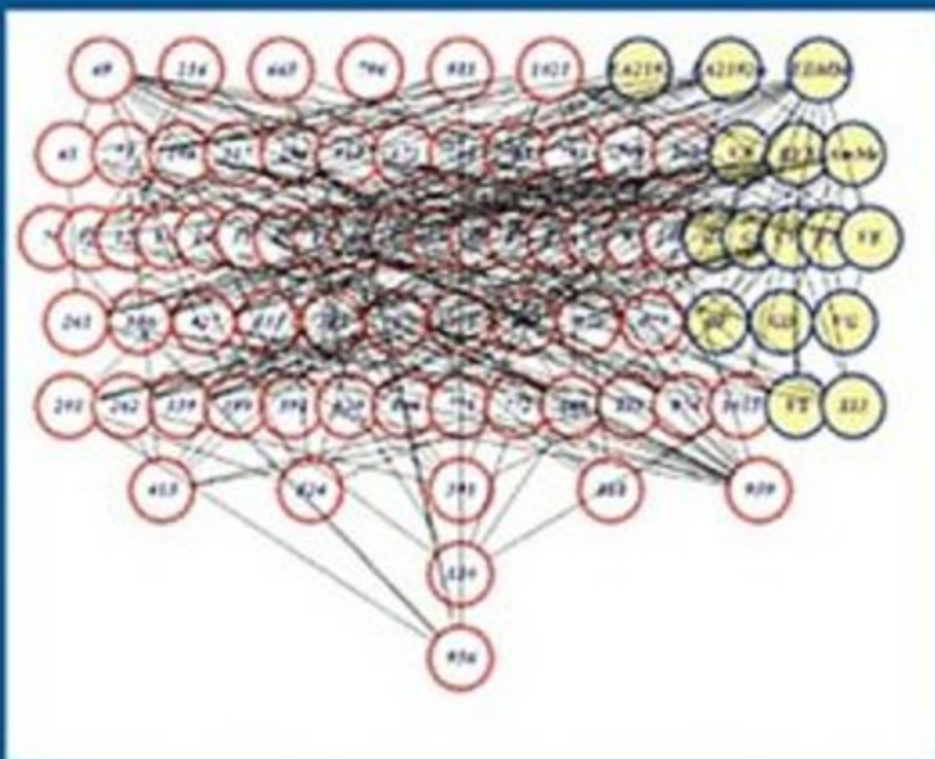




# Scientific Data Ranking Methods: Theory and Applications

MANUELA PAVAN  
ROBERTO TODESCHINI



---

*Data Handling in*  
**SCIENCE AND TECHNOLOGY**

**SCIENTIFIC DATA RANKING METHODS:  
THEORY AND APPLICATIONS**

VOLUME **27**

---

Edited by

MANUELA PAVAN

*Institute for Health and Consumer Protection  
Joint Research Centre  
European Commission  
Ispra, Italy*

ROBERTO TODESCHINI

*Department of Environmental Sciences  
University of Milano-Bicocca  
Milano, Italy*



ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford  
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier  
Linacre House, Jordan Hill, Oxford OX2 8DP, UK  
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2008

Copyright © 2008 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: [permissions@elsevier.com](mailto:permissions@elsevier.com). Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

#### Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

#### British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

#### Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-53020-2

ISSN: 0922-3487

For information on all Elsevier publications  
visit our website at [elsevierdirect.com](http://elsevierdirect.com)

Printed and bound in Hungary

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER

BOOK AID  
International

Sabre Foundation

## CONTRIBUTORS

*Numbers in parenthesis indicate the pages on which the authors' contributions begins.*

*Davide Ballabio, Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (111,169,193)*

*Rainer Brüggemann, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Mueggelseedamm 310, 12587 Berlin, Germany (73)*

*Sergio Canobbio, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (169)*

*Lars Carlsen, Awareness Center, Hyldeholm 4, Veddelev, DK-4000 Roskilde, Denmark (97, 139)*

*Viviana Consonni, Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (193)*

*Alberto Manganaro, Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (193)*

*Andrea Mauri, Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (111, 193)*

*Valeria Mezzanotte, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy (169)*

*Wayne L. Myers, School of Forest Resources, The Pennsylvania State University, University Park, PA 16802, USA (159)*

*M. Cruz Ortiz, Department of Chemistry, Faculty of Science, University of Burgos, Spain (1)*

*Ganapati P. Patil, Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA (159)*

*Manuela Pavan, Consumer Products Safety & Quality, Institute for Health and Consumer Protection, Joint Research Centre, European Commission, Via E. Fermi 2749, 21027 Ispra (VA), Italy (51, 169, 193)*

Stefan Pudenz, *Westlakes Scientific Consulting Ltd., Moor Row, Cumbria CA24 3LN, United Kingdom* (73)

M. Sagrario Sánchez, *Department of Mathematics and Computation, Faculty of Science, University of Burgos, Spain* (1)

Luis A. Sarabia, *Department of Mathematics and Computation, Faculty of Science, University of Burgos, Spain* (1)

Roberto Todeschini, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milan, Italy* (51, 193)

Kristina Voigt, *GSF-National Research Centre for Environment and Health, Ingostaedter Landstr. 1, 85764 Neuahrberg, Germany* (73)

## PREFACE

The intrinsic complexity of the systems analysed in scientific research together with the significant increase of available data require the availability of suitable methodologies for multivariate statistics analysis and motivate the endless development of new methods. Moreover, the increase 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. The different kinds of order ranking 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. Total and partial ordering methods are described in several mathematical books, requiring different degrees of mathematical skills.

Our intention in writing this book has been to provide a comprehensive and widely accessible overview of the basic mathematical aspects of the total and partial ordering methods by a didactical approach and to explain their use by examples of relevant applications in different scientific fields.

In fact, in recent years, ranking methods have been applied in several different fields such as decision support, toxicology, chemical prioritization, environmental hazard, proteomics and genomics, analytical chemistry, food chemistry and quantitative structure–activity relationship (QSAR). Moreover, new researches based on ranking methods are under investigation providing new perspectives for DNA sequence comparisons and for analytical data pattern recognition.

Being usually based on simple algorithms, ranking methods (both total and partial rankings) can be easily understood and successfully applied, resulting in a new appealing multivariate computation tool.

The integration of the theory and application of ranking methods has been of central concern to us in writing this book, based on the idea that the constant development of the ranking field strongly depends on this synergy.

Thus, the first chapter provides an extensive overview of the basic theory of ranking methods in statistics; it gives clear definition of order relations and it covers different aspects of the “order” concept in statistics, including random variable, order statistics, non-parametric methods and rank-based methods. The direct link between order and graphs and the one between order and optimization problems is also presented. The following two chapters are intended to

illustrate the fundamental basis of the mostly known total and partial ordering methods.

A number of different and up-to-date interest applications of order ranking methods are described in the following chapters. Examples on the use of the Hasse diagrams partial ranking method for the evaluation of chemicals and of databases are provided in more detail together with its use for prioritizing polluted sites. A new similarity/diversity measure is also described as a new approach for the analysis of sequential data, where useful information is obtained by the ordering relationships between the sequence elements. The advantageous interplay between partial order ranking and QSAR is illustrated by selected examples from recent studies, including risk assessment, selecting safer alternatives and as a tool in the process of suggesting new substances with specific characteristics.

Furthermore, a case study on the application of total order ranking methods to river functionality assessment is illustrated with the purpose to generate information and to provide further understanding as a basis for of river restoration strategies.

Finally, a software tool called **DART (Decision Analysis by Ranking Techniques)** that implements most of the different ranking methods described above is presented.

We believe that the book can be of value to various researchers in several scientific fields such as

- Research centres and universities
- Governmental organizations
- Pharmaceutical companies
- Health control organizations
- Environmental control organizations

We hope this book will expand the **knowledge and application of order ranking methods**.

Manuela Pavan and Roberto Todeschini  
*April 2008*

## Introduction to Ranking Methods

L.A. Sarabia, M.S. Sánchez, and M.C. Ortiz

---

Contents	1. Definition of order relations	1
	2. Order in Statistics	3
	2.1 Random variables	3
	2.2 Order statistics	7
	2.3 Non-parametric methods	9
	2.4 Rank-based methods	17
	3. Order in graphs	42
	4. Order in optimization problems	43
	References	46
	Appendix A	48

---

### 1. DEFINITION OF ORDER RELATIONS

From the mathematical point of view, an ordering (order relation)  $R$  in a set  $E$  is a binary relation among the elements in  $E$  that verifies:

- (i)  $u R u$ , for each  $u$  in  $E$  (reflexive);
- (ii) if  $u R v$  and  $v R u$ , then  $u = v$  (antisymmetric);
- (iii) if  $u R v$  and  $v R w$ , then  $u R w$  (transitive).

A *total ordering* (or linear ordering or simple ordering) is, in addition, connected (complete), that is, every two members of the set are comparable (either  $u R v$  or  $v R u$ , for all  $u, v$  in  $E$ ), and thus enables every member to be ordered relative to every other, and that generates a unique “linear” chain.

If this is not so,  $R$  is called a *partial ordering* (that is, it is transitive and antisymmetric but not necessarily connected) and thus generates possibly different chains of comparable elements; members of distinct chains may be incomparable.

Typical examples are the relation “less than or equal to” in the real numbers which is a total order whereas the set inclusion is a partial order. Hence, the real



numbers form a unique chain of comparable elements (which is usually represented by the real line), whereas, for instance, the set of even natural numbers and the set of odd natural numbers are two incomparable sets; neither the set of odd numbers is included in the set of even numbers nor vice versa, so that they will be in different chains of comparable elements.

Let us consider an order and denote it as  $\leq$ , for simplicity. There are some special elements within such an order. One of the most important is the *least element* of a (sub)set  $S$ , which is an element  $u$  such that  $u \leq v$ , for all elements  $v \in S$ . A value that is less than or equal to all elements of a set of given values is called a *lower bound*. The infimum or greatest lower bound is the unique largest member of the set of lower bounds for some given set, and it is equal to its *minimum* if the given set has a least element.

Analogously, the *greatest element* of a subset  $S$  of a partially ordered set (poset) is an element of  $S$ , which is greater than or equal to any other element of  $S$ , that is,  $u$ , such that  $v \leq u$ , for all elements  $v$  of  $S$ ; an *upper bound* is a value greater than or equal to all of a set of given values. The unique smallest member of the set of upper bounds for a given set is the supremum or least upper bound, and it is equal to its *maximum* if the given set has a greatest member.

In that respect, note the difference between minimum (resp. maximum) and minimal (resp. maximal) element. An element in an ordered set is *minimal* (resp. *maximal*) when there is no element smaller (resp. greater) than it, that is, it is the least (resp. greatest) element of a chain. If we do not need to specify, we use the generic term *optimal*.

Least and greatest elements may fail to exist but, if they exist, least and greatest elements are always unique. However, there can be many optimal elements in a set and some elements may be both maximal and minimal; thus a minimal (resp. maximal) element may not be the unique least (resp. greatest) element unless the order relation is a total order. Under total order relations, both terms coincide.

In that sense, an order in which every non-empty subset has at least one minimal element is an *inductive order*, whereas an ordering is a *well ordering* if every non-empty subset has a least member under the ordering, i.e. a unique minimal member that has the given relation to all members of the subset. A well order is an inductive order but not necessarily an inductive order is a well order.

There are some more properties that characterize different kinds of orders, see, for instance (Frank and Todeschini, 1994), and there are also some variations in the use of the terms. We have restricted ourselves to the most standard uses and define only the terms that we will use later on.

The fact is that it is not always possible to define a total order in a given set, and this fact affects in many scopes, because rankings (orderings) are used to compare nearly all variables that can be quantified in the interest of demonstrating differences. In general, many approaches have been made to jointly consider the information contained in the quantified variables and *summarize* it into one unique real number, giving rise to the so-called ranking methods. These can be a weighting of the characteristics, scaling and computing some statistics on them, etc. In (Allen and Sharpe, 2005) a case study is used to demonstrate the challenges

of creating a valid ranking structure, and references to different ranking methods are given.

## 2. ORDER IN STATISTICS

The statistical analysis based on the distribution of the ranks (order of the experimental values) has had an increasing development, passing from being a subject treated in the last chapter of books on applied statistics to being object of monographs. In 1956, Siegel published the first book (Siegel, 1956) dedicated to methods not based on normality, one of the most referenced books in statistics. From 1970, it is estimated that annually at least a book on non-parametric methods is published in which the methods based on ranks are a central subject. Some of these books are of interest for the users of the non-parametric statistics and have served as inspiration to write up this paragraph. Among the “classic ones”, advisable books are the aforementioned by Siegel and the one by Kendall (1975), whose first edition is of 1948. Of practical character, usable by researchers with a basic formation in statistics, are the books by Sprent (1989) and by Conover (1999). The books by Lehmann (1975) and the one by Hettmansperger (1984) are centred exclusively in the methods based on ranks. A recent text that includes the last investigations in the subject but with an advanced level is the one by Govindarajulu (2007).

### 2.1 Random variables

Outcomes associated with an experiment may be numerical in nature, such as quantity in an analytical sample. The types of measurements are usually called *measurement scales* and are, from the weakest to the strongest, nominal, ordinal, interval and ratio scale.

The *nominal scale* of measurement uses numbers merely as a means of separating the properties or elements into different classes or categories, for example, the sites of a study about contamination.

The *ordinal scale* refers to measurements where only the comparisons “greater”, “less” or “equal” are relevant, for example, the level of contamination: contamination in *A* is higher than in *B*, and contamination in *B* is higher than in *C*. If some of the values are equal to each other, we say ties exist.

The *interval scale* considers not only the relative order of the values but also the size of the interval between measurements as pertinent information. The *interval scale* involves the concept of a unit distance, and the distance between any two measurements may be expressed as a number of units, for example, the temperature. The actual value is merely a comparison with a reference value (the zero in scale) measured in units. A change in scale or location or both does not alter the principle of interval measurements.

The *ratio scale* is used when not only the order and interval size are important but also the ratio between two measurements has significance. It has sense to say that a measurement is twice or three times greater than another one. The ratio

scale is appropriate for measurements such as yields, quantities, weights and so on. The only distinction between the ratio scale and the interval scale is that the ratio scale has a natural measurement that is called zero, while the zero is arbitrarily defined in the interval scale.

Most of the usual parametric statistical methods require an interval (or stronger) scale of measurement. Most non-parametric methods assume either the nominal or the ordinal scale to be appropriate. Of course, each scale has all of the properties of the weaker measurement scales, therefore statistical methods requiring only a weaker scale may be used with the stronger scales also.

A *random variable* is a function that assigns real numbers to the outcomes of an experiment or observation. We will usually denote random variables by capital letters,  $X, Y, T$ , with or without subscripts. The real numbers attained by the random variables will be denoted by lowercase letters. For example, if we have a sample of wastewater and we apply an analytical procedure to determine the content of triazines, the result is a random variable. If the procedure is applied to  $n$  aliquot samples, we obtain  $n$  outcomes,  $x_1, x_2, \dots, x_n$  that are not equal. The variability of the results caused by the analytical procedure is a characteristic of it and is modelled by means of a random variable.

A random variable is completely specified by its *cumulative distribution function* (cdf)  $F_X(x)$ , that is, the probability of the random variable being less than or equal to  $x$  for any value  $x$ . Symbolically, this is written as  $F_X(x) = \text{pr}\{X \leq x\}$  for any real value  $x$ . In most of the applications, it is assumed that  $F_X(x)$  is differentiable, which implies, among other things, that none of the possible outcomes has positive probability, that is, the probability of obtaining exactly a specific value is zero.

In the case of a differentiable distribution function, the derivative of  $F_X(x)$  is the *probability density function* (pdf)  $f_X(x)$ . Any function  $f(x)$  such that: (i) it is positive,  $f(x) \geq 0$  and (ii) the area under the function is one,  $\int_{\mathbb{R}} f(x)dx = 1$ , is the *probability density function* of a random variable.

The probability of the random variable  $X$  being in interval  $[a, b]$  is the area under the pdf over the interval  $[a, b]$ , that is

$$\text{pr}\{X \in [a, b]\} = \int_a^b f(x)dx \quad (1)$$

and the mean and the variance of  $X$  are written as

$$E(X) = \int_{\mathbb{R}} xf(x)dx \quad (2)$$

$$V(X) = \int_{\mathbb{R}} (x - E(X))^2 f(x)dx \quad (3)$$

These expressions are adapted in the obvious way for discrete random variables, that is, a random variable  $X$  that takes discrete values,  $x_i, i \in I$ . The set  $I$  can be a finite set, for example,  $I = \{0, 1, 2, 3\}$ , or an infinite one (a numerable set),

$I = N = \{1, 2, 3, \dots\}$ , but always totally ordered. In both cases, we speak about the *probability function* instead of the *probability density function*. The probability function is greater than zero only for the values that the random variable takes, i.e.  $f_X(x_i) = p_i = \text{pr}\{X = x_i\} > 0$  while  $f_X(x) = 0$  if  $x \neq x_i \forall x_i$ . Also,  $\sum_i p_i = 1$  must hold. The probability of a discrete variable  $X$  being in interval  $[a, b]$  is thus the sum of the probabilities associated with the values  $x_i$  in  $[a, b]$ , that is

$$\text{pr}\{X \in [a, b]\} = \sum_i p_i \text{ for } x_i \in [a, b] \quad (4)$$

and the mean and the variance of  $X$  are

$$E(X) = \sum_i x_i p_i \quad (5)$$

$$V(X) = \sum_i (x_i - E(X))^2 p_i \quad (6)$$

### Example 1

Consider the finite (discrete) random variable  $X$  that takes the values  $x_i$  with probabilities  $p_i$  written in Table 1. Its cumulative probability function is right continuous in  $x_i$  and constant in the intervals  $[x_i, x_{i+1})$

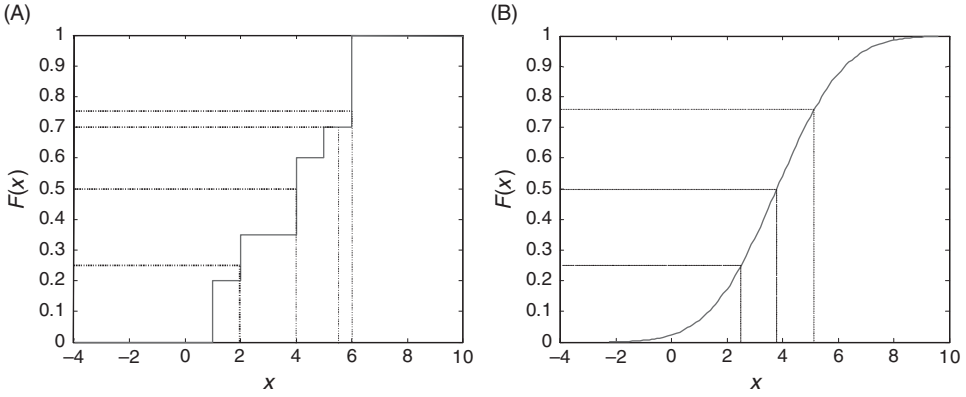
$$F(x) = \begin{cases} 0, & \text{if } x < 1 \\ 0.20, & \text{if } 1 \leq x < 2 \\ 0.35, & \text{if } 2 \leq x < 4 \\ 0.60, & \text{if } 4 \leq x < 5 \\ 0.70, & \text{if } 5 \leq x < 6 \\ 1.00, & \text{if } 6 \leq x \end{cases} \quad (7)$$

This function is drawn in Figure 1A where the “jumps” corresponding to the values  $x_i$  have been joined with vertical lines. A simple calculation applying Eqs (5) and (6) to the data in Table 1 gives  $E(X) = 3.8$ ,  $V(X) = 3.66$  and, thus, the standard deviation is  $\sqrt{3.66} = 1.91$ .

Figure 1B shows the cdf of a normal distribution (continuous distribution) with the same mean and standard deviation as  $X$ , that is, a  $N(3.8, 1.9)$ . Its cdf is given by

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right) du, \text{ where } \mu = 3.8 \text{ and } \sigma = 1.9 \quad (8)$$

When several random variables are defined jointly or when several experiments are considered as a combined experiment, each with its own one or more random



**Figure 1** (A) Cumulative probability function of the discrete random variable of Table 1 (Eq. (7)). (B) Cumulative distribution function of a normal distribution with the same mean, 3.8, and standard deviation, 1.91.

**Table 1** Values  $x_i$  of a discrete random variable and probabilities  $p_i = \text{pr}\{X = x_i\}$

$x_i$	1	2	4	5	6
$p_i$	0.20	0.15	0.25	0.10	0.30

variables, it becomes useful to consider joint distributions, described by *joint probability functions* (discrete case), which are defined as follows:

$$f(x_1, x_2, \dots, x_n) = \text{pr}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad (9)$$

The *joint distribution function*  $F(x_1, x_2, \dots, x_n)$  of the continuous random variables  $X_1, \dots, X_n$  is defined by means of the following equation:

$$F(x_1, x_2, \dots, x_n) = \text{pr}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \quad (10)$$

The random variables  $X_1, \dots, X_n$  are independent if

$$f(x_1, x_2, \dots, x_n) = f_{x_1}(x_1) \times \dots \times f_{x_n}(x_n) = \text{pr}(X_1 = x_1) \times \dots \times \text{pr}(X_n = x_n) \quad (11)$$

or, for continuous random variables

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= \text{pr}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \\ &= \text{pr}\{X_1 \leq x_1\} \times \text{pr}\{X_2 \leq x_2\} \times \dots \times \text{pr}\{X_n \leq x_n\} \\ &= F_{X_1}(x_1) \times F_{X_2}(x_2) \times \dots \times F_{X_n}(x_n) \end{aligned} \quad (12)$$

Given two continuous random variables,  $X$  and  $Y$ , the covariance between them is defined as

$$\text{cov}(X, Y) = E((X - E(X))(Y - E(Y))) = \iint_{\mathbb{R}^2} (x - E(X))(y - E(Y))f(x, y)dx dy \quad (13)$$

where  $f(x, y)$  is the joint density function. If the two random variables  $X$  and  $Y$  are independent, then  $\text{cov}(X, Y) = 0$  because the double integral in Eq. (13) becomes the product of two integrals, both equal to zero.

The cdf of a random variable contains all the information of interest, that is, the values it takes and the probabilities of taking them. However, in many occasions, some parameters are used to have a “summary description” of the random variable.

The use of *quantiles* is rather frequent. The  $p$ th quantile of a random variable  $X$  is the value  $x_p$  such that  $\text{pr}\{X < x_p\} \leq p$  and  $\text{pr}\{X > x_p\} \leq 1 - p$ . If more than one value satisfies the conditions, we will avoid confusion adopting the convention that  $x_p$  equals the average of the largest and the smallest numbers that satisfy the definition.

Special interest, among the quantiles, is to the upper and lower *quantiles*, which are the 0.75 and the 0.25 quantiles, respectively, and to the median which is the 0.50 quantile. Figure 1A shows that the lower quartile is  $x_{0.25} = 2$  for the discrete variable  $X$  because  $\text{pr}\{X < 2\} = 0.20 < 0.25$  and  $\text{pr}\{X > 2\} = 0.65 < 0.75$  (Table 1). Analogously, the median is  $x_{0.50} = 4$  and  $x_{0.75} = 6$ . In contrast,  $x_{0.25} = 2.52$ ,  $x_{0.50} = 3.8$  and  $x_{0.75} = 5.08$  for the normal distribution (Figure 1B). Because of the symmetry of the normal distribution, the mean and the median coincide.

When the distribution is continuous, the quantiles always have a unique value and are obtained as  $x_p = F^{-1}(p)$ . For discrete distributions, the value is not necessarily unique; for example, in Figure 1A, we can see that any value  $x$  in the interval  $(5, 6)$  verifies that  $\text{pr}\{X < x\} < 0.7$  and  $\text{pr}\{X > x\} = 0.3$ , which means that any value of this interval is the 0.70 quantile. According to the convention aforementioned,  $x_{0.70} = 5.5$ .

## 2.2 Order statistics

Let  $X_1, X_2, \dots, X_n$  be a random sample (independent and identically distributed random variables as a given  $X$ ) of size  $n$ . If these variables are arranged in the order of magnitude, but not in the order in which they come, as

$$X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$$

then  $(X_{1,n}, X_{2,n}, \dots, X_{n,n})$  is called the *order statistics*. Note that the  $X_{i,n}$  is neither independent nor identically distributed.

If the random sample of  $n$  observations is arranged in order of increasing magnitude and if the smallest observation is assigned the value of 1, second smallest the value of 2, and the largest the value of  $n$ , then we have  $(r_1, r_2, \dots, r_n)$  where  $r_i$  is the position that  $X_i$  occupies in the ordered sequence. The  $r_i$  are called the *ranks*.

### Example 2

Consider  $(X_1, X_2, X_3, X_4) = (2, -1.2, 0.4, 4.2)$ . The ordered observations will be  $-1.2 < 0.4 < 2 < 4.2$ . Hence,  $r_1 = 3$ ,  $r_2 = 1$ ,  $r_3 = 2$  and  $r_4 = 4$ .

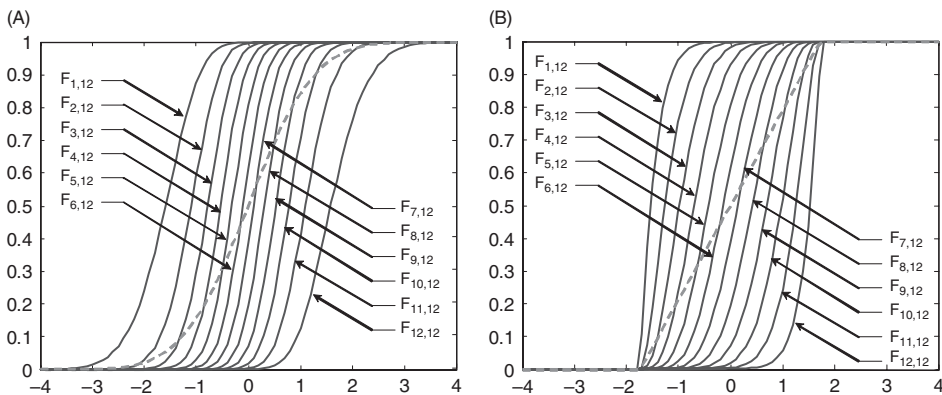
Note that  $(r_1, r_2, \dots, r_n)$  is a permutation of the integers  $(1, 2, \dots, n)$ . If  $(X_1, X_2, \dots, X_n)$  is replaced by  $(R_1, R_2, \dots, R_n)$ , where  $R_i$  is the rank of  $X_{i,n}$ , then  $(R_1, R_2, \dots, R_n)$  is called the *rank order* and it is a random variable.

Order statistics play a dominant role in non-parametric statistics and it is possible to compute their distribution function. Let  $X$  be a random variable with cdf  $F_X(x) = \text{pr}\{X \leq x\}$ , and it is supposed to be continuous, that is,  $\text{pr}\{X = x\} = 0$  for any real number  $x$ . If the random variables  $X_1, X_2, \dots, X_n$  are equally distributed and equal to  $X$ , the distribution function of the order statistics  $X_{i,n}$  is

$$F_{X_{i,n}}(x) = \text{pr}\{X_{i,n} \leq x\} = \sum_{k=i}^n \binom{n}{k} F^k(x) (1 - F(x))^{n-k} \quad (14)$$

Figure 2 shows two cases. If the distribution of  $X$  is a standard normal,  $N(0,1)$ , with sample size  $n=12$ , the distributions of  $X_{i,12}$ ,  $i=1, 2, \dots, 12$  are drawn in Figure 2A, where for comparative purposes, the  $N(0,1)$  has also been drawn. Clearly, the distributions are different. For Figure 2B, a uniform distribution in the interval  $[-\sqrt{3}, \sqrt{3}]$  has been assumed for  $X$  in such a way that  $E(X) = 0$  and  $V(X) = 1$  as in Figure 2A. The difference in the distributions of the order statistics in relation to the previous case is very appreciable.

The joint distribution for any pair  $X_{i,n}$  and  $X_{j,n}$  (that are not independent), the one for all of them, and those of the *sample range* (defined as  $R = X_{n,n} - X_{1,n}$ ), *sample mid-range* and *sample median* can be consulted in Chapter 2 of Govindarajulu (2007). When the random variable  $X$  is discrete, the expression for the distribution function (that now is discontinuous) is formally equal to the one of Eq. (14). A compilation of results and the representation of the order statistics by means of exponential functions can be consulted in the same chapter of the book by



**Figure 2** Distribution functions of order statistics  $X_{1,12}, X_{2,12}, \dots, X_{12,12}$ . (A) For a standard normal distribution  $N(0,1)$ , dashed line. (B) For a uniform distribution in the interval  $[-\sqrt{3}, \sqrt{3}]$ , dashed line.

Govindarajulu just mentioned. It also contains the details of the solution to the Angel's problem empirically solved by Youden (1953) and analytically by Kendall (1954), who posed it in the following way: "A number of laboratory assistants are given a standard experiment to perform. They replicate it and, knowing what the true result ought to be, each submits only its best result (the nearest to the true value). What effect does this have on estimates of experimental error?" If the values are distributed as  $N(0,1)$ , the required variance is approximately equal to  $\pi/((N-1)(N+4))$ , for example if  $N=5$  the variance is 0.087 instead of 1.

## 2.3 Non-parametric methods

Frequently, the statistic courses are limited to study models under the normality hypothesis. Equation (8) showed that the normal distribution depends solely on two parameters  $\mu$  and  $\sigma$ , which are its mean and standard deviation, respectively. Among the discrete distributions, the binomial is also characterized by two parameters: the number of observations,  $n$ , and the probability,  $p$ , of one of the two possible outcomes at each observation (often called success and failure). Given a random sample (independent observations,  $x_1, x_2, \dots, x_n$ ) of a random variable that belongs to a family of distributions described by parameters (parametric distribution), the problem of deciding which is the distribution is reduced to estimate the parameters that define it from the sample. For example, the sample mean

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

is an estimate of the (population) mean  $\mu$  of a normal distribution. With a sample from a normal distribution, the  $t$ -test may be used to decide if the sample is consistent with an a priori hypothesized population mean  $\mu_0$ . The related  $t$ -distribution lets us establish a *confidence interval*, i.e. an interval in which we are reasonably confident that the true although unknown mean  $\mu$  lies.

If in  $n$  observations there are  $r$  "successes", the ratio  $r/n$  is a *point estimate* of parameter  $p$  in a binomial distribution, and we may test whether that estimate supports an a priori hypothesized value  $p_0$  or we may obtain a confidence interval for the true value of  $p$ .

Sometimes it can be reasonable to suppose that the experimental values come from a parametric family of distributions. But sometimes inferences that do not have any relation with a parameter are desired, may be because only the order of some observations is at hand, not their quantitative values. For example, the level of damage in a tissue when a pathological analysis is made, the level of contamination of a territory, the proportion of products that possess a property in a production process in batches, etc. Even if quantitative measures in interval or ratio scale are available, maybe little is known about the distribution, perhaps that it is skew or symmetric or some other characteristic. In these situations, the so-called non-parametric methods of inference, better called distribution-free, are appropriate, many of which are based on ranks.

Although Spearman proposed in 1904 a rank correlation coefficient, that takes his name, the research on non-parametric and distribution-free methods received



a great impulse with the works by Fisher, Pitman and Welch on randomization or permutation tests, tests that at that time (the 1930s) were computationally too demanding for general use. At the same time, the interest for working with data in nominal or ordinal scales grows so that the inferential techniques demand little computational effort. Few years later, [Wilcoxon \(1945\)](#) and [Mann and Whitney \(1947\)](#) published the first non-parametric procedures. For the next decade, non-parametric tests for location were studied using Pitman's asymptotic efficiency to assess their local power properties. In a series of papers, Hodges and Lehmann discovered the surprising result that rank tests suffer negligible loss when compared to the  $t$ -test at the normal model and may be much more efficient at heavy-tailed models. As we have mentioned, the first book on applied non-parametric statistics was published in 1956 by [Siegel \(1956\)](#). In the 1960s, Hodges and Lehmann derived point estimates and confidence intervals for rank test statistics. They further showed that the estimation methods inherit their efficiency properties from the parent test statistics. It was also found that these estimates are robust according to the new criteria proposed by Tukey, Huber and Hampel for assessing stability of estimates. Aligned rank test for analysis of designed experiments was introduced in the early 1960s by Hodges and Lehmann. Adchie proposed and studied rank tests and the corresponding estimates for simple regression models. In the 1970s, the previous work was consolidated and extended to rank-based tests and estimates in the linear model.

A disadvantage of non-parametric methods in the pre-computer era was that simplicity applied only to basic procedures and non-parametric methods lacked the flexibility of much linear models and least-squares theory, which are the key of normal distribution-based inference. The advent of computers has revolutionized this aspect of using non-parametric methods, for many advanced and flexible methods are tedious only in that they require repeated application of simple calculations, a task for which computers are admirably suited and easily programmed.

### 2.3.1 Hypothesis tests. The signs test

The basic elements of a hypothesis test will be introduced by an example.

#### Example 3

In a geographic zone, 12 samples of water have been taken. We want to test the hypothesis that the median,  $\theta$ , of the concentration of triazines is  $50 \mu\text{g L}^{-1}$ , which is the permitted limit (maximum concentration allowed) by some European legal norms for industrial residual waters. We check in each of the samples whether or not the quantity of triazines exceeds  $50 \mu\text{g L}^{-1}$ . In nine samples it does, while in the remaining three it does not. We record this as nine pluses (+) and three minuses (−).

Formally, the hypothesis is stated as follows:

$$\begin{aligned} H_0: \theta &= 50 \mu\text{g L}^{-1} \\ H_a: \theta &> 50 \mu\text{g L}^{-1} \end{aligned} \tag{15}$$

The statement  $H_0: \theta = 50 \mu\text{g L}^{-1}$  in [Eq. \(15\)](#) is called the *null hypothesis*, and the statement  $H_a: \theta > 50 \mu\text{g L}^{-1}$  is called the *alternative hypothesis*. As the

alternative hypothesis specifies values of  $\theta$  that are greater than  $50 \mu\text{g L}^{-1}$ , it is called *one-sided alternative*. In some situations, we may wish to formulate a *two-sided alternative* hypothesis to specify values of  $\theta$  that could be either greater or less than  $50 \mu\text{g L}^{-1}$  as in

$$\begin{aligned} H_0 : \theta &= 50 \mu\text{g L}^{-1} \\ H_a : \theta &\neq 50 \mu\text{g L}^{-1} \end{aligned} \quad (16)$$

The hypotheses are not affirmations about the sample but about the distribution from which these values come. In particular, if the median of the distribution is  $50 \mu\text{g L}^{-1}$ , half of the samples must have a greater concentration and the other half less concentration than this quantity, according to the definition of median (quantile  $x_{0.50}$ ).

In general, to conduct a *test of a hypothesis*, the researcher must consider the experimental objective and decide upon a null hypothesis for the test, as in Eq. (15). Hypothesis-testing procedures rely on using the information in a random sample; if this information is inconsistent with the null hypothesis, we would conclude that the hypothesis is false. If sufficient evidence does not exist to prove falseness, the test defaults to the conclusion that the null hypothesis cannot be rejected but does not actually prove that it is correct. It is therefore critical to carefully choose the null hypothesis in each problem.

In practice, to test a hypothesis, we must take a random sample, compute an appropriate test statistic from the sample data and then use the information contained in this statistic to make a decision. However, as the decision is based on a random sample, it is subject to error. Two kinds of potential errors may be made when testing hypothesis. If the null hypothesis is rejected when it is true, then a *type I error* occurs. A *type II error* is made when the experimenter fails to reject the null hypothesis when it is false. The situation is described in Table 2.

In Example 3, if the experimental data lead to reject the null hypothesis  $H_0$  being true, our (wrong) conclusion is that the site surpasses the legal threshold of contamination. A type I error has been made and an unnecessary cost in anti-pollutants actions will be derived. If, on the contrary, the experimental data lead to accept the null hypothesis when it is false, the researcher will not alert to administration when in fact the content in triazines surpasses the legal limit, a type II error is made. Note that both types of error have to be considered because their consequences are very different. In the case of type I error, an acceptable situation is not recognized with the corresponding extra cost. In contrast, the type II error implies that an unsuitable situation is accepted, with the later damages

**Table 2** Decisions in hypothesis testing

Researcher's decision	The unknown truth	
	$H_0$ is true	$H_0$ is false
Accept $H_0$	No error	Type II error
Reject $H_0$	Type I error	No error

that it may cause (e.g. sanitary). It is clear that the researcher has to specify the risk that assumes to commit these errors by fixing the probability at which they occur.

These probabilities of occurrence of type I and type II errors are denoted by specific symbols:

$$\begin{aligned}\alpha &= \text{pr}\{\text{type I error}\} = \text{pr}\{\text{reject } H_0/H_0 \text{ is true}\} \\ \beta &= \text{pr}\{\text{type II error}\} = \text{pr}\{\text{accept } H_0/H_0 \text{ is false}\}\end{aligned}\quad (17)$$

The probability  $\alpha$  of the test is called the *significance level*, and the *power of the test* is  $1-\beta$ , which measures the probability of correctly rejecting the null hypothesis.

The next element in the hypothesis test is the *statistic*, a function of the sample data, thus a random variable whose distribution is computed under the null hypothesis. In our case, the statistic is

$$S = \#\{x_i > 50\} = \sum_i S(x_i) \quad (18)$$

where  $\#\{ \}$  denotes the cardinal, that is, the number of elements of set  $\{ \}$ , and  $S(X)$  is the sign function

$$S(x_i) = \begin{cases} 1, & \text{if } x_i - 50 > 0 \\ 0, & \text{otherwise} \end{cases}$$

Note that, if  $x_i = 50$ , the observation is ignored and dropped from the sample. Therefore,  $S$  counts the number of samples, among the 12 obtained, with concentration of triazines greater than  $50 \mu\text{g L}^{-1}$ . In our example,  $S = 9$ .

To make a decision, it is necessary to consider the *critical region* at significance level  $\alpha$ , that is

$$\text{CR} = \{x_i, i = 1, \dots, n/S \geq k\} \quad (19)$$

where  $k$  is the value (known as *critical value*) such that  $\text{pr}\{S \geq k/H_0\} = \alpha$ .

If the null hypothesis is true,  $50 \mu\text{g L}^{-1}$  is the median of the distribution and the probability that the concentration of a sample will be greater than the median is 0.5. Hence, under  $H_0$ ,  $S$  follows a binomial distribution with parameters  $n = 12$  and  $p = 0.5$ ,  $B(12, 0.5)$ . Table 3 shows the probability function of this binomial distribution.

Then, the critical value  $k$  at  $\alpha = 0.05$  must verify that  $0.05 = \text{pr}\{S \geq k/B(12, 1/2)\}$ . Taking into account that the distribution is discrete, not all the values are available. According to Table 3,  $\text{pr}\{S \geq 10\} = 0.019$  for  $k = 10$ , whereas  $\text{pr}\{S \geq 9\} = 0.073$  for  $k = 9$ . Therefore,  $k = 9$  will be taken as the critical value at 5% nominal significance level and 7.3% actual significance level. In practice, as the nominal and actual levels are not equal, it would be better to choose the actual level as close as possible to the nominal  $\alpha$ . In our case, the value of the statistic belongs to the critical region, so we reject the null hypothesis, that is, the median of the distribution of the content of triazines is greater than  $50 \mu\text{g L}^{-1}$ .

**Table 3** Binomial probabilities  $p_k = \text{pr}(B(n, p) = k)$  for  $k$  successes when  $n = 12$  and  $p = 1/2$ 

$k$	0	1	2	3	4	5	6
$p_k$	0.000	0.003	0.016	0.054	0.121	0.193	0.226
$k$	7	8	9	10	11	12	
$p_k$	0.193	0.121	0.054	0.016	0.003	0.000	

The critical region of a test is related to the form of the alternative hypothesis. For example, in the case of the two-sided alternative hypothesis (Eq. (16)), the statistic is the same but the critical region is

$$\text{CR} = \{x_i, i = 1, \dots, n/S \leq k_1 \text{ or } S \geq k_2\} \quad (20)$$

Because of the binomial distribution of  $S$  and the fact that the probabilities of such a binomial are symmetric ( $p = 0.5$ ), it is reasonable to take  $k_1 = k$  and  $k_2 = n - k$  and  $k$  is chosen such that

$$\frac{\alpha}{2} = \text{pr}\{B(n, 1/2) \leq k\} \text{ and } \frac{\alpha}{2} = \text{pr}\{B(n, 1/2) \geq n - k\}$$

for a  $100 \times \alpha\%$  significance level. In the case  $\alpha = 0.05$ , consulting Table 3 we see that the set from 3 to 9 has associated a total probability of 0.962, so, excluding these outcomes, reduces the probability below 0.05. The remaining outcomes  $\{0, 1, 2, 10, 11, 12\}$  are the CR at nominal level 0.05 and actual level 0.038. If the two-tail test were applied to the data of Example 3, as  $9 \notin \text{CR}$ , there would be no experimental evidence to reject the null hypothesis. In any case, note that in this problem, the alternative  $\theta \neq 50 \mu\text{g L}^{-1}$  has no sense.

In order to calculate the probability of the type II error, it is necessary to specify the alternative hypothesis, that is to say, what amount makes me saying that the median is greater than  $50 \mu\text{g L}^{-1}$ . The statistic  $S$  in fact depends on the quantile that we wish to specify, not on the value associated with this quantile. Thus, to specify what we call different from the median, we must do it in terms of the quantiles, for example,  $p = 2/3$  instead of  $p = 1/2$ . In other words, we wish to detect a quantile  $2/3$  of the median of the distribution with the data of Example 3. In these conditions,  $\beta$  is the probability of obtaining a value of the statistic outside the critical region when in fact the distribution is a  $B(12, 2/3)$ . The probability is 0.61, that is, there is a 61% of probability of accepting the null hypothesis when it is false.

As it is known, there is a relation between  $\alpha$ ,  $\beta$  and the sample size  $n$ . For example, by maintaining  $\alpha = 0.05$ , for  $n = 50$ ,  $\beta = 0.29$ , whereas for  $n = 100$ ,  $\beta$  decreases until 0.04. In both cases, it was supposed that the quantile  $x_{2/3}$  is to be distinguished from the median.

*The Pitman efficiency.* The efficiency of a test  $T_2$  relative to test  $T_1$  is the ratio  $n_1/n_2$  of the samples sizes needed to attain the same power for the two tests with

the chosen  $\alpha$ . In practice, we usually fix  $\alpha$ ; then  $\beta$  depends on the particular alternative as well as the sample sizes. Fresh calculations of relative efficiency are required for each difference that we desire to detect. Pitman (1948) considered sequences of tests  $T_1$  and  $T_2$  in which we first fix  $\alpha$  but allow the alternative to vary in such a way that  $\beta$  remains constant as the sample  $n_1$  increases. For each  $n_1$ , we determine  $n_2$  such that  $T_2$  has the same  $\beta$  for the particular alternative considered. Pitman found that in these sequences of tests, under very general conditions,  $n_1/n_2$  tended to a limit as  $n_1 \rightarrow \infty$ , and more importantly, that this limit was the same for all choices of  $\alpha$  and  $\beta$ . The limit is the *asymptotic relative efficiency* (ARE) of the two tests.

If the distribution is normal and we use the  $t$ -test, the ARE of the signs test is 0.64, that is, we need a sample size of 100 to attain the same  $\alpha$  and  $\beta$  that would be obtained with 64 data if we were using the more powerful  $t$ -test for the same data. If the parent distribution is uniform, the ARE is 0.33, but if we have a double exponential (that accumulates more probability in the tails than a normal), the ARE is 2.00, that is, we need half of the sample size with the signs test than for the double exponential distribution. In practice, we often have little idea of the distribution we are sampling from, but it is broadly true that if we suspect our sample comes from a long-tail distribution (i.e. with tails longer than the normal), we may do better for location tests or estimates with non-parametric tests.

### 2.3.2 Confidence intervals

A way of obtaining a  $100 \times (1-\alpha)\%$  confidence interval on a location parameter  $\theta$  is to define it as a set of all values,  $\theta$ , which would be accepted using a hypothesis test at significance level  $100 \times \alpha\%$ . For the two-tail signs test on the median using a sample size of 12 and a 5% level, we have already seen that we would accept  $H_0$  if we got between 3 and 9 plus signs, both included. If we know the sample values, we can establish a confidence interval.

#### Example 4

The data of [Example 3](#) are, in increasing order, 28, 32, 41, 52, 63, 64, 79, 83, 88, 92, 98 and  $103 \mu\text{g L}^{-1}$ . It is quite obvious that we would have between 3 and 9 plus signs for any value of  $\theta$  greater than 41 (rank 3) or less than 92 (rank 10). The open interval (41, 92) is thus a nominal 95% (actual 96.2%) confidence interval on the median. It is easy to compute the one-sided intervals in an analogous way.

The meaning of a  $100 \times (1-\alpha)\%$  confidence interval is that if confidence intervals are computed with samples of the same size and from the same distribution, on average, the  $100 \times (1-\alpha)\%$  of these intervals will contain the true median value.

### 2.3.3 Tolerance intervals

Given a random variable  $X$ , an interval  $[l, u]$  is a  $\beta$ -content tolerance interval at confidence level  $\gamma$  if the following is fulfilled:

$$\text{pr}\{\text{pr}\{X \in [l, u]\} \geq \beta\} \geq \gamma \quad (21)$$

In words,  $[l, u]$  contains at least  $100 \times \beta\%$  of the  $X$  values with  $\gamma$  confidence level. Given a sample of  $n$  observations  $x_1, x_2, \dots, x_n$  from a continuous distribution, the  $\beta$ -content tolerance interval is used to answer questions like how large should  $n$  be so that the probability is at least  $\gamma = 0.95$  that  $\beta = 0.90$  of the population lies between the smallest and the largest sample value? For answering this question, it is necessary (Kendall and Stuart, 1979) that  $n$  fulfils (approximately) the equation  $\log(n) + (n-1)\log(\gamma) = \log(1-\beta) - \log(1-\gamma)$ . If we wish  $\beta = \gamma = 0.95$ , the value of  $n$  has to be 89. The  $\beta$ -content tolerance intervals for both continuous and discrete random variables can be one-sided, or two-sided and can be obtained by controlling the centre or both tails. A revision of all variants can be seen in Patel (1986). Other details can be found in the book by Conover (1999) (Section 3.3).

### 2.3.4 Test for significance of changes and for trends

We have used the signs test to introduce the basic elements of a hypothesis test; nevertheless, the signs test has its own great importance because of its versatility. In essence, this test allocates a sign (of there its name) to each observation depending on whether it is greater or less than the hypothesized value and considers whether this is substantially different from what we would obtain just by chance. Therefore, it needs the data to be in ordinal scale. In the following, we will show two very useful variants of the signs test.

*Case 1. The McNemar test for changes.* Suppose data in nominal scale with two categories 0 and 1. We take a sample of size  $n$  from the two independent random variables  $X$  and  $Y$ , both taking only two possible values 0 and 1, which represent the two categories. The question is to decide if the probability of obtaining (1, 0) is different from the probability of (0, 1)

#### Example 5

Two analytical methods have been applied to 100 samples to decide whether they contain a specific analyte or not. Table 4 shows the results.

Reflection on the data of Table 4 immediately reveals that when both methods simultaneously detect or do not detect the analyte, there is no information on possible differences in their capability of detection. Nevertheless, in four samples, method B does not detect the analyte but method A does, so we can write down four plus signs for method A. In contrast, method A has not detected the presence of the analyte in 21 samples in which method B has detected the analyte, so we can

**Table 4** Contingency table for the detection of an analyte in 100 samples with two methods A and B

Method B	Method A	
	Detect	Non detect
Detect	63	21
Non detect	4	12

assign 21 minus signs for the method A. If the method A has less capability of detection, we expect less number of detections with it than with method B.

Formally, the test is posed as:

$H_0$ : Both methods have the same capability of detection.

$H_a$ : Method A has less capability of detection.

The procedure is the one of the signs test, now  $n = 25$  and there are 4 pluses and 21 minuses. It is a one-tail test,  $S = 21$  and  $CR = \{B(25, 0.50) \geq k\}$ . Consulting the  $B(25, 0.50)$  distribution, at 5% nominal significance level,  $k = 18$  (actual level 2.16%). As the value of the statistic belongs to the critical region CR, the null hypothesis is rejected, and the conclusion is that method A has less capability of detecting the analyte. Other details of this test, as well as approximations for large sample sizes, can be consulted in [Conover \(1999\)](#). The McNemar test is frequently used to evaluate the efficacy of a treatment to diverse items.

*Case 2. Cox and Stuart test for trend.* Suppose  $n'$  observations of a random variable ordered with a criterion,  $x_1, x_2, \dots, x_{n'}$ , for example, the order in which they were observed. The hypothesis that there is a trend in the sequence is to be tested. For that, the values are grouped in pairs  $(x_1, x_{1+c}), (x_2, x_{2+c}), \dots, (x_{n'+c}, x_n)$ , where  $c = n'/2$  if it is even and  $c = (n' + 1)/2$  if it is odd (if  $n'$  is odd, the value at the centre is eliminated). Each pair  $(x_i, x_{i+c})$  is substituted by a plus sign if  $x_i < x_{i+c}$  or by a minus sign if  $x_i > x_{i+c}$  removing ties. The number of untied pairs is called  $n$  and then the statistic  $S$  is equal to the number of pluses.

**Example 6**

The amount of nitrogen in the limb of the grapevine is an indicator of the nutritional state of the plant, which varies seasonally with the physiological state. [Table 5](#)

**Table 5** Annual evolution of the nitrogen content (percentage on dry matter) in the limb of the leaves of the grapevine

Year	Physiological state			
	Separated floral bellboys	Blooming	Fruit formation	Large grain great pea
2000	4.20	3.42	3.09	2.59
2001	4.03	3.67	3.13	2.53

Year	Physiological state			
	Veraison 50%	Veraison 100%	Maturity (harvesting)	Post-harvesting (whitering)
2000	2.48	2.26	2.09	1.83
2001	2.34	2.00	2.01	1.95

collects the content of nitrogen in 16 consecutive physiological states during years 2000 and 2001. The experiment tries to decide if the physiological cycle of the plant produces a significant, at 5%, diminution in the nitrogen availability for the following year. Therefore, the hypothesis to be tested is

$H_0$ : The quantity of nitrogen is the same in both cycles.

$H_a$ : The quantity of nitrogen is smaller in the second cycle.

The quantity of nitrogen is known to follow a yearly cycle related to the physiological state of the plant, so that nothing is learned by pairing the nitrogen quantities for two different physiological states. However, by pairing the same physiological state in two successive cycles, the existence of a trend can be investigated.

The test statistics  $S$  equals the number of pairs where the second cycle had a higher quantity of nitrogen than the first cycle, which is 3 in this example. Because the test is to detect a downward trend, the critical region is  $CR = \{S \leq k\}$  and for a  $B(8, 0.5)$  with  $k = 1$ , the probability is 0.0352. Therefore, at 5% nominal level (actual 3.5%), there is no evidence to reject the null hypothesis, that is, there is no decrease in the amount of nitrogen in the limb of the plant. The  $p$ -value is given by  $\text{pr}\{S \leq 3/H_0 \text{ is true}\} = 0.3633$ , which is too large to be an acceptable significance level  $\alpha$ .

The Cox Stuart test to detect trends may be used to detect correlation between two treatments or to test the presence of a predicted pattern (Conover, 1999) among several different applications.

The signs test has a normal approximation that works well for large sample size (say  $n > 20$ ). In this case, the value of the statistic (with a continuity correction applied) is

$$Z = \frac{S^* - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \quad (22)$$

where  $S^*$  refers to  $S - 1/2$  if there are more “plus” signs than “minus” signs or  $S + 1/2$  in the opposite case. This statistic follows a standard normal distribution,  $N(0, 1)$ , which simplifies the computation of the  $p$ -value (or to look for the corresponding critical value).

## 2.4 Rank-based methods

The procedures described in the previous section can be used with data in nominal scale. Even though the data were initially in ratio or interval scale, only dichotomous values are considered, for example, to be above or below a value (the signs test) or to belong to one of two classes (test of McNemar). This loss of information is usually translated in a loss of power. This paragraph presents statistical methods of among the most powerful for data in ordinal scale; they are the test of ranks.



Even in case the data came from a normal distribution (ratio scale), the loss of efficiency when using the test of ranks instead of the parametric counterpart is surprisingly small. The tests of ranks are valid for data with continuous, discrete or both continuous and discrete distributions.

### 2.4.1 One-sample case

The signs test is intuitive and extremely simple to perform but does not take the magnitude of the observation into account. An alternative that accounts for the magnitude of the observations is the Wilcoxon signed rank test. This test uses the concept of distribution  $X$  that is symmetric around a constant  $c$ , that is,  $\text{pr}\{X \leq c - x\} = \text{pr}\{X \geq c + x\}$  for each  $x$ . Supposing that a distribution is symmetric implies two things: (1) the mean and the median coincide and (2) the data must be at least in an interval scale. When the data are in ordinal scale, we only know whether a value is greater or less than the median. If the assumption of symmetry has sense, the distance to the median also does, and consequently, the distance between two observations makes sense. Hence, the data are not in an ordinal scale but in an interval scale.

*Wilcoxon signed rank test.* Let  $x_1, x_2, \dots, x_n$  be a sample of size  $n$  from the random variables  $X_1, \dots, X_n$ , which are independent, symmetric and with the same median. The following hypothesis is to be tested

$H_0$ : The median is  $\theta_0$ ,  $\theta = \theta_0$ .

$H_a$ : The median is greater than  $\theta_0$ ,  $\theta > \theta_0$  (one-sided alternative).

The statistic  $R$  (Siegel, 1956) and (Sprent, 1989) is computed according to the following steps:

1. Subtract  $\theta_0$  from the numerical sample values.
2. Ignore any observations that are equal to the hypothesized value (the difference in step 1 is zero) and reduce the sample size accordingly.
3. Assign ranks to the absolute value of the differences (i.e. rank all observations in increasing order of magnitude, ignoring their sign). If two observations have the same magnitude, regardless of sign, then they are given an average ranking.
4. Add the ranks corresponding to positive differences,  $R^+$ .
5. Add the ranks corresponding to negative differences,  $R^-$ .
6. Compute  $R = \min \{R^+, R^-\}$

The critical region at  $\alpha$  significance level is defined by

$$\text{CR} = \{R < R_{\alpha, \text{one-sided}}\} \quad (23)$$

There are no closed formulas for the distribution of this statistic, so the critical values written above as  $R_{\alpha, \text{one-sided}}$  are tabulated in Table A1. Other statistics have been used for the Wilcoxon signed rank test, for example, in

Conover (1999),  $R^+$  is used and also in Hettmansperger (1984), which includes a recurrent formula for the computation of  $\text{pr}\{R^+ = k\}$  for a given  $k$ . In general, this test is more efficient than the signs test, but it is very sensitive to the lack of symmetry.

### Example 7

With the data of Example 4, we want to decide if the median of the content of triazines is  $50 \mu\text{g L}^{-1}$  or higher. It is the same test as in Example 3, but now the rank of the observations listed in Example 4 is used. Table 6 shows the signed ranks that correspond to step 3 of the procedure so that

$$R^+ = 1 + 3 + 4 + 7 + 8 + 9 + 10 + 11 + 12 = 65$$

$$R^- = 2 + 5 + 6 = 13$$

$$R = \min \{13, 65\} = 13$$

For  $n = 12$ ,  $R_{0.05, \text{one-sided}} = 17$ , thus at 5% significance level, the null hypothesis is rejected.

If the alternative hypothesis was a two-sided alternative,  $H_a: \theta \neq \theta_0$ , the statistic would be the same but the critical region would be  $\text{CR} = \{R \leq R_{\alpha, \text{two-sided}}\}$ . In our example, if the hypothesis to be tested were that the median of the content of triazines is different from  $50 \mu\text{g L}^{-1}$  (what does not have much real sense), then  $R_{0.05, \text{two-sided}} = 13$  and the statistic still belongs to the critical region, thus the null hypothesis will be rejected.

The Wilcoxon signed rank test has a normal approximation that works well for large sample size (say  $n > 20$ ). In this case, the value of the corresponding statistic is defined as

$$Z = \frac{R + \frac{1}{2} - \frac{1}{4}n(n+1)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad (24)$$

which follows a standard normal distribution,  $N(0,1)$ , also known as a *Z distribution*.

**Table 6** Calculations to obtain the Wilcoxon signed rank test statistic; data of Example 7

	Values $d_i =  x - 50 $ in increasing order <sup>a</sup>					
$d_i$	2 (+)	9 (−)	13 (+)	14 (+)	18 (−)	22 (−)
Rank( $d_i$ )	1 (+)	2 (−)	3 (+)	4 (+)	5 (−)	6 (−)
$d_i$	29 (+)	33 (+)	38 (+)	42 (+)	48 (+)	53 (+)
Rank( $d_i$ )	7 (+)	8 (+)	9 (+)	10 (+)	11 (+)	12 (+)

<sup>a</sup> In parentheses, the sign of the difference  $x - 50$ .

Theoretically, when the sample comes from a continuous distribution, the probability of observing repeated values (ties) is zero; the same applies to the median: the probability that one of the observed values is equal to the median is zero. Actually, because of the round off error, equal values (or equal to the median) may be obtained. In this case, the statistic used is

$$Z_{\text{calc}} = \frac{\left| R + \frac{1}{2} - \frac{1}{4}n(n+1) \right|}{\sqrt{\frac{1}{4} \sum_{i=1}^n R_i^2}} \quad (25)$$

where  $R_i$  denotes the  $i$ th rank. Remember that the repeated differences  $|x_i - \theta|$  are ranked as the average of the corresponding ranks, all of them. For example, if instead of the values of [Example 4](#), we had recorded 28, 32, 41, 52, 59, 64, 72, 88, 88, 88, 98 and  $103 \mu\text{g L}^{-1}$ , then the differences  $|x_i - 50|$  (with sign) would have been: 2(+), 9(-), 9(+), 14(+), 18(-), 22(-), 22(+), 35(+), 35(+), 35(+), 48(+), and 53(+). Therefore, with the aforementioned rule, the ranks are 1(+), 2.5(-), 2.5(+), 4(+), 5(-), 6.5(-), 6.5(+), 9(-), 9(+), 9(+), 11(+) and 12(+), where the ranks 2 and 3 have been substituted by the mean 2.5, the same with 6 and 7 or 8, 9 and 10 that correspond to ties. Hence,  $R^+ = 1 + 2.5 + 4 + 6.5 + 9 + 9 + 9 + 11 + 12 = 64$  and  $R^- = 2.5 + 5 + 6.5 = 14$ , and consequently,  $R = 14$ . Using the approximation in Eq. (25)  $Z_{\text{calc}} = 1.9264$  and the critical region at 5% level is  $\text{CR} = \{Z_{\text{calc}} > Z_{0.05} = 1.645\}$ , the null hypothesis is rejected. Note that, with the exact test  $R = 14$  is less than 17 so  $H_0$  will be rejected.

A particularly important case of application of the Wilcoxon signed rank test is for analyzing paired samples. Now the data are  $n$  pairs of observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  on the respective bivariate random variables  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ .  $D_i$  are the  $n$  differences  $Y_i - X_i$ . These differences are the values of  $n$  one-sample problem. Admitting that the random variables  $D_i$  are independent, symmetrical and with the same median, it is desired to test the hypothesis that  $H_0$ : the median of  $D$  is zero as against the two-sided alternative  $H_a$ : the median of  $D$  is different from zero. Also the one-sided alternatives (the median is greater than zero, or the median is less than zero) are possible.

*Confidence intervals based on signed ranks.* The theory that justifies the procedure to calculate confidence intervals on the median (equal to the mean because of the symmetry of the distribution) is given by [Lehmann \(1975\)](#). The basic idea is to take every pair of observations and calculate the mean of each pair as an estimate of the random variable mean. Since each sample value is equally likely to be above or below the mean (unknown) of the random variable and equally likely to come from either of a pair of values symmetrically placed about that mean, these are sensitive estimates. It is expected that some of these values are underestimations and other overestimations, the reason why it is very intuitively reasonable to reject the greatest and the smallest values. In order to obtain the  $100 \times (1 - \alpha)\%$  confidence interval, it is enough to use the value  $R$  tabulated in Table A1, which defines the critical region of the test, thus its complement is the looked confidence interval. [Example 8](#) shows the procedure.

### Example 8

The exercise consists of obtaining a 95% confidence interval on the median of the data of the content of triazines in [Example 4](#). As  $n = 12$ , the critical value for the two-tail test at 5% is 13. Then, we need the  $n(n + 1)/2$  possible semi-sums of the all pairs of data. In our case, there are 78 semi-sums, from which we must discard the 13 smallest and the 13 greatest. The semi-sum in position 14 is 52 and the one in position 65 is 88, thus the 95% confidence interval on the content of triazines is (52, 88). The confidence interval obtained with the signs test for the same data was (41, 92) – [Example 4](#). As the procedure of the Wilcoxon signed rank test uses more information about the sample and further makes the assumption of symmetry, the interval is much shorter at the same confidence level.

When the sample size is greater than 20, [Eq. \(24\)](#) can be used to estimate the critical value of  $R$ . For instance, if  $n = 50$  and a 95% confidence interval is desired,  $Z_{0.025} = -1.96$  and  $R \approx 434$ , so that from the 1275 semi-sums, the interval is obtained by discarding the first and the last 434. There are some tricks ([Sprent, 1989](#); [Conover, 1999](#)) to reduce the computational effort because in fact it is not necessary to compute all the semi-sums.

For purely didactic effects, these intervals can be compared to the ones obtained with the corresponding parametric procedure based on normality, which is computed as

$$\bar{x} \pm t_{0.025, 11} \frac{s}{\sqrt{12}}$$

for the same 95% confidence level. With the same data, the interval is  $68.58 \pm 16.44$ , that is, (52.14, 85.02) as against (52, 88) of the signed rank test or (41, 92) of the signs test. Although the increase of information between the signs test and the signed rank test supposes a remarkable modification, this is not so when adding the hypothesis of normality of data, that a priori is not justified. This is an illustration of the efficiency of the procedures based on ranks.

*Estimation of the median.* An appropriate point estimate is the Hodges–Lehmann estimator, which is the median of the  $n(n + 1)/2$  paired estimates. This estimate possesses important robustness properties. With the same data of [Example 4](#), there are 78 semi-sums, thus the median is any number between the semi-sum in position 39 and the one in position 40, once ordered from smallest to greatest. These two semi-sums are 67.5 and 69.5, respectively, thus the median is 68.5 (the average as indicated in [Section 2.1](#)). Note that this value is not the centre of the confidence interval. Again, it can be compared to the mean value of the sample which is 68.58.

It is interesting to know that the Pitman's efficiency, ARE, of the Wilcoxon signed rank test related to the  $t$ -test when the data are normal is 0.955. If, instead, we assume that the variables have a uniform distribution, the ARE is 1.00. For a double exponential distribution, the ARE is 1.5. The most surprising thing is that the ARE is bounded below by 0.864, that is, the Wilcoxon test never can be too bad, but it can be very good as compared with the usual  $t$ -test when the

distribution is not normal. A detailed bibliography on several aspects of the Wilcoxon test, its multivariate generalization, its power in small samples or its use as a test for symmetry of a distribution can be consulted in Section 5.7 of [Conover \(1999\)](#).

2.4.2 Location test for two independent samples

When there are two independent samples and their locations (medians or means) are to be compared, the question cannot be reduced to a one-sample problem. It is possible to extend the signs test to a test on the median with null hypothesis  $H_0$ : the two random variables have the same median. The procedure consists of determining the median,  $\theta$ , of the series of data formed by joining both samples; later the number of values of each sample that are above or below  $\theta$  is obtained, which makes up a contingency table like [Table 7](#).

The statistic is

$$\chi^2_{\text{calc}} = \frac{(2a_1 - n_1)^2(n_1 + n_2)}{n_1n_2}$$

which should be compared to the critical value of a  $\chi^2$  distribution with one degree of freedom at the fixed significance level. This test is easy to apply but, in a similar manner to the signs test, is less efficient than a test that uses more information on data.

*The Wilcoxon–Mann–Whitney test.* It is based on similar ideas to the signed rank test. The data of the two samples are joined, the rank of each value in the joint series is determined and the rank of the data of each sample is analyzed. If they come from the same distribution, great and small ranks in both samples will be found. If, on the contrary, a distribution differs from the other in the location, there will be a tendency to find greater ranks in a sample than in the other. Formally, we have a random sample of size  $m$ ,  $x_1, x_2, \dots, x_m$ , coming from random variables  $X_1, X_2, \dots, X_m$ , which are independent, and another sample of size  $n$ ,  $y_1, y_2, \dots, y_n$ , coming from random variables  $Y_1, Y_2, \dots, Y_n$ , which are also independent among them and independent of the  $X$ s. It is assumed that if the distributions of  $X$ s and  $Y$ s are different, they only differ in the location (mean or median). In addition, the variables are at least in ordinal scale. The formalization of the two-tail test is

- $H_0$ : the two samples come from the same distribution.
- $H_a$ : the samples come from two distributions differing only in location.

**Table 7** Median test for two samples; contingency table

	Sample 1	Sample 2
Above $\theta$	$a_1$	$a_2$
Below $\theta$	$n_1 - a_1$	$n_2 - a_2$

One-sided alternatives are obtained if we specify in  $H_a$  the direction of any possible shift (positive or negative). The procedure for computing the statistic is as follows:

1. Rank all observations in increasing order of magnitude, ignoring which group they come from. If two observations have the same magnitude, regardless of group, then they are given an average ranking.
2. Sum the ranks of the samples  $X_i$ ,  $S_m$ .
3. Sum the ranks for samples  $Y_j$ ,  $S_n$ .
4. Compute

$$U_m = S_m - (1/2)m(m+1)$$

$$U_n = S_n - (1/2)n(n+1)$$

$$U_{\text{calc}} = \min(U_m, U_n)$$

The critical region at significance level  $\alpha$  is

$$\text{CR} = \{U_{\text{calc}} < U_{\alpha, m, n, \text{two-sided}}\} \quad (26)$$

Some critical values  $U_{\alpha, m, n, \text{two-sided}}$  are tabulated in Table A2. For the one-tail test, the statistic and the critical region are the same but using the corresponding critical value  $U_{\alpha, m, n, \text{one-sided}}$ .

### Example 9

In [Example 4](#), the content of triazines in wastewater in 12 places of a geographic zone, A, was written down, obtaining 28, 32, 41, 52, 63, 64, 79, 83, 88, 92, 98 and  $103 \mu\text{g L}^{-1}$ . Now, the content in 16 places of a second zone, B, with similar characteristics, are obtained with the following results: 12, 14, 15, 20, 21, 26, 27, 46, 48, 51, 70, 73, 74, 77, 78 and  $110 \mu\text{g L}^{-1}$ . Is there evidence that these samples come from populations differing in the median of the content of triazines? If we suppose that a possible shift in location is the only difference between both distributions, the test of Wilcoxon–Mann–Whitney can be applied.

