Rainer Brüggemann · Lars Carlsen
Jochen Wittmann *Editors*

Multi-indicator Systems and Modelling in Partial Order

Springer

Multi-indicator Systems and Modelling in Partial Order

Rainer Brüggemann • Lars Carlsen
Jochen Wittmann
Editors

Multi-indicator Systems and Modelling in Partial Order

Springer

Editors
Rainer Brüggemann
Leibniz-Institut für Gewässerökologie und
Binnenfischerei
Berlin, Germany

Lars Carlsen
Awareness-Center
Roskilde-Veddelev
Denmark

Jochen Wittmann
HTW Berlin,
University of Applied Sciences
Berlin, Germany

ISBN 978-1-4614-8222-2 ISBN 978-1-4614-8223-9 (eBook)
DOI 10.1007/978-1-4614-8223-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013950377

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Overview

The theory of partially ordered set is dynamically evolving, as demonstrated by the number of publications in the mathematical journals. Also the rather special binary relation used for evaluating data matrices with respect to possible rankings is rapidly developing. The special order relation used is the reason why often Hasse Diagram Technique (HDT) is referred to instead of partial order technique or partial order ranking. This restriction on the possible variety of binary relations fulfilling the axioms of partial order is immediately related to its use in the evaluation of data matrices, namely the component-wise order or, as it is also often called, product order. The restriction to the component-wise order has a severe consequence: The application of partial order becomes now a discipline in multivariate statistics too.

Because of the pretty quick development of HDT, the international workshops about HDT become an important platform to exchange ideas not only in theoretical questions but also within the field of applications.

The book "Multi-indicator Systems and Modelling in Partial Order" contains the newest theoretical concepts as well as new applications or even applications, where standard multivariate statistics fail. Some of the presentations have their counterpart in the book; however, there are many contributions, which are completely new in the field of applied partial order.

Why do we use the term "modelling" in the title? It turns out that complex processes or complex properties may simply be understood, if the ranking that is induced can be analyzed in terms of multi-indicator systems. In the field of chemistry this kind of analysis has a long-lasting tradition; however, the techniques can be applied by far more generally. Even the outcome of ranking indices, as calculated from traditional decision support systems (such as PROMETHEE or ELECTRE), may be of interest in terms of modelling.

Partial order applied to multi-indicator systems gets an additional quality because the relations, which may be drawn as arrows or edges in a Hasse diagram, do not only tell us the order between any two objects, but also that a data pattern is behind this finding. Hence, Chap. 3 makes this statement clear. Incomparability in a general Hasse diagram is an inherent feature. Thus, where a data vector is in the background, the incomparability as well as the comparability can get different degrees. This aspect becomes also clear in Chap. 18 of this book, where chain and antichain analysis are specifically based on the multivariate character, which is behind the posetic analysis of matrices suitable for evaluation. Another statement was already several times mentioned: partial order theory provides analytical tools to understand results arising from multicriteria techniques. The composite indicator as the most simple and most transparent variant is studied in several chapters of this book.

In sum the book informs about recent developments in theory and in applications of partial order under the special relation as used in HDT.

The Book Chapters

It is clear that with so many different topics, theoretical and applied, one cannot easily give a logical sequence. Thus, a not necessarily innovative concept is followed to present first the chapters, which are theoretical, then chapters with applicational character and finally chapters where software aspects are considered.

The book takes care for the different aspects attributed to partial order theory. It is organized in the following sections:

(a) Theory
(b) Partial Order as Analysis Tool of Composite Indicators
(c) New Trends in Partial Order
(d) Applications
(e) Software Aspects

Hence, the first section provides an introduction to partial order theory in a general sense, however aiming at evaluation. It starts with a chapter about basics of evaluation where not only the inherent character of evaluation, namely order theory is regarded, but also a unifying concept is represented taking into account that evaluations, even based on a single dimension, can be unsharp. Whereas the first chapter clarifies the interaction between order, fuzziness and evaluation, the second chapter asks whether a given data matrix is suitable for evaluation or not. Obviously, in full generality this is not the case. Only after a series of additional inputs, for example giving the columns of a data matrix (quantifying the attributes) an orientation, a data matrix can be considered as evaluation matrix. Even if partial order is an analytical tool by which results of conventional multicriteria decision systems can be discussed, as shown in the last part of the second chapter, the most typical striking fact is the existence of incomparability. The third chapter shows how the dual character "incomparable"—yes or no—can give a quantification. How incomparable is one object with another, which is a nice example as to how far modelling within partial order can be perfomed. A multi-indicator system demands for such a modelling step.

By incomparability the graph of order relations gets a structure: Whereas a linear order would lead to a Hasse diagram consisting of a single chain, the appearance of incomparability leads to branching points, isolated points, in brief, to all the diversity directed acyclic graphs can have. Then it is a natural question, as to how far different partially ordered sets (posets) and their graphs can be compared. In the fourth chapter, the dissimilarity of posets is modelled by an embedding into a lattice and from that new measures are derived. By this contribution the methods of quantifying distances among posets get an additional sharp instrument. When we discuss similarity or dissimilarity among posets, then it is pretty natural to characterize posets by new measures, which themselves could be a basis for quantification of dissimiarity. With the concept of complexity as a poset-characterizing quantity, the first step is done. Similar to the lattice concept of Chap. 4, another lattice concept, that of Young diagrams, is the vehicle to derive a measure of complexity.

The second section is more specifically contributing to the application of partial order as an analytical tool in the decision process. A chapter about comparative knowledge discovery shows how far stakeholders and decision makers may get support by the instruments partial order is offering. When an evaluation matrix is at hand, then often weights are considered as additional and subjective information. One may take the other way round and see data-based weights, i.e. weights derived from the evaluation matrix as inherent posetic information about the importance of the indicators, which are the columns of the evaluation matrix. In the same direction aims the next chapter. Here, weights as needed to construct composite indicators, widely used in the multicriteria-decision scene, are seen as objects of a modelling process: Which weight system allows the closest coincidence between linear (weak) orders as a result of partial order theory and of composite indicators?

The next section has the title “New Trends in Partial Order”.

> One trend is directed toward partial order itself and its development beyond the graphical representation by Hasse diagrams, another is complementary: What does all the theoretical innovation help, when the applicational sciences do not use the new theoretical findings.

In that sense this section contributes to new developments within partial order theory; especially it shows how the concept of Hasse diagrams, which is limited to only a few objects, can be replaced by more general—data-mining-suitable—concepts. Although the concept is not verbally mentioned in that chapter, but practically it shows how one can use posetic coordinates instead of the Hasse diagram. The other chapter discusses the potential use of partial order concepts especially in socioeconomics. Poverty is a topic of general relevance. Although it has so many facets, the one-dimensional scale is still most often intrinsically assumed. Instead, poverty as many other concepts should be seen as multi-indicator system and hence a suitable object of partial order studies.

The following two sections are devoted to applications and sotware aspects. Both application and software are mutually stimulating. Hence, the separation into more applicational and more software-oriented sections is somewhat arbitrary. Many of the applicational chapters suggest or describe new theoretical developments besides the specific applicational field. So the first chapter in this section (Chap. 10) offers an heuristic solution, for what often can be seen as a conundrum in Hasse diagram technique, namely the isolated vertices. Although the fact of isolatedness indicates a data pattern of specific interest, they are not facilitating a practical evaluation. In this chapter an idea is offered to remedy this problem. The heuristic idea is now—with some modifications—realized by a module of PyHasse. The application field of this chapter is taken from technical chemistry and the risk of accidents attributed to single chemicals. The following chapter combines concepts of Geographical Information Systems and Geostatistics with partial order theory and shows how chains in poset can be helpful in the interpretation of monitoring results. Here the data matrix is almost rectangular, i.e. there are many indicators and relatively few objects. This is a situation which the authors of the next chapter face too. Their question is, how to prioritize waste disposal sites for a remediation? There are many criteria that are taken into account. The authors show how a concept, already published, namely the Hierarchical Partial Order Ranking, can be applied to systematically reduce the criteria in order to arrive at a handsome priority list. In another chapter (Chap. 13) it is shown how indicators in project management can be evaluated by tools offered by partial order theory: As to how far the single indicators are responsible for the structure of the Hasse diagram and hence for the position of the objects. So, skilled processes to study sensitivity of indicators are applied (global sensitivity analysis). Furthermore, the role of the "Local Partial Order Model" was illustrated with exciting results.

When a trend should be extracted from the more applicational section, then it seems as if more and more partial order is applied together with other typical multivariate concepts, so for example partial order and geostatistics. Hence, the next chapter combines partial order with neural networks, where the neural networds are applied as a preprocessing tool to condense the data matrix into a form which is managable within conventional partial order graphical displays, i.e. with Hasse diagrams. The applicational field is taken from monitoring of sediments.

The section, "software aspects", could equally well be a part of the former one. However, in the chapters in this section software aspects play a slightly more central role.

Section "software aspects" starts with the usage of the software **R**. Here it is shown how macros can be written, which facilitate the programming work. In the same direction aims Chap. 16. The applicational field is the evaluation of geographic units with respect to the landscape inventory and applies cluster analysis besides the already detailed described posetic coordinates. The next two chapters apply the software package PyHasse. In Chap. 17, a typical modelling aspect is on the focus, namely as to how far the proximity of one poset can be used to explain causally the results of the other poset: The next chapter deals with the topic of Failed Nations and shows in a pretty systematic way, how different tools of PyHasse can be applied. The main focus is on the role the different indicators have. As the two chapters (17 and 18) widely apply modules of the software PyHasse, it may be a good idea to conclude the book with a short description of PyHasse itself (Chap. 19).

With this chapter the book finds its end. It spanned a spectrum from theoretical concepts to applicational ones and finally to software aspects. What is the summarizing statement we can draw and what are the trends aiming into the future?

What are the trends in the future?

Starting with environmental chemistry, the domain of applicability was steadily increasing, so it is hoped that new fields of applications are opened; perhaps the chapter "Partial orders in socio-economics: an applicative challenge for poset theorists or a cultural challenge for social scientists" is initiating new interests.

Almost all theoretical investigations, explained here, are the germ for further developments, for example the question as to how far partial order can be helpful for stakeholders seems to direct towards algebraic topology, the concepts of dissimilarity and complexity may get a unifying theoretical framework and the modelling by proximity of posets may get a safe background by suitable methods of statistical test theories. We could elongate this list easily; here only some few examples are mentioned. However there is still a main deficit in partial order theory: Incomparabilities are used as in explorative statistics: They inform us about pretty specific data configurations. This trend was inforced by publications about separability and dominance; nevertheless, the practioner, i.e. the decision maker, is not happy with that. What still and most urgently is needed is a transparent procedure to perform the trade-off among contradicting indicator values. It is to be hoped that in the future we can add corresponding concepts into our tool box.

Finally what about software? Theoretical development without its practical applicability by means of software is in the long run not helpful. Indeed software is developing and the interested reader may find either R-software practicable for her/his needs or other packages such as DART, WHASSE, PRORANK or PyHasse. The development of software is (or should be) teamwork. Nevertheless, most developments of PyHasse are on the shoulders of one single scientist. A necessary condition to get a broader personal basis is the visibility of the software. Some first steps in this direction, i.e. representation of the posetic software in the Internet, are already started. For the future, it is hoped that these web-based steps can be intensified.

Berlin, Germany	Rainer Brüggemann
Roskilde, Denmark	Lars Carlsen
Berlin, Germany	Jochen Wittmann

Acknowledgement

This book could not have come to reality without the enthusiasm of all our contributing authors. We are truly grateful and thank each of them cordially. We thank also all our reviewers ensuring the scientific quality of the contributions. Last but not least, we thank Melinda Paul and Clinton Meredith for all their assistance and especially for their patience.

A cross-review was performed among the (co-)authors of this book; however, the expertise of other scientists was extremely helpful and we thank

Dr. P. Annoni, JRC, Italy
Prof. J. Newlin, Brunswick University, USA
P. Koppatz, SUDILE Potsdam
Prof. J. Owsinsky, IBSPAN, Warsaw, Poland
Prof. B. Page, University Hamburg, Germany
Dr. S. Pudenz, Basel, Switzerland
Dr. P.B. Sørensen, Department of Bioscience, Århus University, Denmark

Contents

Part I Theory

Part II Partial Order as Tool to Analyse Composite Indicators

Contributors

G. Achari Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

G. Al-Sharrah Department of Chemical Engineering, Kuwait University, Safat, Kuwait

A. Arcagni Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

M. Bariotakis Department of Biology, University of Crete, Heraklion, Greece

H.-G Bartel Department of Chemistry, Humboldt University Berlin, Berlin, Germany

R. Brüggemann Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

L. Carlsen Awareness Center, Roskilde, Denmark

I. Cok Department of Toxicology, Faculty of Pharmacy, Gazi University, Ankara, Turkey

M. Fattore Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

R. Grassi Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

S.W. Joshi Department of Computer Science, Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA

V. Kalogrias Department of Biology, University of Crete, Heraklion, Greece

S. Katsogianni Department of Biology, University of Crete, Heraklion, Greece

A. Kerber Department of Mathematics, University of Bayreuth, Bayreuth, Germany

F. Maggino Department of Statistics, Computer science, Applications "G. Parenti", University of Florence, Florence, Italy

B. Mazmanci Department of Biology, Faculty of Sciences and Letters, University of Mersin, Mersin, Turkey

M.A. Mazmanci Department of Environmental Engineering, Faculty of Engineering, University of Mersin, Mersin, Turkey

H.-J. Mucha Weierstrass Institute of Applied Analysis and Stochastics, Berlin, Germany

W. L. Myers Penn State Institutes of Energy and Environment, The Pennsylvania State University, University Park, PA, USA

G. P. Patil Center for Statistical Ecology and Environmental Statistics, Department of Statistics, The Pennsylvania State University, University Park, PA, USA

S. Pirintsos Department of Biology, University of Crete, Heraklion, Greece

G. Restrepo Interdisciplinary Research Institute, Universidad de Pamplona, Bogotá, Columbia

Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

C.M. Rocco Facultad de Ingeniería, Universidad Central de Venezuela, Caracas, Venezuela

H. Scherb Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

K.-W Schramm Helmholtz Zentrum München, German Research Center for Environmental Health, Molecular EXposomics (MEX), Neuherberg, Germany

Department fuer Biowissenschaften, Wissenschaftszentrum Weihenstephan fuer Ernaehrung und Landnutzung, TUM, Freising, Germany

V. Simeonov Group of Chemometrics and Environmentrics, Faculty of Chemistry and Pharmacy, University of Sofia, Sofia, Bulgaria

S. Tarantola Institute of the Protection and Security of the Citizen, JRC, European Commission, Ispra, Italy

R.J. Thiessen Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

S. Tsakovski Group of Chemometrics and Environmentrics, Faculty of Chemistry and Pharmacy, University of Sofia, Sofia, Bulgaria

C. Turgut Faculty of Agriculture, Adnan Menderes University, Aydin, Turkey

K. Voigt Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

R. Wieland Leibniz Centre for Agricultural Landscape Research (ZALF), Institute of Landscape Systems Analysis, Muencheberg, Germany

J. Wittmann HTW Berlin, University of Applied Sciences, Environmental Informatics, Berlin, Germany

Part I
Theory

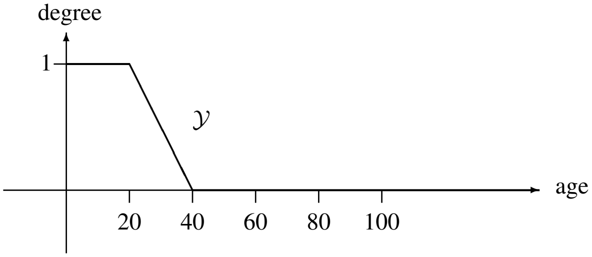
Chapter 1
Evaluation as a General Approach to Problem Driven Mathematical Modeling

Adalbert Kerber

Abstract Many modern applications need *evaluation* of statements, the truth values "true" and "false" alone may not suffice, a statement can be neither true nor false, it may be true (or false) "in a certain sense." They also need modeling of *linguistic expressions* and of *fuzzy situations*. "Binary thinking" does not suffice in many cases. Moreover, the choice of methods might better be *problem driven*, depending, for example, if we better use a pessimistic or an optimistic reasoning. Here is a brief introduction of how we can choose tools that are appropriate for mathematically modeling this kind of problems.

1.1 Easy Examples

First we consider an example of fuzzy modeling the linguistic expression "young." We may use the following mapping $\mathcal{Y}:\mathbb{R}_{\geq 0} \rightarrow [0,1]$ that allows to evaluate the degree of somebody being young,



obtaining the following "evaluation" of statements of the form "x is young":

A. Kerber (✉)
Department of Mathematics, University of Bayreuth, Bayreuth, Germany
e-mail: kerber@uni-bayreuth.de

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_1, © Springer Science+Business Media New York 2014

$$\begin{aligned}\mathcal{Y}(\text{15 years old}) &= 1\\ \mathcal{Y}(\text{30 years old}) &= 0.5\\ \mathcal{Y}(\text{40 years old}) &= 0\end{aligned}$$

The second example shows a fuzzy situation where a partial order and therefore incomparableness occurs: reading and writing devices for CDs and DVDs run under two different technical norms ±R so that we may obtain results like

$$\oplus\oplus\,/\,\ominus\ominus,$$

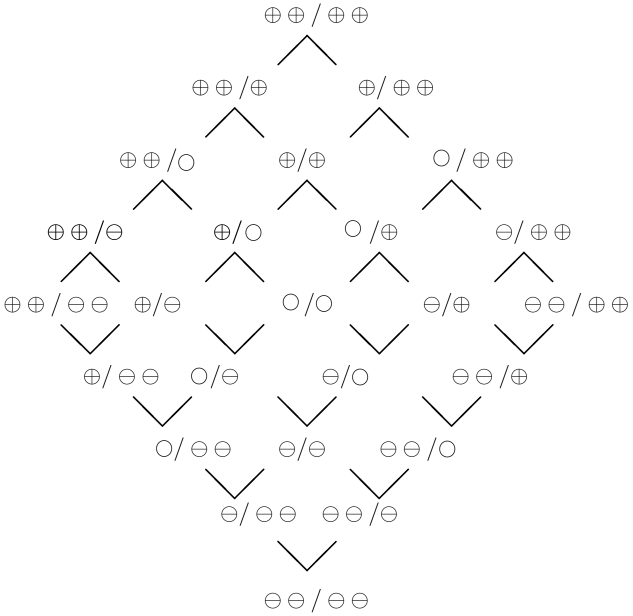
in the journal "c't," for example, i.e., pairs of values, where the first value is the evaluation of the device with respect to the norm $+R$ while the second one the value of the quality under $-R$, say. Such values can be comparable, e.g., $\oplus\oplus\,/\,\ominus\ominus \le \oplus\oplus\,/\,\ominus$, but we may also meet incomparable evaluations like

$$\oplus\oplus\,/\,\ominus\ominus \not\le \ominus\,/\,\oplus\oplus \text{ and } \oplus\oplus\,/\,\ominus\ominus \not\ge \ominus\,/\,\oplus\oplus.$$

The set of values is taken from the partial order

$$L_R = \{\ominus\ominus, \ominus, \bigcirc, \oplus, \oplus\oplus\}^2 = \{\ominus\ominus\,/\,\ominus\ominus, \ominus\ominus\,/\ominus, \ldots, \oplus\oplus\,/\,\oplus\oplus\},$$

the cartesian square of the total order $\ominus\ominus < \ominus < \bigcirc < \oplus < \oplus\oplus$. This partial order L_R, or, more explicitly, the pair $(L_R, \le)$, looks as follows:



The third example is the evaluation of refrigerants. The paper (Brüggemann et al. 2011) was motivated by the thesis (Restrepo 2008), where 40 refrigerants are evaluated using triples

$$(K(\text{ODP}), K(\text{ALT}), K(\text{GWP})) \in [0,1]^3$$

of real numbers. The paper was written in order to demonstrate the embedding of the use of real-valued parameters into fuzzy mathematics.[1] The evaluation of 18 refrigerants described in Brüggemann et al. [2011] used the following table of normalized and oriented values (so that they are contained in the interval $[0, 1]$ and that the smaller the values the better it is) of the parameters ODP (*ozone depletion potential*), GWP (*global warming potential*), and ALT (*atmospheric lifetime*):

Refrigerant	ODP	GWP	ALT
1 (R11)	0.19608	0.31622	0.01406
2 (R12)	0.16078	0.72432	0.03125
6 (R141b)	0.02353	0.04818	0.00290
7 (R142b)	0.01275	0.15338	0.00559
8 (R23)	0.00008	0.96689	0.08437
16 (R290)	0	0.00135	0.00001
21 (R744)	0	0.00007	0.03750
22 (R1281)	1	0.08784	0.00343
23 (RC318)	0	0.67568	1
29 (HFE-125)	0	1	0.05156
32 (R40)	0.00392	0.00108	0.00040
33 (R113)	0.17647	0.40541	0.02656
35 (R114)	0.16667	0.66216	0.09375
36 (R13/1)	0	0.00007	0.00003
37 (–)	0	0.00007	0
38 (R717)	0	0	0.00007
39 (HFE-143)	0	0.04432	0.00178
40 (HFE-245)	0	0.04709	0.00125

The set of *all* triples of real numbers contained in the interval $[0, 1]$ is a partial order, a *lattice*, as is L_R that was used before. So let us introduce the notion of complete lattice next.

1.2 Complete Lattices

We shall assume that the values of the evaluations are taken from a set L that is a *complete lattice* in the following sense:

- *Lattices* are sets L such that for each $\lambda, \mu \in L$ there is an infimum $\lambda \wedge \mu$ and a supremum $\lambda \vee \mu$ contained in L.

[1] A terrible name since not the mathematics is fuzzy but the situations that are modeled.

- They are called *complete* if each subset of L has both an infimum and a supremum, in particular the subset L itself, they are denoted as follows:

$$\wedge L = 0, \vee L = 1.$$

- Hence $\{0,1\}$ as well as $[0,1]$, $[0,1]^3$ and

$$L_R = \{\ominus\ominus / \ominus\ominus, \ldots, \oplus\oplus / \oplus\oplus\}$$

  are complete lattices.
- A lattice L, or, more explicitly, the triple $(L, \wedge, \vee)$, yields a partial order $\leq$ by

$$\lambda \leq \mu \Leftrightarrow \lambda \wedge \mu = \lambda$$

  and so we can describe the lattice either as the triple $(L, \wedge, \vee)$ or as the pair $(L, \leq)$. But we will not hesitate to indicate it briefly as L if no confusion is to be expected.

Thus, the complete lattices, in particular $[0,1]$, $[0,1]^3$ and L_R, contain an element 0 and an element 1, and so they can be considered as generalizations of the complete lattice $\{0,1\}$. Therefore, *the use of a suitable complete lattice L is a first possibility to do a problem driven choice of methods in order to attack evaluation problems, i.e., to model an evaluation problem.* However, this is only the starting step, others are necessary. The second step is the choice of a suitable set theory.

1.3 Set Theory over *L*

Assume that L was chosen in a proper way and consider a crisp set X, i.e., a set in the classical sense where an x either belongs to X or not.

- An *L-subset* of X is an element of the set of mappings from X to L

$$L^X = \{\mathcal{M} | \mathcal{M} : X \to L\},$$

  for short: $\mathcal{M} \in L^X$.
- The value $\mathcal{M}(x) \in L$ of M at $x \in X$ is an element in L that evaluates the statement "x belongs to $\mathcal{M}$."
- For $\mathcal{M}, \mathcal{N} \in L^X$ we introduce the *L-inclusion* by:

$$\mathcal{M} \subseteq_L \mathcal{N} \Leftrightarrow \forall\ x \in X : \mathcal{M}(x) \leq \mathcal{N}(x).$$

For example, if $X = \{\text{ODP}, \text{GWP}, \text{ALT}\}$ and $L = [0,1]$ then we may consider the first and the third row of the table of parameter values of the refrigerants as L-subsets of X:

$$\mathcal{R}11 = (0.19608, 0.31622, 0.01406) \in [0,1]^{\{\text{ODP},\text{GWP},\text{ALT}\}},$$

and

$$\mathcal{R}141b = (0.02353, 0.04818, 0.00290) \in [0,1]^{\{\mathrm{ODP},\mathrm{GWP},\mathrm{ALT}\}}.$$

We obtain the inclusion

$$(0.02353, 0.04818, 0.00290) \subseteq_L (0.19608, 0.31622, 0.01406),$$

i.e.,

$$\mathcal{R}141b \subseteq_L \mathcal{R}11.$$

In words: refrigerant R141b is better than R11, with respect to ODP, GWP, and ALT.

This allows to embed the evaluation of refrigerants mentioned above into the theory of L-subsets. We deduce from the equivalence that refrigerant K_0 is better than refrigerant K_1 if and only if the corresponding evaluation $\mathcal{K}_0$ is a $[0,1]$-subset of the evaluation $\mathcal{K}_1$. For short: if $\mathcal{K}_0$ is contained in $\mathcal{K}_1$ (as an L-subset, of course). *Thus, L-subsets can serve as models for evaluation!*

Our next step is the definition of various intersections of L-subsets using the notion of *t-norm* (see, e.g., Klir and Yuan 1995) which means a mapping $\tau : L \times L \to L$ with the following properties:

- τ is *symmetric,* $\tau(\lambda,\mu)=\tau(\mu,\lambda)$,
- τ satisfies the *boundary conditions* $\tau(\lambda,1)=\lambda$,
- τ is *monotonous*, $\mu \le \nu$ implies $\tau(\lambda,\mu) \le \tau(\lambda,\nu)$,
- τ is *associative*, $\tau(\lambda,\tau(\mu,\nu))=\tau(\tau(\lambda,\mu),\nu)$.

We can now introduce the *τ-intersection* of $\mathcal{M},\mathcal{N} \in L^X$ as the L-subset $\mathcal{S} \in L^X$ with

$$\mathcal{S}(x) = (\mathcal{M} \cap_\tau \mathcal{N})(x) = \tau(\mathcal{M}(x), \mathcal{N}(x)).$$

The most important t-norms are the following ones:

- The *standard norm* is the minimum,

$$s(\lambda,\mu) = \lambda \wedge \mu, \text{ so that } (\mathcal{M} \cap_s \mathcal{N})(x) = \mathcal{M}(x) \wedge \mathcal{N}(x).$$

- The *drastic norm* is defined by

$$d(\lambda,\mu) = \begin{cases} \lambda, & \mu = 1_L, \\ \mu, & \lambda = 1_L, \\ 0_L, & \text{otherwise}, \end{cases} \qquad (\mathcal{M} \cap_d \mathcal{N})(x) = \begin{cases} \mathcal{M}(x), & \mathcal{M}(x) = 1_L, \\ \mathcal{N}(x), & \mathcal{N}(x) = 1_L, \\ 0_L, & \text{otherwise}. \end{cases}$$

- If $L=[0,1]$, then there is the *algebraic product*

$$a(\lambda,\mu) = \lambda\bullet\mu, \text{ in which case } (\mathcal{M}\cap_a \mathcal{N})(x) = \mathcal{M}(x)\bullet\mathcal{N}(x).$$

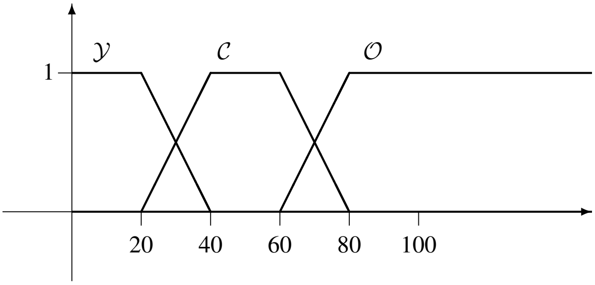
- If again $L=[0,1]$, we have the *bounded difference*

  $$b(\lambda,\mu) = \text{Max}\{0,\lambda+\mu-1\},$$

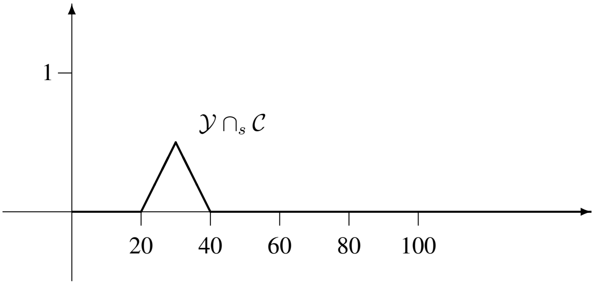
  which gives the following intersection:

  $$(\mathcal{M}\cap_b \mathcal{N})(x) = \text{Max}\{0,\mathcal{M}(x)+\mathcal{N}(x)-1\}.$$

Here is an easy example that shows the difference between the standard and the drastic norm and it gives an idea why it is useful to have a choice between these two. We extend the linguistic expression "young" mentioned above by also modeling the ages "of central age" and "old" as follows:



For example, we obtain as standard intersection of $\mathcal{Y}$ and $\mathcal{C}$ the L-subset



while the drastic intersection yields the zero function! Thus, if we use the standard norm, we accept that there are persons, for example, of age 35, that are *both* young and of central age, while the use of the drastic norm implies that we do not allow that any person may be called young as well as of central age. Although in both cases we evaluate the statement "a 30 years old person is young" and the statement "a 30 years old person is of central age" by 0.5.

Thus, a problem driven choice might be useful, in particular since always

$$d(\lambda,\mu) \le s(\lambda,\mu).$$

We may in fact say that the choice of the drastic intersection is advisable if a more pessimistic thinking might be better, while the choice of the standard intersection is the most optimistic one (the others are in between these two)!

So this is the second chance to do a problem driven choice. The first one was the choice of L, the second one is the choice of a suitable t-norm τ.

Unions of two L-subsets can be introduced similarly, using the notion of t-conorm, cf. Klir and Yuan [1995].

1.4 A Corresponding Logic

In addition to the choice of a set L of values and a set theory that allows to evaluate intersections, unions or other compositions of sets, we need to have a logic that allows to evaluate statements that are composed from statements that are evaluated already. For example, we need to evaluate an implication between two statements and of the negation of a given statement.

It is interesting to see that a logic is quite often provided or even determined by the chosen norm, i.e., by the chosen set theory, in the following way: assume a t-norm τ. A mapping $\tilde{\tau}: L \times L \rightarrow L$ is called a *residuum* corresponding to τ, if it satisfies the following condition: for all $\lambda, \mu, \nu \in L$,

$$\tau(\lambda,\mu) \le \nu \Leftrightarrow \lambda \le \tilde{\tau}(\mu,\nu).$$

There may be several residua, but in many cases, depending on the chosen L, the residuum is uniquely defined, in which case we call τ a *residual t-norm*. For example, the residua of the standard norm s, the drastic norm d, the algebraic product a, and the bounded difference b are unique and they have the following values:

$$\begin{aligned}
\tilde{s}(\lambda,\mu) &= \begin{cases} 1, & \textit{if } \lambda \le \mu, \\ \mu, & \text{otherwise}, \end{cases} \\
\tilde{d}(\lambda,\mu) &= \begin{cases} \mu, & \textit{if } \lambda = 1, \\ 1, & \text{otherwise}, \end{cases} \\
\tilde{a}(\lambda,\mu) &= \begin{cases} \mu / \lambda, & \textit{if } \lambda \ne 0, \\ 1, & \text{otherwise}, \end{cases} \\
\tilde{b}(\lambda,\mu) &= \operatorname{Min}\{1, 1-\lambda+\mu\}.
\end{aligned}$$

Let us now recall the classical binary logic, using the *tertium non datur*. It is obtained from $L = \{0, 1\}$. The table of truth values of statements that are composed from statement A (with given truth value α) and statement B (with its truth value β) is

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
α	β	$1-\alpha$	$\mathrm{Min}\{\alpha,\beta\}$	$\mathrm{Max}\{\alpha,\beta\}$	$\mathrm{Min}\{1,1-\alpha+\beta\}$

.

In more detail:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
1	1	0	1	1	1
1	0	0	0	1	0
0	1	1	0	1	1
0	0	1	0	0	1

Besides this binary logic, there is also a three-valued logic, where $L = \{0,1/2,1\}$ and the value 1/2 can be interpreted as "uncertain." Here is, for example, a weather forecast for a week, using this interpretation of 1/2:

	Warm	Cold	Dry	Windy
Mo	1/2	1/2	1	1
Tue	1	0	1	1
Wed	1/2	1/2	1	0
Thur	1/2	1/2	0	0
Fr	0	1	0	0
Sa	0	1	1/2	0
Su	0	1	1	1

The general situation is now as follows: we assume a complete lattice L, a residual t-norm τ, and its residuum $\tilde{\tau}$, obtaining the *evaluation algebra*

$$(L, \wedge, \vee, \tau, \tilde{\tau})$$

and the following table for the evaluation of compositions of statements:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
α	β	$\tilde{\tau}(\alpha,0)$	$\alpha \wedge \beta$	$\alpha \vee \beta$	$\tilde{\tau}(\alpha,\beta)$

.

1.5 Summary

In order to summarize briefly, we can say that evaluation generalizes binary thinking. In this way it opens an approach to further applications, for example, to modeling of linguistic expressions and their use. In particular we can do a problem driven (or at least problem oriented) mathematical modeling in many cases.

The method is to

- choose a suitable lattice L of values,
- take a suitable residual t-norm τ in order to introduce a set theory over L,
- and finally use the corresponding residuum $\tilde{\tau}$ in order to get a logic that can be applied for the evaluation of composite statements.

Acknowledgment Thanks are due to the referees for helpful hints.

References

Brüggemann R, Kerber A, Restrepo G (2011) Ranking objects using fuzzy orders with an application to refrigerants. MATCH Commun Math Comput Chem 66:581–603

Klir GJ, Yuan B (1995) Fuzzy sets and fuzzy logic (theory and applications). Prentice Hall, Englewood Cliffs

Restrepo G (2008) Assessment of the environmental acceptability of refrigerants by discrete mathematics: cluster analysis and Hasse diagram technique. Doctoral thesis, University of Bayreuth, Germany. Available via the address http://opus.ub.uni-bayreuth.de/frontdoor.php?source_opus=393\&la=de

Chapter 2
Multivariate Datasets for Inference of Order: Some Considerations and Explorations

Ganapati P. Patil, Wayne L. Myers, and Rainer Brüggemann

Abstract Ideal formulation of a multi-indicator system (MIS) would be to define, design, and acquire the entire construct with complete consensus among all concerned. However, such would be an extreme rarity in actuality. Experts have differing views. Factors may not express monotonically, as when either extreme is unfavorable. The entirety cannot be assessed and must be sampled. Empirical experience to validate expectations is inadequate. Consequently, exploratory examination of any available datasets collected for collateral purposes can augment insights relative to suitable surrogates for ideal indicators, with particular attention to ordering relations for subsets of quantifiers and ensembles of entities (objects, cases, instances, etc.).

Multivariate datasets are comprised of several quantifiers (variates or variables) as columns recorded for multiple entities as rows. The data matrix thus realized is not necessarily directly useful nor fully informative for analytically inferring order among entities. In this chapter, some approaches are discussed which may be helpful in extracting insights on ordering properties that are embodied in multivariate

Expanded from the keynote lecture by, Patil (2012), at the International Workshop on Partial Order Theory and Modeling held in Berlin, Germany during September 27–29, 2012. This research and outreach effort have been partially supported by US NSF Project 0307010 for Digital Governance and Hotspot Geo-Informatics of Detection and Prioritization.

G.P. Patil (✉)
Center for Statistical Ecology and Environmental Statistics, Department
of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: gpp@stat.psu.edu

W.L. Myers
Penn State Institutes of Energy and Environment, The Pennsylvania State University,
University Park, PA 16802, USA

R. Brüggemann
Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology
and Inland Fisheries, Müggelseedamm 310, Berlin, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_2, © Springer Science+Business Media New York 2014

datasets and applicable in configuring suites of indicators. These procedures may be particularly helpful in finding suitable surrogates and applying partial order theory when expediency is essential. We consider orientation, crispness of data, and culling of candidates according to importance in respect of some desirable criteria.

2.1 Introduction

The ideal is often not attainable in formulating multiple indicator systems (MIS), and ad hoc adoption of multivariate datasets as suites of surrogate indicators is not directly defensible when contested. The proponents of an MIS have an obligation to analyze the comparative capacity for objective ordering contained in the constituents of a data matrix before advocating adoption as a suite of surrogates for addressing more abstract aspects. Exploratory examination of any available datasets collected for collateral purposes can augment insights relative to suitable surrogates for ideal indicators, with particular attention to ordering relations for subsets of quantifiers and ensembles of entities.

A multivariate data matrix consists of several quantifiers (variates or variables) as columns recorded on multiple entities (objects, cases, instances, etc.) as rows. The given form of data matrix is not necessarily directly useful nor fully informative for analytically inferring order among entities. In this chapter, we consider some approaches which may be helpful in extracting insights on ordering properties that are embodied in multivariate datasets and applicable in configuring suites of indicators.

Commonality of orientation is crucial, and we begin by exploring implications of alternate orientations. Incomparability is adverse with regard to ordering and provides an initial basis for such assessment. We seek succinctness and consider culling candidates regarding redundancy. When suppressing a candidate who has little impact on comparability, it can be considered relatively redundant. When relatively redundant candidates have been suppressed, interest shifts to the incomparability induced by each of the remaining indicators. An indicator having substantial specificity with regard to comparability is speaking to possible distinctive ensembles of entities (objects) that can be further investigated otherwise, such as through clustering. An indicator that induces considerable general incomparability and isolation suggests a distinctive separate dimension of comparison.

Scaling and crispness of data are concerns, with some aspects being considered and/or illustrated. Compositing and weighting of indicators are also discussed. Empirical versus inferential aspects and approximation are of interest as constrained by computational complexity and practicality of graphical depiction. The Hasse Diagram Technique (HDT) has constraints of practicality due to rapid loss of interpretability with increasing number of object entities. Some inferential aspects can be introduced by repeated sampling and compiling frequencies of incomparability by sample size along with constructing representative Hasse diagrams for different sample sizes. Approximations and diagrammatic representations can be obtained from local partial order models and posetic features of up-sets and down-sets. When the entities to be ordered are samples from a larger universe, questions arise as to the global meaning of orderings within samples—particularly so for a single sample or time series of samples.

In overview, the task of determining a sequence, i.e., a linear or weak order or meta-sequence (sequence of sequences, as for example a poset) usually begins with definition of the entities (objects) of interest, i.e., in defining the **object set**. Then the purpose and nature of the intended sequencing is to be defined. Often the purpose has no counterpart in a measurable quantity, so surrogates are formulated to provide a **suite of indicators** as a proxy. If this is done in a knowledge-based manner, we call it here the (preferred) a priori approach. Each of several such indicators should reflect as least a difference of nuance so that redundancy is avoided. The alternate a posteriori approach of primary focus here is to exploit existing multivariate dataset(s) collected for some parallel purpose such as monitoring, diagnosis, or investigation. Questions then arise as to which (column) variables in the dataset can be considered as indicative for the current purpose and in what manner.

After having defined the object set and selected a suite of candidate indicators, additional aspects must be addressed before sequencing is useful:

- Is the data representation appropriate, i.e., how does one handle numerical differences? Is there any slight numerical difference indicative of order within a pair of objects?
- How should one consider candidate indicators of different scaling levels? Whereas in natural sciences metric information is often available; in other sciences the information is ordinal as numbers are assigned to linguistic descriptors. Should all indicators then be expressed in ranks as the "lowest common denominator" of scaling strength?
- Can indicators be culled (suppressed, set aside, eliminated) on the basis of relative redundancy as reflected in marginal impact on comparability? Do certain candidates speak primarily to small ensembles of specific objects, thus being optional for "fine tuning?"

These questions are among those considered in this chapter with the intent of suggesting methodological approaches when the context of sequencing does not give thorough guidance.

The chapter is organized as follows:

1. Matters of orientation and reducing redundancy with the theory of partially ordered sets
2. Formulation of matrices for retrospective indicator systems
3. Analytical aspects of structure and structuring (since many chapters in this book are concerned with these aspects, this is only in overview)
4. Conflict analysis (this is also in overview as a subject of several chapters)

2.2 Orientation and Partitioning

2.2.1 Theoretical Framework of Partially Ordered Sets (Posets)

Much of what follows is framed in the theory of partially ordered sets. There are many references available, as for example, Brüggemann and Patil (2011) or Brüggemann and Voigt (2008).

2.2.1.1 Axioms of Partial Orders

Suppose that objects of interest comprise an "object set" *X* and further that *X* is a finite set. For five objects *a*, *b*, *c*, *d*, and *e*:

$$X = \left\{a, b, c, d, e\right\}.$$

Objects of the set *X* are to be compared and arranged (meta)-sequentially. The symbol ≤ is used as a binary relation among the objects. The role of this relation is specified by axioms:

$$\text{Axiom 1}: \quad \text{Reflexivity}: \quad x \in X : x \le x \tag{2.1}$$

$$\text{Axiom 2}: \quad \text{Anti-symmetry}: \quad x \le y, y \le x \text{ implies } y = x \tag{2.2}$$

$$\text{Axiom 3}: \quad \text{Trannsitivity}: \quad x \le y \, and \, y \le z \text{ implies } x \le z \tag{2.3}$$

Reflexivity: An object can be compared with itself.

Antisymmetry: If both comparisons are valid, i.e., $y \le x$ and at the same time, $x \le y$, then this axiom demands that *x* is identical with *y*.

Transitivity: Transitivity is present if the objects are characterized by properties which are at least ordinal scaled.

These axioms give the set *X* an algebraic structure, namely that of a partially ordered set (poset). For comparison purposes, the selection of ≤ must be considered as a narrowing. A more general theory of comparisons is the tournament theory (see for instance Bartel and Mucha 2014).

2.2.1.2 Quotient and Object Sets

Several objects may have the same numerical values as different individuals (ties). In such case the objects are considered as equivalent, expressing that they have identical rows in the data matrix, but are nevertheless distinct. These objects form an equivalence class, and one object from the equivalence class is often selected to represent all the others.

To make a clear distinction:

- The set of equivalence classes under an equivalence relation R is called quotient set, denoted by X/R.
- From any equivalence class one object may be selected as representative of the others for purposes of comparison.
- The object set retains identical objects as individual elements.
- It is common practice to consider the data matrix, consisting of n rows/objects and m columns/indicators.

2.2.1.3 Comparative Structure Induced by Hasse-Diagram Technique

Let x, y be two different objects of the object set X, with $q(x)$ being the data row for x and $q(y)$ for y. Then

$$\begin{aligned} &x \le y \text{ if and only if } q(x) \le q(y), \\ &q(x) \le q(y) \text{ if and only if } q(x) \le q(y) \text{ for all i.} \end{aligned} \tag{2.4}$$

If x, y are different objects but $q(x) = q(y)$, i.e., $q_i(x)=q_i(y)$ for all i, then the objects x and y are equivalent, denoted as: $x \cong y$

When some $q_i(x) < q_i(y)$ but some others $q_i(x) > q_i(y)$, then x and y are incomparable, denoted as: $x \parallel y$.

When mutual incomparability appears for all objects of a subset $X' \subseteq X$, then X' is an "antichain."

When for the objects x, y, it is valid that $q(x) \le q(y)$ or $q(x) \ge q(y)$, then x and y are comparable, denoted as: $x \perp y$.

$$An\, object\, x\, generates\, a\, down-set\, as\, x \in X : O\left(x\right) := \left\{y \in Y : y \le x\right\} \tag{2.5}$$

$$An\, object\, x\, generates\, an\, up-set\, as\, x \in X : F\left(x\right) := \left\{y \in Y : y \ge x\right\} \tag{2.6}$$

$$\begin{aligned} &An\, object\, x\, may\, generate\, \mathrm{a}\, set\, of\;\, x-incomparables\, as : x \in X : U\left(x\right): \\ &\qquad = \left\{y \in Y : y \parallel x\right\} \end{aligned} \tag{2.7}$$

We call $q_i(.)$ an attribute or indicator interchangeably.

2.2.1.4 Difficulties in Dealing Directly with Hasse Diagrams and Underlying Issues

Equation (2.4) can be applied to every dataset insofar as there are no data gaps, giving rise to an associated Hasse diagram representing the order relations among objects. There are some issues of interpretability for Hasse diagrams, however, and as to how far application of (2.4) is practically informative.

The difficulty that most often becomes apparent is that the graphical representation of a poset by a Hasse diagram increases very rapidly in its perceptual complexity with increasing number of objects, becoming a blur of intersecting lines that is not visually informative. The more complex the comparative structure, the more rapid the degradation of interpretability for a Hasse diagram. The diagram also does not reveal comparative substructure that may reside in column-wise subsets of the indicative data matrix, which is an underlying issue for the Hasse Diagram Technique (HDT).

Isolation of an object is a foundational feature of HDT that stymies further direct interpretation of that object relative to other objects since it lacks comparability to any other (see for a heuristic Al-Sharrah 2014). The extreme degenerate case is with every object as an isolate whereby there is complete lack of both comparability and sequencing. Since Eq. (2.4) encompasses the entire set of prospective indicators, it then offers no insights regarding comparative substructure that may reside within the candidate data. The difficulties with Hasse diagrams arise from the presumption that such diagrams directly provide in every case a foundational representation of the comparative content of an informational context. There is need for supplemental processing and/or alternative depictions that effectively allow generalized views of comparative content and substructure for the candidate data.

Partial orders can be visualized in coordinate systems such as scatter plots, with the basic idea being to extract from the posets coordinate values for each object containing comparative information of interest. Such information can often be obtained from frequencies of features in the Hasse diagram. For example the difference between number of objects in a down-set and number of objects in the up-set may be one coordinate, whereas the other coordinate counts $U(x)$. One may even differentiate among the elements of $U(x)$, according to whether they are isolated elements or not. For the concept of posetic coordinates, see also Myers and Patil (2013a).

Formal concept lattices with their symmetric view on indicators and objects constitute a theoretical approach for exploration beyond HDT that is relevant but not presently pursued (for details see for instance, Annoni and Brüggemann (2008) or Brüggemann and Patil (2010, 2011)].

2.2.2 Reorientation of Candidate Indicators

Unless it is apparent from the context, attention should be given to orientation of column variables as candidate indicators. Statistical evidence of orientation issues in the data come as substantial negative rank correlations and affect HDT by increasing incomparables $U(x)$. The statistical signal can be observed by inspection of a rank correlation matrix, but the effect on $U(x)$ may not be readily detectable by simple inspection.

2.2.2.1 Benchmarking

If several experts are involved who can reach consensus regarding appropriate order for a subset of objects, an expedient approach is to compare original and reversed orientation for each column separately to see which best reproduces the order of the experts for those particular objects. If this can be satisfactorily accomplished, then the more laborious combinatorial approach below can be avoided.

2.2.2.2 Comparative Combinatorial Reorientation

Exploratory analysis can be conducted with regard to amount of incomparability in relation to changing orientation. The "reorientation" is to be exemplified by first considering two fictitious datasets and then a sample of actual data.

Dataset 1:
As one can easily see, quantifier q_2 is just the reverse of quantifier q_1. If one of the first two columns is reversible to accord better with the third, then the first two columns become redundant. The third column differs from the first by pair-wise reversals for object *a* with *b* and object *e* with *f*, along with a tie for objects *c* and *d*. The third column differs more relative to the second, since reversed pairs are reversed. If reorientation is not admissible, then this dataset offers no context of crisp comparability.

An indicator function of incomparability (2.8) as a variant of $U := |U(x)|$ serves here as the criterion for choosing among combinations of reversals:

$$Ugr(X) := \left| \left\{ (x,y) : x, y \in X, x \parallel y \right\} \right| \tag{2.8}$$

with (2.8) being a function on the set of all data matrices resulting from the original matrix by reversing a subset of indicative quantifiers. This set of data matrices has 2^m members including the original data matrix. Bit patterns are used to designate combinations of columns, with a 1 in the string of binary digits (bits) designating that the column is reversed. Examples of bit patterns are as follows:

[0, 0, 0] no data column is reversed

[1, 0, 0] first data column is reversed

[1, 1, 0] first and second data columns are reversed

There are some redundancies among the bit-patterns for Table 2.1 with respect to comparability of objects since reversing an entire table does not change comparability; for instance [1, 0, 0] represents the same set of order relations as [0, 1, 1] in an upside-down manner. The indicator function Ugr([0, 1, 1]) is the same as Ugr([1, 0, 0]), except that the Hasse diagrams of [1, 0, 0] and of [0, 1, 1] are dual (see Brüggemann and Patil 2011). Note also that the pattern [1, 0, 0] is the complement of [0, 1, 1] in Boolean algebra.

Applying (2.8) for all possible eight data matrices of Table 2.1 leads to the following graph (Fig. 2.1).

Table 2.1 First fictitious dataset

	Quantifiers		
Objects	q_1	q_2	q_3
a	1	6	2
b	2	5	1
c	3	4	4
d	4	3	4
e	5	2	6
f	6	1	5

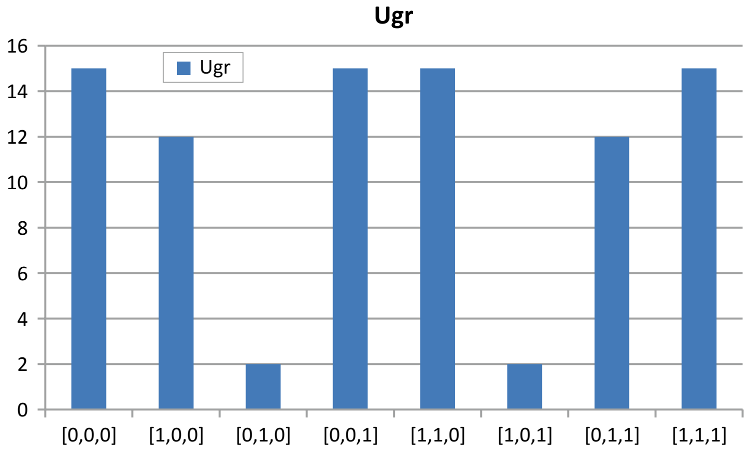
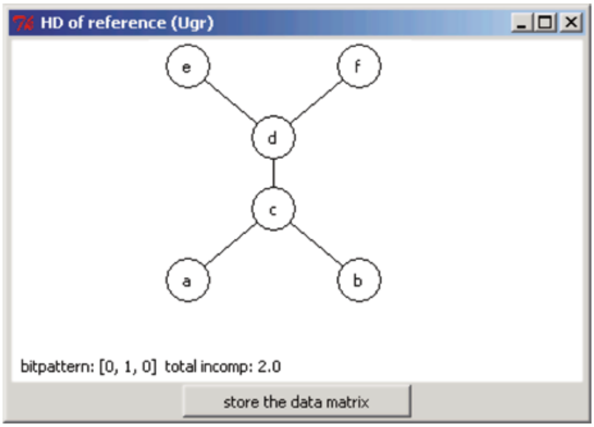


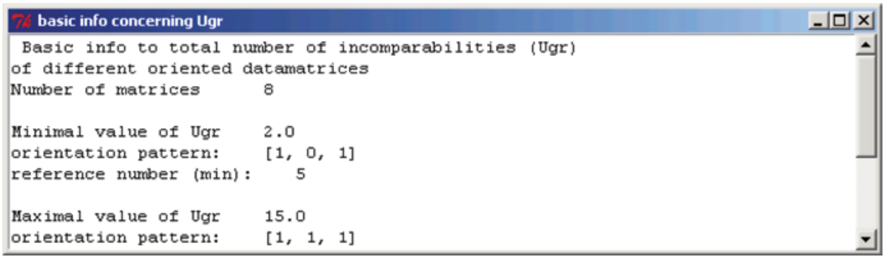
Fig. 2.1 Ugr [Eq. (2.8)] as a function of the eight different reorientation patterns



bitpattern: [0, 1, 0] total incomp: 2.0

Fig. 2.2 Hasse diagram from data matrix having the second column reversed (PyHasse module: orientation1.py)

One can see that reversing the second column [0, 1, 0] would lead to the lowest Ugr and to a Hasse diagram richest with respect to the order relations ⊥ (Fig. 2.2). For the reason given above, the complement of this pattern [1, 0, 1] also carries the same comparability but with the sense of the entire table being reversed. Likewise,



```
 Basic info to total number of incomparabilities (Ugr)
of different oriented datamatrices
Number of matrices      8

Minimal value of Ugr    2.0
orientation pattern:    [1, 0, 1]
reference number (min):    5

Maximal value of Ugr    15.0
orientation pattern:    [1, 1, 1]
```

Fig. 2.3 PyHasse interface for basic information about the set of data matrices with all combinations of attributes reversed

Table 2.2 Information about U-scanning through all possible (Table 2.1) column orientations

Basic info on total number of incomparabilities (Ugr) of differently oriented data matrices	
Number of matrices	8
Minimal value of Ugr	2.0
Orientation pattern:	[1, 0, 1]
Reference number (min):	5
Maximal value of Ugr	15.0
Orientation pattern:	[1, 1, 1]
Reference number (max):	7
Mean value of Ugr	11.0
Reference numbers of minimal values of Ugr:	2, 5

reversing all columns is as bad as not reversing any. Reversing the first column is not the same as reversing the second column, since the third column is more like (higher rank correlation) the first than the second. Reversing both, second and third column, compounds conflict with the third column. It should also be noted with reference to Fig. 2.2 that presence of a completely redundant column does not increase the incomparability of objects, since one conflict is sufficient to make objects incomparable under the product-order relation used in the definitions above.

However, in general, it is not a tractable proposition to draw figures such as Fig. 2.1 since combinatorial expansion would entail a diagram of 1,024 bars for as few as ten quantifiers. Tabular condensation is needed as shown in Fig. 2.3 where the minimal and maximal values of Ugr are reported together with their bit-patterns (extracted from full information provided by the PyHasse module “orientation1.py” as given in Table 2.2). In order to refer to a bit-pattern, a reference number is introduced as follows.

Let $[a_m, a_{m-1}, \ldots, a_1]$ be components of a bit-pattern with a_k referring to contribution of 2^{k-1}, then

$$Reference\,number = \sum a_i * 2^{i-1}\left(i = 1, \ldots, m\right)\ and\ a_i \in \{0,1\}. \tag{2.9}$$

Table 2.2 includes a reference number for maximal value of Ugr which was truncated from Fig. 2.3.

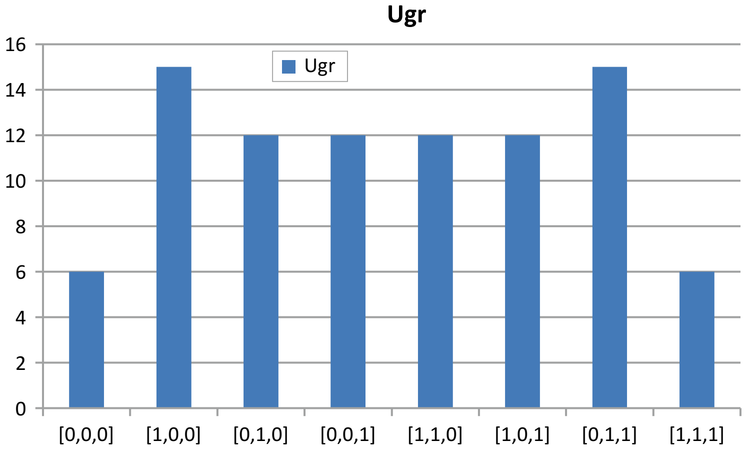


Fig. 2.4 Variation of the Ugr-function [Eq. (2.8)] in the case of data matrix of Table 2.3

Table 2.3 Second fictitious dataset

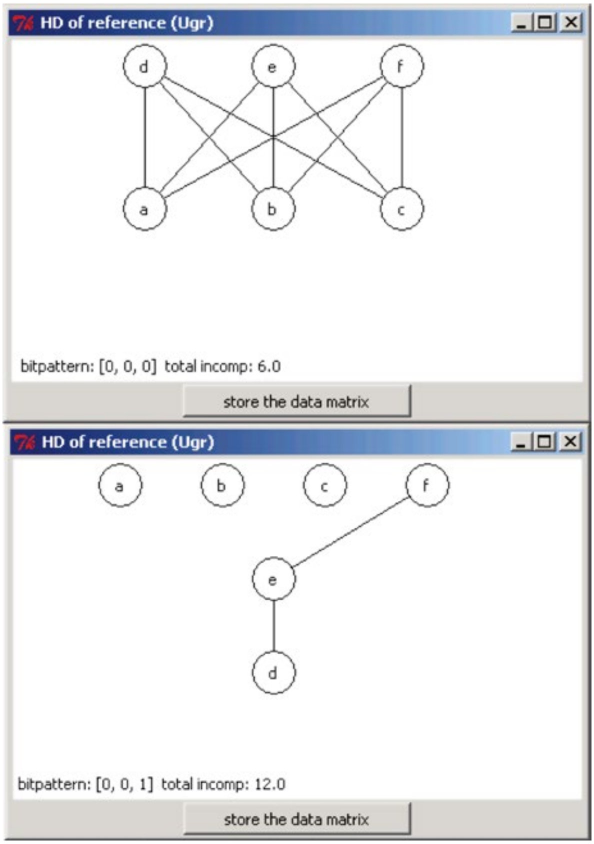
	q_1	q_2	q_3
a	1	3	1
b	2	2	2
c	3	1	3
d	4	5	6
e	5	6	5
f	6	7	4

Table 2.4 Basic results of orientation1.py applied on data of Table 2.3 (second fictitious dataset)

Basic info to total number of incomparabilities (Ugr) of different oriented data matrices	
Number of matrices	8
Minimal value of Ugr	6.0
Orientation pattern:	[1, 1, 1]
Reference number (min):	7
Maximal value of Ugr	15.0
Orientation pattern:	[1, 0, 0]
Reference number (max):	4
Mean value of Ugr	11.25
Reference numbers of minimal values of Ugr:	0, 7

Dataset 2:
Here there are six objects and three quantifiers without any completely countercurrent columns. The first column is the same as for dataset 1. The value 4 does not appear in column q_2.

The result of scanning column reversals for dataset 2 is given in Table 2.4. In this dataset, column reversals change Ugr by a factor of 2.5 (= Ugr_{max})/Ugr_{min}). However, the minimum is for the original orientation with any change increasing the number of incomparabilities—so none should be done. Figure 2.4 shows the bar diagram of the Ugr values for different combinations of column orientations.



```
bitpattern: [0, 0, 0] total incomp: 6.0
```

```
bitpattern: [0, 0, 1] total incomp: 12.0
```

Fig. 2.5 Two examples of Hasse diagrams based on Table 2.3

In Fig. 2.5 two examples of Hasse diagrams are shown: one for the original configuration [0, 0, 0] and one for the configuration [0, 0, 1]. Note that reversal of the third indicator has made half of the objects isolates with an ostensible complete lack of comparability. Here, the possible analysis of antichains and other variants (Carlsen and Brüggemann 2014) are not shown.

Dataset 3:
It may be of interest to check a sample of actual data. For this, a data matrix is selected which is a part of a larger matrix discussed by Myers and Patil (2013a). There are ten indicative attributes as percentages of different kinds of land cover. For regions defined in terms of biology and geomorphology, different kinds of land cover have intrinsic implications as indicators of landscape ecology in a particular region. Indicator Pct40 is related to the fact that a given region has forest as a natural land cover if ecological succession is allowed to proceed without intervention. Natural disturbance dynamics have transitional cover (Indicator Pct33) where disturbed patches are again reverting to forest. Humans are a major (non-natural) disturbance agent in this region. The most extreme and enduring scars of

humanization in these landscapes are quarries, strip mines, and gravel pits (Pct32). Development (Pct20) along with urban/recreational grasses (Pct85) constitutes a prevalent aspect of humanization as long-term disturbance. Agriculture comprises a second prominent aspect of humanization comprised of crops (Pct82) and hay/pasture (Pct81) with crops involving annual exposure of the soil surface to erosion in this relatively steep terrain. Loss of natural wetlands to humanization has led to restrictions on further conversion of woody wetlands (Pct91) and emergent herbaceous wetlands (Pct92) which together currently comprise a small fraction of total landscape area. Water (Pct11) naturally occurs primarily as rivers and streams, but a large impoundment called Raystown Lake is also present. The objects are geographical units called OCTIVs (OCTagonal Integrating Vicinities). The number of objects is restricted to a sample of 12 for a local context with the matrix of reduced rows being shown in Table 2.5. This data matrix exemplifies the degenerate case wherein all objects are isolates with none being comparable to any others. Table 2.6 shows the result of reorientation, whereby some comparability can be induced.

When we take the algebraic complement of [1, 1, 1, 0, 1, 0, 0, 1, 1, 1] namely [0, 0, 0, 1, 0, 1, 1, 0, 0, 0], we see that three attributes reversed would give an enriched poset, where Ugr is reduced by a factor of 1.2 (66/55). There seems to be no single attribute which is completely countercurrent to the others. The number Ugr = 66 corresponds to the maximal possible U. The Hasse diagram, corresponding to the bit pattern [1, 1, 1, 0, 1, 0, 0, 1, 1, 1] is shown in Fig. 2.6, having only three isolates.

The attributes for which reversal is indicated are:

PCT33: Transitional
PCT 81: Hay/pastures
PCT 82: Row crops

These three attributes belong to those land cover characteristics which are considered as neither pro nor con (see Myers and Patil 2013a). Hence the reorientation study suggests reversing three attributes in order to get a more enriched Hasse diagram (which would be the dual poset of that shown in Fig. 2.6).

Additional computational and structural considerations

For practical purposes, exploration of orientation also entails consideration of computation. The Ugr scan is comprehensively combinatorial and therefore fully informative. However, it is computationally most intensive. On way to ease such computational burden is to make the approach somewhat sequential rather than completely combinatorial. Accordingly, an initial sweep can invert each column individually and retain the one inversion that has most impact. The remaining columns in this modified matrix can then each be examined individually for additional improvement, with the best of these being retained in a twice modified matrix. This progression of successive modification continues until no improvement is obtained. Still another strategy would be to reduce computation by investigating random samples from datasets having large numbers of objects (rows).

If the burden of combinatorial computation is reduced in the foregoing manner, then it also becomes feasible to incorporate disjoint features of the Hasse diagrams in the progressive scans. The most obvious of these features are isolates, since HDT structure that is disjoint will preclude any sort of inter-subset inference.

Table 2.5 Twelve geographical units called OCTIVs (OCTagonal Integrating Vicinities) by Myers and Patil (2013a, b) with ten IVI (Integrative Vicinity Indicator) attributes as percentages of different kinds of land cover

ID	Pct11	Pct20	Pct32	Pct33	Pct40	Pct81	Pct82	Pct85	Pct91	Pct92
25	0.007961	0.007961	0.000000	0.732425	32.20285	59.159302	7.889499	0.000000	0.000000	0.000000
26	0.000000	0.007959	0.000000	0.127358	86.93783	10.745840	1.265621	0.000000	0.875587	0.039799
36	42.629608	0.326511	0.000000	0.238910	56.31918	0.015927	0.103527	0.000000	0.000000	0.366329
37	1.751871	0.055741	0.000000	0.541487	95.54069	1.600573	0.445930	0.000000	0.000000	0.063704
38	0.047778	0.000000	0.000000	0.000000	90.38063	8.862876	0.668896	0.000000	0.000000	0.039815
39	0.055719	0.047759	0.000000	0.692509	89.74767	7.689246	1.767093	0.000000	0.000000	0.000000
48	0.063704	0.509635	0.000000	0.597228	54.32394	36.765408	7.684344	0.000000	0.031852	0.023889
49	4.872223	0.063689	0.000000	0.000000	93.25691	1.249900	0.326407	0.000000	0.055728	0.175145
50	4.139468	0.103486	0.000000	1.010985	92.38975	1.807037	0.167170	0.000000	0.151249	0.230854
51	1.480655	1.369208	1.098551	0.175131	94.77790	0.573157	0.310460	0.000000	0.000000	0.214933
59	0.055736	2.508161	0.000000	0.127398	62.49701	26.690023	4.912811	2.508161	0.270722	0.429970
60	3.495222	2.866242	0.000000	0.987261	87.67516	2.300955	1.815286	0.000000	0.597133	0.262738

Table 2.6 Reorientation of the data matrix of Table 2.5

Basic info on total number of incomparabilities (Ugr) of differently oriented data matrices	
Number of matrices	1,024
Minimal value of Ugr	55.0
Orientation pattern:	[1, 1, 1, 0, 1, 0, 0, 1, 1, 1]
Reference number (min):	935
Maximal value of Ugr	66.0
Orientation pattern:	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Reference number (max):	1023
Mean value of Ugr	65.45
Reference numbers of minimal values of Ugr:	88, 935

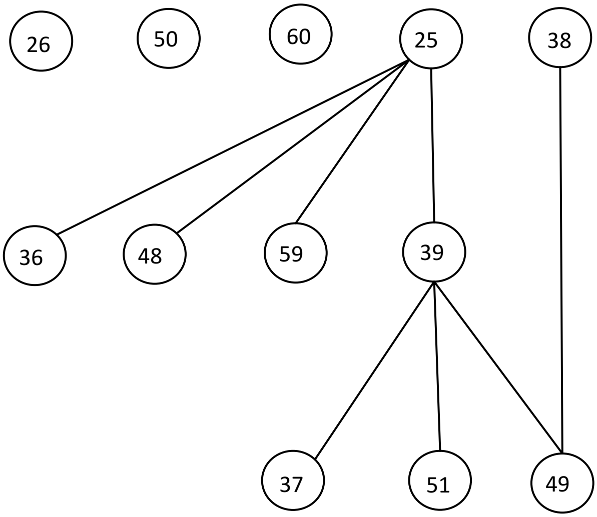


Fig. 2.6 Twelve regions in Pennsylvania under ten attributes (partially reoriented data matrix)

2.2.3 *Elimination of Individual Candidate Indicators*

One may ask as to how far single indicative quantifiers are individually crucial with respect to order relations. Therefore another kind of exploratory analysis can be performed whereby eliminating the effect of one candidate indicator is considered. The number of different matrix outcomes is far less for such single eliminations than for combinatorial reorientation. The procedure is similar to the sensitivity analysis with the central concept of the matrix W (see Brüggemann and Patil 2011; Brüggemann et al. 2001), however, the indicative quantities are not the entries of the matrix W but $|U(x)|$.

Since the scope of scanning is different, another variant of $U := |U(x)|$ is denoted by Uli as the analytical indicator (or sensitivity) function. The set of matrices for which Uli is calculated contains $m+1$ data matrices with these being the original and m data matrices where one single column as a candidate indicator is eliminated. Uli has a value corresponding to elimination of each individual column indicator

Table 2.7 Basic information on (Uli) incomparability for single elimination in Table 2.1

Indicator	1	Not considered: Uli=	12.0
Indicator	2	Not considered: Uli=	2.0
Indicator	3	Not considered: Uli=	15.0

Table 2.8 Basic information on (Uli) incomparability for single elimination in Table 2.2

Indicator	1	Not considered: Uli=	6.0
Indicator	2	Not considered: Uli=	3.0
Indicator	3	Not considered: Uli=	3.0

which is the number of incomparabilities remaining after elimination of the respective column. This is exemplified by applying the PyHasse module on the data matrices of Tables 2.1 and 2.2.

2.2.3.1 Dataset 1 (Table 2.1)

An investigator would normally explore elimination after doing any indicated reorientation. For comparison with the results of the earlier reorientation study, however, we work directly on the data of Table 2.1. The result for the dataset of Table 2.1 is shown in Table 2.7 and is confirmatory to and consistent with prospective reorientation. Eliminating the second indicator reduces the incomparabilities to 2. This is the same number obtained by reversing the second indicator to be completely redundant with the first indicator. This verifies that a completely redundant indicator has no effect on the number of incomparabilities. Hence the finding of reorientation and appropriateness of elimination are both confirmed.

2.2.3.2 Dataset 2 (Table 2.3)

The result for column (candidate indicator) elimination for dataset 2 (Table 2.3) is shown in Table 2.8. For this dataset, the exploration of orientation was not suggestive of any change. Here the order structure is most sensitive to elimination of the first column. It is less sensitive, but equally so, to elimination of either column 2 or 3. This illustration sets the stage for what follows.

2.2.3.3 Sample of Actual Data (Table 2.5)

For the sample of actual data in Table 2.5 (without reorientation) the result of single elimination is not informative. No elimination of a single attribute from the original indicator matrix reduces Uli. The complete isolation of objects for the full dataset remains so for all single-column eliminations.

Table 2.9 Impact values from PyHasse module (sensitivity18_3.py) for reoriented land cover matrix

Property	Sensitivity	Property	Sensitivity	Property	Sensitivity
Pct32	0	Pct92	0	Pct40	3
Pct81	0	Pct11	2	Pct33	7
Pct85	0	Pct20	2		
Pct82	0	Pct91	2		

2.2.4 Culling of Candidate Indicators

Analysis of the impact of each single attribute on the structure of a poset (see Sect. 2.2.1) has attracted considerable interest (see Brüggemann et al. 2001). On that basis, attention is focused here on application of the cumulative ambiguity graph concept (Brüggemann and Patil 2011, 2010). This entails a sensitivity analysis using incomparability as a sensitivity criterion. The attributes are ordered from high to low impact on the poset and successively incorporated beginning with the one that is most influential. Therefore, m data matrices are to be checked. Again $U := |U(x)|$ provides an indicator function for ambiguity and a plot is obtained with U (scaled as fraction of maximum ambiguity) on the ordinate and number of included attributes, "natt," on the abscissa. The indicator function can attain values "near" the maximum either when natt approaches m or even at values of natt = m^* << m. In the latter case, a segregation of the indicative attributes is motivated. The first m^* attributes are considered as important for retention, whereas the remaining $m{-}m^*$ attributes are considered as relatively redundant. Those attributes having relative redundancy are either eliminated or set aside as optional "fine tuning" (nuance) attributes for later inclusion if needed.

This method of sensitivity analysis is to be applied considering that the suggested column reversals provide an oriented matrix (odm) of land cover data. Table 2.9 shows results obtained from PyHasse module (sensitivity18_3.py) for sensitivity in terms of HDT mismatches induced by deleting the respective indicator. The cumulative ambiguity maximum (CAM) graph is shown in Fig. 2.7.

There are some ties with respect to the attribute sensitivity data, with ties being ordered arbitrarily:

Pct33 > Pct40 > Pct91 = Pct20 = Pct11 > Pct92 = Pct82 = Pct85 = Pct81 = Pct32.

This graph and Table 2.9 is suggestive of the first five (ordered) attributes being important for retention and the rest being essentially disposable. The user interface of PyHasse module (sensitivity19_1.py) is shown in Fig. 2.8. A sensitivity analysis for this subset of five (oriented) attributes is given graphically in Fig. 2.9.

From Fig. 2.9:

"Emergent Herbaceous Wetlands" (Pct 92) > "reoriented Hay Pasture" (Pct81) > "reoriented Row Crops" (Pct82) > "Quarries etc." (Pct32) > "Urban/Recreational Grasses" (Pct32).

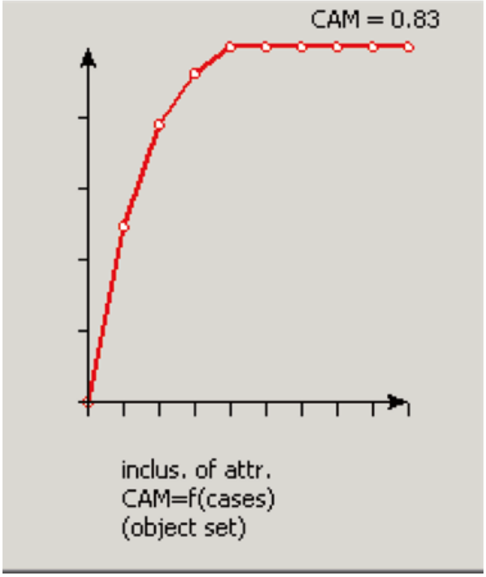
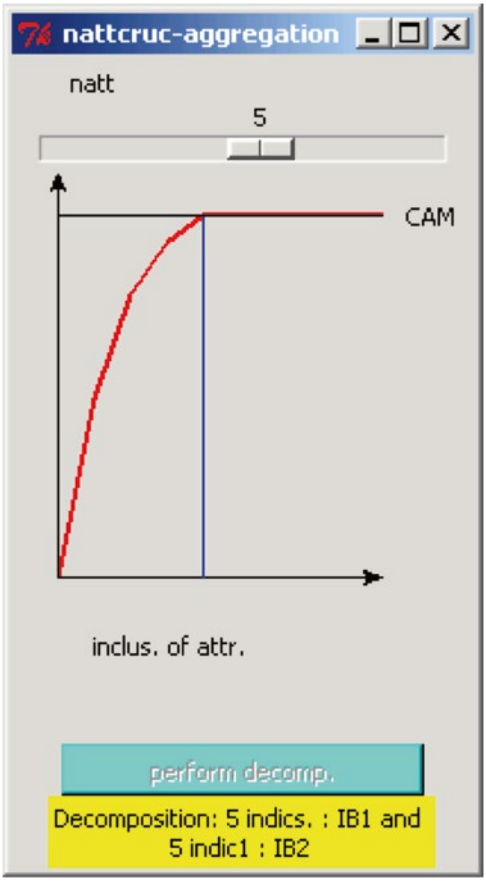


Fig. 2.7 CAM (cumulative ambiguity maximum) graph of the partially reoriented land cover data with ordinate as fraction of maximum ambiguity (incomparability)



Decomposition: 5 indics. : IB1 and 5 indic1 : IB2

Fig. 2.8 Interactively selecting the most important indicative attributes on the basis of a CAM-graph

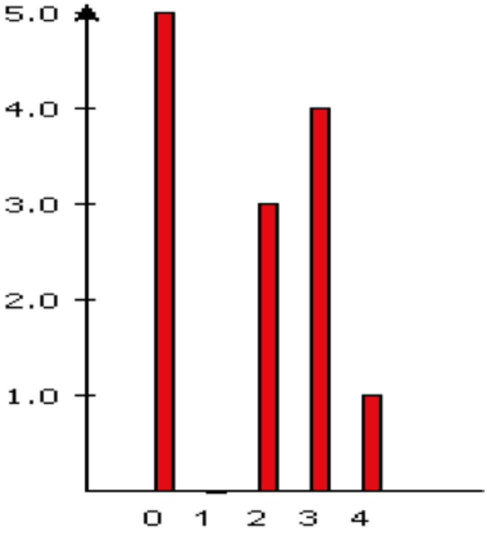


Fig. 2.9 Sensitivity (HDT impact) of five indicators: Pct92 (0), Pct85 (1), Pct82 (2), Pct81 (3), Pct32 (4)

Taking these five most important (reoriented) attributes, a poset results with $U = 23$, whereas U of the reoriented (odm) complete matrix is 55. Seven HDT levels are obtained (Fig. 2.10), indicating substantial gain of order relations. Object 25 is the single maximal element (i.e., a greatest element), with objects 36, 51, and 59 being minimal elements. There are now no isolated elements. We call this data matrix with the five most important (reoriented) indicators as a redacted or "refined" data matrix (rdm).

Explorations of culling attributes need not be limited to the foregoing dichotomy of "major" and "minor" indicators. Two indicators that order the objects in the same manner are order-redundant, and with only a few perturbations there is still a high degree of such redundancy. Given that one such indicator is in the suite, another will only marginally alter the perspective on order relations. The statistical signal of nearly redundant indicators is a strong rank correlation. Conversely, indicators with weak or negative rank correlation will order objects differently and thus introduce appreciable ambiguity. This is essentially the same perspective as above, but framed somewhat differently. This framing would suggest rearranging the indicators so that higher degrees of redundancy are reflected in a gradational block diagonal pattern for the rank correlation matrix. If such a pattern emerges, it can motivate culling of the indicators by selecting one that is central to each block. This reduced suite of indicators can then be compared in terms of effect to the major ("important") indicators emerging from the foregoing method. Carrying this line of exploration further, each block of indicators can be segregated and separately analyzed to obtain its characteristic structural signals in HDT. If distinctive structural patterns emerge in this manner, then the implication is that there are distinct domains of induced order that merit separate characterization followed by study of how those domains interact.

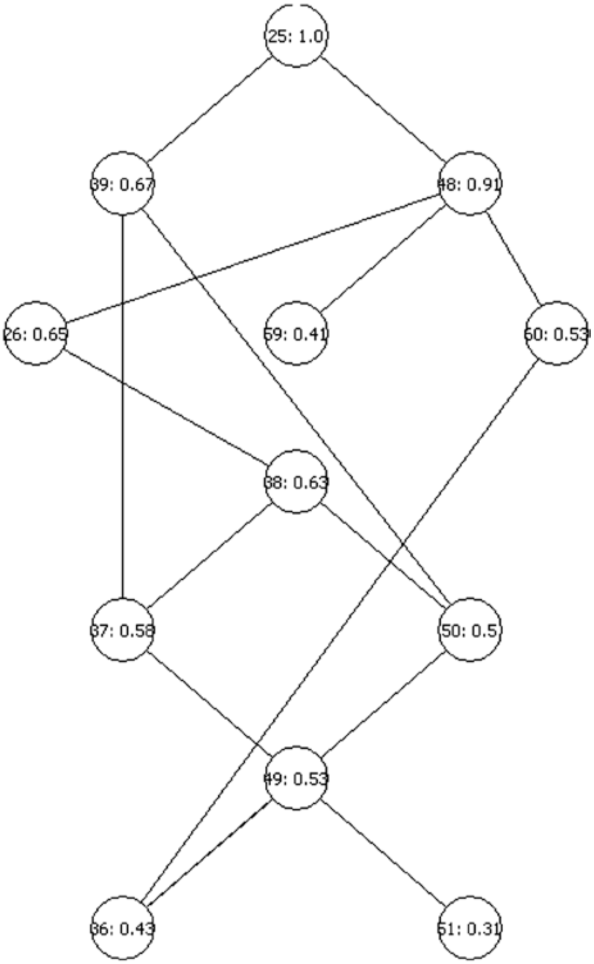


Fig. 2.10 Hasse diagram for land cover data using five major indicators after reorientation and culling of the indicator set. Note: second values inscribed in *circles* are referenced subsequently

2.3 Quantification and Generalization of Indication

2.3.1 Shared Sense of Sequence and Diminutive Differences

In the foregoing section the focus was on reorientation, elimination, and partition of attributes. However, the result of a partial order study also depends heavily on the data representation. One might raise a question regarding which numerical differences of metric attributes are meaningful for sequencing (in terms of ranks). Implicit in such a question, however, is the assumption that magnitudes of differences are meaningful. This assumption holds only for data that are scaled at a level of interval strength, whereas sequencing as ranks need only entail ordinal quantification. Once again, this can be explored with the land cover data of Table 2.5, which is especially suited to this purpose because it addresses compositional components necessarily

Table 2.10 Attribute value differences according to their distribution in 0.1 intervals, based on [0, 1]-normalized data matrix

	Pct11	Pct20	Pct32	Pct33	Pct40	Pct81	Pct82	Pct85	Pct91	Pct92
0.0…0.1:	43	25	55	9	24	19	20	55	28	16
0.1…0.2:	12	12	0	11	7	18	13	0	9	5
0.2…0.3:	0	0	0	5	0	1	4	0	2	3
0.3…0.4:	0	2	0	5	3	3	2	0	8	8
0.4…0.5:	0	8	0	10	7	9	3	0	0	11
0.5…0.6:	0	1	0	8	11	4	4	0	1	9
0.6…0.7:	0	0	0	4	6	3	2	0	9	2
0.7…0.8:	0	2	0	4	0	0	4	0	0	3
0.8…0.9:	1	9	0	6	2	3	2	0	1	4
0.9…1.0:	4	4	0	2	5	5	11	0	2	3

ranging from 0 to 100 %. For compositional data (before any reorientation) the column indicators are additive across a row with upper limit on the total of 100 %. Table 2.10 shows the distribution of differences in values for the compositional indicators in fractional terms; or, more generally, on the basis of a (0, 1)-normalized data matrix. It can be seen that Pct11 has data differences mainly concentrated on the lower values and some few near the maximum possible difference. Pct32 is still more extremely distributed over the ten intervals, since there are only differences in the lowest interval (0.0…0.1). Pct82 has differences spread out over all ten intervals, and the lowest interval is only slightly more frequently represented.

A client might tend to question, for example, whether the slight difference for Pct11 between objects 25 (0.007961 %) and 26 (0.000000 %) in the non-normalized original land cover data matrix should really be taken into account for a ranking, when at the same time differences up to around 42 % are possible. A seemingly apparent approach, namely to declare two objects as equivalent when the numerical difference of an attribute is less than ε (ε arbitrarily selected), is not even possible because of the transitivity of the equivalence relation. Consider three objects *a*, *b*, and *c* with data values 0.1, 0.2, and 0.3 and select $\varepsilon = 0.15$, then certainly a and b are equivalent. In like manner *b*, *c* are equivalent; hence *a*, *b*, and *c* must be in the same equivalence class, although the difference between *a* and *c* is larger than ε! In Brüggemann and Patil (2011) some possibilities for overcoming this problem are described in detail. Three such methods are discretization, cluster analysis, and application of fuzzy concepts. Cluster analysis and fuzzy concepts lead to additional interesting aspects of collective interaction among objects as seen through a lens of multiple indicators.

2.3.2 *Discretization*

The seemingly simplest variant would be to select a range of data for each of the attributes and to select the number of equidistant intervals covering the data range. Instead of $q_i(x)$, object *x* gets a score, $s_i(x)$, indicating to which interval $q_i(x)$ belongs.

In statistics this is generalization in class intervals, and in numerical analysis it is the strategy of generalization by "binning."

It can be argued that such discretization carries drawbacks of arbitrariness in selecting data range and number of classes and that data values having only slight numerical difference can be assigned to different classes (see Brüggemann and Welzl 2002). Although these cautions are well founded, they have not deterred application in either statistics or numerical analysis for achieving generalization. Likewise, it is ill-advised to dismiss the approach because of these cautions in connection with HDT. A major issue for HDT as raised earlier is difficulty in filtering fine structure of a poset so that coarse structure becomes evident. Some occasional perturbations at cutoffs will affect fine structure rather than coarse structure, and generalization should be tunable with regard to its severity. Doing HDT on quotient sets with different bin widths should therefore serve to help elucidate coarse structure as is needed. The variant that is actually simplest is generalization by orders of magnitude through rounding or truncation of successive digits in multi-digit values.

2.3.3 *Clustering, Collectivity, and Posetic Features*

Cluster analysis is one of the several modes of data analysis in the statistical genre that merit greater usage in conjunction with posetic information (Myers and Patil 2012, 2013c). As is often the case, some of the more common configurations for general purposes are not particularly suitable in this regard as with the version called Kmeans which requires specific knowledge about the number of clusters (Bock 1974, 1979; Bock et al. 2003; Diday 1979; Brüggemann et al. 2013a). However, others among the many clustering modalities (Luther et al. 2000) invite imaginative adaptation (see Myers and Patil 2013b).

Since clustering most often addresses objects through (dis)similarity of their attributes, coupling these analytical domains is most readily done through posetic features of individual objects such as down-set for $x \in X: O(x) := \{y \in Y: y \leq x\}$, up-set for $x \in X: F(x) := \{y \in Y: y \geq x\}$, and the x-incomparables $x \in X: U(x) := \{y \in Y: y || x\}$. This is stressed by Myers in several papers (Myers et al. 2006; Myers and Patil 2008, 2010, 2013a), because in areas like data mining the Hasse diagram loses its attractiveness quickly . This extends to more flexible graphic depictions than Hasse diagram itself as is well illustrated by the FOU-plot. The FOU-plot uses as one coordinate the difference of the contents of the down-set ($|O(x)|$) and the up-set ($|F(x)|$) and as second coordinate the content of the incomparable elements with $x: |U(x)|$ (see also Brüggemann et al. 2013c). A diagram of this type which corresponds to the Hasse diagram of Fig. 2.10 is shown in Fig. 2.11.

Leftward in a FOU-plot is less favorable and rightward is more favorable, with higher positions carrying greater incomparability. Thus, the most definitely favorable circumstance is for object 25 at the lower right marked in blue. Similarly, the most definitely unfavorable circumstance is for object 36 marked in blue nearest the lower left.

With regard to clustering, the hierarchical agglomerative mode is most adaptable to the present purpose. It gives rise to a dendrogram by which to view evolution of

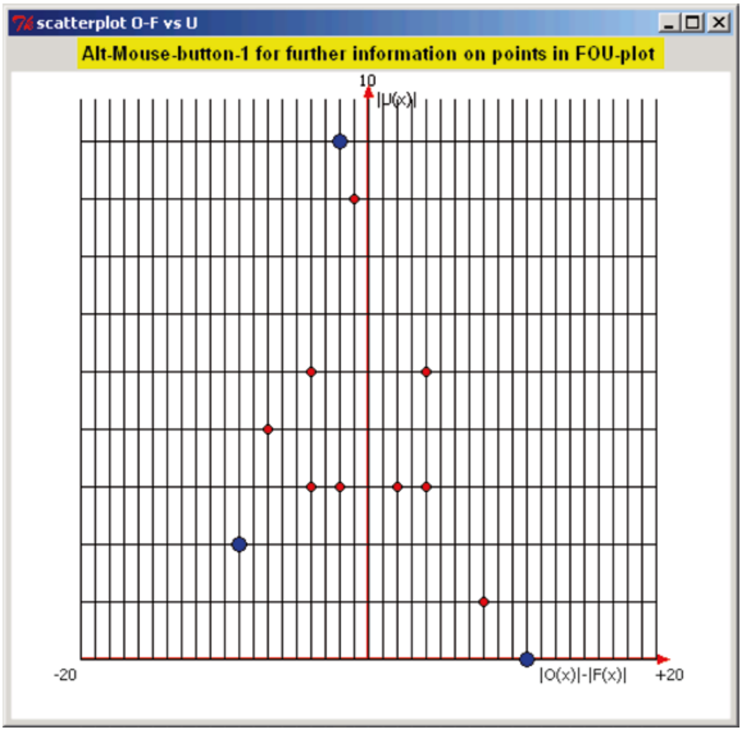


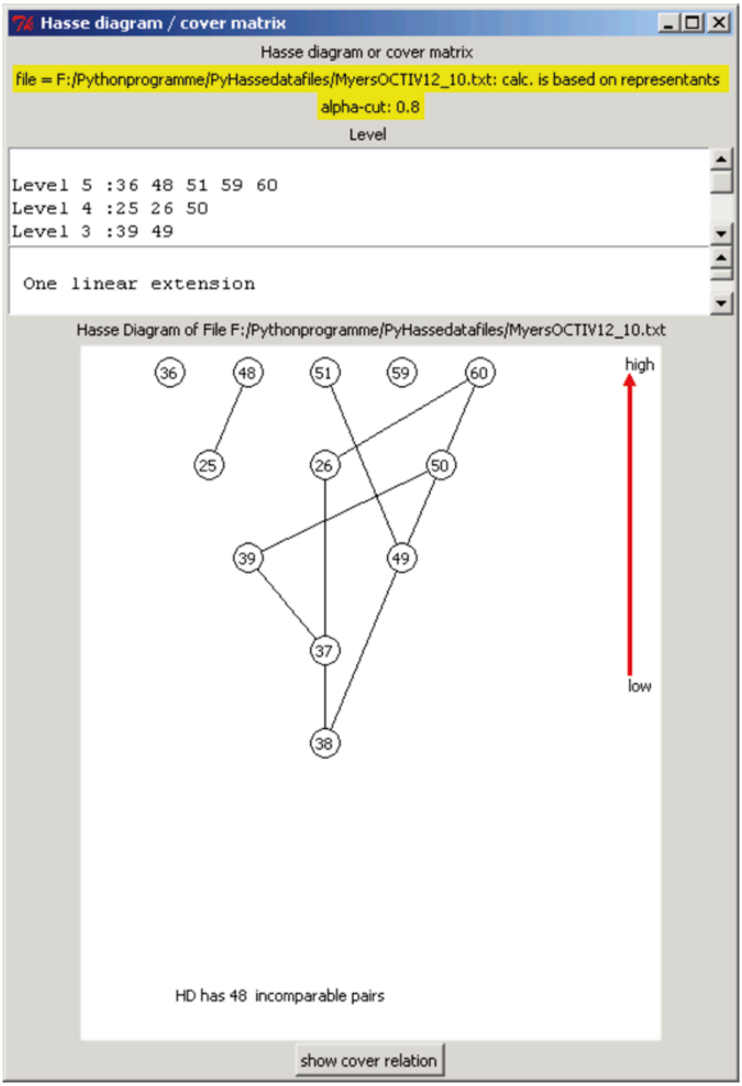
Fig. 2.11 FOU-plot corresponding to the Hasse diagram of Fig. 2.10. *Upper blue point* marks object 59, *left-most blue point* marks object 36, and *right-most blue point* marks object 25

grouping and either graphical or tabular summarization of the objects at any interesting level of grouping can be rendered. The posetic features as illustrated above as well as posetic rankings can be attached to member objects of particular groups with other groups being suppressed for purposes of simplified pattern perception. In this way, posetic characteristics can be carried into the zoom-in and zoom-out generalization of a dendrogram (see Myers and Patil 2013b).

Depicting posetic and attribute characteristics of clusters in lattices of pair-wise plots can be particularly effective for heuristic examination of degrees of joint variation (or what may be called "collectivity") of the indicators. This in turn can suggest exploratory grouping of indicators into different domains of propensity for ordering.

2.3.4 Fuzzy Concept

Fuzzy concepts are described in Brüggemann and Patil (2011) and Brüggemann et al. (2011) and are basically derived from a fuzzy subsethood, where the <-relation among objects is transformed into a fuzzy subsethood (see also Van de Walle et al. 1995).



Hasse diagram / cover matrix

Hasse diagram or cover matrix

file = F:/Pythonprogramme/PyHassedatafiles/MyersOCTIV12_10.txt: calc. is based on representants

alpha-cut: 0.8

Level

```
Level 5 :36 48 51 59 60
Level 4 :25 26 50
Level 3 :39 49
```

```
 One linear extension
```

Hasse Diagram of File F:/Pythonprogramme/PyHassedatafiles/MyersOCTIV12_10.txt

HD has 48 incomparable pairs

Fig. 2.12 Fuzzy results for initial (unmodified) land cover data in Table 2.5

The Fuzzy concept is illustrated here to show possibilities. Exemplification can be done in terms of the land cover data, first using the unmodified data in Table 2.5 with the fuzzyHD13.py module of PyHasse software and then performing defuzzification to obtain the Hasse diagram shown in Fig. 2.12.