[Table 8](#) contains the rank that corresponds to each data in the joint series and also its origin A or B.

If  $m = 16$  and  $n = 12$ ,  $S_m = 1 + 2 + 3 + 4 + \dots + 21 + 28 = 187$ ;  $U_m = 51$ ;  $S_n = 8 + 9 + 10 + \dots + 25 + 26 + 27 = 219$ ;  $U_n = 141$ , thus  $U_{\text{calc}} = 51$ .

By consulting Table A2,  $U_{0.05, 16, 12, \text{two-sided}} = 53$  and the value of the statistic belongs to the critical region defined in Eq. (26), thus the null hypothesis is rejected and the medians of the two distributions are different at 5% significance level.

If the additional hypothesis of the data coming from a normal distribution is assumed, the  $F$ -test on the equality of variances says that there is no evidence to reject the null hypothesis at 5% level and the  $t$ -test on the equality of means (unknown and equal variances) either does not reject the null hypothesis at 5%;

**Table 8** Computation of the statistic for the Wilcoxon–Mann–Whitney test; data of [Example 9](#)

Value	12	14	15	20	21	26	27	28	32	41
Rank	1	2	3	4	5	6	7	8	9	10
Site	B	B	B	B	B	B	B	A	A	A
Value	46	48	51	52	63	64	70	73	74	77
Rank	11	12	13	14	15	16	17	18	19	20
Site	B	B	B	A	A	A	B	B	B	B
Value	78	79	83	88	92	98	103	110		
Rank	21	22	23	24	25	26	27	28		
Site	B	A	A	A	A	A	A	B		

which means that both values (47.625 and 68.583) should be considered as significantly equal.

Hence, the two-tail Wilcoxon–Mann–Whitney test detects a difference that is undetected in the  $t$ -test. The key of the result of the  $t$ -test may be the observation  $110 \mu\text{g L}^{-1}$  of series B, which is very large with respect to the others of that series; the immediately previous one is  $78 \mu\text{g L}^{-1}$  (rank 21 in [Table 8](#)). This observation can be very influential and the  $t$ -test is not robust, the reason why the presence of outlier data makes it very insensitive. In [Section 2.4.5](#) the equality of variances by means of techniques based on ranks will be studied.

When the sample size of the two samples is greater than 20, the normal approximation is used, with statistic:

$$Z_{\text{calc}} = \frac{U_{\text{calc}} + \frac{1}{2} - \frac{1}{2}mn}{\sqrt{\frac{mn(m+n+1)}{12}}} \quad (27)$$

Then, the critical region at  $\alpha$  significance level for the null hypothesis  $H_0$ : “no location difference” as against the two-sided alternative is  $\text{CR} = \{Z_{\text{calc}} < -Z_{\alpha/2} \text{ or } Z_{\text{calc}} > Z_{\alpha/2}\}$ , where  $Z_{\alpha}$  is the critical value of a standard normal distribution, that is, the value such that  $\alpha = \text{pr}\{N(0,1) > Z_{\alpha}\}$ . The critical regions for the one-sided alternatives are  $\text{CR} = \{Z_{\text{calc}} < -Z_{\alpha}\}$  or  $\text{CR} = \{Z_{\text{calc}} > Z_{\alpha}\}$ .

If there are only a few ties, the mid-rank method and the standard test or the normal approximation in [Eq. \(27\)](#) are adequate. If there are many ties, [Eq. \(27\)](#) should be modified (see [Sprent, 1989](#) and [Conover, 1999](#)).

*Wilcoxon–Mann–Whitney confidence intervals and parameter estimation.* Similar to the Wilcoxon signed rank test, all the possible differences  $x_i - y_j$  for  $x_i, i = 1, 2, \dots, 12$ , and  $y_j, j = 1, 2, \dots, 16$ , should be computed and sorted in ascending order. As the critical value at 5% significance level is  $U_{0.05,12,16,\text{two-sided}} = 53$ , the 95% confidence interval will be obtained by discarding the first and last 53 differences. The total number of differences is  $m \times n = 16 \times 12 = 192$ . Once sorted, the difference in position 54 is 1 and difference in position 139 is 47, so the 95% confidence interval on the difference of medians  $\theta_X - \theta_Y$  is  $(1, 47) \mu\text{g L}^{-1}$ . The point estimator,

called the Hodges–Lehmann estimator, is the median of the paired differences; in our problem it is  $20 \mu\text{g L}^{-1}$ .

The efficiency of the Wilcoxon–Mann–Whitney test is the same as that of the Wilcoxon signed rank test when it is compared with the  $t$ -test computed under the assumption that the distributions of  $X$  and  $Y$  are identical except for their means. If the populations are normal, the ARE is 0.955; if the populations are uniform, the ARE is 1.0; and if they are a double exponential, the ARE is 1.5.

### 2.4.3 Location test for several independent samples

A test proposed by Kruskal and Wallis is a direct extension of the Wilcoxon–Mann–Whitney test. Suppose  $t$  independent samples of sizes  $n_1, n_2, \dots, n_t$  ( $n = \sum_{i=1}^t n_i$ ). The hypothesis to be tested is:

$H_0$ : All the samples come from the same random variable.

$H_a$ : At least one sample comes from a random variable with a different location.

Frequently, a shift in location is called *treatment effect* as in experimental design, and each of the  $t$  samples is called *treatment level* (factor level). In fact, it can be said that this test is the analogous to the one-factor ANOVA.

The procedure consists of the following steps:

1. Compute the mean of the sum of squares (corrected) of all the ranks

$$S_r^2 = \frac{1}{n-1} \left( \sum_{ij} r_{ij}^2 - \frac{n(n+1)^2}{4} \right) \quad (28)$$

2. Compute the mean sum of squares (corrected) in each level

$$S_t^2 = \left( \sum_{i=1}^t \frac{s_i^2}{n_i} \right) - \frac{1}{4} n(n+1)^2, \quad \text{with } s_i = \sum_{j=1}^{n_i} r_{ij} \quad (29)$$

3. The statistic  $T_{\text{calc}}$  is

$$T_{\text{calc}} = \frac{S_t^2}{S_r^2} \quad (30)$$

The critical region of the test at  $100 \alpha \%$  significance level is

$$\text{CR} = \left\{ T_{\text{calc}} > \chi_{1-\alpha, \gamma}^2 \right\} \quad (31)$$

where  $\chi_{1-\alpha, \gamma}^2$  is the quantile of a  $\chi^2$  distribution with  $\gamma = t-1$  degrees of freedom, i.e. the value such that  $1 - \alpha = \text{pr}\left\{ \chi_{\gamma}^2 < \chi_{1-\alpha, \gamma}^2 \right\}$ .



When there are no ties,  $T_{\text{calc}}$  can be simplified to obtain

$$T_{\text{calc}} = \frac{12}{n(n+1)} \left( \sum_{i=1}^t \frac{s_i^2}{n_i} \right) - 3(n+1) \tag{32}$$

For very small values of  $n$  it is necessary to use tables for the critical values of the statistic  $T$  (Conover, 1999), but for moderate values of  $n$  the approximation of the  $\chi^2$  in Eq. (31) is sufficient.

**Example 10**

Table 9 contains six data with the amount of triazines measured in a new geographical zone C, further to those obtained in zones A and B, Example 9. Also, the rank of each value in the joint series is in Table 9.

Now  $t = 3$ ,  $n_1 = 12$ ,  $n_2 = 16$  and  $n_3 = 6$ . As the value 63 appears in both zones A and C, their mid-rank (20.5) is assigned. From Eq. (30)  $T_{\text{calc}} = 7.48$ .  $\chi^2_{0.05,2} = 5.99$ , so the value of the statistic is in the critical region, and the null hypothesis must be rejected at 5% significance level.

**2.4.4 Location test for several related samples**

In the cases in which we have raised the study of two samples, that is, to compare the effect of two treatments, the underlying idea is that under the null hypothesis, the samples are more homogenous than under the alternative one when the distributions under each one of the treatments are different. In the case of paired samples, the underlying idea is the same. Also, this is the basic assumption in the parametric treatment of this question with normal distributions: the ANOVA. However, there are many experimental or observational situations in which this working hypothesis cannot be maintained and is necessary to admit that important differences will appear within the treatment though equal or almost equal for all of them. This problem is an extension of the problem of matched pairs, or

**Table 9** Computation to obtain the statistic of the test of Kruskal–Wallis; data of Example 10

<i>Site A</i>								
Values	28	32	41	52	63	64	79	83
Rank	11	13	14	18	20.5	22	28	29
Values	88	92	98	103				
Rank	30	31	32	33				
<i>Site B</i>								
Values	12	14	15	20	21	26	27	46
Rank	1	3	4	6	7	9	10	15
Values	48	51	70	73	74	77	78	110
Rank	16	17	23	24	25	26	27	34
<i>Site C</i>								
Values	13	18	22	29	56	63		
Rank	2	5	8	12	19	20.5		

related samples, examined in the [Section 2.4.1](#) as application of the Wilcoxon signed rank test. First, we will present the Friedman test, which is an extension of the signs test of [Section 2.3.1](#). Then we will present the Quade test, which is an extension of the Wilcoxon signed rank test of [Section 2.4.1](#). The Friedman test is the better-known test of the two and requires fewer assumptions, but it suffers from lack of power when there are only three treatments, just as the signs test has less power than the Wilcoxon signed rank test when there are only two treatments. When there are four or five treatments, the Friedman test has about the same power as the Quade test, but when the number of treatments is six or more, the Friedman test tends to have more power. See Iman et al. (1984) and [Hora and Iman \(1988\)](#) for power and ARE comparisons.

*The Friedman test* ([Friedman \(1937\)](#)). In this case, the data consist of  $b$  mutually independent  $t$  values, called  $b$  blocks or treatments. The random value  $x_{ij}$  is in block  $i$  ( $i = 1, \dots, b$ ) and is associated with treatment  $j$  ( $j = 1, 2, \dots, t$ ). The  $b$  blocks are arranged as in [Table 10](#).

Let  $r_{ij}$  denote the rank (from 1 to  $t$ ) assigned to  $x_{ij}$  within block (row)  $i$ . This means that in the  $i$ th block, the values  $x_{i1}, x_{i2}, \dots, x_{it}$  are sorted in increasing order, and the corresponding rank is assigned to each one. Again, in the case where ties exist, the mid-rank is assigned. The hypothesis test is

$H_0$ : There is no effect of treatment (each ranking of values within a block is equally likely).

$H_a$ : At least one of the treatments tends to yield larger values than at least one other treatment.

As in the case of the test of Kruskal–Wallis, the statistic is

$$T_{1,\text{calc}} = \frac{S_t^2}{S_r^2} \quad (33)$$

but now

$$S_r^2 = (1/(b(t-1))) \left( \sum_{ij} r_{ij}^2 - (1/4)bt(t+1)^2 \right) \text{ and } S_t^2 = (1/b) \sum_{j=1}^t s_j^2 - \frac{1}{4}bt(t+1)^2$$

with  $s_j = \sum_{i=1}^b r_{ij}$ . With no ties,  $S_r^2 = (1/6)bt(t+1)(2t+1)$ .

**Table 10** Arrangement of data to apply the Friedman test

Block	Treatment				
	1	2	3	...	$t$
1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1t}$
2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2t}$
3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3t}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$b$	$x_{b1}$	$x_{b2}$	$x_{b3}$	...	$x_{bt}$

For  $b$  and  $t$  not too small,  $T_{1,\text{calc}}$  has approximately a  $\chi^2$  distribution with  $t-1$  degrees of freedom. Therefore, the critical region is

$$\text{CR} = \left\{ T_{1,\text{calc}} > \chi^2_{\alpha,\gamma} \right\} \tag{34}$$

Again,  $\chi^2_{\alpha,\gamma}$  is the  $\alpha$ -quantile of a  $\chi^2$  distribution with  $\gamma = t - 1$  degrees of freedom. Iman and Davenport (1980) suggested that a better approximation is given by the following equation:

$$T_{\text{calc}} = (b-1) \frac{\frac{1}{b} \sum_{j=1}^t s_j^2 - \frac{1}{4} b t (t+1)^2}{\sum_{i,j} r_{ij}^2 - \frac{1}{b} \sum_{j=1}^t s_j^2} \tag{35}$$

which, under null hypothesis of no treatment difference, has approximately an  $F$  distribution with  $t-1$  and  $(b-1)(t-1)$  degrees of freedom. The critical region at  $100\alpha\%$  significance level is then

$$\text{CR} = \left\{ T_{\text{calc}} > F_{1-\alpha,\gamma_1,\gamma_2} \right\} \tag{36}$$

with  $F_{1-\alpha,\gamma_1,\gamma_2}$  being the  $(1-\alpha)$  quantile of an  $F$  distribution with  $\gamma_1 = t-1$  and  $\gamma_2 = (b-1)(t-1)$  degrees of freedom.

**Example 11**

Six different regression procedures have been applied to data from excitation–emission fluorescence to calibrate naphthalene in the range  $1\text{--}19\text{ }\mu\text{g L}^{-1}$ . With each one, the recovery has been evaluated with new (not used for calibration) samples of 7, 10, 13 and  $16\text{ }\mu\text{g L}^{-1}$  of naphthalene. Table 11 shows the recoveries obtained (mean of five samples) in percentage, the ranks  $r_{ij}$  and values  $s_j$ . For this

**Table 11** Recovery (%) obtained with six different calibration models

Concentration of naphthalene ( $\mu\text{g L}^{-1}$ )	Calibration method					
	Univariate <sup>a</sup>	2-PLS <sub>exc</sub> <sup>b</sup>	2-PLS <sub>em</sub> <sup>c</sup>	PARAFAC <sup>d</sup>	MCR-ALS <sup>d</sup>	3-PLS <sup>d</sup>
7	95 (1)	98 (4)	96 (2)	100 (6)	99 (5)	97 (3)
10	97 (2.5)	96 (1)	97 (2.5)	100 (4.5)	102 (6)	100 (4.5)
13	96 (1.5)	96 (1.5)	97 (3.5)	103 (6)	98 (5)	97 (3.5)
16	95 (1.5)	95 (1.5)	100 (5.5)	100 (5.5)	96 (3)	99 (4)
$s_j = \sum_{i=1}^b r_{ij}$	6.5	8	13.5	22	19	15

Mean of five samples for each concentration of naphthalene. In parentheses, the rank of each value to apply the Friedman test.

<sup>a</sup> Fluorescence intensity recorded at  $\lambda_{\text{exc}} = 219\text{ nm}$  and  $\lambda_{\text{em}} = 330\text{ nm}$ .

<sup>b</sup> Vector of fluorescence intensity recorded at  $\lambda_{\text{exc}}$  between 210 and 250 nm with  $\lambda_{\text{em}}$  fixed at 330 nm.

<sup>c</sup> Vector of fluorescence intensity recorded at  $\lambda_{\text{exc}}$  fixed at 219 nm and  $\lambda_{\text{em}}$  between 290 and 400 nm.

<sup>d</sup> EEM matrix of fluorescence intensity at  $\lambda_{\text{exc}}$  between 210 and 250 nm and  $\lambda_{\text{em}}$  between 290 and 400 nm.

problem,  $t = 6$  and  $b = 4$ . To decide if there are differences among calibrations, we will use the variant of the Friedman test of Eq. (35).

As

$$\begin{aligned}\sum_{ij} r_{ij}^2 &= 361, \quad (1/b) \sum_{j=1}^t s_j^2 = (1/4)(6.5^2 + 8^2 + \dots + 15^2) \\ &= 339.625 \text{ and } (1/4)bt(t+1)^2 = 294\end{aligned}$$

then,

$$T_{\text{calc}} = 3 \frac{339.62 - 294}{361 - 339.62} = 6.40$$

At 5% significance level, the 0.95 quantile of the distribution  $F_{5,15}$  is 2.90, the value of the statistic is in the critical region and therefore, the null hypothesis should be rejected. This means that the different calibration models yield significantly different recoveries; in fact, the  $p$ -value is 0.002, thus they are much significantly different.

*Multiple comparisons with the Friedman test.* The following method for comparing individual treatments may be used only if the Friedman test results in the rejection of the null hypothesis. Treatments  $i$  and  $j$  are considered different at  $100 \times \alpha\%$  significance level if the following inequality holds:

$$|s_i - s_j| > t_{1-(\alpha/2)} \sqrt{\frac{2b}{(b-1)(t-1)} \left( \sum_{ij} r_{ij}^2 - \frac{1}{b} \sum_{j=1}^t s_j^2 \right)} \quad (37)$$

where  $t_{1-(\alpha/2)}$  is the quantile of the  $t$  distribution with  $(b-1)(t-1)$  degrees of freedom. As  $t_{0.975,15} = 2.1314$  and the other factor in Eq. (37) is 3.3764, all the differences that, in absolute value, are greater than 7.1964 will be significantly different at 5%. All the comparisons are in Table 12, where we can see that in six cases the differences are significant.

Note that a summary of the multiple comparisons procedure may be presented by listing the treatments (calibrations in our example) and underlining the groups of treatments that are not significantly different from a single underline as follows (which shows an interesting pattern).

Univariate	2-PLS <sub>exc</sub>	2-PLS <sub>em</sub>	3-PLS	MCR-ALS	PARAFAC
<hr/>					
	<hr/>				
		<hr/>			
			<hr/>		
				<hr/>	
					<hr/>

**Table 12** Results of the test of multiple comparisons among the calibration models in [Example 11](#).

Calibrations	$ s_i - s_j $	Calibrations	$ s_i - s_j $	Calibrations	$ s_i - s_j $	Calibrations	$ s_i - s_j $	Calibrations	$ s_i - s_j $
1–2	1.5	2–3	5.5	3–4	1.5	4–5	4.0	5–6	3.0
1–3	7.0	2–4	7.0	3–5	5.5	4–6	7.0		
1–4	8.5 <sup>a</sup>	2–5	11.0 <sup>a</sup>	3–6	8.5 <sup>a</sup>				
1–5	12.5 <sup>a</sup>	2–6	14.0 <sup>a</sup>						
1–6	15.5 <sup>a</sup>								

Codification of the calibration models after ranking by  $s_i$  values of table 11: 1, Univariate; 2, 2-PLS<sub>exc</sub>; 3, 2-PLS<sub>em</sub>; 4, 3-PLS; 5, MCR-ALS and 6, PARAFAC.

<sup>a</sup>The recoveries obtained with these calibration models are significantly different at 5%.

The Friedman test may be used with data given only as ranks within each block. In this latter circumstance, it becomes a test for consistency of ranking rather than the one for a location parameter; the test in this context was first developed by M.G. Kendall in the first edition (1962) of book [Kendall \(1975\)](#). This version of the Friedman test is useful because there are many problems (sensory valuation, degree of difficulty or degree of contamination), which are naturally measured in ordinal scale.

*The Quade test.* The procedure starts by finding the ranks within treatments (blocks) as described in the previous test. The next step again uses the original observation  $x_{ij}$ . Ranks are assigned to the blocks themselves according to the size of the sample range in each block. Note that the range in  $i$ th block is the maximum of  $x_{ij}$ ,  $j = 1, \dots, t$ , minus the minimum of the same values  $x_{ij}$ ,  $j = 1, \dots, t$ . Assign ranks to blocks according to the sample ranges and use average ranks in case of ties. Let  $q_1, q_2, \dots, q_b$  be the ranks assigned to blocks 1, 2, ...,  $b$  respectively, then

$$s_{ij} = q_i \left( r_{ij} - \frac{t+1}{2} \right) \quad (38)$$

is a statistic that represents the relative size of each value within the block, adjusted to reflect the relative significance of the block in which it appears. Note that  $(t+1)/2$  is the mean of the ranks within block. The statistic is that of the following Eq. (39) which is the same as in [Eq. \(35\)](#) without the correction for the mean that appears in the numerator, because the values have already been weighted by the range of the block. Therefore

$$T_{3, \text{calc}} = (b-1) \frac{\frac{1}{b} \sum_{j=1}^t s_j^2}{\sum_{i,j} s_{ij}^2 - \frac{1}{b} \sum_{j=1}^t s_j^2} \quad (39)$$

The exact distribution of  $T_3$  is difficult to find, so the  $F$  distribution with  $t-1$  and  $(b-1)(t-1)$  degrees of freedom as in the Friedman test is used as an approximation.

### Example 12

With the data of [Example 11](#), we will apply the Quade test. [Table 13](#) collects the ranges per block, the ranks  $q_i$  assigned to each block, and the  $s_{ij}$  values.

Now

$$(1/b) \sum_{j=1}^t s_j^2 = (1/4) \left( (-17.75)^2 + (-17.75)^2 + \dots + (3)^2 \right) = 314.81 \text{ and } \sum_{ij} s_{ij}^2 = 489.00$$

thus, the statistic is

$$T_{3, \text{calc}} = 3 \frac{314.81}{489.00 - 314.81} = 5.42$$

**Table 13** Values of  $s_{ij}$  for the Quade test; data of Table 11

Concentration of naphthalene ( $\mu\text{g L}^{-1}$ )	Block range	Rank ( $q$ )	Calibration method					
			Univariate	2-PLS <sub>exc</sub>	2-PLS <sub>em</sub>	PARAFAC	MCR-ALS	3-PLS
7	5	1.5	−3.75	0.75	−2.25	3.75	2.25	−0.75
10	6	3	−3.00	−7.50	−3.00	3.00	7.50	3.00
13	7	4	−8.00	−8.00	0.00	10.00	6.00	0.00
16	5	1.5	−3.00	−3.00	3.00	3.00	−0.75	0.75
$s_j = \sum_{i=1}^b s_{ij}$ (Eq. (38))			−17.75	−17.75	−2.25	19.75	15.00	3.00

At 5% significance level, the 0.95 quantile of the distribution  $F_{5,15}$  is 2.90, then, again, the value of the statistic is in the critical region and we must reject the null hypothesis. The  $p$ -value now is 0.005 greater than the one of the Friedman test.

Also, it is possible to make multiple comparisons with the same inequality as in Eq. (37) but applied to the values  $s_{ij}$  defined in Eq. (38).

### 2.4.5 Test for equal variances

Although the tests to detect location shifts are the most common, sometimes it is desired to test the homogeneity of the variances or better the homogeneity of the dispersion of the data.

*Case 1. Two samples.* The data consist of two independent random samples. Let  $x_1, x_2, \dots, x_n$  denote the  $n$  values of a random variable  $X$  and let  $y_1, y_2, \dots, y_m$  denote the  $m$  values from a second random variable  $Y$ . Both samples are supposed to be independent and mutually independent. The measurement scale is at least interval. We wish testing the following hypothesis:

$H_0$ :  $X$  and  $Y$  are identically distributed, except for possibly different location.

$H_a$ :  $V(X) \neq V(Y)$ .

The Siegel–Tuckey procedure parallels the Wilcoxon–Mann–Whitney location test and uses the same tabulated values for significance. First, we align the two samples by subtracting an estimation of location difference to the sample with the higher location (or to add this estimate to the values in the other sample). With this new data set, the Wilcoxon–Mann–Whitney procedure, explained in Section 2.4.2, is applied.

### Example 13

With the data of triazines in site A and site B studied in Example 9, we established that the Hodges–Lehmann estimate of location difference was  $20 \mu\text{g L}^{-1}$ . After adding 20 to the values of site B we obtain the values in Table 14 which have been already joined and sorted in increasing order. Now, summing up the corresponding ranks, is  $m = 12$ ,  $S_m = 174.5$ ,  $n = 16$ ,  $S_n = 231.5$ ,  $U_m = 96.5$  and  $U_n = 95.5$ . Therefore,  $U_{\text{calc}} = 95.5$ . Eq. (26) and Table A2 indicate that  $U_{\text{calc}}$  must

**Table 14** Ranking to test the equality of variances by the Siegel–Tuckey test; data of [Example 13](#)

Value	28	32	32	34	35	40	41	41	46	47
Rank	1	2.5	2.5	4	5	6	7.5	7.5	9	10
Site	A	A	B	B	B	B	A	B	B	B
Value	52	63	64	66	68	71	79	83	88	90
Rank	11	12	13	14	15	16	17	18	19	20
Site	A	A	A	B	B	B	A	A	A	B
Value	92	93	94	97	98	98	103	130		
Rank	21	22	23	24	25.5	25.5	27	28		
Site	A	B	B	B	A	B	A	B		

not exceed 53 for significance. The conclusion is that we do not reject the null hypothesis that the variables have the same variance at 5% significance level.

The Siegel–Tuckey test is simple to carry out but not very powerful. For that reason, the following is the squared rank test ([Conover 1999](#), pp. 300–302), which is also called Conover test.

The data and assumptions are the same as in the previous case. The equality of variances hypothesis implies  $E((X - \mu_x)^2) = E((Y - \mu_y)^2)$ . Conover proposes a test for equality of variances based on joint ranks of  $(x_i - \mu_x)^2$ ,  $i = 1, 2, \dots, n$ , and  $(y_j - \mu_y)^2$ ,  $j = 1, 2, \dots, m$ . In practice,  $\mu_x$  and  $\mu_y$  are unknown, so it is reasonable to replace them by sample estimates  $\bar{x}$  and  $\bar{y}$ . Further, for computing the ranks it is not necessary to square the deviations to obtain the required ranking, because the same order is achieved by ranking the absolute deviations.

If we denote the squares of the ranks of these absolute deviations by  $u_i(x) = |x_i - \bar{x}|$ ,  $i = 1, \dots, n$ , and  $u_j(y) = |y_j - \bar{y}|$ ,  $j = 1, \dots, m$ , then the test statistic is as follows:

$$Z_{\text{calc}} = \frac{T - n\bar{u}}{s} \quad (40)$$

where

$$T = \sum_{i=1}^n (u_i(x))^2 \quad (41)$$

$$\bar{u} = \frac{1}{n+m} \left( \sum_{i=1}^n (u_i(x))^2 + \sum_{j=1}^m (u_j(y))^2 \right) \quad (42)$$

$$s = \sqrt{\frac{mn \left( \sum_{i=1}^n (u_i(x))^2 + \sum_{j=1}^m (u_j(y))^2 - (m+n)\bar{u}^2 \right)}{(m+n)(m+n-1)}} \quad (43)$$

If there are no ties, the test statistic  $T$  in [Eq. \(41\)](#) will be used. Quantiles of the exact distribution of  $T$  and a large sample approximation are given in Table A9 of book [Conover \(1999\)](#) but for reasonably large sample sizes,  $Z_{\text{calc}}$  in [Eq. \(40\)](#), is



approximately a standard normal distribution and the critical region at 100  $\alpha\%$  significance level is

$$\text{CR} = \{Z_{\text{calc}} < -Z_{1-(\alpha/2)} \text{ or } Z_{\text{calc}} > Z_{1-(\alpha/2)}\} \quad (44)$$

where  $Z_{1-(\alpha/2)}$  is the  $1-(\alpha/2)$  quantile of the standard normal distribution.

The one-tail tests are obtained by modifying the critical region according to the alternative hypothesis  $H_a$  of the test.

*Case 2. More than two samples.* There are three or more samples,  $t$ , with the same assumptions as in the case of two samples and we wish testing:

$H_0$ : All  $t$  variables are identical, except for possibly different means.

$H_a$ : Some of the population variances are not equal to each other.

The squared rank test given in the previous section extends easily to several independent samples. Suppose we have  $t$  independent samples with sizes  $n_1, n_2, \dots, n_t$  ( $n = \sum_{i=1}^t n_i$ ).

As in the two-sample case, we first replace the observations by the absolute deviations from the sample mean for each sample,  $u_{ij} = |x_{ij} - \bar{x}_j|$ ,  $i = 1, \dots, t$  and  $j = 1, 2, \dots, n_i$ . These  $n$  deviations are ranked across all samples,  $r_{ij}$ ,  $i = 1, \dots, t$  and  $j = 1, 2, \dots, n_i$ , and the squares of the ranks are obtained. The statistical test is based on these squared ranks. The test statistic is identical in form to [Eq. \(30\)](#) except that now  $s_i = \sum_{j=1}^{n_i} r_{ij}^2$ . Therefore,

$$S_r^2 = \frac{1}{n-1} \left( \sum_{ij} r_{ij}^4 - \frac{(\sum_{ij} r_{ij}^2)^2}{n} \right) \quad (45)$$

If there are no ties,  $S_r^2 = (1/180)n(n+1)(2n+1)(8n+11)$  and  $\sum_{ij} r_{ij}^2 = (1/6)n(n+1)(2n+1)$ .

$$S_t^2 = \left( \sum_{i=1}^t \frac{s_i^2}{n_i} \right) - \frac{\left( \sum_{ij} r_{ij}^2 \right)^2}{n} \quad (46)$$

$$T_{\text{calc}} = \frac{S_t^2}{S_r^2} \quad (47)$$

The critical region of the test at 100  $\alpha\%$  significance level is

$$\text{CR} = \{T_{\text{calc}} > \chi_{1-\alpha, \gamma}^2\} \quad (48)$$

As usual,  $\chi_{1-\alpha, \gamma}^2$  denotes the quantile of a  $\chi^2$  distribution with  $\gamma = t-1$  degrees of freedom such that  $1 - \alpha = \text{pr}\{\chi_\gamma^2 < \chi_{1-\alpha, \gamma}^2\}$ .

**Example 14**

With the data of the content of triazines in three sites, A, B and C (Example 10, Table 9), we wish to test the equality of variances. The means of the data of sites A, B, and C are 68.58, 47.62 and 33.50 respectively. Table 15 shows the values  $u_{ij} = |x_{ij} - \bar{x}_j|$ ,  $i = 1, \dots, n_j$ ,  $j = 1, 2, 3$ , their ranks,  $r_{ij}$ , and also the squares of the ranks.

The sums of squared ranks for each site are 4902, 7469 and 1314, respectively, therefore,  $S_t^2 = (1/12)4902^2 + (1/16)7469^2 + (1/6)1314^2 - 5508213 = 268643$  and  $S_r^2 = 129095.2$  and thus,  $T_{\text{calc}} = 2.08$ . As  $\chi_{0.05,2}^2 = 5.99$ , there is no evidence to reject the null hypothesis on the equality of variances.

**2.4.6 Rank correlation**

A measure of correlation is a random variable that is used when the data are  $n$  pairs of numbers  $(x_i, y_i), i = 1, 2, \dots, n$ , obtained from a bivariate distribution  $(X, Y)$ . In general, the correlation measures take values in the interval  $[-1, 1]$ . If large values of  $X$  correspond to large values of  $Y$  and therefore small values of  $X$  correspond to small values of  $Y$ , the correlation is positive and near 1 if the tendency is strong. If, on the contrary, large values of  $X$  correspond to small values of  $Y$  and vice versa, then the correlation measure will be negative, and as nearer to  $-1$  as stronger the correlation.

**Table 15** Absolute deviations,  $|x_{ij} - \bar{x}_j|$ , ranks and squared ranks for the amount of triazines

<i>Site A</i>								
$ x_{i1} - 68.58 $	40.58	36.58	27.58	16.58	5.58	4.58	10.42	14.42
Rank	33	32	22	11	6	5	7	9
Squared rank	1089	1024	484	121	36	25	49	81
$ x_{i1} - 68.58 $	19.42	23.42	29.42	34.42				
Rank	12	18	25	30				
Squared rank	144	324	625	900				
<i>Site B</i>								
$ x_{i2} - 47.62 $	35.63	33.63	32.63	27.63	26.63	21.63	20.63	1.63
Rank	31	29	28	23	21	15	14	2
Squared rank	961	841	784	529	441	225	196	4
$ x_{i2} - 47.62 $	0.38	3.38	22.38	25.38	26.38	29.38	30.38	62.38
Rank	1	3	16	19	20	24	27	34
Squared rank	1	9	256	361	400	576	729	1156
<i>Site C</i>								
$ x_{i3} - 33.50 $	20.50	15.50	11.50	4.50	22.50	29.50		
Rank	13	10	8	4	17	26		
Squared rank	169	100	64	16	289	676		

The most used measure of correlation is the Pearson correlation coefficient defined as

$$r = \frac{\text{cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} \quad (49)$$

where the covariance  $\text{cov}(X, Y)$  is defined in Eq. (13) and the variances in Eqs (3) and (6). The Pearson correlation coefficient is a measure of the strength of the linear relation between  $X$  and  $Y$ . It can be used with any numerical data type without any requirement on the scale of measurement or the type of underlying distribution, but it is of difficult interpretation if the scale is not at least of interval. Unfortunately, the distribution function of  $r$  depends on the joint distribution function of  $(X, Y)$ . Many correlation measures exist; we will centre ourselves to some of them based on ranks, which have distribution functions that do not depend on the bivariate distribution  $(X, Y)$ , provided both are continuous. These correlation measures can be used with non-numerical data, for example, in ordinal scale.

*Spearman rank correlation* (Spearman's  $\rho$ , Spearman (1904)). The procedure consists of substituting the values  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , by their ranks  $(r(x_i), r(y_i))$  or the mid-rank if ties exist. Afterwards, the Pearson correlation coefficient, Eq. (49), is computed with these ranks. Spearman's  $\rho$  is insensitive to some types of dependence; hence it is better to specify what types of dependence can be detected. The two-tail test is as follows:

$H_0$ : The variables  $X$  and  $Y$  are mutually independent.

$H_a$ : Either there is a tendency for the larger values of  $X$  to be paired with the larger values of  $Y$ , or there is a tendency for the smaller values of  $X$  to be paired with the larger values of  $Y$ .

The test statistic is the coefficient  $\rho$  of Spearman, and the critical values are tabulated in Table A9 of Sprent (1989) or Table A10 of Conover (1999). However, a good approximation of the  $p$ th quantile of  $\rho$  is given by

$$w_p = \frac{Z_p}{\sqrt{n-1}} \quad (50)$$

where  $Z_p$  is the standard normal  $p$ -quantile. At level  $100 \alpha \%$ , the critical region is

$$\text{CR} = \{|\rho| > w_{1-(\alpha/2)}\} \quad (51)$$

The approximate two-sided  $p$ -value is

$$p\text{-value} = 2\text{pr}\{Z \geq |\rho|\sqrt{n-1}\} \quad (52)$$

**Example 15**

Every Tuesday of 15 consecutive weeks, the values of pH, conductivity and concentration of chloride in water of human consumption were determined. [Table 16](#) contains the results and the ranks assigned to each variable.

The value of the Spearman's  $\rho$  for the three pairs of variables together with the corresponding  $p$ -values are given in [Table 17](#). For  $(X, Y)$  and  $(X, Z)$  the null hypothesis cannot be rejected (5% significance level) and thus the variables should be considered as independent. The contrary happens with pair  $(Y, Z)$ ;  $H_0$  must be rejected and hence  $H_a$  is accepted.

For comparative purposes, the Pearson correlation coefficients have also been written down with the corresponding  $p$ -values in the same table. The results are completely analogous. In any case, this example illustrates the need for evaluating the  $p$ -values to avoid give some meaning to the sign of correlations that are not significant.

**Table 16** Values of pH, conductivity and concentration of chlorides; data of [Example 15](#)

pH (X)	Conductivity (Y)	Conc. of chlorides (Z)	R(X)	R(Y)	R(Z)
7.33	54.5	3.65	2	15	14
7.40	52.7	3.14	3	11.5	7.5
7.50	52.2	2.70	8	5.5	2
7.53	51.7	2.84	11	1	3
7.49	51.9	2.57	6	2	1
7.50	52.1	3.30	8	3.5	11
7.47	52.3	3.14	5	7.5	7.5
7.50	52.7	3.65	8	11.5	14
7.31	52.1	2.98	1	3.5	4
7.46	52.4	3.65	4	9	14
7.51	52.2	3.47	10	5.5	12
7.60	52.3	3.14	13	7.5	7.5
7.65	52.5	3.14	14.5	10	7.5
7.58	53.0	3.14	12	13	7.5
7.65	53.4	3.14	14.5	14	7.5

**Table 17** Spearman's  $\rho$ , Pearson's  $r$  and the corresponding  $p$ -values for the three pairs of variables in [Example 15](#)

	Spearman's $\rho$	$p$ -value	Pearson's $r$	$p$ -value
(X, Y)	0.090	0.750	-0.1437	0.609
(X, Z)	-0.119	0.673	-0.1642	0.559
(Y, Z)	0.549	0.034	0.5206	0.047

The one-tail test for negative correlation is

$H_0$ : The variables  $X$  and  $Y$  are mutually independent.

$H_a$ : There is a tendency for the smaller values of  $X$  to be paired with the larger values of  $Y$  and vice versa

The critical region is  $CR = \{\rho < -w_{1-\alpha}\}$ , where  $w_{1-\alpha}$  is determined from Eq. (50) and the  $p$ -value is  $\text{pr}\{Z \leq \rho\sqrt{n-1}\}$ .

The one-tail test for positive correlation is

$H_0$ : The variables  $X$  and  $Y$  are mutually independent.

$H_a$ : There is a tendency for the larger values of  $X$  and  $Y$  to be paired together.

The critical region is  $CR = \{\rho > w_{1-\alpha}\}$ , where  $w_{1-\alpha}$  is determined from Eq. (50) and the  $p$ -value is  $\text{pr}\{Z \geq \rho\sqrt{n-1}\}$ .

*Kendall's  $\tau$* . The data are  $n$  pairs of values  $(x_i, y_i)$ ,  $i = 1, \dots, n$ . Two observations  $(x_k, y_k)$  and  $(x_l, y_l)$  are said concordant if  $(y_k - y_l)/(x_k - x_l) > 0$  and discordant if  $(y_k - y_l)/(x_k - x_l) < 0$ . Let  $N_c$  and  $N_d$  denote the number of concordant and discordant pairs, respectively. When  $(y_k - y_l)/(x_k - x_l) = 0$ ,  $1/2$  is added to both  $N_c$  and  $N_d$ . If  $x_k = x_l$ , no comparison is made.

The correlation measure proposed by Kendall (1938) in the version of Goodman and Kruskal (1963) is

$$\tau = \frac{N_c - N_d}{N_c + N_d} \quad (53)$$

where all pairs  $(x_k, y_k)$  and  $(x_l, y_l)$  with  $x_k \neq x_l$  are compared.

If it is computed for the pair  $(Y, Z)$  of Table 16,  $N_c = 70.5$  and  $N_d = 30.5$ , thus  $\tau = 0.3960$ . For the two-sided alternative, the two-tailed  $p$ -value is twice the smaller of the one-sided  $p$ -values given approximately by

$$p(\text{lower-sided}) = \text{pr}\left\{Z \leq -\frac{(N_c - N_d + 1)\sqrt{18}}{\sqrt{n(n-1)(2n+5)}}\right\} \quad (54)$$

$$p(\text{upper-sided}) = \text{pr}\left\{Z \geq \frac{(N_c - N_d - 1)\sqrt{18}}{\sqrt{n(n-1)(2n+5)}}\right\} \quad (55)$$

In our case, Eq. (54) yields 0.0212 and Eq. (55) yields 0.0268, so that the  $p$ -value is 0.0424 and the  $\tau$  coefficient is different from zero at 5%.

Daniels (1950) proposed the use of Spearman's  $\rho$  test for trend by pairing measurements with the time (order) at which the measurements were taken. In Section 2.3.4 case 2, the Cox–Stuart test for trend is presented. Tests of trend based on Spearman's  $\rho$  or Kendall's  $\tau$  are generally considered to be more powerful than the Cox and Stuart test. When it is applied to normal data, the

ARE is about 0.78 with respect to the test based on the correlation coefficient, while the ARE of these tests using Spearman's  $\rho$  or Kendall's  $\tau$  is about 0.98 under the same conditions. However, these tests are not as widely applicable as the Cox–Stuart test, in the particular case when the trend is a periodical one.

*Jonckheere–Terpstra test.* Either Spearman's  $\rho$  or Kendall's  $\tau$  can be used in the case of several independent samples to test the null hypothesis that all of the samples came from the same distribution against the ordered alternative that the distributions of variables  $X_1, X_2, \dots, X_t$  differ in a specified direction.

$$H_0: F_{X_1} = F_{X_2} = \dots = F_{X_t}.$$

$$H_a: F_{X_1} \geq F_{X_2} \geq \dots \geq F_{X_t} \text{ with at least one inequality.}$$

Sometimes the above alternative hypothesis is written as  $H_a: E(X_1) \leq E(X_2) \leq \dots \leq E(X_t)$  with at least one inequality. For the present test, the data and the null hypothesis are the same as for the Kruskal–Wallis test (Section 2.4.3), but the Kruskal–Wallis test is sensitive to any difference in means, while this usage of Spearman's  $\rho$  or Kendall's  $\tau$  is sensitive against only the order specified in the alternative hypothesis.

### Example 17

After a polluting spill, the content of a toxic analyte in five parcels at an increasing distance (1, 4, 6 and 8 km) of the contamination point has been analyzed. It is desired to confirm the hypothesis that the contamination level is decreasing with the distance. In Table 18, the obtained values have been written down, along with the rank of each measurement. It can be observed that the results obtained for each distance belong to intervals [6, 11.5], [5.6, 9.6], [4.7, 8.3] and [4.6, 7.9] in the parcels that are at 1, 4, 6 and 8 km, respectively. Evidently, the intersection of these intervals is not at all empty.

The alternative hypothesis gathers the experimenter's opinion, that is, the distribution of the values in the parcels follow the order  $F_{X_1} \leq F_{X_2} \leq F_{X_3} \leq F_{X_4}$ . Hence, the procedure is to use a variable  $X$  that takes the values 1, 2, 3 and 4 for the parcels at distance 1, 4, 6 and 8 km, respectively. As there are 4 measurements per parcel the mid-ranks assigned according to the distance are used.

**Table 18** Data for Example 17, application of the Jonckheere–Terpstra test

Distance in km to the contamination point			
1	4	6	8
9.8 (19)	8.3 (15.5)	4.7 (2)	6.3 (7)
11.5 (20)	6.6 (9)	5.0 (3.5)	5.0 (3.5)
8.6 (17)	6.8 (10)	8.3 (15.5)	7.9 (13)
6.0 (6)	5.6 (5)	7.1 (11)	7.4 (12)
8.1 (14)	9.6 (18)	6.4 (8)	4.6 (1)

Values in ppt of a pollutant agent in parcels, which are at increasing distances of the contamination point. In parentheses, the rank of the measure.

These mid-ranks would be 3, 8, 13 and 18 for the parcels at distance 1, 4, 6 and 8 km, respectively. The Spearman's  $\rho$  is equal to  $-0.5278$ , which is significantly non-null, thus the null hypothesis should be rejected and we must admit that the distribution of the polluting agent follows an increasing order with the distance (decreasing mean values).

Kendall (1975) has shown that the Kendall's  $\tau$  can be extended to the case of partial correlations. The analogous result for the Spearman's  $\rho$  can be found in Section 5.4 of the book (Conover, 1999).

### 2.4.7 Non-parametric linear regression method

Only one method of univariate linear regression will be summarized to show how the arguments based on the order stay, but the subject of the non-parametric regression and its relation with the robust methods of regression has had an enormous development, and the reader will have to consult specific monographs on these questions. The books (Sprent, 1989; Conover, 1999; Govindarajulu, 2007) contain chapters on the subject of the non-parametric regression with abundant commented bibliography. Particularly extensive is Chapter 5 dedicated to the linear model in Hettmansperger (1984).

In the following, we will assume  $n$  pairs of data  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , and the problem is to find a regression line that "approximates" (fits) these data. Also, we assume that the data follow a linear model of the form

$$Y = \alpha + \beta X + \varepsilon \quad (56)$$

where  $\varepsilon$  is a random variable (which is not observable), thus the values  $y_i$  are also a random variable, although, in principle,  $\alpha$  and  $\beta$  are constants. Consequently, any estimate of  $\alpha$  and  $\beta$  is indeed a random variable. It is assumed that the values  $\varepsilon_i$ ,  $i = 1, \dots, n$ , are independent and equally distributed.

The solution given by least squares, LS, determines the values  $a$  and  $b$ , which minimize the sum of the squares of the residuals, that is, the sum of the squared differences between the experimental data and those calculated by the fitted equation (Draper and Smith, 1998). Formally

$$\min_{a,b} \left\{ \sum_{i=1}^n (y_i - (a + bx_i))^2 \right\} \quad (57)$$

This method was first published by Legendre in 1805, but the question of the priority between Gauss and Legendre over this discovery is still being discussed. Its optimality is stated by the following theorem (Gauss-Markoff (Scheffé, 1959), p. 14): If the hypotheses about  $\varepsilon$  are fulfilled, the LS estimate has the minimum variance in the class of all unbiased linear estimates. If, in addition, the errors  $\varepsilon$  are normally distributed, then this estimator has minimum variance among all unbiased estimators.

There are hundreds of monographs on LS regression, the one already cited by Draper and Smith (1998) is adequate for an introduction to the subject.

**Table 19** Data for the regression analysis by Theil's method

$x$	0	1	2	3	4	5	6
$y$	3.0	3.1	3.4	4.0	4.6	5.1	8.1

*Theil's regression method.* It is a method proposed in 1950 by Theil to estimate the slope of a regression line as the median of the slopes of all the lines that join pairs of points with different values in the abscissa. The estimate,  $\tilde{b}$ , of the slope  $\beta$  is obtained by

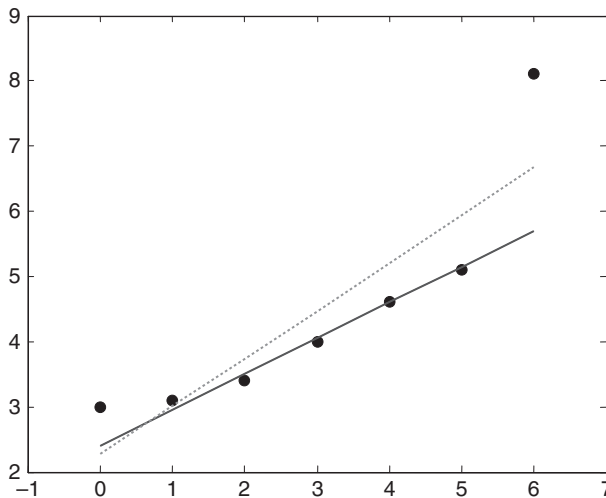
$$\tilde{b} = \text{median} \left\{ b_{ij} = \frac{y_j - y_i}{x_j - x_i}, x_j \neq x_i \right\} \quad (58)$$

The process is somewhat reminiscent of the one for obtaining the Hodges–Lehmann estimator in the Wilcoxon signed rank procedure. In fact, we will see that it is related to the Kendall's  $\tau$  coefficient. Theil suggested estimating  $\alpha$  by

$$\tilde{a} = \text{median} \{ y_i - \tilde{b} x_i \} \quad (59)$$

In order to show the calculations and with purely didactic intention, the data of [Table 19](#) will be analyzed.

The set of 21 slopes has as its median the one in position 11 that equals 0.55, therefore  $\tilde{b} = 0.55$ . Applying [Eq. \(59\)](#),  $\tilde{a} = 2.40$ , so that the Theil regression fitted to data of [Table 19](#) is  $y = 2.40 + 0.55x$ ; its representation is in [Figure 3](#). This figure also shows the LS regression line, which is  $y = 2.275 + 0.7321x$ , rather different from the non-parametric estimation, and shows the effect of datum (6, 8.1), which much influences the regression causing that the straight line does not follow the tendency of the majority of data, as the Theil's does.



**Figure 3** Data of [Table 19](#). The continuous line is the regression line obtained with the Theil's method; the dotted line is the LS regression line.



It is interesting to show the argument that serves as support to the calculation of the confidence interval on the slope  $\hat{b}$  because it shows the relation to Kendall's  $\tau$ . A good estimate of  $\beta$  would verify that residuals associated with each observation should be equally likely positive or negative. This implies an assumption that the  $e_i$  are randomly distributed with a zero median and independent of the  $x_i$ . Now

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} = \frac{(a + bx_j + e_j) - (a + bx_i + e_i)}{x_j - x_i} = b + \frac{e_j - e_i}{x_j - x_i} \quad (60)$$

Equation (60) implies that none of the slopes  $b_{ij}$  of the straight lines between two points will be greater than  $b$  if  $(x_i, e_i)$  and  $(x_j, e_j)$  are concordant; and the slope  $b_{ij}$  will be less than  $b$  if  $(x_i, e_i)$  and  $(x_j, e_j)$  are discordant, in the sense defined in the introduction to Kendall's  $\tau$  (Section 2.4.6).

As the estimate of Theil is the median of the slopes  $b_{ij}$ , half of the pairs of data are discordant and the other half are concordant; consequently any value of  $b$  that is compatible with correlation zero between the residuals and the values of the abscissas (in the sense of Kendall's  $\tau$ ) will be in the looked confidence interval. For computing the  $100(1-\alpha)\%$  confidence interval, it suffices to adapt Eqs (53)–(55), so that

$$r = \frac{1}{2} (N - w_{1-(\alpha/2)}) \quad (61)$$

$$s = \frac{1}{2} (N + w_{1-(\alpha/2)}) + 1 = N + 1 - r \quad (62)$$

where  $w_{1-(\alpha/2)} = Z_{1-(\alpha/2)} \sqrt{n(n-1)(2n+5)/18}$ ,  $N$  is the number of computed slopes  $b_{ij}$  and  $n$  is the number of data for the regression. With the data of the previous regression,  $N = 21$ ,  $n = 7$ , for 95% confidence,  $Z_{1-(\alpha/2)} = 1.96$ , thus  $r = 3$  and  $s = 19$ . Therefore, the endpoints of the confidence interval are the slopes with ranks 3 and 19, that is, 0.30 and 1.36. Hence, the estimated value is 0.55 and the 95% confidence interval is (0.30, 1.36). Evidently, the same procedure applies to test the hypothesis  $\beta = \beta_0$  against a two-sided or a one-sided alternative. There are non-parametric procedures adequate for comparing several regression lines and also methods based on ranks for non-linear regression, specially for monotone regression, that is, when it is known that the values of  $y$  increase (decrease) with those of  $x$  without imposing any functional model, like the one in Eq. (56); for these and other extensions, consult the references in Sprent (1989) and Conover (1999).

### 3. ORDER IN GRAPHS

Following the introduction of Gross and Yellen (1999) we can say that arrangements of nodes and connections are in various fields. Formally, these arrangements can be modelled by combinatorial structures called graphs. Graphs are

highly versatile models for analyzing many practical problems in which points and connections between them have some physical or conceptual meaning.

Formally, a graph  $G = (V, E)$  is a mathematical structure consisting of two sets  $V$  and  $E$ . The elements of  $V$  are called vertices (or nodes) and the elements of  $E$  are called edges. Each edge has a set of one or two vertices associated with it, which are called end points. An edge between two vertices creates a connection in two opposite senses. A line drawing the choice of the direction is indicated by placing an arrow on an edge. In that sense we will have a directed edge (or arc); and a directed graph (or digraph) is a graph whose edges are all directed.

It is usual to draw the graphs in form of points and lines. However, this is useful only for small graphs. In general, a formal specification of a graph is required, which is a finite, non-empty list of vertices, a finite list of its edges and a two-row incidence table whose columns are indexed by the edges. The entries in the column corresponding to edge  $e$  are the end points of  $e$ .

Another interesting representation of graphs is the incidence matrix, which again is linked to some order relations defined in the set of vertices and edges. Formally, the incidence matrix  $I$  of a graph  $G = (V, E)$  is the matrix whose rows and columns are indexed by some orderings of  $V$  and  $E$ , respectively, such that the  $ij$ -element  $I_{ij} = 0$  if vertex  $i$  is not an end point of edge  $j$ ,  $I_{ij} = 1$  if vertex  $i$  is an end point of edge  $j$  and  $I_{ij} = 2$  if edge  $j$  is a self-loop at vertex  $i$ . If the graph is directed, in the definition, distinction between sense of connection should be made, so that  $I_{ij} = 1$  if the connection in edge  $j$  starts in vertex  $i$  whereas  $I_{ij} = -1$  if the connection goes in the reverse sense.

Also, there is another common matrix representation for graphs, which is called the adjacency matrix  $A$ . It is the matrix whose rows and columns are both indexed by identical orderings of the set of vertices, and it contains the number of edges (respectively arcs if the graph is directed) between the considered two vertices and the number of self-loops in the principal diagonal.

The incidence and adjacency matrices of most graphs usually contain many (in fact, mostly) zeros, so there are more efficient representations that are more efficient, computationally speaking. Reference ([Gross and Yellen, 1999](#)) revises some of them.

A partial order, in fact a poset, may be represented in form of diagrams, such as the Hasse diagrams and Young diagrams. For details on the construction of these diagrams and their applications in Chemistry and Environmental Sciences, the reader is referred to [Brüggemann and Carlsen \(2006\)](#).

Ranking is used in many other fields, such as Chemistry (Pavan, 2003). In that case, graphical representations are of use.

#### 4. ORDER IN OPTIMIZATION PROBLEMS

We have already said that the lack of a “good” order for many sets affects in many scopes, one of the most important being optimization, where ranking is essential.

Optimization refers to finding one or more feasible solutions that correspond to extreme values of one or more objectives or criteria. When an optimization problem involves only one objective, the task of finding the optimal solution is called single-objective optimization, whereas if the problem involves more than one objective, it is known as multi-objective optimization.

Although it is clear that single-objective optimization is a particular case of multi-objective optimization, in practice, there is a rather big difference between a single- and a multi-objective optimization. This difference is related to the order relations that can be defined. If the  $n$  objectives are quantified as scalar-valued real functions –which is the most usual case– “less than or equal to,  $\leq$ ”, is a total order defined in  $\mathbb{R}$  but not in  $\mathbb{R}^n$  in such a way that the single-objective optimization is clearly posed but the meaning of better solution in terms of multiple objectives is not necessarily so. For example, let us suppose that we have two objectives quantified as  $f_1$  and  $f_2$ , which we need to minimize, and we have two points in the decision space (it does not matter in the discussion where it is)  $x_1$  and  $x_2$  such that  $f_1(x_1) = 1, f_1(x_2) = 2, f_2(x_1) = 3, f_2(x_2) = 0$ . We see that  $x_1$  is better than  $x_2$  in the first objective  $f_1$  but worse in the second objective  $f_2$ .

This is probably the reason why not enough emphasis is usually given to multi-objective optimization. These usual approaches avoid the complexities involved in a true multi-objective optimization problem and transform it into a single-objective optimization by using some user-defined parameters or weighting functions in the form of preference functions (Sawaragi et al., 1985; Ríos et al., 1989) or desirability functions (Lewis et al., 2000).

However, there is an essential difference between single- and multi-objective optimization, which is ignored when using the transformation method. If the objective functions are conflicting, we have a set of trade-off solutions where a gain in one objective calls for a loss in other, so that none of these trade-off solutions is the best for all objectives.

The general formulation of a multi-objective optimization problem for  $M$  real (objective) functions depending on  $n$  variables with  $J + K$  restrictions is as follows (Sawaragi et al., 1985; Deb, 2001):

$$\begin{aligned}
 &\text{Minimize/maximize} && f_m(\mathbf{x}) && m = 1, 2, \dots, M \\
 &\text{subject to} && g_j(\mathbf{x}) \geq 0 && j = 1, 2, \dots, J \\
 &&& h_k(\mathbf{x}) = 0 && k = 1, 2, \dots, K \\
 &&& x_{L_i} \leq x_i \leq x_{U_i} && i = 1, 2, \dots, n
 \end{aligned} \tag{63}$$

where  $x_{L_i}$  and  $x_{U_i}$  are the lower and upper bounds of each variable  $x_i$ , respectively. These bounds define a decision variable space  $D$  inside the intersection of the domains of each  $f_m$ . A solution  $\mathbf{x}$  is an  $n$ -dimensional vector of variables in  $D$ ,  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ . If it further satisfies all the  $J + K$  constraints in Eq. (63), then it is called a feasible solution.

If we define function  $\mathbf{f}$  to be the function whose components are  $f_m$ , i.e.  $\mathbf{f} = (f_1, f_2, \dots, f_M)^T$ , we have a vector-valued function that represents the objective function and whose image is in a multi-dimensional space, called the objective space,  $O$ .

In this way, for each  $x$  in the decision space  $D$ , there is a point in the objective space  $O$ ,  $f(x) = (f_1(x), f_2(x), \dots, f_M(x))^T$ , therefore mapping takes place between an  $n$ -dimensional space (solutions) and an  $M$ -dimensional space (objectives). That is the reason why sometimes multi-objective optimization is known as vector optimization. In our example,  $f = (f_1, f_2)$  and we have obtained  $f(x_1) = (1, 3)$  and  $f(x_2) = (2, 0)$ .

In any case, to solve the optimization problem is to find a feasible solution  $x$  in  $D$  that provides the best  $f(x)$  in  $O$ . Here best refers to the least value if we are minimizing or the greatest one if we are maximizing. For ease of reading and without losing any generality, let us suppose that we are solving a minimization problem; we lose no generality because the duality principle states that any maximization problem can be converted into a minimization problem by simply changing the sign of the corresponding objective function.

The question is that the selection of the best feasible solution (that which provides the *smallest* value of the objective function  $f$ ) implies necessarily the definition of an order relation in  $O$  that allows comparing different solutions. It sounds reasonable to say that  $x_1 \in D$  is better than  $x_2 \in D$ , if

$$f_m(x_1) \leq f_m(x_2) \text{ for each } m=1, 2, \dots, M \quad (64)$$

i.e.  $x_1$  is better than  $x_2$  in all the objectives  $f_m$ . This is in fact a quite common order relation, which we will still denote as  $f(x_1) \leq f(x_2)$  and which is known as Pareto order.

However, this is a partial-order relation because not all pairs of points in  $O$  are comparable with this relation. Remember  $(1, 3)$  and  $(2, 0)$  in the example, they are incomparable with the order in Eq. (64), because neither  $(1, 3) \leq (2, 0)$  nor  $(2, 0) \leq (1, 3)$ . What we have are chains of comparable elements and we realize that the notion of optimality does not necessarily exist for partial orders.

That means that (unlike single-objective optimization problems) an optimal solution in the sense that one that simultaneously minimizes all the objective functions does not necessarily exist and consequently, we are troubled with conflicts among objectives, in such a way that different solutions may produce trade-offs among different objectives, i.e. a set of points in the domain of feasible solutions that is extreme with respect to one objective requires a compromise in other objectives.

Hence, in a multi-objective optimization problem, it is rather difficult to obtain a unique optimal solution. Solving the problem often leads to a set of feasible solutions. So instead of speaking about optimal solution, the notion of efficiency is introduced. An element  $o$  of the objective space is said to be an efficient (non-inferior, non-dominated) element with respect to the order being considered if there does not exist an element  $o'$ , which is better than  $o$ . In the (partial) order previously defined, this means that if  $o = (o_1, o_2, \dots, o_M)$ , there is no  $o' = (o'_1, o'_2, \dots, o'_M)$  with  $o'_m \leq o_m$  for all  $m$ . In other words, these are the minimal elements in each chain of comparable elements.

The definition can be extended to the points in the decision variable space by saying that a solution  $x$  in  $D$  is an efficient solution if there is no solution  $x'$  such

that  $f(x) \leq f(x')$ , and in this context an efficient solution is also known as a Pareto optimal solution.<sup>1</sup>

Formally, a solution  $x$  to the problem defined in Eq. (63) is said to dominate another solution  $x'$  if both conditions (i) and (ii) hold

- (i) The solution  $x$  is not worse than  $x'$  in all the objectives,  $m = 1, \dots, M$ .
- (ii) The solution  $x$  is strictly better than  $x'$  in at least one objective function.

Note that the dominance relation is not an order relation, not even a partial order relation because it is not transitive, not reflexive, not symmetric and not anti-symmetric.

Among a set of solutions,  $\mathcal{P}$ , the non-dominated set of solutions  $\mathcal{P}'$  is composed of those solutions that are not dominated by any member of the set  $\mathcal{P}$ . The Pareto optimal set is the non-dominated set of the entire decision space. The goal of a multi-objective optimization problem might be to find the set of efficient solutions, in other words, the set of Pareto optimal solutions or Pareto front.

This approach is usual in the context of multi-criteria decision making (Ríos et al., 1989; Yu, 1985) and less usual in Chemistry. However, some works have been done; for instance, in Cela et al. (2003), the methodology is used to develop automated optimization of gradient separations in reversed-phase high-performance liquid chromatography (HPLC); in Sarabia et al. (2003), the conflict between bias and variance in regression is studied; the trade-off between sensitivity and specificity in modelling problems is tackled in Sánchez et al. (2005); and in Ortiz et al. (2006) and Díez et al. (2008) a more general approach to find Pareto optimal solutions in the context of multiple responses fitted by an experimental design (Response Surface Methodology) is presented.

## REFERENCES

- Frank, I., Todeschini, R. (1994). *The Data Analysis Handbook, in Data Handling in Science and Technology*, Elsevier, Amsterdam.
- Allen, I.E., Sharpe, N.R. (2005). Demonstration of ranking issues for students: A case study, *J. Stat. Educ.* 13(3). Available in [www.amstat.org/publications/jse/v13n3/sharpe.html](http://www.amstat.org/publications/jse/v13n3/sharpe.html).
- Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*, Mac Graw Hill, New York.
- Kendall, M. (1975). *Rank Correlation Methods, Fourth Edition Second Impression*, Charles Griffin & Company, Ltd., London.
- Sprent, P. (1989). *Applied Nonparametric Statistical Methods*, Chapman and Hall, Ltd., London.
- Conover, W.J. (1999). *Practical Nonparametric Statistics* 3rd ed., John Wiley & Sons, Inc., New York.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden Day Inc., San Francisco.
- Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*, John Wiley & Sons, New York.
- Govindarajulu, Z. (2007). *Nonparametric Inference*, World Scientific Publishing Co., Singapore.
- Youden, W.J. (1953). Sets of three measurements, *Sci. Mon.* 57, 143.
- Kendall, M.G. (1954). Two problems in sets of measurements, *Biometrika* 41, 500–562.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* 1, 80–83.

<sup>1</sup> If we replace  $\leq$  by  $<$  in Eq. (64), the efficient solutions will be a weak Pareto optimal solutions.

- Mann, H.B., Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Ann. Math. Stat.* 18, 50–60.
- Kendall, M., Stuart, A. (1979). *The Advanced Theory of Statistics, Vol. 2, Inference and relationship, Section 32.11*, Charles Griffin & Company Limited, pp. 547–548.
- Patel, J.K. (1986). Tolerance limits. A review, *Communications in statistics, Theory Met.* 15(9), 2716–2762.
- Iman, R.L., Hora, S.C., Conover, W.J. (1984). A comparison of asymptotically distribution-free procedures for analysis of complete blocks, *J. Am. Stat. Assoc.* 79, 674–685.
- Hora, S.C., Iman, R.L. (1988). Asymptotic relative efficiency of the rank-transformation procedure in randomized complete blocks, *J. Am. Stat. Assoc.* 83, 462–470.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *J. Am. Stat. Assoc.* 32, 675–701.
- Iman, R.L., Davenport, J.M. (1980). Approximation of the critical region of the Friedman statistics, *Commun. Stat.* A9, 571–595.
- Spearman, C. (1904). The proof and measurement of association between two things, *Am. J. Psychol.* 15, 72–101.
- Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika* 30, 81–93.
- Goodman, L.A., Kruskal, W.H. (1963). Measures of association for cross-classifications, III. Approximate sample theory, *J. Am. Stat. Assoc.* 58, 310–364.
- Daniels, H.E. (1950). Rank correlation and population models, *J. R. Stat. Soc.* 12, 171–181.
- Draper, N., Smith, H. (1998). *Applied Regression Analysis*, 3rd ed., John Wiley & Sons Inc., New York.
- Scheffé, H. (1959). *The Analysis of Variance*, John Wiley & Sons, New York.
- Gross, J., Yellen, J. (1999). *Graph Theory and its Applications*, CRC Press, Boca Raton, Florida.
- Brüggemann, R., Carlsen, L. (ed.) (2006). *Partial Order in Environmental Sciences and Chemistry*, Springer-Verlag, Berlin.
- Pavan, M. (2003). Total and Partial Ranking Methods in Chemical Sciences, PhD Thesis University of Milano Bicocca. <http://michem.dista.unimib.it/chm/>
- Sawaragi, Y., Nakayama, H., Tanino, T. (1985). *Theory of Multiobjective Optimization, Mathematics in Science and Engineering*, vol. 176, Academic Press Inc., London.
- Ríos, S., Ríos-Insua, M.J., Ríos-Insua, S. (1989). *Procesos de decisión multicriterio*, Eudema S.A. (Ediciones de la Universidad Complutense S.A.), Madrid.
- Lewis, G., Mathieu, D., Phan-Tan-Luu, R. (2000). *Pharmaceutical Experimental Design*, Marcel Decker, New York.
- Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons Ltd., Chichester.
- Yu, P.L. (1985). *Multiple-Criteria Decision Making. Concepts, Techniques and Extensions, Mathematical Concepts and Methods in Science and Engineering*, vol. 30, Plenum Press, New York.
- Cela, R., Martínez, J.A., González-Barreiro, C., Lores, M. (2003). Multiobjective optimisation using evolutionary algorithms: its application to HPLC separations, *Chemom. Intell. Lab. Syst.* 69, 137–156.
- Sarabia, L.A., Sarabia, D., Ortiz, M.C. (2003). Performance of RP (Regression in Prediction) method. Effect of resampling procedure. In: *PLS and Related Methods* (Vilares, M., Tenenhaus, M., Coelho, P., Vinzi, V.E., Morineau, A., eds), DECISIA-CERESTA, Montreuil, pp. 473–484.
- Sánchez, M.S., Ortiz, M.C., Sarabia, L.A., Lletí, R. (2005). On Pareto-optimal fronts for deciding about sensitivity and specificity in class-modelling problems, *Anal. Chim. Acta* 544, 236–245.
- Ortiz, M.C., Sarabia, L.A., Herrero, A., Sánchez, M.S. (2006). Vectorial optimization as a methodological alternative to desirability function, *Chemom. Intell. Lab. Syst.* 83, 157–168.
- Díez, R., Sarabia, L.A., Sánchez, M.S., Ortiz, M.C. (2008). How to search the experimental conditions that improve a Partial Least Squares calibration model. Application to a flow system with electrochemical detection for the determination of sulfonamides in milk, *Chemom. Intell. Lab. Syst.* 92, 71–82.