Figure 2.12 may be compared to Fig. 2.6 which shows the Hasse diagram obtained after modifying the initial data matrix by orienting columns, while keeping in mind that the data matrix of Table 2.5 gave a complete antichain (all objects as

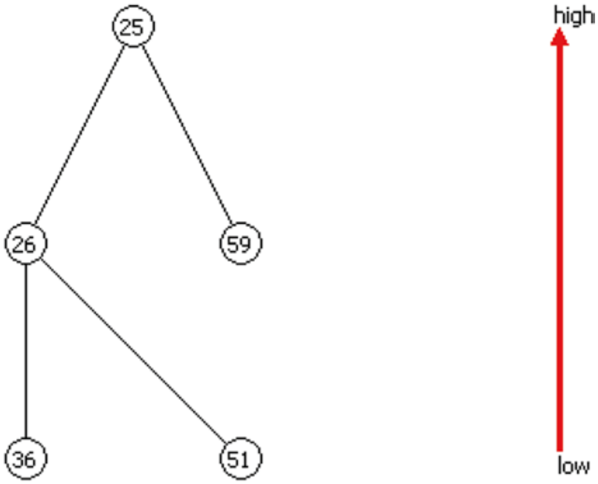


Fig. 2.13 Fuzzy poset for five-column-refined land cover data (compare to Fig. 2.10)

Table 2.11 Equivalence classes of fuzzy poset for five-column-refined land cover data

object 25:	48
object 26:	37 38 39 49 50 60

isolates) before orientation. The fuzzy approach has clearly found structure that was not evident in the crisp approach.

It is of further interest to continue with the fuzzy versus crisp approaches on the five-column-refined data matrix and compare the result to the Hasse diagram of Fig. 2.10. The results of doing so are shown in Fig. 2.13 and Table 2.11.

It is of interest that the indicator function U is now 4, signaling that (a) many conflicts are due to slight numerical differences and (b) the reorientation eliminated many conflicts. Table 2.12 shows that a cost of the reconciliation comes as ties. Comparative results are shown in Table 2.12.

A careful mathematical description based on an example of refrigerants can be found in Brüggemann et al. (2011). Along with questions of combining attribute values in the Kosko-formula (see Van de Walle et al. 1995), the main challenge facing a researcher becomes how to defuzzify the result.

2.4 Weak Order Provided by Partial Order Approaches

In partial order theory, the linear extensions play a large role. A linear extension is a linear order which preserves the order relation of a poset (for details, see Trotter 1992). Scanning all possible linear extensions, objects will be located at different

Table 2.12 Comparative results for fuzzy and crisp approaches

Data matrix	*U*	Ties	Remarks
Initial data (Table 2.5)	Gets its maximum possible value as the poset is a complete antichain	No ties	Many incomparabilities may be caused by small numerical differences as can be deduced from Table 2.10
Oriented (Fig. 2.6)	A reduction from $U = 66$ (initial) to $U = 55$ (oriented)	No ties	The poset shows some structure There are three isolated vertices and a vertex (no 38) loosely connected with the set {25, 36, 37, 39, 48, 49, 51, 59}
Five-column refined (Fig. 2.10)	$U = 23$	No ties	Seven levels, hence at least one chain with seven elements, i.e., a subset of *X* which can be ranked without any additional information
Initial data fuzzy (Fig. 2.12)	$U = 48$	No ties	Five levels, several graph–theoretical components, which indicate peculiarities in the data
Five-column-refined fuzzy (Fig. 2.13)	$U = 4$	Many ties	A weak order can be deduced from *X*, however, the number of ties needs additional steps to break them

heights. Therefore the average height (hav) is a measure for the rank and is called the "average rank" (Rkav). The exact calculation is computationally difficult (NP-hard), however, certain approximations are derived (see Brüggemann and Carlsen 2011) and the most simple one, the "Local Partial Order Model 0" (LPOM0) is just based on the number of objects and the two coordinates also used in the FOU-plot.

Using the rather rough expedient of treating *U*(*x*) as a single super-object gives the approximation:

$$hav(x) = \left[O(x) + U(x)\right]\left[O(x) / \left\{O(x) + F(x)\right\}\right] + O(x)\left[F(x) / \left\{O(x) + F(x)\right\}\right]. \tag{2.10}$$

Therefore we use this extremely simple method to get an approximate measure regarding the weak order of the 12 objects as land-use regions, with results contained in Table 2.13. See also the chapter of Rocco and Tarantola (2014) for an application of the LPOM-concept. Interestingly, Eq. (2.10) is also obtainable as the mean value of a Binomial statistical model with parameters *n*−1 and [*O(x)*/{*O(x)*+F(x)}].

We find

36 < 51 < 59 ≅ 50 < 49 < 37 ≅ 60 < 38 < 26 < 39 < 48 < 25

The result of the exact method, i.e., counting the heights of any object in linear extensions, whereby the linear extensions are obtained as paths in a lattice of

Table 2.13 LPOM0 of F:/Pythonprogramme/PyHasse/pdt-files/Myersdecomp2_29.10.2012

Object	Rkav	$O(x)$	$U(x)$
25	12.0	12	0
26	9.1	7	3
36	1.182	1	2
37	5.2	4	3
38	7.8	6	3
39	9.75	6	5
48	10.833	10	1
49	3.9	3	3
50	3.25	2	5
51	1.444	1	4
59	3.25	1	9
60	5.2	2	8

Table 2.14 Average ranks (Rkav) according to method of Oliver Wienand

25:	12.0
26:	9.212
36:	1.538
37:	5.693
38:	7.63
39:	9.024
48:	10.793
49:	4.086
50:	4.419
51:	2.043
59:	5.397
60:	6.166

down-sets of the original poset (“lattice theoretical method,” K. De Loof et al. 2006, program code: O. Wienand) is displayed in Table 2.14.

This sequence is

36 < 51 < 49 < 50 < 59 < 37 < 60 < 38 < 39 < 26 < 48 < 25

With the exact Wienand method the ties within LPOM0 disappear, and there are some rank inversions. From a methodological point of view this shows that the focus on $|O(x)|$, $|F(x)|$, and $|U(x)|$ is not sufficient. However, it is not claimed that sequence obtained from purely order theoretical aspects should replace all rankings obtained from other conventional multi-criteria decision methods (MCDM),since additional knowledge such as weights of each indicator come into play. Nevertheless, information extracted from partial order theory extended to incorporation of indicator weights may be useful for discussion of MCDM-rankings (see Patil and Joshi 2014), where a concept of reconciliation plays a major role. The theoretical framework is explained in that chapter.

2.5 Weighted Composite Indicators

2.5.1 Weighting for Indicators and Linear Extensions

As indicated above, a generalization of the concept of linear extensions by including weights for any single linear extensions (see Patil and Taillie 2004) opens a wide variety of analysis tools. Tools based on this generalization are subsumed under the concept of comparative knowledge discovery (CKD). “Reconciliation of weights of indicators” as one such tool also belongs to CKD, where it is of interest how far empirical weights of indicators can be adjusted to be coincident with subjective indicator weights of stake holders (see Patil and Joshi 2014).

Here, however, discussion will be limited to less sophisticated tools; namely two relatively simple methods to compare partial order findings with results of formulating composite indicators determined as weighted combinations of basic indicators.

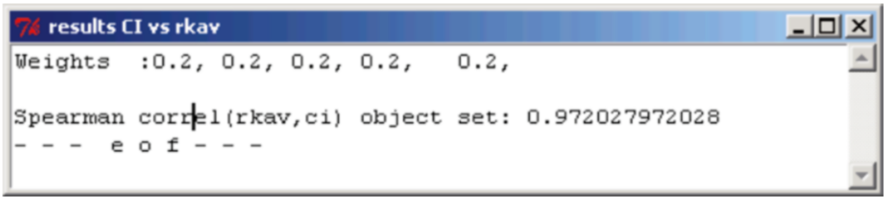
2.5.2 Weight Intervals

The basis of this method is discussed by Brüggemann et al. (2013b) and follows the idea that exact weights may not be known but can be treated as intervals. Instead of stating a certain weight value for an indicator such as Pct11, we argue that the weight may be in an interval [weight$_{\text{min}}$, weight$_{\text{max}}$]. After giving all five indicators a weight interval, a Monte Carlo simulation can be performed by picking a certain weight value out of each interval (renormalizing if necessary so that the sum is one) and calculating a composite indicator (CI).

Clearly there are many CIs possible; however, some may vary in their numerical values without yielding different rankings. The central idea is that the set of CIs derived from the weight intervals constitutes a new multi-indicator system (MIS), which may now also take into account (through the weights) knowledge of stakeholders, decision makers, etc. If the weights are exactly known then clearly only one CI is obtained and a linear or weak order results. In general, however, several candidate CIs are obtained which render different ranks and the question is as to how far common order relations can be found. Thus the new MIS, consisting of the set of CIs, corresponding to the weight intervals is used to get a new and in general enriched poset. The refined data matrix (rdm) of five important reoriented land cover indicators is used to illustrate the procedure by applying weight intervals.

2.5.2.1 Weight Values Being Known

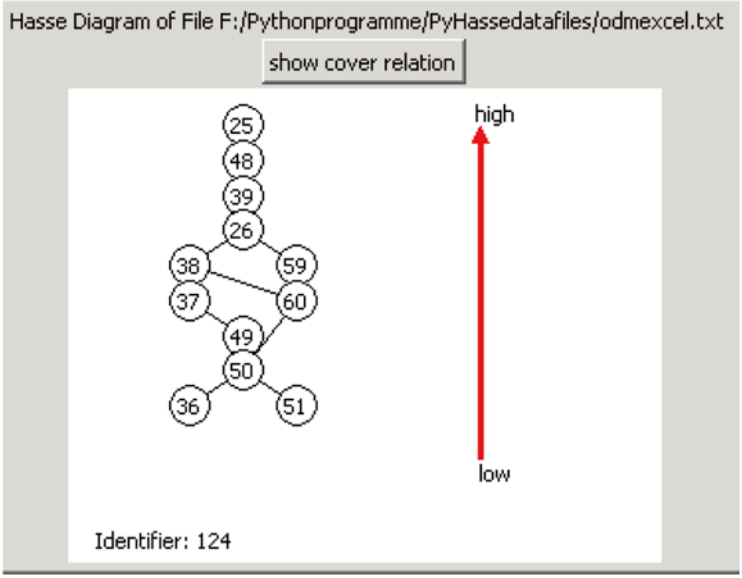
The PyHasse module (conflict7.py) is first applied to compare the average ranks, Rkav, (of the LPOM0-approximation) with the CI for equal weights. From the result given in Fig. 2.14 it is seen that CI and Rkav are not in conflict.



```
Weights  :0.2, 0.2, 0.2, 0.2,   0.2,

Spearman correl(rkav,ci) object set: 0.972027972028
- - -  e o f - - -
```

Fig. 2.14 Result of conflict-analysis based on equal weights of the CI and the LPOM0-approximation


Hasse Diagram of File F:/Pythonprogramme/PyHassedatafiles/odmexcel.txt

Identifier: 124

Fig. 2.15 Hasse diagram of CI's based on weight intervals [0.1, 0.3] for each of the five indicators

Returning to Fig. 2.10, the circles of that Hasse diagram contain the corresponding values of the CI(x) along with the object identifiers. Note that region 50 is covering 36, and there is an artificial overlap of the covering line with the circle representing object 49.

2.5.2.2 Weight Values as Intervals

Next, knowledge about weights is relaxed and an interval is assumed for each weight of [0.1, 0.3].

As the set of all CIs consistent with the weight intervals, the result shown in Fig. 2.15 is obtained whereby it is seen that even this narrow range of weight intervals leads to conflicts.

If the weight interval is further narrowed to [0.15, 0.25] for all five weights, a similar Hasse diagram is obtained with both objects 59 and 60 being quite "resistant" to change with regard to their incomparability.

2.5.3 Weight Evolution

Weighted composite indicators as well as a set of multiple indicators can be built up in a stepwise manner as follows:

Let Γ(*j*) be the composite indicator obtained after calculating the weighted sum for indicators 1,…,*j*.

$$\Gamma_j(x) = \sum_{k=1}^{j} g_k \cdot q_k(x). \tag{2.11}$$

Let |*O*(*x*,*j*)| be the number of elements in the down-set of *x*, taking the indicators $q_1,\ldots,q_j$ as multi-indicator system into account. Using a criterion of entry order such as impact on orderings, the sequence of indicators is fixed. Thus as the first indicator in (2.11) that indicator is selected having the largest, whereas as q_j the indicator with the relative least impact on the partial orders chosen. Then the evolution of the ordering is tracked in terms of *O*(*x*) and/or *F*(*x*) during the progression (*j* = 1,…,*m*). When this is done with the refined subset of land cover indicators, the following evolutionary behavior based on

$$\chi_j(x) = \frac{\Gamma_j(x)}{\left|O(x,j)\right|} \tag{2.12}$$

is observed, whereby we are differentiating pretty qualitatively between "smooth" behavior and "crucial" changes in χ:

Region-objects 25, 26, 36, 48 evolve smoothly.

Region-objects 37, 39 have a crucial step of marked change at entry of indicator Pct81.

Region-object 60 has a crucial step of marked change at entry of Pct32.

Region-objects 49, 50, 51 have crucial steps of marked change at entry of Pct81 and Pct82.

Region-object 59 has crucial steps of marked change for three indicators—Pct81, Pct82 and Pct32.

Pronounced shifts in patterns of progression are signposts of interesting interactions between objects and indicators. Critical steps usually indicate that a particular indicator has pronounced effects on certain objects. This is notably true in this excerpt from a larger dataset, since the Pct32 indicator is non-zero only for object 51. Thus, critical steps give clues to special subsets among the objects as seen through the lens of particular indicators.

2.6 Insights, Implications, and Inference

A multivariate dataset that is completely lacking in context should not be of much interest for any serious seeking of sequences. Formal statement of context for a dataset comes as metadata, and the metadata should be fully exploited in assessing prospects as a multi-indicator system (MIS). However, it is seldom that a multivariate dataset collected for some collateral or parallel purpose such as monitoring will serve well for direct use as an MIS. It will generally be necessary to engage in exploratory extraction of latent information on joint order that may reside in the dataset and that has been the topic of current consideration. There is a parallel with the subject of image enhancement wherein the purpose is to strengthen and bring to the foreground the imaging information that resides in digital image data, so it is hereby proposed to consider this undertaking as Joint Order Enhancement (JOE).

It has been repeatedly shown here that incomparability under partial order theory constitutes a useful criterion for pursuit of joint order enhancement, starting with joint orientation. A different scoping for the tally of incomparability enables assessment of joint effect attributable to each candidate indicator, as being either relatively minor or more major. Selecting the subset of (oriented) major ones will give a refined matrix of candidates for further exploration. That exploration can include probing groups of indicators taking into account the signals provided by the rank correlation matrix. It is only reasonable to couple statistical signals with protocols of partial order theory throughout the JOE endeavor.

It is an expedient aspect of exploration to have the capacity for examining order relations in overview, which the conventional Hasse diagram technique does not provide. Posetic features of "down-set" and "up-set" can be coupled with incomparability to provide an "FOU" space for exploring broader aspects of ordering, beginning with an FOU-plot. Objects positioned high on the ordinate exhibit considerable incomparability. Objects at the lower right are definitely prominent prospects in the order endeavor, and objects at the lower left are definitely subordinate. Thus, "marker" objects can be identified in the respective positions to serve as comparators for other objects that are less distinctive. The identifier of an object might be prefaced by one of the symbols ˥ ˦ ˧ ˨ ˩ to signify its incomparability and have as a trailing (subscript) one of the symbols →↔← to signify whether or not it has favorability for the balance of $O(x)$ and $F(x)$. With reference to the highlighted objects in Fig. 2.11, the symbology would be: ˥59$_{\leftrightarrow}$, ˩25→, ˩36$_{\leftarrow}$. The coordinates of an FOU-plot also provide opportunity for progressive hierarchical clustering to suggest posetic proximity groups of objects.

Generalization of ordering can be induced by directly generalizing the scales of individual attributes, but considerable caution is warranted in this regard, and it should be done with full cognizance of any metadata that is available. The simplest version for multi-digit values (other than ranks) is rounding or truncation to effect generalization by an order of magnitude. This is particularly appropriate in circumstances where intermediate computer processing has induced additional digits. Binning is intentional introduction of equivalence classes that may occasionally perturb transitivity for objects in terms of the raw data and is essentially declaration that the original scaling is inappropriate. Rank data can be binned, but it is then incumbent to rank the bins as the values to put forward. Commensurate generalization for different attributes is problematic for this approach, so each generalization of an individual attribute should be evaluated before proceeding to generalize another attribute.

Fuzzy concepts provide a different genre of generalization that entails greater sophistication and complexity. An experiment on actual data shows herein that a fuzzy approach can be informative with regard to joint order enhancement in a manner that is not readily accomplished otherwise.

Formulating and evaluating composite indicators obtained as weighted summations of original attributes provide a flexible and fertile avenue of joint order enhancement. The flexibility of weighting also allows exogenous information to be incorporated in the evaluation of order relations. Venturing further along such directions leads into the domain of comparative knowledge discovery.

It is all too easy when delving into order relations to treat a dataset as fully encompassing a context. If this were generally the situation, then there would have been no incentive to intellectualize regarding inferential statistics as an extension of descriptive statistics. Particularly when working with opportunistically acquired data, it is not usually a tenable assumption that a context is fully embedded within a dataset. This is patently not true for many settings of environmental sampling such as water, air, and soil. Replicate sampling quickly makes evident the almost instantaneous variability in time and space. To treat an order extracted from a set of sample data as being immutable truth is problematic at best and more often fundamentally deceptive. If replicate observations are not available by which to explore order variation in terms of sampling variation, then some recourse to distributional modeling of perturbations in the data is needed. Trying to cling tenaciously to every distinction of order that arises from nth digits is simply misguided.

As a minimum with regard to exploration of joint order enhancement, we would recommend:

(a) Investigating evidence of conflicting orientation.
(b) Investigating evidence of relative redundancy among indicators for consideration of culling.
(c) Using local partial order models as approximations for computing average height.
(d) Depict general features of comparability using posetic graphics such as FOU plots.

More sophisticated aspects can then be pursued as deemed appropriate, such as fuzzy analysis and weighted composite indicators.

References

Al-Sharrah G (2014) Ranking hazardous chemicals with a heuristic approach to reduce isolated objects in Hasse diagrams. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 10. Springer, New York, NY

Annoni P, Brüggemann R (2008) The dualistic approach of FCA: a further insight into Ontario Lake sediments. Chemosphere 70:2025–2031

Bartel HG, Mucha J (2014) Measures of incomparability and of inequality and their applications. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 3. Springer, New York, NY

Bock HH (1974) Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen

Bock HH (1979) Clusteranalyse mit unscharfen Partitionen. In: Bock HH (ed) Studien zur Klassifikation, Bd 6 Klassifikation und Erkenntnis III - Numerische Klassifikation. Gesellschaft für Klassifikation, Frankfurt, pp 137–163

Bock HH, Gaul W, Schader M (2003) Studies in classification, data analysis, and knowledge organization. Springer, Berlin

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partially orders. Match Commun Math Comput Chem 65:383–414

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems – introduction to partial order applications. Springer, New York, NY

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Brüggemann R, Welzl G (2002) Order Theory Meets Statistics -Hassediagram technique-. In: Voigt K, Welzl G (eds) Order theoretical tools in environmental sciences – order theory (Hasse diagram technique) meets multivariate statistics. Shaker-Verlag, Aachen, pp 9–39

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (2001) Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. J Chem Inf Comp Sci 41:918–925

Brüggemann R, Kerber A, Restrepo G (2011) Ranking objects using fuzzy orders, with an application to refrigerants. Match Commun Math Comput Chem 66:581–603

Brüggemann R, Mucha HJ, Bartel HG (2013a) Ranking of polluted regions in South West Germany based on a multi-indicator system. Match Commun Math Comput Chem 69(2):433–462

Brüggemann R, Restrepo G, Voigt K, Annoni P (2013b) Weighting intervals and ranking, exemplified by leaching potential of pesticides. Match Commun Math Comput Chem 69(2):413–432

Brüggemann R, Carlsen L, Voigt K, Wieland R (2013c) PyHasse software for partial order analysis: scientific background and description of some modules. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 19. Springer, New York, NY

Carlsen L, Brüggemann R (2014) Indicator analyses, what is important – and for what? In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 18. Springer, New York, NY

De Loof K, De Meyer H, De Baets B (2006) Exploiting the lattice of ideals representation of a poset. Fundam Inform 71:309–321

Diday E (1979) Problems of clustering and recent advances. In: Bock HH (ed) Klassifikation und Erkenntnis III. Gesellschaft für Klassifikation, Frankfurt, pp 3–16

Luther B, Brüggemann R, Pudenz S (2000) An approach to combine cluster analysis with order theoretical tools in problems of environmental pollution. Match 42:119–143

Myers WL, Patil GP (2008) Semi-subordination sequences in multi-measure prioritization problems. In: Todeschini R, Pavan M (eds) Data handling in science and technology, vol 27. Elsevier, New York, NY, pp 161–170

Myers WL, Patil GP (2010) Preliminary prioritization based on partial order theory and R software for compositional complexes in landscape ecology, with applications to restoration, remediation, and enhancement. Environ Ecol Stat 17:411–436

Myers WL, Patil GP (2012) Multivariate methods of representing relations in R for prioritization purposes: selective scaling, comparative clustering, collective criteria and sequenced sets. Springer, New York, NY

Myers WL, Patil GP (2013a) Coordination of contrariety and ambiguity in comparative compositional contexts: balance of normalized definitive status in multi-indicator systems. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 8. Springer, New York, NY

Myers WL, Patil GP (2013b) Higher-order indicator with rank-related clustering in multi-indicator systems. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 16. Springer, New York, NY

Myers WL, Patil GP (2013c) Statistical geoinformatics for human environment interface. CRC/Taylor & Francis, Boca Raton, FL

Myers WL, Patil GP, Cai Y (2006) Exploring patterns of habitat diversity across landscapes using partial ordering. In: Brüggemann R, Carlsen L (eds) Partial order in environmental sciences and chemistry. Springer, Berlin, pp 309–325

Patil GP (2012) Keynote lecture on partial orders and composite indicators for multivariate ranking in multivariate nonparametric statistics at the international workshop on partial order theory and modeling held in Berlin, Germany

Patil GP, Joshi S (2014) Comparative knowledge discovery with partial order and composite indicator. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 6. Springer, New York, NY

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

Rocco C, Tarantola S (2014) Evaluating ranking robustness in multi-indicator uncertain matrices: an application based on simulation and global sensitivity analysis. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 13. Springer, New York, NY

Trotter WT (1992) Combinatorics and partially ordered sets dimension theory. The Johns Hopkins University Press, Baltimore, MD

Van de Walle B, De Baets B, Kersebaum KC (1995) Fuzzy multi-criteria analysis of cutting techniques in a nuclear dismantling project. Fuzzy Set Syst 74:115–126

Chapter 3
Measures of Incomparability and of Inequality and Their Applications

Hans-Georg Bartel and Hans-Joachim Mucha

Abstract Usually, there are only two stages of comparability between two objects: they are comparable or incomparable (see, for instance, the theory of partially ordered sets). The same holds with respect to equality/inequality. In this publication, measures of incomparability u_{ij} and of inequality v_{ij} between two objects g_i and g_j with m attributes with respect to the relation $\leq$ are introduced. Based on these definitions the (non-metric) distance measure $a_{ij} = \frac{1}{2}\left(u_{ij} + v_{ij}\right)$ with maximal possible values $m+1+\left[\frac{m}{2}\right]\cdot\left(m-\left[\frac{m}{2}\right]\right)$ is proposed. The distance matrix $\mathbf{A} = (a_{ij})$ will be used for clustering starting from the corresponding complete graph $\langle g \rangle$ (g – number of objects), whose edges g_i–g_j are valued by a_{ij}. The result of the classification consists of a set of complete subgraphs, where, for instance, the objective function of compactness of a cluster is based on all pairwise distances of its members. The same edge-valued graph is used to construct a transitive-directed tournament. Thus, a unique seriation of the objects can be obtained which can also be used for further interpretation of the data. For illustrative purposes, an application to environmental chemistry with only a small data set is considered.

3.1 Introduction and Aim

In *Webster's Encyclopedic Dictionary*, the following definition can be found:

> 'incomparable […] *adj.* 1 : *having no equal (as in quality or worth)* : *matchless*
> 2 : *not suitable for comparison*'

H.-G. Bartel (✉)
Department of Chemistry, Humboldt University Berlin,
Brook-Taylor-Straße 2, 12489 Berlin, Germany
e-mail: hg.bartel@yahoo.de

H.-J. Mucha
Weierstrass Institute of Applied Analysis and Stochastics,
Mohrenstraße 39, 10117 Berlin, Germany
e-mail: mucha@wias-berlin.de

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_3, © Springer Science+Business Media New York 2014

(Webster 1994). As can be shown with examples, both comparative and superlative of this adjective exist, if it has the first meaning. The categories of forms produced by comparison of adjectives and adverbs are called degrees of comparison. Obviously, these two categories are produced by comparison. Therefore, they are called the degrees of comparison in grammar. This means that there are cases where incomparability is not incomparable – at in linguistic context.

Some examples of the use of the degrees of comparison of the English adjective 'incomparable' and of the corresponding Latin one '*incomparabilis*, *-le*', from which 'incomparable' was etymologically directly derived, are listed here. For the positive form, two phrases will be quoted, one from William Shakespeare's (1564–1616) poem *The Rape of Lucrece* (1594): '*Collatinus extolled the incomparable chastity of his wife Lucretia*'. (Shakespeare 1594), and a Latin one from the Vulgate Bible: '[…] *et ideo Dominus hanc in illam pulchritudinem ampliavit ut inconparabili* [sic] *decore omnium oculis appareret.* ([…] *and therefore the Lord increased this her beauty, so that she appeared to all men's eyes incomparably lovely.*)' (Judith 10, 4). The comparative form is found in Oscar Wilde's (1854–1900) letter *De Profundis* and the homiletic commentary *De Isaac vel anima* by Saint Ambrose of Milan (339–397): '*I know of nothing in all drama more incomparable from the point of view of art* […] *than Shakespeare's drawing of Rosencrantz and Guildenstern*'. (Wilde 1911) and '*Et quanto incomparabilior est illa divinitatis Gloria* […]*?* (*And how much more incomparable is that glory of divinity?*)' (Ambrosius 1897: VIII, 78). As examples for the superlative form, uses of the epithet 'most incomparable' in connection with persons can be mentioned. So Henry King (1592–1669) wrote *An Elegy Upon the Most Incomparable K. Charles the I* (1649), and a Roman grave inscription is dedicated '*fratri inconparabilissimo* [sic] (*to the most incomparable Friar*)' (CIL VI 1886: 15947).

The statement about a possible graduation of incomparability obtained in linguistic-grammatical field motivated the proposition of a mathematical measure of incomparability. Of course, this only makes sense if there are conditions for which the relation 'is incomparable' corresponds to the meaning (1) of the adjective 'incomparable'.

The above-mentioned meaning (2) of the property 'incomparable' is however to a greater degree related to the incomparability relation, which is commonly used in mathematics and can be defined as follows:

(D1) If (S, π) is a partially ordered set or poset (π is a partial order), then any elements $s, t \in S$ are called **incomparable** (denoted by $s \| t$) if neither $s\pi t$ nor $t\pi s$ is true.

Indeed, the elements s and t are in that case '*not suitable for comparison*' or – more accurately – cannot be compared. Therefore, the question of whether s and t are comparable or incomparable leads here to a yes/no-response, i.e. there are no degrees of comparability or incomparability except of 1 and 0.

The poset $(\mathbb{N}, |)$ where $\mathbb{N}$ is the set of integers and $|$ = 'is a divisor of' is an example to illustrate this. As both $3|5$ and $5|3$ are false, the integers 3 and 5 are (absolutely) incomparable. On the other hand, the integers 3 and 6 are (totally) comparable due to the validity of $3|6$.

Table 3.1 Comparison of 5-tuples

Comparison			Number			
of			$q^<$	$q^>$		
$\mathbf{x}_1$	=	(1, 2, 3, 4, 5)	of cases			
with			$x_{1h} < x_{jh}$	$x_{1h} > x_{jh}$	min($q^<$, $q^>$)	Incomparability
$\mathbf{x}_2$	=	(2, 3, 4, 5, 6)	5	0	0	No
$\mathbf{x}_3$	=	(3, 4, 5, 6, 2)	4	1	1	Yes
$\mathbf{x}_4$	=	(4, 5, 6, 2, 3)	3	2	2	Yes
$\mathbf{x}_5$	=	(5, 6, 2, 3, 4)	2	3	2	Yes
$\mathbf{x}_6$	=	(6, 1, 2, 3, 4)	1	4	1	Yes
$\mathbf{x}_7$	=	(0, 1, 2, 3, 4)	0	5	0	No

A poset (X, $\leq$) is used to investigate the possibility of indicating the degrees of incomparability between elements of a set X. Here, $\leq$ is the ordinary $\leq$-relation between numbers. The elements $\mathbf{x}_i$ of X are ordered m-tuples of (real) numbers:

$$\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{im}\right) \in X,\ i = 1, 2, \ldots, |X|.$$

Equality of $\mathbf{x}_i$ and $\mathbf{x}_j$ (denoted by $\mathbf{x}_i = \mathbf{x}_j$) is fulfilled, if all elements of $\mathbf{x}_i$ and $\mathbf{x}_j$ are pairwise equal: $(x_{ih} \in \mathbf{x}_i) = (x_{jh} \in \mathbf{x}_j)$ for $h = 1, 2, \ldots, m$.

In the following, it is assumed that there are no two $\mathbf{x}_i$, $\mathbf{x}_j \in X$ with $\mathbf{x}_i = \mathbf{x}_j$ if $i \neq j$. Therefore, each pair $\mathbf{x}_i \neq \mathbf{x}_j$ with $i \neq j$ must differ at least in one of its elements:

$$\text{(A1)} \quad \forall \mathbf{x}_i, \mathbf{x}_j \in X, i \neq j:\ \mathbf{x}_i \neq \mathbf{x}_j \Leftrightarrow \exists \left(x_{ih} \in \mathbf{x}_i\right) \neq \left(x_{jh} \in \mathbf{x}_j\right), h \in \{1, 2, \ldots, m\}$$

This way it is guaranteed that from $\mathbf{x}_i = \mathbf{x}_j$ always follows $i = j$ and *vice versa*. However, it is possible that an element is a representative of a class of equivalent (pairwise equal, but by other way distinguishable) elements.

In the context of order theory, following definition is commonly accepted:

(D2) Let be the ordered m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$ elements of the poset (X, $\leq$). If $x_{ih} \leq x_{jh}$ ($x_{jh} \leq x_{ih}$) is true for all pairs $\{x_{ih} \in \mathbf{x}_i, x_{jh} \in \mathbf{x}_j\}$ ($h = 1, 2, \ldots, m$) then $\mathbf{x}_i \leq \mathbf{x}_j$ ($\mathbf{x}_j \leq \mathbf{x}_i$) holds. Then $\mathbf{x}_i$ and $\mathbf{x}_j$ are called 'comparable' (denoted by $\mathbf{x}_i \perp \mathbf{x}_j$). Otherwise, they are called 'incomparable' (denoted by $\mathbf{x}_i \| \mathbf{x}_j$).

Obviously, the definition (D2) is a special case of (D1). Therefore again, the question whether two m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$ are incomparable or not, can only be answered with yes or no. Concerning incomparability, only two states exist: '1 ≡ incomparable/not comparable' or '0 ≡ not incomparable/comparable'.

The comparison of some 5-tuples in Table 3.1 shows a graduation of incomparability clearly demonstrated by the min($q^<$, $q^>$) values. For example, it can be said that $\mathbf{x}_4$ is more incomparably with $\mathbf{x}_1$ than $\mathbf{x}_3$.

The value min($q^<$, $q^>$) = 0 indicates no incomparability (i.e. comparability). Obviously, maximal incomparability is reached if the value 2 is present. The value 1 makes clear that a degree of incomparability exists between its minimum and maximum.

Later we are going to propose and discuss a measure of incomparability and other related measures. First we want to point the reader to the following.

In analogy to (D2) it is possible to define (see the above statement about equality of $\mathbf{x}_i$ and $\mathbf{x}_j$):

(D3) Let be the ordered m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$ elements of the poset $(X, \leq)$. If $x_{ih}=x_{jh}$ is true for all pairs $\{x_{ih}\in\mathbf{x}_i, x_{jh}\in\mathbf{x}_j\}$ $(h=1,2,\ldots,m)$ then $\mathbf{x}_i = \mathbf{x}_j$ holds. Then $\mathbf{x}_i$ and $\mathbf{x}_j$ are called ‘equal’. Otherwise, they are called ‘unequal’ (denoted by $\mathbf{x}_i \neq \mathbf{x}_j$).

In the case of ordered m-tuples, a graduation of inequality can also be specified. Obviously, the degree of inequality increases in the order $\mathbf{x}_8 = (1, 2, 3, 4, 6)$, $\mathbf{x}_9 = (1, 2, 3, 5, 4)$ to $\mathbf{x}_{10} = (4, 5, 1, 2, 3)$ when compared with $\mathbf{x}_1 = (1, 2, 3, 4, 5)$.

Furthermore, it is important to note that comparability $\perp$, incomparability $\|$, and inequality $\neq$ are symmetric relations (of course, only the first one is reflexive):

If (S, σ), $\sigma = \perp, \|$, or $\neq$, $s, t \in S$ then $s\sigma t \Leftrightarrow t\sigma s$.

Therefore, the measures assigned to those three relations must be symmetric.

Let us consider a set $G=\{g_1,g_2,\ldots,g_g\}$ of $g = |G|$ objects and let us recall that a set $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_g\}$ of $|X| = g$ m-tuples is associated with a data matrix $\mathbf{X}(G, M)$ in the following way (M is a set of $m = |M|$ attributes):

$$\mathbf{X}(G,M)=\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{g1} & \cdots & x_{gj} & \cdots & x_{gm} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_g \end{pmatrix}, \tag{3.1}$$

i.e. the g m-tuples are the rows of the matrix $\mathbf{X}(G, M)$.

The set X is associated with the set of objects G: $\mathbf{x}_i \leftrightarrow g_i$ $(i = 1,2,\ldots,g)$. This relationship enables a connection with data analysis.

Note that Brüggemann and Voigt (2012) calculate degrees of incomparabilities based on a decomposition of an antichain into pairs of objects and pairs of pairwise different columns of matrix (3.1).

3.2 About Three Suggested Measures of Inequality and Incomparability

3.2.1 Some Important Numbers

Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be ordered m-tuples whose elements are pairwise compared: $(x_{ih}\in\mathbf{x}_i)\pi_\tau(x_{jh}\in\mathbf{x}_j)$ $(h=1,2,\ldots,m)$, where π_τ denotes the result of the comparison. Because of trichotomy of real numbers $(\pi_\tau\in\{=,<,>\})$, this result is either $x_{ih}=x_{jh}$, $x_{ih}<x_{jh}$, or $x_{ih}>x_{jh}$.

Therefore, three numbers can be determined with respect to the comparison of $\mathbf{x}_i$ with $\mathbf{x}_j$:

- $q_{ij}^{=}$, the number of the results of the comparison, in which $x_{ih}=x_{jh}$
- $q_{ij}^{<}$, the number of the results of the comparison, in which $x_{ih}<x_{jh}$
- $q_{ij}^{>}$, the number of the results of the comparison, in which $x_{ih}>x_{jh}$

Of course, the following relations hold for the resulting numbers $q_{ji}^{=}$, $q_{ji}^{<}$, and $q_{ji}^{>}$, which correspond to the comparison of $\mathbf{x}_j$ with $\mathbf{x}_i$:

$$q_{ji}^{=} = q_{ij}^{=}, q_{ji}^{<} = q_{ij}^{>}, q_{ji}^{>} = q_{ij}^{<}, \tag{3.2}$$

from which and from assumption (A1) for $i=j$ follows:

$$q_{ii}^{=} = m, q_{ii}^{<} = q_{ii}^{>} = 0. \tag{3.3}$$

The sum of the three numbers is

$$q_{ij}^{=} + q_{ij}^{<} + q_{ij}^{>} = m. \tag{3.4}$$

If it is clear which comparison is considered, these three numbers can be denoted by $q^{=}$, $q^{<}$, and $q^{>}$, respectively. So, for instance, it can be written in general $q^{=}+q^{<}+q^{>}=m$.

3.2.2 The Three Measures

Before introducing the three distance measures, which correspond to incomparability and inequality, the well-known definition of the distance function is repeated as a reminder (Späth 1980).

(D4) A **distance function** d on a given set S is a function d: $S\times S \rightarrow \mathbb{R}$ ($\mathbb{R}$ is the set of real numbers) that fulfils the following conditions (r, s, $t \in S$):

(a) A number $d_0 \in \mathbb{R}$ exists, so that for all s, $t \in S$ the relation $d(s,t) \geq d_0$ holds. (Usually, $d_0 = 0$ is used.)
(b) It is $d(s,s) = d_0$ for all $s \in S$.
(c) The distance function d is symmetric: $d(s,t) = d(t,s)$ for all s, $t \in S$.

If, in addition, the following two conditions are met, the distance function is known as a metric.

(d) If $d(s,t) = d_0$ then $s = t$.
(e) The triangle inequality $d(s,t) \leq d(s,r) + d(r,t)$ is fulfilled for all r, s, $t \in S$.

Now let us introduce the three measures mentioned.

Table 3.2 Cases of complete incomparability $u_{ij}=1$

m	$[m/2]$	$q_{ij}^{=}$	$q_{ij}^{<}$	$q_{ij}^{>}$
Even	$m/2$	0	$m/2$	$m/2$
Odd	$(m-1)/2$	0	$(m-1)/2$	$(m+1)/2$
			$(m+1)/2$	$(m-1)/2$
		1	$(m-1)/2$	$(m-1)/2$

(A) Measure of Incomparability $u_{ij}=u(\mathbf{x}_i,\mathbf{x}_j)$ between the m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$
The values of u_{ij} are calculated according to the equation

$$u_{ij}=\min\left(q_{ij}^{<},q_{ij}^{>}\right)\cdot\left[\frac{m}{2}\right]^{-1} \tag{3.5}$$

This measure is normalized: $0\leq u_{ij}\leq 1$, as $\left[\frac{m}{2}\right]$ is the maximum value of min $(q_{ij}^{<},q_{ij}^{>})$.

The measure u_{ij} is a non-metric distance function: (D4a): Of course, the minimal value of $\min(q_{ij}^{<},q_{ij}^{>})$ is 0 (zero), so that $u_{ij}\geq 0$. (D4b): For $i=j$, $q_{ii}^{<}=q_{ii}^{>}=0$ (3.3) holds, so that $u_{ii}=0$. (D4c): Because of (3.2), the measure is symmetric: $u_{ij}=u_{ji}$. (D4d): This condition is not met, since it is sufficient that only one of the two numbers $q_{ij}^{<}$ or $q_{ij}^{>}$ is equal to 0, so that $\min(q_{ij}^{<},q_{ij}^{>})$ is zero. Therefore, it is possible that $u_{ij}=0$, although $\mathbf{x}_i$ and $\mathbf{x}_j$ are different ($i\neq j$). (D4e): The triangle inequality $u_{ik}\leq u_{ij}+u_{jk}$ is not fulfilled in every case. So it is possible, for example, that $u_{ij}=u_{jk}=0$, but $u_{ik}>0$, although all the three m-tuples $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ are pairwise distinct from each other (e.g., $\mathbf{x}_i=(1, 2, 3)$, $\mathbf{x}_j=(1, 2, 4)$, $\mathbf{x}_k=(1, 1, 4)$, where $u_{ij}=u_{jk}=0$, $u_{ik}=1$).

The equation $u_{ij}=0$ ($i\neq j$) means that $\mathbf{x}_i$ and $\mathbf{x}_j$ are comparable ($\mathbf{x}_{\mathrm{i}}\perp\mathbf{x}_{\mathrm{j}}$). If $u_{ij}=1$ then $\mathbf{x}_i$ and $\mathbf{x}_j$ are called totally incomparable ($\mathbf{x}_i\|_{\mathrm{tot}}\mathbf{x}_j$). This complete incomparability may occur in the cases mentioned in Table 3.2.

The definition (D1), that is usually used for incomparability of $\mathbf{x}_i$ and $\mathbf{x}_j$ ($\mathbf{x}_i\|\mathbf{x}_j$), is now given by $u_{ij}>0$. Here, it is quite obvious that the values u_{ij} express different degrees of incomparability: $[m/2]^{-1}\leq u_{ij}\leq 1$. The smallest value $[m/2]^{-1}$ of $u_{ij}>0$ is obtained if

$$\left.\begin{matrix} q_{ij}^{<} \\ q_{ij}^{>} \end{matrix}\right\}=1,\quad \left\{\begin{matrix} q_{ij}^{>} \\ q_{ij}^{<} \end{matrix}\right\}+q_{ij}^{=}=m-1,\quad \left\{\begin{matrix} q_{ij}^{>} \\ q_{ij}^{<} \end{matrix}\right\}>0.$$

The difference $c_{ij}=1-u_{ij}$ can be regarded as a measure of comparability between $\mathbf{x}_i$ and $\mathbf{x}_j$. It is a non-metric similarity function. $c_{ij}=1$ means complete comparability, while $c_{ij}=0$ means incomparability. The values $0<c_{ij}\leq 1$ are the degrees of comparability.

Let $\mathbf{C}(\mathbf{X}(G,M))=(c_{ij})$ be the matrix of comparability based on the data matrix $\mathbf{X}(G, M)$ given in (3.1). X is the set of $g=|G|$ m-tuples ($m=|M|$): $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots\mathbf{x}_g\}$. A modified form $\mathbf{C}'$ of this matrix $\mathbf{C}$ is related to the adjacency matrix $\alpha(X,\perp)=(\alpha_{ij})$

($\perp$ is the relation of comparability), which in turn can be calculated using the so-called ζ-matrix $\zeta(X,\leq) = (\zeta_{ij})$. (Regarding the last two matrices and the relation between them, see for instance (Bartel and Mucha 2011).) If the poset $(\{\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots\mathbf{x}_g\},\leq)=(X,\leq)$ is given, the ζ-matrix is calculated by: $\zeta_{ij}\begin{cases}=1 & \text{if } \mathbf{x}_i \leq \mathbf{x}_j\\ =0 & \text{otherwise}\end{cases}$. The relationship between the adjacency matrix and the ζ-matrix is

$$\boldsymbol{\alpha}(X,\perp)=\boldsymbol{\zeta}(X,\leq)+\boldsymbol{\zeta}^{\mathrm{T}}(X,\leq),$$

where $\boldsymbol{\zeta}^{\mathrm{T}}(X,\leq)$ is the transposed matrix of the matrix $\boldsymbol{\zeta}(X,\leq)$.

Using the elements of the matrix **C**, the elements of the matrix **C'** are determined as follows: $c'_{ij}\begin{cases}=1 & \text{if } c_{ij}=1\\ =0 & \text{if } c_{ij}<1\end{cases}$. Obviously, this leads to:

$$\mathbf{C}'\big(\mathbf{X}(G,M)\big)=\boldsymbol{\alpha}(X,\perp)-\mathbf{I}=\boldsymbol{\zeta}(X,\leq)+\boldsymbol{\zeta}^{\mathrm{T}}(X,\leq)-\mathbf{I},$$

where $\mathbf{I}=(\delta_{ij})$ is the unit matrix (δ_{ij}: Kronecker's delta). Conversely, the equation

$$\mathbf{C}\big(\mathbf{X}(G,M)\big)+\mathbf{I}=\boldsymbol{\alpha}'(X,\perp) \tag{3.6}$$

leads to a modified adjacency matrix $\boldsymbol{\alpha}'(X,\perp)$, which indicates the degree of comparability between the objects. The application of equation (3.6) will be published in a follow-up publication.

It has also to be noted that $\mathbf{C} = \mathbf{C}'$ holds in the case of $m = 3$.

(B) Measure of Inequality $v_{ij}=v(\mathbf{x}_i,\mathbf{x}_j)$ between the m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$

The definition of this measure is proposed as:

$$v_{ij}=\frac{m-q_{ij}^{=}}{m}=\frac{q_{ij}^{<}+q_{ij}^{>}}{m}. \tag{3.7}$$

The measure v_{ij} is a metric distance function: (D4a): The minimal value of $m-q_{ij}^{=}$ is 0 (if $q_{ij}^{=}=m$), so that $v_{ij}\geq 0$. (D4b): Obviously, for $i = j$, $q_{ij}^{=}=m$ holds (see (D3)), so that $v_{ii}=0$. (D4c): Because of (3.2), the measure is symmetric: $v_{ij}=v_{ji}$. (D4d): Because of (D3), (A1), and the validity of $v_{ii}=0$ (see above), $(v_{ij}=0)\Leftrightarrow(i=j)$ is true. (D4e): As can be shown easily, the triangle inequality $v_{ik}\leq v_{ij}+v_{jk}$ results after a few changes in the condition $m+q_{ik}^{=}\geq q_{ij}^{=}+q_{jk}^{=}$ for its validity. This latter condition is always true which can be shown easily and is, therefore, omitted here.

Since the maximum value of $m-q_{ij}^{=}$ is equal to m, the measure v_{ij} given in (3.7) is normalized: $0\leq v_{ij}\leq 1$. The value $v_{ij}=1$ ($q_{ij}^{=}=0$, $q_{ij}^{<}+q_{ij}^{>}=m$) means that $\mathbf{x}_i$ and $\mathbf{x}_j$ are totally unequal. (C) Distance of Combined Inequality and Incomparability $a_{ij}=a(\mathbf{x}_i,\mathbf{x}_j)$ between the m-tuples $\mathbf{x}_i$ and $\mathbf{x}_j$ This measure is a combination of the measures u_{ij} (3.5) and v_{ij} (3.7):

$$a_{ij}=\frac{1}{2}\left(v_{ij}+u_{ij}\right). \tag{3.8}$$

Table 3.3 The three measures introduced and their corresponding relations

	Measure of	Corresponding pair of relationships for the determination of the distance measurement		Type of corresponding relation
(A)	Incomparability	$<$	$>$	Strict order
(B)	Inequality	$\neq$	$=$	Equivalence
(C)	Incomparability/inequality	$\leq$	$>$	Partial order

The factor ½ in (3.8) is due to the fact that both u_{ij} and v_{ij} have the maximum value of 1. So, a_{ij} has the same maximum value and is normalized: $0 \leq a_{ij} \leq 1$.

The motivation for establishing this combined measure (3.8) can be derived from Table 3.3. There the three proposed measures and the underlying relations are mentioned.

So, the measure a_{ij} that is composed of the measures of incomparability and of inequality corresponds to the relation ≤. The latter is of fundamental significance for the theory of partial order.

The measure a_{ij} is also a non-metric distance function: (D4a): Because the minimum values of u_{ij} and v_{ij} are each zero, this is also true for a_{ij}: $a_{ij} \geq 0$. (D4b): Because of $u_{ii}=0$ and $v_{ii}=0$, $a_{ii}=0$ holds. (D4c): Because of the symmetry of both u_{ij} and v_{ij}, the measure a_{ij} is symmetric: $a_{ij}=a_{ji}$. (D4d): $(a_{ij}=0) \Leftrightarrow (i=j)$ follows from the relationships $(v_{ij}=0) \Leftrightarrow (i=j)$, $u_{ii}=0$, and $a_{ii}=0$ (see above).

(D4e): An example will be help to demonstrate that the triangle inequality $a_{ik} \leq a_{ij} + a_{jk}$ is not fulfilled in every case: If $m \geq 3$ and

$q_{ik}^{<} = q_{ik}^{>}$	$=$	1			
$q_{ij}^{<}$(or $q_{ij}^{>}$)	$=$	1	$q_{ij}^{>}$(or $q_{ij}^{<}$)	$=$	0
$q_{jk}^{<}$(or $q_{jk}^{>}$)	$=$	1	$q_{jk}^{>}$(or $q_{jk}^{<}$)	$=$	0

then $a_{ij} = a_{jk} = \frac{1}{2m}$ and $a_{ik} = \frac{1}{m} + \frac{1}{2}\left[\frac{m}{2}\right]^{-1}$, so that $a_{ik} = \frac{1}{m} + \frac{1}{2}\left[\frac{m}{2}\right]^{-1} > \frac{1}{m} = a_{ij} + a_{jk}$. Hereafter, an example with $m = 3$ is shown. With the 3-tuples $\mathbf{x}_i = (b-1, b, b+2)$, $\mathbf{x}_j = (b-1, b, b+1)$, and $\mathbf{x}_k = (b-1, b+1, b+1)$, $q_{ik}^{<} = q_{jk}^{<} = 1$, $q_{ij}^{<} = 0$, $q_{ik}^{>} = q_{ij}^{>} = 1$, $q_{ji}^{>} = q_{jk}^{>} = 0$ are obtained and from that $a_{ij} = a_{jk} = \frac{1}{6}, a_{ik} = \frac{5}{6}$, and $a_{ik} = \frac{5}{6} > \frac{1}{3} = a_{ij} + a_{jk}$.

The number $N^{(a)}(m)$ of possible states for the distance $a_{ij}(m, q_{ij}^{<}, q_{ij}^{>})$ is given by the following formula

$$N^{(a)}(m) = \sum_{i=0}^{\left[\frac{m}{2}\right]} \left(m - (2i-1)\right) = m + 1 + \left[\frac{m}{2}\right]\left(m - \left[\frac{m}{2}\right]\right). \tag{3.9}$$

Table 3.4 gives some examples of $N^{(a)}(m)$.

Two cases for m can be distinguished:

Table 3.4 Number $N^{(a)}(m)$ of possible states of $a_{ij}(m,q_{ij}^{<},q_{ij}^{>})$

m	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$N^{(a)}(m)$	1	2	4	6	9	12	16	20	25	30	36	42	49	56	64	72	81	90	100

m	28	…	38	…	48	…	58	…	98	…	998	…	9998
$N^{(a)}(m)$	225	…	400	…	625	…	900	…	2.5×10^3	…	2.5×10^5	…	2.5×10^7

- The number m is even: $N^{(a)}(m)=\frac{1}{4}(m+2)^2$ is a square number
- The number m is odd: $N^{(a)}(m)=\frac{1}{4}\left((m+2)^2-1\right)=\frac{1}{4}(m+1)(m+3)$ is an even number.

For the following exposition, the indices ij etc. are suppressed, and the notations

$$q_- = \min\left(q^<,q^>\right) \text{ and } q_+ = \max\left(q^<,q^>\right) \tag{3.10}$$

are used. Obviously, $q_- \le q_+$ holds.

Using (3.5), (3.7), and (3.10), the Eq. (3.8) can be written as:

$$a\left(m,q_-,q_+\right)=\frac{1}{2}\left(\frac{q_-+q_+}{m}+\left[\frac{m}{2}\right]^{-1}\cdot q_-\right)=\frac{1}{2}\left(\frac{q_+}{m}+\left(\frac{1}{m}+\left[\frac{m}{2}\right]^{-1}\right)\cdot q_-\right). \tag{3.11a}$$

Let m be an even number. Because of $\left[\frac{m}{2}\right]=\frac{m}{2}$, then the Eq. (3.11b) results from (3.11a):

$$a_{\text{ev}}=\frac{q_+ + 3q_-}{2m}. \tag{3.11b}$$

Therefore, the condition for two distances $a_{\text{ev}}^{(1)}$ and $a_{\text{ev}}^{(2)}$ having the same value is:
if $q_+^{(1)}-q_+^{(2)}=3(q_-^{(2)}-q_-^{(1)})$ then $a_{\text{ev}}^{(1)}=a_{\text{ev}}^{(2)}$.
This can generally be satisfied in many ways if $m \ge 4$.

In the case where m is odd and $m \ge 3$, all the $N^{(a)}(m)$ values, which the distance a_{od} can have, are different from each other. Now $\left[\frac{m}{2}\right]=\frac{m-1}{2}$ holds. When this is used in (3.11a), the Eq. (3.11c) for determining a_{od} is:

$$a_{\text{od}}=\frac{(m-1)q_+ + (3m-1)q_-}{2m(m-1)}. \tag{3.11c}$$

If the condition

$$q_+^{(1)}-q_+^{(2)}=\frac{3m-1}{m-1}\left(q_-^{(2)}-q_-^{(1)}\right) \tag{3.12}$$

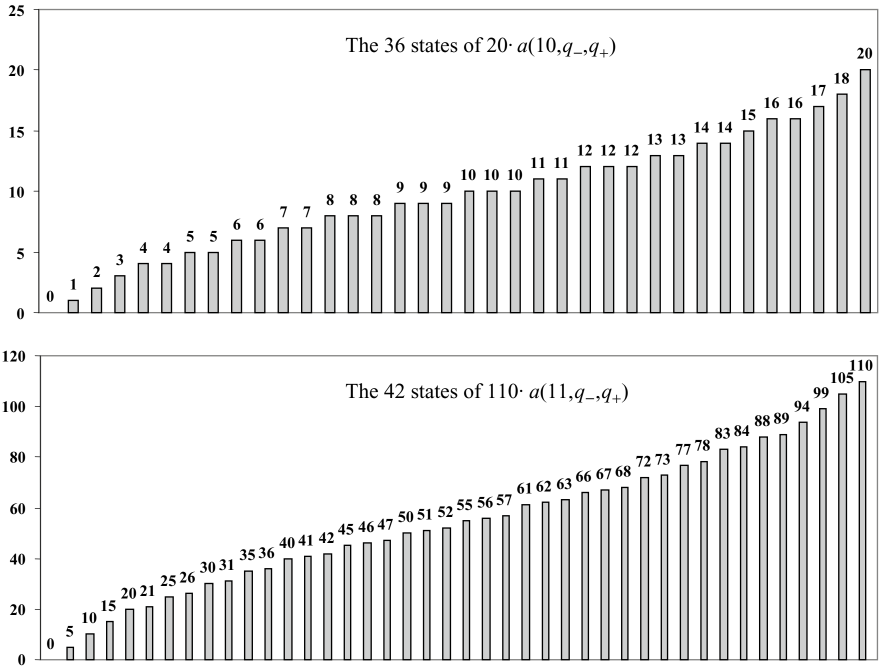


Fig. 3.1 The possible states of the measures $a(10,q_-,q_+)$ and $a(11,q_-,q_+)$ and their values ((11) times corresponding scale factor 20 and 110, respectively, see the ordinate)

would be fulfilled, the equality $a_{\mathrm{od}}^{(1)} = a_{\mathrm{od}}^{(2)}$ would be true. But the difference on the left side of (3.12) is an integer in contrast to that on the right, if $m > 3$. Namely the factor $f = \frac{3m-1}{m-1}$ is not an integer, if $m > 3$. This can easily be seen if m is given as $m = 2n + 1$. Then $f = \frac{3n+1}{n} = 3 + \frac{1}{n}$, so that the factor f is an integer only in the case of $n = 1$ and $m = 3$, respectively. In this case, the distance $a(3,q_-,q_+)$ has six different values:

q_-	0	0	0	0	1	1
q_+	0	1	2	3	1	2
$a(3,q_-,q_+)$	0	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{5}{6}$	1

In Fig. 3.1 the possible states of the functions $a(10,q_-,q_+)$ and $a(11,q_-,q_+)$ and their corresponding values are shown.

Because of its connection to the partial order ≤, only the measure a_{ij} will be used in the following.

3.3 Generating Directions in a Distance Matrix: Tournaments

3.3.1 About Tournaments

As mentioned above, incomparability (or inequality) between pairs of objects $\{\mathbf{x}_i,\mathbf{x}_j\}$ of a set $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_i,\mathbf{x}_j,\ldots,\mathbf{x}_{|X|}\}$ is a symmetrical problem. Therefore, the situation of all the relations between the $g=|G|$ elements of a set G associated with X can be described with the graph theory as a complete (simple undirected) graph $\langle g\rangle$ where the elements $g_i \in G$ are its vertices. (Concerning graph theoretical terms and relationships see for instance (Sachs 1970; Balakrishnan and Ranganathan 2000; Chartrand and Zhang 2005; Bondy and Murty 2008; Bartel and Mucha 2011).) It is possible to assign one of the distance values u_{ij} (3.5), v_{ij} (3.7), or a_{ij} (3.8, 3.11) introduced in Sect. 3.2.2 to the corresponding edge g_i–g_j, which is one of the $g(g-1)/2$ edges of the graph $\langle g\rangle$. But, as mentioned above, only the measure a_{ij} is considered in this chapter.

The introduction of directions into the graph $\langle g\rangle$ can be achieved by converting it into a tournament T_g. This directed graph (digraph) is defined as follows:

(D5) A (g-)**tournament** T_g is a directed graph whose underlying graph is the complete graph $\langle g\rangle$. For each pair of distinct vertices $g_i, g_j \in \mathrm{T}_g$, either $(g_i,g_j)\equiv g_i \rightarrow g_j$ or $(g_j,g_i)\equiv g_j \rightarrow g_i$ is an arc (directed edge), but not both.

In social life, a tournament is a competition involving a number of competitors (individuals or teams), all participating in a sport or in a game. The graph-theoretic image of g competitors is the set of vertices G with $|G|=g$. The complete graph $\langle g\rangle$ means that each competitor has to compete against every other competitor.

In a tournament, rules must exist that allow determining the winner of two competitors in their joint 'fight'. Let $g_i, g_j \in G$ be two competitors. Furthermore, it is agreed that $g_i \rightarrow g_j$ means that g_i has defeated his opponent g_j. The application of the rules to the contests of all pairs assigns directions to all edges. In this way, the graph $\langle g\rangle$ is transformed into a tournament T_g. This transformation $\langle g\rangle \rightarrow \mathrm{T}_g$ will be formalized for the general graph-theoretical aspect to be used in the application of interest in this chapter.

Before doing this, it has to be noted that there are exactly two classes of 3-tournaments as shown in Fig. 3.2: the transitive directed 3-tournaments (Fig. 3.2a) and the cyclic directed 3-tournaments (Fig. 3.2b). Now it can be defined:

(D6) A n-tournament (tournament with n vertices) T_n with $n>3$ is called a **transitive directed tournament** if all its 3-subtournaments are transitive directed. There are no cycles in such a tournament, and if there is no cycle in a tournament it is a transitive directed one.

Let t be a vertex of a tournament T_n. Then $\alpha^+(t)$ denotes the number of incoming (in t) arcs and $\alpha^-(t)$ the number of outgoing (of t) arcs. Of course, it is

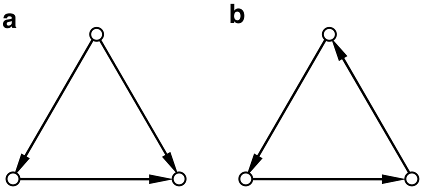


Fig. 3.2 The transitive directed 3-tournament (**a**) and the cyclic directed 3-tournament (**b**)

$$\alpha^{+}(t)+\alpha^{-}(t)=n-1 \tag{3.13}$$

for all vertices of T_n. Using these terms, the following theorem can be formulated (Sachs 170: 167):

(T1) In a transitive directed tournament T_n, there are no two vertices s and t with the same number of outgoing arcs: $\alpha^{-}(s)\neq\alpha^{-}(t)$.

Taking into account equation (3.13) and theorem (T1) it follows:

(T2) In a transitive directed tournament T_n, there are no two vertices with the same number of incoming arcs: $\alpha^{+}(s)\neq\alpha^{+}(t)$.

From theorems (T1) and (T2) and Eq. (3.13), it follows that the vertices of a transitive directed tournament T_n can always be renumbered or sorted in a unique way, so that the following sequences of inequalities hold:

$$0\le\alpha^{-}(t_n)<\alpha^{-}(t_{n-1})<\cdots<\alpha^{-}(t_2)<\alpha^{-}(t_1)\le n-1, \tag{3.14}$$

$$0\le\alpha^{+}(t_1)<\alpha^{+}(t_2)<\cdots<\alpha^{+}(t_{n-1})<\alpha^{+}(t_n)\le n-1. \tag{3.15}$$

Regarding competitions (matches or games) the competitor t_1 is the absolute winner, i.e., is the first ($\alpha^{-}(t_1)=n-1$), the competitor t_2 is the second ($\alpha^{-}(t_2)=n-2$), …, and the competitor t_n is the last (nth), i.e., is the absolute loser ($\alpha^{-}(t_n)=0$).

It can also be shown (Sachs 1970: 168):

(T3) If the vertices of a transitive directed tournament T_n are numbered in the way, so that the inequalities (3.14) are fulfilled, then, in each case, exactly one arc from each vertex $t_i\in\mathrm{T}_n$ $(1\le i\le n-1)$ is directed to each vertex t_j with a higher index $j>i$.

If the vertices are numbered, so that (3.14) (and hence the theorem (T3)) is satisfied, it then follows directly:

(T4) A transitive directed tournament T_n has exactly one directed path of length n where all its $n-1$ arcs have the same orientation.

Using the numbering of the vertices given in (3.14) or (3.15), the directed path mentioned in (T4) is $t_1\to t_2\to\cdots\to t_{n-1}\to t_n$.

3.3.2 Transformation of a Complete Graph into a Tournament

For the transformation of a complete graph $\langle g\rangle$ into a transitive tournament T_g, the following approach is proposed: It is assumed that a data matrix $\mathbf{X}(G, M)$ (with $g=|G|$, $m=|M|$) is given [see Eq. (3.1)]. Thus, it is possible to determine the distance matrix $\mathbf{A} = (a_{ij})$ whose elements a_{ij} are calculated according to (3.8), (3.5), and (3.7).

In the next step the sum (total incomparability of object i with respect to all other objects)

$$s_i^{(a)} = \sum_{h=1}^{g} a_{ih} \text{ with [+ blank]} \frac{g-1}{2m} \le s_i^{(a)} \le g-1 \tag{3.16}$$

is calculated for $i=1,2,\ldots,g$.

The orientation of the arcs, which are generated by transformation from edges, is defined by the following 'rules':

$$\text{if}\begin{Bmatrix} s_i^{(a)} < s_j^{(a)} \\ s_i^{(a)} = s_j^{(a)} \begin{Bmatrix} \text{but}\, i < j \\ \text{but}\, j < i \end{Bmatrix} \\ s_i^{(a)} > s_j^{(a)} \end{Bmatrix} \text{then} \begin{Bmatrix} g_i \to g_j \\ \begin{Bmatrix} g_i \to g_j \\ g_j \to g_i \end{Bmatrix} \\ g_j \to g_i \end{Bmatrix}. \tag{3.17}$$

(For the transformation of a complete undirected graph into a transitive tournament, see for instance (Bartel 1989, 1990a, 1996).)

It is to be noted that in case of equality $s_i^{(a)}=s_j^{(a)}$, the two vertices g_i and g_j are to be regarded as equivalent. In this case, it is decided randomly whether g_i or g_j is the 'winner'. Otherwise, the smaller total incomparability (3.16) ($s_i^{(a)}$or $s_j^{(a)}$) is crucial for this decision.

Taking into account the rules in (3.17) there are obviously three classes of 3-subtournaments in terms of the number (3, 1, or 0) of pairs of equal sums. As demonstrated in Fig. 3.3, they are all transitive directed. The numbers of cases are 3! for Class I, $3\cdot 2$ for Class II, and again 3! for Class III. For any three vertices $\{g_i,g_j,g_h\}\subseteq G$ exactly one in a total of 18 transitive directed 3-subtournaments is existent. Thus, all the $\frac{g(g-1)(g-2)}{6}$ 3-subtournaments of T_g are transitive directed. Therefore, T_g is also a transitive directed tournament (see (D6)).

Based on the latter, the four theorems (T1) to (T4) apply to the tournament $T_g : \langle g\rangle \overset{rules(3.17)}{\rightarrow} T_g$. Because of the theorems (T1) and (T3), it will be advantageous to renumber the vertices $\{g_1,\ldots,g_g\}\in T_g$ in the following way (instead of g_i, the vertices sorted in the order of the inequalities (3.14) are denoted by t_i ($i = 1,2,\ldots,g$)):

$$for\, i = 1,2,\ldots g:\ g_i \to t_h \ \ if\, \alpha^-(g_i) = g-h\big(h\in\{1,2,\ldots,g\}\big). \tag{3.18}$$

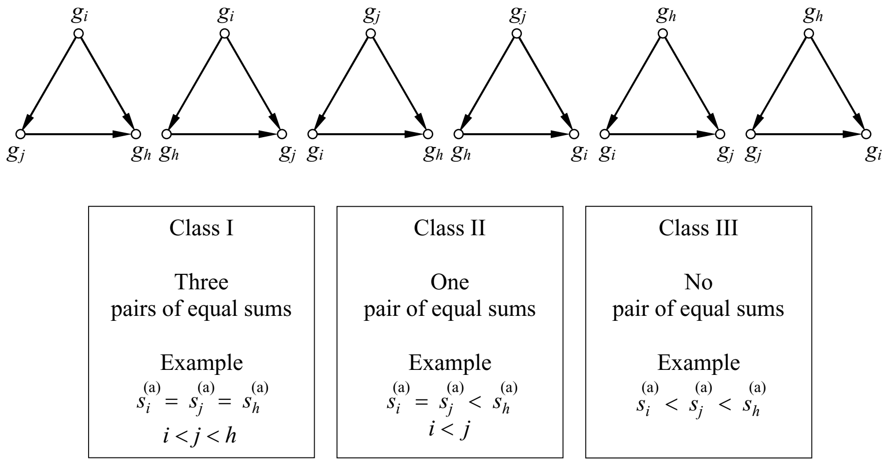


Fig. 3.3 The six (3!) possible types of 3-subtournaments, which are generated applying the rules (3.17) to any three vertices $\{g_i, g_j, g_h\} \subseteq G$ (above), and the three classes, in which they are each realized (The three examples are in accordance with the tournament that is shown above on the left)

With the help of this sorting, the directed path mentioned in Theorem (T4) can be specified:

$$t_1\left(\alpha^- = g-1\right) \to t_2\left(\alpha^- = g-2\right) \to \cdots \to t_i\left(\alpha^- = g-i\right) \to \cdots \to t_{g-1}\left(\alpha^- = 1\right) \to t_g\left(\alpha^- = 0\right), \qquad (3.19)$$

where the arcs → are evaluated by the corresponding element of the distance matrix $\mathbf{A}^{(s)}$, in which the rows and columns are arranged according to the sorting (3.18):

$$t_i \overset{a_{i,i+1} \in \mathbf{A}^{(s)}}{\to} t_{i+1}\left(i = 1, 2, \ldots, g-1\right). \qquad (3.20)$$

The rules for transforming a complete undirected graph $\langle g \rangle$ (edge-evaluated (g_i–g_j by the distance $a_{ij} \in \mathbf{A}$)) into a transitive directed tournament T_g (arc-evaluated ($t_k \to t_l$ by the distance $a_{kl} \in \mathbf{A}^{(s)}$) and vertex-evaluated ($t_k$ by $s_k^{(a)}$)) have been chosen so that this tournament T_g is completely described by the matrix $\mathbf{A}^{(s)}$, whose columns and rows are arranged in accordance with (3.18):

$$\langle g \rangle \equiv \mathbf{A} \overset{(3.17),(3.18)}{\to} \mathbf{A}^{(s)} \equiv \mathrm{T}_g. \qquad (3.21)$$

Using only the lower triangular matrix of $\mathbf{A}^{(s)}$, the equivalence between this matrix and the tournament T_g is demonstrated in Fig. 3.4. Beneath the triangular matrix, the directed path (3.19) is described, whose arcs are evaluated as given in (3.20). Furthermore, the evaluation of the vertices is specified there. The value $a_{kl} \in \mathbf{A}^{(s)}$ represents the evaluation of the arc $t_k \to t_l$ $(k = 1, 2, \ldots, g-1, k < l)$.

$$\begin{array}{c|ccccccc}
t_1 & 0 & & & & & & \\
t_2 & a_{12} & 0 & & & & & \\
\vdots & \vdots & \vdots & \ddots & & & & \\
t_i & a_{1i} & a_{2i} & \cdots & 0 & & & \\
t_{i+1} & a_{1,i+1} & a_{2,i+1} & \cdots & a_{i,i+1} & 0 & & \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \\
t_g & a_{1g} & a_{2g} & \cdots & a_{ig} & a_{i+1,g} & \cdots & 0 \\
\hline
 & t_1 & \to t_2 & \cdots & \to t_i & \to t_{i+1} & \cdots & \to t_g \\
 & s_1^{(a)} & \le s_2^{(a)} & \cdots & \le s_i^{(a)} & \le s_{i+1}^{(a)} & \cdots & \le s_g^{(a)}
\end{array} \equiv \mathrm{T}_g$$

Fig. 3.4 Equivalence of the matrix $\mathbf{A}^{(s)}$ and the tournament T_g (see text)

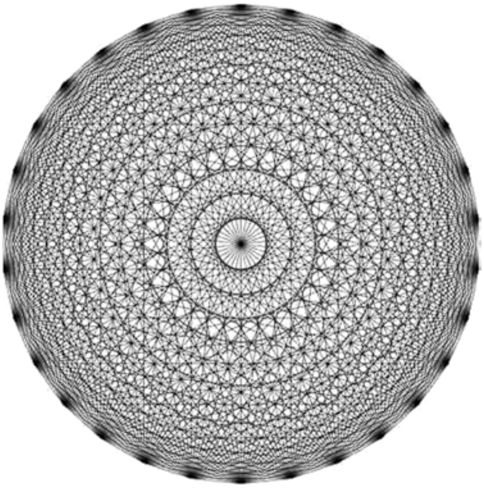


Fig. 3.5 The complete graph ⟨30⟩ (435 edges)

The number of values in the lth column under $a_{ll} = 0$ corresponds to the number $\alpha^{-}(t_l) = g - l$ of outgoing arcs of the vertex t_l.

The use of the matrix representation of the tournament instead of its graphical representation will be advantageous or even necessary, when the number of objects under consideration is not small. In this case, the graphical representation is confusing and provides little or no information. The complete graph ⟨30⟩ in Fig. 3.5 confirms this statement.

An analogous statement also applies to the relation graphs named after Helmut Hasse (1898–1979), whose graphical representation consists of an unmanageable and information empty 'tangle' of lines even for a not very large number of objects (see, for example, the Fig. 2 in (Halfon and Brüggemann 1998: 21) and the corresponding comment on the Hasse diagram shown there).

The method described in this section for generating directions in a symmetric matrix (with main diagonal elements equal to zero) can be applied to each distance matrix $\mathbf{D} = (d_{ij})$.

Table 3.5 Data matrix $\mathbf{X}_{\mathrm{CaBr}}(G,M)$ for demonstration [from (Carlsen and Brüggemann 2011: 126)]

Name of chemical	ID		Volat	Sedim	Advec
Phenanthrene	A_ph	g_1	3	2	4
Pyrene	A_py	g_2	3	3	4
Fluoranthene	A_fl	g_3	2	3	4
Chloroform	C_ch	g_4	4	1	2
Tetrachlormethane	C_tt	g_5	4	1	3
Trichlorethene	C_tr	g_6	4	2	2
Tetrachlorethene ('Per')	C_pe	g_7	3	2	3
Atrazine	D_at	g_8	1	2	4
Nitrilotriacetic acid	D_nt	g_9	1	1	1
EDTA	D_ed	g_{10}	1	1	3

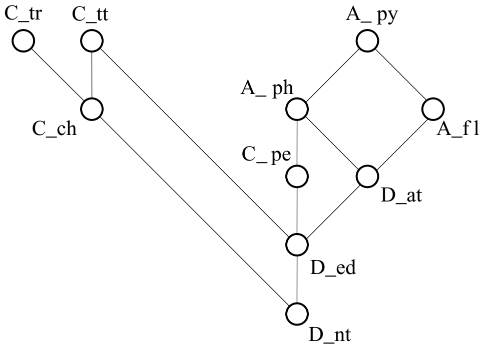


Fig. 3.6 Hasse diagram corresponding to the data matrix $\mathbf{X}_{\mathrm{CaBr}}(G,M)$ [from (Carlsen and Brüggemann 2011: 133)]

3.3.3 An Example for Demonstration

As a small example, the same data matrix is used as in (Bartel and Mucha 2011), which was taken from the publication of Lars Carlsen and Rainer Brüggemann (2011). However, the four chemicals of class B (polychlorinated biphenyls), which are contained in the original data matrix, were not taken into account. Thus, the 10×3 data matrix $\mathbf{X}_{\mathrm{CaBr}}(G,M)$ is considered as given in Table 3.5. Figure 3.6 shows the Hasse diagram derived from this data matrix.

Using the relations (3.8), (3.5), and (3.7) the distance matrix $\mathbf{A}^{(\mathrm{s})}{}_{\mathrm{CaBr}} \equiv \langle 10 \rangle_{\mathrm{CaBr}}$ given in Table 3.6 is obtained. In its last row, the sums $s_i^{(\mathrm{a})}$ calculated according to (3.16) and sorted according to (3.18) are given.

As explained, the matrix $\mathbf{A}_{\mathrm{CaBr}}{}^{(\mathrm{s})}$ corresponds to the arc-evaluated tournament $\mathrm{T}_{10}{}^{\mathrm{CaBr}}$ which is derived from the complete graph $\langle 10 \rangle_{\mathrm{CaBr}}$ by means of the rules (3.17).

Table 3.6 The distance matrix $\mathbf{A}_{\mathrm{CaBr}}{}^{(s)}$

		D_nt	D_ed	A_py	A_ph	D_at	C_pe	A_fl	C_tt	C_ch	C_tr
t_1	D_nt	**0**	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{2}$
t_2	D_ed	$\frac{1}{6}$	**0**	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{6}$	1
t_3	A_py	$\frac{1}{2}$	$\frac{1}{2}$	**0**	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	1	$\frac{1}{6}$	1
t_4	A_ph	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	**0**	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{5}{6}$	1	$\frac{1}{6}$	$\frac{5}{6}$
t_5	D_at	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	**0**	$\frac{5}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	1	$\frac{5}{6}$
t_6	C_pe	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{5}{6}$	**0**	1	$\frac{5}{6}$	1	$\frac{5}{6}$
t_7	A_fl	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{1}{3}$	1	**0**	1	$\frac{1}{6}$	1
t_8	C_tt	$\frac{1}{3}$	$\frac{1}{6}$	1	1	1	$\frac{5}{6}$	1	**0**	$\frac{1}{6}$	$\frac{5}{6}$
t_9	C_ch	$\frac{1}{3}$	$\frac{1}{2}$	1	1	1	1	1	$\frac{1}{6}$	**0**	$\frac{1}{6}$
t_{10}	C_tr	$\frac{1}{2}$	1	1	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	1	$\frac{5}{6}$	$\frac{1}{6}$	**0**
$s^{(a)}$		$3\frac{2}{3}$	$4\frac{1}{3}$	5	$5\frac{1}{6}$	$5\frac{1}{6}$	$5\frac{5}{6}$	$6\frac{1}{3}$	$6\frac{1}{3}$	$6\frac{1}{2}$	7

Figure 3.7 illustrates this connection. The arc evaluations missing in the graphical representation can be found in the triangular matrix. The evaluation of the vertices by the sums (3.16) is written given beneath the triangular matrix (as in Fig. 3.4).

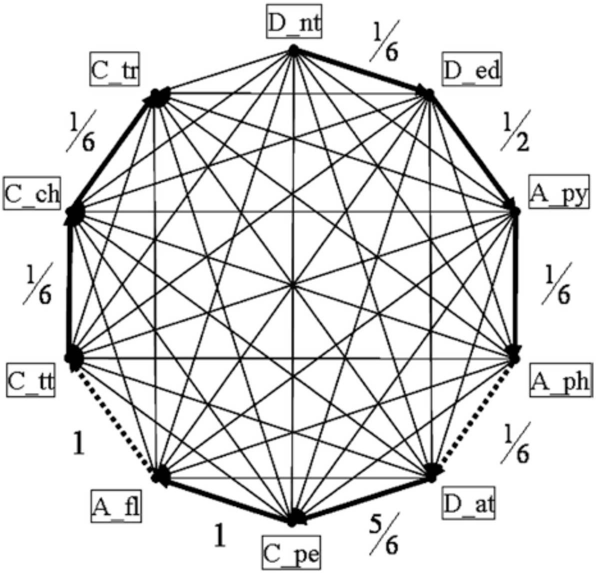
With the help of the matrix $\mathbf{A}^{(s)}{}_{\mathrm{CaBr}}$ (Table 3.6) and the tournament $\mathrm{T}_{10}{}^{\mathrm{CaBr}}$ (Fig. 3.7), respectively, all statements regarding the combined incomparability/inequality (size and direction) between given objects can be obtained because of the different values of their variables. Therefore, information can be obtained complementary to the one obtained by partial order theory (Hasse Diagram Technique, see for instance (Halfon and Reggiani 1986; Brüggemann and Halfon 1995).

A further discussion of these opportunities with more appropriate examples must be postponed for now. In the present case of only three variables, the function u_{ij} (3.5) can only have two values (0 or 1). Therefore, it is equivalent to the yes/no decision about incomparability (D2) that is used in partial order theory.

3.4 Cluster Analysis Using the Measure of Incomparability/Inequality

In this publication, the example ($\mathbf{X}_{\mathrm{CaBr}}(G,M)$, Table 3.5) described above will be used to classify objects G according to their attributes M using the distance matrix $\mathbf{A}^{(s)}{}_{\mathrm{CaBr}}$ (Table 3.6). For the decomposition of the set of objects in classes, a method

Fig. 3.7 Matrix and graphical representation of the tournament T_{10}^{CaBr} (In the graphical representation, only the evaluation of the arcs (bold (dotted if $s_i^{(a)} = s_j^{(a)}$) arrows) of the directed path [see (T4)] is given)



of partitioning cluster analysis is used namely Helmut Späth's **ex**change **m**ethod TIHEXM (Späth 1985), see also (Bartel et al. 2003). The TIHEXM procedure appears particularly suitable, since no restrictions on the nature of the distance function exist.

In the language of graph theory, this method is described generally as: Let $\mathbf{X}(G,M)$ be a given data matrix, from which the distance matrix $\mathbf{D} = (d_{ij})$ is calculated. To this matrix, the complete graph $\langle g \rangle$ ($g = |G|$: number of objects), whose edges g_i–g_j are evaluated by the distances $d_{ij} \in \mathbf{D}$, is assigned. In the TIHEXM method, the decomposition of the set of objects G into $c \geq 2$ classes $C_i \subset G$ ($i = 1, 2, \ldots, c$) corresponds to the splitting of the graph $\langle g \rangle$ into the graph $\Delta_g{}^c$ with c components $\Delta_g{}^c = \{\langle g_1 \rangle, \langle g_2 \rangle, \ldots, \langle g_c \rangle\}$, where $\sum_{i=1}^{c} g_i = g$. This splitting is achieved by deleting edges so that the sum of the sums of the edge evaluations of all c complete subgraphs $\langle g_1 \rangle, \langle g_2 \rangle, \ldots, \langle g_c \rangle \subset \langle g \rangle$ is minimal. Here, the evaluation of the point graph $\langle 1 \rangle$, which represents an isolated vertex and a single object, respectively, is set to zero. (For the underlying exchange algorithm of the TIHEXM method and its three versions see references mentioned above.)

The results for the decomposition of the given example (data matrix $\mathbf{X}_{\mathrm{CaBr}}$: Table 3.5, distance matrix $\mathbf{A}^{(s)}{}_{\mathrm{CaBr}}$: Table 3.6) into two, three and four classes using version 1 of TIHEXM procedure are shown in Table 3.7. The observed number of objects in the classes after decomposition is shown schematically in Fig. 3.8.

When considering and discussing the results, it must be noted that the similarity between objects of a class results from small values of the incomparability/inequality and not from small differences between the values of the attributes. Therefore, it is expected that the classification results are reflected in the corresponding relation graph. The Hasse diagrams shown in Fig. 3.9, in which the vertices are coloured in accordance with classes (see coloured circles in Table 3.7), confirm what has been just said, at least for the studied example.

Table 3.7 Result of the classification (cluster analysis) using the method TIHEXM (The different classes are distinguished by coloured circles)

Partitions of minimum criterion out of 50 reruns		
	Decomposition into	
$c = 2$	$c = 3$	$c = 4$
	Classes	
C_ch●	C_ch●	C_ch●
C_tr	C_tr	C_tr
C_tt	C_tt	C_tt
D_nt	D_nt	D_nt○
D_ed	D_ed○	D_ed
D_at	D_at	D_at◌
A_fl○	A_fl	A_fl
A_ph	A_ph	A_ph
A_py	A_py◍	A_py◍
C_pe	C_pe	C_pe

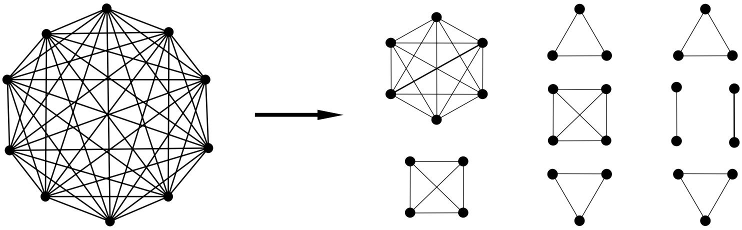


Fig. 3.8 Schematic representation of the splitting of the graph ⟨10⟩ in two, three, and four complete subgraphs according to the results of the cluster analysis with TIHEXM

As shown in Table 3.7 and Fig. 3.9, there are three subsets of objects which are stable with respect to their togetherness: {C_ch, C_tr}, {D_at, A_fl}, {A_ph, A_py, C_pe}. Of course, the reason for this togetherness is a small incomparability/inequality among each other.

The sequence of objects found in Sect. 3.3.3 does not fully coincide with the sequence of their class membership:

	Sequence									
object	D_nt	D_ed	A_py	A_ph	D_at	C_pe	A_fl	C_tt	C_ch	C_tr
class	□	□	▓	▓	░	▓	░	□	■	■

However, this was not to be expected. Rather, this difference can be used for the application of a previously developed classification method that combines classification and seriation (Bartel 1990b, 1991). This would be interesting to investigate further in the future.

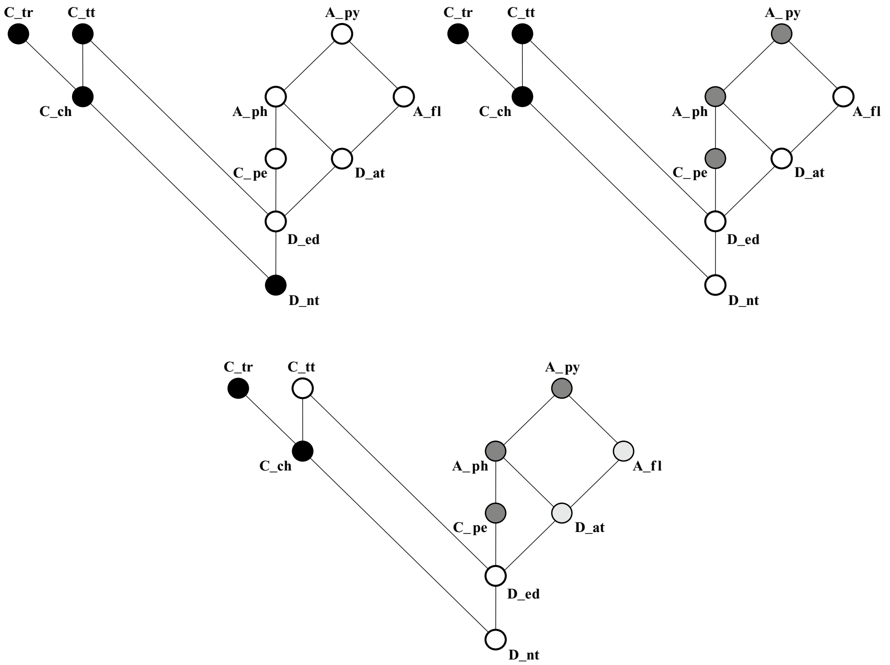


Fig. 3.9 The classes found and their location in the Hasse diagram corresponding to data matrix $\mathbf{X}_{\text{CaBr}}$

3.5 Future Tasks

This publication is meant as a first introduction of the problem regarding measures of inequality and incomparability.

Therefore, several issues will have to be investigated further. Besides the above already mentioned tasks, these include the classification and comparison of the proposed measures with other measures and distance functions, respectively. A modification of the distance functions introduced here is also possible.

In particular, attention will be paid to the application of cluster analysis using the measures of incomparability and inequality. During the process, known methods will be examined for their applicability and modified if necessary. In addition, the development of new methods can be considered. The connection of classification and seriation has already mentioned.

Of course, relations to the theory of partial order (Hasse Diagram Technique) should be studied.

References

Ambrosius (1897) De Isaac vel anima. In: Schenkel K (ed) Sancti Ambrosii opera I. G. Freytag, Leipzig, pp 639–700

Balakrishnan R, Ranganathan K (2000) A textbook of graph theory. Springer, New York, NY

Bartel H-G (1989) Was ist, was kann die Seriation? Wissenschaft und Fortschritt 39:321–324

Bartel H-G (1990a) Seriation auf graphentheoretischer Basis. Wissenschaftliche Zeitschrift der Humboldt-Universität zu Berlin, Reihe Gesellschaftswissenschaften 39(3):251–257

Bartel H-G (1990b) Seriation to describe some aspects of generalized evolution and its application in chemical informatics. Syst Anal Model Simul 7(7):557–566

Bartel H-G (1991) A modified Kretschmer complexity index for selecting end partitions in cluster analysis. Syst Anal Model Simul 8(2):139–145

Bartel H-G (1996) [Diskrete] Mathematische Methoden in der Chemie. Spektrum Akademischer Verlag, Heidelberg, pp 139–141

Bartel H-G, Mucha H-J (2011) Finding incomparable pairs of subsets by using formal concept analysis. Statistica & Applicazioni – Special Issue (Partial orders in applied sciences): 61–79

Bartel H-G, Mucha H-J, Dolata J (2003) Über eine Modifikation eines graphentheoretisch basierten partitionierenden Verfahrens der Clusteranalyse. Match Commun Math Comput Chem 48:209–223

Bondy JA, Murty USR (2008) Graph theory. Springer, New York, NY

Brüggemann R, Halfon E (1995) Theoretical base of the program "Hasse". GSF-Bericht 20/95, München-Neuherberg

Brüggemann R, Voigt K (2012) Antichains in partial order, example: pollution in a german region by lead, cadmium, zinc and sulfur in the herb layer. Match Commun Math Comput Chem 67:731–744

Carlsen L, Brüggemann R (2011) Risk assessment of chemicals in the River Main (Germany) – application of selected partial order ranking tools. Statistica & Applicazioni – Special Issue (Partial orders in applied sciences): 125–140

Chartrand G, Zhang P (2005) Introduction to graph theory. McGraw-Hill Education, Boston, MA

CIL VI (1886) Corpus inscriptionum latinarum, vol. VI, pars III: Inscriptiones urbis Romae latinae. consilio et auctoritate Academiae litterarum regiae Borussicae col-legerunt G. Henzen et I. B. de Rossi, ediderunt E. Bormann, G. Henzen, Chr. Huelsen. G. Reimer, Berlin, p. 1814

Halfon E, Brüggemann R (1998) On ranking chemicals for environmental hazard – comparison of methodologies. Berichte des IGB, Heft 6 (Sonderheft I): 11–48

Halfon E, Reggiani MG (1986) On ranking chemicals for environmental hazard. Environ Sci Technol 20:1173–1179

Sachs H (1970) Einführung in die Theorie der endlichen Graphen, Teil 1. Teubner, Leipzig

Shakespeare W (1594) Lucrece. I. Harrison, London, p [iv]

Späth H (1980) Cluster analysis algorithms for data reduction and classification of objects. E. Horwood, Chichester, p 15

Späth H (1985) Cluster dissection and analysis: theory, FORTRAN programs, examples. E. Horwood, Chichester, pp 84–88, 132–135

Webster (1994) Webster's new encyclopedic dictionary. Könemann, Cologne, p 506

Wilde O (1911) In: Ross RB (ed) De Profundis. G. P. Putnam's Sons, New York, NY, p 112

Chapter 4
Measuring Structural Dissimilarity Between Finite Partial Orders

Marco Fattore, Rosanna Grassi, and Alberto Arcagni

Abstract In this paper, we address the problem of measuring structural dissimilarity between two partial orders with n elements. We propose a structural dissimilarity measure, based on the distance between isomorphism classes of partial orders, and propose an interpretation in terms of graph theory. We give examples of structural dissimilarity computations, using a simulated annealing algorithm for numerical optimization.

4.1 Introduction

The issue of measuring the degree of dissimilarity of two finite partially ordered sets (posets) is interesting from many points of view. Consider, for example, the evaluation of countries against different socio-economic indicators (e.g., democracy, economic freedom, and human development) and the way the resulting partial order may change as different national policies are implemented. In other cases, one may want to compare the way two different populations partially order a set of common alternatives (e.g., personal or social values, and quality-of-life dimensions). In both examples, the problem reduces to measuring the degree of dissimilarity of two or more posets. The issue already found its interest in discrete mathematics and in chemistry (see Brüggemann and Patil 2011, Klein 1995, Monjardet 1981, Voigt et al. 2011) (and it will be a future task to relate these methods to the procedure, shown here) and can be addressed in many ways. For example, one may compare partial orders comparing some of their basic features: the number of comparabilities or incomparabilities, the number of chains, their dimensions, their heights or widths

M. Fattore (✉) • R. Grassi • A. Arcagni
Department of Statistics and Quantitative Methods,
University of Milano-Bicocca, Milano, Italy
e-mail: marco.fattore@unimib.it

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_4, © Springer Science+Business Media New York 2014

(see, for example, Annoni et al. 2011). If the partial orders are defined on the same finite set, one may compute the fraction of shared comparabilities or compare the principal down-sets or up-sets. Our approach, however, is somehow different, in that we attempt to define a "global" indicator to measure the structural dissimilarity of two posets (for other approaches to this goal, see Brüggemann and Bartel 1999, Brüggemann et al. 2001). Although the notion of a "structure" is a little bit vague, the meaning of the term "structural" should be clear: we look for a measure which does not change if the posets to be compared are transformed by order isomorphisms. This way we obtain a measure that does not depend upon the labeling of the posets and that better captures deep differences in relational patterns. Finite posets may be effectively depicted as Hasse diagrams, which are directed acyclic graphs. A dissimilarity measure invariant under poset isomorphisms may therefore be seen as a measure invariant under graph isomorphisms, that is as a "topological" measure. This link between posets and graphs is the key to our approach, which specializes to posets analogous results developed for graphs (Fattore and Grassi 2012). The paper is organized as follows. Section 4.2 provides some basic definitions of graph and poset theory; Sect. 4.3 introduces the structural dissimilarity measure; Sect. 4.4 discusses some examples; Sect. 4.5 concludes.