APPENDIX A

Table A1 Critical values for the Wilcoxon signed rank test

		Sample size														
		6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
One-tail test																
$\alpha = 0.05$		2	3	5	8	10	13	17	21	25	30	35	41	47	53	60
$\alpha = 0.01$	<sup>a</sup>	0	1	3	5	7	9	12	15	19	23	27	32	37	43	
Two-tail test																
$\alpha = 0.05$		0	2	3	5	8	10	13	17	21	25	29	34	40	46	52
$\alpha = 0.01$	<sup>a</sup>	<sup>a</sup>		0	1	3	5	7	9	12	15	19	23	27	32	37

Source: Adapted from P. Sprent, Applied Non-Parametric Statistical Methods, Chapman and Hall Ltd, London, 1989.

<sup>a</sup>Sample too small for test at this level.

**Table A2** Critical values for the Wilcoxon–Mann–Whitney  $U$  statistic (unequal sample sizes)

<i>One-tail test</i>																
<i>m</i>	<i>n</i>															
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6	2		8	10	12	14	16	17	19	21	23	25	26	28	30	32
7	3	4		13	15	17	19	21	24	26	28	30	33	35	37	39
8	4	6	7		18	20	23	26	28	31	33	36	39	41	44	47
9	5	7	9	11		24	27	30	33	36	39	42	45	48	51	54
10	6	8	11	13	16		31	34	37	41	44	48	51	55	58	62
11	7	9	12	15	18	22		38	42	46	50	54	57	61	65	69
12	8	11	14	17	21	24	28		47	51	55	60	64	68	72	77
13	9	12	16	20	23	27	31	35		56	61	65	70	75	80	84
14	10	13	17	22	26	30	34	38	43		66	71	77	82	87	92
15	11	15	19	24	28	33	37	42	47	51		77	83	88	94	100
16	12	16	21	26	31	36	41	46	51	56	61		89	95	101	107
17	13	18	23	28	33	38	44	49	55	60	66	71		102	109	115
18	14	19	24	30	36	41	47	53	59	65	70	76	82		116	123
19	15	20	26	32	38	44	50	56	63	69	75	82	88	94		130
20	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	



**Table A2** (Continued)

<i>Two-tail test</i>																
<i>m</i>	<i>n</i>															
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	1		6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	1	3		10	12	14	16	18	20	22	24	26	28	30	32	34
8	2	4	6		15	17	19	22	24	26	29	31	34	36	38	41
9	3	5	7	9		20	23	26	28	31	34	37	39	42	45	48
10	4	6	9	11	13		26	29	33	36	39	42	45	48	52	55
11	5	7	10	13	16	18		33	37	40	44	47	51	55	58	62
12	6	9	12	15	18	21	24		41	45	49	53	57	61	65	69
13	7	10	13	17	20	24	27	31		50	54	59	63	67	72	76
14	7	11	15	18	22	26	30	34	38		59	64	69	74	78	83
15	8	12	16	20	24	29	33	37	42	46		70	75	80	85	90
16	9	13	18	22	27	31	36	41	45	50	55		81	86	92	98
17	10	15	19	24	29	34	39	44	49	54	60	65		93	99	105
18	11	16	21	26	31	37	42	47	53	58	64	70	75		106	112
19	12	17	22	28	33	39	45	51	57	63	69	74	81	87		119
20	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	

Source: Adapted from P. Sprent, *Applied Non-Parametric Statistical Methods*, Chapman and Hall Ltd, London, 1989.  
 The maximum value of the lesser of  $U_m, U_n$  indicates significance in a one- or two-tail tests for unequal samples sizes between 5 and 20. Values above and to the right of the diagonal apply to significance at a nominal 5% level; values below and to the left of the diagonal apply to significance at a nominal 1% level.

## CHAPTER 2

# Total-Order Ranking Methods

**M. Pavan and R. Todeschini**

---

Contents	1. Introduction	51
	2. Total-Order Ranking Methods	53
	2.1 Pareto optimality	55
	2.2 Simple additive ranking	56
	2.3 Multiattribute value theory	58
	3. Conclusions	70
	References	70

---

## 1. INTRODUCTION

Total-order ranking methods belong to multicriteria decision making (MCDM), a discipline in its own right, which deals with decisions involving the choice of a best alternative from several potential candidates in a decision, subject to several criteria or attribute that may be concrete or vague.

Typically, a decision problem is a situation where an individual has alternative courses of action available and has to select one of them, without an a priori knowledge of which is the best one. A decision process can be organized in three phases. The first phase is problem identification and structuring, which consists in the identification of the purpose of the decision, in the recognition of the problem to be solved, in the diagnosis of the cause–effect relationships for the decision situation and in the identification of the judgment criteria. The second phase is the so-called model development and use, which consist in the development of formal models, decision maker preferences, values, trade-offs, goals to compare the alternatives or actions under consideration to each other in a systematic and transparent way. The third phase is the development of action plans since the analysis does not solve the decision. The decision process, which results in the selection of the best solution, i.e. the solution where the positive outcomes

outweigh possible losses, is efficient if the procedure to reach the solution is optimal. The aims of a decision process are to effectively generate information on the decision problem from available data, to effectively generate solutions and to provide a good understanding of the structure of a decision problem.

Multicriteria decision-making techniques are used for helping people in making their decision according to their preferences, in cases where there is more than one conflicting criterion, finding the optimal choice among the alternatives. Making a decision is not just a question of selecting a best alternative. Often the need is to rank all the alternatives for resource allocation or to combine the strengths of preferences of individuals to form a collective preference.

Mathematics applied to decision making provides methods to quantify or prioritize personal or group judgments that are typically intangible and subjective. Decision making requires comparing different kinds of alternatives by decomposing the preferences into many properties that the alternatives have, determining their importance, comparing and obtaining the relative preference of alternatives with respect to each property and synthesizing the results to get the overall preference. Therefore, the strategy consists in breaking down a complex problem into its smaller components and establishing importance or priority to rank the alternatives in a comprehensive and general way to look at the problem mathematically.

The key starting point of MCDM lies in attempting to represent often intangible goals in terms of the number of individual criteria. A challenging feature of MCDM methods is the identification of the set of criteria by which alternatives are to be compared. The selection criteria is part of the modeling and problem formulation, a significant phase often under-emphasized. A useful general definition of a criterion is the one provided by [Bouyssou \(1990\)](#) as a tool allowing comparison of alternatives according to a particular axis or point of view. It is generally assumed that each criterion can be represented by a surrogate measure of performance, represented by some measurable attribute of the consequences arising from the achievement of any particular decision alternative.

Some thoughts are to be considered in identifying the criteria: their value relevance, i.e. their link with the decision maker concept of their goals; their understandability and their measurability, i.e. the performance of the alternative against the criteria should be measurable; their non-redundancy in order to avoid that the concept they represent is in attributed greater importance; their judgmental independence, i.e. the preferences with respect to a single criterion should be independent from the level of another; their balancing between completeness and conciseness.

Subjectivity is intrinsic in all decision making and in particular in the choice of the criteria on which the decision is based and in their relative weight. Multicriteria decision making does not dissolve subjectivity, but it makes the need for subjective judgments explicit and the whole process by which they are considered is made transparent.

Over the years, several MCDM methods have been proposed (Hobbs and Horn, 1997) in different areas, with different theoretical background and facing different kinds of questions and providing different kinds of results (Hobbs and Meier, 1994).

Some of these methods have been developed to fulfill the need of specific problems; other methods are more general and have been used in different areas. The different MCDM methods are distinguished from each other in the nature of the model, in the information needed and in how the model is used. They have in common the aim to create a more formalized and better-informed decision-making process, the need to define alternatives to be considered, the criteria to guide the evaluation and the relative importance of the different criteria.

In this chapter, the theory of the mostly known total-order ranking techniques is described.

## 2. TOTAL-ORDER RANKING METHODS

Once the decision problem identification phase has generated a set of alternatives, which can be a discrete list of alternatives as well as be defined implicitly by a set of constraints on a vector of decision variables, and once the set of criteria against which the alternatives have to be analyzed and compared are defined, then a decision model has to be built to support decision makers in searching the optimal or the set of satisfactory solutions to the multicriteria decision problem. The decision model is made of two main components as described by Belton and Stewart (2003):

- Preferences in terms of each individual criterion, i.e. models describing the relative importance or desirability of achieving different levels of performance for each identified criterion. In addition, 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 criteria setting is a crucial point since it requires the mathematization of decision criteria, which are often not completely defined or explicit.
- An aggregation model, i.e. a model allowing inter-criteria comparisons (such as trade-offs), in order to combine preferences across criteria. 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. Multicriteria decision-making methods are often based on an aggregation function  $\Gamma$  of the criteria  $f_j$ , where  $j = 1, \dots, p$ :

$$\Gamma \equiv \gamma(\phi_1, \phi_2, \dots, \phi_p) \quad (1)$$

Thus, if an alternative is characterized by  $p$  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  $\Gamma$ . Several evaluation methods, which define a ranking parameter generating a total-order ranking, have been proposed in the literature.

The purpose of the decision model is to create a view of decision maker preferences based on a defined set of assumptions and to guide the decision maker in the optimum solution search.

Before reviewing some total-order ranking methods, some further terms and basic principles are introduced.

A  $p$ -dimensional system is generally considered, with an associated  $(n \times p)$  data matrix  $X$ . To each of the  $n$  alternatives, a set of  $p$  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 of alternatives  $A(a, b, c, d)$ , a ranking (order) on  $A$  is a relation with the following properties:

$$a \leq a \quad (\text{reflexivity}) \quad (2)$$

$$a \leq b \quad \text{and} \quad b \leq a \Rightarrow b = a \quad (\text{antisymmetry}) \quad (3)$$

$$a \leq b \quad \text{and} \quad b \leq c \Rightarrow a \leq c \quad (\text{transitivity}) \quad (4)$$

A set  $A$  equipped with the relation  $\leq$  is said to be an ordered set. An MCDM method can generate:

- a complete or total ranking:  $a > b > c > d$  (or linear order)
- the best alternative:  $a > (b, c, d)$
- a set of acceptable alternatives:  $(a, b, c) > d$
- an incomplete ranking of alternatives:  $a > (b, c, d)$  or  $(a, b) > (c, d)$

The correct definition of a criterion may imply a more or less objective ordering of the alternatives according to this criterion, the orientation of the criterion, i.e. the direction of preference for the criterion. For each criterion, it has to be explicitly established whether the best condition is satisfied by a minimum or a maximum value of the criterion. Where this well-defined measure of performance exists, the performance level or the attribute value of the alternative  $a$  according to criterion  $j$  can be represented by  $f_{aj}$ . To simplify the discussion, all criteria are supposed to be defined in such a way that increasing values are preferred.

The preference function is a measure of performance for criterion  $j$  and is a partial preference function in the sense that alternative  $a$  is strictly preferred to  $b$  in terms of criterion  $j$  if and only if  $f_{aj} > f_{bj}$ . These preference functions may correspond to natural attributes on a cardinal scale or may be built on ordinal or categorical scales. The only property that these functions need to satisfy is the ordinal preferential independence, i.e. it must be possible to rank order

alternatives on one criterion independently of performances in terms of the other criteria. Once partial preference functions have been associated with each criterion, a check is performed on the existence of any pairs of alternatives  $a$  and  $b$  for which  $a$  is at least as good as  $b$  on all criteria (i.e.  $f_{aj} \geq f_{bj}$  for all  $j$ ) and is strictly preferred to  $b$  on at least one criterion (i.e.  $f_{aj} > f_{bj}$  for at least one  $j$ ). In this case, the vector of performance measures  $\mathbf{f}_a$  is said to dominate  $\mathbf{f}_b$ .

## 2.1 Pareto optimality

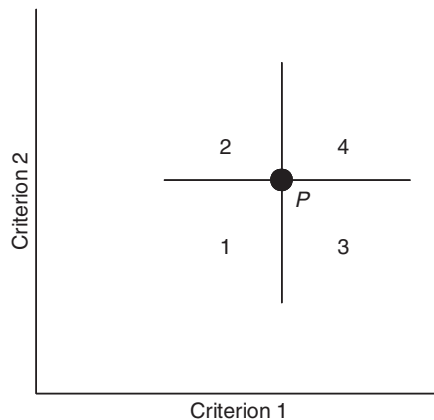
Pareto optimality is a MCDM method introduced into chemometrics by Smilde et al. (Smilde et al., 1986; Smilde et al., 1987; Boer et al., 1988; Doornbos et al., 1990; Keller and Massart, 1990).

If for two alternatives  $a$  and  $b$ ,  $f_{aj} \geq f_{bj}$  for all the criteria ( $1 \leq j \leq p$ ), with at least one inequality, then we say that the alternative  $a$  dominates  $b$ . Alternatives that are not dominated by any other are termed Pareto optimal (PO) points (efficient points). The Pareto frontier is the set of PO points.

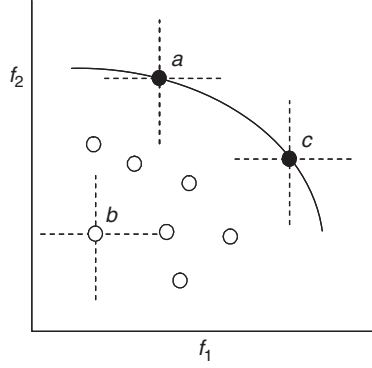
The Pareto optimality technique selects the PO points and the points that are not PO points are inferior to the PO points with respect to at least one criterion. Figure 1 illustrates a two-dimensional criterion space.

A point corresponds to a setting of two criteria, the criterion values of which are plotted against each other. The space around the point  $P$  can be divided into four quadrants. In the case of two criteria, both to be maximized, 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 PO point is superior to all other comparable points; thus in the case of Figure 2 representing the space of two criteria  $f_1$  and  $f_2$ , both to be



**Figure 1** Representation of the four quadrants in a two-dimensional criterion space around point  $P$ .



**Figure 2** Bivariate representation of the criteria  $Y_1$  and  $Y_2$ . Points  $a$  and  $c$  are Pareto optimal (PO) points.

maximized, a point  $a$  is superior to another point  $b$  if the following conditions are verified:

$$f_{a1} > f_{b1} \quad \text{and} \quad f_{a2} > f_{a2} \quad \text{or} \quad (5)$$

$$f_{a1} > f_{a1} \quad \text{and} \quad f_{a1} = f_{a1} \quad \text{or} \quad (6)$$

$$f_{a1} = f_{b1} \quad \text{and} \quad f_{a2} > f_{a2} \quad (7)$$

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

If the system under study is described by more than two criteria, the  $p$ -dimensional criterion space ( $p > 2$ ) containing the PO points can 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.

## 2.2 Simple additive ranking

The simple additive ranking (SAR) method (Zimmermann and Gutsche, 1991; Eisenführ et al., 1986; French, 1988) is a very intuitive approach to MCDM, based on the ranking of the alternative with respect to each criterion separately and the subsequent aggregation of the weighted ranks by arithmetic mean, finally normalized. The resulting score, which defines the performance of the alternative, is computed as:

$$S_i = \frac{\sum_{j=1}^p w_j r_{ij}}{n}, \quad 1/n \leq S_i \leq 1 \quad (8)$$

with the following constraints:

$$\sum_{j=1}^p w_j = 1, \quad 0 \leq w_j \leq 1, \quad j = 1, p \quad (9)$$

where  $r_{ij}$  is the rank of the  $i$ th alternative for the  $j$ th criterion and the number  $n$  of alternatives a normalization factor, under the weight constraint:

$$\sum_{j=1}^p w_r = 1$$

It is assumed that the best and the worst values are rank  $n$  and rank 1, respectively. To obtain an opposite ranking, i.e. assuming the best and the worst values are rank 1 and  $n$ , respectively,  $S_i$  should be simply transformed as  $1 - S_1$ .

Finally, in order to obtain scores between 0 and 1, a further scaling can be performed as:

$$S'_i = \frac{S_i - 1/n}{1 - 1/n}, \quad 0 \leq S'_i \leq 1 \quad (10)$$

Being based on the separate ranking of each criterion, this method is robust with respect to anomalous values obtained for some alternatives in some criterion.

To help better understand the method, consider the decision problem to select the best alternative (e.g. the best phone) out of the three alternatives whose performances judged by five criteria are illustrated in Table 1. Let  $c_1$  and  $c_5$  be criteria to be minimized (e.g. price and weight), while  $c_2$ ,  $c_3$  and  $c_4$  be criteria to be maximized (e.g. standby battery duration, number of functionalities and appeal). Information about the criteria, their relative importance and their ranges are provided in Table 2. The results of the SAR method applied on the example data of Tables 1 and 2 are provided in Table 3.

**Table 1** Values of alternatives for criteria

Alternatives	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$
1	125	60	2	2	90
2	275	48	4	4	110
3	415	32	5	3	200

**Table 2** Criteria, weights and ranges

Criteria	Weights	Ranges Minimum	Maximum
$C_1$ Price (Euro)	0.35	60	600
$C_2$ Standby battery duration (hour)	0.23	10	72
$C_3$ Functionalities (no. of functionalities)	0.18	1	5
$C_4$ Appeal (1= ugly; 5= beautiful)	0.12	1	5
$C_5$ Weight (g)	0.12	50	300



**Table 3** Estimated simple additive ranking (SAR) scores

Alternatives	SAR score
1	0.700
2	0.560
3	0.240

2.3 Multiattribute value theory

Multiattribute value function methods provide a synthesis of the performances of alternatives against individual criteria, together with inter-criteria information, which reflect the relative importance of the different criteria, in order to provide an overall estimate of each alternative reflecting the decision makers’ preferences. Therefore, the purpose of this approach is to associate a number, a global ranking index, to each alternative in order to produce a preference order of the alternatives according to the numerical value of the ranking index. Therefore, a number (score)  $S_a$  is associated to each alternative  $a$  in such a way that  $a$  is judged to be preferred to  $b$  ( $a > b$ ), taking into account all criteria, if and only if  $S_a > S_b$ , which implies indifference between  $a$  and  $b$  if and only if  $S_a = S_b$ .

As pointed out by [Belton and Stewart \(2003\)](#), the preference order implied by any value function should provide a complete or total order. This means that preferences are complete, i.e. for any pair of alternatives, either one is strictly preferred to the other or there is indifference between them. In addition, preferences and indifferences are transitive, i.e. for any three alternatives  $a$ ,  $b$  and  $c$ , if  $a \leq b$  and  $b \leq c$  then  $a \leq c$ .

Once the initial model structure has been defined together with the set of alternatives, then the next step consists in drawing the information required by the model. In the case of multiattribute value function models, there are two types of information needed, sometimes referred to as intra-criterion information and inter-criterion information, or as well scores and weights.

2.3.1 Intra-criteria information (scores)

Scoring is the process of assessing the partial value functions for each criterion in the decision model. Partial value functions are defined so that alternative  $a$  is preferred to  $b$  in terms of criterion  $j$  if and only if  $f_{aj} > f_{bj}$ , while indifference between  $a$  and  $b$  in terms of this criterion exists if and only if  $f_{aj} = f_{bj}$ . The score values need to be defined on an interval scale of measurement, on which the key factor is the difference between points. Therefore, a ratio of values will have meaning only if the zero point on the scale is absolutely and unambiguously defined. Thus, to build such a scale it is necessary to define two reference points and to distribute numerical values to these points. The reference points are often taken at the bottom and top of the scale, to which are assigned values such as 0 and 100 or 0 and 1. The minimum and maximum points on the scale can be defined on a local scale or on a global scale.

The local scale is defined by the set of alternatives under consideration, and the minimum and maximum points correspond to the alternative that does least well on a particular criterion (score = 0) and the one that does the best (score = 1). All other alternatives obtain intermediate scores ( $0 < \text{score} < 1$ ), according to their performance relative to the two reference points. The use of local scales has the advantage that it allows a quite quick assessment of values and thus it is useful under time constraints.

A global scale is defined by taking into account a wider set of possibilities. The reference points correspond to the ideal and the worst possible performance on each specific criterion or to the best and the worst performance that could realistically take place. The definition of a global scale requires additional information and in general more work than a local scale. However, the advantage of the global scale is that it is more general and therefore it can be defined before the examination of the specific alternatives. Another way to describe a global scale consists in specifying reference points in terms of neutral and good performance levels. More details on this approach can be found in Bana e Costa and Yansnick (Bana e Costa and Vansnick, 1999).

In cases of both local and global scales, it is important that all subsequent analyses, like the assessment of the weights, are consistent with the chosen scaling.

For the illustrative example of Tables 1 and 2, the local scale for the first criterion is 125 and 415, i.e. the minimum and maximum values of the actual data, while the global scale is 60 and 600, i.e. the minimum and maximum values corresponding to the ideal and the worst possible performance on that criterion as defined by the decision maker.

Once the reference points have been defined, the next step, which consists in assigning the other scores, can be performed in three ways:

- defining a partial value function
- building a qualitative value scale
- providing a direct rating of the alternatives

#### 2.3.1.1 Defining a partial value function

To define a partial value function, a measurable attribute scale closely related to the decision maker values needs to be identified. The partial value function can be assessed directly or by using indirect questioning. In the case of direct assessment, which is often supported by a visual representation, the decision maker should evaluate whether:

- The value function is monotonically increasing against the natural scale: the highest value of the criterion is most preferred, while the lowest is least preferred.
- The value function is monotonically decreasing against the natural scale: the lowest value of the criterion is most preferred, while the highest is least preferred (cost criteria).
- The value function is non-monotonic: an intermediate point of the scale corresponds to the most preferred or the least preferred point (pH = 7).

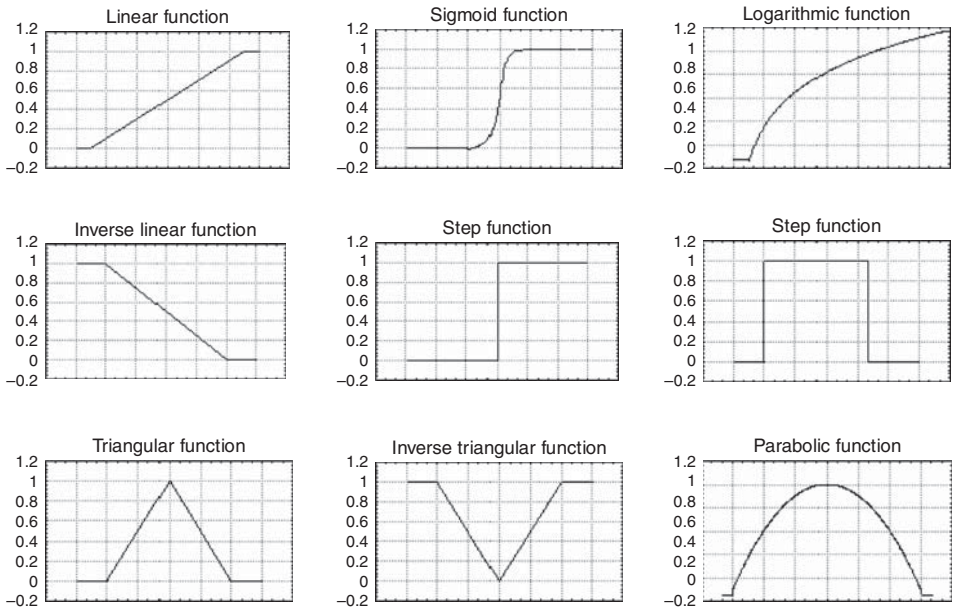
As demonstrated by [Stewart \(1993, 1996\)](#), the final results of the decision-making analysis are strongly dependent on these assumptions, and the defaults assumption of linearity, which is commonly made, may generate confusing results.

Examples of monotonically increasing, monotonically decreasing and non-monotonic value functions are illustrated in [Figure 3](#).

The partial value functions can also be assessed indirectly, assuming that the value function is monotonically increasing or decreasing. Two main methods are widely used for this purpose, called “bisection” and “difference” as described by [von Winterfeldt and Edwards \(1986\)](#) and [Watson and Buede \(1987\)](#).

Using the bisection method, the decision maker has to identify the point on the criterion scale that is halfway, in value terms, between the two minimum and maximum reference points.

The difference approach includes a set of methods all requiring the decision maker to consider increments on the objectively measured scale and to relate them to differences in value. One of these methods, described by Watson and Buede, requires the partition of the criterion scale into a defined number of equal intervals. The decision maker then has to rank the specified differences according to increase in associated value. Another indirect approach, described by [von Winterfeldt and Edwards \(1986\)](#), consists in defining a unit level on the criterion scale (suggested between one-tenth and one-fifth of the difference between the minimum and maximum reference



**Figure 3** Examples of monotonically increasing, decreasing and non-monotonic value functions.

points) and questioning on the unit levels. Examples of these approaches are also described in [Belton and Stewart \(2003\)](#).

### 2.3.1.2 Building a qualitative value scale

An appropriate qualitative scale is needed in all those circumstances where it is not possible to find a measurable scale for an identified criterion. Like for the other approaches, at least the two reference points (minimum and maximum) have to be already defined. Qualitative scales should show evidence of some requirements:

- Operational: they should permit the decision makers to rate alternatives not used in the definition of the scale.
- Reliable: two independent ratings of an alternative should lead to the same score.
- Value relevant: they should be related to the decision maker's objective.
- Justifiable: they should be capable of convincing an independent observer that the scale is reasonable.

An example of this process is described in [Belton and Stewart \(2003\)](#).

### 2.3.1.3 Direct rating of the alternatives

The direct rating approach is another way that can be used for the construction of a value scale; it requires only the definition of the reference points of the scale. Both local and global scales can be used. In the case of a local scale, the best alternative is given the highest rate (usually 100, 10 or 1) and the worst alternative is given a score of 0. All the other alternatives are then located directly on the scale reflecting their performance with respect to the two reference points. The main disadvantage in using local scales concerns the fact that if new alternatives are subsequently introduced into the decision problem, then these scales have to be redefined, and consequentially the weighting criteria.

## 2.3.2 Inter-criteria information (weights)

An important piece of information, which also strongly affect the final results of the decision-making process, is the one related to the relative importance of the criteria. The weight assigned to a criterion  $j$  is essentially a scaling factor that relates scores on that criterion to scores on all other criteria. Thus if a criterion  $j$  is assigned a weight, which is twice that of another criterion  $k$ , this should be interpreted that the decision maker values 10 value points on criterion  $j$ , which is the same as 20 value points on criterion  $k$ . These weights are referred to as swing weights to make a distinction to the concept of importance weights ([Belton and Stewart, 2003](#)). It is rather common to make the error of assuming that weights are independent of the measurement scales used, while the effect of the weight parameter  $w_j$  is directly linked to the scaling used for the partial preference function and the two are closely connected.

To assign swing weights, the method commonly followed is that from the worst value to the best value on each criterion. The decision maker may be asked

to consider all bottom-level criteria at the same time. The swing that gives the greatest increase in overall value is the one that will have the highest weight. The process is then applied on the remaining set of criteria and proceeds until a ranking of the criteria weights has been determined. Once a ranking has been defined, they need to be assigned values. One of the possible ways to do so is that the decision maker is asked directly to compare each criterion with the highest ranked one. Thus, the decision maker is asked to define the increase in overall value resulting from an increase from score 0 to a score of 100 on the selected criterion as a percentage of the increase in overall value resulting in an increase from score of 0 to 100 on the highest ranked criterion. Weights are then generally normalized to sum up to 1 or 100. The normalization allows an easier interpretation of the importance of the criteria.

Weights for ranked criteria can be calculated as explained below. Without loss of generality, it can be assumed that the  $p$  criteria are ranked from 1 to  $p$ , where the criterion 1 is ranked 1, criterion 2 is ranked 2 and criterion  $p$  is ranked  $p$ .

If the decision maker is able to rank the criteria, a simple consistent weighting of the criteria can be obtained with respect to the first one as  $w_j = w_1/r_j^k$ , where  $r_j$  is the  $j$ th criterion rank,  $k$  is a smoothing parameter, assuming positive values in order to preserve the criterion ranking. The smoothing parameter  $k$  influences the relative differences between the criterion weights; in particular, for  $k=0$ , all the criteria are equally weighted; while increasing  $k$ , the weight of the first criterion becomes more and more relevant with respect to the other lower-ranked criteria, as well as the weight of the second criterion becomes more and more relevant with respect to the other lower-ranked criteria. In practice, increasing  $k$  increases the differences of the criterion weights. Negative  $k$  values can also be used to explore a reverse criterion ranking.

The normalized weights  $w'$  for ranked criteria can be calculated as:

$$w_j' = \frac{Q/r_j^k}{\sum_{j=1}^p Q/r_j^k} \quad (11)$$

where  $Q$  is defined as:

$$Q = \prod_{j=1}^p r_j^k = \exp \left[ \sum_{j=1}^p k \ln(r_j) \right] \quad (12)$$

Some examples of weights are given in [Table 4](#) for different  $k$  values and criterion ranks, for four criteria ( $p=4$ ).

### 2.3.2.1 Sensitivity and robustness

Once the intra-criterion information and inter-criterion information have been defined, a good practice is to check whether the preliminary assumptions are robust or if they are sensitive to changes in the model. Technically a sensitivity analysis consists in examining the effect on the output of a model resulted by

**Table 4** Examples of normalized weight calculations from ranks

$k$	Criterion	1	2	3	4
0.5	Rank	1	2	3	4
	Weight	0.359	0.254	0.207	0.180
1.0	Rank	1	2	3	4
	Weight	0.480	0.240	0.160	0.120
1.5	Rank	1	2	3	4
	Weight	0.598	0.212	0.115	0.075
1.0	Rank	1.5	1.5	3	4
	Weight	0.348	0.348	0.174	0.130
1.5	Rank	1.5	1.5	3.5	3.5
	Weight	0.390	0.390	0.110	0.110

changes in input parameters of the model. In the case of decision-making model, the parameters are the partial value functions, the scores and weights assigned by the decision maker, while the output is the overall evaluation of the alternatives. Therefore, a sensitivity analysis is useful to identify which, in case there is any, of the input parameters provide a crucial influence on the overall evaluation. The sensitivity analysis is also helpful for the decision maker to confirm his understanding of the problem, and when it is performed in a group context, it provides the opportunity to consider and/or explore alternative views of the problem.

Different methods belong to the multiattribute value function approach; three most commonly applied methods are illustrated below. More detailed discussion on the value measurement theory can be found in [Keeney and Raiffa \(1976\)](#), [Roberts \(1979\)](#), von Wintefeldt and Edwards (1986), [French \(1988\)](#), [Keller and Massart \(1991\)](#), [Hendriks et al. \(1992\)](#) and [Lewi et al. \(1992\)](#).

Several variants of the multiattribute value function approach have been proposed; among these there is the so-called interactive methods based on value function approach, which comprises methods based on trade-off information ([Geoffrion et al., 1972](#)), methods using direct comparisons ([Zionts and Wallenius, 1976](#); [Zionts and Wallenius, 1983](#)) and the so-called convex cone approach ([Korhonen et al., 1984](#)).

Another way to evaluate the influence of the weights assigned to the criteria is proposed here; it explores different  $k$  values of the function (11).

The proposed approach is based only on the definition of the criterion ranks, and it does not require the definition of the numerical criterion weights and therefore is based on a less level of arbitrariness. The numerical weights are subsequently derived from the criterion ranks.

A plot of the Spearman rank correlation function and the Pearson correlation calculated on the ranked scores and on the scores, respectively, obtained using two successive  $k$  values versus the  $k$  values gives a simple and immediate view of the behaviour of the weights in scoring the objects.

To illustrate this approach a simple example is given.

The considered data set is composed of seven objects described by four criteria (C1–C4) (Table 5).

It has been assumed that for criterion C2,  $A > E > W$  and for criterion C3,  $A > F$ . Moreover, maximum values for each criterion are the optimal solution. Therefore, the ranks obtained from each criterion are given in Table 6.

Using the SAR approach, the scores are obtained from formula (10), assuming that the best and the worst values are rank 1 and  $n$  (Table 7).

**Table 5** Original data

Object	C1	C2	C3	C4
1	300	E	F	10
2	250	E	A	10
3	250	A	F	10
4	200	A	F	5
5	200	A	A	10
6	200	W	F	10
7	100	W	A	5

**Table 6** Ranked data

Object	C1	C2	C3	C4
1	1	6.5	5.5	5
2	2.5	6.5	2	5
3	2.5	4	5.5	5
4	5	4	5.5	1.5
5	5	4	2	5
6	5	1.5	5.5	5
7	7	1.5	2	1.5

**Table 7** Scores obtained by simple additive ranking method

Object	C1	C2	C3	C4
1	0.000	0.917	0.750	0.667
2	0.250	0.917	0.167	0.667
3	0.250	0.500	0.750	0.667
4	0.667	0.500	0.750	0.083
5	0.667	0.500	0.167	0.667
6	0.667	0.083	0.750	0.667
7	1.000	0.083	0.167	0.083

The Spearman rank correlation (triangles) and the Pearson correlation (squares) are calculated on each pair of adjacent ranks; the numerical results

Weights	$K$								
	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
$w1$	0.034	0.059	0.100	0.162	0.250	0.360	0.483	0.603	0.708
$w2$	0.136	0.167	0.200	0.230	0.250	0.255	0.241	0.213	0.177
$w3$	0.415	0.387	0.350	0.304	0.250	0.193	0.138	0.092	0.058
$w4$	0.415	0.387	0.350	0.304	0.250	0.193	0.138	0.092	0.058

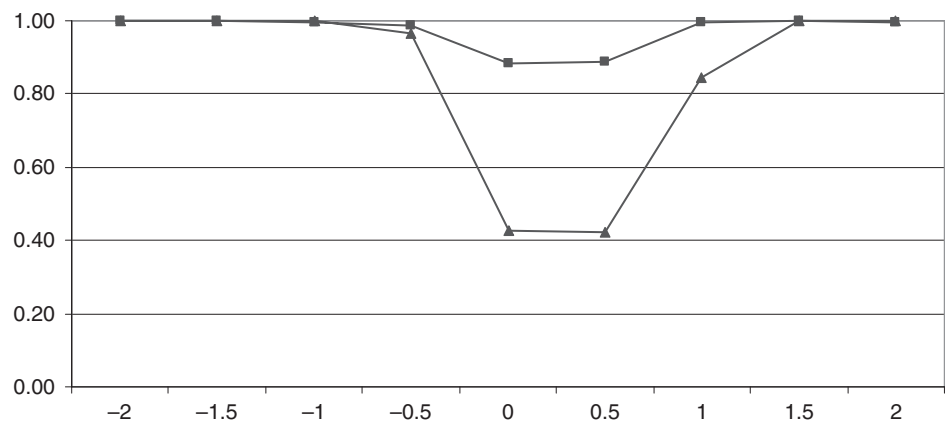
Objects	$K$								
	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
1	0.713	0.701	0.679	0.641	0.583	0.506	0.417	0.326	0.244
2	0.479	0.490	0.500	0.504	0.500	0.484	0.457	0.423	0.387
3	0.665	0.646	0.621	0.586	0.542	0.490	0.437	0.388	0.347
4	0.436	0.445	0.458	0.476	0.500	0.528	0.557	0.585	0.608
5	0.436	0.445	0.458	0.476	0.500	0.528	0.557	0.585	0.608
6	0.622	0.601	0.579	0.558	0.542	0.534	0.537	0.550	0.568
7	0.149	0.170	0.204	0.258	0.333	0.430	0.537	0.644	0.737

Objects	$K$								
	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
1	1	1	1	1	1	4	7	7	7
2	4	4	4	4	5	6	5	5	5
3	2	2	2	2	2	5	6	6	6
4	5	5	5	5	5	2.5	1.5	2.5	2.5
5	5	5	5	5	5	2.5	1.5	2.5	2.5
6	3	3	3	3	2	1	3.5	4	4
7	7	7	7	7	7	7	3.5	1	1



**Table 11** Spearman and Pearson correlations

$k$	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
Spearman	1.00	1.00	1.00	0.97	0.43	0.42	0.84	1.00	1.00
Pearson	1.00	1.00	1.00	0.99	0.88	0.89	1.00	1.00	1.00



**Figure 4** Spearman (triangles) and Pearson (squares) correlation plot.

are given in Table 11 and the  $k$ - $\rho$  plot in Figure 4. Note that both the Spearman and Pearson correlations for  $k=2$  have been assumed to be equal to 1.

The region for reverse ranks is between  $-2$  and  $0$ . As it can be easily observed from the graph, different ranking are obtained when  $k$  value increases from  $0$  (equal weights for all the criteria) to  $0.5$ , from  $0.5$  to  $1$  and from  $1$  to  $1.5$ , i.e. the sensible region for the criteria weighting is between  $k$  values  $0.5$  and  $1.5$ , where different ranks are obtained. If reverse ranks are given, from  $k = -2$  to  $k = -0.5$ , the same ranks are always obtained. However, the Pearson correlation, which is calculated on the scores, reveals that in this region, only small differences in the score values are obtained.

### 2.3.3 Utility and desirability

Utility functions and desirability functions are well-known MCDM methods. The approach is the form most simply and easily understood by decision makers from a variety of backgrounds, since it does not require any stronger restrictions on the preference structures than the aggregation formula. They are based on the definition of partial value function, i.e. a transformation function  $t_i$ , for each criterion in order to standardize the partial value functions transforming values of the criteria to the same scale. Typically the best and worst conditions need to be defined for each criterion. This can be done locally, taking simply the best and worst of the available alternatives, or more generally as the best and worst possible conditions in similar contexts. For this purpose, different kinds of

functions can be used, the more common ones being linear, sigmoid, logarithmic, exponential, step, normal, parabolic, Laplace, triangular and box (Figure 3). Each criterion is independently transformed into a utility/desirability function  $t_{ij}$  by an arbitrary function, which transforms the actual value  $f_{ij}$  of each  $i$ th alternative for the  $j$ th criterion into a value between 0 and 1.

Once the kind of function and its trend for each criterion has been defined, the overall utility/desirability of each  $i$ th alternative is computed. Utility and desirability functions differ only for the aggregation form of the overall utility  $U$  and desirability  $D$ .

The overall utility  $U_i$  of each  $i$ th alternative is defined, for the unweighted and weighted cases, as arithmetic mean:

$$U_i = \frac{\sum_{j=1}^p t_{ij}}{p}, U_i = \sum_{j=1}^p w_j t_{ij}, \quad 0 \leq U_i \leq 1 \quad (13)$$

A particular case of the utility function is the so-called simple additive weighting (SAW) method (Zionts and Wallenius, 1983), which consists in the utility function approach performed by using a linear transformation on a local scale for each  $j$ th criterion.

In the case of the desirability method, first presented by Harrington (1965) and then generalized by Derringer (Derringer and Suich, 1980), the overall desirability  $D_i$  of each  $i$ th alternative is defined, for the unweighted and weighted cases, as geometric mean:

$$D_i = \sqrt[p]{t_{i1} t_{i2} \dots t_{ip}}, \quad D_i = t_{i1}^{w_1} t_{i2}^{w_2} \dots t_{ip}^{w_p}, \quad 0 \leq D_i \leq 1 \quad (14)$$

In all the cases, the weight constraint is assumed:

$$\sum_{j=1}^p w_j = 1$$

It can be noticed that the overall desirability is calculated more severely than the utility: in fact, 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 1 only if all the desirabilities have the maximum value of 1.

Once the overall utility  $U_i$  or desirability  $D_i$  for each alternative has been calculated, all the alternatives can be totally ranked according to their  $U$  or  $D$  values, and the element with the highest  $U$  or  $D$  can be selected as the best one, if its value is considered acceptable.

A desirability scale, given in Table 12, was developed by Harrington (1965).

Both utility and desirability functions are affected by arbitrariness related to the a priori selection of the partial value functions and corresponding upper and

**Table 12** Harrington's qualitative definition of the desirability scale

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

lower limits. Moreover, these functions are very easy to calculate, and no specific software is required.

The critical feature of these approaches to MCDM problems is the establishment of the relation between criteria and partial value functions values, which must be performed by the decision maker.

The simplicity of the additive aggregation makes the utility function approach particularly appealing. Only relatively minor assumptions are needed, and these are primarily related to the criteria definition and to the interpretation of partial value functions and weights. Three additional requirements have been illustrated by [Belton and Stewart \(2003\)](#), derived by simple algebraic properties of the additive form, such as preferential independence, interval scale property and weights as scaling constants.

### 2.3.4 Dominance

The dominance function method is based on the comparison of the state of the different criteria for each pair of alternatives. This approach does not require the transformation of each criterion into a quantitative partial value function; it only requires establishing whether the best condition is satisfied by a minimum or a maximum value of the selected criterion. For each pair of alternatives  $(a, b)$  three sets of criteria are determined:

$P^+(a, b)$  is the set of criteria where  $a$  dominates  $b$ , i.e. where  $a$  is better than  $b$ ,  $P^0(a, b)$  is the one where  $a$  and  $b$  are equal and  $P^-(a, b)$  is the set of criteria where  $a$  is dominated by  $b$ .

The dominance function between two alternatives  $a$  and  $b$  is calculated considering—separately—the weights for the criteria in the  $P^+$  and  $P^-$  sets as follows:

$$C(a, b) = \frac{1 + \sum_{j \in P^+(a, b)} w_j}{1 + \sum_{j \in P^-(a, b)} w_j}, \quad 0.5 \leq C(a, b) \leq 2 \quad (15)$$

where  $P^+(a, b) \cup P^-(a, b) \subseteq F$  and with the usual constraint  $\sum_{j=1}^p w_j = 1$ .

A  $C(a, b)$  value equal to 1 means equivalence of the two alternatives;  $C(a, b) > 1$  means that the alternative  $a$  is, on the whole, superior to the alternative  $b$ ,

whereas  $C(a,b) < 1$  means that the alternative  $a$  is, on the whole, inferior to the alternative  $b$ . The obtained values can be normalized according to the following equation:

$$C'(a,b) = \frac{C(a,b) - 0.5}{2 - 0.5}, \quad 0 \leq C'(a,b) \leq 1 \quad (16)$$

A global score of the alternative  $a$  is then calculated as:

$$\Phi_a = \sum_{i=1}^n C'(a,i), \quad i \neq a, \quad 0 \leq \Phi_a \leq n-1 \quad (17)$$

and the corresponding scaled value is:

$$\Phi'_a = \frac{\Phi_a}{n-1}, \quad 0 \leq \Phi'_a \leq 1 \quad (18)$$

A total ranking is obtained on  $\Phi$  and the highest values are the best alternatives.

The results of desirability, utility and dominance functions applied to the illustrative example of [Tables 1 and 2](#) on the global scale and by using linear transformations for all the partial value functions are illustrated in [Table 13](#).

In case the local scale is used instead of the global one, the following results illustrated in [Table 14](#) are provided. It can be noticed the more severity provided by the desirability approach with respect to the utility one. The second and third alternatives are judged as not desirable because of their low performance on the third and first criteria, respectively, no matter what their performances are on the other criteria. The dominance results are independent of the scale used, being based on a pair comparison approach.

**Table 13** Estimated values of alternatives for criteria computed on global scale

Alternatives	Desirability	Utility	Dominance
1	0.592	0.675	0.538
2	0.671	0.676	0.436
3	0.466	0.510	0.138

**Table 14** Estimated values of alternatives for criteria computed on local scale

Alternatives	Desirability	Utility	Dominance
1	0	0.700	0.538
2	0.618	0.639	0.436
3	0	0.240	0.138

### 3. CONCLUSIONS

Total-order ranking methods are MCDM techniques used for the ranking of various alternatives on the basis of more than one criterion. We reviewed a number of total-order ranking methods, which have been widely used to facilitate the structuring and understanding of the perceived decision problem. We presented here the simplest approaches, like the Pareto optimality and the SAR approach followed by the approaches belonging to the so-called multiattribute value theory, i.e. utility, desirability and dominance functions. In these models, numerical scores are constructed to represent the degree to which an alternative may be preferred to another. These scores are developed initially for each criterion and aggregated into a higher level of preference models. Several other methods have been proposed over the years. Among these an important one is the outranking model category (Roy and Bouyssou, 1993), which includes PROMETHEE (Brans, and Vincke, 1985; Brans et al., 1984), ELECTRE (Roy and Bouyssou, 1993) and ORESTE method (Rubens, 1980). In these methods, alternatives are compared pairwise, initially in terms of each criterion and then the preference information is aggregated across all the criteria. These methods attempt to set up the strength of evidence in favor of one alternative over the others.

Another group of total-order ranking methods is the one of the so-called goal programming approach (Charnes and Cooper, 1961), which includes linear goal programming, interactive goal programming, STEM method, interactive multiple goal programming and the reference point method. These methods based on the establishment of desirable or satisfactory levels of achievement for each criterion search for the alternative that is closest to these desirable goals or aspirations. Other methods are the analytic hierarchy process (AHP) method developed by Saaty (1980, 1990), which emphasizes the role of the weights of the criteria, the fuzzy set theory (Zimmermann, 1983) which attempts to solve the main problem in MCDM field related to the inevitable ambiguity in defining human preferences. Finally, Bayesian analysis is a widely used approach for knowledge representation and reasoning under uncertainty in intelligent systems (Pearl, 1988; Russell and Norvig, 1995). A complete review of the theoretical background of each of these models has been recently published (Pavan and Todeschini, 2008).

### REFERENCES

- Bana e Costa, C.A., Vansnick, J.C. (1999). The MACBETH approach: Basic ideas, software and an application. In: *Advances indecision Analysis* (Meskens, N., Roubens, M. eds.), Kluwer Academic Press (Boston/Dordrecht/London), pp. 131–157.
- Belton, V., Stewart, T.J. (2003). *Multiple Criteria Decision Analysis, An Integrated Approach*, Kluwer Academic Publisher (Boston/Dordrecht/London).
- Boer, J.H., Smilde, A.K., Doornbos, D.A. (1988). Introduction of multi-criteria decision making in optimization procedures for pharmaceutical formulations, *Eur. J. Pharm. Biopharm.* 34, 140–143.
- Bouyssou, D. (1990). Building criteria: A prerequisite for MCDA. In: *Readings in Multiple Criteria Decision Aid* (Bana e Costa, C.A., ed.), Springer-Verlay: Berlin, pp. 58–80.
- Brans, J.P., Vincke, Ph. (1985). A preference ranking organization method (the PROMETHEE Method for Multiple Criteria Decision Making), *Manage. Sci.* 31, 647–656.

- Brans, J.P., Mareschal, B., Vincke, P.H. (1984). PROMETHEE a new family of outranking methods in multicriteria analysis. In: *Oper. Res.* (Brans J.P., ed.), Springer: North Holland, Dordrecht, pp. 477–490.
- Charnes, A., Cooper, W.W. (1961). *Management Model and Industrial Applications of Linear Programming*, John Wiley & Sons, New York.
- Derringer, G.C., Suich, R. (1980). Simultaneous optimization of several response variables, *J. Qual. Technol.* 12, 214–219.
- Doornbos, D.A., Smilde, A.K., Boer, J.H., Duineveld, C.A.A. (1990). Experimental design, response surface methodology and multicriteria decision making in the development of drug dosage forms. In: *Scientific Computing and Automation (Europe)* (Karjalainen, E.J., ed.), Elsevier: Amsterdam, Chapter 8.
- Eisenführ, F., Weber, M. (1986). Zielstrukturierung: Ein kritischer Schritt im Entscheidungsprozeß, *Zeitschrift für betriebswirtschaftliche Forschung* 38, 907–929.
- French, S. (1988). *Decision Theory: An Introduction to the Mathematics of Rationality*, Ellis Horwood, Chichester.
- Geoffrion, A.M., Dyer, J.S., Feinberg, A. (1972). An interactive approach for multicriterion optimization with an application to the operation of an academic department, *Manage. Sci.* 19, 357–368.
- Harrington, E.C. (1965). The desirability function, *Ind. Qual. Control.* 21, 494–498.
- Hendriks, M.M.W.B., Boer, J.H., Smilde, A.K., Doornbos, D.A. (1992). Multicriteria decision making, *Chemom. Intell. Lab. Syst.* 16, 175–191.
- Hobbs, B.F., Horn, G.T.F. (1997). Building public confidence in energy planning: A multimethod MCDM approach to demand-side planning at BC gas, *Energy Policy* 25(3), 357–375.
- Hobbs, B.F., Meier, P.M. (1994). Multicriteria methods for resource planning: An experimental comparison, *IEEE Trans. Power Syst.* 9(4), 1811–1817.
- Keller, H.R., Massart, D.L. (1990). Program for Pareto-optimality in multicriteria problems, *Trends Analyt. Chem.* 9, 251–253.
- Keller, H.R., Massart, D.L. (1991). Multicriteria decision making: A case study, *Chemom. Intell. Lab. Syst.* 11, 175–189.
- Keeney, R.L., Raiffa, H. (1976). *Decision with Multiple Objectives*, J. Wiley & Sons, New York.
- Korhonen, P., Wallenius, J., Zionts, S. (1984). Solving the discrete multiple criteria problem using convex cones, *Manage. Sci.* 30, 1336–1345.
- Lewi, P.J., Van Hoof, J., Boey, P. (1992). Multicriteria decision making using pareto optimality and PROMETHEE preference ranking, *Chemom. Intell. Lab. Syst.* 16, 139–144.
- Pavan, M., Todeschini, R. (2008). Optimisation: Multicriteria decision making methods. In: *Comprehensive Chemometrics*, Elsevier. In press.
- Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference, Morgan Kaufmann Publishers, San Francisco, ISBN 1-55860-479-0.
- Roberts, F.S. (1979). *Measurement Theory with Applications to Decision Making, Utility and the Social Sciences*, Addison-Wesley, London.
- Roy, B., Bouyssou, D. (1993). *Aide Multicritere d'Aide à la Décision: Méthodes et Cas*, Economica, Paris.
- Rubens, M. (1980). Analyse et aggrégation des préférences: Modélisation, ajustement et résumé de données relationnelles, *Revue Belge de Statistique, d'Informatique et de Recherche Operationelle* 20, 36–67.
- Russell, S., Norvig, P. (1995). *Artificial Intelligence: A modern approach*, Prentice Hall, ISBN 0-13-103805-2.
- Saaty, T.L. (1980). *The Analytic Hierarchy Process*, MacGraw-Hill, New York.
- Saaty, T.L. (1990). How to make decision: The analytic hierarchy process, *Eur. J. Oper. Res.* 48, 9–26.
- Smilde, A.K., Knevelmann, A., Coenegracht, P.M.J. (1986). Introduction of multicriteria decision making in optimisation procedures for high-performance liquid chromatographic separations, *J. Chromatogr.* 369, 1–10.
- Smilde, A.K., Bruins, C.H.P., Doornbos, D.A., Vinck, J. (1987). Optimisation of the reversed-phase high-performance liquid chromatographic separation of synthetic estrogenic and progestogenic steroids using the multi-criteria decision making method, *J. Chromatogr.* 410, 1–12.
- Stewart, T.J. (1993). Use of piecewise linear functions in interactive multicriteria decision support: A monte carlo study, *Manage. Sci.* 39, 1369–1381.
- Stewart, T.J. (1996). Robustness of additive value function methods in MCDM, *J. Multi-Criteria Decis. Anal.* 5, 301–309.

- von Winterfeldt, D., Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.
- Watson, S.R., Buede, D.M. (1987). *Decision Synthesis. The Principles and Practise of Decision Analysis*, Cambridge University Press, Cambridge.
- Zimmermann, H.J., Gutsche, L. (1991). *Multi-Criteria Analyse*, Springer, Berlin.
- Zionts, S., Wallenius, J. (1976). An interactive programming method for solving the multiple criteria problem, *Manage. Sci.* 22, 652–663.
- Zimmermann, H.J. (1983). Using fuzzy sets in Operational Research, *Eur. J. Oper. Res.* 13, 201–216.
- Zionts, S., Wallenius, J. (1983). An interactive multiple objective programming method for a class of underlying nonlinear utility functions, *Manage. Sci.* 29, 519–529.

# Partial Ordering and Hasse Diagrams: Applications in Chemistry and Software

R. Brüggemann, K. Voigt, and S. Pudenz

---

Contents	1. Introduction	73
	2. Partial-Order Theory	74
	2.1 Basics of partial order	74
	2.2 Hasse Diagram Technique	75
	3. Software for Hasse Diagram Technique	77
	4. Ranking of Chemicals as An Example	78
	4.1 The Hasse diagram of 12 chemicals	78
	4.2 Concepts to describe and to characterize a Hasse diagram using the example test set	79
	5. Applications of Hasse Diagram Technique to the Data Availability of Chemicals	83
	5.1 Overview	83
	5.2 Stability fields and crucial weights using a data matrix of 17 databases and 4 pharmaceuticals	86
	6. Summary, Outlook and Conclusion	90
	References	91

---

## 1. INTRODUCTION

It is well-known that graph theory plays an important role in chemistry. One of the main applications of graph theory in chemistry is the derivation of topological indices as graph theoretical invariants. Topological indices in turn are the basis for many quantitative structure–activity relationships (QSAR). For details see Todeschini and Consonni (2000), Basak et al. (2000), Basak and Mills (2001) and Sabljic and Trinajstic (1981). However, graph theory in its application on chemistry is not restricted on QSAR. Reaction networks is another example where



graph theory has powerful applications; here the graphs are often directed graphs, i.e. the lines connecting the reacting molecules indicate a direction: The reaction from substance A yielding substances B and C (see e.g. [Glass, 1975](#); Nemes, et al., 1977; Zeigarnik, et al. 1996). Another example of directed graphs arises from the comparison of vectorial quantities. For example, alkanes may be ordered by a list of modified Zagreb indices (Vukicevic et al., 2007) or by properties of chemicals relevant for risk assessment: As a pre-processing step for the evaluation of chemicals (or other objects), it is useful to keep the relevant properties separated (i.e. do not crunch them into a single ranking index), while analyzed simultaneously (see e.g. [Brüggemann et al., 1999a](#)). In that cases the objects are connected if they are comparable, and the direction of the line indicates whether one object is better (whatever “better” may mean) as another one. In this chapter, we explain the nature of Hasse diagrams, give some examples from the area of environmental chemicals and their data availability. Finally, we briefly describe the software, by which partial order from the point of view of applications can be analyzed and by which partial orders can be visualized—for example, by Hasse diagrams.

Hasse diagrams got the name from the German mathematician H. Hasse, who lived from 1898 to 1979 and who used them to represent algebraic structures ([Hasse, 1967](#)). As Hasse diagrams are the visualization of a mathematical concept, namely of partial order, one has to go back until the end of the nineteenth century, where Dedekind and Vogt (see [Rival, 1985](#)) made the first important investigations. Parallel to H. Hasse, the American mathematician G. Birkhoff worked on partial orders and made this mathematical structure popular by his famous book “Lattice theory” ([Birkhoff, 1984](#)). From the pioneering work of E. Halfon, the concept of partial order was introduced in environmental sciences and chemistry ([Halfon, 1979, 1983, 2006](#); [Halfon and Reggiani, 1986](#)). The usefulness of partial order in evaluation problems was then recognized by the authors of this chapter and extended in several directions. Since 1998 regularly workshops about partial order and its Hasse diagram take place, see e.g. [<http://www.criteri-on.de/1/index.html>] as well as the reference work edited by Rainer Brüggemann and Lars Carlsen (2006). They gave a first state of art of application of partial order in environmental sciences and chemistry. The charm of the visualization technique of partial order gives an additional attractiveness, so this chapter is specifically widowed graphical representations, especially the Hasse diagram.

## 2. PARTIAL-ORDER THEORY

### 2.1 Basics of partial order

Partial order is a discipline of Discrete Mathematics; we give here only a brief overview and an introduction of the notation.

*First step:* We need a set of objects. We call this set of objects the ground set, and denote it as  $G$ . Objects can be chemicals (which are to be compared) ([Brüggemann et al., 2001b](#); [Lerche et al., 2002](#)), strategies (for example, water management) ([Simon et al., 2004a,b, 2006a](#)), geographical units ([Brüggemann et al., 1994, 1999b, 2001a](#);

Münzer et al., 1994; Sørensen et al., 2003), databases (Brüggemann and Voigt, 1996; Voigt et al., 2004b, 2006a) etc.

*Second step:* We need an operation between any two objects. As an evaluation is our aim, we must compare the objects. Is object “ $a$ ” better than object “ $b$ ”? It is practice to use the sign  $\perp$  to express that  $a$  and  $b$  are comparable and we write  $a \perp b$ .

*Third step:* We not only want that two objects are comparable, but also would like to know the orientation: Is “ $a$ ” better or worse than “ $b$ ”? Therefore, the signs  $\leq$  and  $\geq$  are introduced:  $a \leq b$  “may” denote that  $a$  is better than  $b$ ,  $a \geq b$  “may” indicate that  $a$  is worse than  $b$ .

*Fourth step:* Why the phrases: “may denote”, “may indicate”? The essential point is that we have to define, when we consider object  $a$  as better than  $b$ . That is, the signs “ $\leq$ ” and “ $\geq$ ” alone do not help in an evaluation procedure, we must give them an appropriate sense.

*Fifth step:* Independently how we define  $\leq$  and  $\geq$  resp., the ground set equipped with, for example, “ $\leq$ ” must obey three axioms, if we want to speak from a partially ordered set (poset):

*Reflexivity:* An object  $a$  can be compared with itself:  $a \leq a$

*Antisymmetry:* If  $a$  is better than  $b$ , and at the same time  $b$  is better than  $a$ , then  $a = b$ . We write:  $a \leq b$  and  $b \leq a \Rightarrow a = b$ . Later we will relax this axiom.

*Transitivity:* If  $a$  is better than  $b$  and at the same time  $b$  is better than  $c$ , then  $a$  is better than  $c$ .  $a \leq b$ ,  $b \leq c \Rightarrow a \leq c$ .

If the  $\leq$  relation is defined properly, then the ground set  $G$  equipped with  $\leq$  is a partially ordered set. A widespread notation is:  $(G, \leq)$

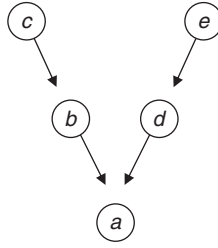
The transitivity axiom is a very important one, and it filters out many situations, which could also be considered as a matter of evaluation: For example, competitions in sports: Crack “ $a$ ” wins about crack “ $b$ ”, in turn crack “ $b$ ” wins about “ $c$ ”, but unexpectedly crack “ $a$ ” does not win about “ $c$ ”! The mathematical analysis of this kind of generalized comparisons is done in tournament theory (Simon and Brüggemann, 2000).

*Sixth step:* Why can a partially ordered set be represented by a directed graph? Consider the objects of a ground set as vertices. Then in any case, where for  $(a, b) \in G^2$  is valid  $a \leq b$  we draw an arrow starting from  $b$  and ending in  $a$ . Because of the transitivity we can omit such arrows, which are represented by a sequence of  $\leq$  relations. Hence, most of the advices on how to construct a Hasse diagram are using the concept of “cover relation”  $\prec$ : Two objects  $a, b$  for which is valid that  $a \leq b$  are in a cover relation, if there is no third object,  $c$ , with  $a \leq c \leq b$ . Then a Hasse diagram is a graph of cover relations with additional conventions of how to locate the objects in the drawing plane.

Example:  $G = \{a, b, c, d, e\}$  and  $a \leq b$ ,  $b \leq c$ ,  $a \leq c$  (necessarily)  $a \leq d$ ,  $d \leq e$ ,  $a \leq e$  (necessarily). Then the directed graph is shown in Figure 1.

## 2.2 Hasse Diagram Technique

As mentioned earlier, we have to define  $\leq$  suitably. This means that on the one side the definition must obey the order axioms, and on the other side it must model the task “evaluation”. One, but not necessarily the only one, definition is:



**Figure 1** Directed graph of  $(\{a, b, c, d, e\}, \leq)$ . The relation  $a \leq c$  and  $a \leq e$  can be derived from the sequences  $a \leq b, b \leq c$  and  $a \leq d, d \leq e$ .

Let  $(a, b) \in G^2$  then  $a \leq b: \Leftrightarrow q_i(a) \leq q_i(b)$  for all  $q_i$ . The  $q_i$  are the attributes by which the objects should be characterized for their evaluation (see below for examples).

This definition obeys the order axioms, indeed it is the well-known “product order” or “component wise” order. Nevertheless the definition contains several traps:

*Trap 1:* If  $q_i(a) = q_i(b)$  for all  $q_i$ , then we can set  $a \leq b$  and at the same time  $a \geq b$ . Nevertheless,  $a$  is not identical with  $b$ . Therefore, the evaluation based on  $q_i$  does not necessarily lead to a partial order, but a quasi- or pre-order (see for explanation De Loof et al., 2007). It is convenient to introduce the equivalence relation. Objects  $a, b$  are equivalent to each other if  $q_i(a) = q_i(b)$  for all  $i$ . By the equivalence relation the ground set can be partitioned into equivalence classes. If one takes one element out of any equivalence class (the representative of an equivalence class) and ignore the others, then we retain the partial orders for the representatives. One has to take care that the conclusions that can be drawn for the representative due to the partial order are valid for all the other elements of that equivalence class, which is represented. The interplay between order and equivalence is described in a systematic manner in Voigt et al. (2004b).

*Trap 2:* If  $a \leq b$ , but  $a$  is not equivalent with  $b$ , then at least for one  $q_i$  a strict relation  $<$  must be valid.

*Trap 3:* Whether  $a$  and  $b$  are comparable (and in which orientation) or not (notation  $a||b$ ) depends crucially on the data representation. Let us think of  $q_1(a)$  slightly larger than  $q_1(b)$ , then  $a \geq b$  although the numerical difference may be very small. In that case, it would be better to state that  $a$  and  $b$  are almost equivalent. Similarly, incomparabilities can appear, just by very small numerical differences. However, when is a numerical difference to be considered as small?

In Helm (2003, 2006), a careful rounding process is described. Often, however, the resulting poset is still very complex and it may be a good strategy to preprocess the data, for example, by a cluster analysis (Luther et al., 2000; Pudenz et al., 2000), where the statement “almost equivalent” can be sharply defined, or by assigning class scores for each attribute (Brüggemann et al., 2001a).

The problems mentioned in trap 3 can also be circumvented by applying fuzzy partial order, see Naessens et al. (2002), de Baets et al. (2002, Van der Walle et al. (1995). An application for chemicals is found in Brüggemann et al.

(2007). It is, however, noteworthy that the introduction of a fuzzy concept in partial orders dislocates the difficulties: the order relation among objects can be adequately modelled; however, one has a new quantity, the “flexibility”, which has to be subjectively selected.

*Trap 4:* In the definition of the product order, the phrase “for all  $q_i$ ” was used. If one applies this phrase literally, one will not get any useful result! The problem is that just those attributes  $q_i$  that are relevant for the evaluation (and clearly which are available) should be selected. This, however, is not a specific problem of partial-order theory but of any multi-criteria evaluation. Nevertheless, methods were developed to analyze the role of the number of attributes, which are included in the evaluation (Brüggemann et al., 2001a; Brüggemann and Carlsen, 2006). In order to express the important role of the selection of the attributes, the partial order in Hasse diagram technique (HDT) is often written as (G, IB), where IB is the information base of evaluation and is the set of all attributes used in the evaluation.

### 3. SOFTWARE FOR HASSE DIAGRAM TECHNIQUE

A good and more detailed overview can be found in a publication by Halfon (2006). Software for drawing simple Hasse diagrams was already written under MS-DOS in the late 1980s by Halfon et al. One of the main difficulties is to avoid crossings of lines and crossing of lines with the circles, denoting the objects. As an example of the rather sophisticated computational efforts, the publication of Halfon et al. (1989) should be mentioned.

In the 1990s, the software was adapted to the MS-Windows platforms (Brüggemann and Halfon, 1995). The functionalities were constantly enhanced and improved in the following years (Brüggemann et al., 1999a,b; Brüggemann et al., 2001a; Brüggemann et al., 2005). The innovative tools called METEOR (Method of Evaluation by Order Theory), which attempts to resolve the incomparabilities among objects by inclusion of external knowledge, are incorporated in the WHasse program (Brüggemann et al., 2008; Voigt and Brüggemann, 2007).

This software named WHasse has always been available for scientific purposes from the first author free of charge.

A second software product based on the HDT background written in Java was developed for commercial use by the third author. The name of this software is ProRank—software for multi-criteria evaluation and decision support. It can be obtained at <http://www.prorank.biz>. An application of this ProRank software including the description of the software tool has recently been published by the authors in the journal *Environmental Modelling & Software* (Voigt et al., 2006a).

Many new software products are developed, for example, by Sørensen et al. to model by partial-order uncertainties (Sørensen et al., 2000, 2001), general correlation between different partial orders (Sørensen et al., 2005), etc.

New programs are also developed by the Milano Chemometrics and QSAR Research Group, headed by Roberto Todeschini. Well documented is the program RANA (Pavan, 2003; Pavan et al., 2004) and DART (see Chapter 9, this book), developed by Talete Srl on behalf of the European Chemicals Bureau (ECB).

Finally, it should be mentioned that the first author Rainer Brüggemann in cooperation with G.P. Patil and several Indian computer scientists is developing a completely new Program RAPID [Ranking and Prioritization Information Delivery (working title)]. It is planned to deliver a test version in 2009.

4. RANKING OF CHEMICALS AS AN EXAMPLE

4.1 The Hasse diagram of 12 chemicals

Let us imagine that we have to evaluate 12 high production volume chemicals (HPVC). As information base, we take attributes describing the exposure: PV (classified) the production volume and  $\log K_{ow}$  as measure for accumulation.

$G_{HPVC} = \{CNB, 4NA, 4NP, \dots, THI\}$   
 $IB_{Exposure} = \{PV, \log K_{ow}\}$   
Orientation: High values indicate a hazard

Hence, for example,  $DIM \geq GLY$  because  $PV(DIM) = 2$ ,  $PV(GLY) = 2$ ,  $\log K_{ow}(DIM) = 0.7$ , whereas  $\log K_{ow}(GLY) = 0.002$ ; and  $LIN \parallel ATR$  because  $PV(LIN) = 1$ ,  $\log K_{ow}(LIN) = 2.7$ , whereas  $PV(ATR) = 2$ ,  $\log K_{ow}(ATR) = 2.5$ .

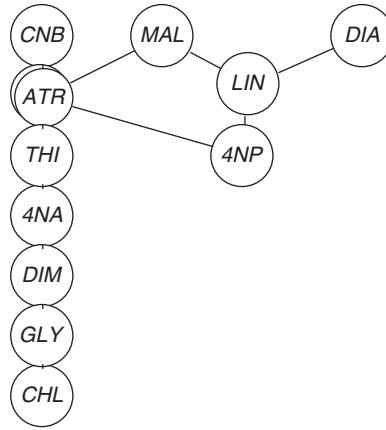
See Table 1 for full information.  
A Hasse diagram is obtained from the directed graph if

- generally all objects  $y \geq x$  are located in the drawing position higher than the object  $x$ ;
- the objects are located as far as possible in the same vertical position;

Table 1 Data of the 12 high production volume chemicals (HPVC)

	PV	negLC	Log $K_{ow}$	negBD
CNB	4	− 1.5	2.6	− 0.2
4NA	2	− 35	1.4	0
4NP	1	− 7	1.9	− 0.1
ATR	2	− 4.3	2.5	− 0.5
CHL	2	− 80	− 2.2	− 1
DIA	1	− 2.6	3.3	0
DIM	2	− 7.5	0.7	0
LIN	1	− 11	2.7	− 0.4
GLY	2	− 52	0.002	− 0.3
ISO	2	− 3	2.5	− 30
MAL	3	− 0.04	2,7	− 100
THI	2	− 0.3	1.7	0

PV, production volume; LC, lethal concentration;  $K_{ow}$ , distribution coefficient of octanol-water; BD, biodegradation.