4.2 Technical Preliminaries

In this section, we collect some technical definitions and results that are essential for subsequent developments. We begin with some basic concepts from graph theory. The presentation is in the spirit of Fattore and Grassi [2012].

A graph $G=(V,E)$ is an ordered pair V,E, where V is a set of n *nodes*, or *vertices*, and E is a set of m pairs of nodes of V ; the pair $(i,j)\in E$ $(i\neq j)$ is called an *edge* of G and i and j are called adjacent $(i\sim j)$; an *undirected* graph is a graph in which $(j,i)\in E$ whenever $(i,j)\in E$, otherwise the graph is called *directed* (or a *digraph*). A *path* is a sequence of distinct adjacent vertices; an $i-j$ *path* is a path starting from vertex i and arriving at vertex j. The *length* of a path is the number of edges in it; a shortest path joining vertices i and j is called an $i-j$ *geodesic*. The *distance* $d\ (i,j)$ between two vertices i and j is the length of an $i-j$ geodesic. If two vertices are connected by no path, then their distance is set to ∞. The *diameter* of G is the greatest distance between any pair of vertices of G. A *cycle* is a path starting from and ending at the same node. A graph is *acyclic* if it has no cycles. In a directed graph $D=(V,E)$, elements of E are called *arcs*. If (i,j) (or $j,i)\in E$, then vertices i and j are *adjacent*. An $i-j$ *directed* path is a sequence from i to j of distinct adjacent vertices; in this case, j is *reachable* from i and the distance $d\ (i,j)$ between two vertices i and j is the length of any shortest directed $i-j$ path. A *cycle* is a directed path with $i=j$; as in the undirected case, D is *acyclic* if it has no cycles (Harary 1969).

We now turn to partial order theory. A partially ordered set (or a *poset*) $P=(X,\leq)$ is a set X equipped with a partial order relation $\leq$, that is a binary relation satisfying the properties of *reflexivity*, *antisymmetry*, and *transitivity* (Davey and Priestley 2002):

1. $p \leq p$ for all $p \in X$ (reflexivity).
2. if $p \leq q$ and $q \leq p$ then $p = q$, $p, q \in X$ (antisymmetry).
3. if $p \leq q$ and $q \leq r$, then $p \leq r$, $p, q, r \in X$ (transitivity).

If $p \leq q$ or $q \leq p$, then p and q are called *comparable*, otherwise they are said *incomparable* (written $p \parallel q$). A partial order P where any two elements[1] are comparable is called a *chain* or a *linear order*. On the contrary, if any two elements of P are incomparable, then P is called an *antichain*. The set of comparabilities of a poset P is written Comp(P).[2] Given $p, q \in P$, q is said to *cover* p (written $p \prec q$) if $p \leq q$ and there is no other element $r \in P$ such that $p \leq r \leq q$. In a finite poset (i.e., a poset defined on a finite set), the cover relation determines the partial order relation, since it can be easily proved that $p \leq q$ if and only if there exists a sequence of elements $r_0, r_1, \ldots, r_k$, such that $p = r_0 \prec r_1 \prec \ldots \prec r_k = q$. A finite poset P can be easily depicted by means of a *Hasse diagram*. Hasse diagrams are directed acyclic graphs representing the cover relation generating the partial order. Hasse diagrams are drawn according to the following two rules: (1) if $p \leq q$, then node q is placed above node p; (2) if $p \prec q$, then an edge is inserted linking node q to node p. These rules do not yield unique representations of posets as Hasse diagrams. In practice, one tries to draws diagrams minimizing the number of crossing edges, to simplify the visual display. By transitivity, $p \leq q$ holds (or $q \leq p$) in P, if and only if in the Hasse diagram there is a descending path linking the corresponding nodes; otherwise, p and q are incomparable. Given $p \in P$, the *up-set* of p is the set of all the elements $q \in P$ such that $p \leq q$. Dually, the *down-set* of p is the set of all the elements $q \in P$ such that $q \leq p$. Given two posets $(P_1, \leq_1)$ and $(P_2, \leq_2)$, an *order isomorphism* between P_1 and P_2 is a bijective correspondence $f(\cdot)$ which preserves order, that is such that

$$p \leq_1 q \Leftrightarrow f(p) \leq_2 f(q).$$

Furthermore, an order isomorphism between two finite posets preserves the cover relation.

Given $Q \subset P$, an *upper bound* of Q is an element $p \in P$ such that $q \leq p$ for any $q \in Q$. Let Q^* be the set of all upper bounds of Q. If $p^* \in Q^*$ exists such that $p^* \leq p$ for any $p \in Q^*$, then p^* is called the *least upper bound* (written $l.u.b.$) of Q. Dually, a *lower bound* of Q is an element $p \in P$ such that $p \leq q$ for any $q \in Q$. Let $\hat{Q}$ be the set of all lower bounds of Q. If $\hat{p} \in \hat{Q}$ exists such that $p \leq \hat{p}$ for any $p \in \hat{Q}$, then $\hat{p}$ is called the *greatest lower bound* (written $g.l.b.$) of Q. A partially ordered set P such that any pair of elements $\{p, q\} \subset P$ has both l.u.b. and g.l.b. is called a *lattice*. Usually, the l.u.b. and the g.l.b. of p and q are written $p \vee q$ and $p \wedge q$, respectively (also known as the *join* and the *meet* of p and q). In a finite lattice L, there are always a *top* element (written 1), such that $p \leq 1$ for any $p \in L$, and a *bottom* element (written 0), such that $0 \leq p$ for any $p \in L$. A partially ordered set where any pair of

[1] For sake of simplicity, in the following, elements of X partially ordered by $\leq$ will be referred directly as elements of P.

[2] As stated above, we say that p and q are comparable if either $p \leq q$ or $q \leq p$, thus $P \subseteq$ Comp(P).

elements has a g.l.b. (but not necessarily a l.u.b.) is called a *meet semilattice*. It is easy to see that a finite meet semilattice always has a bottom element.

Given their central role in the paper, we end this technical section giving some basic definitions about the concept of distance. Let V be a (nonempty) set and let $d(\cdot,\cdot)$ be a non-negative real function on $V \times V$. Function $d(\cdot,\cdot)$ is a *distance* (or a *metric*) on V, if it satisfies the following properties:

1. $d(a,b) \geq 0, \forall\, a,b \in V$.
2. $d(a,b) = 0 \Leftrightarrow a = b, \quad a,b \in V$.
3. $d(a,b) = d(b,a), \forall\, a,b \in V$.
4. $d(a,b) \leq d(a,c) + d(c,b), \forall\, a,b,c \in V$.

The last condition is known as *triangle inequality*. The ordered pair (V,d) is called a *metric space*.

4.3 Structural Dissimilarity Measure Between Posets

The definition of a measure of structural dissimilarity between finite posets is based upon the identification of a suitable way to compute the distance between partial orders. We thus begin the section discussing this preliminary issue.

4.3.1 Distance Between Finite Posets

Let P_1 and P_2 be two finite posets with n elements. We may compute a distance between them, introducing a metric $d(\cdot.\ \cdot)$ on the set Π_n of all labeled posets with n elements:

$$d \quad : \quad \Pi_n \times \Pi_n \mapsto \mathbb{R}^+ \quad : \quad (P_1,P_2) \to d(P_1,P_2).$$

Various metrics may be defined on Π_n. One choice in this context is to define $d(P_1,P_2)$ in terms of adding or deleting comparabilities, analogously to the so-called *editing distance* (Axenovich et al. 2008, Zeng et al. 2009) developed in graph theory. Let $(P, \leq_P)$ be a poset and let $a \| b$ in P. We indicate with $A_{a\,b}$ the addition of comparability $a \leq_P b$ to Comp(P). Similarly, if $s \leq_P t$, we indicate with $D_{s\,t}$ the deletion of this comparability from Comp(P). Clearly, P_1 may be turned into P_2 applying a sequence[3] S of operators $A_{..}$ and $D_{..}$.[4] The number of operators in S is called the *length* of S. We may now give the formal definition of $d(\cdot,\cdot)$

[3] The sequence is to be read from right to left.

[4] The dots "$_{..}$" stand for an unspecified pair of elements of the poset.

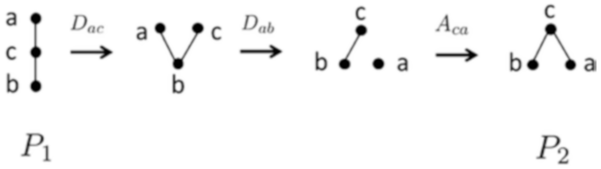


Fig. 4.1 Two posets with three elements, with connecting sequence $D_{ac} - D_{ab} - A_{ca}$

Definition 4.3.1.

The distance $d(P_1, P_2)$ between two posets P_1 and P_2 with n elements is defined as the length of a shortest sequence S of $A_{\cdot\cdot}$ and $D_{\cdot\cdot}$ operators turning P_1 into P_2 (or vice versa).

Some remarks to the above definition are in order: (i) the function $d(\cdot, \cdot)$ defined above is indeed a metric, as it will be proved later; (ii) in general, the sequence S of minimum length connecting P_1 and P_2 is not unique, since any permutation of S accomplishes the same task; (iii) while S turns a poset into another, a subsequence of S may not, since in general adding and deleting comparabilities may affect transitivity[5]; (iv) as shown in the following, given S, one can always find a permutation S^* of S such that, for any $1 \leq k \leq length(S)$, $S_{1:k}{}^*$ turns P_1 into a poset, where $S_{1:k}{}^*$ is the subsequence of S^* composed by the first k elements of S^*, counted from the right; (v) since any operator $A_{\cdot\cdot}$ or $D_{\cdot\cdot}$ adds or deletes just one comparability, it is clear that

$$d(P_1, P_2) = |\operatorname{Comp}(P_1) \setminus \operatorname{Comp}(P_2)| + |\operatorname{Comp}(P_2) \setminus \operatorname{Comp}(P_1)|.$$

It is also of interest to notice that $d(P_1, P_2)$ equals the city-block distance between the matrices representing P_1 and P_2, respectively (for a discussion on matrix representations of posets, see Patil and Taillie 2004). Figure 4.1 provides an example of two posets connected by a shortest sequence of additions and deletions of comparabilities (see also Brüggemann et al. 2001, for similar considerations).

The Poset of Posets with n **Elements.** An alternative way to interpret the distance between posets and the action of a minimum length connecting sequence employs the notion of *poset of posets*. This will also provide an effective way to visually display the distance concept. The set Π_n comprising all labeled posets on a set of n elements can be turned into a poset, defining the partial order relation $\leq_{\Pi_n}$ according to:

$$P_1 \leq_{\Pi_n} P_2 \Leftrightarrow \operatorname{Comp}(P_1) \subseteq \operatorname{Comp}(P_2).$$

[5] For example, suppose P is a poset and suppose a, b, and c constitute an antichain in P. Then, $A_{\mathrm{a}\,\mathrm{b}} A_{\mathrm{b}\,\mathrm{c}} P$ is not a poset, unless also $A_{\mathrm{a}\,\mathrm{c}}$ is applied to P.

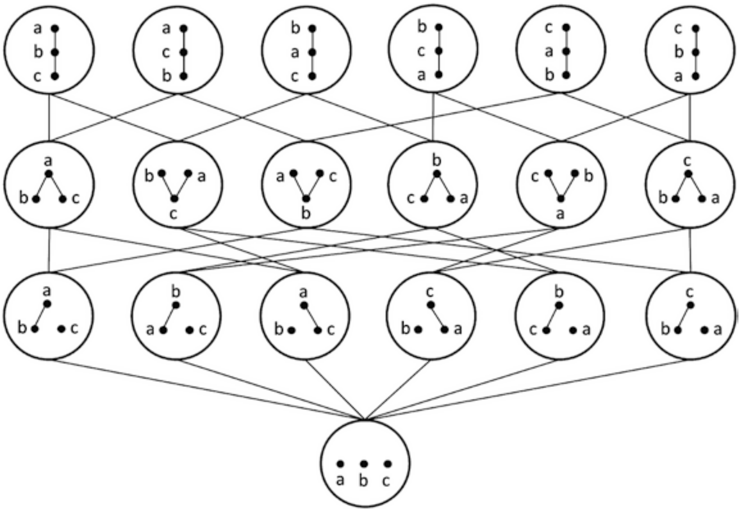


Fig. 4.2 The poset of posets with three elements

The set Π_n endowed with the partial order $\leq_{\Pi_n}$ is known as the *poset of posets of n elements*.[6] As a mathematical object, it has been widely studied (Brualdi et al. 1994). In particular, it may be proved that[7]:

1. Π_n is a meet semilattice, where the meet is given by ordinary set intersection $\cap$. Explicitly, $P_1 \wedge P_2$ is defined by

$$\mathrm{Comp}(P_1 \wedge P_2) = \mathrm{Comp}(P_1) \cap \mathrm{Comp}(P_2).$$

2. The bottom element of Π_n is the antichain on n elements.
3. Π_n has $n!$ maximal elements which are the linear orders on n elements.
4. Poset P_2 covers poset P_1 in Π_n if and only if $\mathrm{Comp}(P_1) \subset \mathrm{Comp}(P_2)$ and $|\mathrm{Comp}(P_2)| = |\mathrm{Comp}(P_1)| + 1$, that is if P_2 may be obtained from P_1 adding a single comparability to it.

From the above properties, it follows that the length of any chain from an element P of Π_n to the bottom of Π_n, that is the *height*[8] of P, equals $|\mathrm{Comp}(P)| - n$ and, more generally, that the absolute difference between the heights of two comparable posets in Π_n equals the absolute difference between the cardinalities of their comparability sets. Figure 4.2 provides an example, reproducing the Hasse diagram of Π_3.

[6] For the problem of determining the number of posets with n elements, see Schröder [2002].

[7] In the following, we write Π_n to mean the partially ordered set $(\Pi_n, \leq_{\Pi_n})$.

[8] The notion of *height* is well defined in Π_n, since the *Dedekind chain condition* holds in meet semilattices (see Davey and Priestley 2002).

As the following proposition shows, there is a close link between the distance between posets introduced in Definition 4.3.1 and the geodesic metric defined on the Hasse diagram of Π_n.

Proposition 4.3.1.

Let P_1 and P_2 be elements of Π_n. The distance $d(P_1,P_2)$ introduced in Definition 4.3.1 coincides with the length of a geodesic d^{gds} (P_1,P_2) between P_1 and P_2 in the Hasse diagram of Π_n.

Proof.

Let P_1 and P_2 be elements of Π_n and let Comp(P_1) and Comp(P_2) be their comparability sets. We have already noticed that

$$d(P_1,P_2) = |\,\mathrm{Comp}(P_1)\setminus\mathrm{Comp}(P_2)\,| + |\,\mathrm{Comp}(P_2)\setminus\mathrm{Comp}(P_1)\,|.$$

Now, consider P_1 and P_2 as elements of the Hasse diagram of Π_n and consider the path **p** (recall that $\wedge$ stands for the meet)

$$\mathbf{p}: P_1 \to P_1 \wedge P_2 \to P_2$$

composed of a descending part ($P_1 \longrightarrow P_1 \wedge P_2$) and an ascending part ($P_1 \wedge P_2 \longrightarrow P_2$). Since any step in **p** links two posets having comparability set whose cardinalities differ by one, we can conclude that:

1. the lengths of the descending part and the ascending part are, respectively, $|\mathrm{Comp}(P_1)\setminus\mathrm{Comp}(P_2)|$ and $|\mathrm{Comp}(P_2)\setminus\mathrm{Comp}(P_1)|$. So the length of **p** equals $d(P_1,P_2)$.
2. any other path between P_1 and P_2 in the Hasse diagram of Π_n must at least comprise $|\mathrm{Comp}(P_1)\setminus\mathrm{Comp}(P_2)|$ descending steps together with $|\mathrm{Comp}(P_2)\setminus\mathrm{Comp}(P_1)|$ ascending steps, thus **p** is a minimum length path, or a geodesic, between P_1 and P_2.

□

Therefore, we have proved that $d(\cdot,\cdot)$ is indeed a metric and that $d(P_1,P_2)$ coincides with the geodesic metric on the Hasse diagram of Π_n.

Since any path between P_1 and P_2 in the Hasse diagram of Π_n defines a sequence S of addition and deletion operators connecting the two posets, we have also proved remark (iv), after Definition 4.3.1. Notice also that, in general, there are several minimal length paths between two posets in Π_n.

Figure 4.3 embeds the example of Fig. 4.1 into Π_3. Linear order $P_1=(b<c<a)$ is connected to poset $P_2=(a<c, b<c, a\,||\,b)$ through their meet ($b<c$, $a\,||\,b$, $a\,||\,c$). The distance between the two posets is 3.

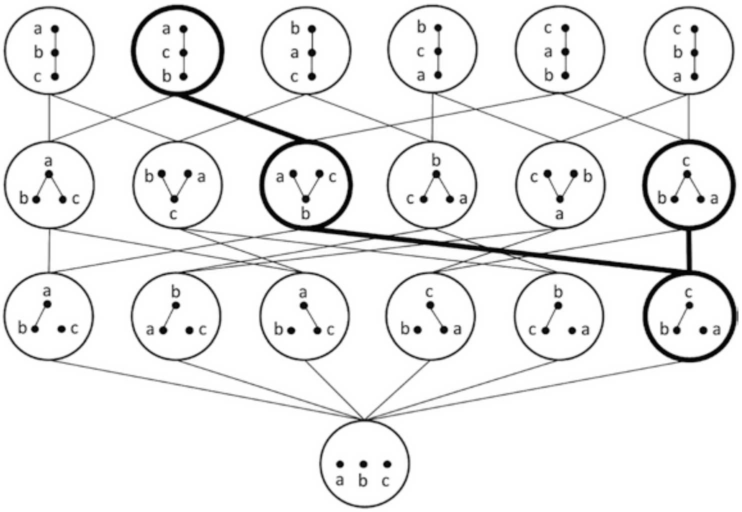


Fig. 4.3 A geodesic between two posets in Π_3

Remark.

The distance $d(\cdot, \cdot)$ may be interpreted as an absolute dissimilarity measure between finite posets. Dividing it by the maximum distance achievable in Π_n, i.e., by the diameter of Π_n, one gets a relative dissimilarity measure (a normalized metric) $d^*(\cdot, \cdot)$. The diameter of Π_n equals the distance between a linear order and its dual (i.e., the reversed linear order) that is $n(n-1)$. The diameter of Π_3 is 6, so in the example of Fig. 4.3, the relative dissimilarity measure between the two posets is $3/6 = 0.5$.

A Brief Digression. There is a simple but interesting connection between the above discussion and Kendall's τ (Kendall 1938), the widespread measure of rank correlation between two linear orders l_1 and l_2 on the same set of n elements. Kendall's τ between l_1 and l_2 may be defined as follows:

$$\tau(\ell_1, \ell_2) = 1 - \frac{4\Delta}{n(n-1)}$$

where Δ is the number of *discordant pairs*, that is the number of pairs of elements where the two linear orders disagree. A path linking in Π_n two linear orders with Δ discordant pairs has length[9] 2Δ, so that we may write

$$\tau(\ell_1, \ell_2) = 1 - 2d^*(\ell_1, \ell_2).$$

where $d^*(l_1, l_2)$ is the relative dissimilarity measure introduced above.

[9] In fact, one has to delete Δ comparabilities and add the corresponding reversed Δ comparabilities.

4.3.2 The Measure of Structural Dissimilarity Between Posets

Looking at the Hasse diagram depicted in Fig. 4.3, it is immediately checked that many of the elements of Π_3 have the same structure, i.e., they are identical up to order isomorphisms (that is, label permutations). In other words, they are topologically equivalent, if the labels are removed. For example, all the maximal elements of Π_3 are linear orders; similarly, all the elements with height 1 are isomorphic. Elements with height 2 split into two isomorphism classes, the first comprising posets with a top element and the second comprising posets with a bottom element. Informally speaking, the dissimilarity between isomorphic posets is due to label permutation only, while dissimilarity between non-isomorphic posets is also due to a structural difference in their relational patterns. It is precisely this *structural dissimilarity* that we want to measure. We thus look for a new distance function $d^{\mathrm{str}}(\cdot, \cdot)$ that captures the structural difference between two posets. Following (Fattore and Grassi 2012), $d^{\mathrm{str}}(\cdot, \cdot)$ is defined as follows:

Definition 4.3.2.

Let P_1 and P_2 be two posets with n elements and let $[P_1]$ and $[P_2]$ be their isomorphism classes. Then

$$d^{\mathrm{str}}(P_1, P_2) = \min_{P_1 \in [P_1], P_2 \in [P_2]} d(P_1, P_2).$$

In other words, the structural dissimilarity is measured as the distance between the isomorphism classes of P_1 and P_2, analogously to what is done in graph theory (Axenovich et al. 2008). As for the distance $d(\cdot, \cdot)$, we may re-state the above discussion involving the poset of posets Π_n.

Let us introduce in Π_n the equivalence relation $\sim$ defined by

$$P_1 \sim P_2 \Leftrightarrow P_1 \text{ is isomorphic to } P_2$$

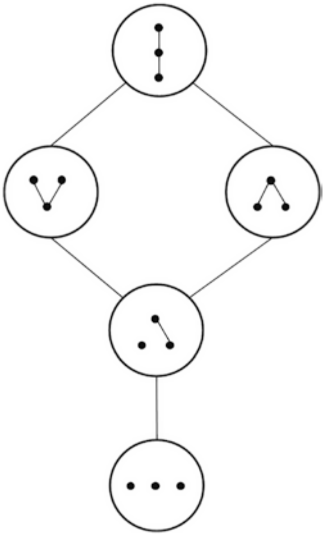
and consider the quotient set $\Pi_n/\sim$, whose elements are equivalence classes. $\Pi_n/\sim$ can be turned into a poset $(\Pi_n/\sim, \leq^*)$, defining a partial order $\leq^*$ as follows:

Definition 4.3.3.

Let $[P]$ and $[Q]$ be two elements of $\Pi_n/\sim$. We put $[P] \leq^* [Q]$ if and only if $P_1 \in [P]$ and $Q_1 \in [Q]$ exist such that $P_1 \leq_{\Pi_n} Q_1$ in Π_n.

Since elements of the same equivalence class are isomorphic, it follows that if $[P] \leq^* [Q]$, then for any $\hat{P} \in [P]$ there exist $\hat{Q} \in [Q]$ such that $\hat{P} \leq \hat{Q}$. Checking that $\leq^*$ satisfies the properties of reflexivity, antisymmetry, and transitivity is a routine task, so $\leq^*$ is a partial order relation. It is also easy to see that the poset $(\Pi_n/\sim, \leq^*)$ has a bottom element (the equivalence class of the antichain on n

Fig. 4.4 Hasse diagram of $(\Pi_3/\sim, \leq^*)$



elements) and a top element (the equivalence class of linear orders on n elements). Figure 4.4 depicts the Hasse diagram of $\Pi_3/\sim$.

The interest in $\Pi_n/\sim$ lies in the fact that it plays with respect to $d^{\text{str}}(\cdot, \cdot)$ the same role that Π_n plays with respect to $d(\cdot, \cdot)$, as the following proposition states.

Proposition 4.3.2.

Let P,Q $\in \Pi_n$. *Then* d^{str} *(P,Q)* = d^{gds} *([P].[Q]), i.e., the structural distance between two posets with n elements equals the geodesic distance between the corresponding equivalence classes in the Hasse diagram of* $\Pi_n/\sim$.

Proof.

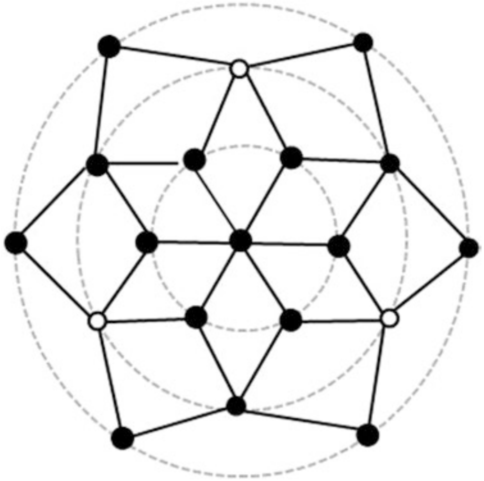
Let $d^{\text{str}}(P, Q)$ be the structural distance between two posets. If $P_1 \sim P$, then there exists $Q_1 \sim Q$ (which depends on P_1) such that $d^{\text{str}}(P, Q) = d(P_1, Q_1)$. Therefore, the structural distance between P and Q may always be computed choosing a representative P_1 of $[P]$ and the corresponding representative Q_1 of $[Q]$ and computing their distance or, equivalently, computing $d^{\text{gds}}(P_1, Q_1)$ in Π_n. Now, suppose elements of Π_n may be grouped in m equivalence classes $C_1, \ldots, C_m$. Chosen a representative P_k in class C_k, a set of representatives $P_i \in C_i$ $(i = 1, \ldots, m)$ may be consequently identified such that for any s, t:

$$d^{\text{str}}([P_s],[P_t]) = d(P_s, P_t) = d^{\text{gds}}(P_s, P_t).$$

From the definition of $\leq^*$, the poset of posets $P_1, \ldots P_m$ is order isomorphic to $\Pi_n/\sim$, so that the cover relation is preserved and we have

$$d^{\text{gds}}(P_s, P_t) = d^{\text{gds}}(C_s, C_t)$$

Fig. 4.5 An alternative representation of Π_3



(where the distance on the left is computed in Π_n and the distance on the right is computed in $\Pi_n/\sim$). Thus

$$d^{\text{str}}([P_s],[P_t]) = d^{\text{gds}}(C_s, C_t).$$

□

To illustrate the idea behind this result, in Fig. 4.5 we give an alternative (and informal) representation of poset Π_3, where equivalence classes and symmetries should be evident. The picture must be read as follows:

1. The node at the center of the picture represents the antichain on three elements, which is the unique representative of its equivalence class.
2. Gray circumferences represent levels of Π_3, that is, points on the same circumference are incomparable, but share the same number of comparabilities and have the same height in the original Hasse diagram. Starting from the center, the number of comparabilities increases by one at any circumference.
3. On each circumference, nodes with the same color (black or white) belong to the same equivalence class (so we see that there is just one equivalence class on the first—the center of the diagram—the second and the fourth circumference, but there are two, represented by white and black nodes, respectively, on the third circumference).
4. Edges represent the partial order relation. They are implicitly directed from outwards to the center of the diagram.

The picture shows that along each ray from the center, one may find a subposet $\Pi_3{}^*$ isomorphic to $\Pi_3/\sim$, composed of equivalence class representatives. By visual inspection, the geodesic distance between two equivalence classes in $\Pi_3/\sim$ is realized by the geodesic distance between the corresponding representatives in the selected $\Pi_3{}^*$.

Remark.

We conclude this paragraph with a final observation. Given the structural distance $d^{\mathrm{str}}(\cdot,\cdot)$, we may decompose the absolute dissimilarity measure as follows:

$$d(\cdot,\cdot) = d^{\mathrm{str}}(\cdot,\cdot) + d^{\mathrm{perm}}(\cdot,\cdot)$$

where $d^{\mathrm{perm}}(\cdot,\cdot)$ is a residual term. Since by definition $d^{\mathrm{str}}(\cdot,\cdot) \leq d(\cdot,\cdot)$, $d^{\mathrm{perm}}(\cdot,\cdot)$ is always non-negative and may be interpreted as the contribution to $d(\cdot,\cdot)$ due to label permutation. Thus we have decomposed the absolute dissimilarity measure between two posets as the sum of two components, accounting for structural and nonstructural changes, respectively (for other discussions on labeling and permutations in posets, see Sœrensen et al. 2005). In this chapter, our main interest is in structural difference between posets, motivated by the theoretical need of identifying a synthetic way to compare partial order patterns. But in concrete applications, also the permutation component may be of interest. It gives insights on the "role exchange" between labels, which may represent countries, market players, individuals, alternatives, or any other kind of interesting entities, whose role in the posets has changed. A high value of this component relative to the absolute dissimilarity suggests that a re-shuffling among labels has occurred, which may be subsequently investigated in greater details, to specify and interpret it. Both the structural and the permutation components are in fact high-level indicators of the way a poset have changed and cannot provide details on the way single elements have modified their "relational position" within the partial order. Further analysis is then needed, if this "micro" level is of interest. We suggest that the structural-permutation decomposition proposed in this paper could be applied not only to the overall poset, but also to some interesting subposets (for example, the subposet of the elements that cover or are covered by a specific element of the poset). Therefore, one could investigate the way the partial order changes at different "observation scales," obtaining more information on the global modifications occurring in it.

4.4 Examples

In this paragraph, we briefly present two examples of structural dissimilarity computations. Due to the number of elements in the posets involved, it is not possible to depict the Hasse diagrams of the corresponding posets of posets. Numerical computations have been performed running a simple version of the simulated annealing algorithm (van Laarhoven and Aart 1987), implemented by an *R* script. The determination of the structural difference, in fact, requires exploring the space of equivalence classes (or, more precisely, the orbits of the automorphism group) of a poset, minimizing a distance function on it. This is not an easy task, since such a space is discrete and discrete optimization algorithms are more complex than algorithms working on continuous spaces. The main problem is to avoid "being captured" in local minima, in a space whose topology may be very complicated and is not known

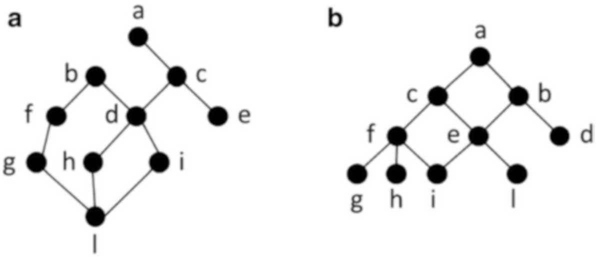


Fig. 4.6 Two posets with ten elements

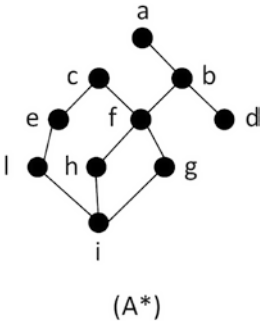


Fig. 4.7 Representative of the equivalence class of poset **A** of Fig. 4.6, estimated as closest to poset **B**

a priori. Therefore, global optimization algorithms (in contrast to the so-called greedy algorithms) are preferable. Simulated annealing is a general purpose algorithm of this kind, not specifically designed for optimizing over equivalence classes of graphs; still it suffices for our examples. Surely, developing ad hoc algorithms for the problem at hand could greatly improve estimation precision and computational performances.

First example pertains to the posets whose Hasse diagrams are reported in Fig. 4.6. They comprise ten elements and are clearly non-isomorphic.

The overall difference between the two Hasse diagrams is 21. Searching among the set of permutations of labels of poset A, the simulated annealing algorithm estimated a minimum distance between the isomorphism classes of the two posets of 11. Thus the permutation component is estimated as 10 and we see that the structural and the permutation components have similar weight, in the decomposition of the overall difference. On a personal computer equipped with an Intel$^{®;}$ Core^{TM}2 Duo CPU E8400 3.00GHz ×2, 1.9 GiB RAM and Linux Ubuntu 12–10 64 bit, the computations took 64.4 s. Given the number of elements, the maximal possible structural dissimilarity is $10 \times (10-1) = 90$. Therefore, the actual structural dissimilarity is 0.12 of the maximum achievable. Figure 4.7 reports the element **A*** of the isomorphism class of poset **A** which is estimated as the closest to poset **B**.

The second example pertains to the posets represented in Fig. 4.8. They comprise 16 nodes and have been extracted from a paper of Patil and Taillie [2004]. In the original paper, each node in the Hasse diagrams represent a Country of Western Europe (poset A) or Latin America (poset B) scored on three variables pertaining to

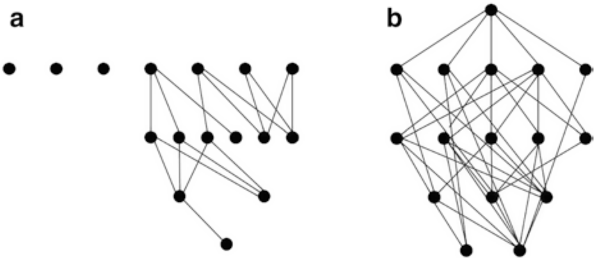


Fig. 4.8 Unlabeled posets from the paper by Patil and Taillie

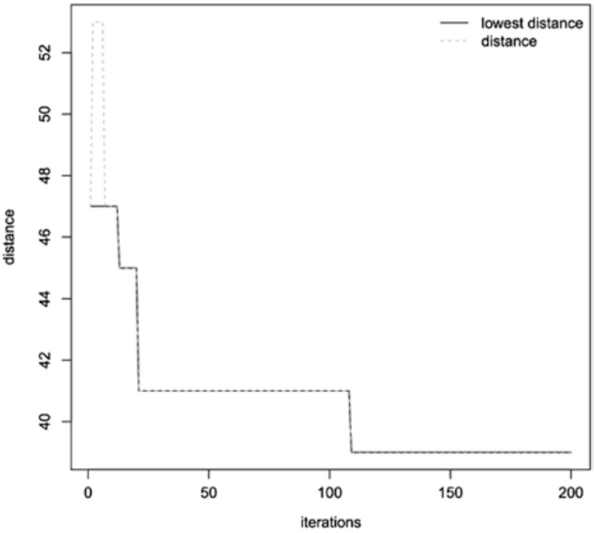


Fig. 4.9 Convergence of the simulated annealing algorithm, for posets represented in Fig. 4.8

environmental quality of land, air, and water. Labels have been removed, since here we are just interested in structural differences.[10] Details about the data and their interpretation may be found in the cited reference. Using the same software procedure applied in the first example, the structural dissimilarity between the two posets has been estimated as 39, a fraction of 0.16 of the maximum achievable ($16 \times (16-1) = 240$). In this example, computations took 6815.8 s, on the same personal computer described above. To give an idea of the optimization process, Fig. 4.9 reports the convergence of the simulated annealing algorithm. As the number of iterations grows, the lowest computed distance between isomorphism classes decreases (while the actual distance may locally increase), until a stable result is achieved.

[10] In this example, the compared posets are defined over different label sets, so we can only focus on the distance between equivalence classes.

4.5 Conclusion

In this paper, we have proposed a way to measure structural and nonstructural dissimilarity between two posets on n elements. The main feature of the proposal is that it is "global," in that it attempts to capture and measure differences in the relational patterns of the compared posets. Its main disadvantage is that its computation is based on numerical optimization, since there is no explicit formula to compute it. The computational problem is not trivial, since optimizing efficiently on the discrete space of all the permutations of n labels requires ad hoc algorithms (Gao et al. 2010). The examples provided have been worked out using simulated annealing, to overcome the limitations of greedy procedures. The basic idea presented in the paper may be extended in many directions. For example, one could consider metrics other than the minimum length of sequences of additions/deletions of comparabilities. It would also be of interest to explore the connection between structural dissimilarity measures and more "classical" indicators, based on changes in number of comparabilities, dimension, width, height, etc. More important, the proposed approach must be applied to real cases, to check it and to verify its effectiveness in extracting new and useful information for real applications.

References

Annoni P, Brüggemann R, Saltelli A (2011) Partial Order investigation of multiple indicator systems using variance - based sensitivity analysis. Environ Model Softw 26:950–958

Axenovich M, Kézdy A, Martin R (2008) On the editing distance of graphs. J Graph Theory 58(2):123–138

Brualdi RA, Junga HC, Trotter WT Jr (1994) On the posets of all posets on n elements. Discrete Appl Math 50(2):111–123

Brüggemann R, Bartel HG (1999) A theoretical concept to rank environmentally significant chemicals. J Chem Inf Comput Sci 39:211–217

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems - introduction to partial order applications. Springer, New York

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (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

Davey BA, Priestley BH (2002) Introduction to lattices and order, Cambridge University Press, Cambridge

Fattore M, Grassi R (2012) Measuring dynamics and structural change of time-dependent socio-economic networks, Qual Quant May 2013

Gao X, Xiao B, Tao D, Li X (2010) A survey of graph edit distance. Pattern Anal Appl 13:113–129

Harary F (1969) Graph theory. Perseus Books, Cambridge

Kendall M (1938) A new measure of rank correlation. Biometrika 30(1–2):81–89

Klein DJ (1995) Similarity and dissimilarity in posets. J Math Chem 18(2):321–348

Monjardet B (1981) Metrics on partially ordered sets - a survey. Discrete Math 35:173–184

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

Schröder BSW (2002) Ordered sets. Birkhäuser, Basel

Sœrensen PB, Brüggemann R, Thomsen M, Lerche D (2005) Applications of multidimensional rank-correlation. Match Comm Math Comput Chem 54:643–670

van Laarhoven PJM, Aart EHL (1987) Simulated annealing: theory and applications. Kluwer, Dordecht

Voigt K, Scherb H, Brüggemann R, Schramm KW (2011) Application of the PyHasse program features: sensitivity, similarity, and separability for environmental health data. Statistica & Applicazioni, Special Issue, 155–168

Zeng Z, Tung AKH, Wang J, Feng J, Zhou L (2009) Comparing stars: on approximating graph edit distance. Proc VLDB Endow 2(1):25–36

Chapter 5
Quantifying Complexity of Partially Ordered Sets

Guillermo Restrepo

Abstract We discuss two complexity indicators reported in the literature for partially ordered sets (posets), the first one based on linear extensions and the second one on incomparabilities. Later, we introduce a novel indicator that combines comparabilities and incomparabilities with a Shannon's entropy approach. The possible values the novel complexity indicator can take are related to the partitions of the number of order relationships through Young diagrams. Upper and lower bounds of the novel indicator are determined and analysed to yield a normalised complexity indicator. As an example of application, the complexity is calculated for the ordering of countries based on their performance in chemical research. Finally, another complexity indicator is outlined, which is based on comparabilities, incomparabilities, and equivalences.

5.1 Introduction

Order relationships are often used in chemistry and environmental sciences, as can be recognised in statements as "less reactive than", "less polluting than", etc. If objects are included in the above claims, then they turn into "x is less reactive than y", "x is less polluting than y", etc. Now, if "is less polluting (reactive) than" is symbolised by $\preccurlyeq$,[1] then we obtain $x \preccurlyeq y$. The binary relation $\preccurlyeq$ satisfies three properties (Trotter 1992), i.e. reflexivity, antisymmetry and transitivity, which means that $x \preccurlyeq x$, that if $x \preccurlyeq y$ and $y \preccurlyeq x$, then $x = y$, and that if $x \preccurlyeq y$ and $y \preccurlyeq z$, then $x \preccurlyeq z$, respectively, with x, y, z being objects to order. If the objects are grouped into a set X, then the

[1] We use $\preccurlyeq$ instead of $\prec$ to allow the relation between one object and itself.

G. Restrepo (✉)
Interdisciplinary Research Institute, Universidad de Pamplona, Bogotá, Colombia

Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia
e-mail: grestrepo@unipamplona.edu.do; guillermorestrepo@gmail.com

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_5, © Springer Science+Business Media New York 2014

couple $(X,\preccurlyeq)$ is called a partially ordered set, abbreviated as *poset* (Trotter 1992). It is called "partially" because it might happen that not every couple of objects in X satisfies the relation $\preccurlyeq$. In general, for any couple $x, y \in X$, one of these possibilities holds: (1) $x \preccurlyeq y$, $y \preccurlyeq x$; (2) $x \not\preccurlyeq y$, $y \not\preccurlyeq x$; in the first case we say that x and y are *comparable* and we write, in general, $x \perp y$ and in the second case we state that x and y are *incomparable* $(x \parallel y)$ (Brüggemann and Bartel 1999). For the particular case where $x \preccurlyeq y$ and $y \preccurlyeq x$, we say that x and y are *equivalent* $(x \sim y)$. This kind of relation is only explored in Sect. 5.2.3. If every pair of objects of X is comparable, the poset is called a *chain* and $\preccurlyeq$ a *linear order* (or *total order*). In contrast, an *antichain* is made of incomparable objects (Brüggemann and Bartel 1999). A poset may contain several chains in it. An incomparable object is a chain as well as an antichain; if the chains of a poset are collected, the chain containing more objects than all the other chains is called a *maximum chain*. Likewise, a *maximum antichain* is the antichain of a given poset having more objects than the other antichains (Trotter 1992).

The derived idea from the above generalisation is that it is not always possible to end up with a total order. A particular popular kind of total order is a ranking, where there is only one first, one second, one third, and so on. Hence, given a set of objects to order, it might occur (according to the ordering constraints) that some couples are equivalent and incomparable and others comparable. Brüggemann, Klein, and their co-workers have shown how these situations abound in daily life and in chemistry and environmental sciences (Brüggemann and Patil 2011; Restrepo et al. 2011).

Once $(X,\preccurlyeq)$ is given, a graphical representation can be generated, namely the Hasse diagram (Halfon 2006; Neggers and Kim 1998). This diagram is a directed graph $\mathcal{H} = (X,E)$ with $(x,y) \in E$ iff $x, y \in X$, $x \preccurlyeq y$, and there exists no $z \in X$ with $x \preccurlyeq z \preccurlyeq y$ (symbolised as $x \preccurlyeq: y$ and called a *cover relation* of y over x) (Trotter 1992); E is called the set of directed edges of $\mathcal{H}$. By convention, $\mathcal{H}$ is drawn in the Euclidean plane whose horizontal/vertical coordinate system requires that the vertical coordinate of $y \in X$ be larger than the one of $x \in X$ iff $x \preccurlyeq: y$ (Restrepo and Brüggemann 2008).

If $X = \{a,b,c,d,e\}$ and the following order relations hold: $a \preccurlyeq: b$, $c \preccurlyeq: b$, $c \preccurlyeq: d$, $b \preccurlyeq: e$, and $d \preccurlyeq: e$, then the corresponding Hasse diagram is shown in Fig. 5.1a. For the ensuing discussion it is important to define linear extensions and intersection of posets. A *linear extension* of the poset $(X,\preccurlyeq)$ is a total order obtained from the poset such that it preserves the order relations contained in $(X,\preccurlyeq)$ (Trotter 1992). The corresponding linear extensions of the Hasse diagram in Fig. 5.1a are shown in Fig. 5.1b.

In spite of the simplicity gained when drawing cover relations rather than comparabilities, Hasse diagrams may become rather entangled as the number of elements and cover relations grows. For practical reasons, mainly related to the readability of a Hasse diagram, it is important to develop indicators of such readability, which may be associated with the complexity of the diagram.[2] Here we

[2] The fact of having entangled diagrams does imply that the order information of the poset is hidden. In fact, there are different approaches to extract information from posets, as discussed by Brüggemann and Patil (see Brüggemann and Patil 2011, pp. 36–38).

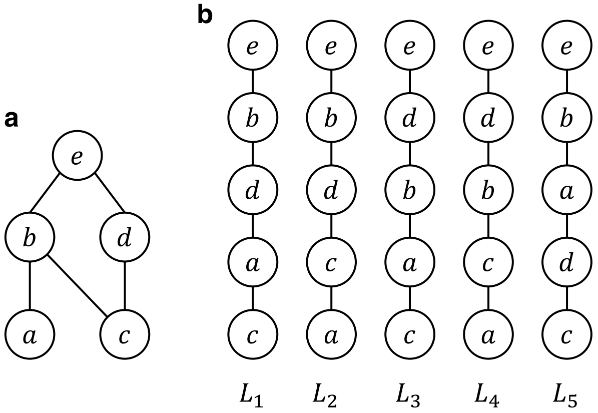


Fig. 5.1 (**a**) Hasse diagram and (**b**) linear extensions of the poset $(\{a,b,c,d,e\},\{a \preccurlyeq: b, c \preccurlyeq: b, c \preccurlyeq: d, b \preccurlyeq: e, d \preccurlyeq: e\})$

discuss some approaches to treating posetic complexity and moot a novel complexity indicator, which is further generalised at the end of the chapter.

5.2 Complexity of Posets

Trotter and Bogart, in 1976, related the complexity of a poset to its dimension (Trotter and Bogart 1976). The dimension of a poset was defined by Dushnik and Miller in 1941 as the minimum number of linear extensions of the poset whose intersection[3] yields the poset. Hence, for the poset $(X,\preccurlyeq)$ shown in Fig. 5.1a, its dimension is 2 since $(X,\preccurlyeq)=L_4 \cap L_5$. Trotter and Borgart gave an interesting justification of dimension as a posetic complexity indicator: "Suppose each of the finite number of observers expresses his individual opinion on the relative merits of a finite set of options by ranking the options in a linear order. A partial ordering on the options is obtained by ranking option x higher than option y when all observers have agreed that x is preferred to y. Conversely, the dimension of a partial order indicates the minimum number of observers necessary to produce the given partial order as a statement of those preferences on which the observers agree unanimously" (Trotter and Bogart 1976). Trotter and Borgat defined two additional complexity indicators, namely interval dimension and semi-order dimension, which are based on the original dimension definition and require the introduction of functions from the poset to the real numbers (Trotter and Bogart 1976). Another dimension complexity indicator

[3] Given the posets $(X,\preccurlyeq')$ and $(X,\preccurlyeq'')$, their intersection yields the poset $(X,\preccurlyeq)$ where $\preccurlyeq=\{(x,y)|x \preccurlyeq' y \wedge x \preccurlyeq'' y; x,y \in X\}$.

is the greedy dimension (Kierstead and Trotter 1985), which follows the same dimension principle explained above but with restrictions on the kind of linear extensions to use, called greedy linear extensions. As explained by Trotter, greedy dimension arose as a solution to the scheduling problem known as the jump number problem: "An ordered set P[4] represents a set of tasks to be performed on a single processor. If $x<y$ in P, then x must be performed before y. An admissible schedule is then a linear extension of P. Suppose that a set-up cost is paid for each pair $x, y \in P$ with x incomparable to y in P and x and y occurring consecutively in the linear extension. Find a linear extension L of P which minimizes the number of consecutive pairs of L which are incomparable in P" (Kierstead and Trotter 1985). The minimum number of those greedy extensions able to reproduce the poset by their intersection is called the greedy dimension of the poset. Some other results on dimension theory are found in Kelly and Trotter (1982) and West (1985).

Although complexity associated with posetic dimension has generated a wealth of studies, its calculation is a difficult task, to the extent that it is an NP-hard problem[5] (Yannakakis 1982; Yáñez and Montero 1999).

Another posetic complexity indicator was developed in 2000 by Luther et al., who developed a heuristic indicator, defined as follows:

$$\mathrm{J} = -\mathrm{D} \times " \, C \, / " \, X$$

where $T(s)$ is given by

$$T(s) := \begin{cases} 0 & 0 \le s \le \beta, \\ \dfrac{s-\beta}{\alpha-\beta} & \beta \le s \le \alpha, \\ \dfrac{1-s}{1-\alpha} & \alpha \le s, \end{cases}$$

with

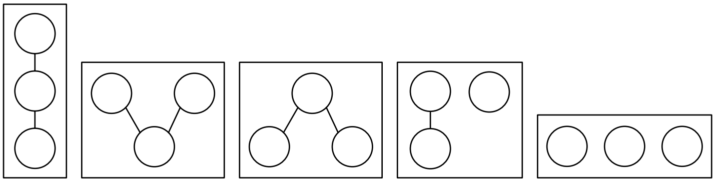
$$s := \frac{\mathit{number\ of\ incomparabilities}}{\mathit{number\ of\ objects} \times (\mathit{number\ of\ objects} - 1)/2}$$

where α and β are steering parameters, which are set by the authors as $\alpha = 0.8$ and $\beta = 0.3$. To illustrate this method, let us consider the unlabelled posets of a set of three objects,[6] the corresponding values, along with the posets, are depicted in Fig. 5.2.

[4] Here, $P = (X, \preccurlyeq)$.

[5] Yannakakis found that for certain posets it is NP-hard to decide if their dimension is equal or lower than a particular natural number (see Yannakakis 1982).

[6] We take unlabelled posets since what is important in the kind of complexity we are considering is the connectivity among the objects on the poset rather than their identity.



$s = 0$	$s = 0.33$	$s = 0.33$	$s = 0.67$	$s = 1$
$T(s) = 0$	$T(s) = 0.07$	$T(s) = 0.07$	$T(s) = 0.73$	$T(s) = 0$
$C = 0$	$C = 0.2$	$C = 0.2$	$C = 2.2$	$C = 0$
$p_\perp = 3/3$	$p_\perp = 2/3$	$p_\perp = 2/3$	$p_\perp = 1/3$	$p_\perp = 0/3$
$p_\parallel = 0/3$	$p_\parallel = 1/3$	$p_\parallel = 1/3$	$p_\parallel = 2/3$	$p_\parallel = 3/3$
$H = 0$	$H = 0.9183$	$H = 0.9183$	$H = 0.9183$	$H = 0$

Fig. 5.2 Unlabelled posets of three objects and their respective values s and $T(s)$ needed to calculate their complexity C according to Luther et al.'s method. Values $p_\perp$, $p_\parallel$, and H are explained in Sect. 5.2.1

The complexity indicator C captures the intuitive idea that a maximum chain and a maximum antichain are not complex posets, while posets holding a mixture of comparabilities and incomparabilities are more complex. However, the method has a subjectivity, namely the setting of α and β.

In the current chapter, we develop a novel posetic complexity indicator following a Shannon's entropy approach based on the number and distribution of order relationships in the poset.

5.2.1 An Entropic Posetic Complexity Indicator

If the system whose complexity is going to be calculated can be described as a discrete random variable with possible values $\{A_1, A_2, \ldots, A_n\}$ and associated probabilities $\{p_1, p_2, \ldots, p_n\}$, then Shannon's entropy (Shannon 1948) (complexity) of the system is $H = -\sum p_i \log_2 p_i$.

The random variable we define on a poset is given by the relationships among the elements of the poset and its values are the number of comparabilities and of incomparabilities.

Definition 1 Let $(X, \preccurlyeq)$ be a poset and let $R \subset X \times X$ be the possible relationships between couples of objects to order. Let $\perp \subseteq R$ and $\parallel \subseteq R$ be the comparabilities and incomparabilities of X, respectively, such that $R = \perp \cup \parallel$. As $|R| = N(N-1)/2$ is the cardinality of R, with $|X| = N$, we define $p_\perp = |\perp|/|R|$ and $p_\parallel = |\parallel|/|R|$ as the probabilities of having a comparability and an incomparability in $(X, \preccurlyeq)$, respectively. We say that $H(X, \preccurlyeq) = -\sum p_i \log_2 p_i$, with $i \in \{\perp, \parallel\}$, is the *complexity of the poset* $(X, \preccurlyeq)$.

For the posets of Fig. 5.2, the respective probabilities and complexities are shown at the bottom of the same figure, where the posets with maximum chain and

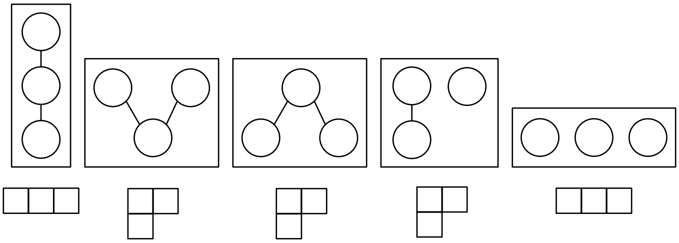


Fig. 5.3 Unlabelled posets of three elements, considering two relationships, and their associated Young diagrams for $|R|$ partitioned into one or two parts

maximum antichain have lowest complexity while all the other posets have maximum complexity. This is so since the distribution of relationships for those with maximum complexity is the same, i.e. 2 and 1, regardless of if they are comparabilities or incomparabilities. Note that for the maximum chain the number of comparabilities is 3 even if the Hasse diagram depicts only 2; this occurs because the diagram shows cover relationships rather than comparabilities.

If both complexity indicators, i.e. C and H, are normalised, then they yield similar results for the posets in Fig. 5.2, except for the one with one comparability and two incomparabilities. In Luther et al.'s indicator, this poset is the most complex one, while in our methodology it has the same complexity as the others that are neither maximum chain nor maximum antichain. This occurs because Luther et al.'s method gives more importance to incomparabilities than to comparabilities, which, as mentioned, is a subjectivity of the method. In our case, both comparabilities and incomparabilities are treated with the same importance.

Some properties of H can be explored on the basis of its relationship with integer partitions and therefore with Young diagrams (Andrews and Eriksson 2004). A Young diagram is a graphical representation of a partition; it is made up of a two-dimensional arrangement of boxes where the kth row has the same number of boxes as the kth term in the partition. If the partition of the integer n is $a+b+\ldots+c$, for a list $a, b, \ldots c$, of r positive integers in such a way that $a \geq b \geq \ldots \geq c$, then the diagram is the arrangement of n boxes in r rows (Andrews and Eriksson 2004).

As R is always an integer and $\perp$ and $\parallel$ partition it, then $|R| = |\perp| + |\parallel|$. Hence, the associated Young diagrams for the posets shown in Fig. 5.2 are depicted in Fig. 5.3. The Young diagrams here described always have two rows, one referring to the number of comparabilities and the other to the incomparabilities. The top row always corresponds to the relation with more cases. Hence, if there are more comparabilities than incomparabilities, then the top row refers to comparabilities. If the incomparabilities are highest, the top row corresponds to incomparabilities. In the case of having the same number of comparabilities and incomparabilities, then there is no distinction as to which relation is at the top or at the bottom.

Since our complexity indicator always requires R to be partitioned into comparabilities and incomparabilities and as extreme cases one has either $|R|$ comparabilities or $|R|$ incomparabilities, then the number of possible Young diagrams for unlabelled

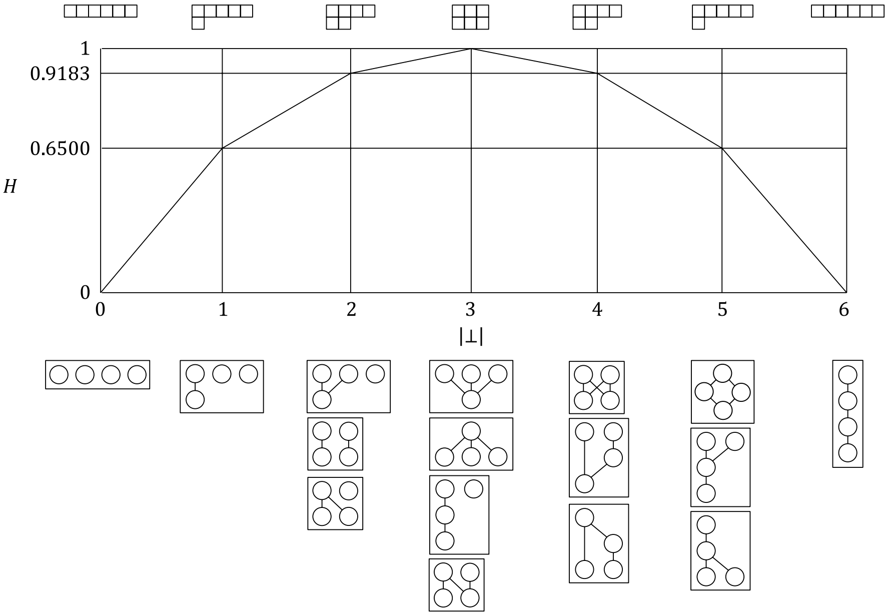


Fig. 5.4 Entropy values H of unlabelled posets of four objects, which are ordered according to their number of comparabilities $|\perp|$. At the top of the plot, the corresponding Young diagram for each set of equivalent posets is depicted

posets of $|R|$ relations is $\lfloor(|R|/2)+1\rfloor$,[7] which is the number of partitions of $|R|$ into one or two parts (Hardy 1920). This implies that there are always $(|R|/2)+1$ different values of H for a given $|R|$, which is the same as there are always $(N(N-1)/4)+1$ different values of H for the set of all unlabelled posets on a set X of cardinality N.

Proposition 1 For an unlabelled poset $(X,\preccurlyeq)$, with $|X|=N$, there are $(N(N-1)/4)+1$ different values of H (Definition 1).

Proof It is already stated in the previous paragraph□

For all 16 posets on a set of four objects, their respective complexities and Young diagrams are shown in Fig. 5.4.

Proposition 1 is important as it states that just by knowing the number of objects to order, one knows the possible complexity values the different unlabelled posets can take.

Now the question that arises is: having calculated the complexity of a poset is it possible to state that the poset is complex or not? The flat answer is no, for the entropic complexity calculated is not an absolute indicator. The problem can be overcome if one knows the total number of unlabelled posets for the N in question, i.e. the number of posets for N objects. It turns out that knowing the total number of

[7] Note that $\lfloor x \rfloor$ is the floor function that maps a real number to the largest previous following integer.

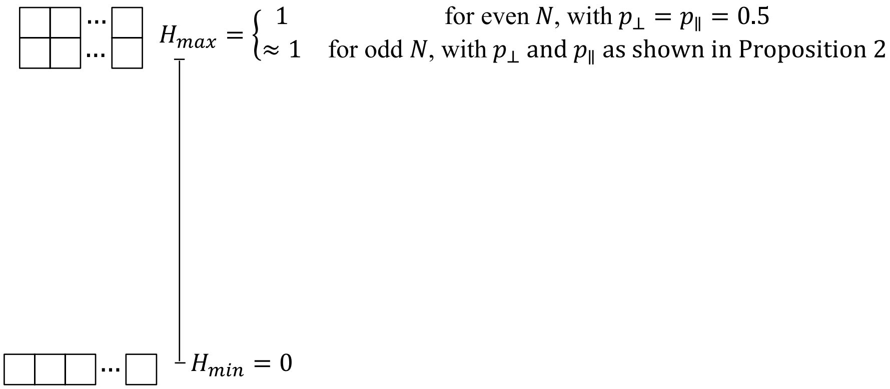

$H_{max} = \begin{cases} 1 & \text{for even } N\text{, with } p_\perp = p_\| = 0.5 \\ \approx 1 & \text{for odd } N\text{, with } p_\perp \text{ and } p_\| \text{ as shown in Proposition 2} \end{cases}$
$H_{min} = 0$

Fig. 5.5 Upper $H_{\max}$ and lower $H_{\min}$ bounds of H, with their respective Young diagrams. The *line* corresponds to the real scale bounded by $H_{\max}$ and $H_{\min}$

unlabelled posets of a given N is an open question in mathematics[8] (McKay and Brinkmann 2002). Fortunately there is a way to overcome the problem of counting posets, namely through knowing the upper and lower bounds of H. Hence, if we know the maximum and minimum values H can take, then the values of H make more sense, as a bounded scale is given for H and then one can know whether the poset and its complexity are close to the H upper bound, in which case one can claim that the poset is complex to a great extent. Or one can state that the poset is not that complex if its complexity is close to the lower bound.

Let us first consider the lower bound of H. According to the Young diagram's representation of the partition of N induced by $\perp$ and $\|$, the minimum value H can occur when there is either only comparabilities or only incomparabilities, i.e. $p_\perp = 1$ and $p_\| = 0$, or $p_\perp = 0$ and $p_\| = 1$, respectively. In that case, the Young diagram is made of a single row containing N boxes. The complexity associated with such a case is $H_{\min} = 0$, which is the lower bound for H.

The upper bound of H occurs when the number of comparabilities and incomparabilities is equal or almost equal, i.e. when $p_\perp = p_\| \cong 0.5$, which implies a Young diagram with two rows evenly or close to evenly populated.

In general, we have H values ranging between the real scale depicted in Fig. 5.5.

The above findings can be formalised as follows:

Proposition 2 For an unlabelled poset $(X, \preccurlyeq)$, with $|X| = N$, the upper ($H_{\max}$) and lower ($H_{\min}$) bounds of H (Definition 1) are given by $H_{\min} = 0$ and [9]

[8] Counting the number of unlabelled posets for a given N is a matter of current research in order theory. McKay and Brinkmann, in 2002, developed an algorithm to count this number up to $N = 16$ (see McKay and Brinkmann 2002). Some of the results they found are that for $N = 4$, there are 16 posets; 16,999 for $N = 8$; 1,104,891,746 for $N = 12$; and 4,483,130,665,195,080 for $N = 16$.

[9] $\lceil x \rceil$ is the ceiling function that maps a real number to the smallest following integer.

$$H_{\max} = \begin{cases} 1 & \mathit{for\,even\,N, with\,} p_{\perp} = p_{\|} = 0.5 \\ -\sum p_i \log_2 p_i & \mathit{for\,odd\,N, with\,} p_{\perp} \mathit{\,and\,} p_{\|} \mathit{\,as\,follows} \end{cases}$$

$$p_{\perp} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}} \mathit{\,and\,} p_{\|} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}}$$

or

$$p_{\perp} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}} \mathit{\,and\,} p_{\|} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}}$$

Proof In the previous paragraph proof was given that $H_{\min} = 0$. The upper bound of H is attained when N is partitioned into two parts $p_{\perp}$ and $p_{\|}$ of the same or almost the same cardinality. It is the same when N is even, in which case the total number of possible relationships $(N(N-1))/2$ is evenly divided into two parts of cardinality $(N(N-1))/4$. Hence, $p_{\perp} = p_{\|} = \frac{(N(N-1))/4}{(N(N-1))/2} = 0.5$; therefore $H = 1$. If N is an odd integer, then $(N(N-1))/4$ is not an integer number and the most homogeneous distribution of comparabilities and incomparabilities is given when the absolute value of $(|\perp| - |\|\,|) = 1$. In that case the respective cardinalities are given by

$$\left\lfloor \frac{N(N-1)}{4} \right\rfloor \mathit{and} \left\lceil \frac{N(N-1)}{4} \right\rceil$$

or

$$\left\lceil \frac{N(N-1)}{4} \right\rceil \mathit{and} \left\lfloor \frac{N(N-1)}{4} \right\rfloor.$$

Hence,

$$P_{\perp} = \frac{\left\lfloor \frac{N(N-1)}{4} \right\rfloor}{\frac{N(N-1)}{2}} \mathit{\,and\,} p_{\|} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}}$$

or

$$p_{\perp} = \frac{\left\lceil \frac{N(N-1)}{4} \right\rceil}{\frac{N(N-1)}{2}} \; and \; p_{\shortparallel} = \frac{\left\lfloor \frac{N(N-1)}{4} \right\rfloor}{\frac{N(N-1)}{2}}$$

The complexity H is then calculated as in Definition 1□

Having found the upper and lower bounds of H for posets of N objects, it is now possible to frame each complexity indicator for a particular poset of N objects into the range defined by the upper and lower bounds. Hence, the closer the complexity to its upper bound, the more complex it is; likewise, the closer to its lower bound, the less complex it is.

5.2.2 An Application to the Ordering of Countries by Performance in Chemistry

A widespread source of posets in chemistry and environmental sciences is the ordering of objects that are characterised by different properties, something that has been further explored in the so-called Hasse diagram technique (Brüggemann and Bartel 1999; Restrepo and Brüggemann 2008; Restrepo et al. 2008a). If objects $x, y \in X$ are characterised by properties $q_1(x), q_2(x), \ldots, q_i(x)$ and $q_1(y), q_2(y), \ldots, q_i(y)$, respectively, x is ordered lower than y ($x \preccurlyeq y$) if all its properties are lower in magnitude than those of y, or if at least one property is lower for x while all others are equal. This gives place to comparabilities. If all properties of x and y are equal, both objects are equivalent. Note that in the entropic complexity approach introduced in this chapter, equivalences are not considered. If at least one property q_j satisfies $q_j(x) < q_j(y)$ while the others are opposite ($q_i(x) \geq q_i(y)$), x and y are incomparable.

Let us consider the ordering of countries according to the circulation of their scientific production in chemistry from 1996 to 2007, i.e. a bibliometric ordering of countries. We took the data from SCImago (2007). The example takes 195 countries characterised by three properties: NDoc, number of citable documents (articles, reviews, and conference papers); ACit, average citations of documents published during 1996–2007[10]; and Hind, the H-index (Hirsch 2005), which takes the value h if the country's number of documents has at least h citations. To make a fair comparison of countries, we took into account the 2011 population (Pop) of the countries listed in the database of the World Bank (2013), which corresponds to 183 countries. Thus, we

[10] SCImago takes a citation window of 4 years less than the observation window. That is why even if the query was performed in 2012, the information corresponds to the period 1996–2007.

Table 5.1 29 countries characterised by NDoc-cap (number of citable articles, reviews, and conference papers, per capita), ACit-cap (average citations of documents published during 1996–2007, per capita), and Hind (H-index, per capita) of chemical documents published between 1996 and 2011

Country	Label	NDoc-cap	ACit-cap	Hind-cap
Austria	AUS	0.00144578	0.02508424	0.17783062
Australia	AUS*	0.00119272	0.02150471	0.19560578
Belgium	BELG	0.00173828	0.03061113	0.23466797
Canada	CAN	0.00133530	0.02683961	0.29109632
Czech Republic	CZE	0.00152683	0.01908544	0.16184449
Denmark	DEN	0.00188572	0.04201383	0.26965788
Estonia	EST	0.00110075	0.01594981	0.06164179
Finland	FIN	0.00170355	0.02878992	0.18398292
France	FRA	0.00143855	0.02438349	0.35819989
Germany	GER	0.00160919	0.03017239	0.49241341
Hong Kong	HKO	0.00125856	0.02582556	0.15480231
Hungary	HUN	0.00121212	0.01410902	0.10909036
Ireland	IRE	0.00127368	0.02455654	0.12864163
Israel	ISR	0.00149800	0.03148791	0.21271566
Italy	ITA	0.00101272	0.01722637	0.20355674
Japan	JAP	0.00125943	0.01948339	0.33878685
Monaco	MONA	0.00285093	0.04843735	0.05416773
Netherlands	NET	0.00147347	0.03455279	0.28585254
Norway	NOR	0.00111611	0.01880653	0.10379867
New Zealand	NZE	0.00115069	0.01883672	0.10816444
Portugal	POR	0.00113227	0.01710866	0.11096287
Singapore	SIN	0.00182592	0.03805209	0.21545807
South Korea	SKO	0.00100585	0.01393096	0.16898210
Slovenia	SLO*	0.00225097	0.02831726	0.15531725
Spain	SPA	0.00146586	0.02543266	0.28291083
Sweden	SWE	0.00226870	0.04827789	0.39021602
Switzerland	SWI	0.00339661	0.08121296	0.69970178
United Kingdom	UK	0.00149308	0.02938381	0.42104845
United States	USA	0.00100059	0.02415426	0.50930071

obtained three properties: NDoc-cap, number of citable documents per capita [NDoc/Pop]; ACit-cap, average citations of documents published during 1996–2007 per capita [ACit × NDoc-cap]; and Hind-cap, the H-index per capita [H-index × (NDoc/Pop)]. For the sake of simplicity, we selected those countries with $\mathrm{NDoc-cap} \geq 0.001$, $\mathrm{ACit-cap} \geq 0.01$, and $\mathrm{Hind-cap} \geq 0.05$. Hence, we ended up with 29 countries, whose information is shown in Table 5.1. The complete information for all countries is shown in Table SP1 of the supplementary material uploaded in the following link: https://docs.google.com/file/d/0B4tX8gtPEjlbNjJFUWsxOTk1c0k/edit.

To assess the effect of each property upon the complexity of the poset, we considered four posets: one considering the three properties and three others where only two out of the three properties are regarded. These posets are depicted in Figs. 5.6, 5.7, 5.8 and 5.9.

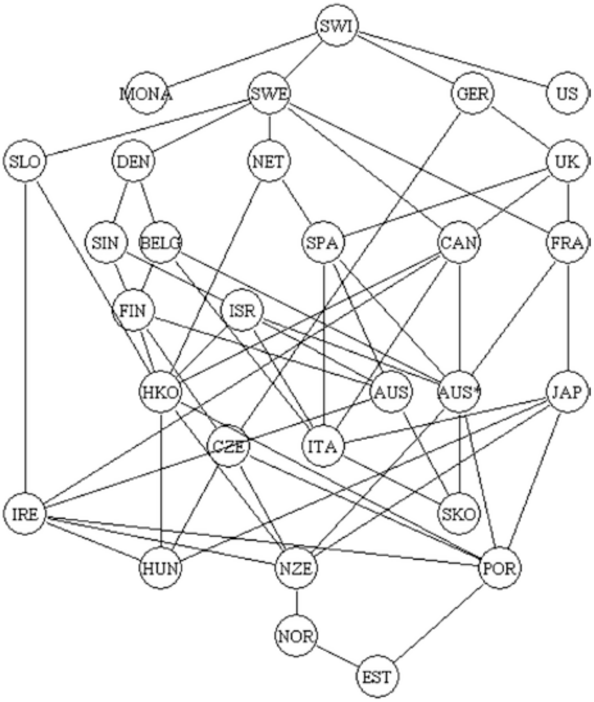


Fig. 5.6 Hasse diagram of 29 countries ordered by three bibliometric properties, with complexity of 0.9831. Country's labels are found in Table 5.1.

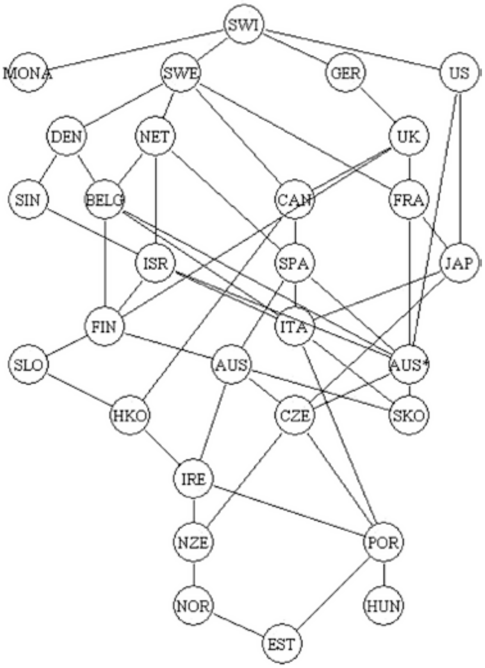


Fig. 5.7 Hasse diagram of 29 countries ordered by two bibliometric properties: ACit-cap (average citations of documents published during 1996–2007, per capita) and Hind-cap (H-index per capita) of chemical documents published between 1996 and 2011, with complexity of 0.9073. Country's labels are found in Table 5.1

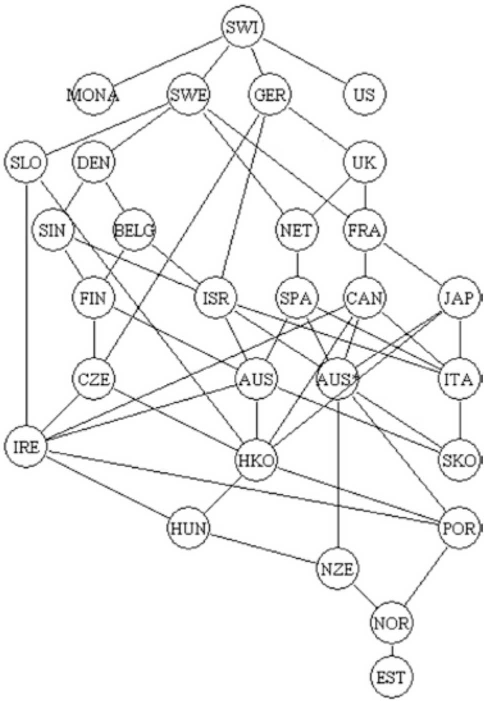


Fig. 5.8 Hasse diagram of 29 countries ordered by two bibliometric properties: NDoc-cap (number of citable articles, reviews, and conference papers) and Hind-cap (H-index per capita) of chemical documents published between 1996 and 2011, with complexity of 0.9593. Country's labels are found in Table 5.1

The three properties are oriented, i.e. high values of them indicate better circulation than low values. The poset of the 29 countries is shown in Fig. 5.6, which was drawn and further analysed with the software WHasse (Brüggemann and Bartel 1999; Brüggemann et al. 1995) available from Rainer Brüggemann. It shows that there is a country with maximum circulation of chemical scientific production, i.e. Switzerland (SWI). There are two countries behaving better than the others except for SWI, i.e. Sweden (SWE) and Germany (GER). Five are the countries with minimum circulation of chemical literature: Monaco (MONA), Hungary (HUN), Estonia (EST), South Korea (SKO), and USA. By inspecting each one of the three properties, we found that SWI always had the maximum score in each of them. Therefore, by removing one or even two properties out of the three discussed, SWI keeps its position as a country with maximum circulation of chemical literature (Figs. 5.7, 5.8 and 5.9).

As $N=29$, H can take 204 different values (Young diagrams). According to Proposition 2, the maximum value H can take is a number close to 1. The poset has 234 comparabilities and 172 incomparabilities, yielding $H=0.9831$, i.e. the poset has a complexity of 98 %, which indicates that it is closer to the maximum allowed complexity of 100 %. The Young diagram depicting this partition would be that with the largest row having 234 boxes and the shortest with 172 ones.

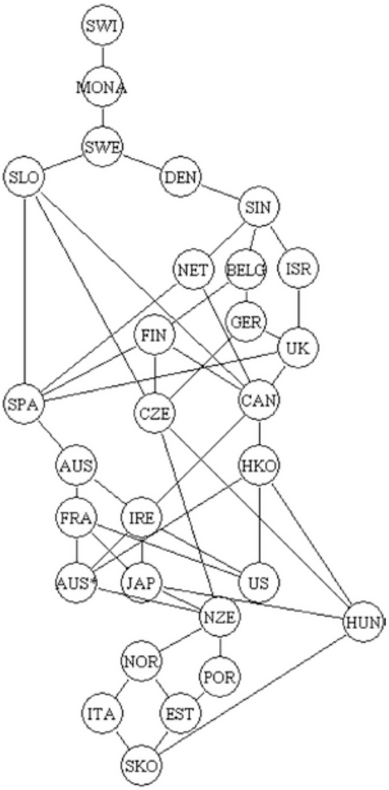


Fig. 5.9 Hasse diagram of 29 countries ordered by two bibliometric properties: NDoc-cap (number of citable articles, reviews, and conference papers, per capita) and ACit-cap (average citations of documents published during 1996–2007, per capita) of chemical documents published between 1996 and 2011, with complexity of 0.5917. Country's labels are found in Table 5.1

The deletion of NDoc-cap and the further ordering of the countries based on the remaining two properties give place to the poset shown in Fig. 5.7 and to its respective complexity ($H = 0.9073$). Sweden (SWE) and Germany (GER) keep their position as countries with better circulation than other countries except Switzerland (SWI). Note that USA now accompanies SWE and GER in their behaviour of good circulation. In fact USA, in this poset, is not part of the countries with minimum circulation as it was in the poset of Fig. 5.6. USA is now behaving better than several other countries, e.g. Japan (JAP), Australia (AUS*), Italy (ITA), and Czech Republic (CZE), among others. The countries with minimum circulation are Monaco (MONA), Hungary (HUN), Estonia (EST), and South Korea (SKO). In the diagram of Fig. 5.6 the only thing we knew about USA was that SWI was better than it in terms of chemical scientific circulation. Hence, by ordering countries disregarding the number of citable documents per capita, the complexity of the poset is reduced.

The deletion of ACit-cap and the ordering of the countries based on the remaining two properties yielded the poset shown in Fig. 5.8 with complexity $H = 0.9593$. Sweden (SWE) and Germany (GER) keep their good behaviour and USA leaves this group by becoming, again, a country with minimum scientific circulation along with Monaco (MONA), Estonia (EST), and South Korea (SKO). By comparing this result with that of Fig. 5.6, Hungary (HUN) is not part of the countries with minimum circulation, as it is now better than New Zealand (NZE), Norway (NOR), and Estonia (EST). Hence, by ordering countries disregarding the average citations of documents published per capita, the complexity of the poset is reduced (Fig. 5.8).

The deletion of Hind-cap and the ordering of the countries based on NDoc-cap and ACit-cap yielded the poset shown in Fig. 5.9 with complexity $H=0.5917$. This poset has several changes, e.g. Sweden (SWE) and Germany (GER) leave their good circulation, for there are some other countries behaving better than them. In the case of SWE, now Monaco (MONA) is better than it. In the case of GER, it turns out that Belgium (BELG), Singapore (SIN), Denmark (DEN), and even SWE now have better circulation than GER. South Korea (SKO) and USA are the countries with minimum chemical scientific circulation. In this case MONA and Hungary (HUN) leave the set of countries with minimum circulation as MONA becomes better than many other countries; in fact the only country with better circulation than MONA is Switzerland (SWI). Regarding HUN, in spite of leaving the set of countries with minimum circulation, it is only better than South Korea (SKO). Thus, by ordering countries disregarding the H-index per capita, the complexity of the poset is largely reduced regarding the complexity considering the three properties. This result indicates that this poset is more ordered or possesses more comparabilities than the previous ones. An explanation of such an effect can be seen in the meaning of NDoc-cap and ACit-cap, properties used to obtain the discussed poset. NDoc-cap refers to the citable documents per capita and ACit-cap to the average citations per capita, which depict similar orderings. In fact, SWI > MONA > SWE shows a complete agreement of both properties for these three countries. Thus, it is the inclusion of Hind-cap which gives complexity to the poset.

5.2.3 *Some Possibilities for Posetic Complexity Indicators*

Coming back to the complexity indicator here presented (Definition 1), such an indicator is based on comparabilities and incomparabilities. However, there is another kind of relationship between objects of a poset, namely equivalence, which occurs when $x \preccurlyeq y$ and $y \preccurlyeq x$, with x and y being objects to order. This relation between the two objects is written $(x \sim y)$. Hence, a general posetic complexity indicator could include the three relationships. In this case the total number of relationships R, given the number of objects to order, would be partitioned into three parts. The number of different possibilities of doing that (Hardy 1920) is given by $(n+3)^2/12$, which yields the number of corresponding Young diagrams for those partitions. However, not all Young diagrams are attainable by partitioning relationships into comparabilities, incomparabilities and equivalences. This can be seen by a set X of three elements, which yields three possible relationships. For example, the Young diagram with three rows indicating one comparability, one incomparability and one equivalence is not a real possibility, as the equivalence of two elements

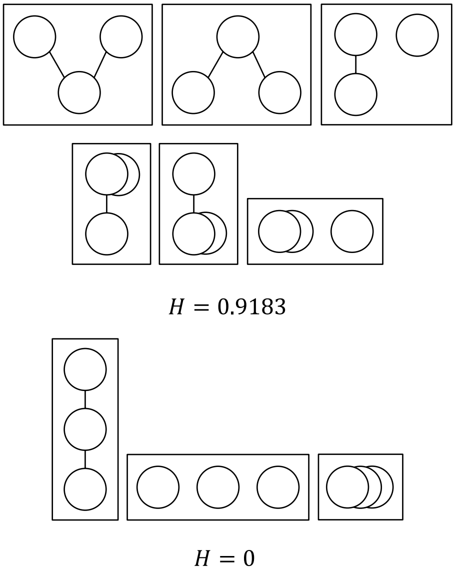


Fig. 5.10 Unlabelled posets of three objects, considering three relationships, and their respective values of complexity *H*. *Multiple circles* indicate equivalence among circles

implies either the comparability or incomparability with the third one, giving place to a partition 2,1, i.e. a Young diagram of only two rows, one with two boxes and another one with one. A mathematical question to be solved is the calculation of the realisable Young diagrams associated with these particular posetic partitions including three parts (three relationships).

Even if such a relation between relation partitions and Young diagrams is not established in this chapter, the complexity indicator settled in Definition 1 can be generalised to the three order relationships.

Definition 2 Let $(X,\preccurlyeq)$ be a poset and let $R \subset X \times X$ be the possible relationships between couples of objects to order. Let $\perp \subseteq R$, $\| \subseteq R$ and $\sim \subseteq R$ be the comparabilities, incomparabilities, and equivalences of X, respectively, such that $R = \perp \cup \| \cup \sim$. As $|R| = N(N-1)/2$ is the cardinality of R, with $|X| = N$, we define $p_{\perp} = |\perp|/|R|$, $p_{\|} = |\,\|\,|/|R|$, and $p\sim = |\sim|/|R|$ as the probabilities of having a comparability, an incomparability and an equivalence in $(X,\preccurlyeq)$, respectively. Hence, we say that $H(X,\preccurlyeq) = -\sum p_i \log_2 p_i$, with $i \in \{\perp, \|, \sim\}$, is the *complexity of the poset* $(X,\preccurlyeq)$.

The posets with three objects and their respective complexities are shown in Fig. 5.10.

Following a similar analysis of the upper and lower bounds found in Proposition 2, we found that here the respective values are $H_{\min} = 0$ and $H_{\max}$ is equal to 1 if N is even and close to 1 if N is odd. In this latter case, the probabilities of each one of the three relationships are almost equal and are of the form

$$p_i \approx \frac{\dfrac{N(N-1)}{6}}{\dfrac{N(N-1)}{2}} \approx 0.33\,, \text{ always satisfying } \sum p_i = 1.$$

5.3 Conclusions and Outlook

The posetic complexity indicator, based on comparabilities and incomparabilities, solves the shortcomings of the complexity indicators reviewed in the chapter, namely the one based on dimension theory and the heuristic one using parameters α and β. When contrasted with the first one, our complexity indicator does not have the problems of dimension calculation for it is based on counting comparabilities and incomparabilities, rather than intersection of linear extensions. These kinds of counting are already included in several statistical packages to treat posets, e.g. WHasse and PyHasse (Brüggemann and Voigt 2009), available from Rainer Brüggemann. Regarding the contrast of our complexity indicator with the heuristic one, our method does not fit any parameter before complexity calculation, which makes it an objective complexity indicator. Additionally, our method does not emphasise the importance of incomparabilities over comparabilities. In the indicator presented in this chapter, both comparabilities and incomparabilities are evenly regarded.

In a recent workshop on posets and their applications,[11] Fattore mooted the idea of calculating the complexity through a Kolmogorov's approach. In this case the complexity is not based on counting comparabilities and incomparabilities but on the treatment of a posetic derived matrix, e.g. a cover matrix (showing cover relationships between couples of objects of the poset). This approach is related to the compressibility of the given matrix. Hence, a quite complex poset is one requiring more bits to be represented in a string code, e.g. a binary string. In fact, for a given set of objects, the poset with maximum complexity is the one that after compression has the longest string. In contrast, the less complex poset is the one that, after compression, is represented by a minimum number of bits. It would be interesting to explore this approach and its mathematical properties, as well as its relationship with other complexity indicators.

The example showing the applicability of the complexity indicator introduced here is of particular importance given the current interest on academic rankings. For several reasons, including distribution of funds for research based on research performance, academic rankings have become popular. Examples of these rankings are the Academic Ranking of World Universities (Shanghai ranking), Times Higher Education and the QS World University Rankings, among others. All these rankings, in the end, yield a total order resulting from the weighted aggregation of indicators (properties); the resulting indicator is called a composite indicator. The difference among those rankings lies on the kind of indicators used, i.e. some more oriented to research, some others to education, etc. The additional difference is the importance each ranking gives to the indicators. It turns out that the aggregation of indicators is customarily a linear combination, whose weights are selected upon the importance of the indicators. That is why, contrary to the popularisation of these

[11] Tenth International Workshop on Partial Order, Theory and Application, Berlin, 27–28 September 2012.

rankings and to their use for decision-making processes, we think they are not "the" best option due to their subjectivities. Changing the subjectivities here mentioned changes the final ranking (Restrepo et al. 2008b). A possible way to overcome the subjectivity on the indicators' weights is to avoid aggregations, something the Hasse diagram technique allows.

The example of countries' ordering based on chemical literature circulation was, then, selected as a way to show an alternative option to academic rankings. The analysis of the posetic complexity shows that, among the properties considered, the H-index per capita constitutes the property introducing more incomparabilities when combined with the other two bibliometric properties. This kind of approach—not a ranking—and several other techniques designed to extract information from posets under the shade of the Hasse diagram technique constitute a novel approach worth studying and using by decision-makers. The results on the ranking of countries show that Switzerland (SWI), regardless of the three kinds of descriptors used, is always the best country in chemical circulation of knowledge. There is not "a" worst country regarding this circulation, which constitutes one of the advantages of posets, i.e. if data do not allow it and if aggregations are not performed, several "firsts", several "seconds", and several "lasts" may result. Among the 29 countries considered, Monaco (MONA), Hungary (HUN), Estonia (EST), South Korea (SKO), and the USA are countries that need to take measures to address (increase) their circulation of chemical knowledge. It is a matter of surprise to find the USA in the group of countries needing action, as it is common to consider this country as one of the best in this kind of circulation of its research. This idea is true only if the bibliometric data are not considered per capita. For the particular case considered in this chapter, it would be interesting to know the investment on chemical research of the countries studied to see how this information maps the ordering found here or how it affects the ordering. Then, considering the population of a country is important as it is not fair to compare, e.g., the USA with SWI based on raw information, where USA has by far more scientists than SWI. If that is done, the USA appears with the best chemical circulation.

References

Andrews GE, Eriksson K (2004) Integer partitions. Cambridge University Press, Cambridge

Brüggemann R, Bartel HG (1999) A theoretical concept to rank environmentally significant chemicals. J Chem Inf Comput Sci 39:211–217

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, New York, NY

Brüggemann R, Voigt K (2009) Analysis of partial orders in environmental systems applying the new software PyHasse. In: Wittmann J, Flechsig M (eds) Simulation in Umwelt- und Geowissenschaften- workshop Potsdam 2009. Shaker-Verlag, Aachen, pp 43–55

Brüggemann R, Halfon E, Bücherl C (1995) Theoretical base of the program "Hasse". GSF-Bericht 20/95: Neuherberg

Dushnik B, Miller EW (1941) Partially ordered sets. Am J Math 63:600–610

Halfon E (2006) Hasse diagrams and software development. In: Brüggemann R, Carlsen L (eds) Partial order in environmental sciences and chemistry. Springer, Berlin, pp 385–392

Hardy GH (1920) Some famous problems of the theory of numbers and in particular Waring's problem. Clarendon, Oxford

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102:16569–16572

Kelly D, Trotter WT (1982) Dimension theory for ordered sets. In: Rival I (ed) Ordered sets. North-Holland, Amsterdam, pp 171–212

Kierstead HA, Trotter WT (1985) Inequalities for the greedy dimensions of ordered sets. Order 2:145–164

Luther B, Brüggemann R, Pudenz S (2000) An approach to combine cluster analysis with order theoretic tools in problems of environmental pollution. Match Commun Math Comput Chem 42:119–143

McKay BD, Brinkmann G (2002) Posets on up to 16 points. Order 19:147–179

Neggers J, Kim HS (1998) Basic posets. World Scientific publications, Singapore

Restrepo G, Brüggemann R (2008) Dominance and separability in posets, their application to isoelectronic species with equal total nuclear charge. J Math Chem 44:577–602

Restrepo G, Weckert M, Brüggemann R, Gerstmann S, Frank H (2008a) Ranking of refrigerants. Environ Sci Technol 42:2925–2930

Restrepo G, Brüggemann R, Weckert M, Gerstmann S, Frank H (2008b) Ranking patterns, an application to refrigerants. Match Commun Math Comput Chem 59:555–584

Restrepo G, Brüggemann R, Klein D (2011) Partially ordered sets: ranking and prediction of substances' properties. Curr Comput Aided Drug 7:133–145

SCImago (2007) SJR—SCImago Journal & Country Rank. Retrieved 13 Dec 2012 from http://www.scimagojr.com

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Trotter WJ (1992) Combinatorics and partially ordered sets, dimension theory. The Johns Hopkins University Press, Baltimore, MD

Trotter WJ, Bogart KP (1976) On the complexity of posets. Discrete Math 16:71–82

West DB (1985) Parameters of partial orders and graphs: packing, covering, and representation. In: Rival I (ed) Graphs and orders. North-Holland, Amsterdam, pp 267–350

World Bank (2013) Retrieved 12 Mar 2013 from http://data.worldbank.org/indicator/SP.POP.TOTL

Yáñez J, Montero J (1999) A poset dimension algorithm. J Algorithm 30:185–208

Yannakakis M (1982) The complexity of the partial order dimension problem. SIAM J Algebraic Discrete Methods 3:351–358

Part II
Partial Order as Tool to Analyse Composite Indicators

Chapter 6
Comparative Knowledge Discovery with Partial Orders and Composite Indicators: Multi-indicator Systemic Ranking, Advocacy, and Reconciliation

Ganapati P. Patil and S.W. Joshi

Abstract In many decision-making situations, ranking of objects with related tasks is a fundamentally important issue. In these situations, a number of objects are ranked on the basis of measurements on a set of several indicators. A prevalent approach is to form a composite index from these several measurements using weights of relative importance for the selected indicators determined by experts and/or stakeholders. An entirely different approach for ranking uses the theory of partially ordered sets (posets). In classical poset ranking derived by average ranks (AR) method, unequal indicator weights of any kind do not play any part in the computation of ranking based on a given data matrix. Here we present a novel method of poset ranking that involves stochastic order of weighted indicator cumulative rank frequency (CRF) distributions. We then investigate how this data-validated evidence-based ranking can be used to construct a composite index reproducing an identical ranking. We further seek reconciliation between databased weighted poset CRF ranking and ranking induced by an arbitrary subjective composite index. This investigation acquires particular importance today in view of issues of trade-offs among indicators, implicit in the apparent advocacy involved in the choice of weights of the composite index. This chapter is based on research conducted in the spirit of start small even for big data. The concept of databased

This is a concept paper based on the inaugural keynote address by the first author at the International Workshop on Partial Order Theory and Modeling held at Berlin, Germany, during September 27–29, 2012. This research and outreach effort has been partially supported by US NSF Project 0307010 for Digital Governance and Hotspot Geo-Informatics of Detection and Prioritization.

G.P. Patil (✉)
Center for Statistical Ecology and Environmental Statistics, Department of Statistics,
The Pennsylvania State University, University Park, PA 16802, USA
e-mail: gpp@stat.psu.edu

S.W. Joshi
Department of Computer Science, Slippery Rock University of Pennsylvania,
Slippery Rock, PA 16057, USA
e-mail: sharadchandra.joshi@sru.edu

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_6, © Springer Science+Business Media New York 2014

weighted poset ranking introduced here may open doors to still other ways of weighting schemes and other reconciliation approaches for comparative knowledge discovery using partial orders and composite indicators. Meaningful ability to deal with big data is an urgent need of comparative knowledge discovery with partial orders and composite indicators in this infometrical computer science and software engineering age of statistical information science and technology. This chapter is prepared in the spirit of a concept paper for digital age infometrics and comparative knowledge discovery critical in several fields, such as document discovery, drug discovery, gene discovery, chemical discovery, criminal discovery, geospatial critical area discovery, etc. The ranking, prioritization, and selection of objects and indicators carrying a variety of names in a variety of contemporary issues of societal and scientific importance based on relevant evidence embodied in data matrices provide insightful leads in these substantive investigations involving variously big data.