**Figure 2** Hasse diagram of  $(G_{HPVC}, IB_{Exposure})$  generated by the software WHASSE (“Hasse for Windows”), see Brüggemann and Halfon, (1995); Brüggemann et al., (1999a).

- the arrows are replaced by simple lines (the orientation is now provided by the position of the object in the plane);
- in the HDT we indicate the vertices by circles and label them with the object names.

So, in Figure 2, a Hasse diagram of  $(G_{HPVC}, IB_{Exposure})$  is shown.

The indication of a second circle behind the full circle reminds us that there are equivalence classes, i.e. the Hasse diagram is based on the representatives of the equivalence relation “same value for PV and  $\log K_{ow}$ ”. Here ATR is the representative of the equivalence class  $\{ATR, ISO\}$ .

## 4.2 Concepts to describe and to characterize a Hasse diagram using the example test set

*Maximal elements:* elements that have no upper neighbour in a Hasse diagram are maximal elements. If there is only one maximal element, then it is called the greatest element. In Figure 2, CNB, MAL and DIA are the maximal elements. There is no greatest element.

*Minimal elements:* elements that have no lower neighbour in a Hasse diagram are minimal elements. If there is only one minimal element, then it is called the least element. In Figure 2, we find two minimal elements, CHL and 4NP.

*Isolated elements:* elements that are at the same time maximal and minimal elements, i.e. which have neither an upper nor a lower neighbour. In Figure 2, there is no isolated element.

*Chain:* subset of  $G$  where each element is comparable with each other. In Figure 2,  $CHL < GLY < DIM < 4NA < THI < \{ATR, ISO\} < CNB$  form an (ascending) chain. There are other chains too. For example,  $4NP < LIN < DIA$ . Counting

chains: by the Redheffer matrices, it is possible under some circumstances to count the number of chains between two endpoints. The Redheffer matrix is related to the well-known Möbius function (see [Wilf, 2004](#)). Note that from a computational point of view, very quick access is possible to

- the existence of chains,
- number of chains and
- shortest chain.

by appropriate use of recurrence relations. A PYTHON program, PyHasse, by the first author is under development for a deepened graph theoretical analysis of the partial order relation. (For the programming language PYTHON, see, for example, [Weigend, 2006](#) or Lutz and Ascher, 2003)

*Level:* If one starts with the maximal elements and gives them the same vertical position, they form the maximal level  $L_{\max}$ . By reducing  $G$  by these maximal elements, the reduced set has new maximal elements (here ATR and LIN). They form the level  $L_{\max}-1$ . By repeating this procedure, the level 1 is formed. Here the level 1 consists of CHL. The procedure can be performed as often as a maximal chain has elements (only the representatives are counted). Therefore, in the case of [Figure 2](#), the number of levels is 7.

*Antichain:* a subset of  $G$  where no element is comparable. For example, {CNB, MAL, DIA} is a subset, which forms an antichain.

The presence of chains indicates a potential correlation among the elements of IB or the case that only one attribute is deciding the position; the presence of antichains can be considered as diversity of the hazard. For example, the antichain, which is formed by the maximal elements, shows that these chemicals are hazardous but in a different way: CNB has the highest value of PV, DIA shows the highest value in the accumulation tendency (expressed by  $\log K_{ow}$ ), MAL has slightly higher and lower values but is still hazardous in both exposure aspects.

The number of chains and the number of elements in the maximal antichain are related to each other: The famous Dilworth theorem states that the ground set can be partitioned into those numbers of chains that are equal to the number of elements in the largest antichain (the width of a poset).

As the Hasse diagram in [Figure 2](#) was obtained from two attributes, a scatter diagram of the 12 chemicals can be obtained, without information loss. Technically, the ability to be mapped onto a  $k$ -dimensional coordinate system is called a projection. Naturally  $k = 2$  is attractive as it allows a simple geometrical presentation. [Figure 3](#) shows the scatter plot.

In a geometrical representation, elements comparable to  $x$  but worse in both aspects ( $q_1$ ,  $q_2$ ) must be located in the shadow (1) (see [Figure 4](#)), whereas elements better in both aspects must be in the shadow (2). Elements in the remaining parts are incomparable.

The chemicals MAL, CNB and DIA in relation to each other are located in the “incomparability” region of the scatter plot.

In [Figure 5](#), the Hasse diagram is shown where all four attributes are taken for the analysis. See [Table 1](#).

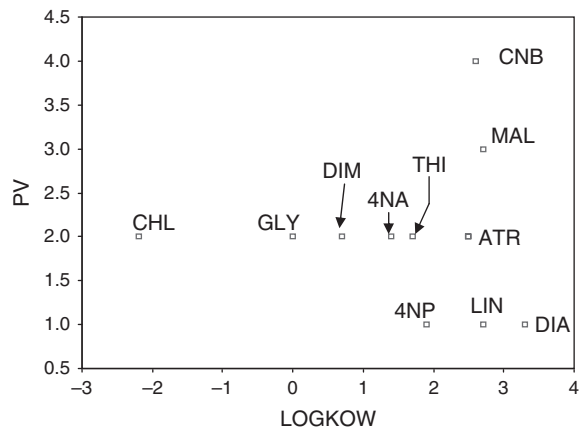


Figure 3 Scatter plot of the substances.

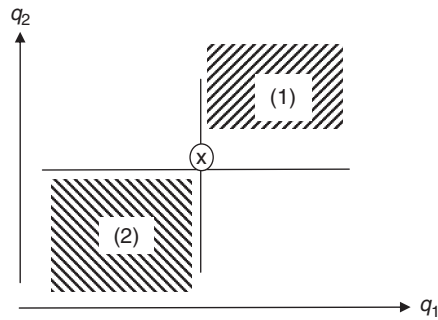


Figure 4 “Shadows of  $x$ ”, High values of  $q_i$  indicate—as before—a high impact on the environment.

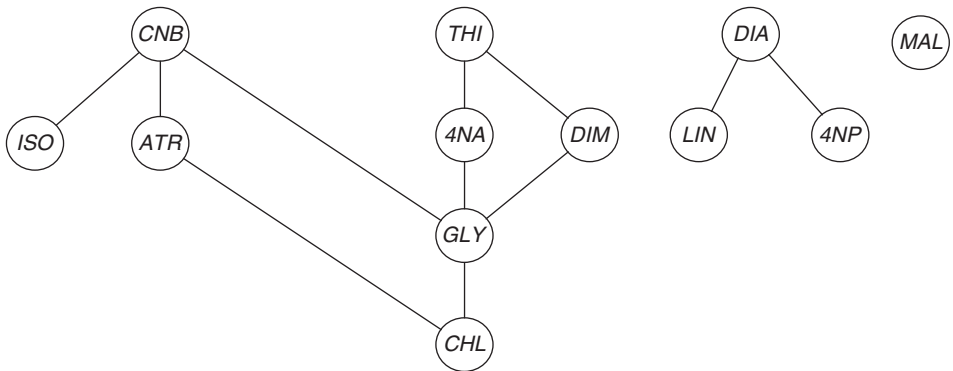


Figure 5 Hasse diagram of  $(G_{\text{HPVC}}, \text{IB}_{\text{Exposure}} \cup \{\text{negLC}, \text{negBD}\})$ .



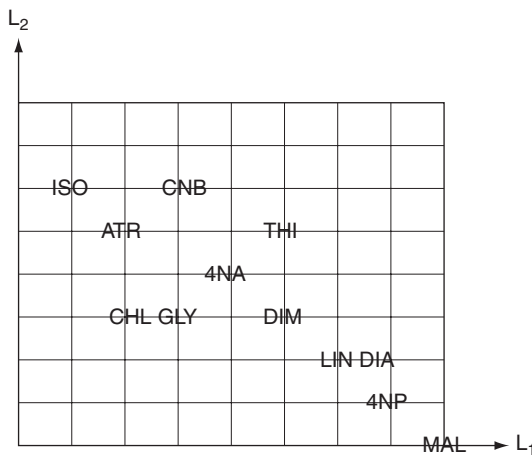
As discussed in step 2 in [Section 2.1](#), we need a common orientation for all attributes. In this case, high values stand for contribution to a high risk either just by the high production volume (PV), or by a high toxicity (LC) or by a distinct accumulation tendency or finally by a high persistency. As the toxicity, for example, is expressed as LC50 value, we need the reverse orientation to let high values represent a high toxicity. The same is valid for biodegradation (BD), the percentage per day degraded.

This diagram has only four levels: Level 1: CHL; Level 2: GLY; Level 3: ISO, ATR, 4NA, DIM, LIN, 4NP} and Level 4: CNB, THI, DIA, MAL. The minimal elements are: ISO, CHL, LIN and 4NP (MAL). The maximal elements are the same as in Level 4. Isolated elements: Only one, namely MAL.

There is a new feature in [Figure 5](#). The directed graph contains three components, i.e. subsets that are not connected. One is speaking of three hierarchies:  
*Hierarchy 1*: CNB, THI, ISO, ATR, 4NA, DIM, GLY, CHL.  
*Hierarchy 2*: DIA, LIN, 4NP  
*Hierarchy 3* (also called a trivial hierarchy because of consisting of an isolated element): MAL.

Certainly this Hasse diagram can be embedded into a space of real numbers of linear space dimension 4, because it is built from four attributes. However, dimension theory (see [Brüggemann and Carlsen, 2006](#)) tells us that this diagram can nevertheless be embedded into a space spanned by two coordinates. There are many possibilities to do this; one model is shown in [Figure 6](#).

A systematic procedure, albeit accepting approximation, is provided by the approach partially ordered scalogram analysis with coordinates (POSAC) ([Shye, 1985](#)), ([Voigt et al., 2004c](#)). Here we make use of the dimension theory, which tells



**Figure 6** Two latent variables  $L_1$ ,  $L_2$ ; with respect to them an embedding is found.

us that the embedding must be possible without any approximations or neglecting of order relations. As the points in the four-dimensional space of the original attributes can be mapped onto a two-dimensional space of two latent variables, one may ask which of the original variables contributes to  $L_1$  or to  $L_2$ . This, however, is a difficult task and still an open problem in HDT. Approximately one may analyze the Spearman correlation or perform a correspondence analysis.

Note that there are still many concepts, which could not be mentioned here, for example, linear extensions, which are a powerful albeit computationally almost intractable concept, to formulate the partial order in the setting of probability theory. As the calculation of linear extensions becomes difficult, if the number of objects is greater than 15, several new concepts were developed:

- The approach due to the set of order ideals, which can be partially ordered and form a lattice (see [De Loof et al., 2006, 2007](#)).
- The approach of MC simulation, which is discussed by Sørensen et al (see e.g. Lerche et al., 2003; Lerche and Sørensen, 2003) and which is similar to the Bubley–Dyers algorithm (see the good explanation in Denoeux et al., 2005).
- The approach of local partial orders, which is developed by the first author and which replaces the partial order of  $n$  objects by  $n$  simplified partial orders, which allows a rather simple analytical approximation (see [Brüggemann et al., 2004, 2005](#)).

Besides the concept of linear extensions, new concepts of antagonism (see Simon et al., 2006b) are useful to explain the occurrence of different hierarchies in a poset.

A recent development assumes that besides the partial order of objects, a classification exists, which is not necessarily derived from the attributes. Hence the partially ordered set of objects is additionally structured by equivalence relations. To perform the analysis, the dominance and separability degree was introduced and analyzed (Restrepo et al., 2007a,b).

## 5. APPLICATIONS OF HASSE DIAGRAM TECHNIQUE TO THE DATA AVAILABILITY OF CHEMICALS

### 5.1 Overview

The topic of ranking chemicals with respect to their data availability or data sources with respect to their content concerning environmental chemicals has been scrutinized for many years. In [Table 2](#), an excerpt of topics, the sizes of the data matrices as well as the references are given.

You can read from the table that plenty of data analysis approaches have been performed during the past 7 years in the field of data availability on chemicals. The size of the data matrices has to be commented. The matrices are all relatively small. For larger data matrices, multivariate statistical methods like cluster analysis have to be combined with the HDT. Taking another aspect from [Table 2](#) “the application of methods”, it can be detected that in the first years, only the HDT

**Table 2** Application of Hasse diagram technique concerning data availability

No.	Topic	Data matrix	Application	Publication
1	Numerical databases evaluated by general and content-specific criteria	$19 \times 9$	HDT	Voigt et al. (2000a)
2	Environmental descriptors found in three meta-databases	$50 \times 3$	MVS	Voigt et al. (2000b)
3	Search engines evaluated by chemical evaluation criteria	$21 \times 15$	HDT, MVS	Voigt et al. (2001)
4	Pesticide databases evaluated by general evaluation criteria	$31 \times 5$	HDT, MVS	Voigt and Welzl (2002a)
5	Drinking water systems in German cities	$16 \times 5$	HDT, MVS	Voigt and Welzl (2002b)
6	Environmental air monitoring systems evaluated by general evaluation criteria	$16 \times 5$	HDT, MVS	Voigt et al. (2002c)
7	Chemical databases evaluated by environmental chemicals	$12 \times 12$	HDT, MVS	Voigt (2003)
8	Databases evaluated by environmental parameters and vice versa	$12 \times 15$ ; $15 \times 12$	HDT, MVS	Voigt and Welzl (2004a)
9	Chemical databases evaluated by parameters and chemicals	$12 \times 27$ ( $12 \times 3$ )	HDT, METEOR	Voigt et al. (2004b)
10	European environmental air pollutant systems evaluated by evaluation criteria	$15 \times 5$	HDT, MVS, POSAC	Voigt et al. (2004c)
11	Evaluation of pharmaceuticals in publications	$12 \times 75$	HDT, METEOR	Voigt and Brüggemann (2005a)

12	Evaluation of publications including pharmaceuticals in the environment by environmental media	$75 \times 7$	HDT	Voigt et al., (2005b)
13	Evaluation of publications including pharmaceuticals in the environment by environmental media	$75 \times 7$	HDT, MVS, POSAC	Voigt et al. (2005c)
14	Evaluation of databases by chemicals and verse visa (environmental chemicals and pharmaceuticals)	$15 \times 24$ ( $24 \times 15$ )	HDT	Voigt et al. (2006b)
15	Evaluation of databases by chemicals and verse visa (environmental chemicals and pharmaceuticals)	$15 \times 24$ ( $15 \times 2$ )	HDT, METEOR	Voigt et al. (2006c)
16	Chemical databases evaluated by chemicals and their parameters	$12 \times 27$	HDT, METEOR	Voigt and Brüggenmann (2006b)
17	Evaluation of databases by high production volume chemicals and pharmaceuticals	$15 \times 24$	HDT, METEOR	Voigt et al. (2006a)
18	Evaluation of Internet databases by pharmaceuticals	$22 \times 16$	HDT, METEOR	Voigt et al. (2006b)
19	Databases evaluated by pharmaceuticals	$16 \times 17$	HDT, METEOR, stability fields, crucial weights	Voigt and Brüggenmann (2007)

---

HDT, Hasse diagram technique; MVS, multi-variate statistics; METEOR, Method of Evaluation by Order Theory.

was used, later it was combined with multivariate statistical methods and then the HDT was further developed in the direction of incorporating a weighting process into the method (METEOR). In the recently published study (see also Brüggemann et al., 2008), we introduced the highly innovative ideas of crucial weights and stability fields.

We will shortly exemplify this procedure with an example in the field of data availability of chemicals (see also Voigt and Brüggemann, 2007).

## 5.2 Stability fields and crucial weights using a data matrix of 17 databases and 4 pharmaceuticals

### 5.2.1 Data-matrix $17 \times 4$

The order theoretical software METEOR allows the participation of stakeholders in the evaluation process, and the software provides the stepwise introduction of weights. We take a closer look at four pharmaceuticals, two cytostatic agents, Cyclophosphamide (CYC) and 5-fluorouracil (FLU), and two contrast media, diatrizoate (DIT) and iopromide (IOP). The chosen databases are all well-known freely available Internet databases, which all comprise environmentally relevant chemicals. The chosen Internet databases are: CIV (Chemicals Information System for Consumer-relevant Substances), CEX (ChemExper Catalog of Chemical Suppliers, Physical Characteristics), CHF (Chemfinder), ECO (ECOTOX), ENV (Envirofacts), ESI (ESIS—European Chemical Substances Information System), GES (GESTIS—Dangerous Substances Database), GSB (GSBL Public), HSB (Hazardous Substances Database), IAR (IARC), ICS (International Chemical Safety Cards), INT (INTOX), OEK (Oekopro), OIH (OECD Integrated HPV Database), RXL (RXList The International Drug Index), SIR (SIRI Material Safety Data Sheets), SRC (SRC PhysProp Database).

The databases are evaluated according to the availability of information on the pharmaceutical  $x$ : If information is available, code = 1; if not, code = 0.

### 5.2.2 Superattribute, stability fields, crucial weights

If some indicators (also called attributes or criteria)  $q(i)$  are linearly combined, taking weights as scalars, then any resulting “superattribute”  $\phi(k)$  is calculated as follows:

$$\phi(k) : = \sum_{i=1}^{n(k)} g(i)q(i)$$

together with the normalization:

$$1 = \sum_{i=1}^{n(k)} g(i)$$

$n(k)$  being the number of indicators, actually combined in order to calculate  $\phi(k)$ .

(Note that we write  $q(i)$  instead of  $q_i$  to facilitate the readability of the text.)

Any superattribute has the “freedom” of  $n(k) - 1$  freely varying scalars  $g(i) \in [0,1]$ . We call  $[0,1]^{n(k)-1}$  the  $g$ -space of the  $k$ th superattribute. Therefore, we associate to any superattribute a space of weights with the dimension  $n(k) - 1$ , and any aggregation step in METEOR is accompanied by the product of all  $g$ -spaces ( $k = 1, \dots, m$ ), which we call the  $G$ -space. In general,  $n(k)$  may vary and may depend on the intuition of the researcher, applying METEOR. Here, however, we restrict ourselves on aggregation schemes with freedom 1, i.e. we analyze a linear space in the subsequent parts of the paper for any superattribute. If we combine, for example, four indicators pairwise to two superattributes, the two linear  $g$ -spaces are combined, forming a two-dimensional  $G$ -space  $[0,1]^* [0,1]$ . As we will see later in the text, the restriction to freedoms = 1 considerably simplifies the procedure and we call a procedure based on a purely pairwise combination of attributes the “orthogonal-METEOR” (o-METEOR).

Imagine that four indicators are aggregated pairwise as follows:

$$\phi(1) = g(1)q(1) + (1 - g(1))q(2) \quad (1a)$$

$$\phi(2) = g(2)q(3) + (1 - g(2))q(4) \quad (1b)$$

Assume that the database  $x$  is incomparable with database  $y$  due to:

$$q(1, x) > q(1, y) \text{ and } q(2, x) < q(2, y)$$

For this case, we write:  $x \parallel_{(q1, q2)} y$ .

If  $x \parallel_{(q1, q2)} y$ , then the result of aggregation (1a) depends on the one weight  $g(1)$ , whether or not  $\phi(x) > \phi(y)$ .

Obviously the equation

$$\phi(1, x) = \phi(1, y) \quad (2)$$

determines that  $g(1)$  value where the character of order relation between  $x$  and  $y$  changes.

For further details of the concept of crucial weights we refer to a paper by Brüggenmann et al. (2008). Stability fields are subspaces of the  $G$ -space where a change of weights does not change the relative positions of any two incomparable objects.

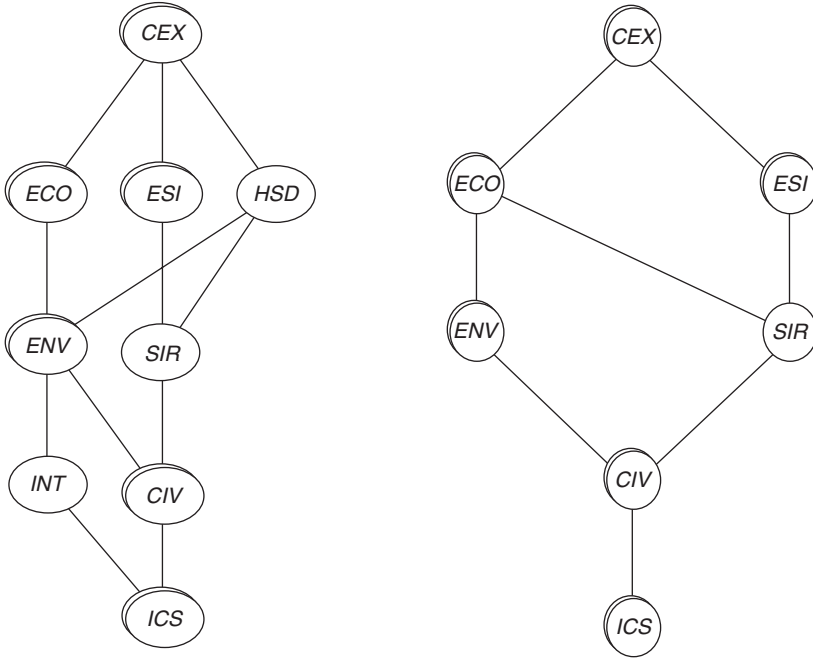
### 5.2.3 Example: 17 databases $\times$ 4 pharmaceuticals

The Hasse diagram for the  $17 \times 4$  data matrix is shown in [Figure 7](#), lhs.

The four pharmaceuticals named above are now aggregated to the following two groups, each comprising two pharmaceuticals

$$\text{CYC} + \text{FLU} = \text{CYTO}, \text{DIT} + \text{IOP} = \text{CONT}$$

The corresponding Hasse diagram of these two superattributes is shown in [Figure 7](#), rhs. It is demonstrated that the number of incomparabilities is



**Figure 7** 17 Databases  $\times$  4 Pharmaceuticals (lhs), 17  $\times$  2 Aggregated Groups CYTO/CONT Hasse diagram (rhs).

considerably reduced by the aggregation step from four pharmaceuticals to two groups of drugs. The maximal objects and minimal objects in both diagrams remain more or less the same.

Following the aggregation strategy described above, the Hasse diagrams of  $\phi(1)$  and  $\phi(2)$  are set up (Figure 8).

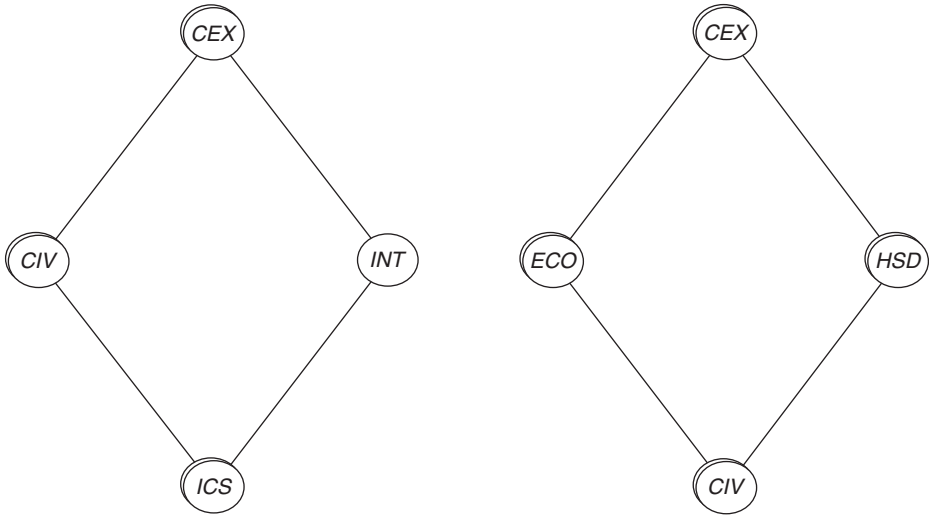
$$\phi(1) = g(1) * \text{CYC} + (1 - g(1)) * \text{FLU}$$

$$\phi(2) = g(2) * \text{DIT} + (1 - g(2)) * \text{IOP}$$

$\phi(1)$  stands for cytostatic agents, that is to say, CYC and FLU,  $\phi(2)$  stands for contrast media, that is to say, DIT and IOP.

In the next step, we identify the incomparable pairs of objects in the CYC/FLU diagram as well as in the DIT/IOP diagram (see Figure 8).

This means that we receive four stability fields. The explanation on the calculation of the stability fields and crucial weights is found in Brüggenmann et al. (2008). To assign the resulting order in any of the four stability fields, one has to compare the  $\phi(1)$  with the  $\phi(2)$  diagram (see Figure 8). In both diagrams (quotient sets), there is only one incomparability  $U$ . By aggregation, either  $\text{CIV} > \text{INT}$  or  $\text{CIV} < \text{INT}$  from the  $\phi(1)$  diagram, as both databases are equivalent in the  $\phi(2)$



**Figure 8** Hasse diagrams for  $17 \times 2$ :  $\phi(1)$  CYC, FLU (lhs),  $\phi(2)$  DIT, IOP (rhs).

Lhs: equivalent objects: {CIV;ESI;GES;OEK;GSB;SIR;SRC}

{CEX;ECO;ENV;HSD;IAR;CHF;RXL} {ICS;OIH}

Rhs: {CIV;ENV;GES;ICS;OEK;OIH;GSB;IAR;INT}

{CEX;ESI;SRC;RXL} {ECO;CHF} {HSD;SIR}

$U_{CYC,FLU} = \{(CIV;ESI;GES;OEK;GSB;SIR;SRC), (INT)\}$

$U_{DIT,IOP} = \{(ECO;CHF), (HSD;SIR)\}$

diagram, the two orientations do not lead to an incomparability in the resulting order. The same applies to the incomparable databases ECO and HSD in the  $\phi(2)$  diagram, whereas there are equivalent objects in the  $\phi(1)$  diagram. However, as, for example,  $SIR < ENV$  in the lhs, whereas  $SIR > ENV$  in the rhs of Figure 8, not necessarily linear orders will appear in all four stability fields.

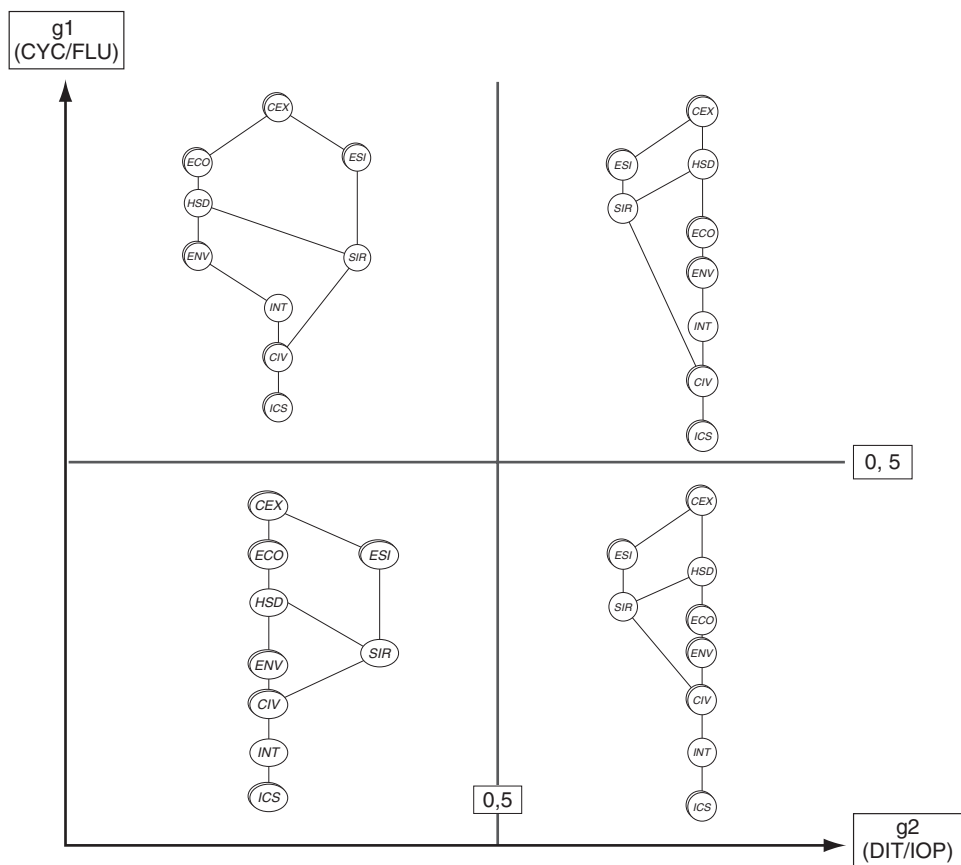
The crucial weights are calculated by one of the sub-modules of PyHasse (see above), which actually contains 22 sub-modules and is as test version available from the first author. For the programming language PYTHON, see Weigend (2006) or Lutz and Ascher (2003).

The crucial weights for  $g(1)$  is 0.5 and for  $g(2)$  is also 0.5. We therefore calculate the Hasse diagrams for the stability fields, which are given in Figure 9.

It can be demonstrated that all diagrams in the four stability fields are different from each other. The maximal and minimal objects, however, are the same in all four Hasse diagrams. The equivalent objects are not listed for reasons of visibility.

All conducted approaches show that the data situation on the chosen test set 17 publicly available in Internet databases concerning their data availability on four well-known and highly produced pharmaceuticals is far from being satisfactory. Only the large databases CEX and RXL comprise all of the chosen pharmaceuticals. It has to be mentioned, however, that neither of these databases contains data on ecotoxicity and/or degradation and accumulation. The issue of





**Figure 9** Stability fields and their different Hasse diagrams.

pharmaceuticals in the environment and the unavailability of data necessitate more research and of course closer communication between science and medical health care and politicians in the future.

## 6. SUMMARY, OUTLOOK AND CONCLUSION

Partial order as a binary mathematical relation is so simple, i.e. so little specialized that it can be widely applied. As in this book, QSAR will play a rather great role; here we mention only the approaches of inverse QSAR to show that partial order in chemistry must not necessarily be restricted to evaluation (Brüggemann et al., 2001b, Carlsen et al., 2002; Carlsen, 2004). The graphical display by Hasse diagrams is very attractive as long as the number of objects is not too high. The resulting digraph allows many insights, which cannot be easily derived if other graphical methods are applied. Here we concentrated on the explanation of

Hasse diagrams with the examples coming from the evaluation of chemicals and of databases.

What will be the development in the future?

One of the most urgent deficits is the still missed field of statistical tests. How sure can we be if we obtain a typical partial-order result, like the element  $x$  is a maximal element or the attribute  $q(i)$  is most sensitive, due to the  $W$ -matrix concept (Brüggemann et al., 2001a)?

Another direction is the further development of the concept of stability fields, where still many questions (e.g. how to relate different stability fields if the  $G$ -space is more than two-dimensional) are open and must quickly be solved, as the application may have direct consequences for better substitutes of refrigerants (Restrepo et al., 2007; Weckert et al., 2007). We conclude that partial order is a general applicable tool just because of its conceptual simplicity, and it seems that the concepts derived from partial order fit very well to chemical problems, as typical questions in chemistry are answered in the form of series, nephelauxetic series, soft and hardness series, series of electronegativity, etc. Here it seems as if the numerical value is not as important as just the position of a chemical entity within a series.

## REFERENCES

- Basak, S.C., Balaban, A.T., Grunwald, G.D., Gute, B.D. (2000). Topological indices: Their nature and mutual relatedness, *J. Chem. Inf. Comput. Sci.* 40, 891–898.
- Basak, S.C., Mills, D.R. (2001). Use of mathematical structural invariants in the development of QSPR models, *MATCH Commun. Math. Comput. Chem.* 44, 15–30.
- Birkhoff, G. (1984). *Lattice theory*. American Mathematical Society, Vol: XXV, Providence, Rhode Island (USA), pp. 1–418.
- Brüggemann, R., Bücherl, C., Pudenz, S., Steinberg, C. (1999a). Application of the concept of partial order on comparative evaluation of environmental chemicals, *Acta hydroch. hydrobiol.* 27, 170–178.
- Brüggemann, R., Carlsen, L. (2006). *Partial Order in Environmental Sciences and Chemistry*, Springer-Verlag, Berlin (GE), pp. 1–406.
- Brüggemann, R., Halfon, E. (1995). *Theoretical base of the program "Hasse"*, GSF-Bericht 20/95, GSF-National Research Centre for Environment and Health, Neuherberg (GE), pp. 1–66.
- Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., Steinberg, C. (2001a). 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.
- Brüggemann, R., Münzer, B., Halfon, E. (1994). An algebraic/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., Pudenz, S., Carlsen, L., Sørensen, P.B., Thomsen, M., Mishra, R.K. (2001b). The use of Hasse diagrams as a potential approach for inverse QSAR. SAR and QSAR, *Environ. Res.* 11, 473–487.
- Brüggemann, R., Pudenz, S., Voigt, K., Kaune, A., Kreimes, K. (1999b). An algebraic/graphical tool to compare ecosystems with respect to their pollution. IV: Comparative regional analysis by Boolean arithmetics, *Chemosphere* 38, 2263–2279.
- Brüggemann, R., Simon, U., Mey, S. (2005). Estimation of averaged ranks by extended local partial order models, *MATCH Commun. Math. Comput. Chem.* 54, 489–518.
- Brüggemann, R., Sørensen, P.B., Lerche, D., Carlsen, L. (2004). Estimation of averaged ranks by a local partial order model, *J. Chem. Inf. Comput. Sci.* 44, 618–625.

- Brüggemann, R., Voigt, K. (1996). Stability of comparative evaluation, – example: Environmental databases, *Chemosphere* 33, 1997–2006.
- Brüggemann, R., Voigt, K., Restrepo, G., Simon, U. (2008). The concept of stability fields and hot spots in ranking of environmental chemicals. *Env. Mod. and Software* 23, 1000–1012.
- Brüggemann, R., Voigt, K., Sørensen, P., De Baets, B. (2007). Partial order concepts in ranking environmental chemicals. In: *Environmental Informatics and Systems Research* Vol. 2 (Hryniewicz, O., Studzinski, J., Szediw, A., eds.), Workshop and Application papers, Shaker Verlag, Aachen, 169–175.
- Carlsen, L. (2004). Giving molecules an identity. On the interplay between QSARs and partial order ranking, *Molecules* 9, 1010–1018.
- Carlsen, L., Sørensen, P.B., Thomsen, M., Brüggemann, R. (2002). QSAR's based on partial order ranking, *SAR QSAR Environ. Res.* 13, 153–165.
- De Baets, B., De Meyer, H., Naessens, H. (2002). On rational cardinality – based inclusion measures, *Fuzzy Set Syst.* 128, 169–183.
- De Loof, K., De Baets, B., De Meyer, H. (2007). On the random generation and counting of weak order extensions of a poset with given class cardinalities, *Inf. Sci.* 177, 220–230.
- De Loof, K., De Meyer, H., De Baets, B. (2006). Exploiting the lattice of ideals representation of a poset, *Fundam. Informaticae* 71, 309–321.
- Denoëux, T., Masson, M.-H., Hébert, P.-A. (2005). Nonparametric rank-based statistics and significance tests for fuzzy data, *Fuzzy set. Syst.* 153, 1–28.
- Glass, L. (1975). Combinatorial and topological methods in nonlinear chemical kinetics, *J. Chem. Phys.* 63, 1325–1335.
- Hasse, H. (1967). *Vorlesungen über Klassenkörpertheorie*, Physica-Verlag, Germany, p. 275.
- Halfon, E. (1979). Computer-based development of large-scale ecological models: Problems and prospects. In: *System Analysis of Ecosystems* (Innis, H.S., O'Neill, R.V. eds.), International Co-operative Publishing House, Burtonsville, Maryland, USA, pp. 197–209.
- Halfon, E. (1983). Is there a best model structure? I: Modelling the fate of a toxic substance in a lake, *Ecol. Model.* 20, 135–152.
- Halfon, E. (2006). Hasse diagrams and software development. In: *Partial Order in Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L. eds.), Springer-Verlag, Berlin (GE), pp. 385–392.
- Halfon, E., Hodson, J., Miles, K. (1989). An algorithm to plot Hasse diagrams on microcomputers and calcomp plotters, *Ecol. Modell.* 47, 189–197.
- Halfon, E., Reggiani, M.G. (1986). On ranking chemicals for environmental hazard, *Environ. Sci. Technol.* 20, 1173–1179.
- Helm, D. (2003). Bewertung von Monitoringdaten der Umweltprobenbank des Bundes mit der Hasse-Diagramm-Technik, *UWSF – Z. Umweltchem. Ökotox.* 15, 85–94.
- Helm, D. (2006). Evaluation of biomonitoring data. In: *Partial Order in Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L. eds.), Springer-Verlag, Berlin (GE), pp. 285–307.
- Lerche, D., Brüggemann, R., Sørensen, P.B., Carlsen, L., Nielsen, O.J. (2002). A comparison of partial order technique with three methods of multicriteria analysis for ranking of chemical substances, *J. Chem. Inf. Comput. Sci.* 42, 1086–1098.
- Lerche, D., Sørensen, P.B. (2003). Evaluation of the ranking probabilities for partial orders based on random linear extensions, *Chemosphere* 53, 981–992.
- Lerche, D., Sørensen, P.B., Brüggemann, R. (2003). Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual probability, *J. Chem. Inf. Comput. Sci.* 53, 1471–1480.
- Luther, B., Brüggemann, R., Pudenz, S. (2000). An approach to combine cluster analysis with order theoretical tools in problems of environmental pollution, *MATCH Commun. Math. Comput. Chem.* 42, 119–143.
- Lutz, M., Ascher, D. (2003). *Learning Python*, O'Reilly Publisher, Beijing (CH). pp. 1–591.
- Münzer, B., Brüggemann, R., Halfon, E. (1994). An algebraic/graphical tool to compare ecosystems with respect to their Pollution II: Comparative regional analysis, *Chemosphere* 28, 873–879.
- Naessens, H., De Meyer, H., De Baets, B. (2002). Algorithms for the computation of T-Transitive closures, *IEEE Trans. Fuzzy Syst.* 10, 541–551.
- Nemes, I., Vidoczy, T., Gal, D. (1977). A possible construction of chemical reaction networks, *Theor. Chim. Acta* 46, 243–250.

- Pavan, M. (2003). Total and Partial Ranking Methods in Chemical Sciences. Thesis at the University of Milan – Bicocca (IT), Cycle XVI, Tutor: Prof. Todeschini, pp. 1–277.
- Pavan, M., Mauri, A., Todeschini, R. (2004). Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority setting, *Anal. Bioanal. Chem.* 350, 430–444.
- Pudenz, S., Brüggemann, R., Luther, B., Kaune, A., Kreimes, K. (2000). An algebraic/graphical tool to compare ecosystems with respect to their pollution V: Cluster analysis and Hasse diagrams, *Chemosphere* 40, 1373–1382.
- Restrepo, G., Brüggemann, R., Voigt, K. (2007a). Partially ordered sets in the analysis of Alkanes Fate in rivers, *Croatia Chemica Acta* 80, 261–270.
- Restrepo, G., Weckert, M., Brüggemann, R., Gerstmann, S., Frank, H. (2007b). Refrigerants ranked by partial order theory. In: *Environmental Informatics and Systems Research* Vol. 2. (Hryniewicz, O., Studzinski, J., Szediw, A., eds.), Workshop and Application papers, Shaker Verlag, Aachen, 209–217.
- Rival, I. (1985). The diagram. In: *Graphs and Order* (Rival, I. ed.), D. Reidel Publishing Company, Dordrecht (NL), 103–133.
- Sabljić, A., Trinajstić, N. (1981). Quantitative structure-activity relationships: The role of topological indices, *Acta Pharm. Jugosl.* 31, 189–214.
- Shye, S. (1985). *Multiple Scaling*, Elsevier Publishers, Amsterdam (NL).
- Simon, U., Brüggemann, R. (2000). Assessment of water management strategies by Hasse diagram technique. In: *Order Theoretical Tools in Environmental Sciences* (Sørensen, P.B., Carlsen, L., Mogensen, B.B., Brüggemann, R., Luther, B., Pudenz, S., Simon, U., Halfon, E., Bittner, T., Voigt, K., Welzl, G., Rediske, F., eds.), *Proceedings of the Second Workshop, October 21st, 1999 in Roskilde, Denmark*, National Environmental Research Institute, Roskilde (DK), pp. 117–134.
- Simon, U., Brüggemann, R., Behrendt, H., Shulenberg, E., Pudenz, S. (2006). METEOR: a step-by-step procedure to explore effects of indicator aggregation in multi criteria decision aiding – application to water management in Berlin, Germany, *Acta hydroch. hydrobiol.* 34, 126–136.
- Simon, U., Brüggemann, R., Pudenz, S. (2004a). Aspects of decision support in water management – example Berlin and Potsdam (Germany) I – spatially differentiated evaluation, *Water Res.* 38, 1809–1816.
- Simon, U., Brüggemann, R., Pudenz, S. (2004b). Aspects of decision support in water management – example Berlin and Potsdam (Germany) II – improvement of management strategies, *Water Res.* 38, 4085–4092.
- Simon, U., Brüggemann, R., Pudenz, S., Behrendt, H. (2006b). Aspects of decision support in water management: Data based evaluation compared with expectations. In: *Partial Order in Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L. eds.), Springer-Verlag, Berlin (GE), pp. 221–236.
- 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, *Envir. Tox. Chem.* 22, 661–670.
- Sørensen, P.B., Brüggemann, R., Thomsen, M., Lerche, D. (2005). Applications of multidimensional rank-correlation, *MATCH Commun. Math. Comput. Chem.* 54, 643–670.
- Sørensen, P.B., Lerche, D.B., Carlsen, L., Brüggemann, R. (2001). Statistically approach for estimating the total set of linear orders. In: *Order Theoretical Tools on Environmental Science and Decision Systems* (Pudenz, S., Brüggemann, R., Lühr, H.-P., eds.), *Proceedings of the third Workshop November 6th – 7th, 2000 in Berlin, Germany*; Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, – 222 – Heft 14, Sonderheft IV, pp. 87–97.
- Sørensen, P.B., Mogensen, B.B., Carlsen, L., Thomsen, M. (2000). The influence on partial order ranking from input parameter uncertainty – Definition of a robustness parameter, *Chemosphere* 41, 595–600.
- Todeschini, R., Consonni, V. (2000). *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim (GE).
- Van der Walle, B., De Baets, B., Kersebaum, K.C. (1995). Fuzzy multi-criteria analysis of cutting techniques in a nuclear dismantling project, *Fuzzy set. Syst.* 74, 115–126.
- Voigt, K. (2003). Evaluation of environmental chemicals in databases: An example for the interdisciplinarity between environmental informatics and environmetrics/chemometrics. In: *The*

- Information Society and Enlargement of the European Union, 17th International Conference Informatics for Environmental Protection* (Gnauck, A., Heinrich, R. eds.), Cottbus 2003, Part 2: Applications, Workshops, Posters, Metropolis Verlag, Marburg (GER), 613–620.
- Voigt, K., Brüggemann, R. (2005a). Water contamination with pharmaceuticals: Data availability and evaluation approach with Hasse diagram technique and METEOR, *MATCH Commun. Math. Comput. Chem.* 54, 3, 671–689.
- Voigt, K., Brüggemann, R. (2006a). Information systems and databases. In: *Partial Order in Environmental Sciences and Chemistry*, (Brüggemann, R., Carlsen, L. eds.), Springer-Verlag, Berlin (GER), 327–351.
- Voigt, K., Brüggemann, R. (2006b). Method of evaluation by order theory applied on the environmental topic of data-availability of pharmaceutically active substances. In: *Development of Methods and Technologies of Informatics for Process Modeling and Management* Vol. 50 (Studzinski, J., Hryniewicz, O., eds.), Polish Academy of Sciences, Systems Research Institute, Warsaw (PL), 107–120.
- Voigt, K., Brüggemann, R. (2007). Data-availability of pharmaceuticals detected in water: An evaluation study by order theory (METEOR). In: *Environmental Informatics and Systems Research* Vol. 2 (Hryniewicz, O., Studzinski, J., Szediw, A., eds.), Workshop and Application papers, Shaker Verlag, Aachen, 201–208.
- Voigt, K., Brüggemann, R., Pudenz, S. (2004b). Chemical databases evaluated by order theoretical tools, *Analytical and Bioanalytical Chemistry* 380, 467–474.
- Voigt, K., Brüggemann, R., Pudenz, S. (2005b). Application of computer-aided decision tools concerning environmental pollution with pharmaceuticals. In: *Applications of Informatics in Environmental Engineering and Medicine* (Studzinski, J., Drelchowski, L., Hryniewicz, O., eds.), Polish Academy of Sciences, Systems Research Institute, Series: Systems Research Vol. 42, Warsaw (PL), pp. 135–146.
- Voigt, K., Brüggemann, R., Pudenz, S. (2006b). Information quality of environmental and chemical databases exemplified by high production volume chemicals and pharmaceuticals, *Online Inf. Rev.* 30, 1, 8–23.
- Voigt, K., Brüggemann, R., Pudenz, S. (2006a). A multi-criteria evaluation of environmental databases using the Hasse diagram technique (ProRank) software, *Environ. Modell. Softw* 21, 1587–1597.
- Voigt, K., Brüggemann, R., Pudenz, S., Scherb, H. (2005c). Environmental contamination with endocrine disruptors and pharmaceuticals: An environmental evaluation approach. In: *EnviroInfo Brno 2005, Informatics for Environmental Protection, Networking Environmental Information* (Hrebicek, J., Racek, J. eds.), Part 2, Masaryk University in Brno (CZ), pp. 858–862.
- Voigt, K., Gasteiger, J., Brüggemann, R. (2000a). Comparative evaluation of chemical and environmental online databases, *J. Chem. Inf. Comput. Sci.* 40, 1, 44–49.
- Voigt, K., Pudenz, S., Brüggemann, R. (2006c). ProRank a software tool used for the evaluation of environmental databases. In: *Proceedings of the iEMSS Third Biennial Meeting: "Summit on Environmental Modelling and Software"* (Voinov, A., Jakeman, A., Rizzoli, A., eds.), International Environmental Modelling and Software Society, Burlington, USA, July 2006. CD ROM. Internet: <http://www.iemss.org/iemss2006/sessions/all.html>.
- Voigt, K., Welzl, G. (2002a). Chemical databases: An overview of selected databases and evaluation methods, *Online Inf. Rev.* 26, 3, 172–192.
- Voigt, K., Welzl, G. (2002b). Drinking water analysis systems in German cities: An evaluation approach combining Hasse diagram technique with multivariate statistics. In: *Order Theoretical Tools in Environmental Sciences, Order Theory (Hasse Diagram Technique) Meets Multivariate Statistics* (Voigt, K., Welzl, G. eds.), Shaker-Verlag, Aachen (GE), pp. 113–140.
- Voigt, K., Welzl, G. (2004a). Data availability on existing substances in publicly available databases – A data analysis approach. In: *Order theory in Environmental Sciences, NERI Technical Report No. 479* (Sørensen, P., Brüggemann, R., Lerche, D.B., Voigt, K., eds.), National Environmental Research Institute, Ministry of the Environment, Copenhagen (DK), pp. 52–67.
- Voigt, K., Welzl, G., Brüggemann, R. (2004c). Data-analysis of environmental air pollutant monitoring systems in Europe, *Environmetrics* 15, 577–596.
- Voigt, K., Welzl, G., Brüggemann, R., Pudenz, S. (2002c). Bewertungsansatz von Umweltmonitoring-Systemen in Ballungsräumen, *UWSF- Umweltwissenschaften und Schadstoff Forschung* 14, 1, 58–64.
- Voigt, K., Welzl, G., Glander-Höbel, C., Brüggemann, R. (2001). Hasse diagram technique meets multivariate statistical methods meet search engines. In: *Order Theoretical Tools on Environmental*

- Science and Decision Systems* (Pudenz, S., Brüggemann, R., Lühr, H.-P., eds.), Proceedings of the third Workshop November 6th – 7th, 2000 in Berlin Germany; Leipzig Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, – 222 – Heft 14, Sonderheft IV, pp. 161–178.
- Voigt, K., Welzl, G., Rediske, G. (2000b). Multivariate statistics applied to the evaluation of environmental and chemical data sources, *Online Inf. Rev.* 24, 2, 116–123.
- Vukicevic, D., Sedlar, J., Rajtmajer, S.M. (2007). A graph theoretical method for partial ordering of Alkanes, *Croatia Chemica Acta* 80, 169–179.
- Weckert, M., Restrepo, G., Gerstmann, S., Frank, H. (2007). Hasse diagram technique – a useful tool for life cycle assessments of refrigerants. In: *Environmental Informatics and Systems Research* Vol. 2 (Hryniewicz, O., Studzinski, J., Szediw, A., eds.), Workshop and Application papers, Shaker Verlag, Aachen, 219–226
- Weigend, M. (2006). *Objektorientierte Programmierung mit Python*, Mitp-Verlag, Bonn (GE), pp. 1–700.
- Wilf, H.S. (2004). The Redheffer matrix of a partially ordered set, *Electron. J. Comb.* 11, 1–5.
- Zeigarnik, A.V., Temkin, O.N., Bonchev, D. (1996). Application of graph theory to chemical kinetics. 3. Topological specificity of multiroute reaction mechanisms, *J. Chem. Inf. Comp. Sci.* 36, 973–981.

# CHAPTER 4

## Partial Ordering and Prioritising Polluted Sites

L. Carlsen

---

Contents	1. Introduction	97
	2. Methodology	98
	2.1 Partial-order ranking	98
	2.2 Linear extensions	99
	2.3 Meta-descriptors	99
	2.4 Hierarchical partial-order ranking	100
	3. Applications	101
	3.1 Generation of meta-descriptors	104
	3.2 Hierarchical partial-order ranking	104
	4. Conclusions	108
	Acknowledgments	108
	References	108

---

### 1. INTRODUCTION

The prioritising of pollutant sites for confinement or possibly remediation or clean up may well be a rather complicated task as a wide range of parameters should be taken into account including both scientifically based factor relating to the environment and the actual location and socio-economic factors. Both of these groups are obviously sub-divided into various rather different topics. Thus, the environmental factors to be taken into account may be the impact of the polluted site on the surrounding areas due to various dimensions, e.g., atmospheric dispersion of emitted substances, pollution of ground water or surface water compartments as a result of either leaching or simple run-off and distance to drinking water resources. The socio-economic factors include direct cost of cleaning up and a cost-benefit analysis of doing something or doing nothing. In practice, this means that we have to take a wide range of factors into account that obviously a priori are incommensurable.

The possible assessment consequently turns into a multicriteria evaluation scheme. However, the number of parameters that should initially be included compared to the number of sites included in the study may well appear to be prohibitive for developing a robust model (Sørensen et al., 2000). To reduce the number of parameters to be taken into account, primary analyses of specific dimensions leading to latent variables, meta-descriptors, are conducted, the idea being similar to that of the well-established of principal component analyses. Thus, a primary assessment of the objects under investigation based on connected parameters will lead to a set of meta-parameters that can be subsequently used for an analysis of the meta-dimension for the final assessment. The studies take its onset in the partial-order theory. Hence, the new methodology, hierarchical partial-order ranking (HPOR) (Carlsen, 2008), takes into account a range of otherwise incomparable parameters disclosing those polluted sites that on a cumulative basis appear to constitute the major risk towards both human and environmental health and thus potentially being those sites that a priori should be subject to appropriate confinement, remediation or cleanup.

## 2. METHODOLOGY

The successful use of HPOR (Carlsen, 2008) obviously depends on the quality of the single techniques. In the following sections, partial-order ranking (POR) including linear extensions (LE) and average rank as well as QSARs, as applied for the studies included in the present paper, will be shortly presented.

### 2.1 Partial-order ranking

The theory of POR is presented elsewhere (Davey and Priestley, 1990; Brüggemann and Carlsen, 2006). In brief, POR is a simple principle, which a priori includes “ $\leq$ ” as the only mathematical relation. If a system is considered, which can be described by a series of descriptors  $p_i$ , a given site A, characterized by the descriptors  $p_i(A)$ , can be compared to another site B, characterized by the descriptors  $p_i(B)$ , through comparison of the single descriptors. Thus, site A will be ranked higher than site B, i.e.,  $B \leq A$ , if at least one descriptor for A is higher than the corresponding descriptor for B and no descriptor for A is lower than the corresponding descriptor for B. If, in contrast,  $p_i(A) > p_i(B)$  for descriptor  $i$  and  $p_j(A) < p_j(B)$  for descriptor  $j$ , A and B will be denoted incomparable. Obviously, if all descriptors for A are equal to the corresponding descriptors for B, i.e.,  $p_i(B) = p_i(A)$  for all  $i$ , the two sites will have identical rank and will be considered as equivalent, i.e.,  $A = B$ . In mathematical terms, this can be expressed as

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



It further follows that if  $A \geq B$  and  $B \geq C$ , then  $A \geq C$ . If no rank can be established between  $A$  and  $B$ , then these sites are denoted as incomparable, i.e., they cannot be assigned a mutual order. Therefore, POR is an ideal tool to handle incommensurable attributes.

In POR—in contrast to standard multidimensional statistical analysis—neither any assumptions about linearity nor any assumptions about distribution properties are made. In this way the POR can be considered as a non-parametric method. Thus, there is no preference among the descriptors. However, due to the simple mathematics outlined above, it must be emphasized that the method a priori is rather sensitive to noise, since even minor fluctuations in the descriptor values may lead to non-comparability or reversed ordering.

A main point is that all descriptors have the same direction, i.e., “high” and “low”. As a consequence of this, it may be necessary to multiply some descriptors by  $-1$  in order to achieve identical directions. Bioaccumulation and toxicity can be mentioned as an example. In the case of bioaccumulation, the higher the number, the higher a chemical substance tends to bioaccumulate and thus the more problematic the substance, whereas in the case of toxicity, the lower the figure, the more toxic the substance. Thus, in order to secure identical directions of the two descriptors, one of them, e.g., the toxicity figures, has to be multiplied by  $-1$ . Consequently, in the case of both bioaccumulation and toxicity, higher figures will correspond to more problematic sites.

The graphical representation of the partial ordering is often given in the so-called Hasse diagram ([Hasse, 1952](#); [Halfon and Reggiani, 1986](#); [Brüggemann et al., 1995, 2001](#)). In practice, the PORs are performed using the WHasse software ([Brüggemann et al., 1995](#)).

## 2.2 Linear extensions

The number of incomparable elements in the partial ordering may obviously constitute a limitation on the attempt to rank, e.g., a series of chemical substances based on their potential environmental or human health hazard. To a certain extent, this problem can be remedied through the application of the so-called linear extensions of the POR ([Fishburn, 1974](#); [Graham, 1982](#)). A linear extension is a total order where all comparabilities of the partial order are reproduced ([Davey and Priestley, 1990](#); [Brüggemann et al., 2001](#)). Due to the incomparisons in the POR, a number of possible linear extensions correspond to one partial order. If all possible linear extensions are found, a ranking probability can be calculated, i.e., based on the linear extensions, the probability that a certain element has a certain absolute rank can be derived. If all possible linear extensions are found, it is possible to calculate the average rank of the single elements in a partially ordered set ([Winkler, 1982, 1983](#)).

## 2.3 Meta-descriptors

On the basis of the linear extensions, the average rank of the single elements can be established. The average rank is simply the average of the ranks in all

the linear extensions. On this basis, the most probable rank for each element can be obtained leading to the most probably linear rank of the elements studied.

The average ranks are subsequently used as meta-descriptors. Thus, if a selection of primary descriptors, like the annual average atmospheric concentrations, the maximum 1-h atmospheric concentrations and the maximum 8-h running mean atmospheric concentrations, is used to rank a given site according to atmospheric dispersion of substances emitted from the polluted site, the average ranks of the single sites constitute a meta-descriptor reflection, the impact according to atmospheric dispersion. Analogously meta-descriptors reflecting, e.g., the impact on the aquatic and terrestrial compartments and the human health, as well as groups of socio-economic factors, may be derived.

The average rank, and thus the meta-descriptors of the single element in the Hasse diagram, can be obtained by deriving a large number of randomly generated linear extensions (Sørensen et al., 2001; Lerche et al., 2002, 2003). Alternatively the generation of the average rank of the single sites in the Hasse diagram is obtained by applying the simple relation recently reported by Brüggemann et al. (2004). The average rank of a specific element,  $c_i$ , can be obtained by the simple relation

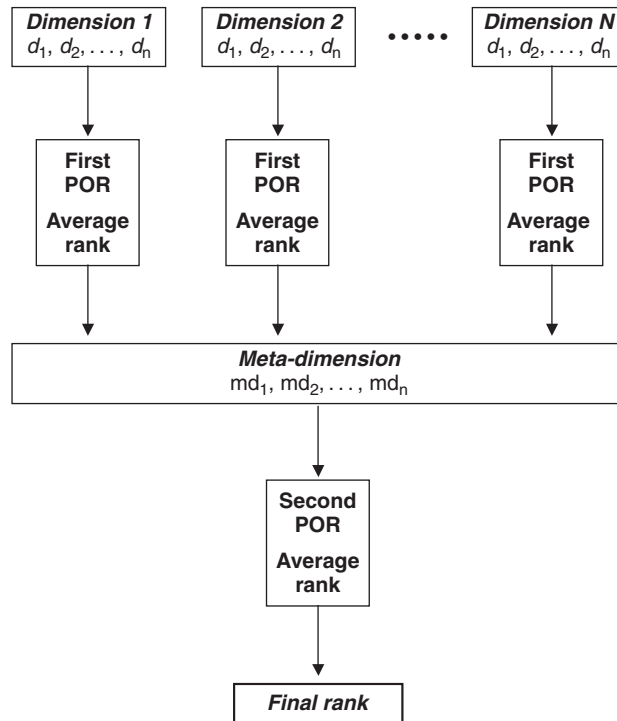
$$Rk_{av} = (N + 1) - (S + 1) \times (N + 1) / (N + 1 - U) \quad (2)$$

where  $N$  is the number of elements in the diagram,  $S$  the number of successors, i.e., comparable element located below, to  $c_i$  and  $U$  the number of elements being incomparable to  $c_i$  (Brüggemann et al., 2004). It should be noted that in the ranking, according to Eq. (2), the lower the number, the higher the levels. Thus, the highest level will be "1". This is reversed compared to the original approach (Brüggemann et al., 2004).

## 2.4 Hierarchical partial-order ranking

In the second stage, the above-mentioned meta-descriptors describe the meta-dimension and are subsequently used as descriptors in a consecutive POR. Thus, the number of descriptors is significantly reduced and the ranking leads to the possible development of a robust model (Sørensen et al., 2000), which in principle will contain all information based on the original set of descriptors.

Since the meta-descriptors, as the descriptors, are ordered with the highest rank being denoted "1", all the meta-descriptors must be multiplied by  $-1$  in order to make sure that the elements with the highest rank, i.e., with the lowest attributed number, will be ranked in the top of the Hasse diagram as a result of the ranking based on the meta-descriptors. In Figure 1, a graphical representation of the HPOR approach is depicted.



**Figure 1** Graphical representation of the hierarchical partial-order ranking (HPOR).

### 3. APPLICATIONS

In different societies, polluted sites cause significant problems, the actual hazard being associated with the type of pollution, the site location, etc. Thus, in Denmark, in approx. 32,000 polluted sites registered, pollution originating from oil/gasoline and from cleaning industries constitute the major part (MST, 2006). A totally different type of pollution is, for example, seen in the area of the Baikonur Cosmodrome, where a significant number of sites are heavily contaminated by 1,1-dimethyl hydrazine and its transformation products, the pollution originating from unburned rocket fuel dispersed after the fall of the burned-out rocket stages (Carlsen et al., 2007; Carlsen et al., 2008).

In both cases, the number of sites by far exceed the number that may be treated simultaneously. Thus, a prioritisation is needed to disclose the specific sites that call for immediate confinement or possibly remediation or clean up. Obviously, a prioritisation taking all sites into account will not take place. Thus, a selection of sites, e.g., within a certain distance from residential areas, would be included in the analysis.

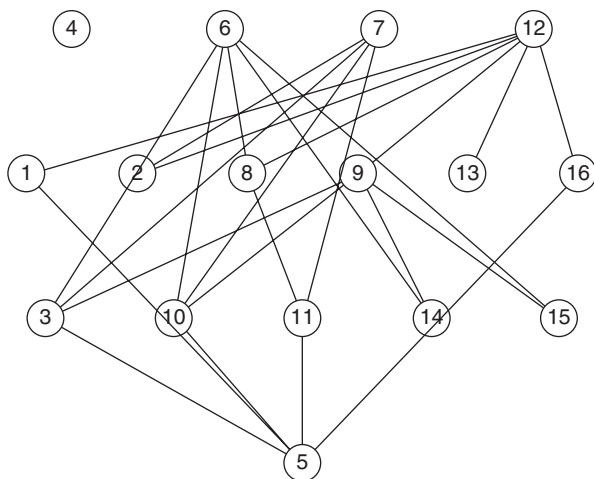
In the present study, we have looked at an arbitrary selection of 16 sites to be prioritized based on the possible impact of three possible dimensions, i.e.,

atmospheric dispersion, possible ground water (drinking water) contamination in combination with socio-economic aspects, on the population.

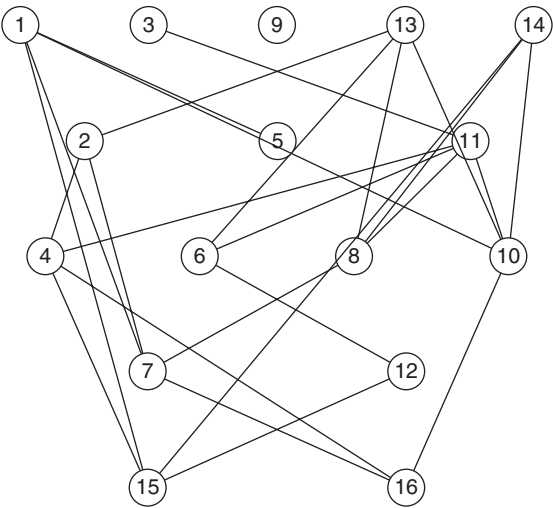
To elucidate the possible impact of atmospheric dispersion of possible volatiles on the population, it is suggested to include yearly average concentrations, maximum hourly concentrations and maximum 8-h running mean concentrations. These values may be calculated applying the OML-Multi model (DMU, 2006). The expected impact from ground water (drinking water) may be elucidated through the migration potential of the substance under investigation as determined by water solubility of the compound, annual precipitation and ground water flow. The socio-economic factors may include the number of people possibly being subject to atmospheric and/or drinking water exposure, the costs associated with confinement or possibly remediation or clean up and the costs associated with an increasing number of health problems among the affected population.

In the present study, for illustrative purposes, we use randomly generated descriptor values for 16 polluted sites, the resulting POR being visualized in the Hasse diagram in Figures 2–4 for three dimensions, i.e., atmospheric impact, the ground water impact and the socio-economic factors.

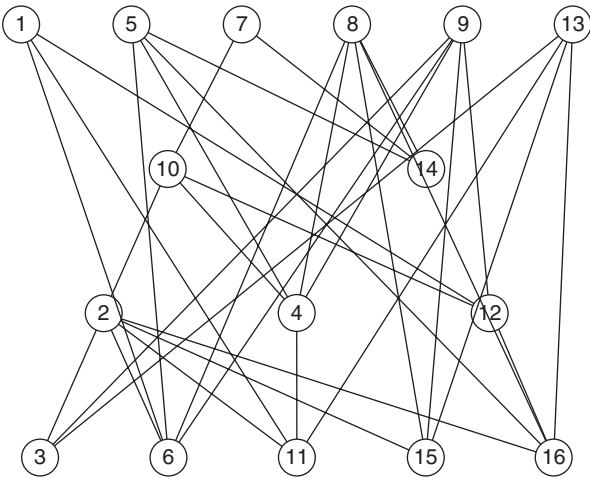
It is immediately seen that in all cases, clear rankings of the 16 sites are achieved. Moreover, it is equivocal that the three rankings are distinctly different as expected. A priori POR gives the opportunity to include all descriptors simultaneously. However, in the present case, the result is a complete anti-chain (Figure 5) reflecting the maximum instability as a result of too many descriptors compared to the number of elements (Sørensen et al., 2000).



**Figure 2** Hasse diagram visualizing the partial-order ranking (POR) based on the impact through the dimension “atmospheric dispersion”.



**Figure 3** Hasse diagram visualizing the partial-order ranking (POR) based on the impact through the dimension “drinking water”.



**Figure 4** Hasse diagram visualizing the partial-order ranking (POR) based on the dimension “socio-economic considerations”.



**Figure 5** Hasse diagram visualizing the partial-order ranking (POR) based on the simultaneous inclusion of impacts through atmospheric dispersion, drinking water dispersion and socio-economic factors.

### 3.1 Generation of meta-descriptors

On the basis of POR visualized in [Figures 1–3](#), the corresponding meta-descriptors in the form of average ranks of the single sites are generated. This may be achieved either through an empirical estimation of the average rank according to [Eq. \(2\)](#) ([Brüggemann et al., 2004](#)) or based on randomly generated linear extensions ([Sørensen et al., 2001](#); [Lerche et al., 2002, 2003](#)). The resulting meta-descriptors, i.e., the estimated average ranks of the 16 sites, are given in [Tables 1 and 2](#).

It can be noted immediately that although some discrepancies apparently prevail between the two methods, which is not surprising ([Brüggemann et al., 2004](#); [Carlsen, 2006a](#)), a nice overall agreement between the two in principle different methodologies prevails.

### 3.2 Hierarchical partial-order ranking

On the basis of the above given meta-descriptors, it is now possible to pursue the overall ranking taking all three dimensions into account, the single dimensions being represented by the corresponding meta-descriptors. Thus, we have a meta-dimension with 16 elements, the polluted sites, and three meta-descriptors, representing three primary dimensions, i.e., atmospheric and ground water impacts and socio-economic factors, respectively.

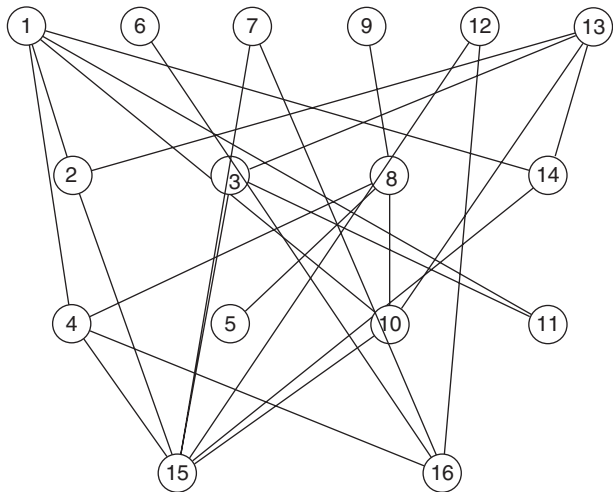
**Table 1** Meta-descriptors as derived based on average linear rank according to [Eq. \(2\)](#)

Site No.	Atmospheric impact	Ground water impact	Socio-economic factors
1	8.5	2.4	2.8
2	13.6	4.9	5.7
3	12.1	1.5	14.6
4	8.5	10.6	12.8
5	15.6	11.3	2.4
6	1.7	9.7	15.1
7	2.4	13.6	1.4
8	8.5	10.6	2.1
9	4.3	8.5	1.9
10	12.1	12.8	3.1
11	12.1	3.1	15.5
12	1.2	12.1	12.1
13	11.3	1.5	2.8
14	13.6	2.4	13.6
15	13.6	15.5	14.9
16	8.5	15.6	15.5

**Table 2** Meta-descriptors as derived based on the average linear rank estimated based on randomly generated linear extensions

Site No.	Atmospheric impact	Ground water impact	Socio-economic factors
1	8.5	4.4	5.3
2	10.7	6.3	7.0
3	10.7	2.2	12.3
4	8.6	10.3	10.8
5	15.5	10.0	4.7
6	2.9	8.2	12.8
7	4.2	12.6	2.1
8	7.6	9.5	4.4
9	4.5	8.5	4.2
10	10.7	11.1	4.2
11	11.5	4.4	14.1
12	1.4	11.3	10.6
13	9.1	2.6	5.5
14	10.8	4.6	11.2
15	10.7	14.6	12.6
16	8.5	15.3	14.1

In [Figure 6](#), the Hasse diagram resulting from the ranking of the 16 sites based on the meta-descriptors estimated by the empirical equation (2) is depicted and the resulting overall linear ranking of the 16 sites, given as average ranks, is given in [Table 3](#), both based on estimation by [Eq. \(2\)](#) and based on randomly generated linear extensions.



**Figure 6** Hasse diagram visualizing the partial-order ranking (POR) of the meta-dimension based on meta-descriptors derived by [Eq. \(2\)](#).

**Table 3** Eventual linear ranking according to Eq. (2),  $Rk_{av}$ , and based on linear extension,  $RLE$ , estimated based on meta-descriptors as derived based on average linear rank according to Eq. (2)<sup>a</sup>

Site No.	$Rk_{av}$	$RLE$
1	1.9	3.2
2	10.2	9.7
3	6.8	7.4
4	9.7	9.9
5	12.8	10.7
6	5.7	7.5
7	4.3	7.1
8	4.3	5.4
9	2.1	2.9
10	12.1	10.9
11	13.6	11.8
12	4.3	7.1
13	2.1	2.9
14	10.2	9.7
15	15.7	15.4
16	15.1	14.3

<sup>a</sup> cf. Table 1.

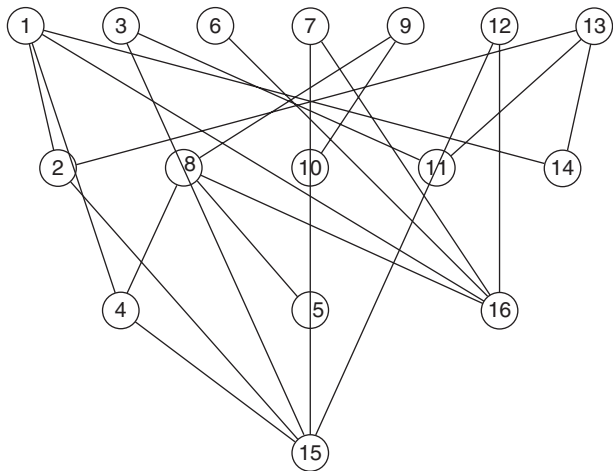
It is seen (Figure 6) that sites 1, 6, 7, 9, 12 and 13 are apparently the top candidates for confinement or possibly remediation or clean up. Since three of these six sites, 1, 9 and 13 (cf. Table 3), apparently have the highest number of successors (cf. Section 2.3), the focus should be on these sites, which according to Eq. (2) will be expected to exhibit a high average rank.

Analogously, Figure 7 displays the Hasse diagram generated based on the meta-descriptors derived from randomly generated linear extensions. Table 4 gives the resulting overall ranking of the 16 sites. Although some differences can be noted immediately by simply looking at the shape of the Hasse diagram (Figure 7), the overall picture remains virtually identical. Thus, here also we find that the top three candidates for confinement or possibly remediation or clean up are the sites 1, 9 and 13.

It should be noted that even an element as site 4 that based to the primary ordering according to the atmospheric impact appears as an isolated element, in the final ranking obtain a definite position connected to the sites 1, 8 and 15.

From the above results it appears clear that the prioritisation of polluted sites can be advantageously done by applying HPOR. Obviously, the initial generation of meta-descriptors by estimating average ranks a priori constitutes an element of uncertainty as the average rank, as determined according to the empirical equation (2) or by using randomly generated linear extensions, is only an approximation to the true values that should be optimally generated based on the total set of linear extensions, which, however, for all practical purposes is not possible.





**Figure 7** Hasse diagram visualizing the partial-order ranking (POR) of the meta-dimension based on meta-descriptors derived from randomly generated linear extensions.

**Table 4** Eventual linear ranking according to Eq. (2),  $Rk_{av}$ , and based on linear extension,  $RLE$ , estimated based on meta-descriptors as derived based on average linear rank estimated based on randomly generated linear extensions<sup>a</sup>

Site No.	$Rk_{av}$	$RLE$
1	2.4	4.0
2	10.2	10.1
3	4.3	6.1
4	11.3	10.7
5	12.8	10.9
6	5.7	7.3
7	4.3	6.9
8	4.9	5.8
9	2.1	2.7
10	11.3	9.5
11	12.8	11.4
12	4.3	6.8
13	2.8	4.2
14	12.8	10.9
15	15.5	15.0
16	14.9	13.8

<sup>a</sup> cf. Table 2.

Further, it can be argued that the average rank is associated with a certain uncertainty, as the single elements possess a distribution of possible ranks (Sørensen et al., 2001; Lerche et al., 2002, 2003; Carlsen, 2006b). However, the above study demonstrates that only minor differences between the two methods to generate meta-descriptors, i.e. average ranks, prevail in agreement with previous studies (Brüggemann et al., 2004). Further, previous studies have demonstrated that the use of randomly generated linear extensions can be applied appropriately to obtain ranking probabilities and thus average ranks (Sørensen et al., 2001; Lerche et al., 2002, 2003). Thus, both methods may apparently be used to generate meta-descriptors for HPOR.

#### 4. CONCLUSIONS

The present study has demonstrated that prioritisation of polluted sites for possible confinement or possibly remediation or clean up can be advantageously achieved by the use of HPOR, thus allowing the inclusion of a higher number of descriptors than will usually be applicable in simple POR as a too high number of descriptors compared to the number of elements typically leads to non-robust models. Thus, the HPOR approach is based on the generation of meta-descriptors, each of which bearing the information of a POR of the elements under investigation based a group of primary descriptors. Subsequently, a second POR, based on the set of meta-descriptors, leads to the final ranking of the elements, possibly being expressed as their mutual average ranks.

#### ACKNOWLEDGMENTS

The author is grateful to the National Environmental Research Institute, Denmark, for making the OML-Multi model for studying atmospheric dispersion available.

The author further wishes to thank the members of the Center of Physical-Chemical Methods of Analysis, Kazakh National University, Almaty, for extended discussions on the dimethyl hydrazine problem associated with the space activities at the Baikonur Cosmodrome.

#### REFERENCES

- Brüggemann, R., Carlsen, L. (Eds.) (2006). *Partial Order in Environmental Sciences and Chemistry*, Springer, Berlin (GER).
- Brüggemann, R., Halfon, E., Bücherl, C. (1995). *Theoretical Base of the Program "Hasse"*, GSF-Bericht 20/95, Neuherberg (GER); The software may be obtained by contacting Dr. R. Brüggemann, Institute of Freshwater Ecology and Inland Fisheries, Berlin, brg@igb-berlin.de, or brg\_home@web.de.
- Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., Steinberg, C.E.W. (2001). 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.

- Brüggemann, R., Lerche, D., Sørensen, P.B., Carlsen, L. (2004). Estimation of average ranks by a local partial order model, *J. Chem. Inf. Comput. Sci.*, 44, 618–625.
- Carlsen, L. (2006a). A combined QSAR and partial order ranking approach to risk assessment, *SAR QSAR Environ. Res.* 17, 133–146.
- Carlsen, L. (2006b). Interpolation schemes in QSAR. In: *Partial Order In Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L. Eds), Springer, Berlin (GER).
- Carlsen, L. (2008). Hierarchical partial order ranking, *Environ. Pollut. In press.* (available online 080104: HYPERLINK "<http://dx.doi.org/10.1016/j.envpol.2007.11.023>"\t"doilink"doi: 10.1016/j.envpol.2007.11.023).
- Carlsen, L., Kenesova, O.A., Batyrbekova, S.E. (2007). A preliminary assessment of the potential environmental and human health impact of unsymmetrical dimethylhydrazine as a result of space activities, *Chemosphere* 67, 1108–1116.
- Carlsen, L., Kenessov, B., Batyrbekova, S.Ye. (2008). A QSAR/QSTR study on the environmental health impact by the rocket fuel 1,1-dimethyl hydrazine and its transformation products, *Environ. Health Insights*, 1, 11–20 (available online: [http://la-press.com/article.php?article\\_id=913](http://la-press.com/article.php?article_id=913)) (accessed Oct. 2008).
- Davey, B.A., Priestley, H.A. (1990). *Introduction to Lattices and Order*, Cambridge University Press, Cambridge (UK).
- DMU(2006). *OML-Multi ver. 5.03*, National Environmental Research Institute (DK), <http://oml-international.dmu.dk>.
- Fishburn, P.C. (1974). On the family of linear extensions of a partial order, *J. Combinat. Theory* 17, 240–243.
- Graham, R.L. (1982). Linear extensions of partial orders and the FKG inequality. In: *Ordered Sets* (Rival, I. Ed.) Reidel Publishing Company, Dordrecht (NL), pp. 213–236.
- Halfon, E., Reggiani, M.G. (1986). On the ranking of chemicals for environmental hazard, *Environ. Sci. Technol* 20, 1173–1179.
- Hasse, H. (1952). *Über die Klassenzahl abelscher Zahlkörper*, Akademie Verlag, Berlin (GER).
- Lerche, D., Brüggemann, R., Sørensen, P., Carlsen, L., Nielsen, O.J. (2002). A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances, *J. Chem. Inf. Comput. Sci.* 42, 1086–1098.
- Lerche, D., Sørensen, P.B., Brüggemann, R. (2003). Improved estimation of ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability, *J. Chem. Inf. Comput. Sci.* 43, 1471–1480.
- MST (2006). Forurene og muligt forurene grunde, <http://www.mst.dk/default.asp?Sub=http://www.mst.dk/affald/02020000.htm> (in Danish).
- Sørensen, P.B., Mogensen, B.B., Carlsen, L., Thomsen, M. (2000). The influence of partial order ranking from input parameter uncertainty. Definition of a robustness parameter, *Chemosphere* 41, 595–601.
- Sørensen, P.B., Lerche, D.B., Carlsen, L., Brüggemann, R. (2001). Statistically approach for estimating the total set of linear orders. A possible way for analysing larger partial order sets. In: *Order Theoretical Tools in Environmental Science and Decision Systems* (Brüggemann, R., Pudenz, S., Lühr, H.-P. eds), Berichte des IGB, Leibniz-Institut of Freshwater Ecology and Inland Fisheries, Berlin, Heft 14, Sonderheft IV, pp. 87–97.
- Winkler, P.M. (1982). Average height in a partially ordered set. *Discrete Mathematic.* 39, 337–341.
- Winkler, P.M. (1983). Correlation among partial orders. *J. Alg. Disc. Meth.*, 4, 1–7.

# Similarity/Diversity Measure for Sequential Data Based on Hasse Matrices: Theory and Applications

A. Mauri and D. Ballabio

---

Contents	1. Introduction	111
	2. Theory	112
	2.1 Principles of partial ordering	112
	2.2 Similarity/diversity measures based on Hasse distance	112
	2.3 Hasse distance between matrices of different size	114
	2.4 Example of Hasse distances	115
	3. Application of the Hasse Distance Approach to Sequential Data	116
	3.1 DNA sequences	118
	3.2 NMR and mass spectra	120
	3.3 Molecular descriptors	124
	3.4 Electronic nose signals	127
	3.5 Proteomic maps	130
	4. Conclusions	137
	References	137

---

## 1. INTRODUCTION

The concept of similarity and its dual concept of diversity play a fundamental role in several fields, such as library searching, virtual screening, QSAR/QSPR modelling as well as in genomics and proteomics. Usually, the concept of similarity/diversity is defined by means of distances that are the quantitative measure of diversity between a pair of objects, thus large distances indicate large diversity. Several distance measures, such as Euclidean, Manhattan, Minkowski, Canberra distances for quantitative variables and Hamming, Tanimoto, Jaccard distances for binary variables have been defined.

In this chapter, a new similarity/diversity measure is described as a new approach for the analysis of sequential data, where useful information can also be obtained by the ordering relationships between the sequence elements. This new similarity/diversity measure is based on the distance evaluated between pairs of Hasse matrices derived from the classical partial ordering rules. It can be naturally standardised, thus allowing the interpretation of these distances as absolute values (e.g. percentage) and deriving simple similarity and correlation indices.

Basically, the similarity/diversity between two sequences is obtained by the definition of a distance between the corresponding Hasse matrices; these distances have some useful properties and seem to show high sensitivity to changes in structure sequences. This methodology can be used for several applications: (a) evaluation of molecular similarity/diversity, using sets of sequential descriptors; (b) evaluation of similarity between spectra or sequential analytical data and (c) evaluation of DNA and protein sequences and, in general, assessment of similarity of sequential data.

## 2. THEORY

The theory of the proposed approach to the similarity/diversity analysis of sequential data is presented by introducing some partial ordering concepts, the Hasse matrix and the calculation of the proposed distance between Hasse matrices; some examples are given, showing its main characteristics and possible different applications.

### 2.1 Principles of partial ordering

Partial ordering is an approach to the ranking where the relationship of “incomparability” is added to the classical relationships of “greater than”, “less or equal than”, etc. (Halfon and Reggiani, 1986; Brüggemann and Bartel, 1999; Brüggemann et al., 2004; Pavan and Todeschini, 2004). Given a set  $Q$  of  $n$  elements, each described by a vector  $\mathbf{x}$  of  $p$  variables (attributes), the two elements  $s$  and  $t$  belonging to  $Q$  are comparable if *for all* the variables  $x_j$  either  $x_j(t) \geq x_j(s)$  or  $x_j(s) \geq x_j(t)$ . If  $x_j(t) \geq x_j(s)$  *for all*  $x_j$  ( $j=1, \dots, p$ ), then  $t \geq s$ . The request “*for all*” is very important and is called the generality principle:

$$t \geq s \Leftrightarrow x_j(t) \geq x_j(s), \quad \forall j \in [1, p] \quad (1)$$

The ordering relationships between all the pairs of elements are collected into the Hasse matrix; for each pair of elements  $s$  and  $t$ , the entry  $H_{st}$  of this matrix is:

$$H_{st} \begin{cases} +1, & \text{if } x_j(s) \geq x_j(t), \quad \forall j \in [1, p] \\ -1, & \text{if } x_j(s) < x_j(t), \quad \forall j \in [1, p] \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

If the entry  $s - t$  contains  $+1$ , the entry  $t - s$  contains  $-1$ ; if the entry  $s - t$  contains  $0$ , the entry  $t - s$  contains  $0$ . Thus, the Hasse matrix is a square  $n \times n$  matrix whose elements take only values  $0$  and  $\pm 1$ ; if pairs of equal elements are not present, the Hasse matrix is antisymmetric. In fact, in the presence of elements having the same values for all the variables, in both the corresponding entries of the Hasse matrix ( $s - t$  and  $t - s$ ), a value equal to  $1$  is stored.

It is interesting to observe that the Hasse matrix contains a holistic view of all the ordering relationships among  $n$  elements belonging to the set  $Q$ . In other words, the Hasse matrix can be assumed as a fingerprint of the ordering relationships among  $n$  elements.

In order to add more information to the Hasse matrix, the augmented Hasse matrix can be defined by adding to the main diagonal (zero in the original Hasse matrix) any property  $P$  of the elements. The property values of each set of  $n$  elements are scaled by dividing each value by the maximum property value:

$$H_{ii} = P_i / P_{\text{MAX}}$$

## 2.2 Similarity/diversity measures based on Hasse distance

Let  $\mathbf{H}^A$  and  $\mathbf{H}^B$  be two  $n \times n$  Hasse matrices obtained by two different realisations of the variables defining  $n$  elements, i.e. representing two partial orderings A and B. The distance between the two partial orderings can be obtained by summing up the differences between the corresponding matrix elements. The distance between A and B can be considered as the contribution of two terms:

$$d_D(A, B) = \frac{\sum_{i=1}^n |H_{ii}^A - H_{ii}^B|}{n}, \quad d_H(A, B) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n |H_{ij}^A - H_{ij}^B|}{n(n-1)/2} \quad (3)$$

where the first term  $d_D$  is the contribution to the distance due to the diagonal terms (the property values), while the second term  $d_H$  is the contribution to the distance due to the off-diagonal terms (the ranking relationships of the Hasse matrix). In both cases, the two distance terms  $d$  range from  $0$  to  $1$ . This is obvious for the diagonal contribution using scaled values but not for the off-diagonal contribution.

In case that only two variables are considered in building the Hasse matrix and that no discrepancy is observed between the ordering provided by the two variables, the corresponding Hasse matrix obtained contains only  $+1$  and  $-1$  values, meaning that a total ranking of the elements exists. If the Hasse matrix is obtained by using a second variable, which provides an inverse ordering with respect to the first one, it will comprise only zero values, meaning that no ordering relationships exist among the elements based on these variables. Then, it is noticeable that the maximum theoretical distance between these two matrices is  $n \times (n - 1)$ .

From the two contributions, a weighted standardised Hasse distance (WSHD) can be defined as a trade-off between the ranking relationships and the property values. Therefore, the WSHD  $d_W$  can be defined as:

$$d_W(A,B) = (1 - w) \times d_H(A,B) + w \times d_D(A,B), \quad 0 \leq d_W \leq 1 \quad (4)$$

where  $w$  is a weighting term ranging between 0 and 1. Using a weight equal to 0, the distance is calculated by taking into account only the ranking relationships, while a weight equal to 1 takes into account only the property values. A weight equal to 0.5 takes into account both terms, resulting in a distance measure where both the ordering relationships among the elements and their property differences are equally considered.

Moreover, WSHD can be seen as a generalized Manhattan distance calculated on the corresponding pairs of elements of two Hasse matrices, thus preserving all the metric properties of the Manhattan distance. This distance is straightforwardly interpretable as an absolute measure of distance (or as percentage  $d \times 100$ ) or as an absolute measure of similarity after the transformation as  $s = 1 - d_W$  or as a correlation measure after the transformation:

$$r_W = (1 - d_W)2 - 1, \quad -1 \leq r_W \leq +1 \quad (5)$$

The rank correlation  $r_H$  calculated for  $w=0$  (i.e.  $d_W = d_H$ ) coincides with the Greiner–Kendall rank correlation index, which is defined as

$$\tau = \frac{4 \sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij}^+}{n(n-1)} - 1, \quad -1 \leq \tau \leq +1 \quad (6)$$

where  $d_{ij}^+$  is defined as

$$d_{ij}^+ = \begin{cases} 1, & \text{if } i < j \text{ and } p_i < p_j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

and  $p$  are the ranks of the samples.

Therefore, the Spearman rank correlation uses more information with respect to the Greiner–Kendall rank correlation: the Spearman index is more suitable if no information has to be discarded, while the Greiner–Kendall index is a more robust statistical index.

### 2.3 Hasse distance between matrices of different size

As explained above, Hasse matrices are square  $n \times n$  antisymmetric matrices able to take into account the partial ordering of  $n$  elements. When two sets of different element size are considered, i.e. the two sets are constituted by  $n_1$  and  $n_2$  elements, respectively, with  $n_1 > n_2$ , two Hasse matrices  $\mathbf{H}^1$  ( $n_1 \times n_1$ ) and

$H^2$  ( $n_2 \times n_2$ ) of different size have to be compared. In this case, the WSHD distance is not univocally defined and the algorithm has to be further developed.

The distance between the two matrices can be calculated by overlapping  $n_1 - n_2 + 1$  times the smaller matrix ( $n_2 \times n_2$ ) to the bigger one ( $n_1 \times n_1$ , the reference matrix), starting from the left-up corner and shifting the smaller matrix diagonally until the right-down corner. Each distance between the pair of matrices is calculated as explained above and the smallest distance among the  $n_1 - n_2 + 1$  distances is taken as the final distance. Basically, this procedure is equivalent to the search of the subset of ordered elements in the bigger matrix, which is more similar to the  $n_2$  ordered elements in the smaller matrix.

## 2.4 Example of Hasse distances

In order to better understand the theory presented in paragraphs 2.1–2.3, a simple example is given. The five sequential intensities (1, 2, ..., 5) of two samples A and B are given in Table 1. The augmented Hasse matrices of samples A and B, obtained by comparing the ordering and the corresponding property variables, are given in Tables 2 and 3.

The diagonal elements of the two Hasse matrices have been scaled with respect to the maximum property values (20 for sample A and 18 for sample B). The differences between the matrices are collected in Table 4. The sums of the diagonal and off-diagonal terms (on the half matrix) are 1 and 3, respectively, and the distances:

$$d_D = \frac{1.00}{5} = 0.20, \quad d_H = \frac{3}{5(5-1)/2} = 0.30$$

Let us now suppose that sample B is represented only by the first four signals. In this case, the fifth row and the fifth column of Table 3 are not considered. The distance between the samples is calculated as the minimum distance between the two distances from the  $4 \times 4$  B Hasse matrix overlapped (1) to the first four rows/columns of the A matrix and (2) to the last four rows/columns of the A matrix. In this example, the two Hasse distances are both equal to  $4/12$ , i.e.  $d_H = 0.34$ .

**Table 1** Five-dimensional profiles of two artificial samples

Ordering variable	Property variable	
	Sample A	Sample B
1	12	15
2	17	18
3	20	16
4	14	10
5	6	12



**Table 2** Augmented Hasse matrix of sample A

A	1	2	3	4	5
1	0.60	−1	−1	−1	0
2	+1	0.85	−1	0	0
3	+1	+1	1.00	0	0
4	+1	0	0	0.70	0
5	0	0	0	0	0.30

**Table 3** Augmented Hasse matrix of sample B

B	1	2	3	4	5
1	0.83	−1	−1	0	0
2	+1	1.00	0	0	0
3	+1	0	0.89	0	0
4	0	0	0	0.56	−1
5	0	0	0	+1	0.67

**Table 4** Matrix of the difference between matrices A and B (reported in [Tables 2 and 3](#))

A−B	1	2	3	4	5
1	0.23	0	0	1	0
2	0	0.15	1	0	0
3	0	1	0.11	0	0
4	1	0	0	0.14	1
5	0	0	0	1	0.37

**3. APPLICATION OF THE HASSE DISTANCE APPROACH TO SEQUENTIAL DATA**

Data including an ordering variable can be considered as sequential data. These can be characterised by an ordering variable (sequential integer numbers, variable X1) and a property variable (real numbers, variable X2). Examples of sequential data are mass spectrometry signals, which are ordered by increasing masses, the intensity of signals being the property variable and their position in the spectrum the ordering variable; IR/UV signals, the signal intensity being the property variable and the wavelength the ordering variable; 1D NMR spectra, the signal intensity being the property variable and the chemical shifts the ordering variable. In general, all the spectra achieved along time are intrinsically ordered and can be analysed as sequential data. Analogously, data based on natural sequences can be also considered as sequential data.

In effect, a sequence of integer numbers representing the positions of the elements in the sequence is the ordering variable, while any property characterising the elements of the sequence is the property variable. A word can be thought as a sequence of characters whose position in the sentence is the ordering variable, while the position in the alphabet is the property variable. In the case of DNA sequences, which are sequences of four nucleic acids, the molecular weight can be chosen as the property characterising the elements of the sequence, i.e. the nucleic acids. For proteins, any physico-chemical property of the 20 amino acids can be used as property variable, while the most relevant protein abundances can be used in the case of proteomic maps.

This kind of data can be easily characterised by Hasse matrices and their similarity/diversity assessed by the previously defined Hasse distance. In this case, the maximum information about the sequence is obtained by using only two variables, i.e. the ordering variable ( $X1$ ) and the property variable ( $X2$ ). In fact, in this case, the incomparabilities between two samples  $s$  and  $t$  can be due to only one condition, i.e. when the two variables  $X1$  and  $X2$  show an opposite rank:

$$X1(s) > X1(t) \text{ and } X2(s) < X2(t) \text{ or } X1(s) < X1(t) \text{ and } X2(s) > X2(t)$$

For example, if three variables are taken into account, the incomparabilities between two samples can be obtained by opposite ranks of  $X1$ – $X2$  or  $X1$ – $X3$  or  $X2$ – $X3$ , with a loss of information. In fact, in this case, the presence of zero values in the Hasse matrix cannot be univocally related to a specific relationship. Examples of sequential data are collected in Table 5 and briefly discussed in the next paragraphs. Modules of MATLAB (Mathworks) dedicated to the calculation of Hasse distances have been produced by the authors of this chapter and can be freely downloaded at [michem.disat.unimib.it/chm](http://michem.disat.unimib.it/chm).

**Table 5** Examples of sequential data for applying the Hasse distance

Sequential data	Ordering variable	Property variable
DNA sequences	1, . . . , sequence length	A, C, G, T property
NMR spectra	1, . . . , 1500 (from spectra resolution)	Signal intensity
E-nose signals	None	Signal intensity
Mass spectra	1, . . . , 250 (from spectra resolution)	Signal intensity
Molecular descriptors	1, . . . , sequence length	Descriptor value
Proteomic maps	Number of considered proteins, . . . , 1	Protein abundance

3.1 DNA sequences

DNA sequences are sequences of four nucleic acid bases (adenine, thymine, guanine and cytosine) can be denoted by the letters A, T, G, C, respectively. Even when sequences are not too long, the search for their similarity/diversity is not usually easy as shown by several sequence comparisons considered in literature papers. Comparisons among DNA sequences can be evaluated using as ordering variable the element positions in the sequence and as property variable the molecular weight of the nucleic bases, as shown in Table 6 (Todeschini et al., 2006).

In order to illustrate the characteristics of the Hasse matrix and the corresponding Hasse diagram, a random 20-length sequence S1 composed of four different elements has been arbitrarily defined:

ATGGTGCACCTGACTCCTGA

The two variables used for building the Hasse matrices are shown in bold characters in Table 7. In Figure 1, the Hasse diagram of this sequence is represented. As it can be easily noted, the information contained in the diagram not only considers the absolute sequence of the elements but also highlights four linear extensions, one for each different element (A, C, G, T).

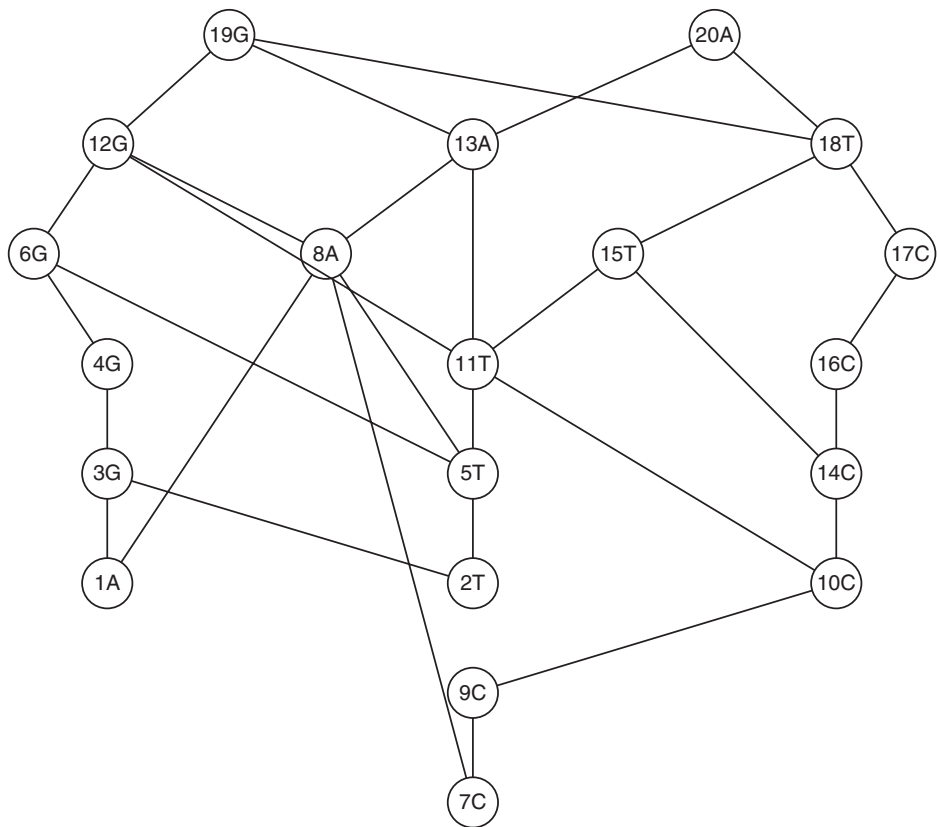
**Table 6** Different representations of the DNA sequences

Label	ID	MW	Scaled ID	Scaled MW
C	1	111.1	0.25	0.735
T	2	126.0	0.50	0.834
A	3	135.13	0.75	0.894
G	4	151.13	1.00	1.000

MW is the molecular weight.

**Table 7** The two variables used for building the Hasse matrices (columns 1 and 4 in bold characters) in order to represent a random 20-length DNA sequence S1

ID	Base	MW	Scaled ID
<b>1</b>	A	135.13	<b>0.75</b>
<b>2</b>	T	126.0	<b>0.50</b>
<b>3</b>	G	151.13	<b>1.00</b>
<b>4</b>	G	151.13	<b>1.00</b>
<b>5</b>	T	126.0	<b>0.50</b>
...	...	...	...
...	...	...	...
<b>18</b>	T	111.1	<b>0.50</b>
<b>19</b>	G	135.13	<b>1.00</b>
<b>20</b>	A	151.13	<b>0.75</b>



**Figure 1** Hasse diagram of the artificial sequence ATGGTGCACCTGACTCCTGA.

For example, the sequence of the element A is characterised by the path 1-8-13-20, while the sequence of the element C is characterised by the path 7-9-10-14-16-17. The links between pairs of nodes represent ordering relationships between the elements, while elements on the same horizontal level are incomparable elements (not linked among them).

A simple example of WSHD calculation showing its sensitivity to small changes is discussed by substituting the fifth element T of the sequence S1 by C (S2), A (S3) and G (S4), respectively. The calculated distances using the three weights 0, 0.5 and 1 are given in [Table 8](#).

As expected, the distances between S1-S2 and S1-S3, calculated by taking into account only the property values, are equal (both differences between the pair of elements T-C and T-A are 0.25); nevertheless, the corresponding distances calculated by using the ordered property are different: S1-S2 is equal to 3.684 and S1-S3 is equal to 2.105. For all the weights, the most dissimilar pair is S1-S4, the substitution being more influent in the global ordering of the sequence. The presence of four A characters and six C characters in the original sequence S1 justifies the greater difference when T is substituted by C.

**Table 8** Distances (as percentages) of the three modified sequences with respect to the original sequence S1, calculated by using three different weights

w	S2-C	S3-A	S4-G
0	3.684	2.105	4.737
0.5	2.467	1.678	3.618
1	1.250	1.250	2.500

### 3.2 NMR and mass spectra

In general, experimental spectra constitute a typical case where an ordering variable (time, masses, chemical shifts, wavelengths, etc.) and a property variable (signal intensities) can be naturally associated. In the case of mass and NMR spectra, the ordering variable is a sequence of integer numbers from 1 to the maximum number of numerically different signals (given by the spectra resolution). For example, in NMR spectroscopy, assuming that the chemical shifts take values from 0.01 to 15.00, with a resolution of 0.01, the total number of resolved signals is 1500. The signals with intensities greater than 0 are registered and embedded into 1500 signals.

The two variables used for building the Hasse matrices for 24 NMR signals whose intensities are greater than 0 are shown in bold characters in Table 9. These signals are successively embedded into a 1500 array. Then the distances are calculated from pairs of Hasse matrices of size  $1500 \times 1500$ . In an analogous way, the ordering variable for mass spectra is constituted by integer numbers ranging from 1 to the spectral resolution (assuming a mass resolution of 0.1 in the range

**Table 9** Example of 24 NMR signals with intensities greater than 0

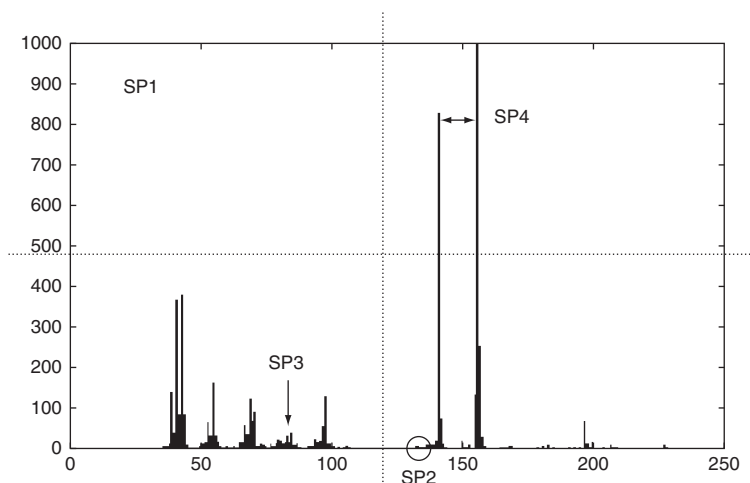
Signal	ppm	ID	Height
1	0.01	<b>1</b>	<b>0.1159</b>
2	1.17	<b>117</b>	<b>0.1278</b>
3	1.18	<b>118</b>	<b>0.1247</b>
4	2.17	<b>217</b>	<b>1</b>
5	2.42	<b>242</b>	<b>0.0314</b>
6	2.43	<b>243</b>	<b>0.0255</b>
...	...	...	...
...	...	...	...
...	...	...	...
21	11.74	<b>1174</b>	<b>0.0709</b>
22	11.88	<b>1188</b>	<b>0.0703</b>
23	13.79	<b>1379</b>	<b>0.0629</b>
24	13.97	<b>1397</b>	<b>0.0668</b>

The ordering variable (ID) and the property variable (height) selected for building the Hasse matrices are shown in bold characters.

0–25, 250 signals are obtained). The sensitivity of WSHD is evaluated on a real mass spectrum of pentobarbital (SP1).

The original mass spectrum (SP1) is shown in Figure 2, where the differences of SP2, SP3 and SP4 are also highlighted. The data are collected in Table 10, where only the signals different from 0 are reported. A total number of 250 signals is considered and three spectra are arbitrarily obtained by performing small modifications of the signal intensities of the first spectrum. In particular, for SP2, only some small signals have been modified (signals 133–135), for SP3, one signal has been modified from 30 to 0 (signal 83) and for SP4, the two greatest signals have been inverted (signals 141 and 156).

The distances calculated by using the weights 0, 0.25, 0.50, 0.75 and 1 are collected in Table 11. As it can be seen, when only the signal intensities are considered ( $w=1$ ), the two most similar spectra are SP1 and SP2, while the most dissimilar spectra are SP3–SP4 (0.164), SP2–SP4 (0.153) and SP1–SP4 (0.151). However, when only the signal ranking is considered ( $w=0$ ), the two most similar spectra are SP1–SP4 (0.004), while the most dissimilar spectra are SP2–SP3 (0.847), SP3–SP4 (0.634) and SP1–SP3 (0.630). It is interesting to observe the opposite behaviour of the distances between SP1 and SP4, where the intensities of two highest signals have been exchanged. In this case, the contribution of the difference of intensities is maximal (case  $w=1$ ), while in the Hasse diagram only the two cells corresponding to the two highest signals take opposite values, the ranking of all the other signals not being influenced. The strong sensitivity of the Hasse distance ( $w=0$ ) to changes of small/medium signals is highlighted by the distances between SP1–SP2 (0.216) and SP1–SP3 (0.630).



**Figure 2** Mass spectrum of SP1, together with the modifications performed for spectra SP2 and SP4 (see also Table 10).

**Table 10** Signal intensities for the pentobarbital (SP1) and three modified simulated spectra (SP2–SP4)

Mass	SP1	SP2	SP3	SP4	Mass	SP1	SP2	SP3	SP4	Mass	SP1	SP2	SP3	SP4
36	3	3	3	3	75	3	3	3	3	139	6	6	6	6
37	3	3	3	3	77	10	10	10	10	140	20	20	20	20
38	12	12	12	12	78	5	5	5	5	141	826	826	826	1000
39	139	139	139	139	79	12	12	12	12	142	71	71	71	71
40	38	38	38	38	80	22	22	22	22	143	11	11	11	11
41	364	364	364	364	81	18	18	18	18	144	1	1	1	1
42	83	83	83	83	82	13	13	13	13	151	1	1	1	1
43	378	378	378	378	83	30	30	0	30	152	1	1	1	1
44	84	84	84	84	84	13	13	13	13	153	6	6	6	6
45	6	6	6	6	85	38	38	38	38	154	0	0	0	0
50	5	5	5	5	86	8	8	8	8	155	133	133	133	133
51	11	11	11	11	87	9	9	9	9	156	1000	1000	1000	826
52	14	14	14	14	88	1	1	1	1	157	253	253	253	253
53	64	64	64	64	91	5	5	5	5	158	29	29	29	29
54	31	31	31	31	92	3	3	3	3	159	3	3	3	3
55	162	162	162	162	93	4	4	4	4	165	1	1	1	1
56	30	30	30	30	94	21	21	21	21	166	1	1	1	1
57	17	17	17	17	95	17	17	17	17	167	2	2	2	2
58	5	5	5	5	96	20	20	20	20	168	3	3	3	3
59	1	1	1	1	97	55	55	55	55	169	5	5	5	5
60	5	5	5	5	98	127	127	127	127	179	1	1	1	1
61	2	2	2	2	99	9	9	9	9	181	4	4	4	4

62	1	1	1	1	100	3	3	3	3	183	7	7	7	7
63	3	3	3	3	101	3	3	3	3	185	2	2	2	2
64	1	1	1	1	102	0	0	0	0	191	1	1	1	1
65	13	13	13	13	103	1	1	1	1	193	1	1	1	1
66	13	13	13	13	105	2	2	2	2	195	2	2	2	2
67	58	58	58	58	106	3	3	3	3	197	65	65	65	65
68	33	33	33	33	107	2	2	2	2	198	10	10	10	10
69	120	120	120	120	133	3	2	3	3	199	2	2	2	2
70	67	67	67	67	134	2	3	2	2	204	1	1	1	1
71	89	89	89	89	135	2	3	2	2	207	6	6	6	6
72	5	5	5	5	136	2	1	2	2	208	2	2	2	2
73	9	9	9	9	137	6	6	6	6	209	1	1	1	1
74	8	8	8	8	138	7	7	7	7	227	6	6	6	6
										228	1	1	1	1

The modified intensities are in bold characters.



**Table 11** The distances (as percentages) between SP1 and SP4 spectra calculated with different weights

$w$	$d(1-2)$	$d(1-3)$	$d(1-4)$	$d(2-3)$	$d(2-4)$	$d(3-4)$
0.00	0.216	0.630	0.004	0.847	0.220	0.634
0.25	0.163	0.476	0.041	0.639	0.203	0.517
0.50	0.109	0.322	0.078	0.431	0.187	0.399
0.75	0.055	0.167	0.114	0.223	0.170	0.282
1.00	0.002	0.013	0.151	0.015	0.153	0.164

### 3.3 Molecular descriptors

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health research and quality control, being obtained when molecules are transformed into a molecular representation allowing some mathematical treatment. Many molecular descriptors have been proposed until now, derived from different theories and approaches (Todeschini and Consonni, 2000). The information content of a molecular descriptor depends on the kind of molecular representation used and on the defined algorithm for its calculation. There are simple molecular descriptors derived by counting some atom types or structural fragments in the molecule, others derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D descriptors, and there are molecular descriptors derived from a geometrical representation, which are called geometrical or 3D descriptors.

In chemistry, molecular descriptors are the basic elements used by all the methods for assessing molecular similarities. In order to apply the proposed approach to the similarity/diversity in QSAR/QSPR problems, a set of convenient molecular descriptors has to be found. Several ordered descriptors, such as autocorrelation descriptors of different lags, connectivity indices of different orders and radial distribution function (RDF) descriptors, are defined in literature. However, not all the ordered descriptors can be properly used in this approach. In fact, when the ranking of the descriptors values depends largely on the descriptor definition and less on the molecular structure, the descriptors cannot be used, since all the Hasse matrices are equal, i.e. the similarity/diversity measure does not depend on the descriptor ranking.

For example, the values of connectivity indices  $\chi_0, \chi_1, \dots, \chi_5$  calculated for different molecules largely differ among them, but their ranking is the same in almost all the cases, i.e. they decrease from  $\chi_0$  to  $\chi_5$ . Then, the information related to the ranking is lost and the differences in similarity/diversity arise only from the differences in descriptor values. A set of ordered descriptors showing a ranking independence is the set of RDF descriptors (Hemmer et al., 1999). They are defined as:

$$\text{RDF}_{R,w} = f \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i w_j e^{-\beta(R-r_{ij})^2} \quad (8)$$

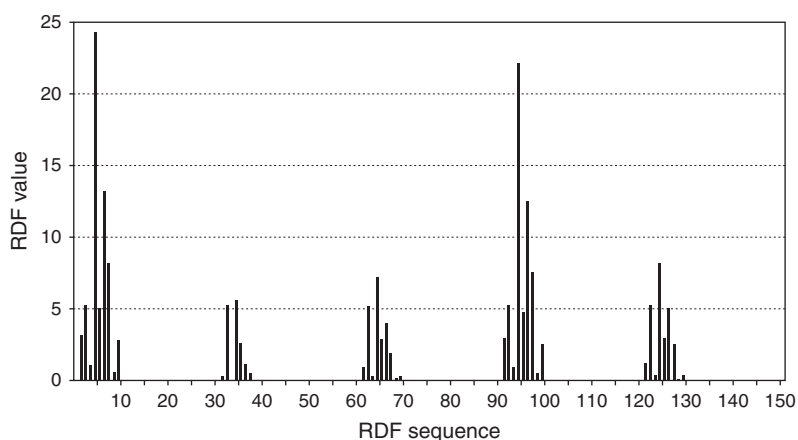
where  $A$  is the number of atoms,  $w$  the atomic property and  $r_{ij}$  the geometric distance between  $i$ th and  $j$ th atoms. The parameter  $f$  is a scaling factor (equal to 1),  $\beta$  a smoothing parameter (assumed equal to 100) and  $R$  represent the radius related to the spherical volume (range of 1–15), with a step assumed equal to 0.5 Å. Five different properties have been used (Table 12) and 30 RDF descriptors for each property have been calculated using DRAGON software (Mauri et al., 2006; Talet srl, 2006), giving a total of 150 descriptors for molecule. An example of RDF spectrum of cyclohexane is shown in Figure 3.

In order to explore the characteristics of the proposed similarity measure, 10 simple selected non-congeneric molecules were considered. The distances between the 10 molecules calculated with  $w=0$  (upper matrix) and  $w=0.5$  (lower matrix) are shown in Table 13.

Being the set of 150 RDF descriptors built by considering all the weights of Table 12, the molecule representation contains several different sources of chemical information (geometric, mass-related, electronic, etc.). The use of selected subsets can obviously highlight different levels of chemical similarity. As it can be noted, only considering the distances calculated by the Hasse

**Table 12** Atomic properties (weights) used for the calculation of the radial distribution function (RDF) descriptors

Weight	Description
$u$	No weight (all weights equal to 1)
$m$	Atomic mass
$v$	van der Waals volume
$e$	Sanderson electronegativity
$p$	Atomic polarizability



**Figure 3** Radial distribution function (RDF) spectrum of cyclohexane. The five blocks of signals correspond to five different properties (Table 12).

**Table 13** WSHD distances (as percentages) for  $w=0$  (upper matrix) and  $w=1$  (lower matrix)

ID	Molecule	1	2	3	4	5	6	7	8	9	10
1	<i>n</i> -Hexane	0	18.398	18.685	14.058	17.575	16.206	16.617	17.888	11.579	14.130
2	Cyclohexane	<i>11.791</i>	0	<b>1.718</b>	8.510	<b>3.436</b>	5.486	4.412	5.754	11.579	27.302
3	Benzene	<i>12.077</i>	<b>2.180</b>	0	7.204	<b>2.291</b>	4.788	4.823	6.309	10.416	26.067
4	Toluene	<i>9.685</i>	<i>6.280</i>	<i>4.512</i>	0	5.736	5.047	6.497	7.427	4.358	21.280
5	F-benzene	<i>11.533</i>	<b>3.270</b>	<b>1.687</b>	<b>3.666</b>	0	<b>3.302</b>	6.738	8.260	9.092	25.011
6	Cl-benzene	<i>11.659</i>	<i>5.540</i>	<i>4.429</i>	<i>4.478</i>	<b>3.571</b>	0	8.591	9.951	8.421	24.931
7	Br-benzene	<i>11.818</i>	<i>4.205</i>	<i>4.345</i>	<i>5.599</i>	<i>5.514</i>	7.293	0	<b>2.720</b>	9.065	25.074
8	I-benzene	<i>13.256</i>	<i>6.034</i>	<i>6.152</i>	<i>6.758</i>	<i>7.266</i>	<i>8.789</i>	<i>3.803</i>	0	10.353	26.488
9	Naphthalene	<i>8.215</i>	<i>7.833</i>	<i>6.516</i>	<b>3.168</b>	<i>5.752</i>	<i>6.572</i>	<i>7.086</i>	<i>8.428</i>	0	17.888
10	Anthracene	<i>10.358</i>	<i>17.619</i>	<i>16.278</i>	<i>13.383</i>	<i>15.643</i>	<i>16.581</i>	<i>16.507</i>	<i>18.245</i>	<i>11.003</i>	0

The five lowest distances of  $w=0$  are in bold and the five lowest distances for  $w=1$  are in bold-italic.

matrices ( $w=0$ ), the most similar pairs of molecules are cyclohexane and benzene, benzene and toluene, Br-benzene and I-benzene, F-benzene and Cl-benzene; moreover, the molecules less dissimilar from anthracene are *n*-hexane and naphthalene. By considering the weighted Hasse distances ( $w=0.5$ ), similar considerations can be done.

### 3.4 Electronic nose signals

During the last decade, electronic nose has increased in uses, capabilities and applications in food science (Gardner and Bartlett, 1993; Bartlett et al., 1997). Basically, the principle involved in the electronic nose is the transfer of the total headspace of a sample to a sensor array, where each sensor has partial specificity to a wide range of aroma molecules. These non-selective gas sensors are theoretically able to simulate human sensing and give an objective tool of detecting aromatic fingerprints. Since electronic noses offer several advantages (cheapness, quickness, simplicity, little or no prior sample preparation), they have been used in food science for various applications: assessment of food properties, detection of adulteration, sensory properties prediction, classification of different food matrices (Llobet et al., 1999; Guadarrama et al., 2000; Brezmes et al., 2001; Cerrato Oliveros et al., 2002; García-González and Aparicio, 2003; Buratti et al., 2004; Vinaixa et al., 2004; Vinaixa et al., 2005; Buratti et al., 2006; Cosio et al., 2006). Unlike traditional analytical methods, electronic nose sensor responses do not provide information on the nature of the compounds under investigation but only give a digital fingerprint of the food product, which can be subsequently investigated by means of multivariate analysis.

Electronic nose data can also be characterised by means of Hasse matrices and their similarity/diversity measures (Ballabio et al., 2006). In fact, even if parameters are usually extracted from the electronic nose spectra, it is not necessary to transform the measured time profile into univariate features (Skov and Bro, 2005). Therefore, if the whole signal among time is considered, i.e. the sequential property of the data is preserved, the Hasse approach can be easily applied.

With respect to all the other applications showed in the previous paragraphs, when dealing with electronic nose data, only the property variable is used: in this way, incomparabilities cannot be present and only Hasse matrices with  $\pm 1$  off-diagonal values can be obtained. Hence, for each sample ( $s$ ) of each electronic nose sensor ( $e$ ), an Hasse matrix ( $\mathbf{H}^{se}$ ) can be calculated, by considering the intensities among time as unique property variable, i.e. the time profile of the sample. If a total number of  $m$  times are taken into account, each  $\mathbf{H}^{se}$  matrix, built on the basis of the ordering relationships of these  $m$  times, has dimensions  $m \times m$ , and each element  $H_{ij}$  represents the relationship between the signal intensities at  $i$ th and  $j$ th times, i.e. it says intensity in time  $i$  is lower ( $H_{ij} = -1$ ) or higher ( $H_{ij} = 1$ ) than intensity in time  $j$ . Consequently a total ordering relationship is obtained, since no incomparabilities will be present. However, in this way, the pairwise ordering relationships are also taken into account when the similarity measure is calculated between two different samples. Concerning the augmented Hasse matrix, the considered property added into the main diagonal

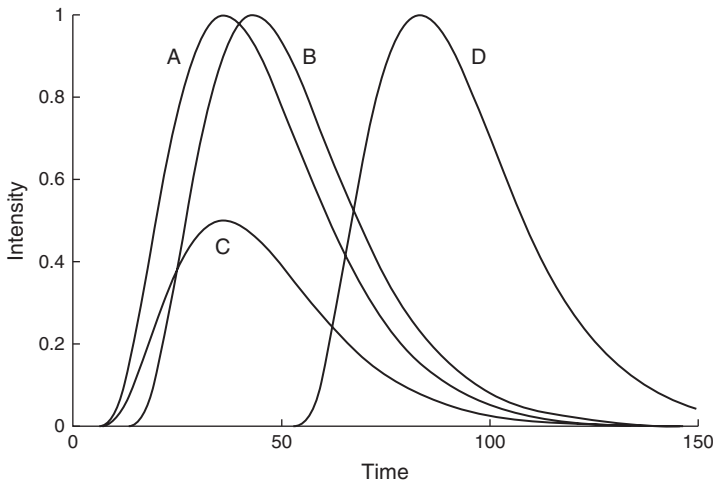
of each Hasse matrix is just the intensity (scaled on the maximum value). Therefore, the achieved Hasse matrix can be interpreted as a fingerprint of the sample time profile, i.e. a mathematical representation of the electronic nose signal, which takes into account both information of the curve shape and the intensity.

In order to clarify the Hasse approach for electronic nose data, four curves (A, B, C and D) have been built as different gamma probability density functions and their Hasse distances have been calculated. These four curves can represent the time profiles of four samples, achieved by means of a singular electronic nose sensor. In Figure 4, each curve is shown as intensity values ( $I$ ) plotted versus 150 times values. The curves B, C and D have been built by a small shift of curve A (B), by its intensity reduction (C) and by a big shift of curve A (D), respectively.

In Table 14, some intensity (not consecutive) values for curve A are reported: the intensity column is the input of the Hasse analysis, as explained above. For example, taking into account the signals 33 (0.98), 34 (0.99) and 41 (0.96), the Hasse matrix elements are the following:  $H^A(33,34) = -1$ , while  $H^A(34,33) = +1$ , since  $I(33)$  is lower than  $I(34)$ ;  $H^A(33,41) = +1$ ,  $H^A(41,33) = -1$ , and so on. In Table 15, a partial augmented Hasse matrix, relative to the data of Table 14 (times 33–43), is shown. In the main diagonal, zero values have been replaced with the scaled intensities.

After the whole Hasse matrices  $H^A$ ,  $H^B$ ,  $H^C$  and  $H^D$  have been calculated, the weighted Hasse distances between the four curves have been carried out (Table 16), by using a weight  $w$  equal to 0 (i.e. taking into account only the ranking relationships) and equal to 1 (i.e. taking into account only the intensities).

In the first case ( $w=0$ ),  $d_W(A,C)$ , i.e. the weighted distance calculated between A and C, is equal to 0, since the curves A and C have exactly the same shape and do not shift;  $d_W(A,B)$  is small (with respect to the range 0–1), since the



**Figure 4** Plot of four simulated electronic nose signals.

**Table 14** Time and intensity values of curve A

Time	Intensity
1	0.00
2	0.00
3	0.00
...	...
33	0.98
34	0.99
35	1.00
36	1.00
37	1.00
38	0.99
39	0.99
40	0.98
41	0.96
42	0.95
43	0.93
...	...
148	0.00
149	0.00
150	0.00

Ranges 1:3, 33:43 and 148:150 are reported.

**Table 15** Augmented Hass matrix, relative to the data of [Table 14](#) (times 33–43)

	33	34	35	36	37	38	39	40	41	42	43
33	0.98	-1	-1	-1	-1	-1	-1	1	1	1	1
34	1	0.99	-1	-1	-1	1	1	1	1	1	1
35	1	1	1	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	1	1
38	1	1	-1	-1	-1	0.99	1	1	1	1	1
39	1	1	-1	-1	-1	1	0.99	1	1	1	1
40	1	-1	-1	-1	-1	-1	-1	0.98	1	1	1
41	-1	-1	-1	-1	-1	-1	-1	-1	0.96	1	1
42	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.95	1
43	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.93

shift between A and B is small and  $d_W(A,D)$  is greater, since the shift between A and D is greater;  $d_W(C,D)$  is equal to 0.53 and is the greatest distance calculated with  $w = 0$ . The sensibility of the Hasse approach can be highlighted by considering the results achieved with  $w = 1$ . In fact, the distance between curves D and C

**Table 16** Weighted standardised Hasse distances between the four simulated curves, calculated with weight  $w = 0$  (non-italic) and  $w = 1$  (italic)

$d_w$	A	B	C	D
A	0.00	0.09	0.00	0.53
B	<i>0.09</i>	0.00	0.09	0.46
C	<i>0.15</i>	<i>0.17</i>	0.00	0.53
D	<i>0.48</i>	<i>0.43</i>	<i>0.38</i>	0.00

should be expected to be the greatest one, since the profiles differ in time and intensity, but this is not confirmed by considering only the intensities, since with  $w = 1$ ,  $d_w(A, D)$  and  $d_w(B, D)$  are greater than  $d_w(C, D)$ . Hence, the combination of the two approaches could represent an optimal solution.

### 3.5 Proteomic maps

The evaluation of complex therapeutic and toxic behaviour of chemicals from their effects on simpler biological systems such as cells is among the most interesting trends in drug discovery, environmental safety studies, molecular pharmacology and hazard assessment. This scientific cross-breeding is going under the name “toxicogenomics”. In these last years, special attention has been focused to the cellular proteome that characterises the different abundance of thousands of proteins belonging to the same cell.

A typical proteomic map is a planar map constituted by two axes representing charge ( $x$ -axis) and mass ( $y$ -axis) where even 2000 cell's proteins may appear as separated spots accordingly to their charge versus mass values; the spot diameters are related to the abundance of the proteins. Toxicological studies on proteomic maps consist in perturbing the control cell with a chemical and evaluate the resulting differences in the abundance of protein expressions with respect to the control cell.

A very interesting mathematical challenge is to understand the complexity of cellular events and then describe and characterise changes in proteomic maps. In the literature there are several attempts to get a numerical characterisation of the proteomic maps and similarity/diversity measures, transforming the maps into mathematical entities such as numbers and graphs (Randic, 2001; Randic et al., 2001a; Randic et al., 2001b; Randic and Basak, 2002; Witzmann, 2002; Bajzer et al., 2003; Randic et al., 2004; Randic et al., 2005). These approaches are based on transforming the proteomic information into convenient graphs and topological matrices such as distance/distance matrices, Euclidean adjacency matrices, path distance matrices (Randic et al., 2001a; Randic and Basak, 2002; Randic et al., 2005), neighborhood matrices (Bajzer et al., 2003; Randic et al., 2004; Randic et al., 2005) and then deriving graph invariants (such as eigenvalues).

Obviously, the theory of the Hasse distances can also be extended to this kind of data and used to characterise proteomic maps. Five proteomic maps were taken from literature (Randic et al., 2001a; Randic and Basak, 2002). They consist

of 29 most abundant proteins, represented by their  $x$ - $y$  coordinates and their abundances (Table 17). The set of proteomic maps is constituted by four peroxisome proliferators, i.e. perfluorooctanoic acid (PFOA), perfluorododecanoid acid (PFDA), clofibrate and diethylhexyl phthalate (DEHP), which characterise their effects on the proteomic map of rat liver cells (control map).

In this case, the two variables used to characterise the proteomic maps by the Hasse matrices are the integer numbers representing the inverse ranking of the detected abundances (common to all the proteomic maps) and the abundance values of each proteomic map, scaled with respect to the maximum abundance, thus obtaining values ranging between 0 and 1 (Table 18). The abundances of the

**Table 17** Twenty-nine most abundant proteins of the five proteomic maps

#	$x$	$y$	1 Control	2 PFOA	3 PFDA	4 Clofibrate	5 DEHP
22	1183.9	959.6	136653	113859	150253	163645	8111
52	2182.2	928.8	127195	99160	73071	76642	112096
20	1527.9	825.5	114929	192437	221567	166080	180590
62	1346.0	1352.5	112251	58669	38915	73159	77075
48	1406.3	1118.1	98224	91147	82963	84196	92942
9	1474.0	665.1	90004	129340	112361	112655	119402
36	2068.4	823.1	84842	73814	45482	71911	97444
2	642.2	669.8	82492	73974	74466	84703	88545
44	2032.7	902.8	80015	77314	80072	76027	100836
15	1053.6	864.3	72173	77982	60376	46808	78121
5	1214.3	620.0	64684	63511	38075	58364	75760
45	2094.5	680.5	58977	142865	46225	48625	146609
1	1021.7	390.2	58001	56547	53473	60224	71654
56	2070.4	929.6	55402	46146	33152	59438	69031
35	1375.7	992.3	49027	42506	52137	46058	69214
19	1623.4	640.8	48976	81452	133705	64580	65976
47	1842.5	885.9	48145	40390	24149	47585	52350
14	1189.5	614.7	42773	49044	77144	46005	59322
26	1465.5	821.1	40923	60359	94014	79981	3838
41	1323.4	993.0	36433	30640	31611	33764	44692
39	1278.8	981.6	35896	31707	21801	29026	32956
24	1433.5	662.3	31194	42226	41489	42432	69142
29	1170.0	862.2	30510	29742	30786	26460	37812
33	1139.2	958.4	29296	30067	39531	39204	27565
18	1167.3	611.7	26155	25182	41604	21039	23170
12	1202.3	495.5	25389	22811	17341	20416	30852
40	1030.2	863.2	24006	28597	36744	50236	46151
30	1122.7	863.0	22344	31904	18559	17418	19410
57	1894.5	903.1	20142	14044	13687	16071	17075

All the spots are sorted on the abundances of the control proteomic map.



**Table 18** Variables used to build the Hasse matrix

#	Rank ID	Control	PFOA	PFDA	Clofibrate	DEHP
22	<b>29</b>	1.000	0.592	0.678	0.985	0.045
52	<b>28</b>	0.931	0.515	0.330	0.461	0.621
20	<b>27</b>	0.841	1.000	1.000	1.000	1.000
...	...	...	...	...	...	...
...	...	...	...	...	...	...
57	<b>1</b>	0.147	0.073	0.062	0.097	0.095

For each proteomic map, the rank ID variable (in bold) is used (for all the maps) together with the variable of the scaled abundances of the corresponding chemical.

proteomic map of the reference cell (control) are ordered in decreasing order. For the reference map, since the two scoring variables (ranking and property variables) give the same ranking, the Hasse matrix is constituted only by  $\pm 1$  values, corresponding to a total ranking.

Although for the considered data all the 29 spots have abundances different from zero, the proposed algorithm can also be easily applied when in a proteomic map new spots appear or spots disappear with respect to the control map. In this case, new spots with zero abundance can be added correspondingly in order to obtain equal length vectors.

**3.5.1 Sensitivity analysis**

In order to check the sensitivity of the proposed approach, a preliminary calculation has been performed producing some artificial proteomic maps, which differ systematically from the control (Table 19).

Following the strategy proposed by Randic et al. (2001b), the abundance of the third spot corresponding to the coordinates (12, 10) is modified iteratively by subtracting 7 from the initial abundance (136.7), obtaining other four different proteomic maps. The four abundance values of these modified proteomic maps (1–4) are 129.7, 122.7, 115.7 and 108.7.

The Hasse distances calculated between the control map and the four modified maps for weights 0, 0.5 and 1 are reported in Table 20. All the distances are presented as percentages, i.e. multiplied by 100. When only the structure of the Hasse matrix is considered ( $w = 0$ ), i.e. the diagonal elements are not taken into account, the distance between the control and the first modified map is equal to 0, because the new value of the spot (129.7) does not modify the ranking of the abundances; in the three other cases, the ranking is modified in increasing manner (1, 2 and 5 positions, respectively) and the distances reflect these modified rankings.

In the case of  $w = 0.5$ , the distances are calculated by taking into account both the Hasse matrix off-diagonal terms and the diagonal terms. In this case, the first modified map also shows a distance greater than 0 from the control. In the last case ( $w = 1$ ), only the diagonal contributions of the Hasse matrix are considered, and the distances from the control differ in a uniform way (0.243).

**Table 19** Coordinates and abundances of 20 proteins of a proteomic map

<i>n</i>	Rank ID	<i>x</i>	<i>y</i>	<i>A</i>
1	20	21	23	144.4
2	19	28	9	143.6
<b>3</b>	<b>18</b>	<b>12</b>	<b>10</b>	<b>136.7</b>
4	17	22	9	125.3
5	16	27	12	118.6
6	15	15	8	114.9
7	14	13	14	112.3
8	13	29	8	108.9
9	12	14	11	98.2
10	11	26	13	94.1
11	10	25	4	93.6
12	9	16	6	90.0
13	8	30	8	86.7
14	7	21	8	84.8
15	6	6	7	82.5
16	5	29	19	82.0
17	4	20	9	80.0
18	3	28	8	79.8
19	2	23	10	72.8
20	1	11	9	72.2

The third entry (12,10 – spot 3, in bold) is taken as reference for artificial modifications, subtracting 7 in each step.

**Table 20** Distances (as percentages) of four artificially modified proteomic maps from the reference one (spot 3), for three different weights

<i>w</i>	1	2	3	4
0	0.000	0.526	1.053	2.632
0.5	0.121	0.506	0.890	1.801
1	0.242	0.485	0.727	0.970

Thus, it can be observed that when the off-diagonal terms are taken into account ( $w=0$  and  $0.5$ ), the distances from the control increase non-linearly, due to the relevant role of the global ordering relationships of the abundances. Moreover, the Hasse distance not only appears sensitive to small changes in abundances but also shows robustness with respect to changes that do not modify the ranking of the abundances.

3.5.2 Comparison of proteomic maps

The comparison of the five proteomic maps (Table 17) has been performed using the Hasse distance approach, with the weights 0 and 0.5. All the distances are

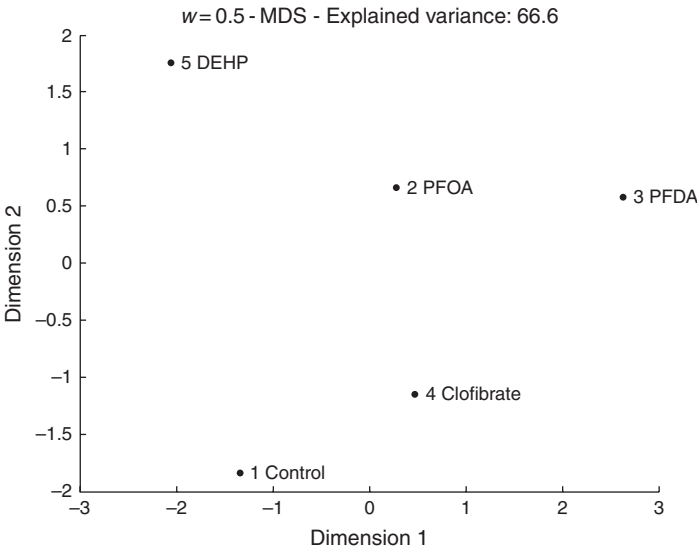
**Table 21** Distances calculated for  $w = 0$  (upper matrix) and  $w = 0.5$  (lower matrix, in italics)

	1 Control	2 PFOA	3 PFDA	4 Clofibrate	5 DEHP
Control	0	<b>14.778</b>	24.384	<b>14.532</b>	<b>18.719</b>
PFOA	<i>14.737</i>	0	18.966	<b>16.010</b>	19.212
PFDA	<i>22.291</i>	<b>14.241</b>	0	<b>18.227</b>	30.296
Clofibrate	<b>12.964</b>	<b>12.168</b>	<b>14.721</b>	0	24.384
DEHP	<i>15.931</i>	<b>14.627</b>	24.325	<i>17.744</i>	0

The five smallest distances are shown in bold characters.

represented as percentages. The results are collected in Table 21: the upper matrix contains the distances calculated using  $w = 0$ , while the lower matrix contains those calculated using  $w = 0.5$  (in italics). The five smallest distances are highlighted in bold for the two cases. From the upper part of Table 21 it can be noted that clofibrate and PFOA show distances from the control lower than 15%, while PFDA is the most dissimilar (i.e. the most perturbed) from the control (24.4%).

Considering the Hasse distance and a weight equal to 0.5, the most similar proteomic maps are PFOA and clofibrate (12.2%); with respect to the control map, clofibrate is the most similar (13.0%), while PFDA remains the most dissimilar from the control (22.3%). In Figure 5, for the case  $w = 0.5$ , a graphical view



**Figure 5** Multidimensional scaling projection of the five proteomic maps by standardised Hasse distances with weight equal to 0.5.