6.1 Introduction and Overview

In many decision-making situations, ranking of objects and related tasks is a fundamentally important issue. Given a collection of objects, the ranking of objects presupposes some abstract latent property of objects. If this latent property is not directly measurable, measurements are made on multiple surrogate indicators believed to be positively oriented with the latent property. These indicator values are then used to rank the objects. A prevalent approach is to rank the objects based on a composite index calculated from indicator values for the individual objects. The index depends on weights determined by experts with insight and/or by stakeholders with interest in the issue. Another approach to arrive at ranking is the application of the theory of partially ordered sets (posets). It recognizes that multiple indicator measurements define a partially ordered set of objects and enumerates linear extensions of the partial order. A linear extension of a poset is a total (linear) order that is an extension of the partial order consistent with the data matrix of objects and indicators. Theoretical and computational aspects of poset ranking, based on average heights together with its practical applications, have been extensively studied by Brüggemann and his associates. In particular, Brüggemann and Patil (2011) give an extensive account of the theory and application of posets for ranking and prioritization.

Patil and Taillie (2004) use the set of linear extensions to find a total (weak) order among the objects on the basis of ranks of the objects in all the linear extensions of the poset. They determine the poset ranking in terms of the stochastic order of cumulative rank frequency (CRF) distributions of objects, obtained iteratively.

The popular composite index-based approach depends on indicator weights usually provided from subjective considerations. In this chapter, we investigate a modification of Patil and Taillie's approach to construct composite indexes corresponding to databased weights for indicators. To accomplish this, we first introduce poset ranking with weighted CRF distributions. Then, we discuss methods to see if there is a composite index that will rank the objects in the same order, or as per some criterion nearly in the same order as the databased weighted poset ranks. That is, if the weighted poset ranking is representable or nearly representable as a ranking due to some composite index. Finally, given a composite index, we will suggest a procedure to determine if that can be considered as a basis to reproduce data-validated poset ranking.

A list of abbreviations/acronyms used in the chapter appears in the appendix at the end of the chapter, together with some illustrative computational detail.

6.2 Poset Ranking and Preliminaries

In this section, we introduce basic vocabulary, notation, and the version of poset ranking described in Patil and Taillie (2004) as needed for this chapter.

6.2.1 Basic Definitions and Notation

n: number of objects to be ranked

m: number of indicators to be used

O_i: ith object, $i = 1, 2, 3, \ldots, n$

$$OS = \{O_i \mid i = 1,2,3,\ldots,n\}$$

q_j: jth indicator, $j = 1, 2, \ldots, m$

IB $= \{q_j \mid j = 1, 2, \ldots, m\}$, where IB stands for "information base," Brüggemann et al. (1995).

x_{ij}: real valued measurement of indicator q_j on object O_i

X: n by m matrix (x_{ij}), called data matrix.

$\prec$: We write $O_i \prec O_j$ if $i = j$ and, for $i \neq j$, if $x_{ik} \geq x_{jk}$ for $k = 1, 2, \ldots, m$ and $x_{ik} > x_{jk}$ for at least one k.

$\parallel$: we write $O_i \parallel O_j$, for $i \neq j$, if $x_{ik} > x_{jk}$ for at least one k, say, $k = k'$ and $x_{jk} > x_{ik}$ for at least one k, say, $k = k''$

When $O_i \prec O_j$, we say O_i precedes O_j.

When $O_i \prec O_j$ or $O_j \prec O_i$ for a given pair of i and j, we say objects O_i and O_j are comparable.

When $O_i \parallel O_j$, we say objects O_i and O_j are incomparable.

If two distinct rows i and j of the data matrix are identical, then the two objects O_i and O_j are equivalent, and we write $O_i \cong O_j$. Given a data matrix with equivalent objects, we deal with the quotient set of objects OS/$\cong$ instead of OS. Having ranked objects of OS/$\cong$, equivalent objects of OS are assigned appropriate tied ranks. To simplify our discussion, we assume throughout this chapter that rows of the original data matrix are all distinct. Thus for our purpose, each pair of objects is related through $\prec$ or $\parallel$.

Further, $\prec$ is assumed to be a binary relation on OS that is reflexive, transitive, and antisymmetric as defined in poset literature. The binary relation $\prec$ is a strict partial order and OS together with $\prec$ is a partially ordered set. We denote it as (OS,$\prec$). Since the partial order $\prec$ is induced by X we may want to denote it by $\prec_X$ and write (OS, $\prec$) as (OS, $\prec_X$). But unless there is a need to show the underlying dataset, we will write $\prec$ instead of $\prec_X$. If no pair of objects from OS is incomparable, then $\prec$ is a total (linear) order, in which case the ranking problem is trivial, since the linear order is the rank order. In the presence of incomparable pairs, we need a nontrivial extension of the partial order $\prec$ to a total order that represents the rank order. This can be done, often only as a weak order, by employing linear extensions. Before proceeding to the details of linear extensions, let us look at two useful tools, namely, Hasse diagram and zeta matrix, that give us graphical and computational insight into data matrix and related partial order, as needed for this chapter.

6.2.2 Hasse Diagram

A Hasse diagram is a graphical representation of a poset. It is a graph drawn with some special rules. The OS is the set of its vertices (nodes) with edges between objects and their *cover* objects. A cover element of an object O in OS is defined to be an object $O' \neq O$ in OS such that (i) $O' \prec O$ and (ii) there is no third object O'' in OS such that $O'' \prec O$and $O' \prec O''$. An object O is a maximal object if there is no object O' in OS such that $O' \prec O$. An object O in OS is a minimal object if there is no object O' in OS such that $O \prec O'$. An object, which is both maximal and minimal, is called an isolated object. A Hasse diagram is drawn in levels numbered 1, 2, ... as described below:

1. Maximal as well as isolated elements are drawn at level 1, the top level.
2. For $i = 2, 3,\ldots$ those elements are drawn at level i whose cover elements occur at levels $j, j < i$.

The following example illustrates a Hasse diagram.

Example 2.2.1

Consider the data matrix in Table 6.1 below. Its Hasse diagram is shown in Fig. 6.1.

If we remove object d from the dataset, object f will still remain at level 3.

Table 6.1 Data matrix

Object	q_1	q_2
a	3	6
b	5	3
c	2	5
d	4	2
e	1	4
f	2	1

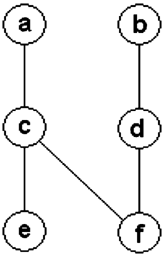


Fig. 6.1 Hasse diagram

For our purpose, a Hasse diagram is useful for manually constructing linear extensions by visual inspection, especially when the data matrix is small. It is also useful for one to understand various aspects of the data matrix as discussed in Brüggemann and Patil (2011).

6.2.3 Zeta Matrix

An n by n matrix, Zeta = (ζ_{ij}) is a tool to represent a partial order. Its elements are 0 or 1. $\zeta_{ij} = 1$ if $O_j \prec O_i$ and $\zeta_{ij} = \zeta_{ji} = 0$ if $O_j || O_i$. Table 6.2 shows the Zeta matrix for the data matrix in Example 2.2.1. Once constructed, it can be conveniently used for many computational purposes. For example, sum of the elements of row *i* of the zeta matrix is the highest position from the top that the object O_i will occupy in any linear extension and *n minus* the sum of the elements in the *i*th column is the lowest position O_i can occupy in any linear extension. These facts are used in constructing linear extensions in the software developed and used for this chapter. Patil and Taillie (2004) give a succinct algorithm for the construction of a Hasse diagram from the zeta matrix.

6.2.4 Linear Extensions and Their Application to Ranking

A linear extension of a partial order $\prec$ is a linear order $\prec$* such that if $O_i \prec O_k$ then $O_i \prec^* O_{k.}$ In other words, a linear extension is a permutation of objects that does not contradict the order of objects implied by $\prec$. Table 6.3 shows all sixteen linear extensions for the data matrix of Example 2.2.1.

Following Patil and Taillie, each linear extension assigns ranks to the objects in the data matrix. All of these different ranks are used to compute the final order, from which a weak order or a linear order can be obtained. In the latter case, we are speaking of poset ranks.

Table 6.2 Zeta matrix

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

Table 6.3 Sixteen linear extensions for the data matrix of Example 2.2.1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*b*	*b*	*b*	*b*	*b*	*b*	*b*
b	*b*	*b*	*b*	*b*	*c*	*c*	*c*	*c*	*a*	*a*	*a*	*a*	*a*	*d*	*d*
c	*c*	*c*	*d*	*d*	*b*	*b*	*b*	*e*	*c*	*c*	*c*	*d*	*d*	*a*	*a*
d	*d*	*e*	*c*	*c*	*d*	*d*	*e*	*b*	*d*	*d*	*e*	*c*	*c*	*c*	*c*
e	*f*	*d*	*e*	*f*	*e*	*f*	*d*	*d*	*e*	*f*	*d*	*e*	*f*	*e*	*f*
f	*e*	*f*	*f*	*e*	*f*	*e*	*f*	*f*	*f*	*e*	*f*	*f*	*e*	*f*	*e*

Table 6.4 Computation of final ranks using AR method for the data matrix of Example 2.2.1

Object	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Average rank	Final rank
a	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	1.563	1
b	2	2	2	2	2	3	3	3	4	1	1	1	1	1	1	1	1.875	2
c	3	3	3	4	4	2	2	2	2	3	3	3	4	4	4	4	3.125	3
d	4	4	5	3	3	4	4	5	5	4	4	5	3	3	2	2	3.750	4
e	5	6	4	5	6	5	6	4	3	5	6	4	5	6	5	6	5.063	5
f	6	5	6	6	5	6	5	6	6	6	5	6	6	5	6	5	5.625	6

There are two versions of poset ranking: (1) method of average heights—often in applications called average ranks—(AR method) (Winkler 1982) and (2) cumulative rank frequency method (CRF method) (Patil and Taillie 2004). The former method computes average height/rank for each object over all linear extensions and assigns final (some times tied) ranks based on the average height/ rank scores. Thus, let

LE (OS, $\prec$) = set of all linear extensions of (OS, $\prec$) and

$\rho_l(O_i)$ = rank of object O_i in the linear extension *l*.

Then the final average rank of O_i is

$$\rho_{av}(O_i) = \sum l \rho_l(O_i) / \mid LE(OS, \prec) \mid, \tag{6.1}$$

where |LE(OS, $\prec$)| is the total number of linear extensions of the poset (OS, $\prec$).

Table 6.4 shows results for Example 2.2.1. It shows rank of each object in each of the possible sixteen linear extensions with the average rank and the final rank of each object in the two rightmost columns.

For the CRF method, instead of averaging ranks of each object from its ranks in linear extensions, a rank frequency distribution is obtained for each object. Thus for each object O_i, we record the count of linear extensions that assign rank r to the object and denote the rank frequency count for object O_i by $f_i(r)$, for $r = 1, 2, \ldots, n$. More formally, if we define for each linear extension l the characteristic function

$\text{chf}_l(r,i) = 1$ if r = rank of O_i in linear extension l, and = 0 otherwise,

so that $\text{chf}_l(r,i)$ is equal to one if and only if the linear extension l assigns rank r to O_i, then

$f_i(r) = \sum_l \text{chf}_l(r,i)$, for $i = 1, 2, \ldots, n$ where summation extends over all linear extensions of (OS, $\prec$) , and the unnormalized cumulative rank frequency (CRF) distribution becomes $E_i(r) = \sum_{t \leq r} f_i(t)$ for $i = 1, 2, \ldots, n$.

The normalized CRF then becomes

$$F_i(r) = E_i(r) / E_i(n) = E_i(r) / \,|\, LE(OS, \prec) \,|, \quad for\, i = 1,2,\ldots,n$$

These are the statistical cumulative distribution functions.

It can be seen that if $O_i \prec O_j$ then $E_i(r) \geq E_j(r)$ or, equivalently, $F_i(r) \geq F_j(r)$ for $r = 1, 2, \ldots, n$. Let us now denote by φ_1 the set $\{F_i \mid i = 1, 2, 3, \ldots, n\}$. We define the equivalence relation $\cong_1$ on φ_1 and write $F_i \cong_1 F_j$ if $F_i(r) = F_j(r)$ for $r = 1, 2, \ldots, n$. Further, we denote by OS$_1$ the quotient set φ_1 and define a relation $\prec_1$ on OS$_1$ as follows:

If ξ_i and ξ_j are equivalence classes of F_i and F_j, respectively, then $\xi_i \prec_1 \xi_j$ if $F_i(r) \geq F_j(r)$ for $r = 1, 2, \ldots, n$.

Clearly, (OS$_1$, $\prec_1$) is a poset. Thus starting with a poset (OS, $\prec$) computation of F_i, $i = 1, 2, 3, \ldots, n$, produces another poset (OS$_1$, $\prec_1$). We denote the entire computation with (OS, $\prec$) yielding another poset (OS$_1$, $\prec_1$) by the operator CRF, that is, CRF(OS, $\prec$) = (OS$_1$, $\prec_1$). From above we conclude that $\xi_i \prec_1 \xi_j$ if $O_i \prec O_j$ where ξ_i and ξ_j are equivalence classes of cumulative rank frequency functions of O_i and O_j, respectively. Further, mapping from OS to OS$_1$ induces a partition of OS such that O_i and O_j belong to an equivalence class if and only if the corresponding cumulative rank frequency functions belong to the same equivalence class in OS$_{1.}$ If $\prec_1$ is a linear order on OS$_1$, then the equivalence classes in OS$_1$ receive appropriate ranks from $\{1, 2, \ldots, n\}$. The maximal element receives the highest rank and individual objects (cumulative rank frequency functions, F_i's) belonging to the same non-singleton equivalence class receive tied ranks. Then the rank of O_i is the same as that of F_i for $i = 1, 2, \ldots, n$. As an example, (non-normalized) CRFs for Example 2.2.1 computed from Table 6.4 are shown in Table 6.5. Figure 6.2 shows the stochastic order to be linear so that the resulting CRF ranking coincides with poset ranking resulting from the AR method.

If (OS$_1$, $\prec_1$) is not linearly ordered, then (OS$_i$, $\prec_i$) is computed recursively as (OS$_i$, $\prec_i$) = CRF(OS$_{i-1}$, $\prec_{i-1}$) for $i = 2, 3, 4, \ldots$ until a linearly ordered (OS$_i$, $\prec_i$) is obtained. If $\prec_i$ is a linear (possibly weak) order, ranks of objects in OS$_{i-1}$ are determined in terms of ranks of objects in OS$_i$ and so forth, until objects in OS are ranked.

Table 6.5 CRF's for data of Example 2.2.1

	Cumulative rank frequency					
Object	1	2	3	4	5	6
a	9	14	16	16	16	16
b	7	12	15	16	16	16
c	0	4	10	16	16	16
d	0	2	6	12	16	16
e	0	0	1	4	10	16
f	0	0	0	0	6	16

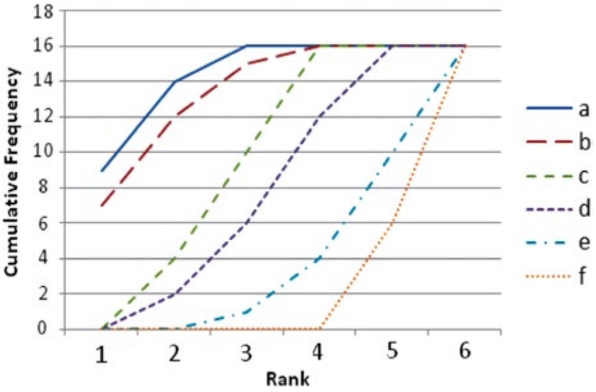


Fig. 6.2 Cumulative rank frequencies frequencies for Example 2.2.1

Table 6.6 Data matrix for Example 2.4.2

Object	q_1	q_2
a	3	6
b	5	3
c	6	1
d	3	2
e	1	4
f	2	1

We present another example, Example 2.4.2 that further brings out some difference between the AR and CRF methods and shows the need for iteration for the CRF method.

Example 2.4.2.

Consider the data matrix shown in Table 6.6. Its Hasse diagram and zeta matrix are shown in Fig. 6.3 and Table 6.7, respectively. There are 33 linear extensions, which are not shown here.

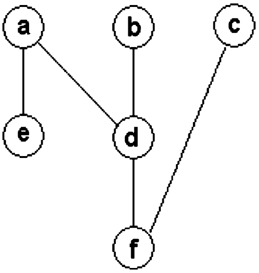


Fig. 6.3 Hasse diagram for Example 2.4.2

Table 6.7 Zeta matrix for Example 2.4.2

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

Table 6.8 CRF matrix for Example 2.4.2

	Rank						Average linear
Object	1	2	3	4	5	6	extension rank
a	15	27	33	33	33	33	1.73
b	11	22	30	33	33	33	2.09
c	7	14	21	28	33	33	2.88
d	0	0	6	21	33	33	4.18
e	0	3	9	17	25	33	4.36
f	0	0	0	0	8	33	5.76

The CRF matrix and average linear extension ranks are shown in Table 6.8. The chart of the CRFs for the six data objects appears in Fig. 6.4. As the figure shows, CRFs for objects *d* and *e* intersect, meaning $\prec_1$ is not a linear order.

In a situation like this, the CRF method proposes to apply the CRF operator to (OS_1, $\prec_1$) and obtain CRF(OS_1, $\prec_1$) = (OS_2, $\prec_2$). If $\prec_2$ is a linear order, the process stops, or else the iterative process CRF(OS_{i-1}, $\prec_{i-1}$) = (OS_i, $\prec_i$) continues as mentioned above. For Example 2.4.2, $\prec_2$ is a linear order, as will be seen below.

Figure 6.5 displays the Hasse diagram for the CRF matrix shown in Table 6.8. There are only two linear extensions: one ranks object *d* over *e* and the other *e* over *d*. The corresponding CRF matrix is shown in Table 6.9. It has two identical rows so that $d \cong e$. The quotient set of objects is a linear order. Accordingly, the objects listed by their ranks are *a*, *b*, *c*, {*d*, *e*}, *f*. The objects *d* and *e* are tied for the fourth place in the final ranking.

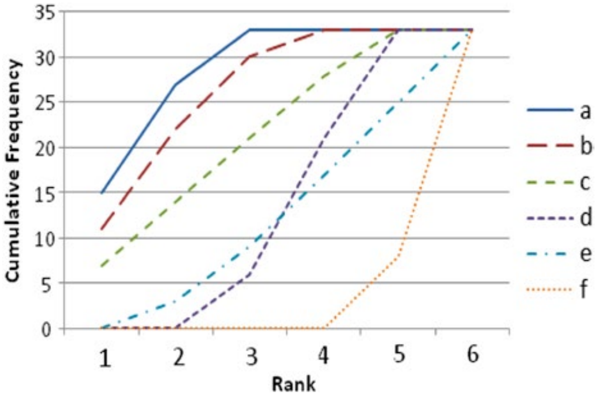


Fig. 6.4 CRF's for Example 2.4.2

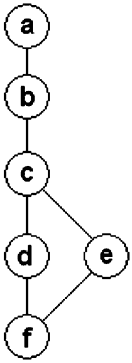


Fig. 6.5 Hasse diagram for CRF matrix in Table 6.8

Table 6.9 CRF matrix computed from linear extensions based on Table 6.8

	Rank					
Object	1	2	3	4	5	6
a	2	2	2	2	2	2
b	0	2	2	2	2	2
c	0	0	2	2	2	2
d	0	0	0	1	2	2
e	0	0	0	1	2	2
f	0	0	0	0	0	2

When the quotient set of objects is nontrivial, individual objects belonging to an equivalent set are tied. If there are p objects in an equivalent set that are tied for the rth place, then each individual object in the set receives the half-integer rank of r+(p−1)/2 and the next available rank is $r+p$.

A problem arises for either of the two methods in a situation when exhaustive enumeration as done for Examples 2.2.1 (Table 6.3) and 2.4.2 above is not possible or practical. The number of linear extensions increases exponentially with the number of incomparabilities in the data matrix. When it becomes impossible to exhaustively enumerate linear extensions, for the CRF method, we must use estimates of the ratios rather than the actual ratios $F_i(r)=E_i(r)/E_i(n)$, for $i=1, 2, \ldots, n$, from randomly selected linear extensions. A convenient algorithm to generate random linear extensions is due to Bubley and Dyer [Bubley and Dyer (1999); Brüggemann and Patil (2011)]. Given a linear extension, it generates a sequence of linear extensions using Monte Carlo Markov chain (MCMC) simulation. Given a linear extension l_i in the sequence, the next linear extension l_{i+1} in the sequence is obtained by switching positions in l_i of a pair of incomparable objects selected using a certain random mechanism. We present the algorithm here in a pseudocode format.

Algorithm Bubley–Dyer

```
Adopted from Bubley and Dyer (1999) and Brüggemann and Patil (2011).
Remark:
  Given a poset (OS,<), |OS| = n, this algorithm generates a sequence {$l_1$, $l_2$, $l_3$,…}of random linear extensions
Let $l_0$ = a starter linear extension constructed in a suitable manner
Set i = 1
loop while (one more linear extension needed)
  generate a random number p between 0 and 1
  if (p < 0.5) then
      $l_i$ = $l_{i-1}$       ($l_i$ is a new linear extension)
  else
      select a random integer j, 1≤j≤n
      k = j+1 mod n
      switch jth and kth elements in $l_{i-1}$ to obtain $l_{test}$
      if ($l_{test}$ is a linear extension ) then
        $l_i$ = $l_{test}$     ($l_i$ is a new linear extension)
  else
      (no new linear extension obtained)
      end if
  end if
  if (a new linear extension was obtained) then
      process the new linear extension - update rank frequency, etc.
  set i = i+1
  end if
end loop
End of Algorithm Bubley-Dyer
```

Bubley and Dyer (1999) show that a uniform and stationary distribution for linear extensions can be obtained and they estimate that the time needed (under certain conditions) is of the order of $|OS|^4$.

As far as some computational aspects of the CRF method are concerned, to ascertain that a reasonably stationary distribution for linear extensions has been reached, one may use a suitable convergence criterion for the assumed convergence conjectured in Patil and Taillie with no counterexample to date. Actually, the computational experience has been that of absolute convergence in practice. The termination of the CRF method requires an eventual total stochastic order. However, random crisscrossing of estimates of CRFs of two objects that are in reality stochastically equivalent (have identical CRF) can continue indefinitely making it impossible to achieve a total stochastic order (ranking without ties). To avoid this situation, one needs to use a rule to declare two CRFs to be identical if their estimates differ from each other by less than a stipulated negligible amount. Finally, for each iteration of the CRF method, one may decide to compute the CRF matrix based on all linear extensions or estimated CRFs using Bubley–Dyer algorithm depending on the number of incomparabilities for the iterative step.

We make a few observations:

1. The CRF method is iterative .The iterative character of the CRF Operator provides a progressive enrichment device. The ranking resulting from each successor iteration is an enrichment of the ranking resulted from its predecessor iteration with CRF square matrix dimension decreasing to its terminating value, resulting from the termination of the iterative process.
2. The CRF method obtains poset ranking from a partial order using stochastic order without computing rank averages.
3. The CRF method is more likely to produce ties.
4. The CRF ranking will be identical to that of AR method if $\prec_1$ is a linear order.

6.2.5 *Poset Ranking with Databased Weights*

Patil (2005) discussess weighting schemes for linear extensions including a possibility of using relative importance of different indicators to assign weights to linear extensions and hence to ranks the linear extensions assign to different objects. Here we explore the idea of obtaining weighting schemes for indicators on the basis of evidence provided by data and use the indicator weights to rank objects. To that end in this section, we do the following: In Sect. 6.2.5.1, we introduce the idea of poset ranking with weights for indicators. In Sect. 6.2.5.2, we conceptualize the idea of databased indicator weights and propose an iterative procedure to compute the same. In Sect. 6.2.5.3, we define databased weighted poset ranking.

6.2.5.1 CRF Method with Weighted Indicators

In this section, we define poset ranking with arbitrary weights for indicators. Assume that we have the object set OS together with the data matrix X as defined earlier and a row vector w = (w_1, w_2, …, w_m) , where $w_j \geq 0$, for j = 1, 2, …, m, and w_1+ w_2 +…+ w_m = 1 representing relative importance of respective indicators. CRF matrices provide a convenient way to introduce indicator weights in the ranking process. Poset ranking with weights for indicators requires weighted indicator CRF matrix (WICRFM). The WICRFM is computed in a way similar to the way the CRF matrix is obtained. The latter is obtained from the rank frequency matrix. The entry in the ith row and the rth column of the rank frequency matrix is the number of *linear extensions* assigning rank r to object O_i. WICRFM is obtained from the weighted rank frequency matrix. The entry in the ith row and the rth column of the weighted rank frequency matrix is the sum of *weights* of those *indicators* that assign rank r to object O_i. A little complication that arises in the computation of the weighted rank frequency matrix is that an indicator can assign ranks that are tied which sometimes can be half-integers. This situation is simplified by treating fractional half-integer ranks as lower integer ranks. Formally, we define for each indicator j, j= 1, 2, … m, the characteristic function

$\mathrm{chf}_j(i,r)$ = 1 if r = the rank of object O_i on the basis of indicator, and = 0 otherwise.

Further, weighted (and normalized) indicator rank frequency (WIRF) function for each object O_i, i = 1, 2, …, n, is defined by

$$\begin{aligned} &\mathrm{f^w i(r)} = \sum \mathrm{jw_j} \mathit{chf}_\mathrm{j}(\mathrm{i,r}) + \sum \mathrm{jw_j} \mathit{chf}_\mathrm{j}(\mathrm{i,r}+1/2), \mathit{summed\ over\ j} \\ &\mathit{for}\, \mathrm{r} = 1,2,\ldots,n \end{aligned} \tag{6.2}$$

where the second summation accounts for tied ranks.

Note that the definition (6.2) of WIRF allows us to use lower integer values for tied ranks with half-integer values.

WIRFM is an n by n matrix ($f^w_i(r)$), i, r = 1, 2, …, n.

Weighted indicator CRF (WICRF)

$$\mathrm{F^w i(r)} = \sum \mathrm{t \leq rf^w i(t)}, \quad \mathit{for}\, \mathrm{i} = 1,2,\ldots,\mathrm{n}. \tag{6.3}$$

In view of the definition (6.2), each $F^w_i(r)$ may have jumps only at integer values of r.

WICRFM is an n by n matrix ($F^w_i(r)$).

In the calculation of WIRFM, we can think of each indicator as a referee assigning ranks to different objects, but each indicator's ranking strength is proportional to its weight, and the entry $f^w_i(r)$ in the ith row and the rth column of WIRFM is the weighted average of the rank, the ith object receives from m indicators.

WICRF ranking is the poset ranking by the CRF method, using WICRFM in place of the original data matrix, without any further reference to the original data matrix. With WICRFM as the data matrix, we have the same n objects, but newly derived n indicators. It may happen that all rows of WICRFM may not be distinct, in which case one needs to work with a quotient set of objects and there may be some ties in the final ranking. Conceptually, a quotient set is not harder to deal with.

Below, we give an explicit algorithm to compute WICRFM. The algorithm uses the rank matrix, R = (r_{ij}), an n by m matrix, where r_{ij} = rank of object O_i with respect to indicator j in the original data matrix. The rank matrix is useful as a descriptive device and a computational tool.

Algorithm 2.5.1.1

Remark:

```
  Algorithm to compute weighted indicator CRF matrix WICRFM(R,
  w)
  Given:
    n = |OS|
    m = |IB|
    rank matrix R = (rij)
    vector w = (w1, w2, …, wm) of indicator weights
  This algorithm computes n by n WICRF matrix F = (fij)
Initialize all elements of F to 0 (zero)
for i = 1 to n {
  for j = 1 to m {
    let r = floor(rij)
    fir = fir + wj
  }
}
Remark: This completes computation of WRFM. Next compute WICRFM
for i = 1 to n {
  for j = 2 to n {
    fij = fi(j-1) + fij
  }
}
End of Algorithm
```

Example 2.5.1 illustrates WICRF ranking using arbitrary indicator weights, including details about how ties are handled.

Example 2.5.1

Consider the data matrix in Table 6.10 and arbitrary indicator weights proportional to 5, 3, and 2 (normalized 0.5, 0.3, 0.2) for indicators q_1, q_2, and q_3, respectively. Table 6.11 shows the rank matrix, and Table 6.12 shows the *unnormalized* weighted rank frequency matrix. In cases of tied ranks with half-integer values, rank frequencies have been calculated using the lower integer values, using expression (6.2) above. If all indicators have equal weights, then unweighted, unnormalized rank frequencies are straight counts of votes. If indicators have unequal weights, then a vote for a given rank for an object by an indicator is counted as many times as the indicator weight indicates. If an object has not received rank r from any of the three

Table 6.10 Data matrix for Example 2.5.1

Object	q_1	q_2	q_3
a	13	9	8
b	8	9	9
c	2	11	7
d	16	14	12
e	5	5	9
f	14	3	13
g	14	16	16

Table 6.11 Rank matrix for Example 2.5.1

Object	q_1	q_2	q_3
a	4	4.5	6
b	5	4.5	4.5
c	7	3	7
d	1	2	3
e	6	6	4.5
f	2.5	7	2
g	2.5	1	1

Table 6.12 Weighted rank frequency matrix with indicator weights of 5, 3, and 2

	Rank						
Object	1	2	3	4	5	6	7
a	0	0	0	8	0	2	0
b	0	0	0	5	5	0	0
c	0	0	3	0	0	0	7
d	5	3	2	0	0	0	0
e	0	0	0	2	0	8	0
f	0	7	0	0	0	0	3
g	5	5	0	0	0	0	0

indicators, then the entry in the corresponding cell in Table 6.12 is zero. If an object has received rank r from at least one indicator then the entry in that cell in Table 6.12 is not zero. The magnitude of a nonzero entry in a cell for an object in the column for rank *r* is the sum of weights of those indicators that give rank r to the object. In view of definition (6.2), half-integer ranks (which might occur in case of a tie) are effectively counted as lower integer ranks. Thus, for example, object a has received rank 4 from indicator q_1 whose weight is 5 and rank 4.5 from indicator q_2 whose weight is 3. We treat rank 4.5 as if it was 4. Thus, effectively, for this calculation, object a has received rank 4 from q_1 and from q_2. Hence the entry in the cell for *a* and rank 4 is 5+3 = 8. Table 6.13 shows the weighted indicator CRF matrix. Entries in Table 6.13 are obtained by calculating cumulative sums of entries in cells, from left to right, for each given row. Each row in Table 6.13 is a weighted cumulative rank frequency function (Fig. 6.6).

Table 6.13 Weighted cumulative rank frequency matrix obtained from Table 6.12

	Rank						
Object	1	2	3	4	5	6	7
a	0	0	0	8	8	10	10
b	0	0	0	5	10	10	10
c	0	0	3	3	3	3	10
d	5	8	10	10	10	10	10
e	0	0	0	2	2	10	10
f	0	7	7	7	7	7	10
g	5	10	10	10	10	10	10

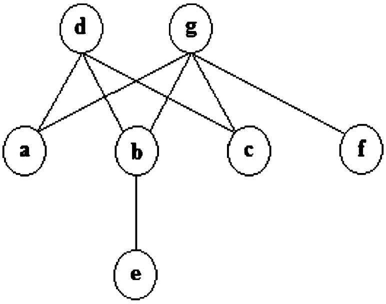


Fig. 6.6 Hasse diagram for Table 6.10

Table 6.14 Ranking for data in Example 2.5.1 by different methods

	Ranks by various methods						
Object	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
a	6	5.5	5	4.5	4.5	4	5.5
b	4	3.5	3	4.5	4.5	5	3
c	5	5.5	6	6	6	6	5.5
d	2	2	2	2	2	2	2
e	7	7	7	7	7	7	7
f	3	3.5	4	3	3	3	4
g	1	1	1	1	1	1	1

(i) CRF-based poset ranks from the original data matrix; uses MCMC sampling
(ii) CRF-based poset ranks from the original data matrix; uses all 132 linear extensions
(iii) WICRF ranks with equal weights
(iv) WICRF ranks with weights 0.5, 0.3, 0.2 for three indicators
(v) DBWP ranks (WICRF ranks with databased weights). Poset ranking with databased weights is discussed below
(vi) Ranks produced by the composite index constructed using indicator weights of 0.3551, 0.3032, and 0.3417 for q_1, q_2, and q_3, respectively. These are databased weights corresponding to DBWP ranks mentioned in (v) above
(vii) Average ranks from the original data matrix; uses all 132 linear extensions

Column (iv) of Table 6.14 is WICRF ranking. It is obtained by the CRF method using Table 6.13 as the data matrix without any more reference to data in Table 6.10. Column (v) of Table 6.14 gives WICRF ranking with databased weights. Other columns of Table 6.14 show rankings by other methods.

In the following section, we will discuss derivation of improved indicator weights, given the original data matrix and WICRFM that was obtained using an arbitrary weight vector.

6.2.5.2 Databased Weights

Poset ranking with arbitrary indicator weights described in Sect. 6.2.5.1 in itself may be viewed as a legitimate approach to ranking, since it makes allowance for a subjective input, utilizing information external to the data matrix. Our aim is here, however, to develop indicator weights that are based on relative importance of indicators as exhibited by their values in the data matrix. We call them databased weights. Poset ranking obtained from WICRFM computed from databased weights rather than arbitrary indicator weights will be called databased weighted poset ranking (DBWP ranking or DBWPR). We will measure importance of an indicator in terms of how close ranks assigned by the indicator to objects are to ranks assigned by linear extensions used in CRF poset ranking. These linear extensions are based on WICRFM. We will measure agreement between ranks assigned by an indicator and ranks collectively assigned by linear extensions by means of cumulative rank correlation between the column vector of the original rank matrix corresponding to the indicator and linear extensions. More specifics of databased weights are detailed below.

We assume we have object set OS, information base IB, a data matrix X, and associated rank matrix R, both n by m. We also assume that with some arbitrary prior weight vector wo (for w old), we have obtained WICRF matrix, say, F, as described in the previous subsection. For simplicity, we assume all n rows of F are distinct. If all rows are not distinct, then we work with the corresponding quotient set. We describe here a method to obtain a new posterior vector wn = (wn_1, wn_2, …, wn_m) (for w new) of weights for indicators from R and F in a way that will take into account contribution of each indicator to ranking the object set. To do this, we consider (OS, $\prec_F$), where $\prec_F$ is the partial order based on WICRFM F and all linear extensions of $\prec_F$. We measure relative importance of an indicator by its closeness to linear extensions of $\prec_F$. Closeness of a particular linear extension lf of $\prec_F$ with an indicator q_j is measured by $(1+\mathrm{corr}(lf,q_j))/2$, where $\mathrm{corr}(lf, q_j)$ is correlation between object ranks defined by lf and the jth column vector of the original rank matrix R. We use $(1+\mathrm{corr}(lf,\mathrm{q}_\mathrm{j}))/2$ to assure nonnegative values for the closeness measure. The overall closeness $\mathrm{CL}(\prec_\mathrm{F}, q_j)$ of $\prec_F$ to indicator q_j is the sum of $(1+\mathrm{corr}(lf,q_j))/2$ over the set of all linear extensions of $\prec_F$. Thus, wn_j is given by

$$wn_j = \frac{CL\left(\prec_F, q_j\right)}{\sum_j CL\left(\prec_F, q_j\right)}.$$

In general, the vector wn may be different from wo. But if wn is identical to wo, then we conclude that the rank matrix, and hence effectively the data matrix evidence, supports wo in the sense that the closeness agreement between the rank matrix and the linear extensions based on WICRFM(R, wo) reproduces the same weight vector. In such a case, we propose that the vector wn computed as above is the cumulative correlation iterate of wo, denoted as CCI(wo), and state the following as the definition of databased weights.

Definition 2.5.2.1

Given the object set OS, information base IB, the data matrix X, and an m-dimensional row vector w = (w_1, w_2, …, w_m), the weights w_1, w_2, …, w_m are databased weights or, synonymously, databased indicator weights (DBIW), if the cumulative correlation iterate of w is equal to w.

Next, given OS, IB, and X, we propose an iterative procedure to compute databased weights when they exist. The procedure starts with computation of WICRFM with equal indicator weights, which is the same as CRFM. Let the initial vector of weights be denoted by w^0. Then the procedure computes iteratively $w\nu^{+1}$ = CCI($w\nu$) for $\nu = 0, 1, 2, \ldots$until two successive iteration weight vectors are equal, subject to some assumed tolerance. First we give the algorithm to compute CCI(w).

Algorithm 2.5.2.1

Remark:

```
      Algorithm to compute iterated indicator weight vector wn = CCI(wo).
      Given:
            n = |OS|
            m = |IB|
            rank matrix R = (rij)
      This algorithm will compute the new weight vector wn = (wn1, wn2, …, wnm)
      For this version, we will use all linear extensions based on F. If number of incomparabilities is prohibitively large, then the algorithm needs to be modified by using Bubley-Dyer sampling together with a convergence criterion, such as Cauchy criterion, for convergence of the weight vector to be computed, instead of using all linear extensions based on F.
  Compute F = WICRFM(R, wo) using Algorithm 2.5.1.1
  Initialize each component wnj of w to 0
  Compute (OS, ≺F) in some form (e.g., a zeta matrix) that can facilitate generation of linear extensions of ≺F.
  for each linear extension lf of ≺F {
      for j = 1 to m {
          wnj = wnj + (1+corr(r.j, lf))/2 where corr(r.j, lf) is Spearman rank correlation coefficient between the jth column r.j of R and lf.
      }
  }
  Compute t = Σj wnj
  for j = 1 to m {
      wnj/t
  }
```

```
End of algorithm
```

Now we present a procedure to compute databased indicator weights.

Procedure 2.5.2.1

Remark:

```
    Procedure to compute databased indicator weights DBIW
    This procedure will compute the compute data based indicator weights
    Given:
          n = |OS|
          m = |IB|
          rank matrix R = (rij)
  Select suitable tolerance
  Set vector w0 = (1/m, 1/m, …, 1/m)
  Set ν = 0
  Repeat {
      wν+1 = CCI (wν)
  until (|wν+1 -wν| < tolerance) where |wν+1 - wν| measures difference between wν+1 and wν.
  wν+1 is the desired vector
  End of procedure.
```

Existence of databased weights as defined by Definition 2.5.2.1 is not guaranteed, since it depends on the numerical entry values of the data matrix. If the databased weights do not exist, the above procedure to compute databased weights will not yield a converging sequence {wν}. However, for all incidental examples presented in this chapter, the procedure has produced convergent sequences of weight vectors. Moreover, computed limiting vectors have been found to be independent of starting vectors of weights. This leads us to believe that, under some general conditions, databased weights exist, and are unique, and that the procedure described above yields databased weights subject to the assumed tolerance. At this point, these general conditions are not known. An alternative course of action when databased weights as defined do not exist for a given numerical data matrix also deserves further investigation.

6.2.5.3 Databased Weighted Poset Ranks

Given the OS, IB, *X*, for which databased weights exist, the WICRFM computed from databased weights as derived by the described procedure above will be denoted by DBWICRFM. We are now in a position to define databased weighted poset ranks.

Definition 2.5.3.2

Given the object set OS, information base IB, and the data matrix *X*, if a vector *w* of databased weights exists, then poset ranks of OS based on the DBWICRFM using CRF method are defined as databased weighted poset ranks (DBWPR) of OS.

Table 6.15 Data matrix for Example 2.5.2

Object	q_1	q_2
a	9	3
b	2	7
c	6	1
d	6	4
e	7	4
f	9	6
g	3	1
h	1	1
k	8	5

Table 6.16 Four iterations of poset ranks for Example 2.5.2

Object	0	1	2	3
a	3	3	3	3
b	5	6	6	6
c	7	7	7	7
d	6	5	5	5
e	4	4	4	4
f	1	1	1	1
g	8	8	8	8
h	9	9	9	9
k	2	2	2	2

Table 6.17 Four iterations of data-based weights for Example 2.5.2

Iteration	0	1	2	3
w_1	0.5000	0.5176	0.5370	0.5370
w_2	0.5000	0.4824	0.4630	0.4630

It may be pointed out that computation of DBWPR is solely based on DBWICRFM as the data matrix without making any reference to the original data matrix *X*. Following examples will illustrate these concepts and the method.

Example 2.5.2

Table 6.15 shows a data matrix with nine objects and two indicators. Table 6.17 shows results of application of Procedure 2.5.2.1 to obtain databased weights with three iterates of the starting vector of equal weights. Databased weights are in rightmost column of Table 6.17. Details of calculations are too lengthy to be included here. However, we illustrate calculations involved in iteration 0 and derivation of indicator weights for iteration 1 in Appendix 2. We also computed poset ranks from WICRFMs based on iterates of the weight vector. They are shown in Table 6.16.

It should be noted that ranks based on the composite index constructed using the databased indicator weights in column 3 of Table 6.17 are not guaranteed to be

Table 6.18 Four iterations of DBWP ranks for Example 2.5.3

Object	0	1	2	3
a	5	4.5	4.5	4.5
b	3	4.5	4.5	4.5
c	6	6	6	6
d	2	2	2	2
e	7	7	7	7
f	4	3	3	3
g	1	1	1	1

Table 6.19 Four iterations of data-based indicator in Table 6.10 weights for Example 2.5.3

Iteration	0	1	2	3
w_1	0.3333	0.3470	0.3551	0.3551
w_2	0.3333	0.3132	0.3032	0.3032
w_3	0.3333	0.3398	0.3417	0.3417

identical to the DBWP ranks. Interestingly, for the current example, however, for the composite index based on the weights in column 3 of Table 6.17 ranks are the same as in column 3 of Table 6.16.

Example 2.5.3

We use the data matrix of Example 2.5.1. It is shown in Table 6.10. Tables 6.18 and 6.19 show, respectively, results of four iterations needed to compute DBWP ranks and associated databased weights. Although conceptually this example is not different from the previous one, its results will be used in later sections on representability and equivalence of composite indicators. Table 6.19 shows results of application of Procedure 2.5.2.1 to obtain databased weights with three iterates of the starting vector of equal weights. Databased weights are in rightmost column of Table 6.19. Poset ranks from WICRFMs based on iterates of the weight vector are shown in Table 6.18.

We shall present another example with 11 by 3 data matrix in the next section in the context of representability.

With the objective of reconciliation between object ranks and indicator weights determined by the two approaches, namely, poset ranking and composite index, we need to first investigate if DBWP ranks can be replicated by a composite index. This is done in the next section.

6.3 Representability of Databased Weighted Poset Ranking

With poset ranking in hand for a set of objects on the basis of a data matrix and the derived databased weights consistent with the data matrix, a natural query arises, as to whether this poset ranking can be reproduced by a composite index.

Also of interest is the kind of weights it is composed of. This is of critical interest because the weights signify the importance of the individual indicators relative to each other and the implicit trade-offs among them. This query acquires particular significance, since a successful response will allow us to compare expert/stakeholder weights of their subjective composite index with the weights of the composite index generating the databased poset ranking, when such a corresponding composite index should exist.

In this section then, we try to answer the following questions: Given a poset ranking of a set of objects on the basis of a data matrix, does there exist a composite index that will induce ranking of objects that is identical to poset ranking? If such a composite index exists, how do we construct it? And, if we can, is it unique? The latter part of the question will be discussed in Sect. 6.4. To address the earlier part, we need the following notation and terminology:

A composite index for the object set OS with data matrix X and the row vector of weights w will be denoted by CI(OS,X, w). Ranking based on the composite index will be denoted by CIR(OS,X, w).

Definition 3.1

Representability of poset ranking: Poset ranking DBWPR(OS,X) is representable if there is a composite index CI(OS,X,w) with ranking CIR(OS,X,w) identical to DBWPR(OS,X).

Definition 3.2

Equivalence of Weights: Given a set of objects OS and data matrix X, two weight vectors w and w^* are equivalent if CIR(OS, X, w) = CIR(OS, X, w^*).

In Sect. 6.4, we explore flexibility to vary weights of a given composite index without altering ranking it determines. These results together with representability help us to know the extent to which reconciliation is possible between databased poset ranking and stakeholder weights based on composite index-induced ranking.

Specifically, we show that DBWPR(OS,X) is representable if and only if a certain linear programming problem admits a solution. In what follows, we provide some discussion and formulation needed for the construction of a desired system of inequalities amenable to linear optimization processes.

Consider a set of objects OS = $\{O_i \mid i=1, 2, 3,\ldots, n\}$ with the associated data matrix $X = (x_{ij})$ with n rows for n objects and m columns, column j representing values for the objects of the indicator q_j, for $j = 1, 2, \ldots, m$. Assume OS has been ranked by DBWP ranking mechanism, assigning rank r_i to object O_i for $i = 1, 2, \ldots, n$. Although the presence of a tie does not essentially change conclusion, just to keep language simple, assume that there are no ties. We can relax this assumption later. By our convention, objects with higher indicator values receive higher ranks, rank 1 being the highest. Suppose now that this ranking is representable, so that there is a composite index CI with indicator weights $w_1, w_2, \ldots, w_m$ which also assigns rank r_i to object O_i for $i = 1, 2, \ldots, n$. Consider then a matrix Y obtained from X by permuting rows of X such that if $r_i = j$, then the ith row of X becomes the jth row of Y. Thus, in particular, the row for the object with DBWP rank of 1 which is also to be ranked 1 by the composite indicator becomes row 1 of Y. In general, if $r_i < r_j$, then the row for object O_i will be placed higher than the row for object O_j in Y. If, now, we denote by w the row vector $(w_1, w_2, \ldots, w_m)$ and consider the column vector Yw^{T}, the n elements of the column vector Yw^{T} are to be the n index values of the composite index under construction for the n objects. They are in descending order.