**Table 22** Euclidean distances calculated by using the leading eigenvalues (upper matrix) and the first five eigenvalues (lower matrix, in italics)

	1 Control	2 PFOA	3 PFDA	4 Clofibrate	5 DEHP
Control	0	4.024	6.422	3.804	4.312
PFOA	<i>5.886</i>	0	<b>2.398</b>	<b>0.220</b>	<b>0.288</b>
PFDA	<i>8.639</i>	<b>3.020</b>	0	2.619	<b>2.111</b>
Clofibrate	<i>5.626</i>	<b>0.618</b>	3.156	0	<b>0.508</b>
DEHP	<i>5.844</i>	<b>0.614</b>	<b>3.136</b>	<b>0.913</b>	0

The five smallest distances are shown in bold characters.

of the inter-distances among the five proteomic maps is shown by means of the metric multidimensional scaling technique. The comparison among the five proteomic maps has been also performed using the matrix invariants (the eigenvalues) obtained from the augmented Hasse matrix, using the leading eigenvalues (EED1), the first 5 (EED5), 10 (EED10) and 20 (EED20) eigenvalues, respectively. In [Table 22](#), the Euclidean distances calculated by using the leading eigenvalues (upper matrix) and the first five eigenvalues (lower matrix, in italics) are collected.

Unlike the previous case, the most similar proteomic maps are PFOA, clofibrate and DEHP among themselves. The most similar to the control map is clofibrate (3.804), while the most perturbed appears PFDA (6.422). Taking into account the first five eigenvalues, the most similar proteomic maps remain the same as those calculated using the leading eigenvalues; however, in this case, PFOA, clofibrate and DEHP appear at the same distance from the control, while PFDA remains the most dissimilar (8.639). Similar results have been obtained taking into account 10 and 20 eigenvalues.

In [Table 23](#), all the obtained results are compared with other results from literature ([Randic et al., 2001a](#); Randic and Basak, 2002; [Randic et al., 2004](#); [Randic et al., 2005](#)). Cases A–B are distances derived from five spots, while cases D–E correspond to distances derived from 20 spots; cases F and G correspond to distances derived from 30 spots; case H corresponds to distances derived from 250 spots; the matrix invariants used for the distance calculations have been the following: (A) Euclidean distance matrix, (B) zigzag path distance matrix, (C) Euclidean distance from 20 spot vectors, (D) Euclidean distance on leading eigenvalues from the distance adjacency matrix (Kronecker matrix product), (E) Euclidean distance on leading eigenvalues from the distance adjacency matrix (standard matrix product), (F) Euclidean distances calculated by six-component vectors derived from the relative abundances, (G) and (H) similarity measure based on the differences between the magnitudes of two-component vectors where each vector represents individual proteomic map. The two-component vectors are constructed as the average value of contributions of individual spots, the contributions of which are obtained considering two components, the first

**Table 23** Five most similar pairs (R1–R5) of the proteomic maps of [Table 17](#), obtained using different approaches

ID	Method	R1	R2	R3	R4	R5
1	$d_W(0)$	1-4	1-2	2-4	3-4	1-5
2	$d_W(0.5)$	2-4	1-4	2-3	2-5	3-4
3	$d_W(1)$	2-4	2-3	2-5	4-5	3-4
4	EED1	2-4	2-5	4-5	3-5	2-3
5	EED5	2-5	2-4	4-5	2-3	3-5
6	EED10	2-5	4-5	2-4	2-3	3-5
7	EED20	2-5	4-5	2-4	2-3	3-5
6	A	3-5	1-3	2-3	1-4	3-4
7	B	2-3	3-4	1-2	4-5	1-3
8	C	2-3	2-4	3-4	1-2	1-5
9	D	2-4	2-5	1-5	4-5	1-2
10	E	2-5	1-5	2-4	4-5	1-2
11	F	2-4	1-4	1-2	2-3	3-4
12	G	1-4	2-4	1-2	2-3	1-5
13	H	1-4	2-3	1-2	2-4	4-5

The first seven rows show results obtained in this work. A–B show results obtained using the first five spots. C–E show results obtained using the first 20 spots. F shows results obtained using 30 spots. G shows results obtained using 30 spots. H shows results obtained using 250 spots.

given by the abundance of the spot and the second one related to weighted adjacency matrix for the graph of the nearest neighbors.

The results obtained when only the ranking information is considered ( $w=0$ ) do not differ significantly from the results obtained for the case H: as it can be observed in [Table 23](#), both methods consider the pair 1–4 as the most similar and the pair 1–5 as the fifth similar pair. Furthermore, it can be observed that when the diagonal contribution is taken into account ( $w=0.5$  and 1), the similarities between pairs 2–4, 2–5 and 2–3 are enhanced with respect to the case  $w=0$ , thus giving results more similar to those obtained for C–E, where Euclidean distances on the spot abundances have been calculated. Finally, PFDA is the most dissimilar proteomic map (i.e. the most perturbed) from the control (pair 1–3) for the methods proposed in this work and the methods C–E, while DEHP is the most perturbed proteomic map (pair 1-5) for the cases A–B as well as for case F, where only five spots have been taken into account. Also in [Bajzer et al. \(2003\)](#), DEHP is the proteomic map evaluated as the most different from the control. In the case A, PFDA is evaluated as the most similar to the control map (pair 1–3).

In any case, the differences between the rankings obtained by the WSHD with weights 0 and 1 are remarkable, thus showing the different role of the ranking contribution with respect to the property contribution, i.e. the protein abundance in this case.

## 4. CONCLUSIONS

In general, the proposed similarity/diversity measure appears as a new approach to sequential data analysis, where useful information can be obtained by evaluating the ordering relationships between the sequence elements. Basically, the similarity/diversity between two sequences is obtained by the definition of a distance between the corresponding Hasse matrices; these distances have some useful properties and seem to show a high sensitivity to changes in structure sequences.

As described in the present chapter, this methodology can be extended to several kinds of sequential data (DNA sequences, proteomics maps, molecular descriptors, analytical data such as NMR and electronic nose signals) and shows the following advantages: (a) the Hasse matrices and the corresponding distances are calculated by a straightforward algorithm; (b) the distance is naturally standardised, allowing a natural interpretation of the obtained values; (c) the distances are able to take into account the whole structure of the ranking relationships of the sequences; (d) the distances can be obtained by an adaptive strategy (the weights) depending on the specific similarity/diversity study and (e) a simple rank correlation measure is derived, also taking into account incomparabilities among sequence elements. Moreover, the differences between the rankings obtained by the WSHD with weights 0 and 1 highlight the different role of the two distance contributions  $d_H$  and  $d_D$ , different aspects of the similarity/diversity being taken into account.

## REFERENCES

- Bajzer, Z., Randic, M., Plavsic, D., Basak, S.C. (2003). Novel map descriptors for characterization of toxic effects in proteomics maps, *J. Mol. Graphics Modell.* 22, 1–9.
- Ballabio, D., Cosio, M.S., Mannino, S., Todeschini, R. (2006). A chemometric approach based on a novel similarity/diversity measure for the characterisation and selection of electronic nose sensors, *Anal. Chim. Acta* 578, 170–177.
- Bartlett, P.N., Elliot, J.M., Gardner, J.W. (1997). Electronic nose and their application in the food industry, *Food Technol.* 51, 44–48.
- Brezmes, J., Llobet, E., Vilanova, X., Ortis, J., Saiz, G., Correig, X. (2001). Correlation between electronic nose signals and fruit quality indicators on shelf-life measurements with pink lady apples. *Sens. Actuators B* 80, 41–50.
- Brüggemann, R., Bartel, H.G. (1999). A theoretical concept to rank environmentally significant chemicals, *J. Chem. Inf. Comput. Sci.* 39, 211–217.
- Brüggemann, R., Franck, H., Kerber, A. (2004). Proceedings of the conference “Partial Orders in Environmental Sciences and Chemistry”, *MATCH* 54, 485–689.
- Buratti, S., Ballabio, D., Benedetti, S., Cosio, M.S. (2006). Prediction of Italian red wine sensorial descriptors from electronic nose, electronic tongue and spectrophotometric measurements by means of Genetic Algorithm regression models, *Food Chem.* 100, 211–218.
- Buratti, S., Benedetti, S., Scampicchio, M., Pangerod, E.C. (2004). Characterization and classification of Italian Barbera wines by using an electronic nose and an amperometric electronic tongue, *Anal. Chim. Acta* 525, 133–139.
- Cerrato Oliveros, C., Pérez Pavón, J.L., García Pinto, C., Fernández Laespada, E., Moreno Cordero, B., Forina, M. (2002). Electronic nose based on metal oxide semiconductor sensors as a fast alternative for the detection of adulteration of virgin olive oils, *Anal. Chim. Acta* 459, 219–228.

- Cosio, M.S., Ballabio, D., Benedetti, S., Gigliotti, C. (2006). Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks, *Anal. Chim. Acta* 567, 202–210.
- Garcia-González, D.L., Aparicio, R. (2003). Virgin olive oil quality classification combining neural network and MOS sensors, *J. Agric. Food Chem.* 51, 3515–3519.
- Gardner, J.W., Bartlett, P.N. (1993). Brief history of electronic nose, *Sens. Actuators B* 18, 211–217.
- Guadarrama, A., Rodríguez-Méndez, M.L., de Saja, J.A., Ríos, J.L., Olías, J.M. (2000). Array of sensors based on conducting polymers for the quality control of the aroma of the virgin olive oil, *Sens. Actuators B* 69, 276–282.
- Halfon, E., Reggiani, M.G. (1986). On ranking chemicals for environmental hazard, *Environ. Sci. Technol.* 20, 1173–1179.
- Hemmer, M.C., Steinhauer, V., Gasteiger, J. (1999). Deriving the 3D structure of organic molecules from their infrared spectra, *Vib. Spectro.* 19, 151–164.
- Llobet, E., Hines, E.L., Gardner, J.W., Franco, S. (1999). Non-destructive banana ripeness determination using a neural network-based electronic nose, *Meas. Sci. Technol.* 10, 538–548.
- Mauri, A., Consonni, V., Pavan, M., Todeschini, R. (2006). DRAGON software: an easy approach to molecular descriptor calculations, *MATCH* 56, 237–248.
- Pavan, M., Todeschini, R. (2004). New indices for analysing partial ranking diagrams, *Anal. Chim. Acta* 515, 167–181.
- Randic, M. (2001). On graphical and numerical characterization of proteomics maps, *J. Chem. Inf. Comput. Sci.* 41, 1330–1338.
- Randic, M., Basak, S.C. (2002). A Comparative study of proteomics maps using graph theoretical biodescriptors, *J. Chem. Inf. Comput. Sci.* 42, 983–992.
- Randic, M., Lers, N., Plavsic, D., Basak, S.C. (2004). On Invariants of a 2-D Proteome Map Derived from Neighborhood Graphs, *J. Proteome Res.* 3, 778–785.
- Randic, M., Witzmann, F.A., Kodali, V., Basak, S.C. (2005). On the dependence of a characterization of proteomics maps on the number of protein spots considered, *J. Chem. Inf. Comput. Sci.* 46, 116–122.
- Randic, M., Witzmann, F.A., Vracko, M., Basak, S.C. (2001a). On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: applications to peroxisome proliferators, *Med. Chem. Res.* 10, 456–479.
- Randic, M., Zupan, J., Novic, M. (2001b). On 3-D graphical representation of proteomics maps and their numerical characterization, *J. Chem. Inf. Comput. Sci.* 41, 1339–1344.
- Skov, T., Bro, R. (2005). A new approach for modelling sensor based data, *Sens. Actuators B* 106, 719–729.
- Taleta srl (2006). DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.4. <http://www.taleta.mi.it>.
- Todeschini, R., Consonni, V. (2000). *Handbook of Molecular Descriptors*, Wiley – VCH, Wei.
- Todeschini, R., Consonni, V., Mauri, A., Ballabio, D. (2006). Characterization of DNA primary sequences by a new similarity/diversity measure based on the partial ordering, *J. Chem. Inf. Modell.* 46, 1905–1911.
- Vinaixa, M., Marín, S., Brezmes, J., Llobet, E., Vilanova, X., Correig, X., Ramos, A., Sanchis, V. (2004). Early detection of fungal growth in bakery products by use of an electronic nose based on mass spectrometry, *J. Agric. Food Chem.* 52, 6068–6074.
- Vinaixa, M., Vergara, A., Duran, C., Llobet, E., Badia, C., Brezmes, J., Vilanova, X., Correig, X. (2005). Fast detection of rancidity in potato crisps using e-noses based on mass spectrometry or gas sensors, *Sens. Actuators B* 106, 67–75.
- Witzmann, F.A. (2002). Proteomics applications in toxicology. In: *Comprehensive Toxicology, Cellular and Molecular Toxicology* (Vanden Heuvel, J.P., Greenbee, W.F., Perdew, G.H., Mattes, W.B., eds), Elsevier, New York, pp. 539–558.

# The Interplay between Partial-Order Ranking and Quantitative Structure–Activity Relationships

L. Carlsen

---

Contents	1. Introduction	139
	2. Methodology	140
	2.1 Partial-order ranking	140
	2.2 Linear extensions	142
	2.3 Average ranks	142
	2.4 Quantitative Structure–Activity Relationships	142
	2.5 “Noise-deficient” Quantitative Structure–Activity Relationships	143
	3. Results and Discussion	144
	3.1 Risk assessment	144
	3.2 Safer alternatives	149
	3.3 Inverse Quantitative Structure–Activity Relationships	151
	4. Conclusions and Outlook	153
	References	154
	Abbreviations	157

---

## 1. INTRODUCTION

The assessment of chemicals probably constitutes one of the most serious environmental challenges today as the number of chemical substances that are in use and that constitute a potential risk to the environment exceeds 100.000 (EEA, 1998). It is immediately obvious that an assessment of all chemicals based on an experimental approach alone is not practically possible. However, information concerning the fate and effects of these substances in the environment as well as the possible influence on human health is crucial. Thus, it is expected that model-generated data, as obtained through Quantitative Structure–Activity



Relationship (QSAR) modelling, will play an increased role in the future assessment work. Thus, within the proposed new risk assessment scheme in Europe, REACH (EC, 2003; EP, 2005), a widespread use of QSAR modelling to retrieve physico-chemical and toxicological data is foreseen.

Obviously, an assessment of a single chemical can be made based on modelled or experimentally derived data. However, this would typically give rise to an assessment of the chemical based on a single criterion only. Obviously, in many cases it is desirable to include more than one criterion simultaneously in the assessment. Furthermore, it might be desirable to extend the assessment to include comparison to other chemicals, e.g., to disclose those substances that on a cumulative basis appear to be environmentally more problematic or alternatively to select chemicals that from an overall judgement may mimic highly toxic substances, however, without possessing the high toxicity or even to disclose substances with specific characteristics. In this respect, partial-order ranking (POR) appears as a highly attractive tool (Brüggemann and Carlsen, 2006).

In the present chapter, the advantageous interplay between POR and QSAR will be illustrated by selected examples from recent studies, including risk assessment (Carlsen, 2006), selecting safer alternatives (Carlsen, 2004; Carlsen, 2005a, b) and as a tool in the process of suggesting new substances with specific characteristics, i.e., the so-called inverse QSAR approach (Brüggemann et al., 2001b).

It is clear that POR will not directly give rise to an absolute ranking of the substances under investigation. Thus, to further elucidate the mutual ranking of the chemicals under investigation, linear extensions (LEs) may be brought into play, leading to the most probable linear (absolute) rank of the substances under investigation (Fishburn, 1974; Graham, 1982; Davey and Priestley, 1990; Brüggemann et al., 2001a). Further the concept of average rank (Brüggemann et al., 2004) will be discussed.

Partial-order techniques have been applied directly for QSAR modelling as illustrated by the use of the QSAR descriptors as input to the ranking (Carlsen et al., 2001; Carlsen et al., 2002). However, this will not be dealt with in the present paper.

## 2. METHODOLOGY

The successful interplay between POR and QSARs obviously depends on the quality of the single techniques. In the following, POR including LEs and average rank as well as QSARs, as applied for the studies included in the present paper, will be shortly presented.

### 2.1 Partial-order ranking

The theory of POR is presented elsewhere (Davey and Priestley, 1990; Brüggemann and Carlsen, 2006). In brief, POR is a simple principle, which a priori includes " $\leq$ " as the only mathematical relation. If a system is considered, which

can be described by a series of descriptors  $p_i$ , a given compound A, characterized by the descriptors  $p_i(A)$ , can be compared to another compound B, characterized by the descriptors  $p_i(B)$ , through comparison of the single descriptors. Thus, compound A will be ranked higher than compound B, i.e.,  $B < A$ , if at least one descriptor for A is higher than the corresponding descriptor for B and no descriptor for A is lower than the corresponding descriptor for B. If, in contrast,  $p_i(A) > p_i(B)$  for descriptor  $i$  and  $p_j(A) < p_j(B)$  for descriptor  $j$ , A and B will be denoted incomparable. Obviously, if all descriptors for A are equal to the corresponding descriptors for B, i.e.,  $p_i(B) = p_i(A)$  for all  $i$ , the two compounds will have identical rank and will be considered as equivalent, i.e.,  $A = B$ .

In mathematical terms this can be expressed as

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

It further follows that if  $A \leq B$  and  $B \leq C$ , then  $A \leq C$ . If no rank can be established between A and B, then these compounds are denoted as incomparable, i.e., they cannot be assigned a mutual order.

In POR—in contrast to standard multidimensional statistical analysis—neither any assumptions about linearity nor any assumptions about distribution properties are made. In this way, the POR can be considered as a non-parametric method. Thus, there is no preference among the descriptors. However, due to the simple mathematics outlined above, it is obvious that the method a priori is rather sensitive to noise, since even minor fluctuations in the descriptor values may lead to non-comparability or reversed ordering. To overcome this problem, the concept of “noise-deficient” QSARs for generation of descriptors was introduced, *vide infra* (see also (Carlsen, 2006)). An alternative approach to how to handle loss of information by using an ordinal instead of a matrix can be found in Pavan et al. (2006).

In contrast to standard multidimensional statistical analysis, POR does not assume linearity of the single descriptor values; nor are any assumptions about distribution properties made.

A main point is that all descriptors have the same direction, i.e., “high” and “low”. As a consequence of this, it may be necessary to multiply some descriptors by  $-1$  to achieve identical directions. Bioaccumulation and toxicity can be mentioned as examples. In the case of bioaccumulation, the higher the number, the higher the substance tends to bioaccumulate and thus the more problematic the substance, whereas in the case of toxicity, the lower the figure, the more toxic the substance. Thus, in order to secure identical directions of the two descriptors, one of them, e.g., the toxicity figures, has to be multiplied by  $-1$ . Consequently, in the case of both bioaccumulation and toxicity, higher figures will correspond to more hazardous compounds.

The graphical representation of the partial ordering is often given in the so-called Hasse diagram (Hasse, 1952; Halfon and Reggiani, 1986; Brüggemann et al., 1995; Brüggemann et al., 2001a). In practice, the PORs are performed using the WHasse software (Brüggemann et al., 1995).

## 2.2 Linear extensions

The number of incomparable elements in the partial ordering may obviously constitute a limitation on the attempt to rank, e.g., a series of chemical substances based on their potential environmental or human health hazard. To a certain extent, this problem can be remedied through the application of the so-called linear extensions, LEs of the POR (Fishburn, 1974; Graham, 1982). An LE is a total order where all comparabilities of the partial order are reproduced (Davey and Priestley, 1990; Brüggemann et al., 2001a). Due to the incomparisons in the POR, a number of possible LEs correspond to one partial order. If all possible LEs are found, a ranking probability can be calculated, i.e., based on the LEs, the probability that a certain compound has a certain absolute rank can be derived. If all possible LEs are found, it is possible to calculate the average ranks of the single elements in a partially ordered set POSET (Winkler, 1982, 1983).

## 2.3 Average ranks

The average rank is simply the average of the ranks in all the LEs. On this basis, the most probable rank for each element can be obtained leading to the most probably linear rank of the substances studied.

The average rank of the single compounds in the Hasse diagram is obtained by applying the simple empirical relation recently reported by Brüggemann et al. (2004). The average rank of a specific compound,  $c_i$ , can be obtained by the simple relation

$$Rk_{av} = (N + 1) - (S + 1) \times (N + 1) / (N + 1 - U) \quad (2)$$

where  $N$  is the number of elements in the diagram,  $S$  the number of successors, i.e., comparable element located below, to  $c_i$  and  $U$  the number of elements being incomparable to  $c_i$  (Brüggemann et al., 2004). It should be noted that in the ranking, according to Eq. (2), the lower the number, the higher the levels. Thus, the highest level will be "1". This is reversed compared to the original approach (Brüggemann et al., 2004).

## 2.4 Quantitative Structure–Activity Relationships

The basic concept of QSARs in its simplest form can be expressed as the development of correlations between a given activity or property (end point), i.e., physico-chemical or biological,  $P$ , and a set of parameters (descriptors),  $D_i$ , that are inherent characteristics of the compounds under investigation

$$P = f(D_i) \quad (3)$$

In general, models that are able to describe/calculate key properties of chemical compounds include three types of inherent characteristics of the molecule, i.e.

structural, electronic and hydrophobic characteristics. Different models may bring few or many of these descriptors into play. Thus, Eq (3) can be rewritten as

$$P = f(D_{\text{structural}}, D_{\text{electronic}}, D_{\text{hydrophobic}}, D_x) + e \quad (4)$$

The descriptors reflecting structural characteristics may, e.g., be elements of the actual composition and the three-dimensional structure of the molecule, whereas descriptors reflecting the electronic characteristics may, e.g., be charge densities and dipole moment. The descriptors reflecting the hydrophobic characteristics are related to the distribution of the compound between a biological, hydrophobic phase and an aqueous phase. The fourth type of characteristics,  $D_x$ , accounts for possible underlying characteristics that may be known or unknown, such as environmental or experimental parameters like temperature, ionic strength and microbial activity. In general, the data may be associated with a certain amount of systematic and non-quantifiable variability in combination with uncertainties. These unknown variations are expressed as “noise”, accounted for by the parameter  $e$ , i.e., the variation in the property that cannot be explained by the model.

In principle, all types of QSAR models can be used to generate descriptors for subsequent use in POR. However, as POR, due to its inherent nature focusing only on the relation “ $\leq$ ” (vide infra), may be hampered by random fluctuations in the descriptors, the so-called “noise-deficient” QSARs (Carlsen, 2004, 2005a, 2005b) can be advantageously applied (vide infra).

## 2.5 “Noise-deficient” Quantitative Structure–Activity Relationships

In the studies referred to in the present chapter, the end-points are generated through QSAR modelling all obtained applying the EPI Suite as the primary tool. The EPI Suite is a collection of QSAR models for physical, chemical and toxicity end point developed by the EPA’s office of Pollution Prevention Toxics and Syracuse Research Corporation (EPA, 2008).

Both experimental and EPI Suite-generated data are typically defective. “Noise-deficient” QSARs eliminate the noise associated with the data. Thus, the subsequent rankings will not be hampered by incidental variations either in the experimental data or in the EPI-generated data.

To generate new, strictly linear “noise-deficient” QSAR models, EPI-generated values for the selected end points such as solubility (log *Sol*), octanol–water partitioning (log *K*<sub>OW</sub>), vapour pressure (log *VP*) and Henry’s law constant (log *HLC*) are further treated by estimating the relationships between the EPI-generated data and available experimental data as being available in the database associated with the EPI Suite (EPA, 2008). Thus, for a series of experimentally well-characterized compounds in the training set, the general formula for the single end points,  $D_i$ , to be used is given as

$$D_i = a_i \times D_{\text{EPI}} + b_i \quad (5)$$

where  $D_{\text{EPI}}$  is the EPI-generated descriptor value and  $a_i$  and  $b_i$  are constants (see Carlsen 2004, 2005a, 2005b).

### 3. RESULTS AND DISCUSSION

In the following sections, three examples of the advantageous interplay between POR and QSAR will be elucidated, i.e., risk assessment, selecting safer alternatives and inverse QSARs.

#### 3.1 Risk assessment

Today the risk assessment within the European Union is performed according to the Technical Guidance Document (TGD) (EC, 1996). The requirements outlined in the TGD have been made operational in the designated software, i.e., the European Union System for the Evaluation of Substances (EUSES) (ECB, 2005).

However, a significant drawback of the EUSES tool is the substantial requirements for input data including 466 input parameters, 961 connections between parameters and 132 defaults (Verdonck et al., 2005). In order to circumvent this, a simple rule-based screening environmental risk assessment tool derived from EUSES has been proposed (Verdonck et al., 2005), based on the information on four parameters only, i.e., octanol–water partitioning, vapour pressure, biodegradability and ecotoxicity, the latter being expressed by the *PNEC* value, in addition to information on the actual release scenario and production tonnage. The method that allows for an immediate identification of substances that should undergo a more thorough risk assessment can be regarded as a simple “yes/no” approach.

However, in order to further elaborate on this simple “yes/no” approach, possibly to obtain details regarding the mutual risk posed by a selected group of substances, the interplay between QSAR modelling and POR may be advantageously brought into play by ranking these substances, which at the same time will disclose the environmentally more hazardous chemicals.

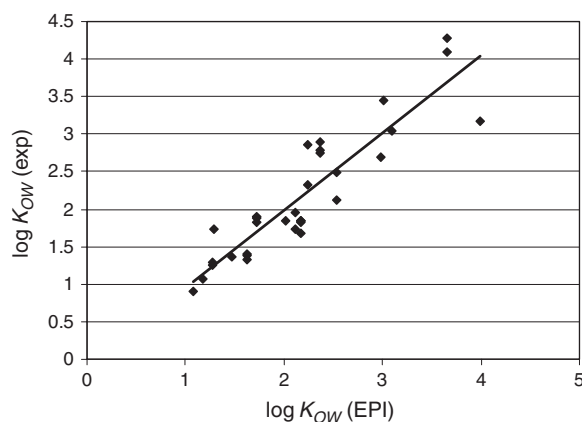
Obviously the rule-based approach (Verdonck et al., 2005) may be used to set up a linear rank of the compounds investigated. In this context, it is worthwhile to note that the rule-based approach applies a look-up table approach based on ranges of log *K*<sub>OW</sub> and log *VP* values and not exact values. Thus, it may well be defective in the ranking of the selected substances according to their potential environmental hazard. In contrast, as will be elucidated in the following, POR (Davey and Priestley, 1990) in combination with LEs (Fishburn, 1974; Graham, 1982) appears as the obvious choice to remedy the problem.

To illustrate the advantageous interplay between the ranking methodology and QSAR modelling, a set of 45 substituted anilines was studied (Carlsen, 2006). For this group of substances, experimental data concerning octanol–water partitioning and vapour pressure are lacking for part of the compounds. The missing data consequently were obtained using the EPI Suite (EPA, 2008) and “noise-deficient” QSARs (Carlsen, 2006). The ready/non-ready biodegradability was derived applying the BioWin software, which is an integrated part of the EPI Suite (EPA, 2008), adopting the non-linear biodegradation probability program (BDP2). Thus, a  $BDP2 \geq 0.5$  indicates that the compound is “readily

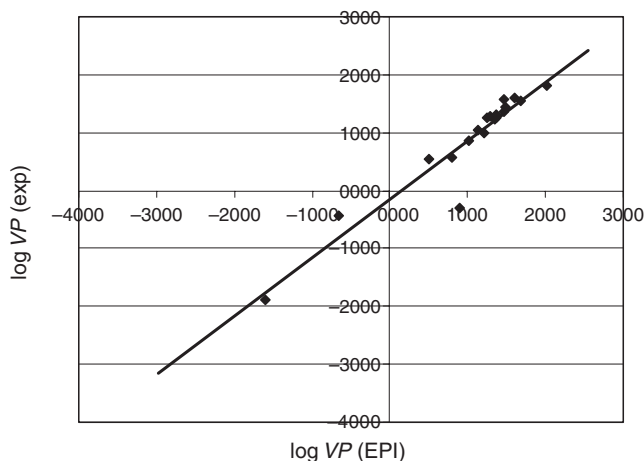
biodegradable”, whereas a  $BDP2 < 0.5$  indicates that the compound is “non-biodegradable” (EPA, 2008). As a measure of the toxicity of the anilines towards aquatic organisms, the population growth impairment of *Tetrahymena pyriformis* (Schultz, 1997) was adopted (Carlsen, 2006).

The generation of “noise-deficient” QSARs can be illustrated by the models derived for  $\log K_{OW}$  (Figure 1) and  $\log VP$  (Figure 2).

The readiness of biodegradation was established through the  $BDP2$  values obtained from the BioWin module of the EPI Suite, which was subsequently transformed into a parameter suitable for the subsequent POR (Carlsen, 2006).



**Figure 1** “Noise-deficient” QSAR for  $\log K_{OW}$  ( $\log K_{OW} = (1.050 \pm 0.090) \cdot \log K_{OW}(EPI) - (0.114 \pm 0.204)$ ;  $n = 31$ ,  $s = 0.363$ ,  $r^2 = 0.824$ ,  $F = 135.5$ ).

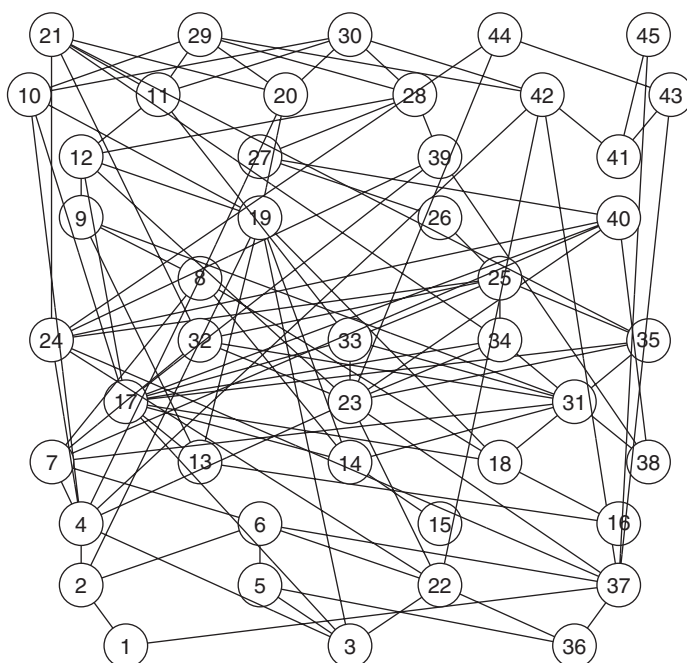


**Figure 2** “Noise-deficient” QSAR for  $\log VP$  ( $\log VP = (1.010 \pm 0.081) \cdot \log VP(EPI) - (0.146 \pm 0.107)$ ;  $n = 19$ ,  $s = 0.291$ ,  $r^2 = 0.902$ ,  $F = 156.0$ ).

The population growth impairment of *T. pyriformis*, expressed as  $IGC_{50}$ , adopted from [Schultz and Netzeva \(2004\)](#) was used as a measure for aquatic toxicity; the corresponding  $PNEC$  values eventually used as a ranking parameter was adopted as the  $IGC_{50}/1000$  values.

In [Figure 3](#), the resulting POR of 45 anilines is depicted. It should be noted that both the descriptors  $\log VP$  and  $PNEC$  have been multiplied by  $-1$  before applying them as descriptors in the POR (cf. the above discussion). Compounds exhibiting the highest environmental impact are given rank 1, the next highest given rank 2, etc. Thus, in the Hasse diagram depicted in [Figure 3](#), the environmentally more hazardous anilines are found in the top of the diagram.

It is immediately noted that the diagram consists of 11 levels. The top level, comprising five anilines, i.e., 2,6-di-iso-propyl aniline, 2,3,5,6-tetrachloro aniline, 2,3,4,5-tetrachloro aniline, 2,4-dinitro aniline and 2,4-dinitro aniline, respectively, comprises environmentally more hazardous species, whereas the bottom level, comprising three anilines, i.e., aniline, 3-methyl aniline and 2-fluoro aniline, comprises the less hazardous species. Thus, in terms of risk assessment, the 11 levels are a priori regarded as a classification of 45 anilines in 11 classes according to their potential environmental impact, simultaneously taking into account water-octanol partitioning, vapour pressure, biodegradability and toxicity ([Carlsen, 2006](#)) and assuming a production/use of 1 tonne in all cases.



**Figure 3** Hasse diagram of 45 anilines based on the four descriptors,  $\log K_{OW}$ ,  $\log VP$ ,  $BDP2$  and  $PNEC$ . The single compounds are identified through their ID (cf. Annex).

Often it appears appropriate to further qualify the POR by estimating the average rank of the single substances being studied, thus approaching the most probable linear ranking of the compounds.

In Table 1, the linear ranking of 45 anilines is disclosed according to their potential environmental impact. The individual ranks are estimated using random linear extensions (*RLE* rank) (Sørensen et al., 2001; Lerche et al., 2003) and the empirical formula (Eq. (2);  $Rk_{av}$ ).

**Table 1** Linear ranks of the single anilines estimated using random linear extensions (*RLE* rank) and the empirical formula (Eq. (2);  $Rk_{av}$ ) (for ID, cf. Annex)

ID	<i>RLE</i> rank	ID	$Rk_{av}$
29	1.8	29	1.1
30	1.8	30	1.1
28	4.1	21	1.4
21	4.3	44	3.5
27	6.3	28	3.7
11	8.6	27	5.8
26	8.6	11	6.3
25	10.8	26	7.4
44	11.1	45	7.7
12	12.3	39	8.0
39	13.7	25	8.9
40	14.3	10	9.2
10	14.4	40	9.2
20	15.9	12	9.5
9	16.1	42	11.5
35	16.7	20	12.3
32	16.7	9	12.4
34	16.7	32	13.1
42	17.2	34	13.1
33	17.3	35	13.1
45	18.1	43	13.1
8	20.9	33	14.3
43	21.0	8	16.0
31	23.6	31	22.3
19	24.6	19	23.0
24	25.5	24	26.6
17	27.0	7	30.1
7	27.2	23	30.7
23	27.6	17	32.6
6	30.0	6	32.9
41	33.5	4	39.9
38	34.1	5	40.3
13	34.5	41	40.3

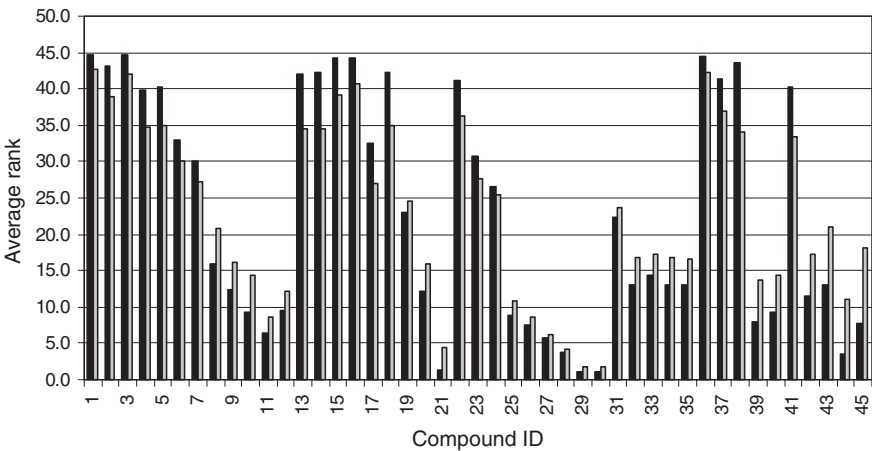


**Table 1** (Continued)

ID	<i>RLE</i> rank	ID	<i>Rk<sub>av</sub></i>
14	34.6	22	41.1
4	34.8	37	41.4
18	35.0	13	42.0
5	35.0	14	42.3
22	36.3	18	42.3
37	37.0	2	43.0
2	39.0	38	43.6
15	39.3	15	44.2
16	40.7	16	44.2
3	42.0	36	44.5
36	42.3	3	44.6
1	42.8	1	44.6

The minor deviations in the linear ranking of the anilines observed for the two methods applied are not surprising (Brüggemann et al., 2004). However, as expected, a convincing agreement between the *RLE* approach and the approach given in Eq. (2) in general prevails (Figure 4).

On the basis of the above combination of POR, supplemented with a study on LEs and average ranks, QSARs appear as an interesting decision support tool in relation to risk assessment of a group of chemicals.



**Figure 4** Average rank of 45 anilines calculated based on random linear extensions (*RLE*) (black) and formula (2) (grey).

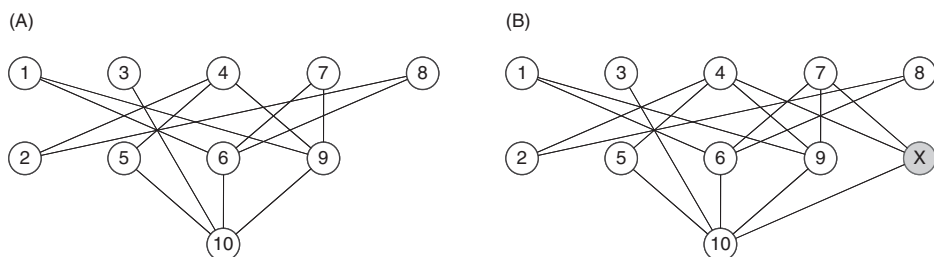
### 3.2 Safer alternatives

The basic idea of using POR for identifying safer alternatives is illustrated in Figure 5. Thus, let us assume that a suite of 10 compounds has to be evaluated and that the evaluation should be based on three pre-selected criteria, e.g., persistence, bioaccumulation and toxicity (cf. Carlsen, 2004). Let the resulting Hasse diagram be the one depicted in Figure 5A. To illustrate this, we apply three descriptors representing  $\log K_{OW}$ ,  $\log VP$ , and toxicity ( $PNEC$ ). It should be noted that both the descriptors  $\log VP$  and  $PNEC$  have been multiplied by  $-1$  before applying them as descriptors in the POR to secure identical ranking order by all used descriptors. Thus, for the single descriptors, the compounds are ranked 1, 2, 3, ..., 10 according to their environmental impact, the compound exhibiting the highest impact being given rank 1, the next highest given rank 2, etc. Consequently, following the POR, taking the three descriptors into account, the environmentally more hazardous compounds are found in the top of the Hasse diagrams.

Figure 5A discloses that the compounds in the top level, i.e., compounds 1, 3, 4, 7 and 8 on a cumulative basis including octanol–water partitioning, vapour pressure and toxicity, are the environmentally more problematic of the 10 compounds studied, whereas compound 10, which is found in the bottom of the diagram, is the less hazardous.

Subsequently, we can introduce compounds solely characterized by QSAR-derived data to give this new compound, X, an identity, e.g., in an attempt to elucidate the possible environmental impact of the latter. Adopting the above-discussed 10 compounds and the corresponding Hasse diagram (Figure 5A), we introduced the compound X. The revised Hasse diagram, now including 11 compounds, is shown in Figure 5B (cf. Carlsen, 2004). It is immediately disclosed that compound X has now obtained an identity in comparison to the originally well-characterized compounds, as it is evaluated as less environmentally harmful than compounds 4 and 7 but more harmful than compound 10. Thus, through the POR, the compound X has obtained an identity in the scenario with regard to its potential environmental impact.

The above analysis of 45 anilines may serve as an example in relation to the substitution of environmentally hazardous substances with less problematic



**Figure 5** Illustrative Hasse diagram of (A) 10 compounds using three descriptors and (B) the same 10 compounds plus one new compound X.

alternative selected, as the present example suggests within the same group of chemicals. Thus, from the Hasse diagram (Figure 3) it is suggested that 3-chloroaniline (ID 23) would be a safer alternative to 3-cyanoaniline (ID 41). However, an analysis based on average ranks discloses that this apparently is not the case. Thus, 3-chloroaniline (ID 23) is found at rank 27.6 based on *RLE* rank and 30.7 based on  $Rk_{av}$ , whereas 3-cyanoaniline (ID 41) is found at rank 33.5 based on *RLE* rank and 40.3 based on  $Rk_{av}$  (cf. Table 1).

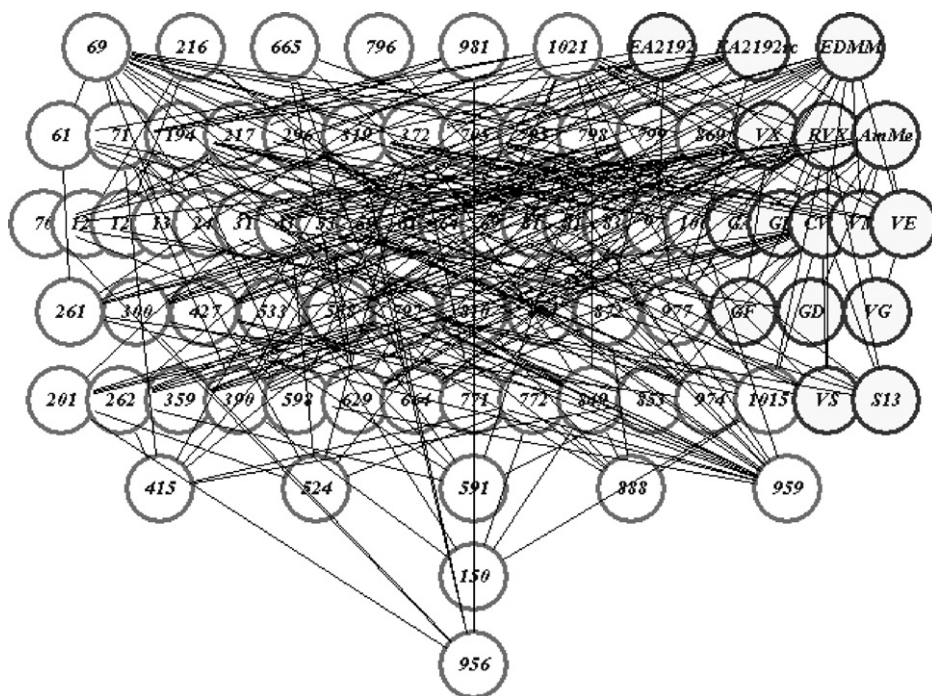
As a further example may serve the recent combined QSAR and POR studies of experimentally well-characterized organo phosphorus (OP) pesticides, the latter possibly serving as substitutes for highly toxic OP compounds like the chemical warfare nerve agents, i.e., the G-agents, like Tabun, Sarin and Soman, and V-agents, like VX (Carlsen, 2005a, b) in experimental studies as they from an overall viewpoint exhibit analogous environmental characteristics, however, without having the same extreme toxicity (Carlsen, 2004, 2005a, 2005b). Thus, 65 OPs together with 16 known or potential OP nerve agents focusing on the physico-chemical characteristics such as solubility (*Sol*), the biodegradation potential (BDP) and the Henry's Law Constants (*HLC*) were included in the study, i.e. descriptors representative of the aquatic persistence of the compounds (Carlsen, 2004, 2005a, 2005b).

The location of compounds at the same level in the Hasse diagram suggests that these compounds are close in their overall characteristics based on the set of descriptors used. However, this is not necessarily true. A further analysis appears to be necessary in order to eventually disclose how close these compounds actually are, as the incomparisons may be an Achilles' heel of the POR method. Hence, the concept of average rank (Lerche et al., 2002; Brüggemann et al., 2004; Carlsen, 2005a, b) was adopted. Thus, it is assumed that if the average ranks,  $Rk_{av}$ , of two compounds are close, the two compounds will on an average basis display similar characteristics as being determined by the set of descriptors applied.

Looking for safer alternative for, e.g., VX a Hasse diagram (Figure 6) disclosed VX being located at the same level as the pesticides anilofos, azinphos methyl, chlorfenvinphos, chlorpyrifos methyl, dialifos, dicrotophos, ditalimfos, monocrotophos, phosalone, phosmet, phosphamidon and pyraclofos in addition to the Russian version of VX (RVX) and the potential nerve agent amiton methyl.

The average ranks for the above-mentioned OPs together with minimum acute oral toxicity and acute percutaneous toxicity in both cases for rats (Carlsen, 2004) are given in Table 2.

It is immediately seen that although the above-mentioned compounds are found at the same level in the Hasse diagram, the true identity of the single compounds is disclosed only through the subsequent analysis of average linear rank. Thus, in the present case, it is obvious that VX ( $Rk_{av}=5.3$ ), which in the present context plays the role of being the unknown compound, may be associated with an identity comparable to phosphamidon ( $Rk_{av}=6.2$ ), the apparent closest counterpart. Thus, with regard to aqueous persistence, the combined QSAR and POR analysis indicates that VX and phosphamidon will display close to identical behaviour. It is immediately noted (cf. Table 2) that the acute oral toxicity of



**Figure 6** Hasse diagram displaying the aqueous persistence of 65 OP insecticides (blank circles) and 16 nerve agents (grey circles). The numbers correspond to the numbering of the OP insecticides in the FADINAP database.

phosphamidon is approx. 200 times lower than that of VX, whereas the acute percutaneous toxicity of phosphamidon appears to be nearly 4000 times lower than that of VX. Hence, phosphamidon appears, within the group of compounds included in the investigation, as the best possible low-toxic substitute for VX in experimental studies where aqueous persistence is a crucial parameter.

### 3.3 Inverse Quantitative Structure–Activity Relationships

Quantitative Structure–Activity Relationships are often based on standard multi-dimensional statistical analyses and apply sophisticated local and global molecular descriptors, assuming linearity as well as implying normal distribution behaviour of the latter.

The above approach using the POR–QSAR approach to retrieve safer alternatives may be further elaborated to an inverse QSAR approach, i.e., an approach to use the methodologies as a tool helpful to define a molecule or a class of molecules that fulfil prescribed properties.

It should be remembered that a pure QSAR approach typically will be based on highly sophisticated descriptors, the structure of potential candidates and thus the actual synthetic pathways possibly being hard to derive. It is, in contrast,

**Table 2** Average ranks for the aqueous persistence as determined by the solubility, the biodegradation potential and the Henry's law constants for a series of OP insecticides and VX (Carlsen, 2004)

Compound	Average rank ( $Rk_{av}$ )	Acute oral toxicity (mg/kg)	Acute percutaneous toxicity (mg/kg)
Anilofos	20.5	472	>2000
Azinphos methyl	25.6	4	220
Chlorfenvinphos	9.6	24	31
Chlorpyrifos methyl	18.2	1630	>3700
Dialifos	41	5	na
Dicrotophos	9.1	17	110
Ditalimfos	19.3	5660	>2000
Monocrotophos	10.3	20	112
Phosalone	35.1	135	>1500
Phosmet	21.9	160	na
Phosphamidon	6.2	17.9	374
Pyraclofos	18.9	237	>2000
VX	5.3	0.088	0.1

na: not available in the applied database.

appropriate to include simple descriptors that may form the basis for the synthesis recipe. Unfortunately, it may turn out that if descriptors simple enough to be used for defining synthesis recipes of chemicals are used, the accuracy of an arithmetic expression may fail. Recently Brüggemann et al. (2001b) suggested a method, based on the theory of POR, to find a qualitative basis for the relationship between such fair descriptors on the one side and a series of ecotoxicological properties on the other side. The obvious advantage in this context to adopt a POR approach is the fact that POR does not assume either linearity or normal distribution of the descriptors.

In the study of Brüggemann et al. (2001b), a series of synthesis-specific descriptors, i.e., simple structural descriptors such as the number of specific atoms and the number of specific bonds, were included in the analyses along with graph theoretical and quantum chemical descriptors. On this basis, a six-step procedure was developed to solve inverse QSAR problems.

1. Verify the correlation between ecotoxicological data and synthesis-specific descriptors and other possible descriptors to be included such as structural, hydrophobic and electronic descriptors (cf. Eq. (4)).
2. If high-positive or negative correlation between descriptor property on the one side and ecotoxicological value on the other side prevails, one or more descriptor values may have to be multiplied by  $-1$  before applying them as descriptors in the POR to secure identical ranking order by all used descriptors.

3. Calculate the Hasse diagrams for the adjusted data matrices as a result of step 4.
4. From the Hasse diagrams based on descriptors being highly correlated with the ecotoxicological end points, those that have the highest similarity to the Hasse diagram for the ecotoxicological end points are selected, the Hasse diagram with highest similarity serving as a “predictor” Hasse diagram for estimating “unknown” ecotoxicological end points of a new chemical, in the above example corresponding to the diagram displayed in [Figure 5A](#).
5. To the set of chemicals included in the “predictor” diagram ([Figure 5A](#)), a new chemical X is added, the latter being characterized by the corresponding descriptors values (cf. step 1), whereas the ecotoxicological data are unknown. This results in a new Hasse diagram (cf. [Figure 5B](#)).
6. From the upper neighbors, the minimum value for the single descriptors is retrieved and from the lower neighbor(s), the maximum value for single descriptors is retrieved. From these values, the ecotoxicological data of X are estimated (Brüggemann et al., 2001b).

#### 4. CONCLUSIONS AND OUTLOOK

The present paper has elucidated the potential advantageous use of the combination of “noise-deficient” QSAR modelling and POR as an effective tool in various areas of chemical sciences. Thus, the interplay between POR and QSAR constitutes an effective decision support tool to assess chemical substances, e.g., in relation to their potential environmental hazard. In that respect, the combined POR–QSAR approach offers the possibility to assess large numbers of chemicals taking several parameters such as persistence, bioaccumulation and toxicity simultaneously into account. Through this approach, environmentally more problematic substances that require immediate attention will unequivocally be disclosed.

The QSAR–POR system further appears as an appropriate tool to give specific molecules an identity in relation to others and thus serves as a support tool in the development of less hazardous and environmentally more friendly alternatives to proved harmful substances.

Further, how the POR–QSAR methodology may serve as a tool to select experimentally well-characterized, low-toxicity compounds as substitutes for highly toxic compounds has been elucidated, as illustrated in the present paper by the highly toxic nerve agents. Hence, this procedure allows that the environmental behaviour of the latter may be studied experimentally using compounds that from an overall viewpoint exhibit analogous environmental characteristics, however, without exhibiting extreme toxicity.

Finally, the potential use of the POR–QSAR approach as a rather strong tool to solve inverse QSAR problems, e.g., to develop suitable synthetic pathways for new substances, has been elucidated.

The combined approach of POR and QSAR modelling constitutes a technique for improving the screening activities of chemicals prior to a comprehensive and thus resource-consuming risk assessment as well as for identifying better

alternatives among chemicals. Research activities in this direction are foreseen to be of increasing importance in the future, e.g., in relation to the new chemical assessment scheme REACH.

## REFERENCES

- Brüggemann, R., Carlsen, L. (ed.) (2006). *Partial Order in Environmental Sciences and Chemistry*, Springer, Berlin.
- Brüggemann, R., Halfon, E., Bücherl, C. (1995). Theoretical base of the program "Hasse", GSF-Bericht 20/95, Neuherberg; The software may be obtained by contacting Dr. R. Brüggemann, Institute of Freshwater Ecology and Inland Fisheries, Berlin (brg@igb-berlin.de).
- Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., Steinberg, C.E.W. (2001a). 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.
- Brüggemann, R., Lerche, D., Sørensen, P.B., Carlsen, L. (2004). Estimation of average ranks by a local partial order model, *J. Chem. Inf. Comput. Sci.* 44, 618–625.
- Brüggemann, R., Pudenz, S., Carlsen, L., Sørensen, P.B., Thomsen, M., Mishra, R.K. (2001b). The use of Hasse diagrams as a potential approach for inverse QSAR, *SAR QSAR Environ. Res.* 11, 473–487.
- Carlsen, L. (2004). Giving molecules an identity. On the interplay between QSARs and Partial Order Ranking, *Molecules* 9, 1010–1018 <http://www.mdpi.org/molecules/papers/91201010.pdf>
- Carlsen, L. (2005a). A QSAR approach to physico-chemical data for organophosphates with special focus on known and potential nerve agents, internet electron, *J. Mol. Design* 4, 355–366 <http://www.biochempress.com>
- Carlsen, L. (2005b). Partial order ranking of organophosphates with special emphasis on nerve agents, *MATCH Commun. Math. Comput. Chem.* 54, 519–534.
- Carlsen, L. (2006a). A combined QSAR and partial order ranking approach to risk Assessment, *SAR QSAR Environ. Res.* 17, 133–146.
- Carlsen, L. (2006b). Interpolation schemes in QSAR. In: *Partial Order in Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L., eds), Springer, Berlin.
- Carlsen, L., Sørensen, P.B., Thomsen, M. (2001). Partial order ranking based QSAR's: Estimation of solubilities and octanol–water partitioning, *Chemosphere* 43, 295–302.
- Carlsen, L., Sørensen, P.B., Thomsen, M., Brüggemann, R. (2002). QSAR's based on partial order ranking, *SAR and QSAR Environ. Res.* 13, 153–165.
- Davey, B.A., Priestley, H.A. (1990). *Introduction to Lattices and Order*, Cambridge University Press, Cambridge.
- EC (1996). Technical Guidance Document in support of Directive 93/67/EEC on Risk Assessment for New and Notified Substances and Directive (EC) No. 1488/94 on Risk Assessment for Existing Substances, European Commission.
- EC (2003). Proposal for a regulation of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) [on Persistent Organic Pollutants] Proposal for a directive of the European Parliament and of the Council amending Council Directive 67/548/EEC in order to adapt it to Regulation (EC) of the European Parliament and of the Council concerning the registration, evaluation, authorisation and restriction of chemicals, COM 2003 644 Final. <http://europa.eu.int/eur-lex/en/com/pdf/2003/act0644en03/1.pdf>
- ECB (2005). The European Union System for the Evaluation of Substances (EUSES), European Chemical Bureau, <http://ecb.jrc.it/existing-chemicals/>
- EEA(1998). *Chemicals in the European Environment: Low Doses, High Stakes?* European Environment Agency, Copenhagen, p.33.
- EP (2005). European Parliament legislative resolution on the proposal for a regulation of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction

- of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) No .../... [on Persistent Organic Pollutants] (COM(2003)0644 – C5-0530/2003 – 2003/0256(COD)) (Codecision procedure: first reading). Available: [http://www.europarl.eu.int/omk/sipade3?L=EN&PUBREF=-//EP//TEXT+TA+20051117+ITEMS+DOC+XML+V0//EN&NAV=S&MODE=XML&LSTDOC=N&LEVEL=2&SAME\\_LEVEL=1](http://www.europarl.eu.int/omk/sipade3?L=EN&PUBREF=-//EP//TEXT+TA+20051117+ITEMS+DOC+XML+V0//EN&NAV=S&MODE=XML&LSTDOC=N&LEVEL=2&SAME_LEVEL=1).
- EPA (2008) Estimation Program Interface (EPI) Suite <http://www.epa.gov/oppt/exposure/pubs/episuite.htm> (accessed Oct. 2008).
- Fishburn, P.C. (1974). On the family of linear extensions of a partial order, *J. Comb. Theory* 17, 240–243.
- Graham, R.L. (1982). Linear extensions of partial orders and the FKG inequality In: *Ordered Sets* (Rival, I. ed), Reidel Publishing Company, Dordrecht, pp.213–236.
- Halfon, E., Reggiani, M.G. (1986). On the ranking of chemicals for environmental hazard, *Environ. Sci. Technol.* 20, 1173–1179.
- Hasse, H. (1952). Über die Klassenzahl abelscher Zahlkörper, Akademie Verlag, Berlin.
- Lerche, D., Brüggemann, R., Sørensen, P., Carlsen, L., Nielsen, O.J. (2002). A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances, *J. Chem. Inf. Comput. Sci.* 42, 1086–1098.
- Lerche, D., Sørensen, P.B., Brüggemann, R. (2003). Improved estimation of ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability, *J. Chem. Inf. Comput. Sci.* 43, 1471–1480.
- Pavan, M., Consonni, V., Gramatica, P., Todeschini, R. (2006). New QSAR modelling approach based on ranking models by genetic algorithms – Variable subset selection (GA-VSS). In: *Partial Order in Environmental Sciences and Chemistry* (Brüggemann, R., Carlsen, L. eds), Springer, Berlin.
- Schultz, T.W. (1997). TETRATOX: Tetrahymena pyriformis population growth impairment endpoint – A surrogate for fish lethality, *Toxicol. Methods* 7, 289–309.
- Schultz, T.W., Netzeva, T.I. (2004). Development and evaluation of QSARs for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity In: *Predicting Chemical Toxicity and Fate* (Cronin, M.T.D., Livingstone, D.J. eds), CRC Press, Boca, pp.265–284.
- Sørensen, P.B., Lerche, D., Carlsen, L., Brüggemann, R. (2001). Statistically approach for estimating the total set of linear orders – A possible way for analysing larger partial order set. In: *Proceeding Order Theoretical Tools in Environmental Science and Decision Systems* (Pudenz, et al., eds), Berichte des IGB, Berlin, pp.87–97.
- Verdonck, F.A.M., Boeije, G., Vandenberghe, V., Comber, M., de Wolf, W., Feijtel, T., Holt, M., Koch, K., Lecloux, A., Siebel-Sauer, A., Vanrolleghem, P.A. (2005). A rule-based screening environmental risk assessment tool derived from EUSES, *Chemosphere* 58, 1169–1176.
- Winkler, P.M. (1982). Average height in a partially ordered set, *Discrete Math.* 39, 337–341.
- Winkler, P.M. (1983). Correlation among partial orders, *Siam. J. Alg. Disc. Meth.* 4, 1–7.



Annex: Forty-five anilines included in the study (nd: noise-deficient)

CAS No	Name: aniline	ID	log $K_{ow}$ (nd)	log VP (nd) (Pa)	BDP2 <sup>a</sup>	PNEC ( $\mu\text{g/L}$ )
62-53-3	H	1	1.03	1.90	1	158.16
95-53-4	2-Methyl	2	1.59	1.55	1	154.89
108-44-1	3-Methyl	3	1.59	1.48	1	204.19
106-49-0	4-Methyl	4	1.59	1.33	1	120.24
578-54-4	2-Ethyl	5	2.09	1.34	2	201.11
587-02-0	3-Ethyl	6	2.09	1.24	2	129.85
589-16-2	4-Ethyl	7	2.09	1.16	2	113.09
643-28-7	2-Iso-Propyl	8	2.53	1.09	2	102.57
99-88-7	4-Iso-Propyl	9	2.53	0.99	2	81.47
104-13-2	4-Butyl	10	3.12	-0.02	1	12.70
30273-11-1	4- <i>sec</i> -Butyl	11	3.03	0.60	2	36.63
769-92-6	4- <i>tert</i> -Butyl	12	3.00	0.63	2	65.15
87-59-2	2,3-Dimethyl	13	2.15	1.07	1	326.16
95-68-1	2,4-Dimethyl	14	2.15	1.24	1	236.28
95-78-3	2,5-Dimethyl	15	2.15	1.24	1	259.08
87-62-7	2,6-Dimethyl	16	2.15	1.22	1	326.16
95-64-7	3,4-Dimethyl	17	2.15	0.66	1	175.16
108-69-0	3,5-Dimethyl	18	2.15	1.12	1	277.61
88-05-1	2,4,6-Trimethyl	19	2.72	0.82	1	151.71
579-66-8	2,6-Diethyl	20	3.17	0.76	1	73.09
24544-04-5	2,6-Di-iso-Propyl	21	4.04	0.25	2	30.81
95-51-2	2-Chloro	22	1.69	1.36	2	188.69
108-42-9	3-Chloro	23	1.69	0.87	2	76.87
106-47-8	4-Chloro	24	1.69	0.36	2	113.70
554-00-7	2,4-Dichloro	25	2.36	0.16	2	44.62
95-82-9	2,5-Dichloro	26	2.36	0.15	2	42.62
626-43-7	3,5-Dichloro	27	2.36	-0.09	2	31.59
636-30-6	2,4,5-Trichloro	28	3.02	-0.74	2	9.85
3481-20-7	2,3,5,6-Tetrachloro	29	3.69	-1.49	2	4.01
634-83-3	2,3,4,5-Tetrachloro	30	3.69	-1.29	2	2.53
615-65-6	2-Chlor-4-Methyl	31	2.26	1.11	2	93.55
95-69-2	4-Chlor-2-Methyl	32	2.26	0.50	2	63.25
95-79-4	5-Chloro-2-Methyl	33	2.26	0.65	2	44.78
95-74-9	3-Chloro-4-Methyl	34	2.26	0.56	2	57.69
87-60-5	3-Chloro-2-Methyl	35	2.26	0.54	2	59.03
348-54-9	2-Fluoro	36	1.23	2.09	2	260.49

(Continued)

CAS No	Name: aniline	ID	log $K_{ow}$ (nd)	log $VP$ (nd) (Pa)	BDP2 <sup>a</sup>	PNEC ( $\mu\text{g/L}$ )
372-19-0	3-Fluoro	37	1.23	1.81	2	139.89
771-60-8	penta-Fluoro	38	2.06	2.42	2	100.61
615-43-0	2-Iodo	39	2.23	−0.18	2	97.84
626-01-7	3-Iodo	40	2.23	0.04	2	49.03
2237-30-1	3-Cyano	41	1.12	−0.76	1	348.66
88-74-4	2-Nitro	42	2.00	−0.81	2	114.89
99-09-2	3-Nitro	43	1.43	−1.76	2	128.91
97-02-9	2,4-Dinitro	44	1.81	−3.16	2	34.89
606-22-4	2,6-Dinitro	45	1.24	−2.88	2	26.47

BDP2 = 1 denotes “readily biodegradable”, BDP2 = 2 denotes “non-biodegradable”.

## ABBREVIATIONS

BDP	BioDegradation Potential (part of the EPI Suite)
EPA	US Environmental Protection Agency
EPI Suite	Estimations Programs Interface for Windows
EUSES	European Union System for the Evaluation of Substances
HLC	Henry’s Law Constant
ID	Identification number for 45 anilines included in the study (cf. Annex)
IGC <sub>50</sub>	Population growth impairment
$K_{OW}$	Octanol–water partitioning coefficient
LE	Linear extension
OP	Organo Phosphorous
PNEC	Prediction No Effect Concentration
POR	Partial-Order Ranking
QSAR	Quantitative Structure–Activity Relation
REACH	Registration Evaluation Authorisation of CHemicals, the future European chemicals assessment scheme
$Rk_{av}$	Average rank according to (Brüggemann et al., 2004)
RLE	Average rank based on random linear extensions
TGD	Technical Guidance Document
VP	Vapour pressure

# Semi-Subordination Sequences in Multi-Measure Prioritization Problems

**W.L. Myers and G.P. Patil**

---

<b>Contents</b>		
	1. Introduction	159
	2. Theory	161
	2.1 Definite dominance	161
	2.2 Subordinate status	161
	2.3 Collective criteria and semi-subordination	163
	2.4 Semi-Subordination series and sib sorting	165
	2.5 Confirmatory considerations	167
	References	167

---

## 1. INTRODUCTION

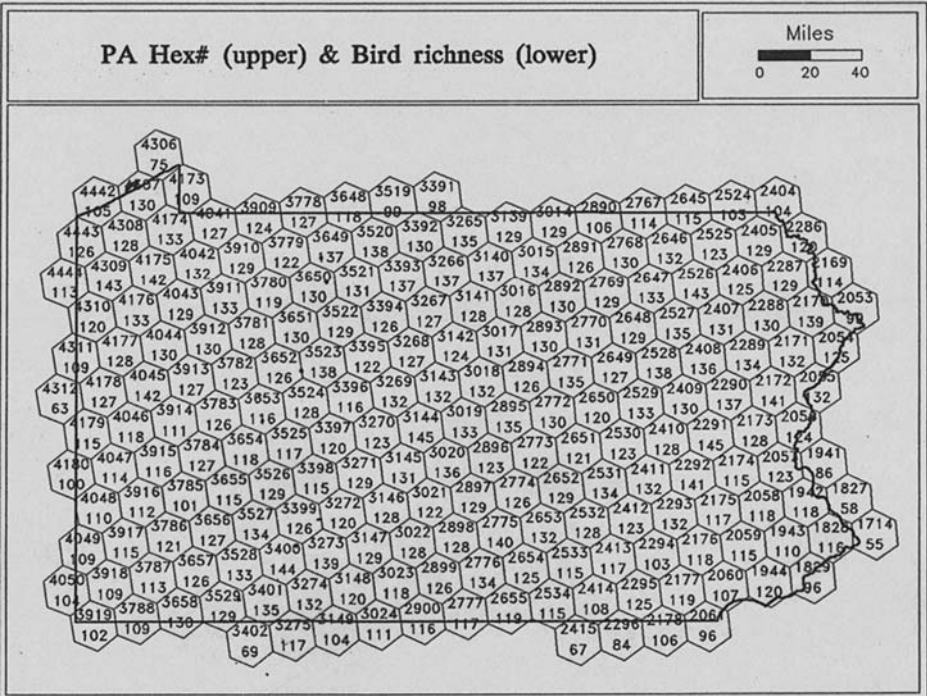
Practical problems of prioritization are often focused on sorting out salient cases among a collective when there are compound criteria that are less than completely consistent. Salient means standing out from the rest, notable, conspicuous or prominent. The problem is that different constellations of cases in multi-measure space may exhibit salience depending on the dimensional directions of interpretive interrogation. The mathematics of partial ordering and partially ordered sets is most often a point of departure for such situations, but divergence develops in attempting to resolve relativities within a hierarchy of partially ordered sets having intrinsic internal incomparabilities if all criteria are considered collectively ([Bruggemann et al., 2004](#); [Bruggemann and Carlsen, 2006](#)).

[Patil and Taillie \(2004\)](#) approach the relativities of intrinsic incomparabilities through linear extensions and rank frequency distributions among them. There are, however, computationally intractable multiplicities of linear extensions with even most modest numbers of cases and criteria. Therefore, they resort to pursuit of Markov Chain Monte Carlo convergence to suggest a specific ranking. An intuitive

assessment of aspects of emphasis among the linear extensions is somewhat illusive for contentious clientele, and the utility of the usual Hasse diagram as a graphical aid to interpretation deteriorates rapidly with increasing number of cases (Bruggemann et al., 1997). Another approach is to consider simplified subspaces of the case configurations with a flavor of multi-dimensional scaling that variously preserve information about comparability and incomparability (Shye, 1985; Bruggemann et al., 2003).

Here we start in the conventional direction of partial ordering (Lerche et al., 2002) and then change to a course of cross-comparisons for semi-subordination whereby we obtain sibsets on salience scaling and some sorting within sibsets based on semi-subordination sequences of stringency. We can readily depict our semi-subordination sorting on an X,Y graphic and can truncate to tail for selection of saliently superior or inferior cases.

As evidence of efficacy, we use a favorite dataset dealing with biodiversity in Pennsylvania that was a product of the GAP Analysis biodiversity assessment (Myers et al., 2000; Myers et al., 2006). In this instance, the state is divided into 211 hexagonal cells (cases) encompassing 635 km<sup>2</sup> each. We use five moderately correlated indicators of habitat diversity on these hexagons as our multiple measures. The first indicator is the number of avian species having potentially viable habitat in the hexagon, as shown in Figure 1. The second is the number of



**Figure 1** Map of Pennsylvania biodiversity hexagons showing avian species as lower value in each cell with hexagon identifier as upper number.

mammalian species having potentially viable habitat in the hexagon. The third is standard deviation for a grid of elevation points in the hexagon as an expression of topographic diversity. The fourth is percent of the hexagon that is forested. The fifth is the percent of the hexagon occupied by the single largest forest patch as an expression of habitat integrity. In this approach it is assumed that all the multiple measures have positive polarity with respect to the common concern.

## 2. THEORY

### 2.1 Definite dominance

Partial ordering typically proceeds to extract poset levels on the basis of definite dominance. The domination perspective on partial order is that one observational unit (case) dominates another if its values on all measures are as good or better, with at least one being better. Extracting levels of domination is a recursive process that begins with removing all undominated units from the case pool and marking them accordingly and then proceeding likewise on the remainder in a cyclic manner until all cases are marked for level of domination. It is common to mark the first (sub)set of undominated cases as level 1, but we will use level 0 instead to signify that they have no dominating (sub)set. It is important to note that each successive level is increasingly dominated, so that increasing level of dominance reflects greater consensus on *inferiority* among the several measures. Members of the same level have an intrinsic incomparability since there is disagreement among the measures in this regard.

These poset levels (of domination) conventionally become the basis for preparation of Hasse diagrams and elaboration in terms of linear extensions, cover relations, etc. Such additional analytical probing is consonant with aspects of the domination perspective. This is the point at which we depart the usual path of pursuit.

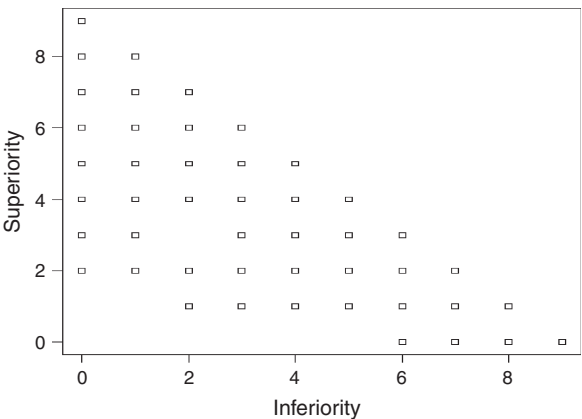
### 2.2. Subordinate status

The conditions of domination can be inverted to express what we will term *simple subordination*. One case is simply subordinate to another if its values on all indicators are less than or equal to those of the other, with at least one being less. Levels of simple subordinate status are determined in a parallel process to that of domination. The process begins with segregation and marking of cases *that have no subordinates* and are thereby simply subordinate to all cases remaining in the pool. These cases are marked as simple subordination level zero, reflecting the fact that they have no subordinates. The remaining pool is then processed recursively. Each successive level has increasingly more subordinates and reflects increasing consensus among the measures on *superiority*.

Definite domination and simple subordination are complementary but not equivalent constructs in partial ordering. This is shown by the plot in Figure 2 of subordination level on Y against domination level on X for the hexagon diversity data in [Figure 1](#).

Each plotted rectangle in [Figure 2](#) represents a *subset* of hexagons having the same domination level and subordination level. We consider such subsets as *poset families* of cases from coupling of complementary parent perspectives and the members of the family as being *siblings* of status in which the sense of incomparability has been refined relative to either type of parent poset levels. It is to be noted that families on the diagonal are conformant with respect to definite domination and simple subordination, whereas increasing departure from diagonal represents increasing difference according to parent perspectives. There are 46 families of hexagons.

We now exploit this differing duality to construct an ordinal place-rating scale of what we term *salience* for sibling subsets (sibsets), with lower numbers being better placed as in sports scenarios. The salience scale starts with 1 at the upper left of [Figure 2](#) and increases across rows then down for occupied places. First place thus goes to cases having the highest superiority and lowest inferiority. Within a given level of superiority, the scale decreases for increasing inferiority before proceeding to the next level of superiority. In the context of a plot such as [Figure 2](#), progression for level of *Y* moves toward the consistent case on the diagonal. The reasoning for this progression is that the diagonal represents cases that are equally “good” and “bad”, whereas leftward has lesser “badness”. Another view that leads to the same progression is to first find the best of both, which is the upper-left corner position. From this best of both, there are two possible moves—downward in the column or on the diagonal. Moving to the diagonal position is worse in both senses, whereas the direct downward is worse in only one sense. Thus, movement should be along a row toward the diagonal. Unoccupied places are simply skipped over in the place numbering since they do not pertain to the data at hand, and there is no claim to interval properties for the scale. It should also be noted that the



**Figure 2** Dual diagram of superiority (subordination level) and inferiority (domination level) for Pennsylvania hexagon biodiversity data.

same results are obtained if the data are first rank-ordered in a consistent sense prior to processing. Therefore, the salience scale makes use of only ordinal information in the data.

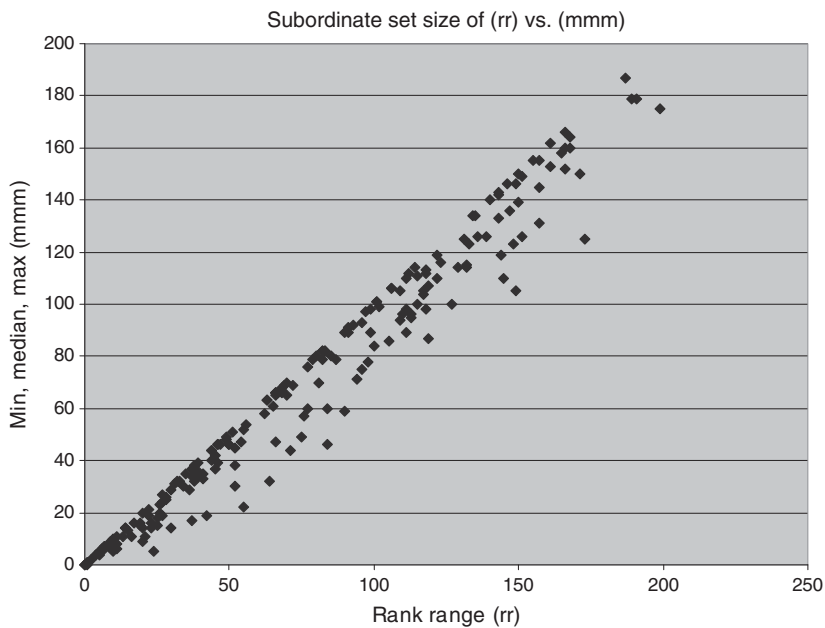
### 2.3 Collective criteria and semi-subordination

There are many kinds of comparison criteria that could be used in attempting to resolve the incomparabilities in a sibset, with subsets (subspaces) of the multiple measures being among these many possibilities. Any kind of subordination scaling that does not use all of the order information in the multiple measures can be considered as *semi-subordination*. This latter term has seen some limited use in the linguistics literature but apparently not in conjunction with principles of partial prioritization. Therefore, we adopt it here for present purposes. In moving toward further sorting within sibsets, we would prefer to retain representation of all measures and further to treat them equally in this regard (without weighting). Additional objectivity would entail not having to move mathematically outside the realm of ordering information into algebraic formulations. For example, the use of rank frequency distributions mentioned in the introduction retains these preferred properties but at the cost of computational complexity and some lack of interpretive transparency.

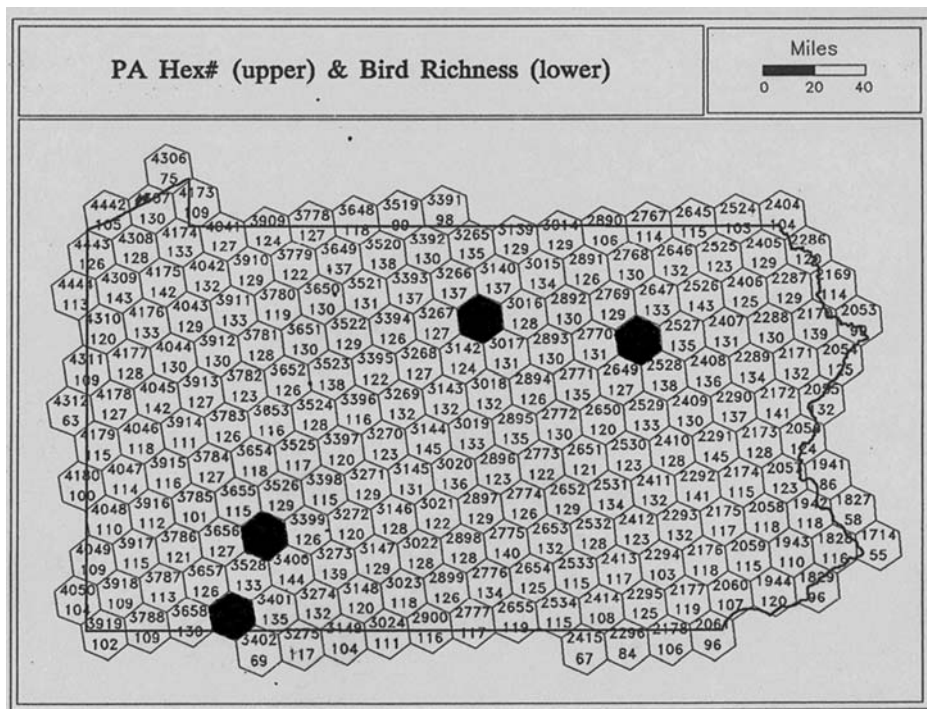
A liberalization relative to rank information is to use order statistics of ranks across indicators, since these stay within the domain of order information and treat all of the measures equally in an a priori sense. The minimum, maximum and median ranks across the measures are intuitively attractive since they are also easy to interpret and can be computed as long as there are three or more measures. Suppose then that ranks are assigned using place protocols so that a lower rank is more favorable than a higher rank. Then one case has a statistical semi-subordinance relative to another if it has either (a) less than or equal minimum rank and lesser maximum rank or (b) lesser minimum rank and less than or equal maximum rank. This can be called a *rank-range (rr) statistical semi-subordination*. A somewhat more stringent collective criterion of this nature for semi-subordination would be to specify a lesser median rank (min–median–max or mmm) also. To avoid combinatorial confusion, we can judge one case relative to another on a semi-subordination criterion by the size of the subordinate set for that criterion. Thus, one unit is semi-subordinate to another if it has a larger subset of semi-subordinate cases. Subordinate rr and mmm set sizes for the biodiversity data are compared in Figure 3. Since mmm is more stringent than rr, points of Figure 3 must lie on or to the right of the diagonal. Cases farther to the right of the diagonal are more sensitive to inclusion of the median criterion.

There is a constellation of four points at the extreme upper right showing collective advantage or high precedence on both of these criteria, and the hexagons corresponding to these four points are indicated in Figure 4. The mmm statistical semi-subordination is compared to simple subordination (all metrics or measures) in Figure 5. Since there is only one diagonal element appearing in Figure 5, the mmm statistical semi-subordination is apparently much less stringent than simple subordination.



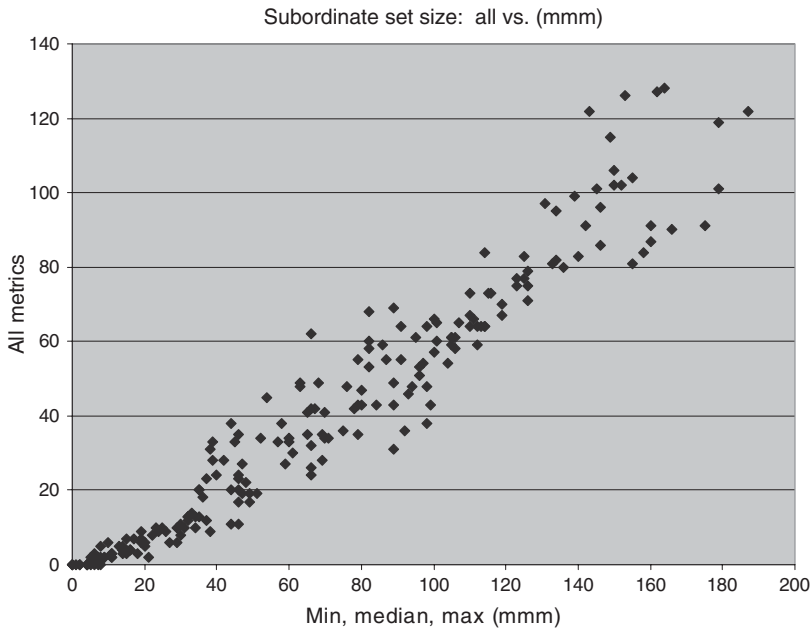


**Figure 3** Subordinate set size of rank-range (rr) and min–median–max (mmm) semi-subordination for hexagon data.



**Figure 4** Hexagons showing high precedence on rank-range (rr) and min–median–max (mmm) semi-subordination.





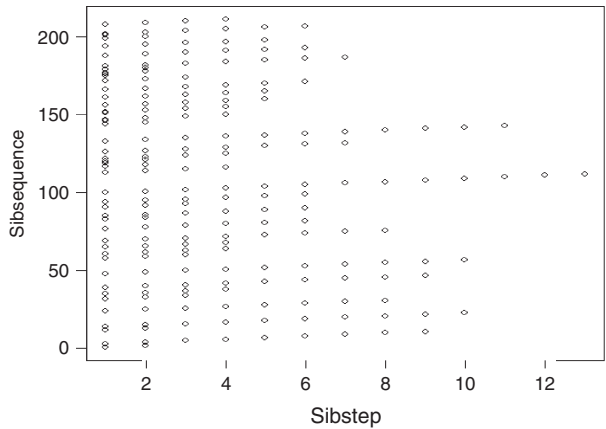
**Figure 5** Statistical (mmm) semi-subordination vs. simple (all metrics) subordination.

## 2.4 Semi-Subordination series and sib sorting

As we progress from rr to mmm and then to strict subordination on all measures, we go through a series of increasingly stringent criteria for (semi-) subordination. To capture collective case characteristics across the series, we sum the subordinate set sizes across the series for each case. These “subsums” are somewhat reminiscent of cumulative rank frequency profiles for the cases, and they can serve to impose a sort order within a family on the salience scale. We break ties on specific components of the subsum. Sibs having equal subsums are further sorted on rr, then on mmm and then on strict subordination with all metrics. For present purposes, any further ties among sibs would be left in the original case order. We then assign subsequence numbers within each (sorted) family of cases, calling these *sib steps*.

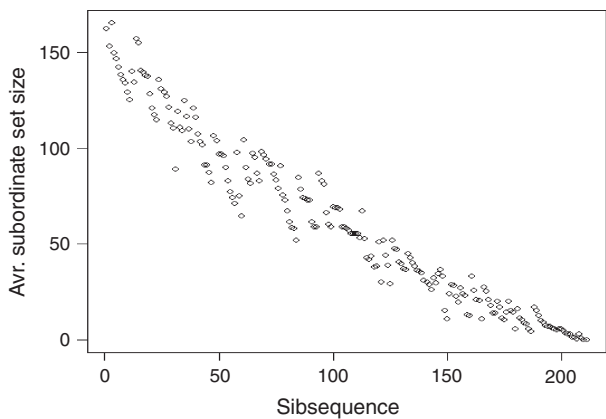
At this stage the entire collection of cases is substantially sorted with regard to (semi-)subordinate status, so we proceed to assign sequence numbers across the entire file to form a “subsequence” rating of prioritization precedence.

Plotting subsequence on Y-axis against sibstep on X-axis provides a “*sibspread*” graphic that shows size of sibsets and superiority sorting along with sibscale sectors. The lower elements on this graphic are prime prospects for superiority selection, whereas the upper elements are indicated for inferiority interventions. The sibspread graphic is shown for the biodiversity-based hexagon data in [Figure 6](#). The “first family” consists of two hexagons—#3527 and #2647—with #2648 being the first sib in the second family. It can be noted that #2647 was not

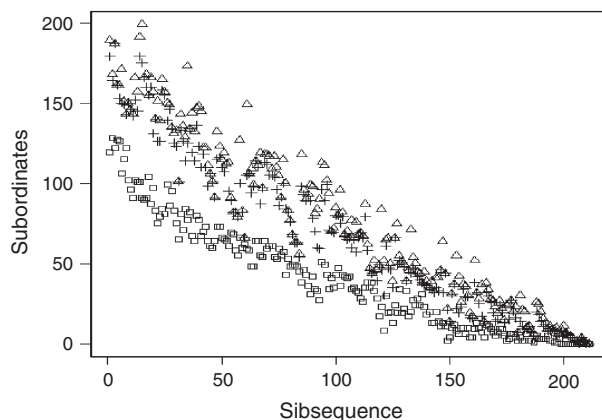


**Figure 6** Sibspread graphic for prioritization of Pennsylvania hexagons with regard to biodiversity.

among the prominent four of Figure 4. It takes precedence over others by virtue of having a particularly large number of simply subordinate cases, and this is what controls the salience scale. This affords evidence that semi-subordination should be used as a supplement to salience scaling rather than separately. However, it is also notable that #2647 is a geographic neighbor of #2678 with associated similarity of landscape context. It is apparent that the “second family” is substantially larger than the first, containing nine hexagons (#2648, #2171, #3143, #3274, #2290, #2527, #3266, #3393 and #2170). These can be examined for geographic distribution using Figures 1 and 4. The other two hexagons of the prominent four in Figure 4 constitute the first and second sibs of the fourth family, which is still well within the top 10% of the 46 families.



**Figure 7** Relationship of average subordinate set size to sibsequence for Pennsylvania hexagon data.



**Figure 8** Co-plotted components of subsum with rank-range (rr) triangles, + for mmm and box for simple (strict) subordination.

For dealing with a multitude of cases, plotting on the Y-axis can be easily censored from either above or below according to the prioritization purpose. Likewise, plotting can be restricted to any particular percentage of cases working from either end. It bears repeating that the sibsequence scale uses a place perspective whereby the lower numbers are better placed.

## 2.5 Confirmatory considerations

If the semi-subordination subsum used to sort sibsteps is generally compatible with salience scaling, then it should be expected to find a strong trend across the entire sibsequence of declining subordinate sizes with increasing sibsequence. This trend is shown in Figure 7 and the expected decline is seen, with the trend being somewhat nonlinear. Further perspective in this regard can be obtained by co-plotting the components of the subsum using different symbols as in Figure 8. It can be seen that the nonlinearity is induced largely by the simple (strict) subordination rather than by the semi-subordination components.

## REFERENCES

- Bruggemann, R., Carlsen, L. (2006). Introduction to partial order theory exemplified by the evaluation of sampling sites. In: *Partial Order in Environmental Sciences and Chemistry* (Bruggemann, R., Carlsen, L. eds), Springer, Berlin, pp. 61–110.
- Bruggemann, R., Oberemm, A., Steinberg, C. (1997). Ranking of aquatic effect tests using Hasse diagrams, *Toxicol. Environ. Chem.* 63, 125–139.
- Bruggemann, R., Sorensen, P.B., Lerche, D., Carlsen, L. (2004). Estimation of averaged ranks by a local partial order model, *J. Chem. Inf. Comp. Sci.* 44, 618–625.
- Bruggemann, R., Welzl, G., Voight, K. (2003). Order theoretical tools for the evaluation of complex regional pollution patterns, *J. Chem. Inf. Comp. Sci.* 43, 1771–1779.

- Lerche, D., Bruggemann, R., Sorensen, P.B., Carlsen, L., Nielsen, O.J. (2002). A comparison of partial order technique with three methods of multicriteria analysis for ranking of chemical substances, *J. Chem. Inf. Comp. Sci.* 42, 1086–1098.
- Myers, W., Bishop, J., Brooks, R., O'Connell, T., Argent, D., Storm, G., Stauffer, J., Jr. (2000). *The Pennsylvania GAP Analysis Final Report*, The Pennsylvania State University, University Park, PA 16802, USA.
- Myers, W., Patil, G.P., Cai, Y. (2006). Exploring patterns of habitat diversity across landscapes using partial ordering. In: *Partial Order in Environmental Sciences and Chemistry* (Bruggemann, R., Carlsen, L. eds), Springer, Berlin, pp. 309–325.
- Patil, G.P., Taillie, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization, *Environ. Ecol. Stat.* 11, 199–228.
- Shye, S. (1985). *Multiple Scaling – The Theory and Application of Partial Order Scalogram Analysis*, North-Holland, Amsterdam.

# Multi-Criteria Decision-Making Methods: A Tool for Assessing River Ecosystem Health Using Functional Macroinvertebrate Traits

**S. Canobbio, V. Mezzanotte, D. Ballabio, and M. Pavan**

---

<b>Contents</b>		
	1. Introduction	169
	2. Biomonitoring Program	171
	2.1 Study area	171
	2.2 Sampling	171
	2.3 Invertebrate analysis	173
	3. Applications of Total-Order Ranking Techniques	175
	4. Environmental Description of Serio River	175
	5. Ecology of Serio River	177
	5.1 Invertebrate assemblages	177
	5.2 Ecosystem attributes	178
	5.3 Ranking of sites	180
	6. What we Obtained Using This Method?	183
	6.1 River evaluation and restoration	183
	6.2 Final considerations	184
	References	185
	Appendix	188

---

## 1. INTRODUCTION

Macroinvertebrate communities are widely used for assessing river ecosystem health, and this has led to the development of a large amount of indices. Most of them are focused on taxonomical diversity and abundance or on the presence/absence of indicator taxa ([Hellawell, 1989](#); [Karr, 1991](#); [Rosenberg & Resh, 1993](#)). However, in the last years, decision makers are asking for new bioassessment

techniques that could dialogue with the emerging priorities of freshwater protection policies. These priorities are (1) integrated approach at watershed scale, considered as a comprehensive and continuous ecosystem affected by different impacts, for more efficient conservation results (Vugteveen et al., 2006; Witter et al., 2006); (2) monitoring protocols to be applied to broad geographical areas (therefore making taxonomical approach partially ineffective) to check the results of the enforcement of supranational rules, such as the Water Framework Directive (WFD) 2000/60 (Dolédec et al., 1999; Statzner et al., 2001, 2005) and (3) assessment tools, which can easily lead to a direct understanding of the aquatic ecosystem and provide the basis for well-aware decision making and restoration planning (Karr & Chu, 1999; Ghilarov, 2000; Bash & Ryan, 2002; Dolédec et al., 2006).

Functional diversity analysis has already proved useful in the study of aquatic ecosystems. Assessment techniques have been developed using the distribution of macroinvertebrate functional groups (i.e. Cummins, 1974; Merritt & Cummins, 1996; Tachet et al., 2000) across the physical heterogeneity of streams and rivers (Malmqvist, 2002), instead of taxonomical analysis. Thus, observations focus on spatial and temporal modifications of ecological traits (Townsend & Hildrew, 1994; Statzner et al., 2005) and, eventually, they can be used to determine ecosystem attribute surrogates. The use of invertebrate traits to determine the most common running water ecosystem attributes has been introduced recently (Merritt et al., 1996, 1999; Stone & Wallace, 1998) and has been successfully adopted for the assessment of river oxbows in Florida (Merritt et al., 2002). Ecosystem attributes are defined starting from traits such as functional feeding groups (FFG), functional habit groups (FHG), voltinism (generation turnover) and drift propensity in river flow.

According to Merritt et al. (2002), ecosystem attributes and criteria adopted to analyze functional groups are focused on energy balance and trophic web. The attributes studied are: ecosystem primary production–respiration ratio (P/R); coarse particulate organic matter–fine particulate organic matter (CPOM/FPOM) ratio; suspended–benthic fine particulate organic matter (SPOM/BPOM) partition; habitat stability; preys control by predators; velocity of community life cycle; water-column-feeding fish food availability; benthic-feeding fish and wading birds food availability. To every ecosystem attribute, a threshold level (studied, proposed and/or adopted by Minshall et al., 1992; Merritt et al., 1996; Stone & Wallace, 1998; Merritt et al., 2002) is given between a good riverine environment and an impaired site.

The intrinsic complexity of the analyzed system and the high number of derived attributes, which can even be conflicting among them, need a proper data treatment to set priorities and define rank order of the studied environments (Pavan, 2003). In recent years, total and partial ranking techniques have been widely and successfully used for different purposes, including evaluation of aquatic toxicological tests (Brüggemann et al., 1997), ranking chemicals for environmental hazard (Newman, 1995) or contaminated sites (Sørensen et al., 1998) and comparison among ecosystems (Pudenz et al., 1997; Brüggemann et al., 1999).

In the present paper, we apply Multi-Criteria Decision-Making (MCDM) methods proposed in literature (Keller & Massart, 1991; Hendriks et al., 1992) to river functionality assessment to generate information and to provide further

understanding as a basis for river restoration strategies. We selected three functions (desirability, utility and dominance) and applied them to Serio River, an Italian impaired river, which is the object of intense studies, planning and restoration efforts.

## 2. BIOMONITORING PROGRAM

### 2.1 Study area

Serio River is 124 km long, draining 1256 km<sup>2</sup> in Lombardy, a Northern Italy region. It takes origin from the Alps and flows into Adda River just before the latter merges into Po River, the main watercourse in Italy. Serio River undergoes several abstractions both for hydroelectric power generation and for irrigation, and its watershed includes wide urban and industrial areas producing high-polluting loads.

The river course can be divided into four sectors, each with different physical habitat:

(S1) Alpine valley (highest point in catchment is 3052 masl), where the river main substratum is chiefly made of cobbles and boulders. Various dams for hydroelectric power generation are already present.

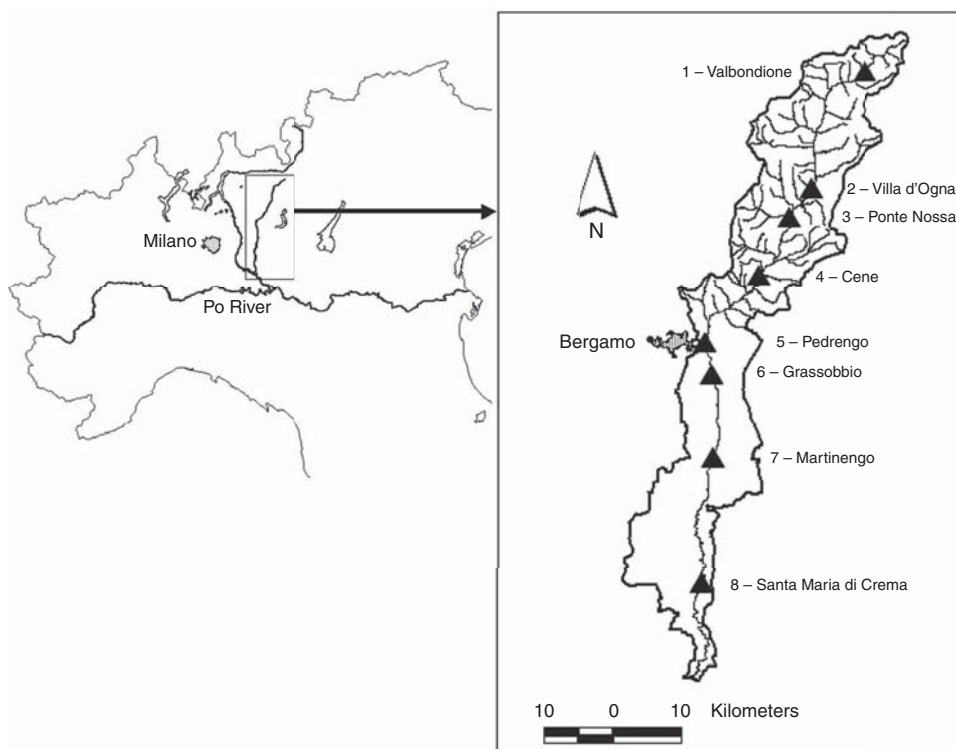
(S2) Mountain and piedmont floodplain, wider and characterized by a lower slope than S1. Substratum is mainly made of cobbles, gravels, and occasionally of boulders and bedrocks. Several dams are present, the last of which abstract water for irrigation, with no return flow. The floodplain area is intensively exploited, with several industrial and urban settlements (Bergamo urban district). Here, the river receives the effluents from the first wastewater treatment plants (WWTPs).

(S3) Main alluvial plain. The river becomes braided and gravel bars and pools appear as typical habitats. Substratum is mainly made of gravel and small cobbles. The riverbed is very permeable and is often completely dry during Summer. Agriculture covers large areas and the remaining land includes urban and industrial settlements whose wastewaters are discharged, after treatment in WWTPs, into the river.

(S4) Thinner texture plain. The river flow turns back to a single channel and is sinuously meandered. Besides gravel and sand, silt is present as substratum component. The land use in this area is mainly agricultural, though the watercourse passes through a city (Crema). Resurgence and return flow from irrigation net increase the river water flow.

### 2.2 Sampling

Macroinvertebrate data were collected from eight selected sites (Figure 1) representative of the above-mentioned sectors in five different seasons, from Spring



**Figure 1** Study site. *Left*: position of Serio River in Northern Italy. *Right*: location of the eight sampling sites along the river. Basin shape is shown as well.

2004 to Spring 2005. Sampling sites were (1) Valbondione; (2) Villa d'Ogna and (3) Ponte Nossia for sector S1; (4) Cene and (5) Pedrengo for sector S2; (6) Grassobbio and (7) Martinengo for sector S3; (8) Santa Maria di Crema for sector S4. Site 8 was included from Summer 2004, during the second sampling campaign. Site 1 was found covered by a thick layer of ice during Winter 2004/2005 campaign and, thus, not sampled. Site 7 was found completely dry during the fifth campaign (Spring 2005).

In each site, invertebrate assemblages were collected by taking samples in different microhabitat locations, possibly along transects. Each sample covered a square area of  $0.05 \text{ m}^2$  where all the invertebrates, substratum and vegetation were removed by hand using a  $500\text{-}\mu\text{m}$  mesh net. Invertebrates and vegetation were preserved in a final solution of formaldehyde 4%.

Environmental variables were measured in field or in laboratory and included mean water depths, flow velocities, temperature, specific electric conductivity, pH, COD, concentrations of dissolved oxygen (DO), total nitrogen (tot-N), ammonia nitrogen ( $\text{N-NH}_4$ ) and total phosphorus (tot-P). Information



about observed flow was compared with expected mean annual flows according to a regionalized model built over a 1929–1991 data set (Paoletti & Becciu, 2006).

## 2.3 Invertebrate analysis

In laboratory, macroinvertebrates were counted and classified to genus or family level, according to Sansoni (1992), Merritt & Cummins (1996) and Tachet et al. (2000). Species were identified when necessary to assign the correct traits. Taxonomical information was used to determine extended biotic index (EBI; Woodiwiss, 1978; adapted to Italian benthic fauna by Ghetti, 1997). Extended biotic index is based upon taxa sensitivity to DO concentrations and taxa richness and is the official bioassessment index in Italy.

Dry mass for each taxon was determined on the basis of the total length of individuals in a subsample (generally made of 10% of the detected individuals, except for low-density taxa for which higher percentages were used), approximated to the nearest 5 mm. Mean lengths were related to dry mass by regression equations, deriving coefficients from various literature (Finogenova, 1984; Meyer, 1989; Towers et al., 1994; Burgherr & Meyer, 1997; Nyström & Pérez, 1998; Benke et al., 1999; Stoffels et al., 2003; Hale et al., 2004) to cover up all the taxonomical diversity of invertebrates sampled.

Traits were assigned to invertebrate taxa according to Merritt & Cummins (1996) and Tachet et al. (2000). Functional feeding groups were classified as follows (Merritt et al., 2002): shredders, divided into detritus (D-SHRED) or live vascular plant eaters (LVP-SHRED); scrapers (SCRAP); filtering collectors (F-COLL); gathering collectors (G-COLL) and predators, including parasites (PRED). Piercers (PIERC) were never found. Functional habit group, used to describe locomotion and attachment to substrate, were: clingers (CLING); climbers (CLIMB); sprawlers (SPRAW); burrowers (BURW) and swimmers (SWIM).

Voltinism is the attitude of the specific taxon to generate less/equal or more than one generation per year, while drift propensity is the attitude of the specific taxon to be a behavioural or accidental drifter. Values of these traits are expressed as percentage of the different options according to Tachet et al. (2000).

Traits were aggregated to express ratios that are related to ecosystem attributes, as shown and explained in Merritt et al. (2002). Thresholds between unimpaired and impaired ecosystems for each of the attributes were also derived from Merritt et al. (2002) and references cited therein and were used as criteria for total ranking. Considering the climatic conditions of Serio river basin, we decided to use a unique threshold for CPOM/FPOM ratio. Thus, we adopted the proposed dry season one. The description, ratio and threshold values of ecosystem attributes are shown in Table 1.

**Table 1** Description of ecosystem attributes based upon invertebrate functional traits

Ecosystem attributes	Description	Traits (FFG, FHG, voltinism and drift) ratios representing ecosystem parameters	General criteria levels of ratios (good functionality for the given attribute)
P/R	Primary production as a proportion of community respiration	LVP-SHRED + SCRAP + PIERC as a proportion of D-SHRED + total COLL	>0.75 (autotrophic system)
CPOM/FPOM	Storage CPOM as a proportion of FPOM in and on the sediments	Total SHRED as a proportion of total COLL	>0.50 (normal shredder riparian system in dry climate)
SPOM/BPOM	Suspended FPOM as a proportion of deposited FPOM	F-COLL as a proportion of G-COLL	>0.50 (enriched in SPOM)
Habitat stability I	Availability of stable surfaces and non-shifting sediments	SCRAP + F-COLL as a proportion of total SHRED + G-COLL	>0.50
Habitat stability II	Availability of stable surfaces and non-shifting sediments	CLING + CLIMB as a proportion of BURW + SPRAW + SWIM	>0.60 (stable substrates not limiting)
Top-down control	Control of predators on prey	PRED as a proportion of all other FFG	0.15 (normal top-down predator control)
Life cycle	Short life cycle versus long-life cycle	Generations/year >1 as a proportion of generations/year $\leq$ 1	>0.75 (pioneer, early successional communities)
Drift food	Food supply for water-column-feeding fish	Behavioural drifters as a proportion of accidental drifters	>0.50 (good food supply)
Benthic food	Food supply for benthic-feeding fish and wading birds	SPRAW as a proportion of CLING + CLIMB + BURW + SWIM	>0.60 (good food supply)

Source: Modified after [Merritt et al. \(2002\)](#)

### 3. APPLICATIONS OF TOTAL-ORDER RANKING TECHNIQUES

The MCDM as described in Chapter 2 of this book was applied to the ecosystem attribute data set, using the downloadable software DART (2007, version 1.0, Talete srl, Chapter 9). Values and thresholds proposed in the cited literature were converted into desirability, utility and dominance functions in order to transform values of the criteria into the same scale. For every ecosystem attribute but top-down control, a sigmoid function was used. This choice aimed at enhancing and pointing out the representativeness of values close to the extremes. The function transformed the value of each element under the threshold level in a value comprised between 0 and 1. Values above the threshold were considered optimal (=1) as the threshold itself, according to ranking method theory and to avoid a drop of the desirability values at threshold level. Top-down control attribute was converted in a parabolic function centred at the threshold of 0.15.

The ranking values obtained were compared qualitatively to EBI and to some environmental variables describing each site. All the criteria have been considered equally important. The different behaviour of functional indicators is, thus, due to the mathematical definition of the indicators themselves.

Multi-criteria decision-making methods have been applied to evaluate the different samples taken in the studied sites. Each sample is identified by the site code (i.e. 1 for Valbondione, see [Figure 1](#)) and a number, representative of the sampling campaign (1: Spring 2004; 2: Summer 2004; 3: Fall 2004; 4: Winter 2004/2005; 5: Spring 2005), as 1.1, 1.2, etc.

### 4. ENVIRONMENTAL DESCRIPTION OF SERIO RIVER

Environmental variables measured in each of the sampled sites of Serio River are summarized in [Table 2](#). From Alps to the plains, an altitude-related increase of temperature takes place. Dissolved oxygen saturation levels are quite good for the whole watercourse, but high variability was observed in sector S2 (sites 4 and 5). Site 4 is strongly affected by water abstraction: this brings to water scarcity episodes and to fluctuations of DO. Site 5 is influenced by the discharge of a WWTP effluent treated with pure oxygen and disinfected with ozone: mean DO saturations are high and a maximum of 159.1% was observed in Spring 2005. Maximum COD was detected in site 5 as well, while nutrients (total phosphorus and total nitrogen) and specific conductivity increase along the river and show their maximum at site 8 (being the river enriched by both point and non-point source pollution). Ammonia is quite low but presents some fluctuations in sites (5 and 6) where WWTP effluents are discharged. Depth and water velocity show some variability due to the different flows characterizing the river in the various seasons. Mean depth was between 0.66 m (site 8) and 0.21 m (site 7), and this reflects the high variability of river geomorphology in plains: in site 8, the channel pattern is straight, 20 m, while in site 7, the channel pattern is braided, 100–150 m.

**Table 2** Environmental variables for eight Serio River study sites

	Valbondione (1)	Villa d'Ogna (2)	Ponte Nossà (3)	Cene (4)	Pedrengo (5)	Grassobbio (6)	Martinengo (7)	S. Maria di Crema (8)
<i>T</i> (°C)	11.5 ± 2.4	11.6 ± 3.8	12.3 ± 4.7	13.5 ± 4.9	17.0 ± 7.0	14.6 ± 5.7	16.1 ± 6.0	16.8 ± 5.2
DO (%)	91.4 ± 2.3	93.0 ± 7.4	95.1 ± 4.3	87.7 ± 20.4	108.5 ± 28.4	95.6 ± 5.5	97.3 ± 10.0	94.2 ± 5.7
pH	7.80 ± 0.15	8.02 ± 0.27	8.13 ± 0.13	8.14 ± 0.06	7.95 ± 0.15	8.23 ± 0.21	8.29 ± 0.21	7.82 ± 0.17
Conductivity (μS/cm)	147 ± 8	211 ± 38	525 ± 319	325 ± 59	486 ± 116	419 ± 73	584 ± 111	792 ± 35
COD (mg/l)	1.3 ± 1.3	0.8 ± 0.9	6.3 ± 5.9	5.1 ± 0.5	10.0 ± 6.9	8.7 ± 3.8	7.0 ± 1.6	6.3 ± 1.7
Tot-P (mg/l)	0.020 ± 0.031	0.023 ± 0.029	0.034 ± 0.028	0.054 ± 0.020	0.144 ± 0.063	0.122 ± 0.053	0.210 ± 0.056	0.205 ± 0.047
Tot-N (mg/l)	1.179 ± 0.231	1.301 ± 0.332	1.711 ± 0.341	1.903 ± 0.434	3.095 ± 0.976	2.897 ± 1.249	3.697 ± 0.811	6.918 ± 1.799
NH <sub>4</sub> (mg/l)	0.000 ± 0.000	0.028 ± 0.044	0.003 ± 0.004	0.041 ± 0.052	0.139 ± 0.119	0.151 ± 0.205	0.120 ± 0.079	0.147 ± 0.093
Depth (m)	0.54 ± 0.06	0.34 ± 0.05	0.38 ± 0.06	0.29 ± 0.12	0.57 ± 0.10	0.21 ± 0.06	0.16 ± 0.04	0.66 ± 0.04
Flow/mean flow (year)	0.89 ± 0.31	1.04 ± 0.43	0.63 ± 0.24	0.46 ± 0.20	0.53 ± 0.22	0.43 ± 0.20	0.40 ± 0.21	0.77 ± 0.30

Data (mean ± SD) were recorded from Spring 2004 to Spring 2005 in the same days of macroinvertebrate sampling.

Measured flows can differ a lot from expected mean annual values, especially downstream water abstractions for irrigation purposes (starting from site 4). During the investigation period, flows were always lower than expected and the combined effect of abstraction and dry season caused even complete droughts (sample 7.5). Physical habitat was better in the upper valley (sites 1, 2 and 3) and decreased to minimum (sites 5 and 6) in the urbanized piedmont area. Physico-chemical variables were significantly different in samples obtained in sites 1–4 when compared with samples obtained in sites 5–8 ( $P < 0.01$ ). Hydrological variable flow/mean expected flow was significantly different between sites 1, 2, 3, 8 and sites 4, 5, 6, 7 ( $P < 0.05$ ). Thus, sites 1–3 (S1) can be considered the least impaired; site 4 (S2) unpolluted but affected by hydrological stress; sites 5–7 (S2 and S3) impaired in both their water quality and hydrology; site 8 (S4) polluted but presenting better flow conditions when compared with upstream sites.

## 5. ECOLOGY OF SERIO RIVER

### 5.1 Invertebrate assemblages

The total number of taxa found in each sample has been taken as an indicator of the site taxa richness. Data about observed assemblages and EBI values are reported in [Table 3](#), while a complete list of collected taxa is shown in [Appendix](#).

Taxa richness and estimated dry mass were higher in montane valley sites (especially in sites 1 and 2), while density was lower. Data variability was high. Starting from site 4, the total number of taxa decreased and density increased significantly ( $P < 0.05$ ). Dry mass was already about 50% less in site 3 than in the upstream sites, even if richness and density were still comparable. Density decreased again in plains, at sites 7 and 8, but this change did not concern richness and mass. Mean EBI values, normally ranging from 0 to 12, were very high in the first three sites (with a maximum of 9.8 at site 1) and decreased thereafter, keeping then around 6.

**Table 3** Mean number of taxa, density of individuals (individuals/0.1 m<sup>2</sup>), total dry mass (mg) and EBI value from samples taken in eight Serio River study sites (mean  $\pm$  SD) from Spring 2004 to Spring 2005 (five sampling campaigns)

Site	Taxa richness	Density	Mass	EBI value
1	15.5 $\pm$ 6.6	205 $\pm$ 137	1,223 $\pm$ 849	9.8 $\pm$ 1.3
2	16.4 $\pm$ 5.1	211 $\pm$ 135	1,254 $\pm$ 881	9.6 $\pm$ 1.1
3	14.0 $\pm$ 4.4	240 $\pm$ 92	885 $\pm$ 163	8.4 $\pm$ 1.7
4	10.4 $\pm$ 1.8	504 $\pm$ 201	884 $\pm$ 297	6.0 $\pm$ 1.0
5	8.8 $\pm$ 2.2	408 $\pm$ 318	623 $\pm$ 304	5.6 $\pm$ 1.1
6	9.4 $\pm$ 2.1	545 $\pm$ 267	943 $\pm$ 384	6.2 $\pm$ 0.8
7	10.3 $\pm$ 2.2	290 $\pm$ 106	674 $\pm$ 488	6.8 $\pm$ 0.5
8	8.5 $\pm$ 1.7	188 $\pm$ 108	618 $\pm$ 323	6.0 $\pm$ 0.8

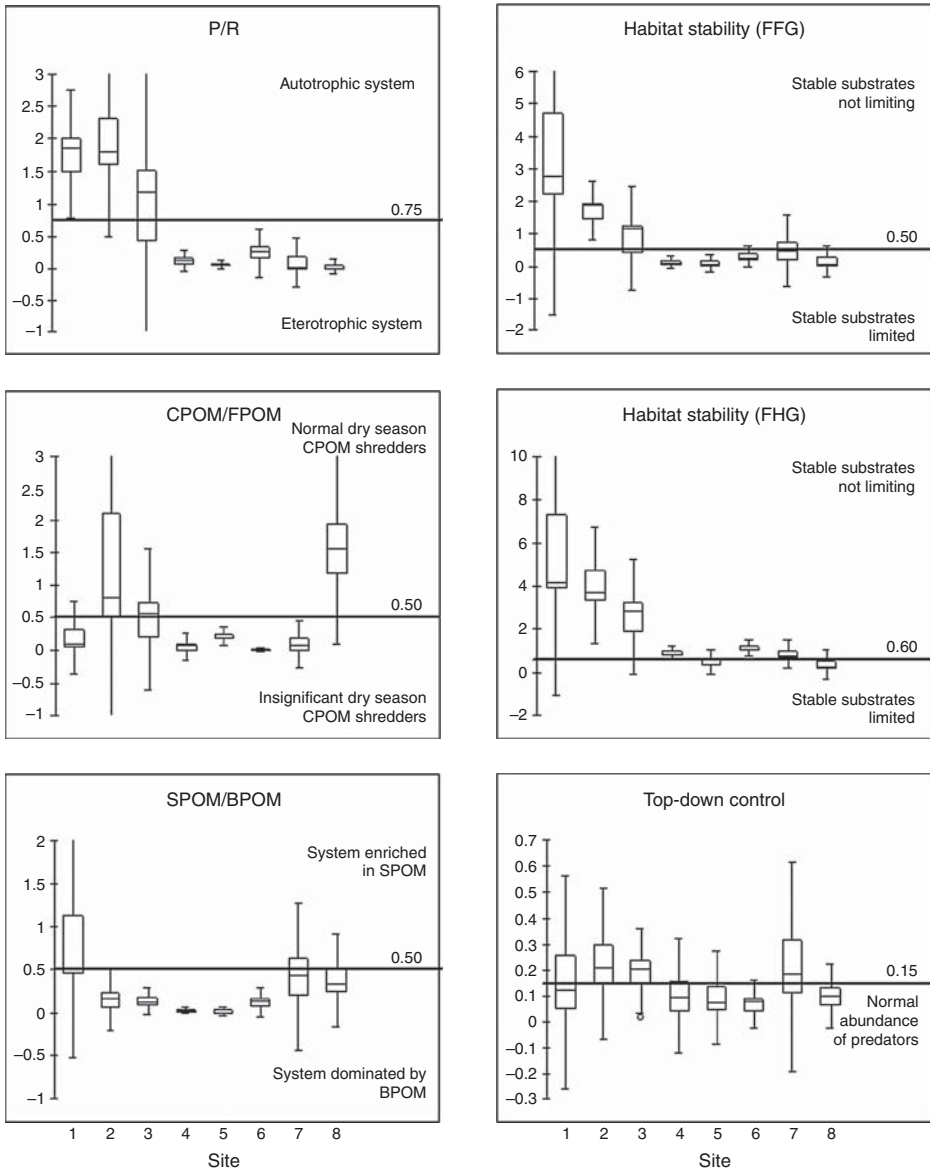
Invertebrate mass composition changed from one site to another and during the different seasons. Generally, in site 1, *Ecdyonurus* and *Rithrogena* accounted for at least half of the total mass. However, in Spring 2004, the most important fraction of dry mass was made of some trichoptera, mainly Limnnaephilidae, Hydropsichidae and Sericostomatidae. In Spring 2005, these trichoptera were not so significant and ephemeroptera accounted again for most of dry mass. In site 2, plecoptera (*Perla*, *Isoperla*) contributed more to total mass than in site 1, along with ephemeroptera (mainly *Ecdyonurus*, *Rithrogena* and *Baetis*) and trichoptera. In site 3, *Ecdyonurus* population was still relevant, but total mass was mostly made of *Baetis* and some trichoptera families (Rhyacophilidae, Limnnaephilidae and Hydropsichidae). In site 4, *Baetis* still accounted for most of dry mass (up to 81% in Winter), while Gastropoda became significant. Site 5 was dominated by *Baetis*, Chironomidae or Gammaridae, alternatively, while site 6 was always dominated by *Baetis* (up to 73% in Summer) with *Lymnaea* (reached 23% in Winter). In site 7, dominant taxa were *Baetis*, Hydropsichidae and Tabanidae. In site 8, Gammaridae were prevailing, with the exception of Spring 2005 when Hydropsichidae was the dominant taxon.

## 5.2 Ecosystem attributes

Ecosystem attribute ratios, shown in [Figure 2](#), allowed to evaluate the overall river functionality. Serio River appeared to be an autotrophic system in the first three sampling sites (Sector 1), where the high value of P/R ratio was due to the high density of scrapers rather than LVP shredders. Site 3 was characterized by high variability: P/R ratio ranged between 0.28 and 1.57, but both mean (0.99) and median (1.17) were over the threshold. P/R attribute dropped below 0.75, to very low values (heterotrophic system), at site 4 and did not rise any more.

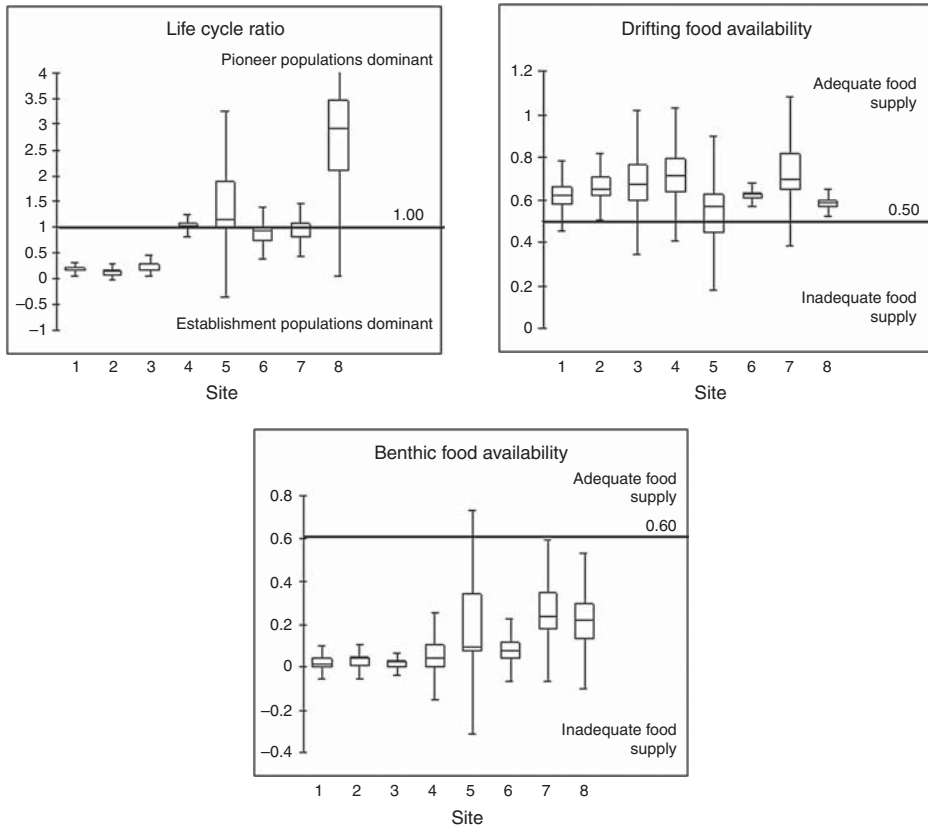
Coarse particulate organic matter–fine particulate organic matter ratio was above the threshold (0.50) at sites 2 and 3 (Sector 1) and at site 8 (Sector 4). Shredders contribution and, thus, coarse organic matter were low in the other sites. Where the river is most affected by the presence of dams (Sector 2, site 4), fine POM increased and the ratio kept below 0.50 (or even 0.25, which is a wet season threshold according to [Merritt et al., 2002](#)) until site 8. At site 5, a maximum value of 6.58 was observed in Spring 2004, but the low ratio of the attribute in the other seasons dropped the median value to 0.19. Conflicting values of P/R and CPOM/FPOM in sites 1 and 8 prove the autotrophy of the first one and the heterotrophy of the latter. The ratio between suspended and benthic (or deposited) FPOM showed low values for all the monitored sites: only site 1 presented a median value of 0.53 (and a mean of 1.05 due to a maximum value of 2.79 observed in Winter), accounting for a high proportion of filters in the total collectors mass.

Using both FFG and FHG, a firm substrate supporting invertebrate attachment was detected in the first sector where, thus, habitat stability (1) and (2) attributes got very high values. Using FFG, median values for sites 1, 2 and 3 were 2.77, 1.91 and 1.18, respectively, while using FHG, median values were even higher (7.14, 4.60 and 2.79, respectively). At site 4, ratios dropped. From there,



**Figure 2** Box plots showing median, first and third quartile and standard deviation of the ecosystem attributes derived from macroinvertebrate trait analysis (see Table 1) for each sampling site on Serio River. The line identifies the threshold value for each ecosystem attribute ratio as suggested in Merritt et al. (2002).

stability, evaluated by FFG, remained below the threshold (0.50) over the entire river (with the exception of site 7 where median was 0.51). In contrast, values obtained by FHG, even if lower than in the first three sites, were generally over the threshold (0.60). Only sites 5 (0.38) and 8 (0.25) showed limiting substrates.



**Figure 2** (Continued).

Top-down control was generally not far from optimum, set at 0.15. The median values of this attribute ranged from 0.08 (sites 5 and 6) to 0.21 (site 2). Rapid turnover species, which normally absolve pioneer function, were predominant in site 5 (1.14) and, especially, in site 8 (2.92). The ratio was close to the threshold at sites 4 (1.04), 6 (0.93) and 7 (1.00), but very low at sites 1 (0.19), 2 (0.14) and 3 (0.29), where well-established and, thus, slow-colonizing populations were dominant. The evaluation of invertebrates as a food source for upper trophic levels gave conflicting results. Invertebrates appeared as a good food supply for water-column-feeding fish (the median value of each site was higher than the adopted threshold of 0.50) but as a very poor food supply for benthic-feeding fish and wading birds. For all sites, the median was very low (from 0.02 at site 1 to 0.24 at site 2) and, thus, did not reach the threshold (0.60).

### 5.3 Ranking of sites

Results for desirability, utility and dominance functions are given in [Table 4](#), compared with EBI results.



**Table 4** Desirability, utility, dominance and EBI values for each site (mean, SD and subsequent rank)

Sites	Desirability			Utility			Dominance			EBI value		
	Mean	SD	Rank	Mean	SD	Rank	Mean	SD	Rank	Mean	SD	Rank
1	0.192	0.239	4	0.627	0.099	2	0.393	0.107	2	9.8	1.3	1
2	0.201	0.187	3	0.645	0.071	1	0.410	0.061	1	9.6	1.1	2
3	0.202	0.195	2	0.585	0.093	4	0.364	0.052	5	8.4	1.7	3
4	0.046	0.064	8	0.423	0.066	8	0.272	0.049	7	6.0	1.0	6
5	0.171	0.177	5	0.476	0.154	7	0.267	0.041	8	5.6	1.1	8
6	0.145	0.156	6	0.490	0.095	6	0.340	0.088	6	6.2	0.8	5
7	0.093	0.108	7	0.590	0.100	3	0.380	0.072	3	6.8	0.5	4
8	0.317	0.216	1	0.585	0.140	4	0.374	0.133	4	6.0	0.8	6

Due to their formulations, desirability function is much more demanding than utility function. For example, samples 1.3, 1.5, 2.3, 2.4 and 7.2 show no desirability at all but show high utility. This is due to the fact that a low value for a single (or few) attribute is enough to strongly affect desirability. Sample 7.2 has low values for CPOM/FPOM and life cycle, while top-down control value is 0.50 and, so, very different from the 0.15 threshold. Sample 1.3 has CPOM/FPOM and benthic food = 0, while sample 1.5 has top-down control and benthic food <0.01. Sample 2.3 has low SPOM/BPOM and benthic food = 0, while in sample 2.4, attributes are far from the threshold in four cases. Generally, desirability values are quite low: 0.611 for the first ranked sample (8.5). According to the desirability scale proposed by Harrington (1965), ranging from 0 to 1, this is “acceptable but poor”. Quality level is then to be considered “borderline” starting from sample 2.2, ranked fifth, and “unacceptable” from sample 1.2, ranked tenth. For all the ranked objects, the most limiting criteria are those derived from CPOM/FPOM, SPOM/BPOM and, especially, benthic food. In fact, CPOM/FPOM mean value for all samples is 0.69, but median is 0.18. Suspended-benthic fine particulate organic matter mean value is 0.27 and median is 0.12. The lowest values are calculated for benthic food, whose mean and median are 0.11 and 0.05, respectively. Thus, samples with a single criterion equal to 0, present at least once in all sites except for site 8, show the same desirability (=0). This, as well as the high frequency of very low values, leads to non-significant differences in desirability between sites (ANOVA,  $P = 0.48$ ).

Evaluation provided by utility function is much less severe than the desirability one: the overall quality of a site can be considered good even if the value of a single criterion is very low. Maximum utility value (sample 1.2) is 0.773 and mean utility value of all samples is 0.549, while mean desirability value is 0.168. Best mean utility belongs to site 2 (0.645), while worst belongs to site 4 (0.423). ANOVA applied to compare utility values among sites shows significant differences ( $F = 2.75$ ,  $P = 0.026$ ).

Results provided by dominance function are based on pair comparison of samples and allow to rank the examined samples but not to give a significative classification. All mean dominance values are comprised in the short range between 0.267 and 0.410. Thus, ANOVA shows that differences in dominance values between sites are not significant ( $F = 2.29$ ,  $P = 0.055$ ).

Extended biotic index values are poorly related to ranking functions: particularly, some samples show no desirability at all (1.5, 2.4 and 3.4) but a very high EBI value and, thus, a high taxonomical quality. In contrast, some samples (8.5, 5.5 and 3.3) have the highest desirability but poor taxonomical quality. Sample 5.5 has the worst EBI result (4) but it is ranked third in desirability, though its value is not very high in the scale (<0.50). Thus, correlation between EBI and desirability is very low ( $r = 0.152$ ). Correlations between EBI and utility ( $r = 0.412$ ) and between EBI and dominance ( $r = 0.519$ ) are higher, but it is still possible to say that the methods provide different information.

From Table 4, it can be observed that site 4 has the lower desirability and utility rank; it is seventh using dominance and sixth according to EBI. In contrast, site 8, having the same mean EBI value as site 4 (EBI rank = 6), is ranked first

using mean desirability function. Generally, sites 4, 6 and 7 show lower desirability and also (Table 2) lower ratio between observed flows and mean expected annual flows, which indicates a hydraulic stress affecting macroinvertebrate assemblages. Site 8, affected by considerable polluting loads, shows a low taxonomical diversity (EBI = 6.0, taxa richness =  $8.5 \pm 1.7$ , see Table 3). However, a greater and more constant flow, due to groundwater resurgence and return flow from irrigation net, causes an increase in the observed/expected flow ratio. This is probably why site 8 gets a good ranking according to functional ecology criteria (desirability = 1; utility = 4; dominance = 4). Upland sites (1, 2 and 3) always show high position ranks. This result is achieved notwithstanding samples presenting Desirability equal to 0.

## 6. WHAT WE OBTAINED USING THIS METHOD?

### 6.1 River evaluation and restoration

Rivers in Lombardy, and particularly Serio River, are object of intense study and restoration efforts, and, thus, decision makers are continuously looking for new assessment tools. Traditional monitoring using macroinvertebrates is based on the evaluation of taxonomical diversity and sensitivity, but these premises can lead to incomplete knowledge of ecological patterns and to questionable decisions. To improve knowledge about the river ecosystem, we applied a method developed by Merritt et al. (2002), based on functional ecology analysis. Ecosystem attributes derived in this way from invertebrate trait analysis are numerous and often conflicting. Thus, an MCDM technique was necessary to rank river sites and set intervention priorities. The use of these tools hopefully can lead to some advantages in river management efforts. As an example, resources allocated to rehabilitation or conservation programs can be driven to sites ranked the best, because a good functionality is a premise of restoration success. Vice versa, knowledge of functional efficiency levels allows establishing how it will be difficult to achieve planned results for a given site and allocating the proper resources for it. Further, comprehension of functional analysis leads to faster decision-making capability. Ecosystem attributes are measurements of nutritional resource availability, trophic structure and habitat suitability, and, thus, recording a deficit in one or more criteria allows designing interventions focused on their recovery.

Ecosystem attribute ranking in Serio River sites suggested that the entire watercourse is more or less impaired. The best conditions can be found in the montane valley sector (S1), represented by the first three sampling sites, and in the lowland terminal sector S4, described by site 8. All these sites presented, however, a few ecosystem attributes with poor quality. Therefore, recommendations can be made based on ranking and ecological attribute analysis. For initial rehabilitation efforts, we suggest selecting those sites ranked the best, previously cited, because they are likely to reach full recovery with low amount of resources and can increase the recolonization potential along the river. Moreover, we

suggest orienting restoration on those interventions that can lead to the recovery of specific, poor-rated ecosystem attributes. In the uplands, efforts should be made to increase benthic food availability (thus improving sprawler FHG typical habitats) and SPOM availability. Invertebrate assemblages appear to be established, slow-colonizing populations, and this should be taken in account when planning interventions requiring great physical alteration. In site 1, intervention should also be made to increase CPOM availability. Downstream interventions should pursue primarily the increase of P/R ratio and benthic food availability. This could be achieved by creating habitat structures in slow flow areas (such as artificial oxbows) to act as refugia and by allowing riparian vegetation colonization. Restoration of central sectors (sites 4–7) appears to be more difficult, because interventions involve increasing stream flow, bank quality and retention structure quantity, as well as upgrading river continuum, broken by too many dams.

Differences between taxonomical and functional approaches involve primarily sites such as 4 and 8. Site 4 presents low pollution but great hydrological alterations. Its EBI value, used to evaluate taxonomical quality, is poor but similar to the other sites (from 5 to 8) presenting evidences of pollution. In contrast, site 4 shows the worst functional quality. Site 8 is polluted and presents low EBI values (equal to site 4), but its hydrology is not so altered. Its ranking based on functional attributes is very good.

Generally, sites affected by pollution (including site 8) show low EBI values, according to the fact that the index was ideated to identify primarily organic pollution. However, ecosystem attributes in Serio River are more suitable to be modified by hydrological perturbations.

Thus, multi-criteria functional ecology analysis can spot out major differences if compared with the taxonomical index: it can quantify alteration in apparently unimpaired sites (like 1, 2 and 3) and it can record valuable characteristics in the ecological condition of apparently heavily impaired sites (like 8). This kind of information is obviously important for river restoration planning.

## 6.2 Final considerations

Functional ecology analysis, based on macroinvertebrate traits and derived ecosystem attributes, is a promising tool to increase river monitoring and assessment capability. Ordering such attributes with MCDM methods is one of the possible ways to analyze data and to get an overview over the elements of the system. Desirability, utility and dominance functions were applied and helped providing different information if compared to the taxonomical index EBI. Desirability, being a geometrical mean, is strict: if a single criterion value is poor, the overall desirability will be poor. This can lead to insignificant differences between monitored sites, as happened in Serio River bioassessment program.

Utility and dominance functions applied to river site lead to similar results: values can differ a lot, but final ranking does not show all the differences from EBI it has with desirability. However, utility showed significant differences between sites. Desirability has been useful in showing alteration in apparently unimpaired sites; it can be used as a warning tool to spot out single ecological

attribute that need attention. Utility seems to be a more suitable method for ranking of sites, due to its lesser exigency.

Ecosystem attributes ranked with MCDM methods have a different answer to hydrological alterations and pollution. This suggests that the method could spot differences between press (such as WWTP-treated effluent discharge) and pulse (such as periodic flow fluctuations and droughts) perturbations. Thus, we suggest to use these techniques in coincidence with taxonomy-based indexes to increase overall explanatory ability.

The critical features of this approach to decision-making problems are three. First of all, it is necessary to establish a correct relationship between criteria and ranking function values. Further application of these methods will be necessary to understand if the proposed thresholds, adopted from literature, and describing functions are the most proper. Second, it will be necessary to weight every criterion in order to take into account their relative importance in the decision rule. Some of the criteria are likely less important than others, and definition of weights could be crucial in better understanding the situation and redirecting restoration efforts. Finally, presented ranking methods do not dialogue with any criterion value that is over the used threshold: thus, very different attribute values can lead to the same result, and this can be somehow misleading, especially in unimpaired and least-altered reference sites. However, ranking of ecosystem attribute ratios obtained by invertebrate traits analysis is a river assessment tool, which is both promising and inexpensive; it is worthy of further research, by using different MCDM techniques and applying multi-variate statistical tools to larger data sets, to explore relationships between obtained values and the corresponding environmental variables.

## REFERENCES

- Bash, J.S., Ryan, C.M. (2002). Stream restoration and enhancement projects: Is anyone monitoring? *Environ. Manage.* 29(6), 877–885. DOI: 10.1007/s00267-001-0066-3.
- Benke, A.C., Huryn, A.D., Smock, L.A., Wallace, J.B. (1999). Length-mass relationships for freshwater macroinvertebrates in North America with particular reference to the southeastern United States, *JNABS* 18(3), 308–343.
- Brüggemann, R., Oberemm, A., Steinberg, C. (1997). Ranking of aquatic effect tests using Hasse Diagrams, *Toxicol. Environ. Chem.* 63, 125–139.
- Brüggemann, R., Pudenz, S., Voigt, K., Kaune, A., Kreimes, K. (1999). An algebraic/graphical tool to compare ecosystems with respect to their pollution IV: Comparative regional analysis by Boolean arithmetics, *Chemosphere* 38, 2263–2279.
- Burgherr, P., Meyer, E.I. (1997). Regression analysis of linear body dimensions vs. dry mass in stream macroinvertebrates, *Archiv für Hydrobiologie* 139(1), 101–112.
- Cummins, K.W. (1974). Structure and function of stream ecosystems, *BioScience* 24, 631–641.
- Dolédec, S., Phillips, N., Scarsbrook, M., Riley, R.H., Townsend, C.R. (2006). Comparison of structural and functional approaches to determining landuse effect on grassland stream invertebrate communities, *JNABS* 25(1), 44–60.
- Dolédec, S., Statzner, B., Bournaud, M. (1999). Species traits for future biomonitoring across ecoregions: Patterns along a human-impacted river, *Freshw. Biol.* 42, 737–758.
- Finogenova, N.P. (1984). Growth of *Stylaria lacustris* (L.) (Oligochaeta, Naididae). *Hydrobiologia* 115, 105–107.

- Ghetti, P.F. (1997). *Indice Biotico Esteso (IBE) – Manuale di Applicazione*, Provincia Autonoma, Trento.
- Ghilarov, A.M. (2000). Ecosystem functioning and intrinsic value of biodiversity, *Oikos* 90, 408–412.
- Hale, C.M., Reich, P.B., Frelich, L.E. (2004). Allometric equations for estimation of ash-free dry mass from length measurements for selected European earthworm species (Lumbricidae) in the western great lakes region, *Am. Midl. Nat.* 151(1), 179–185.
- Harrington, E.C. (1965). The desirability function, *Industrial Quality Control*, 21, 494–498.
- Hellawell, J.M. (1989). *Biological Indicators of Freshwater Pollution and Environmental Management*, Elsevier Applied Science, London. Chemom. Intell. Lab. Syst.
- 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.
- Karr, J.R. (1991). Biological integrity: A long-neglected aspect of water resource management, *Ecol. Appl.* 1, 66–84.
- Karr, J.R., Chu, E.W. (1999). *Restoring Life in Running Waters*, Island Press, Washington (DC).
- Keller, H.R., Massart, D.L. (1991). Multicriteria decision making: A case study, *Chemom. Intell. Lab. Syst.* 11, 175–189.
- Malmqvist, B. (2002). Aquatic invertebrates in riverine landscapes, *Freshw. Biol.* 47, 679–694.
- Merritt, R.W., Cummins, K.W. (eds) (1996). *An Introduction to the Aquatic Insects of North America* 3rd ed., Kendall-Hunt, Dubuque, Iowa.
- Merritt, R.W., Higgins, M.J., Cummins, K.W., Van den Eeden, B. (1999). The Kissimmee River-riparian marsh ecosystem, Florida: Seasonal differences in invertebrate functional feeding group relationships. In: *Invertebrates in Freshwater Wetlands in North America: Ecology and Management* (Batzler, D.P., Rader, R.B., Wissinger, S. eds), John Wiley and Sons, New York.
- Merritt, R.W., Wallace, J.R., Higgins, M.J., Alexander, M.K., Berg, M.B., Morgan, W.T., Cummins, K.W., VandenEeden, B. (1996). Procedures for the functional analysis of invertebrate communities of the Kissimmee River-Floodplain ecosystem, *Florida Scientist* 59, 216–274.
- Merritt, R.W., Cummins, K.W., Berg, M.B., Novak, J.A., Higgins, M.J., Wessell, K.J., Lessard, J.L. (2002). Development and application of a macroinvertebrate functional-group approach in the bioassessment of remnant river oxbows in Southwest Florida, *JNABS* 21(2), 290–310.
- Meyer, E.I. (1989). The relationship between body length parameters and dry mass in running water invertebrates, *Archiv für Hydrobiologie* 117(2), 191–203.
- Minshall, G.W., Petersen, R.C., Bott, T.L., Cushing, C.E., Cummins, K.W., Vannote, R.L., Sedell, J.R. (1992). Stream ecosystem dynamics of the Salmon River, Idaho: an 8th-order system, *JNABS* 11, 111–137.
- Newman, A. (1995). Ranking pesticides by environmental impact, *Environ. Sci. Technol.* 29, 324–326.
- Nyström, P., Pérez, J.R. (1998). Crayfish predation on the common pond snail (*Lymnaea stagnalis*): The effect of habitat complexity and snail size on foraging efficiency, *Hydrobiologia* 368, 201–208.
- Paoletti, A., Becciu, G. (2006). *Stima delle portate e delle precipitazioni e strumenti per la loro regionalizzazione – Allegato 2 alla Relazione generale del Programma di tutela e uso delle acque della Regione Lombardia*, Regione Lombardia-Direzione Generale Reti e Servizi di Pubblica Utilità, Milano.
- Pavan, M. (2003). Total and Partial Ranking Methods in Chemical Sciences. Ph.D. Thesis in Chemical Sciences, University of Milano-Bicocca, Milano.
- Pudenz, S., Brüggemann, R., Komoßa, D., Kreimes, K. (1997). An algebraic/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.
- Rosenberg, D.M., Resh, V.H. (1993). *Freshwater Biomonitoring and Benthic Macroinvertebrates*, Chapman and Hall, New York.
- Sansoni, G. (1992). *Atlante per il riconoscimento dei macroinvertebrati dei corsi d'acqua italiani*, Provincia Autonoma, Trento.
- Sørensen, P.B., Mogensen, B.B., Gyldenkerne, S., Rasmussen, A.G. (1998). Pesticides leaching assessment method for ranking both single substances and scenarios of multiple substance use, *Chemosphere* 36, 2251–2276.
- Statzner, B., Bady, P., Dolédec, S., Scholl, F. (2005). Invertebrate traits for the biomonitoring of large European rivers: An initial assessment of trait patterns in least impacted river reaches, *Freshw. Biol.* 50, 2136–2161. DOI: 10.1111/j.1365-2427.2005.01447.x.

- Statzner, B., Bis, B., Dolédec, S., Usseglio-Polatera, P. (2001). Perspectives for biomonitoring at large spatial scales: A unified measure for the functional composition of invertebrate communities in European running waters, *Basic Appl. Ecol.* 2, 73–85.
- Stoffels, R.J., Karbe, S., Paterson, R.A. (2003). Length-mass models for some common New Zealand littoral-benthic macroinvertebrates, with a note on within-taxon variability in parameter values among published models, *N.Z.J. Mar. Freshw. Resour.* 37, 449–460.
- Stone, M.K., Wallace, J.B. (1998). Long-term recovery of a mountain stream from clear-cut logging: The effects of forest succession on benthic invertebrate community structure, *Freshw. Biol.* 39, 151–169.
- Tachet, H., Richoux, P., Bournaud, M., Usseglio-Polatera, P. (2000). *Invertébrés d'eau douce*, CNRS Editions, Paris.
- Towers, D.J., Henderson, I.M., Veltman, C.J. (1994). Predicting dry weight of New Zealand aquatic macroinvertebrates from linear dimensions, *N.Z.J. Mar. Freshw. Resour.* 28, 159–166.
- Townsend, C.R., Hildrew, A.G. (1994). Species traits in relation to a habitat templet for river systems, *Freshw. Biol.* 31, 265–275.
- Vugteveen, P., Leuven, R.S.E.W., Huijbregts, M.A.J., Lenders, H.J.R. (2006). Redefinition and elaboration of river ecosystem health: Perspective for river management, *Hydrobiologia* 565, 289–308. DOI: 10.1007/s10750-005-1920-8.
- Witter, J.V., van Stokkom, H.T.C., Hendriksen, G. (2006). From river management to river basin management: A water manager's perspective, *Hydrobiologia* 565, 317–325. DOI: 10.1007/s10750-005-1922-6.
- Woodiwiss, F.S. (1978). *Comparative Study of Biological-Ecological Water Quality Assessment Methods. Second Practical Demonstration*. Summary Report. Commission of the European Communities.

## APPENDIX

Traits for invertebrate taxa sampled in Serio River

Taxon (usual recon. level)	Dry mass <sup>a</sup>	Trait		Voltinism		Drift	
		FFG	FHG	GEN/Y≤1	GEN/Y>1	Accidental	Behavioural
Plecoptera (genus)							
<i>Isoperla</i>	1	PRED	CLING	1	0	0.5	0.5
<i>Perla</i>	1	PRED	CLING	1	0	0.33	0.67
<i>Cloroperla</i>	1	PRED	CLING	1	0	0.6	0.4
<i>Leuctra</i>	1	LVP-SHRED	CLING	1	0	0.5	0.5
<i>Protonemura</i>	1	D-SHRED	SPRAW	1	0	0.5	0.5
<i>Siphonoperla</i>	1	PRED	CLING	1	0	0.5	0.5
<i>Capnia</i>	2	LVP-SHRED	CLING	1	0	0.33	0.67
<i>Amphinemura</i>	1	D-SHRED	SPRAW	1	0	0.5	0.5
<i>Nemoura</i>	1	D-SHRED	SPRAW	1	0	0.5	0.5
Ephemeroptera (genus)							
<i>Ecdyonurus</i>	1	SCRAP	CLING	1	0	0.75	0.25
<i>Rithrogena</i>	1	SCRAP	CLING	1	0	0.6	0.4
<i>Ephemerella</i>	1	G-COLL	CLING	0.75	0.25	0.4	0.6
<i>Procleon</i>	1	G-COLL	SWIM	0	1	0.33	0.67
<i>Epeorus Alpicola</i>	1	G-COLL	CLING	1	0	0.67	0.33
<i>Habroleptoides</i>	1	G-COLL	CLING	1	0	0.67	0.33
<i>Caenis</i>	3	G-COLL	SPRAW	0.25	0.75	0.67	0.33
<i>Baetis</i>	1	G-COLL	SWIM/CLING	0.40	0.60	0.6	0.4
<i>Habrophlebia</i>	2	D-SHRED	CLING	1	0	0.5	0.5
<i>Paraleptophlebia</i>	1	G-COLL	CLING	1	0	0.67	0.33



Trichoptera (family)							
Hydropsichidae	1	F-COLL	CLING	0.5	0.5	0.6	0.4
Limnaephilidae	3	D-SHRED	CLIMB	1	0	0.5	0.5
Rhyacophilidae	1	PRED	CLING	0.75	0.25	0.6	0.4
Sericostomatidae	1	SHRED/PRED	SPRAW	0.83	0.17	0.67	0.33
Philopotamidae	1	F-COLL	CLING	1	0	0.5	0.5
Glossosomatidae	1	SCRAP	CLING	0.67	0.33	0.33	0.67
Goeridae	3	SCRAP	CLING	1	0	0.5	0.5
Coleoptera (family)							
Dytiscidae	4	PRED	SWIM	1	0	0	1
Hydrophilidae	4	G-COLL	SWIM	1	0	0	1
Elminthidae	4	SCRAP	CLING	1	0	0.67	0.33
Hydraenidae	1	SCRAP	CLING	0.75	0.25	0.67	0.33
Odonata (genus)							
<i>Onychogomphus</i>	2	PRED	BURW	1	0	0.67	0.33
<i>Orthetrum</i>	2	PRED	SPRAW	1	0	0.5	0.5
<i>Crocothemis</i>	2	PRED	SPRAW	0.33	0.67	0	1
Diptera (family)							
Chironomidae	2	G-COLL/PRED	BURW	0.25	0.75	0.67	0.33
Simuliidae	2	F-COLL	CLING	0.40	0.60	0.5	0.5
Limoniidae	2	SHRED/PRED	BURW	0.75	0.25	0	1
Tabanidae	2	PRED	SPRAW	1	0	0	1
Tipulidae	2	SHRED/PRED	BURW	0.80	0.20	0	1row>
Athericidae	2	PRED	SPRAW	1	0	0	1
Psychodidae	2	G-COLL	BURW	0.25	0.75	0.75	0.25
Dixidae	2	F-COLL	SWIM	1	0	0.75	0.25
Rhagionidae	2	PRED	SPRAW	1	0	0	0
Anthomyiidae	2	PRED	SPRAW	0.25	0.75	0	1

Traits for invertebrate taxa sampled in serio River (*Continued*)

Taxon (usual recon. level)	Dry mass <sup>a</sup>	Trait		Voltinism		Drift	
		FFG	FHG	GEN/Y $\leq$ 1	GEN/Y $>$ 1	Accidental	Behavioural
Heteroptera (genus)							
<i>Nepa</i>	2	PRED	CLIMB	1	0	0.25	0.75
Crustacea (family)							
Gammaridae	2	D-SHRED	SWIM	0	1	0.6	0.4
Asellidae	2	D-SHRED	SPRAW	0.25	0.75	1	0
Gastropoda (genus)							
<i>Lymnaea</i>	5	SCRAP	CLING	1	0	0.67	0.33
<i>Physa</i>	5	SCRAP	CLING	1	0	0.67	0.33
<i>Ancylus</i>	5	SCRAP	CLING	1	0	0.67	0.33
<i>Planorbis</i>	5	SCRAP	CLING	0.75	0.25	0.67	0.33
Triclada (genus)							
<i>Dugesia</i>	3	PRED	SPRAW	1	0	0	1
<i>Polycelis</i>	3	PRED	SPRAW	1	0	0.33	0.67
Hirudinea (genus)							
<i>Erpobdella</i>		PRED	SPRAW	1	0	0.5	0.5
<i>Helobdella</i>		PRED	SPRAW	0.75	0.25	0.5	0.5
<i>Hemiclepsis</i>		PRED	SPRAW	1	0	0.75	0.25
<i>Dyna</i>		PRED	SPRAW	1	0	0.5	0.5
<i>Haemopis</i>		PRED	SPRAW	1	0	0.5	0.5

Oligochaeta (family)							
Naididae	6	G-COLL	BURW	0	1	0.67	0.33
Lumbricidae	7	G-COLL	BURW	1	0	1	0
Tubificidae	8	G-COLL	BURW	0	1	1	0
Lumbiculidae	7	G-COLL	BURW	0	1	1	0
Nematoda (family)							
Mermithidae	8	PRED	BURW	1	0	0.67	0.33

Functional feeding groups (FFG): D-SHRED = CPOM shredders; LVP-SHRED = live vascular plant shredders; SCRAP = scrapers; F-COLL = filtering collectors; G-COLL = gathering collectors; PIERC = piercers; PRED = predators and parasites. Functional habit group (FHG): CLING = clingers; CLIMB = climbers; SPRAW = sprawlers; BURW = burrowers; SWIM = swimmers. Voltinism:  $GEN/Y \leq 1$  = univoltine or less;  $GEN/Y > 1$  = multi-voltine. Drift: Accidental = propension to accidental drifting; behavioural = propension to behavioural drifting.

<sup>a</sup>Literature used for length/mass conversion: (1) [Burgherr & Meyer \(1997\)](#); (2) [Benke et al. \(1999\)](#); (3) [Meyer \(1989\)](#); (4) [Towers et al. \(1994\)](#); (5) [Nyström & Pérez \(1998\)](#); (6) [Finogenova \(1984\)](#); (7) [Hale et al. \(2004\)](#); (8) [Stoffels et al. \(2003\)](#).

## The DART (Decision Analysis by Ranking Techniques) Software

**A. Manganaro, D. Ballabio, V. Consonni, A. Mauri, M. Pavan,  
and R. Todeschini**

---

Contents	1. Introduction	193
	2. The DART Software	194
	2.1 Data management and transformation	194
	2.2 Data pre-processing	195
	2.3 Total ranking	197
	2.4 Partial ranking	197
	3. Example of Application of the Dart Software	197
	4. Conclusions	207
	References	207

---

### 1. INTRODUCTION

As more and more chemical and toxicological data are available, both derived from experimental analysis or calculated by means of quantitative structure–activity relationship (QSAR) approaches, there is a strong need for computer-aided tools designed to understand and interpret these data: this is the well-known issue of multicriteria decision making. It is often impossible to catch the meaning of complex and multivariate datasets and consequently to take critical decisions deriving from their analysis. This is true in particular for what concerns data aimed at environmental and human hazard assessment, a topical field as the new EU legal framework named REACH (Registration Evaluation Authorization of Chemicals) brings the strong need of reliable methods to evaluate chemicals ([European Commission, 2003](#); [Worth et al., 2004](#)).

The field of multicriteria decision making is mostly based on ranking methodologies, which allow to obtain a simple ranking of objects from a dataset

containing several criteria for each object. Thus, it is possible to evaluate the objects and take decisions based on their ranking, whereas this would be difficult considering only the original dataset, due to the complexity of reading and understanding a multivariate set.

Ranking methods, as already shown in other chapters of this book, can be roughly divided into total and partial ranking methods. Both of them are based on elementary methods of Discrete Mathematics. The main difference is the introduction of the “not comparable” relationship between objects in the partial-order techniques, while in the total-order techniques, the result is always a ranking value for each object. This is made clear in the Hasse diagram, the graphical representation of the mathematical results of the partial-order method, which allows to easily understand the relationships between objects.

Furthermore, data pre-processing is required to establish an adequate data matrix and obtain a useful evaluation from a ranking technique; obviously pre-processing may influence and change the results significantly. Well-known statistical techniques, like clustering or principal component analysis (PCA), can provide a satisfactory solution to those drawbacks related to noise and measurement error and can help to perform a suitable and useful ranking analysis of the desired data.

Talete srl developed on behalf of the Joint Research Center a software called DART (Decision Analysis by Ranking Techniques), which implements several ranking methods, useful in the work of extracting information for assessment purposes from any kind of dataset. Besides applying ranking methods, DART also allows to perform several pre-processing methods. The Joint Research Center made DART available for free on its website <http://ecb.jrc.it/>.

The theory for total and partial ranking methods being used in DART is not reported, as it has already been deeply discussed in other chapters of this book.

## **2. THE DART SOFTWARE**

DART is a toolbox designed to import and pre-process data and to obtain a ranking of the dataset thus gaining useful information on decision making. It implements several total ranking methods and a partial ranking method (the Hasse diagram technique). It is important to underline that in the work flow of the application, the management and pre-processing of data is a relevant issue, as this part of the software allows the user to have great flexibility on the analysis and to apply his or her knowledge of the meaning of data.

### **2.1 Data management and transformation**

The first step for using DART consists in the loading and setting of data. The import procedure reads plain-text files representing the dataset fields separated by tab, commas or other special characters and then allows the user to select the data to be imported (thus cutting text string with non-numeric information) as a

matrix of  $n$  objects (rows) and  $p$  variables (columns). Once a dataset is imported, it can be saved and loaded with the DART format, which stores information on setting of the data along with the dataset itself.

After a proper dataset is loaded, a setup must be performed. This section is necessary before performing any ranking evaluation. The setup form is made of several tabs. The first one reports a summary of the current settings. The second one allows to select which variables shall be used for the subsequent analysis and which are to be excluded or considered as class variables; in this tab, if some missing values are present, a further frame will be visible giving the chance to replace them with a chosen value (random value, average, minimum or maximum value). The third tab similarly shows all objects and allows to select which ones are to be included for the analysis. The fourth tab is dedicated to the relevant issue of transform functions. In fact, for mathematical calculation of the ranking methods, it is necessary to scale all values of the original dataset, and this is made by applying a transform function, which takes the values from the original ones into the domain between 1 and 0. By choosing functions different from the linear one, it is possible to assign a different “meaning” to the selected variable, depending on its physical meaning. Furthermore, in this tab it is possible to set different weights to variables, resulting in a different importance of the variables.

## 2.2 Data pre-processing

DART gives the chance to apply several pre-processing methods, which can be fundamental to obtain useful results from the ranking analysis. Pre-processing methods are statistical techniques that have to be applied to the dataset before proceeding to the ranking analysis. Their purpose is to produce a better dataset without any relevant information loss. The concept of “better dataset” is strictly related to the type of the desired ranking analysis and to the peculiarities of the dataset itself. For example, PCA is a good solution to reduce the number of variables; clustering is instead a good way to reduce the number of elements; rounding or partitioning into bins can help to reduce incomparable objects in the Hasse diagram. The pre-processing menu can be accessed once a dataset is imported or loaded. The methods are divided into two categories: methods working on variables and methods working on objects.

### 2.2.1 Significant digits

The form for this method allows the user to choose how many decimal digits retain for each variable, thus rounding its values. This is useful especially for Hasse diagram analysis; in fact, by rounding values, it is possible to reduce the number of incomparable objects.

### 2.2.2 Bins partition

The method of bins partition consists in dividing the variable range into an arbitrary number of bins, i.e. regular intervals, chosen by the user and then

each value is set to the average value of the bin in which it falls. This operation makes the dataset more homogeneous, as similar values are set to an exact and equal value, thus resulting very useful for Hasse diagram.

### 2.2.3 Principal component analysis

The form for this method allows to apply a classical PCA on the raw data. Principal component analysis is a well-known procedure in multivariate statistics and transforms  $p$ -correlated variables into a set of orthogonal new variables that reproduces the original variance/covariance structure. This means rotating a  $p$ -dimensional space to achieve independence between variables. The new variables, called principal components (PCs), 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. Because of their properties, PCs 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. This means that it is possible, after having calculated the PCs, to choose only few of them and to project the original elements in their new space, obtaining a reduced dataset.

The results of the analysis are shown in different tabs, making possible to see the scree plot, the diagram of the explained variance of the PCs, the loadings and scores values and the loading and score plots for all PCs. In the last tab, the number of PCs to be retained can be chosen. This means that instead of performing analysis on the original dataset, a new dataset built with the projection of all the objects into the selected PCs will be used, thus allowing a reduction of dimensions without a loss of relevant information.

### 2.2.4 K-means clustering

Clustering consists in a partition of the elements into  $k$  clusters. This allows to retain representative data points (e.g. cluster centroids) of each cluster and to use them in order to represent all the elements belonging to that cluster, thus reducing the number of elements used in the ranking method.

DART performs clustering by means of  $K$ -means algorithm, which is an iterative procedure for the division of elements into  $k$  arbitrary clusters, starting from a random partition and then moving each element, during each iteration, to the closest cluster.

The form for this method allows the user to choose  $k$  number of clusters and the distance method (Euclidean, Mahalanobis and Chebyshev), then the results of the clustering are shown in a grid that reports all the centroid elements and the objects belonging to each cluster. If the results are accepted, a new dataset is built with all the centroids as objects, thus resulting in a reduction of the original object number. After the setup has been performed, it is possible to access to the ranking forms.

## 2.3 Total ranking

The total ranking form shows the results for all the total ranking methods implemented in DART, together with several charts and statistics that help the user in better understanding the obtained results.

The first tab of the form reports the numerical results for the implemented methods: desirability, utility, dominance, concordance, simple additive ranking (SAR), Hasse average ranking (HAR) and absolute reference ranking.

The second and third tabs report the histogram and the line plot for all the methods, in order to have a visual outlook of the results. The fourth tab shows the pareto plot for the desired methods, and the fifth tab shows a scatter plot in which it is possible to select two ranking methods and view how they behave. Finally, the last tab shows some useful statistics for each method, such as its stability and their degeneracy indices.

## 2.4 Partial ranking

Partial ranking form shows the mathematical results for partial ranking analysis, which leads to the Hasse diagram. In the first tab, some indices to evaluate the analysis performed are reported: CHI (comparability degree), YdbyR (discrimination power by ranking), selectivity  $T$  and diversity  $d$  (Pavan, 2003). Together with them, the list of maximal, minimal and isolated elements is reported. In the second tab, stability and degeneracy indices are reported. The third tab reports the level structure as it will be displayed in the Hasse diagram, i.e. the levels and the objects belonging to each level are shown. Finally, the fourth tab reports the complete Hasse matrix.

### 2.4.1 Hasse diagram

The Hasse diagram form shows the Hasse diagram together with its legend. When this form is active, it is possible to access the *diagram* menu, which consists of several visualization options for the diagram. Different draw modes are available, allowing to align objects in different ways (left align or symmetric); furthermore, it is possible to optimize the diagram, by automatically reducing the number of lines (links) crossing each object. Other options include the visualization of a grid, of the object labels, and the possibility of draw samples with different color on the basis of a defined class. The Hasse diagram can be also exported to the clipboard or saved as a jpeg image.

## 3. EXAMPLE OF APPLICATION OF THE DART SOFTWARE

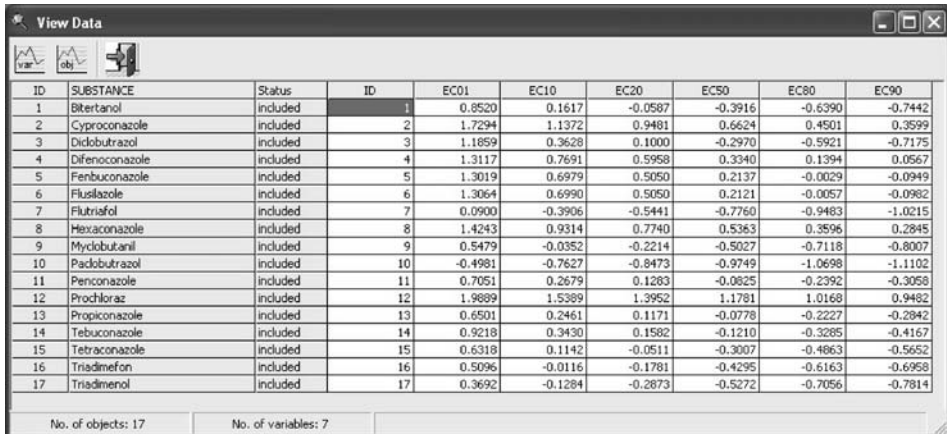
In order to show how DART works, a short example is explained on a dataset of 17 chemicals tested experimentally for their toxicity at 01, 10, 20, 50, 80 and 90 concentrations on *Scenedesmus vacuolatus*, extracted from a dataset by the BEAM EU project. Data are collected in Table 1 (European Communities, 2000).



**Table 1** Dataset of 17 chemicals with toxicity experimental data

Substance	ID	EC01	EC10	EC20	EC50	EC80	EC90
Bitertanol	1	0.8520	0.1617	−0.0587	−0.3916	−0.6390	−0.7442
Cyproconazole	2	1.7294	1.1372	0.9481	0.6624	0.4501	0.3599
Diclobutrazol	3	1.1859	0.3628	0.1000	−0.2970	−0.5921	−0.7175
Difenoconazole	4	1.3117	0.7691	0.5958	0.3340	0.1394	0.0567
Fenbuconazole	5	1.3019	0.6979	0.5050	0.2137	−0.0029	−0.0949
Flusilazole	6	1.3064	0.6990	0.5050	0.2121	−0.0057	−0.0982
Flutriafol	7	0.0900	−0.3906	−0.5441	−0.7760	−0.9483	−1.0215
Hexaconazole	8	1.4243	0.9314	0.7740	0.5363	0.3596	0.2845
Myclobutanil	9	0.5479	−0.0352	−0.2214	−0.5027	−0.7118	−0.8007
Paclobutrazol	10	−0.4981	−0.7627	−0.8473	−0.9749	−1.0698	−1.1102
Penconazole	11	0.7051	0.2679	0.1283	−0.0825	−0.2392	−0.3058
Prochloraz	12	1.9889	1.5389	1.3952	1.1781	1.0168	0.9482
Propiconazole	13	0.6501	0.2461	0.1171	−0.0778	−0.2227	−0.2842
Tebuconazole	14	0.9218	0.3430	0.1582	−0.1210	−0.3285	−0.4167
Tetraconazole	15	0.6318	0.1142	−0.0511	−0.3007	−0.4863	−0.5652
Triadimefon	16	0.5096	−0.0116	−0.1781	−0.4295	−0.6163	−0.6958
Triadimenol	17	0.3692	−0.1284	−0.2873	−0.5272	−0.7056	−0.7814

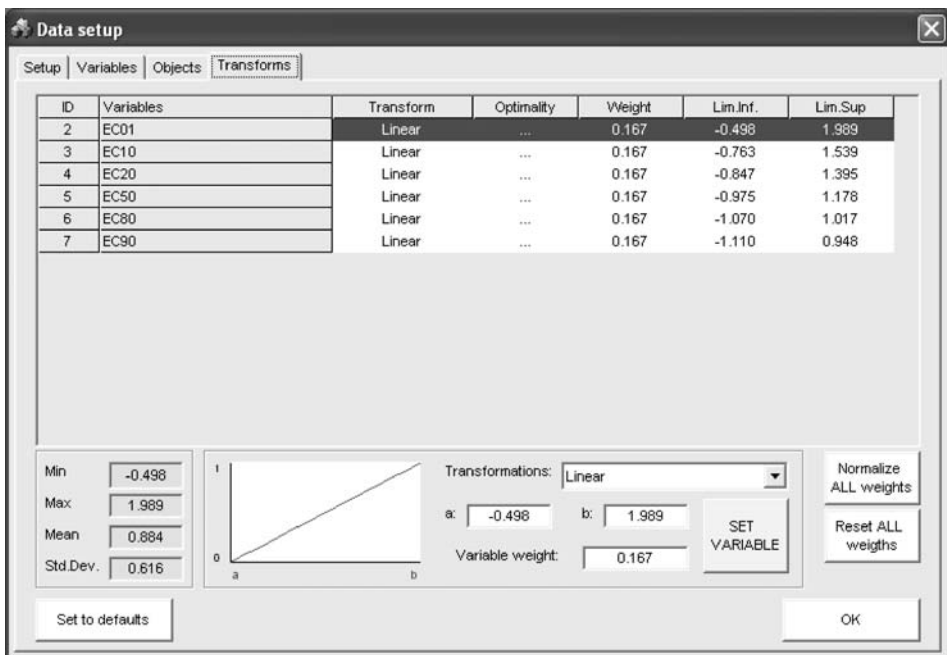
After performing the import procedure, the dataset is loaded into DART, and it is possible to see it by accessing the View voice in the Data menu (Figure 1). In the data setting, the variable “id” is removed from the dataset—as it just collects the sample label. Then, proper transform functions must be set (Figure 2). In



ID	SUBSTANCE	Status	ID	EC01	EC10	EC20	EC50	EC80	EC90
1	Biteranol	included	1	0.8520	0.1617	-0.0587	-0.3916	-0.6390	-0.7442
2	Cyproconazole	included	2	1.7294	1.1372	0.9481	0.6624	0.4501	0.3599
3	Didobutrazol	included	3	1.1859	0.3628	0.1000	-0.2970	-0.5921	-0.7175
4	Difenoconazole	included	4	1.3117	0.7691	0.5958	0.3340	0.1394	0.0567
5	Fenbuconazole	included	5	1.3019	0.6979	0.5050	0.2137	-0.0029	-0.0949
6	Flusilazole	included	6	1.3064	0.6990	0.5050	0.2121	-0.0057	-0.0982
7	Flutriafol	included	7	0.0900	-0.3906	-0.5441	-0.7760	-0.9483	-1.0215
8	Hexaconazole	included	8	1.4243	0.9314	0.7740	0.5363	0.3596	0.2845
9	Myclobutanil	included	9	0.5479	-0.0352	-0.2214	-0.5027	-0.7118	-0.8007
10	Paclobutrazol	included	10	-0.4981	-0.7627	-0.8473	-0.9749	-1.0698	-1.1102
11	Penconazole	included	11	0.7051	0.2679	0.1283	-0.0825	-0.2392	-0.3058
12	Prochloraz	included	12	1.9889	1.5389	1.3952	1.1781	1.0168	0.9482
13	Propiconazole	included	13	0.6501	0.2461	0.1171	-0.0778	-0.2227	-0.2842
14	Tebuconazole	included	14	0.9218	0.3430	0.1582	-0.1210	-0.3285	-0.4167
15	Tetraconazole	included	15	0.6318	0.1142	-0.0511	-0.3007	-0.4863	-0.5652
16	Triadimefon	included	16	0.5096	-0.0116	-0.1781	-0.4295	-0.6163	-0.6958
17	Triadimenol	included	17	0.3692	-0.1284	-0.2873	-0.5272	-0.7056	-0.7814

No. of objects: 17      No. of variables: 7

Figure 1 View data form.



ID	Variables	Transform	Optimality	Weight	Lim.Inf.	Lim.Sup
2	EC01	Linear	...	0.167	-0.498	1.989
3	EC10	Linear	...	0.167	-0.763	1.539
4	EC20	Linear	...	0.167	-0.847	1.395
5	EC50	Linear	...	0.167	-0.975	1.178
6	EC80	Linear	...	0.167	-1.070	1.017
7	EC90	Linear	...	0.167	-1.110	0.948

Min: -0.498    Max: 1.989    Mean: 0.884    Std.Dev.: 0.616

Transformations: Linear  
a: -0.498    b: 1.989    Variable weight: 0.167

Set to defaults    OK    Normalize ALL weights    Reset ALL weights

Figure 2 Transform tab in the setup form.

this example, all the variables are intended to have the same importance and consequently all weights were set to equal values; as no further knowledge is available about particular trends and physical meaning of the variables, a linear transformation is set for all of them; since the ranking here is intended from the less toxic to the most toxic chemicals, the linear transformations are set so that high values of the original variables bring high values in ranking (in this example, this means “more toxic”).

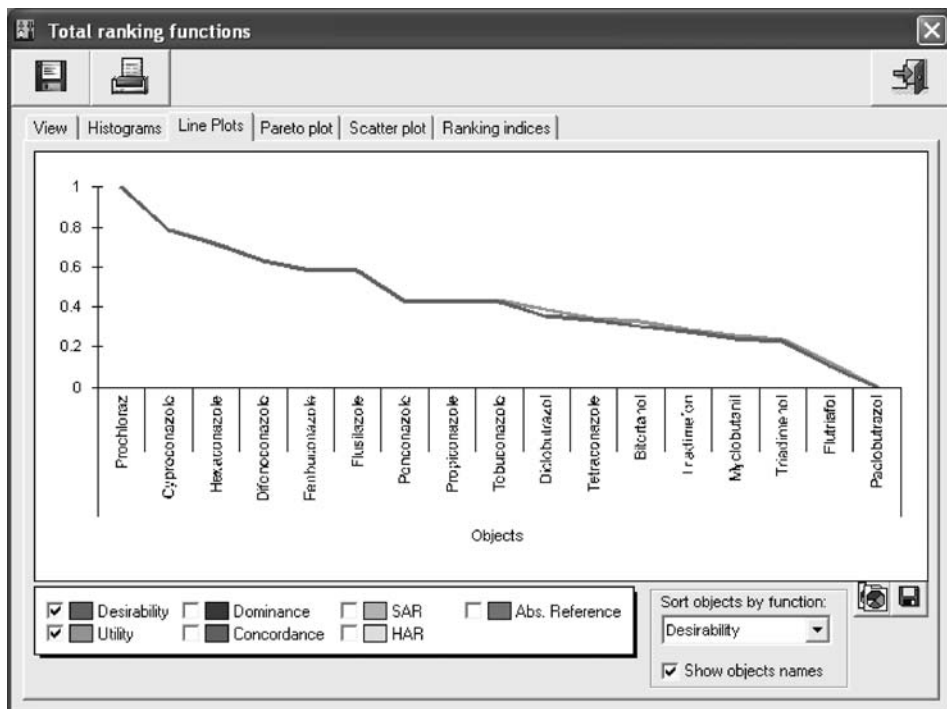
Once the setup has been performed, it is possible to access the ranking analysis. In Figure 3, the main tab of the total ranking form is reported, where all the results for different methods are reported. In the form, chemicals are sorted using values of the highlighted method (desirability); just by clicking on the desired method, chemicals are sorted on its values. In the following tabs, it is possible to check different graphs to evaluate and better understand the numerical values. For example, in Figure 4, it is reported the line plot tab, having chosen in the legend to show only desirability and utility results. It is noted that results are almost similar for both methods, and this is not always true. Desirability is more strict, as it is calculated as geometrical mean (thus a single bad ranking strongly contributes to a low desirability). In this example, just by viewing the

View | Histograms | Line Plots | Pareto plot | Scatter plot | Ranking indices

☐ sorting function: Desirability      reference object : Prochloraz

Rank	Objects	Desirability	Utility	Dominance	Concord	SAR	HAR	Abs.Ref.
1	Prochloraz	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	Cyproconazole	0.785	0.787	0.938	1.000	0.937	0.938	0.779
3	Hexaconazole	0.715	0.716	0.875	1.000	0.875	0.875	0.714
4	Difenoconazole	0.630	0.632	0.812	1.000	0.812	0.813	0.628
5	Fenbuconazole	0.581	0.586	0.714	0.845	0.729	0.732	0.579
6	Flusilazole	0.581	0.586	0.704	0.845	0.708	0.732	0.579
7	Penconazole	0.427	0.428	0.493	0.060	0.530	0.500	0.427
8	Propiconazole	0.425	0.426	0.493	0.062	0.530	0.500	0.425
9	Tebuconazole	0.424	0.431	0.510	0.206	0.541	0.538	0.426
10	Diclobutrazol	0.353	0.387	0.426	0.183	0.468	0.457	0.365
11	Tetraconazole	0.335	0.341	0.356	0.041	0.374	0.359	0.338
12	Bitertanol	0.302	0.325	0.305	0.181	0.332	0.339	0.314
13	Triadimefon	0.276	0.284	0.250	0.031	0.269	0.238	0.280
14	Myclobutanil	0.244	0.260	0.162	0.023	0.175	0.148	0.254
15	Triadimenol	0.228	0.236	0.138	0.025	0.144	0.136	0.233
16	Flutriafol	0.103	0.121	0.062	0.007	0.061	0.063	0.119
17	Paclobutrazol	0.000	0.000	0.000	0.000	0.000	0.000	0.000

**Figure 3** View tab in the total ranking form.



**Figure 4** Line plots tab in the total ranking form.

line plot it is clear that the two methods bring the same information, so no further analysis on which variables contributed most in a potential difference is needed.

In the partial ranking form, it is possible to view all mathematical information on the Hasse diagram and some indices to evaluate its stability and degeneracy. In most of the cases, it is more easy to look directly at the Hasse diagram (Figure 5). The quality of Hasse diagrams is strongly related to data pre-processing; in this example, the obtained diagram is good (i.e. brings a clear information) even without any form of pre-processing, as it shows a ranking with few incomparable objects (e.g. 5 and 6 or 9 and 16). There are a maximal and a minimal element, thus it is possible to define the most toxic (12, prochloraz) and the less toxic (10, paclobutrazol) chemicals.

However, a pre-processing of the dataset to obtain an even more clear diagram can be performed. For example, it is possible to apply a clustering to group chemicals with similar toxicity values and consequently reduce the number of objects to be ranked. The K-means algorithm was applied by accessing the K-means Clustering voice in the pre-processing menu; the resulting clustering, having chosen eight clusters, is summarized in the form shown in Figure 6.

After applying the method, the resulting clustering is shown in Table 2. The ranking analysis is thus made only on centroids, and the dataset is reduced to eight objects. The Hasse diagram for this new dataset results is more clear, as shown in Figure 7.

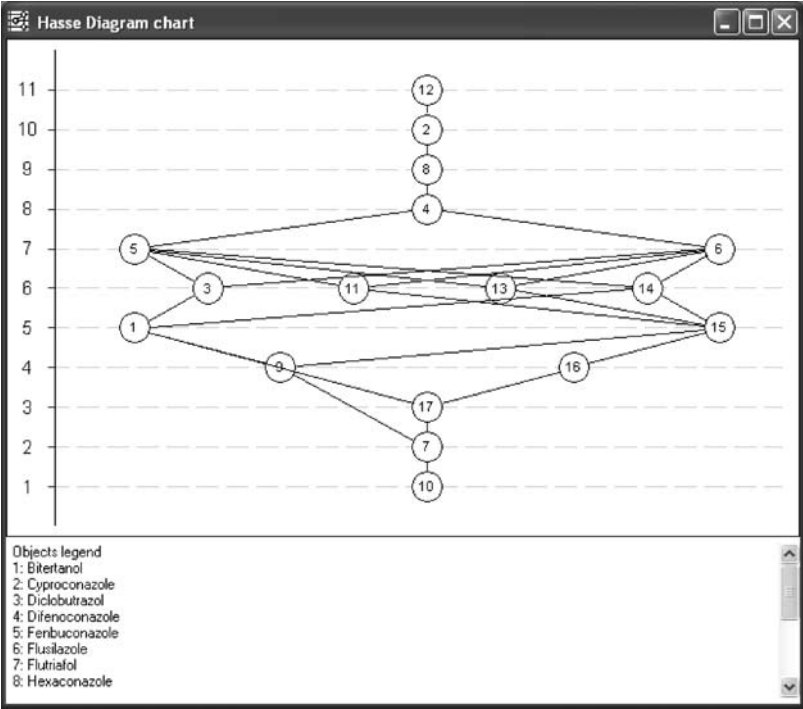


Figure 5 Hasse diagram chart form for the original example dataset.

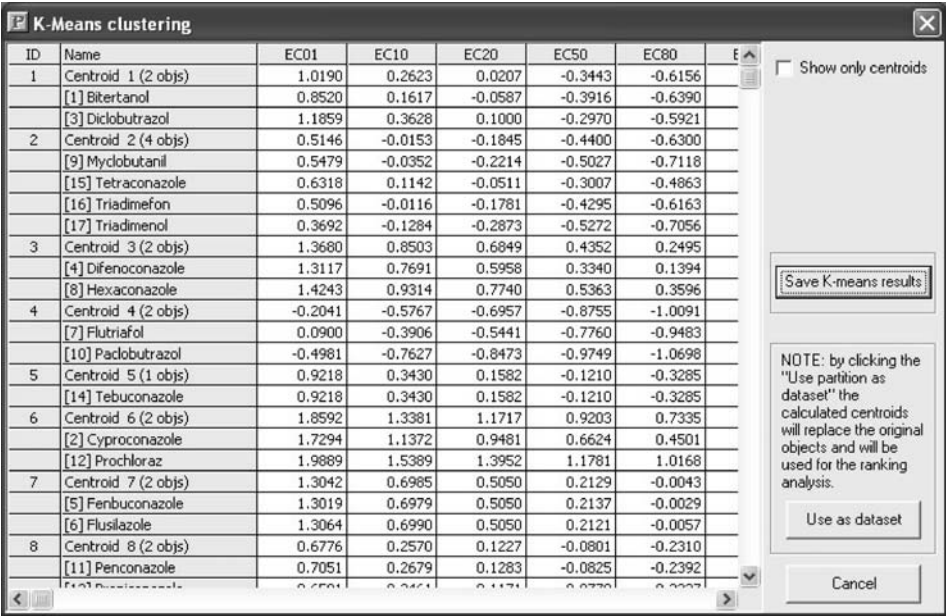


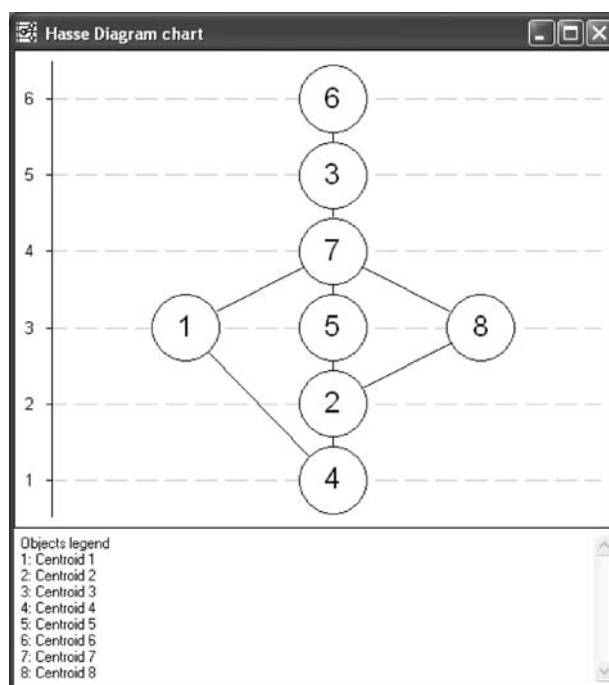
Figure 6 K-means clustering form.

**Table 2** Results of the K-means clustering

Object	EC01	EC10	EC20	EC50	EC80	EC90
<i>Cluster no. 1 (two objects)</i>						
Centroid 1	1.0190	0.2623	0.0207	−0.3443	−0.6156	−0.7309
Bitertanol	0.8520	0.1617	−0.0587	−0.3916	−0.6390	−0.7442
Diclobutrazol	1.1859	0.3628	0.1000	−0.2970	−0.5921	−0.7175
<i>Cluster no. 2 (four objects)</i>						
Centroid 2	0.5146	−0.0153	−0.1845	−0.4400	−0.6300	−0.7108
Myclobutanil	0.5479	−0.0352	−0.2214	−0.5027	−0.7118	−0.8007
Tetraconazole	0.6318	0.1142	−0.0511	−0.3007	−0.4863	−0.5652
Triadimefon	0.5096	−0.0116	−0.1781	−0.4295	−0.6163	−0.6958
Triadimenol	0.3692	−0.1284	−0.2873	−0.5272	−0.7056	−0.7814
<i>Cluster no. 3 (two objects)</i>						
Centroid 3	1.3680	0.8503	0.6849	0.4352	0.2495	0.1706
Difenoconazole	1.3117	0.7691	0.5958	0.3340	0.1394	0.0567
Hexaconazole	1.4243	0.9314	0.7740	0.5363	0.3596	0.2845
<i>Cluster no. 4 (two objects)</i>						
Centroid 4	−0.2041	−0.5767	−0.6957	−0.8755	−1.0091	−1.0659
Flutriafol	0.0900	−0.3906	−0.5441	−0.7760	−0.9483	−1.0215
Paclobutrazol	−0.4981	−0.7627	−0.8473	−0.9749	−1.0698	−1.1102
<i>Cluster no. 5 (one object)</i>						
Centroid 5	0.9218	0.3430	0.1582	−0.1210	−0.3285	−0.4167
Tebuconazole	0.9218	0.3430	0.1582	−0.1210	−0.3285	−0.4167

**Table 2** (Continued)

Object	EC01	EC10	EC20	EC50	EC80	EC90
<i>Cluster no. 6 (two objects)</i>						
Centroid 6	1.8592	1.3381	1.1717	0.9203	0.7335	0.6541
Cyproconazole	1.7294	1.1372	0.9481	0.6624	0.4501	0.3599
Prochloraz	1.9889	1.5389	1.3952	1.1781	1.0168	0.9482
<i>Cluster no. 7 (two objects)</i>						
Centroid 7	1.3042	0.6985	0.5050	0.2129	−0.0043	−0.0966
Fenbuconazole	1.3019	0.6979	0.5050	0.2137	−0.0029	−0.0949
Flusilazole	1.3064	0.6990	0.5050	0.2121	−0.0057	−0.0982
<i>Cluster no. 8 (two objects)</i>						
Centroid 8	0.6776	0.2570	0.1227	−0.0801	−0.2310	−0.2950
Penconazole	0.7051	0.2679	0.1283	−0.0825	−0.2392	−0.3058
Propiconazole	0.6501	0.2461	0.1171	−0.0778	−0.2227	−0.2842



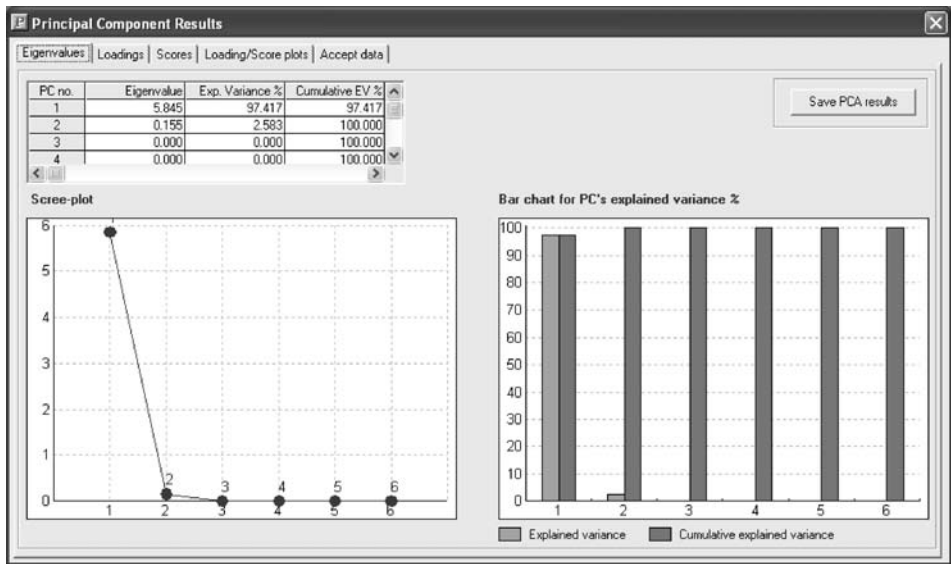
**Figure 7** Hasse diagram chart form for the example dataset after clustering.

Moreover, other approaches can be used to pre-process the data and obtain a better ranking output: reducing decimal digits or partitioning values into bins can help to reduce the number of incomparable objects. Furthermore, PCA can be performed to explore the dataset and better understand it.

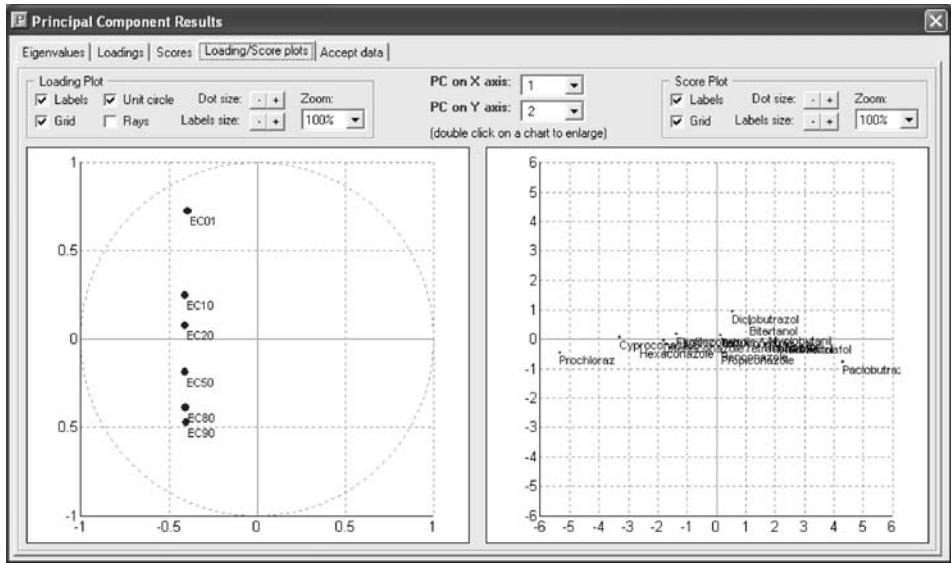
For example, the results of PCA made on the original dataset are shown in [Figures 8 and 9](#), reporting two tabs of the PCA form in which values of eigenvalues and explained variance of PCs and the scatter plot for loadings and scores of the first two components are summarized. From this analysis, it is seen that most of the information are retained in the first component (explained variance equal to 97%). Looking at the loading plot, the first component seems to take into account all the variables almost in the same measure, meaning that they are strongly correlated and bring the same kind of information. Thus, in the score plot, we directly have a good picture of the relationship between chemicals: in fact, the maximal and minimal elements highlighted by the Hasse diagram are already separated from the other objects along the first PC.

On the basis of this analysis, it would be possible to use only the first component as a new dataset, obtaining a radical reduction of variables. In this case, the Hasse diagram would result in a linear rank, as it is not possible to have incomparable objects when working with only one variable.





**Figure 8** Eigenvalues tab in the principal component (PC) results form.



**Figure 9** Loading/score plots tab in the principal components (PC) results form.

## 4. CONCLUSIONS

The DART software is able to perform a ranking analysis by means of several total ranking methods and a partial ranking method, which leads to a chart known as Hasse diagram. It also allows the user to improve the quality of the analysis by performing pre-processing methods, such as PCA and cluster analysis. All the results can be easily understood, thanks to the presence of several charts and a user-friendly graphic interface. The DART software can be downloaded for free at the European Chemicals Bureau website <http://ecb.jrc.it/>.

## REFERENCES

- European Commission (2003). Proposal for a regulation of the European parliament and of the council concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European chemicals agency and amending directive 1999/45/EC and regulation (EC) on persistent organic pollutants. COM (2003) 644 final, Brussels.
- European Communities (2000). *IUCLID CD-ROM Year 2000 Edition, Public Data on High Volume Chemicals*, EUR 19559EN, European Communities, Luxembourg.
- Pavan, M. (2003). Total and Partial Ranking Methods in Chemical Sciences. Ph.D. Thesis. University of Milano-Bicocca (Milano).
- Worth, A.P., Van Leeuwen, C.J., Hartung, T. (2004). The prospects for using (Q)SARs in a changing political environment—high expectations and a key role for the European Commission's joint research centre, *SAR and QSAR Environ. Res.* 15, 331–343.

# INDEX

- Accumulation tendency, 80, 82
- Acute oral toxicity, 150–1, 152
- Aggregation model, 53
- Alternatives, 20, 23, 24, 52–3, 54, 55, 57, 58, 59, 61, 63, 66, 67, 68, 69, 70, 140, 144, 149, 151
- Analysis of variance (ANOVA), 25, 26, 182  
*see also* Kruskal-Wallis test; Friedman test, multiple comparisons
- Antagonism, 83
- Antichain, 80
- Antisymmetry, 54, 75
- Aquatic persistence, 150
- Aquatic and terrestrial compartments, 100
- Asymptotic Relative efficiency, 14
- Atmospheric concentrations, 100
- Average rank, 18, 23, 31, 98, 99–100, 101, 104, 105, 106, 108, 140, 142, 147, 148, 150, 152, 197
- BDP2, 144–5, 146, 156, 157
- Bioaccumulation, 99, 141, 149
- Biodegradability, 144, 146
- Biodegradation, 78, 82, 144, 145, 150, 152
- Biodegradation potential, 150, 152
- Birkhoff, G., 74
- Bubley-Dyers algorithm, 83
- Chain, 79
- Cluster analysis, 76
- Complete ranking, 54
- Component wise order, 76
- Confidence interval, 9, 10, 14, 20, 21, 24, 42  
 based in sign test, 14, 21  
 for the slope in regression, 41  
 Wilcoxon-Mann-Whitney, 24  
 Wilcoxon signed ranks, 24
- Confinement, 97, 98, 101, 102, 106
- Connections, 42–3, 144
- Conover test, 33
- Consecutive POR, 100
- Contingency table, 15, 22
- Correlation coefficient, 9, 36, 37  
 Kendall's  $\tau$ , 38–9, 40, 41, 42  
 Pearson, 36, 37  
 Spearman rank correlation, 36, 63, 65, 114
- Covariance, 7, 36, 196
- Cover relation, 75, 161
- Cox and Stuart test for trend, 16
- Criteria, 10, 44, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 68, 69, 84, 86, 149, 163, 165, 169–86
- Critical region, 12, 13, 16, 17, 18, 19, 20, 23, 24, 25, 26, 28, 29, 32, 34, 36, 38
- Crucial weights, 85, 86, 87, 88, 89
- Cumulative distribution function (cdf), 4
- DART, 77, 175, 193–207
- DART software, 193–207
- Data  
 Concordant, 38, 42  
 Discordant, 38, 42
- Data representation, 76
- Decision making techniques, 52
- Decision model, 53–4, 58
- Decision variable space, 44, 45–6
- Descriptors, 84, 97–108, 124–7, 141, 142, 143, 146, 149, 150, 151, 152, 153
- Desirability function, 44, 66, 67, 182, 183
- Desirability *see* Total Ranking techniques
- Dilworth theorem, 80
- Dimension theory, 82
- Directed graph (or digraph), 43, 74, 75, 76, 78, 80, 82
- Direct rating, 59, 61
- Discrete Mathematics, 74, 194
- Distribution:  
 binomial, 9, 12, 13  
 cumulative, 4  
 discrete, 7, 9, 12, 18  
 F-distribution, 28, 31  
 exponential, 14, 21  
 Joint, 6, 8, 36  
 normal, 5, 6, 7, 8, 9, 10, 14, 17, 18, 19, 23, 24, 26, 34, 151, 152  
 t-distribution, 9  
 uniform, 8, 14, 21
- Distribution-free, 9  
*see also* Nonparametric methods
- Distribution function, 4, 6, 8, 36, 124, 125
- Dominance, 68–9
- Dominance function, 68, 69, 175, 180, 182, 184

- Dominance relation, 46
- Dominance *see* Total Ranking techniques
- Dominant, 8, 178, 180
- Duality principle, 45
- Ecosystem attribute ranking, 183
- Ecotoxicological data, 152, 153
- Effect, 9, 21, 25, 26, 27, 41, 51, 61, 62, 117, 130, 131, 177
- Environmental impact, 146, 147, 149
- EPI Suite, 143, 144, 145
- Equivalence relation, 76, 79, 83
- European Union System for the Evaluation of Substance (EUSES), 144
- Evaluation, 54, 63, 68, 74, 75, 76, 77, 84, 85, 86, 91, 98, 112, 130, 144, 149, 170, 180, 182, 183, 195
- Freedom, 22, 25, 28, 29, 31, 34, 87
- Friedman test, 27, 28, 29, 31, 32
- Functional biodiversity analysis, 160
- Global ranking index, 58
- Global scale, 58, 59, 61, 69
- Graphs, 1
  - arc, 43
  - connections, 42–3
  - edge, 43
  - node, 42–3
- Ground set, 74, 75, 76, 80
- G-space, 87, 91
- Halfon, E., 74, 77, 79, 99, 112, 141
- Hasse diagram, 43, 73–92, 99, 100, 102, 103, 105, 106, 107, 118, 119, 121, 141, 142, 146, 149, 150, 151, 153, 160, 161, 194, 195–6, 197, 201, 202, 205
- Hasse distance
  - application on DNA sequences, 118–20
  - application on electronic nose, 127–30
  - application on mass spectra, 120–4
  - application on molecular descriptors, 124–7
  - application on NMR spectra, 120–4
  - application on proteomic maps, 130–2
  - example of, 115
  - between matrices of different size, 114–5
  - weighted standardised, 114, 130
- Hasse, H., 99, 141
- Hasse matrix
  - augmented, 113, 115, 116, 127, 128, 129, 135
  - diagonal terms, 113, 115, 132
  - off-diagonal terms, 113, 115, 132
  - ranking relationships, 113
- Henry's Law Constants, 150, 152
- Hierarchical partial-order ranking (HPOR), 100–1
- High production volume chemical, 78, 85
- Hodges–Lehmann estimator, 21, 25, 41
- HPVC *see* High production volume chemical
- Hypothesis tests, 10, 14
  - actual significance level, 12
  - alternative hypothesis, 10–11, 13, 34
  - critical region, 13, 34
  - nominal significance level, 12, 16
  - null hypothesis, 10, 11, 12, 23
  - power of test, 12
  - p-value, 17, 29, 32, 36, 37, 38
  - type I error, 11, 12
  - type II error, 11, 12, 13
- Incomparable elements, 99, 119, 142
- Incomplete ranking, 54
- Independent random variables, 15
- Inferiority, 161, 162, 165
- Information base, 77, 78
- Inter-criteria information, 61–6
- Internet database, 85, 86, 89
- Interval scale, 3, 4, 17, 18, 58, 68
- Intra-criteria information, 58–61
- Inverse QSARs, 144
- Isolated elements, 79, 82, 197
- Joint distribution function, 6, 36
- Joint probability functions, 6
- Jonckheere-Terpstra test, 39
- Kendall's  $\tau$ , 38–9, 40, 41, 42
- K-means clustering, 196, 201, 202, 203
- Kruskal-Wallis test, 39
- Level, 3, 9, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 32, 34, 36, 37, 39, 52, 53, 54, 59, 60, 61, 62, 63, 70, 80, 82, 100, 119, 125, 142, 146, 149, 150, 161, 162, 170, 173, 174, 175, 180, 182, 183, 197
- Linear extension (LE), 98, 99, 140
- Linear ordering *see* Total ordering
- Linear rank, 100, 104, 105, 106, 107, 142, 144, 147, 148, 150, 205
- Local scale, 58, 59, 61, 67, 69
- Location test
  - one sample case, 18–22
  - several independent samples, 25–6
  - several related samples, 26–32
  - two independent samples, 22–5
- Log  $K_{ow}$ , 78, 79, 80, 143, 144, 145, 146, 149, 156, 157
- Log VP, 143, 144, 145, 146, 149, 156, 157

- MacNemar test for changes, 17
- Macroinvertebrate functional groups, 170
- Matrix
  - of adjacency, 43, 130, 135, 136
  - of incidence, 43
- Maximal elements, 79, 80, 82
- MCDM Methods *see* Multi Criteria Decision-Making methods
- Mean, 4, 5, 6, 7, 9, 18, 20, 21, 22, 25, 28, 31, 34, 40, 56, 67, 74, 100, 102, 172, 173, 175, 177, 178, 181, 182, 184, 200
- Median, 7, 8, 10, 11, 12, 13, 14, 18, 19, 20, 21, 22, 23, 25, 41, 42, 163, 164, 165, 178, 179, 180, 182
- Median estimator, 21, 24–5, 41
- Meta-descriptors, 99–100, 104
- Meta dimension, 98, 101, 105, 107
- METEOR, 77, 84, 85, 86, 87
- Migration potential, 102
- Minimal elements, 45, 79, 82, 205
- Multiattribute value theory, 58
- Multi-Criteria Decision-Making methods
  - applied to river ecosystem health study, 169–85
- Multicriteria decision making methods (MCDM), 53, 170
- Multicriteria evaluation, 98
- Multi-objective optimization, 44, 45
- Multiple comparisons, 29, 32
- Mutual order, 99, 141
  
- Noise deficient QSARs, 141, 143, 145
- Nominal scale, 3, 15, 17
- Nonparametric methods, 3, 4, 10, 99, 141
- Null hypothesis, 10, 11, 12, 13, 16, 19, 20, 22, 23, 24, 26, 28, 29, 32, 35, 37, 39, 40
  
- Objective space, 44–5
- O-METEOR *see* Orthogonal-METEOR
- OML-Multi, 102
- One-factor ANOVA, 25
- Optimal, 2, 44, 45, 46, 52, 53, 54, 55, 56, 64, 130, 175
  - optimal solution, 44, 45, 46, 64, 130
  - Pareto optimal front, or Pareto front, 46, 55
  - Pareto optimal set, 46
  - Pareto optimal solution, 46
- Order, 1
  - Inductive, 2
  - partial, 1, 43, 45, 46, 74–7, 80, 83, 98, 100, 104, 139–54, 140–1, 142, 161, 194
  - greatest, 2
  - infimum, 2
  - least, 2
  - lower bound, 2
  - maximal, 2
  - maximum, 2
  - minimal, 2
  - minimum, 2
  - optimal, 2
  - supremum, 2
  - upper bound, 2
  - relation, 1, 2
  - statistics, 3–9
  - total (linear or simple), 1
  - well ordering, 2
- Ordinal scale, 3, 4, 10, 15, 17, 18, 22, 31, 36
- Orientation, 54, 75, 76, 78, 79, 82, 89
- Orthogonal-METEOR, 87
- Outlier data, 24
- Overall ranking, 104, 106
  
- Paired samples, 20, 26
- Pareto optimal front, 46, 55
- Pareto optimality, 55, 56
- Pareto optimal point, 55, 56
- Pareto optimal set, 46
- Pareto optimal solutions, 46
- Pareto order, 45
- Partial ordering, 1, 73–91, 97–108, 112, 141, 142, 159, 160, 161
- Partial-order ranking (POR), 98–9
- Partial-order theory, 74–7, 98
- Partial preference function, 54–5, 61
- Pearson correlation, 36, 37, 63, 65, 66
- Persistence, 149, 150, 151, 152
- Pesticides, 150
- Pharmaceuticals, 84, 85, 86, 87, 88, 89–90
- Pitman efficiency, 13–14
- PNEC, 144, 146, 149, 156, 157
- Point estimates, 9–10
- Pollutant sites, 97
- POR-QSAR approach, 151, 153
- POSAC, 82, 84, 85
- Predictor, 153
- Preference function, 44, 54–5, 61
- Primary dimensions, 104
- Principal Components Analysis, 56, 194, 196
- Prioritization, 78, 159–67
- Probability density function (pdf), 4, 5, 128
- Probability function, 5, 6, 12
- Probable linear absolute rank, 99, 142
- Probable linear ranking, 147
- Probable rank, 100, 142
- Production volume, 78, 82, 85
- Product Order, 76, 77

- ProRank, 77
- PyHasse, 89
- PYTHON, 80, 89
  
- Quade test, 27, 31, 32
- Qualitative value scale, 59, 61
- Quantil, 7, 11, 13, 25, 28, 29, 32, 33, 34, 36
- Quantitative structure-activity relationships (QSARs), 73, 139–54
- Quantil, 7, 179
  
- RANA, 77
- Randomly generated linear extensions, 100, 104, 105, 106, 107, 108
- Random variable
  - continuous, 6, 15
  - discrete, 15
- Ranking, 8
- Ranking index, 53–4, 58, 74
- Ranking methods, 1–46, 51–70, 185, 194, 195, 197
- Ranking probabilities, 108
- Ranking of river sites, 183
- Rank-range, 163, 164, 167
- Ranks
  - rank order, 8, 54–5, 163, 170
  - rank test, 10, 18, 19, 20, 21, 22, 24, 27, 33, 34
  - signed rank test or Wilcoxon
  - signed test, 18, 19, 20, 21, 24, 27
- RAPID, 78
- Ratio scale, 3–4, 9, 18
- REACH, 140, 193
- Redheffer matrix, 80
- Reflexivity, 54, 75
- Regression
  - least squares, 10, 40
  - Theil's method, 41
- Risk assessment, 74, 144
- Risk assessment scheme, 140
- River bioassessment techniques, 169
- River restoration, 171, 184
- $Rk_{av}$ , 100, 106, 107, 142, 147, 148, 150, 152
- Robust, 10, 24, 40, 57, 62, 98, 100, 114
- Robustness, 21, 62, 133
- Rule-based approach, 144
  
- Salience, 159, 160, 162, 163, 165, 166, 167
- Salient, 159
- Sample range, 8, 31
- Scales, 3, 4, 10, 54, 59, 61
- Scoring, 58, 63, 132
- Semi-subordinate, 163
- Sensitivity, 46, 62, 63, 112, 119, 121, 132, 137, 173, 183
- Siblings, 162
- Sibsequence, 165, 166, 167
- Sibspread, 165, 166
- Siegel–Tuckey procedure, 32
- Significance level, 12, 13, 14, 16, 17, 18, 19, 22, 23, 24, 25, 26, 29, 32, 34, 37
- Simple additive ranking, 56, 58, 64, 197
- Simple ordering *see* Ordering
- Socio-economic factors, 97, 100, 102, 104, 105
- Solubility, 102, 143, 150, 152
- Spearman correlation, 83
- Spearman rank correlation, 36, 63, 65, 114
- Stability fields, 85, 86, 87, 88, 89, 90, 91
- Statistic *see* order statistic, confidence interval and hypothesis test
- Subordinate, 161, 163, 164, 165
- Superattribute, 86, 87
- Superiority, 161, 172, 165
- Supremum (or least upper bound), 2 *see also* Maximum
  
- Technical Guidance Document (TGD), 144
- Test
  - Conover, 33
  - Cox and Stuart test for trend, 16
  - for equality of variances, 23, 24, 33, 35
  - more than two samples, 34
  - for significance of changes and for trends, 15
  - two samples, 32
  - Friedman, 27, 28, 29, 31, 32
  - Jonckheere–Terpstra test, 39
  - Kruskal–Wallis test, 39
  - MacNemar, 17
  - parametric test, 26
  - Quade, 27, 31, 32
  - rank test, 10, 18, 19, 20, 21, 22, 24, 27, 33, 34
  - sign test, 10
  - squared rank test, 33, 34
  - t-test, 9, 10, 14, 21, 24, 25
  - Wilcoxon–Mann–Whitney, 22, 23, 24, 25, 32
  - Wilcoxon signed rank, 18, 19, 20, 21, 24, 27, 41
- Ties, 3, 16, 20, 24, 26, 27, 31, 34, 36, 165
- Tolerance interval, 14–15
- Topological index, 73
- Total-order ranking, 53–69, 175
- Total ranking techniques, 194, 197
- Tournament theory, 75
- Toxicity, 82, 99, 140, 141, 143, 144, 145, 146, 149, 150, 151, 152, 153, 197, 198, 201
- Transitivity, 75

Utility function, 66, 67, 68, 182

Vapour pressure, 143, 146, 149

Variance, 4, 5, 9, 33, 40, 46, 134, 196, 205

Vectorial optimization

*see also* Multi-objective optimization

Vertices (or nodes), 43

Water-octanol partitioning, 146

Weights, 61, 86

Well-ordering, 2

WHASSE, 77, 79, 99, 141

WHasse software, 99, 141

Wilcoxon signed rank test, 18, 19,  
20, 21, 24, 27