Now, define matrix $D = (d_{ij})$ with $n-1$ rows and m columns where

$$d_{ij} = y_{ij} - y_{(i+1)j} \quad for\, j = 1,2,\ldots,m \;\; and\, i = 1,2,\ldots,n-1,$$

and call D, the difference matrix or just the D matrix.

Let us now consider the column vector Dw^{T} whose elements are

$$\sum d_{ij} w_j = \sum (y_{ij} - y_{(i+1)j}) w_j = \sum y_{ij} w_j - \sum y_{(i+1)j} w_j > 0 \quad for\, i = 1,2,\ldots,n-1.$$

Since the elements of Yw^{T} are in descending order, $Dw^{\mathrm{T}} > 0$ where 0 is an $(n-1)$ dimensional vector whose each component is 0 and the inequality indicates componentwise inequality.

Thus we are in search of $w_i \geq 0$, $i = 1, 2, \ldots, m$, that satisfy the constraints $w_1 + w_2 + \ldots + w_m = 1$, and $Dw^{\mathrm{T}} > 0$, which can be relaxed to $Dw^{\mathrm{T}} \geq 0$ to be able to exploit linear programming capabilities.

From this, we clearly see that, if DBWP ranking is representable by a composite indicator CI, then the weights for the composite indicator need to be a solution to a linear programming problem. On the other hand, if optimization of an objective function, subject to the constraints:

$$Dw^{\mathrm{T}} \geq 0 \tag{6.4}$$

$$w_i \geq 0, \quad i = 1,2,\ldots,m \tag{6.5}$$

$$w_1 + w_2 + \ldots + w_m = 1, \tag{6.6}$$

admits a solution, then there exists a composite indicator whose ranking agrees with DBWP ranking to one or more ties in view of the relaxed constraints $Dw^{\mathrm{T}} \geq 0$. Thus, to seek a composite indicator that replicates DBWP ranking, we need to attempt to solve a linear programming problem as defined above with a suitable objective function. However, agreement is only up to a tie since an optimizing solution occurs at a vertex of the solution space, where at least one of the constraints is met with equality. Why an optimizing solution will produce a composite index with at least one tie is explained in Sect. 6.4, where we study equivalence of composite indicators. Our success of finding a composite indicator in complete agreement with our ranking depends on our ability to find several optimizing solutions with different objective functions, so that their convex combinations will provide a composite indicator with complete agreement with DBWPR. Incidentally, it turns out that if there is one composite indicator with desired properties, there are many equivalent indicators with identical rankings, but with different weight vectors. This is discussed in Sect. 6.4.

If the DBWP ranking contains one or more ties, then it implies that one or more components of Dw^{T} are zeros. Because the weight vector being sought needs to produce identical composite indicator values for objects that are tied and the corresponding rows of matrix X will be next to each other in the matrix Y. In this case to obtain a composite indicator CI in total agreement with DBWP ranking, we will need to stipulate one or more specific constraints of the type $\Sigma d_{ij}w_j = 0$ for a single suitable i, or more corresponding to pairs of objects with tied ranks in DBWPR.

Importance of results obtained above is that it makes us available a number of existing powerful computational tools in our investigation of representability and can help one choose a particular weighting scheme from among equivalent weights that suits one's requirements by using a suitable objective function.

We will illustrate the idea with two examples below. For both examples, multiple solutions are obtained using a variety of linear objective functions. These multiple solutions help us in two ways. First, they help us identify the entire representability region (discussed in Sect. 6.4). Secondly, they can be used by stakeholders as an aid in their overall analysis regarding an appropriate choice of weights for constructing a composite index, such as constructing a composite index that assigns maximal weight to a given indicator. Nonlinear objective functions ($-\sum w_j$*$\ln(w_j)$, etc.) listed in Tables 6.22 and 6.25 can also be used when we are interested in constructing a composite index with most equal or most unequal weights for indicators.

Example 3.1

Table 6.20 shows the 9 by 2 data matrix of Example 2.5.2 sorted by DBWP ranks with databased weights. Its D matrix is shown in Table 6.21. Its DBWP ranking is representable. Weights, appropriately normalized, corresponding to different objective functions are tabulated in Table 6.22.

Solutions in the first four rows of Table 6.22 were obtained under the constraints, listed below as (6.7):

$$\begin{array}{cl}
w_1 + w_2 & \geq 0 \\
-w_1 + 2w_2 & \geq 0 \\
2w_1 - w_2 & \geq 0 \\
w_1 & \geq 0 \\
4w_1 - 3w_2 & \geq 0 \\
-4w_1 + 6w_2 & \geq 0 \\
3w_1 & \geq 0 \\
2w_1 & \geq 0 \\
w_1 & \geq 0 \\
w_2 & \geq 0 \\
w_1 + w_2 & = 1
\end{array} \tag{6.7}$$

Table 6.20 Data matrix for Example 3.1 sorted by DBWP Ranks

Object	q_1	q_2	DBWPR
f	9	6	1
k	8	5	2
a	9	3	3
e	7	4	4
d	6	4	5
b	2	7	6
c	6	1	7
g	3	1	8
h	1	1	9

Table 6.21 *D* Matrix for data in Table 6.20

q_1	q_2
1	1
−1	2
2	−1
1	0
4	−3
−4	6
3	0
2	0

Table 6.22 Weights for data in Table 6.20 for various objective functions

Objective function	w_1	w_2
w_1, maximize	0.6000	0.4000
w_1, minimize	0.4286	0.5714
w_2, maximize	0.4286	0.5714
w_2, minimize	0.6000	0.4000
$-\sum w_j{*}\ln(w_j)$, max	0.5000	0.5000
$-\sum w_j{*}\ln(w_j)$, min	0.6000	0.4000
$\sum w_j$^2, minimize	0.5000	0.5000
$\sum w_j$^2, maximize	0.6000	0.4000

First eight of the above constraints are just $Dw^{\mathrm{T}} \geq 0$. Three of these constraints happen to be redundant. The last three constraints are standard constraints to assure weights are nonnegative and that they add up to 1.

Example 3.2

Table 6.23 shows the 7 by 3 data matrix of Example 2.5.3 sorted by DBWP ranks with databased weights, the matrix *Y*. Its *D* matrix as defined in (6.4) is shown in Table 6.24. Its DBWP ranking is representable. Weights, appropriately normalized, corresponding to different objective functions are tabulated in Table 6.25.

Table 6.23 Data matrix for Example 3.2 sorted by DBWP ranks

Object	q_1	q_2	q_3	DBWPR
g	14	16	16	1
d	16	14	12	2
f	14	3	13	3
a	13	9	8	4.5
b	8	9	9	4.5
c	2	11	7	6
e	5	5	9	7

Table 6.24 *D* Matrix for data in Table 6.23

q_1	q_2	q_3
−2	2	4
2	11	−1
1	−6	5
5	0	−1
6	−2	2
−3	6	−2

Table 6.25 Weights for data in Table 6.23 for various objective functions

Objective function	Objective	w_1	w_2	w_3	Ref. Fig. 6.9
w_1	Maximize	0.418605	0.302326	0.27907	*R*
w_1	Minimize	0.096774	0.419355	0.483871	*S*
w_2	Maximize	0.096774	0.419355	0.483871	*S*
w_2	Minimize	0.122449	0.265306	0.612245	*Q*
w_3	Maximize	0.122449	0.265306	0.612245	*Q*
w_3	Minimize	0.418605	0.302326	0.27907	*R*
$-\sum w_j{*}\ln(w_j)$	Maximize	0.33334	0.33333	0.33333	*C*
$-\sum w_j{*}\ln(w_j)$	Minimize	0.122449	0.265306	0.612245	*Q*
$\sum w_j$^2	Minimize	0.33334	0.33333	0.33333	*C*
$\sum w_j$^2	Maximize	0.122449	0.265306	0.612245	*Q*

For Example 3.2, constraints $Dw^{\mathrm{T}} \geq 0$ become

$$\begin{aligned}
-2w_1 + 2w_2 + 4w_3 &\geq 0\\
2w_1 + 11w_2 - w_1 &\geq 0\\
w_1 - 6w_2 + 5w_3 &\geq 0\\
5w_1 - w_3 &\geq 0\\
6w_1 - 2w_2 + 2w_3 &\geq 0\\
-3w_1 + 6w_2 - 2w_3 &\geq 0
\end{aligned} \tag{6.8}$$

Solutions in the first six rows of Table 6.25 were obtained with all above constraints.

Actually, DBWPR contains one tie, namely, the tie between objects *a* and *b* which are tied at rank 4.5. To obtain the vector w that will define a composite indicator producing a tie between objects *a* and *b*, we need to set up constraints as

$$\begin{array}{ll} -2w_1 + 2w_2 + 4w_3 & \geq 0 \\ 2w_1 + 11w_2 - w_1 & \geq 0 \\ w_1 - 6w_2 + 5w_3 & \geq 0 \\ 5w_1 - w_3 & = 0 \\ 6w_1 - 2w_2 + 2w_3 & \geq 0 \\ -3w_1 + 6w_2 - 2w_3 & \geq 0 \end{array} \tag{6.9}$$

These constraints yield solutions 2 through 5 (vectors *S* and *Q*) in Table 6.25, of which two are distinct. Any nontrivial convex combination of these two distinct solutions gives us a weight vector for a composite indicator whose ranking is identical to DBWPR.

As the above examples indicate, when a weighted poset ranking is representable, multiple solutions to the problem exist. This prompts us to seek all solutions, thus possibly affording us a wider choice of weights in support of a particular weighting scheme. More generally, it is of interest to investigate all equivalence classes of weights. This investigation is presented in the next section.

6.4 Geometry of Composite Indicators and Equivalent Weights

Consider an *n* (objects) by *m* (indicators) data matrix $X = (x_{ij})$ and two *m-dimensional* row vectors *w* and *w** of weights. To begin with, suppose that elements of the vector Xw^{T} and those of $Xw^{*\mathrm{T}}$ are distinct. Recall *w* and *w** are equivalent if the composite indicators CI(OS, *X*, *w*) = CI(OS, *X*, *w**). Given any vector of weights *w*, is it possible to find another vector of weights *w** that is equivalent to *w*? Answer is yes, except in an intriguing posetic situation, for example, when the entire poset is an antichain with all n row vectors (objects) of the data matrix being coplanar in an (*m*-1)-dimensional hyperplane of the *m*-dimensional indicator space.

Interestingly, to investigate the answer to the question posed, consider all possible rankings of *n* objects. There are *n*! of them. If each of them is representable, an unlikely case, there would be a maximum of *n*! composite indicators. However, since the weight space is infinite, one can imagine infinitely many composite indicators. Hence, it is possible that some composite indicators will have infinitely many equivalent surrogates. Actually, if given a composite indicator, it has another equivalent composite indicator, and then it has infinitely many equivalents, since the composite indicator is continuous with respect to each individual weight. We wish to identify all weight vectors *w** that are equivalent to *w*.

To do this, it is helpful to interpret geometrically the vector Xw^T in the m-dimensional Euclidean space. Let n rows of X be denoted by row vectors $X_1, X_2, \ldots, X_n$. These n vectors are points in the m-dimensional Euclidean space, representing n objects. Each point $w = (w_1, w_2, \ldots, w_m)$ is a point in the hyperplane or the 'weight space' W defined by $w_1+w_2+,\ldots, +w_m = 1$. For each X_i, the composite indicator value for the ith object in the dataset, $X_i \bullet w^T$ is the projection (shrunk by the factor $|w|$) of X_i on the radius vector passing through the point w.

To see how a change in w affects the composite indicator, consider any two object vectors X_i and X_j that are not comparable. We see that $X_i \bullet w^T = X_j \bullet w^T$ defines a hyperplane passing through the origin such that for any w in W composite indicator values for objects i and j are equal. Further, this hyperplane divides W into two regions such that for w in one region $X_i \bullet w^T > X_j \bullet w^T$ and in the other $X_i \bullet w^T < X_j \bullet w^T$. The collection of all such dividing hyperplanes partitions W into at most 2^c convex mutually disjoint subregions where c is the number of incomparable pairs of objects.

For any w belonging to a given subregion in the partition, composite indicator values for all objects are distinct. Moreover, any two points w_1 and w_2 that are interior to such a subregion are equivalent. A point w along a border of a subregion will produce a tie with respect to the composite indicator. There is no dividing hyperplane corresponding to any pair of objects when the two objects are comparable. Thus for any data matrix with m indicators, it is always possible to partition the m-dimensional hyperplane $w_1 + w_2 + \ldots + w_m = 1$ with $w_j \geq 0$ into regions of equivalent weights, such that composite indices based on two weight vectors belonging to the same equivalent region will produce identical rankings of the n objects, and the composite indices based on two weight vectors belonging to two different regions will produce different rankings.

Figure 6.7 illustrates the concept for the number of indicators, $m=2$. Here the line segment RS represents the line $w_1+w_2=1$. X_i and X_j are vectors of indicator values of two incomparable objects i and j, respectively. $V=(v_1, v_2)$ is an arbitrary weight vector. With v_1 and v_2 as weights, the value of the composite indicator for the object i is the length OP, the projection of OX_i on OA, shrunk by the factor $|V|$, and that for object j is the length OQ, the projection of OX_j on OA, shrunk by the same factor $|V|$. Thus the composite indicator will rank object j higher than the object i. The ray OB is perpendicular to the line joining X_i and X_j. It intersects the line segment RS at $U=(u_1,u_2)$. It is clear by construction that with u_1 and u_2 as weights, values of the composite indicator for object i and j will be equal.

Further, the composite indicator with weights defined by any point between U and R will rank object j higher than object i and the composite indicator with weights defined by any point between U and S will rank object i higher than object j. Point U is the critical point for objects i and j that serves as the point of reversal of ranking of the two objects. If two objects were comparable, no such critical point would exist, since one object would always be ranked higher than the other for any set of weights. One can spot a point of reversal of ranking for each pair of incomparable objects in the data matrix. The collection of such points partitions the segment RS in intervals of equivalent weights. Two composite indicators based on any pair of weight vectors belonging to a given interval will produce identical ranking of the objects.

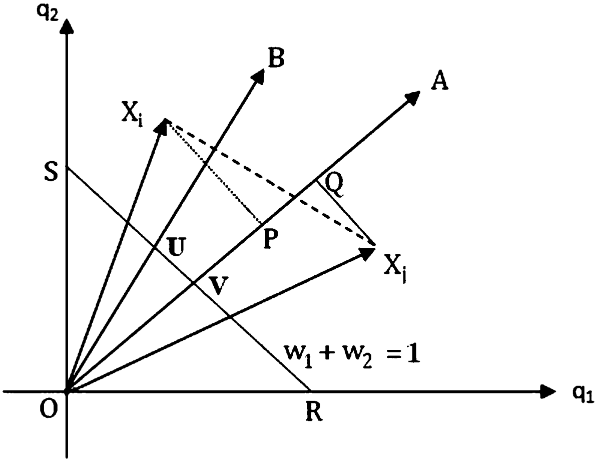


Fig. 6.7 Division of weight space for a two-indicator data matrix

When DBWP ranking is representable as a composite indicator, then the region consisting of all equivalent weights producing the composite indicator will be referred to as the representability region.

We present below two examples of such partitions, one with a two-column data matrix and another with a three-column data matrix.

Example 4.1

Consider the 9 by 2 data matrix of Example 3.1. Points X_2, X_3, and X_4 represent objects *b*, *c*, and *d*, respectively, of the data matrix. Other unidentified objects of Example 3.1 are represented by cross marks in the area near X_2, X_3, X_4. In order to keep in view indicator values as well as the weight space conveniently in a reasonably sized figure, indicator values were transformed by dividing values in Table 6.20 by 10 and then by adding 1. Thus, for example, coordinates of X_2 (object *b*) are (1.2, 1.7). For the data matrix, databased weighted poset ranking is representable. Figure 6.8 shows the partition of the weight space $w_1 + w_2 = 1$ defined by cross marks on the line segment RS. Each cross mark on the line segment RS is the point of intersection of RS with the perpendicular (not shown) to the line segment (not shown) joining a pair of incomparable objects. For example, *U* is the intersection of the perpendicular to the line segment connecting X_2 and X_4 with RS and *V* is the intersection of the perpendicular to the line segment joining the data points X_2 and X_3 with RS.

The composite index defined by weights at a cross mark produces a tie between the corresponding pair of objects. The open interval (*U*, *V*) is the representability region. Composite indicator with weight vector *U* (0.4286, 0.5714) produces a tie between *b* and *d*. Composite vector with weight vector *V* (0.6000, 0.4000) produces a tie between objects *b* and *c*.

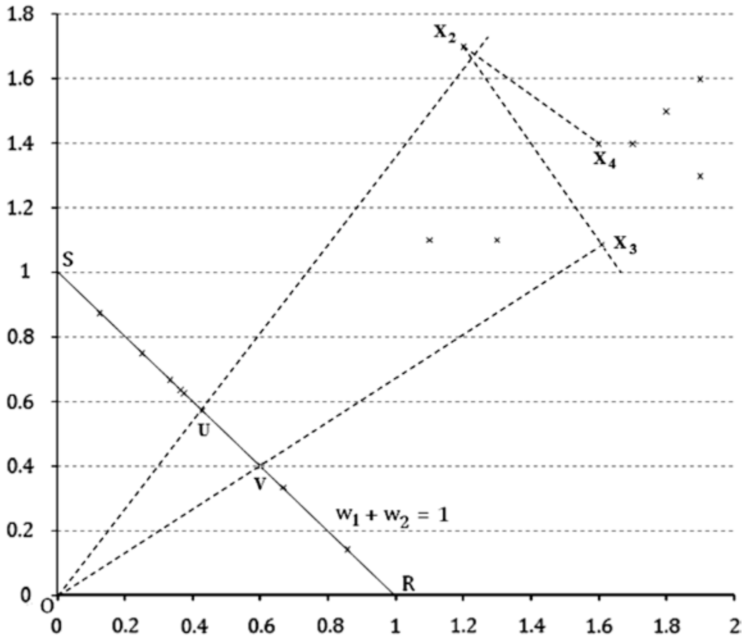


Fig. 6.8 Partition of weight space for data matrix of Example 3.1. Data points X_2, X_3, X_4 represent objects *b*, *c*, and *d* of Example 3.1. Other objects are represented by *cross marks* in the area near X_2, X_3, and X_4. *Hash marks* on the segment RS are points of intersection of perpendiculars to line segments joining points representing objects that are not comparable. The ray passing through points *O* and *U* is perpendicular to the line segment joining X_2 to X_4. Open interval (*U*, *V*) is the representability region. Composite indicator with weight vector *U* ties *b* with *d*. Composite indicator with weight vector *V* ties *b* with *c*

Example 4.2

Figure 6.9 shows the partition of the weight space for the 7 by 3 data matrix of Example 3.2. It is in three dimensions since data matrix has three indicators. The triangle $w_1 + w_2 + w_3 =1$ represents the weight space. The weight space is seen divided by many lines. Each line is the intersection of the plane $w_1 + w_2 + w_3 =1$ with another plane that is perpendicular to line segment that joins two points and passes through the origin. These two points represent indicator values of two incomparable objects. If we form a composite indicator with weights on this line of intersection then the composite indicator will assign equal ranks to the two objects involved. For example, any composite indicator with weights on the line that contains points *Q* and *S* will guarantee tied rank of 4.5 for two objects *a* and *b*. The two objects *a* and *b* themselves could not be depicted in the diagram. As is well known in discrete geometry of convex polytopes and their properties (Grunbaum 1967), all these lines of intersection divide the triangle forming the weight space into convex regions. Any composite indicator with weights that are in the interior of any convex region will assign distinct ranks to all n objects without any tie. Moreover any two

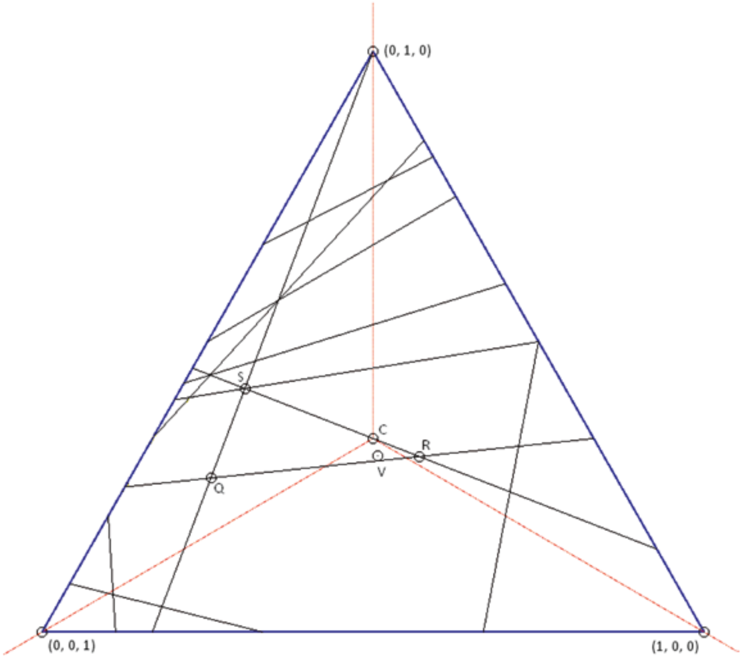


Fig. 6.9 Partition of the weight space into sets of equivalent weight vectors for the data matrix of Example 3.2

composite indicators whose weights are in the interior of the same convex region will produce mutually identical rankings. For geometric reference, point *C* is the centroid of the weight –space with $w_1 = w_2 = w_3 = 1/3$. The segment QS represents the set of weight vectors each of which would produce the tie as per DBWPR. Points *Q* and *S* are solutions to the linear programming problems to maximize the objective functions w_3 and w_2, respectively, subject to constraints (6.9) of Sect. 6.3. To again identify points in Table 6.25, *Q* is (0.1225, 0.2653, 0.6122), *R* is (0.4186, 0.3023, 0.2791), and *S* is (0.0968, 0.4193, 0.4839). Point *V* is the vector (0.3551, 0.3032, 0.3417) of databased weights. *Q* is where weights are most unequal. *C* is where weights are most equal if one does not insist that objects *a* and *b* need be tied. *S* is where weights are most equal if we stipulate objects *a* and *b* be tied.

6.5 Approximate Representability and Reconciliation

A major objective of this chapter has been to seek reconciliation between databased poset CRF ranking and ranking induced by an arbitrary subjective composite index. The concept of representability of DBWP ranking can be helpful in this respect because it widens the choice of potentially satisfactory indicator weights. Further, the representability of ranking as defined is able to provide a nice visual geometric representation in two- and three-dimensional cases and provides a decent algebraic multidimensional vector space structure dealing with hyperplanes based on half-spaces and polytopes. Representability makes identification of ranking with an equivalent region in the Euclidian weight space. However, if DBWP is not representable, we may wish to seek approximate representability for DBWP ranking. And thus, we may wish to find a composite indicator that ranks objects close to the DBWP ranking. Hence, we investigate approximate representability for potential reconciliation. In the situation, where DBWPR is not representable, we wish to construct a composite indicator that will produce ranking having high correlation with DBWPR. Below we propose an approach to construct a composite indicator that is closest to DBWPR in terms of correlation coefficient between the two rankings involved.

6.5.1 Approximate Representability Using Constraint Relaxation

Recall from Sect. 6.3 that given DBWPR(OS, X) there exists a composite indicator that induces ranking CIR(OS, X, w) identical to DBWPR(OS, X) if and only if linear programming problem subject to constraints $Dw^{\mathrm{T}} \geq 0$ admits a solution. Then it is natural to investigate, in the absence of an exact solution to the linear programming problem, to see if relaxation of one or more constraints from $Dw^{\mathrm{T}} \geq 0$ will lead to an approximate solution leading to a "satisfactory" composite indicator. Since we do not know what our choices are at this stage, it is not possible to define what satisfactory means a priori. But we know there is at least one solution since, in the extreme case, removal of all constraints guarantees a solution.

Theoretically, consider the following scenario. $Dw^{\mathrm{T}} \geq 0$ specifies at most $n-1$ constraints. One can relax any of the at most 2^{n-1} combinations of these constraints to obtain one or more solutions each time. In view of equivalency of composite indicators, there are finitely many such solutions. Of these solutions, one chooses a solution whose ranking has maximum correlation with DBWPR. We define this as an approximate solution and say that DBWPR is approximately representable as the corresponding composite indicator. A method such as branch and bound (Aho et al. 1983) can be used to search for this approximate solution. The following example illustrates the point.

Example 5.1.1

Consider the data matrix in Table 6.26. Its DBWP ranking, given in the fifth column, is not representable. Figure 6.10 shows partition of the weight space in regions of equivalent composite indicators for the data matrix. W in Fig. 6.10 is the vector of databased indicator weights. Table 6.27 shows rows of the data matrix sorted according to DBWPR. Table 6.28 shows the D matrix.

Having seen that DBWPR is not representable, a plausible strategy to relax constraints is to remove as few constraints as possible since relaxing more constraints is likely to arrive at composite indicator whose ranking will differ more from

Table 6.26 Data matrix of Example 5.1 DBWP ranking is not representable

Object	q_1	q_2	q_3	DBWP ranks
1	12	14	14	3
2	6	2	12	9
3	11	15	15	2
4	10	2	9	8
5	5	6	7	11
6	12	14	16	1
7	14	10	4	7
8	15	9	9	5
9	12	7	14	5
10	12	11	9	5
11	2	4	8	10

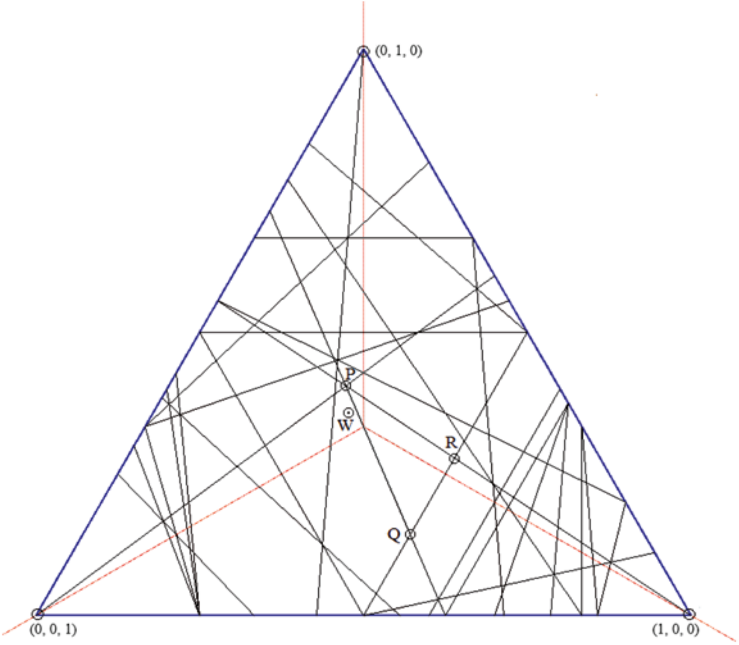


Fig. 6.10 Partition weight space for data in Table 6.26

DBWPR. Thus, to start with, we relax single constraints, then constraints in pairs, then in triples, and so forth. For this small example, it can be verified that relaxing the constraint $-3w_1-2w_2+1w_1\geq 0$ produces three solutions identified in Fig. 6.10 as points *P*, *Q*, and *R*.

Rankings of objects by three composite indicators with weight vectors *P* (0.2703, 0.4054, 0.3243), *Q* (0.5000, 0.1429, 0.3571), and *R* (0.5000, 0.2778, 0.2222) are shown in Table 6.29. Correlation coefficients of DBWPR with ranks defined by

Table 6.27 Data matrix in Table 6.26 sorted by DBWPR

Object	q_1	q_2	q_3	DBWP
6	12	14	16	1
3	11	15	15	2
1	12	14	14	3
8	15	9	9	5
9	12	7	14	5
10	12	11	9	5
7	14	10	4	7
4	10	2	9	8
2	6	2	12	9
11	2	4	8	10
5	5	6	7	11

Table 6.28 *D* matrix for data matrix in Table 6.26

q_1	q_2	q_3
1	−1	1
−1	1	1
−3	5	5
3	2	−5
0	−4	5
−2	1	5
4	8	−5
4	0	−3
4	−2	4
−3	−2	1

composite indicators with weight vectors P, Q, and R, respectively, are 0.983931, 0.981577, and 0.981577.

Table 6.29 contains ranking due to different composite indicators based on different weight vectors appearing in Fig. 6.10. CIR(OS, X, DBIW) is ranking due to the composite index obtained from DBIW (databased indicator weights) ranking.

6.5.2 Reconciliation with Stakeholders Index

In many practical situations ranking is based on composite index with weights proposed by stakeholders on some subjective basis. Such a ranking may need to be reconciled with poset ranking, which is based on evidence supported by data. A favorable situation may occur when DBWP ranking is representable and stakeholder ranking is identical with DBWP ranking. This is possible only when stakeholder weights belong to the representability region. In this, reconciliation is implicit and already accomplished and the stakeholders have further choice to choose weights that are most consistent with their consideration without altering ranking. Similarly, if DBWP ranking is not representable and stakeholder weights

Table 6.29 Comparison of non-representable DBWPR with ranking by constraint relaxation

	Indicators			DBWPR	CIR(OS, *X*,	Vertices obtained by constraint relaxation		
Object	q_1	q_2	q_3		DBIW)	*P*	*Q*	*R*
1	12	14	14	3	3	3	2.5	2.5
2	6	2	12	9	9	9	9	9
3	11	15	15	2	2	2	2.5	2.5
4	10	2	9	8	8	8	8	8
5	5	6	7	11	10	10	10	10
6	12	14	16	1	1	1	1	1
7	14	10	4	7	7	7	7	7
8	15	9	9	5	6	4	4.5	4
9	12	7	14	5	5	5.5	4.5	5.5
10	12	11	9	5	4	5.5	6	5.5
11	2	4	8	10	11	11	11	11
Correlation coefficient with DBWPR					0.98169	0.98393	0.98158	0.98158

DBWPR, Ranking induced by composite indicator with databased indicator weights and induced by composite indicators based on approximate representability. *P*, *Q*, and *R* are the vertices of approximate representability region by constraint relaxation. Bottom row shows correlation of respective rankings with DBWPRW is the databased indicator weight vector

belong to the approximate representability region, we have approximate reconciliation. If stakeholder weights belong to a region adjoining the representability region, then stakeholders may be persuaded to use weights on the common border. If that is done, then too reconciliation is accomplished. A more difficult situation is when stakeholder weights are in a region distant from the region of representability or approximate representatability. In this case, we measure closeness of stakeholder weights with DBWP ranking by means of the correlation coefficient between stakeholder ranking and DBWP ranking. If the correlation coefficient is statistically significant then we may consider the conflict resolvable and consider the situation as statistically significant reconciliation. However, if the correlation is not statistically significant then stakeholder weights to be recommended need to be moved following the shortest path towards the representability region through intervening equivalent regions. The shortest path is defined as the straight line joining the centroid of the region of the stakeholder's weights and the centroid of the representability region. Each move along the straight path is from one region into the next region.

As a linear move is made into a new region, the correlation between the DBWP ranking and ranking defined by the new position of stakeholder's proposed weights is monitored for statistical significance. When for the first time correlation becomes significant in the region, the movement stops. At that point, the stakeholder has a choice of weights in the stopping region as reconciliation. If no weight in the stopping region is acceptable, then it is necessary for advocates of both approaches to revisit and reexamine the data matrix, data reliability, and the appropriateness of indicators and their weights for the intangible latent concept underlying the desired basis of ranking and prioritization. Figure 6.11 illustrates the concept. It uses the data of Example 2.5.3. Point *A* represents the initial stakeholder weights. The points *B*, *C*, and *D* show the progress of the moving stakeholder weights. *M* is the midpoint of the line segment QS, which is the

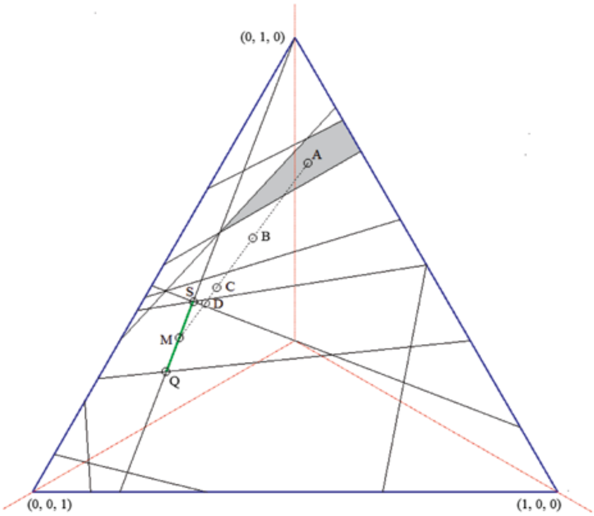


Fig. 6.11 Reconciliation process. Stakeholder equivalent region is *shaded light*. DBWP ranking representability region is the segment QS

Table 6.30 Reconciliation process

Position	w_1	w_2	w_3	Corr. coeff.
A	0.1608	0.7242	0.1150	0.7207*
B	0.1784	0.5889	0.2327	0.7748*
C	0.2207	0.4701	0.3092	0.8829**
D	0.2468	0.3893	0.3639	0.9370

An asterisk (*) and a double asterisks (**) in the right-most column of Table 6.30 indicate that respective correlation coefficients are significant at 5% and 1% level of significance, respectively.

representability region. Weight vectors as well as the corresponding correlations for the respective positions are shown in Table 6.30.

6.6 Conclusion

In this chapter, we have proposed a new method of poset ranking to rank and prioritize a set of objects on the basis of multiple indicator values. While it avoids some common shortcomings of a purely index-based approach requiring subjective weights for indicators, it also goes beyond average rank ordering approach implicitly assuming equal weights for indicators. It attempts to objectively assess relative importance of different indicators in the ranking process. This evaluation is used to compute indicator weights in ranking objects based on indicator values.

Further, we have investigated exact and approximate representability of desired data-validated ranking in terms of a composite index. This investigation leads to the concept of equivalence of composite indicators. This in turn allows us to seek reconciliation of index-based ranking with poset ranking. The concept of databased weighted poset ranking introduced here may open doors to still other ways of weighting schemes and other reconciliation approaches for comparative knowledge discovery using partial orders and composite indicators.

Knowledge Discovery in Databases (KDD) results from exploring data in order to discover previously unknown patterns. Comparative Knowledge Discovery (CKD) needs to result when interest lies in discovering previously unknown ranking patterns and related comparative issues of consequence underlying multivariate datasets/ matrices for purposes of preferred selection, decision, prioritization, policy, etc.

In this research and outreach chapter, we have introduced some basic concepts and methods, illustrating with small datasets/matrices in the spirit of start small even for big data. Meaningful ability to deal with big data concept wise, methods wise, computations wise, and visualizations wise in a next best challenge for comprehensive research with live case studies is an urgent need of comparative knowledge discovery with partial orders and composite indicators in this infometrical computer science and software engineering age of statistical information science and technology. This chapter is prepared in the spirit of a concept paper for digital age infometrics and comparative knowledge discovery critical in several fields, such as document discovery, drug discovery, gene discovery, chemical discovery, criminal discovery, critical area discovery, etc. The ranking, prioritization, and selection of objects and indicators carrying a variety of names in a variety of contemporary issues of societal and scientific importance based on relevant evidence embodied in corresponding data matrices provide insightful leads in these substantive investigations involving variously big data.

6.7 Appendix 1

The following is a list of abbreviations/acronyms used in this chapter:

AR	Average rank
CCI	Cumulative correlation iterate
CI	Composite index
CIR	Composite index induced ranks or ranking
CKD	Comparative knowledge discovery
CRF	Cumulative rank frequency
DBIW	Databased indicator weights
DBWICRFM	Databased weighted indicator CRF matrix
DBWP	Databased weighted poset
DBWPR	DBWP rank(s)
KDD	Knowledge discovery in databases
LE	Set of linear extension
MCMC	Markov Chain Monte Carlo
WIRF	Weighted indicator rank frequency
WICRF	Weighted indicator cumulative rank frequency
WICRFM	Weighted indicator cumulative rank frequency matrix

6.8 Appendix 2

In this appendix we show detailed calculations of poset ranks for iteration 0 where we assume equal indicator weights. Indicator weights for iteration 1 are computed from the cumulative rank frequency matrix that was used to compute poset ranks for iteration 0 (Tables 6.31, 6.32, 6.33, and 6.34). We use the data from Example 2.5.2

Number of Objects: 9

Number of Attributes: 2

Applying the CRF operator to the above matrix we obtain the poset ranks for iteration 0 as shown in Table 6.35.

Indicator weights for iteration 1 are computed from the 9×9 data matrix in Table 6.34 above. To do this, we compute all ten CRF-linear extensions shown in Table 6.36:

Ranks of objects in each of these linear extensions as shown in Table 6.37.

Table 6.38 shows computed correlation coefficients between ranks assigned by individual linear extensions to objects and columns of the rank matrix.

Table 6.31 Data matrix for Example 2.5.2

Object	q_1	q_2
a	9	3
b	2	7
c	6	1
d	6	4
e	7	4
f	9	6
g	3	1
h	1	1
k	8	5

Table 6.32 Rank matrix for Example 2.5.2

Object	q_1	q_2
a	1.5	6
b	8	1
c	5.5	8
d	5.5	4.5
e	4	4.5
f	1.5	2
g	7	8
h	9	8
k	3	3

Table 6.33 Rank frequency matrix with equal weights

	Rank								
Object	1	2	3	4	5	6	7	8	9
a	1	0	0	0	0	1	0	0	0
b	1	0	0	0	0	0	0	1	0
c	0	0	0	0	1	0	0	1	0
d	0	0	0	1	1	0	0	0	0
e	0	0	0	2	0	0	0	0	0
f	1	1	0	0	0	0	0	0	0
g	0	0	0	0	0	0	1	1	0
h	0	0	0	0	0	0	0	1	1
k	0	0	2	0	0	0	0	0	0

Table 6.34 Cumulative rank frequency matrix with equal indicator weights

	Rank								
Object	1	2	3	4	5	6	7	8	9
a	1	1	1	1	1	2	2	2	2
b	1	1	1	1	1	1	1	2	2
c	0	0	0	0	1	1	1	2	2
d	0	0	0	1	2	2	2	2	2
e	0	0	0	2	2	2	2	2	2
f	1	2	2	2	2	2	2	2	2
g	0	0	0	0	0	0	1	2	2
h	0	0	0	0	0	0	0	1	2
k	0	0	2	2	2	2	2	2	2

Table 6.35 Poset ranks

Object	Rank
a	3
b	5
c	7
d	6
e	4
f	1
g	8
h	9
k	2

Table 6.36 Linear extensions

	Linear extension number									
Rank	1	2	3	4	5	6	7	8	9	10
1	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*f*
2	*k*	*k*	*k*	*k*	*k*	*k*	*a*	*a*	*a*	*a*
3	*a*	*a*	*a*	*e*	*e*	*e*	*k*	*k*	*k*	*b*
4	*e*	*e*	*b*	*a*	*a*	*d*	*e*	*e*	*b*	*k*
5	*b*	*d*	*e*	*b*	*d*	*a*	*b*	*d*	*e*	*e*
6	*d*	*b*	*d*	*d*	*b*	*b*	*d*	*b*	*d*	*d*
7	*c*	*c*	*c*	*c*	*c*	*c*	*c*	*c*	*c*	*c*
8	*g*	*g*	*g*	*g*	*g*	*g*	*g*	*g*	*g*	*g*
9	*h*	*h*	*h*	*h*	*h*	*h*	*h*	*h*	*h*	*h*

Table 6.37 Ranks assigned by linear extensions to objects

Object	Rank									
a	3	3	3	4	4	5	2	2	2	2
b	5	6	4	5	6	6	5	6	4	3
c	7	7	7	7	7	7	7	7	7	7
d	6	5	6	6	5	4	6	5	6	6
e	4	4	5	3	3	3	4	4	5	5
f	1	1	1	1	1	1	1	1	1	1
g	8	8	8	8	8	8	8	8	8	8
h	9	9	9	9	9	9	9	9	9	9
k	2	2	2	2	2	2	3	3	3	4

Table 6.38 Computation for indicator weights for iteration 1

	Correlation with columns of rank matrix		(1 + Correlation)/2	
Linear extension #	q_1	q_2	q_1	q_2
1	0.8656	0.7321	0.9328	0.8660
2	0.9076	0.6725	0.9538	0.8362
3	0.7983	0.7917	0.8992	0.8958
4	0.8236	0.7576	0.9118	0.8788
5	0.8656	0.6980	0.9328	0.8490
6	0.7983	0.7236	0.8992	0.8618
7	0.8908	0.6810	0.9454	0.8405
8	0.9328	0.6214	0.9664	0.8107
9	0.8236	0.7406	0.9118	0.8703
10	0.7395	0.7746	0.8698	0.8873
Total	8.4457	7.1931	9.2228	8.5966
Normalizer				17.8194
Weights for next iteration			0.5176	0.4824

References

Aho AV, Hopcroft JE, Ullmann JD (1983) Data structures and algorithms. Addison-Wesley, Reading, MA, 427 pp

Brüggemann R, Patil GP (2011) Ranking and prioritization with multi-indicator systems, introduction to partial order and its applications. Springer, New York, NY, 337 pp

Brüggemann R, Schwaiger J, Negele RD (1995) Applying Hasse diagram technique for the evaluation of toxicological fish tests. Chemosphere 30(9):1767–1780

Bubley R, Dyer M (1999) Faster random generation of linear extensions. Discrete Math 201:81–88

Grunbaum B (1967) Convex polytopes. Wiley, New York, NY, 456 pp

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multi-criteria ranking and prioritization. Environ Ecol Stat 11:199–228

Patil and Tailie (2004) Geo-informatic hotspot systems (GHS) for detection, prioritization, and early warning. In: Proceedings of the national conference on digital government research 2005, DGRC

Winkler P (1982) Average height in a partially ordered set. Discrete Math 39:337–341

Chapter 7
A Software Platform Towards a Comparison of Cars: A Case Study for Handling Ratio-Based Decisions

Jochen Wittmann and Rainer Brüggemann

Abstract Environmental aspects often are in conflict with the criteria for a best/optimal behavior under technical aspects only. In these cases, common methods to compare different options and to come to a decision show methodological disadvantages. In this situation, this chapter intends to demonstrate the situation firstly by giving a typical example, secondly to show, how a software platform might support the decision, thirdly, to provide different methods for comparison to demonstrate the effects of the method chosen, and finally to sensitize the users for the interdependencies between the comparison method and the resulting ranking. The example will deal with the decision to find a new car according to individually scalable ratios. General data on different cars mainly are in conflict with the ratio of CO_2 expressing the environmental aspects of the cars to select. The chapter proposes a software platform that allows dealing with these conflicting parameters by individually weighting and a flexible interface for comparison.

7.1 The Motivation: Growing Relevance of Environmental Ratios for Decisions

In many areas the environmental impacts of a decision have to be taken into account. Therefore, the decision maker has to perform complicated comparisons. These comparisons are executed on the base of data on the objects under observation.

J. Wittmann (✉)
HTW Berlin, University of Applied Sciences, Environmental Informatics,
Wilhelminenhofstraße 75A, 12459 Berlin, Germany
e-mail: Jochen.Wittmann@HTW-Berlin.de

R. Brüggemann
Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology
and Inland Fisheries, Müggelseedamm 310, 12587 Berlin, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_7, © Springer Science+Business Media New York 2014

Very often, the data on the objects are aggregated by ratios and a simple order concerning the resulting ratios shows the way to the optimal decision. Common examples for this tendency are the growing number of applications of ratios like the CO_2 footprint, the financial rating, the index I_{geo}, etc.

However, environmental aspects often are in conflict with the criteria for a best/optimal behavior under technical aspects only. In these cases, common methods to compare different options and to come to a decision show methodological disadvantages. Optimization theory deals these problems under the topic of "multicriterial optimization" and tries to handle the problem by introducing sophisticated weights and special algebraic operations within the objective function [see, e.g., Munda (2008), and to mention one typical and well-known method, PROMETHEE: Brans and Vincke (1985)]. Common intention of all these approaches is to reduce the decision making to the one-dimensional value of an objective function. With this as precondition, the common optimization methods can be applied. Even the modern concepts for heuristic optimization such as genetic strategies and evolutionary algorithms depend on this assumption (see, e.g., the original papers of Goldberg (1989) concerning the genetic algorithms and Schwefel (1995) for the evolutionary algorithms).

In this situation, the critic of the authors focuses on the fact that there are often technical features under observation with values that tend to be good, if they are bigger and faster but with the very simple effect that any increase in these values the ecological impact will increase too. The decision depends strongly on the different weights; the ecological contribution factors of the objective function get in comparison to the weights of the technical criteria (having composite indicators (CI), weighted sums of attribute values in mind). At the end, the optimization result offers the single value ratio as the objective function produces, but for the user the origin of the optimization result and its dependence and its sensitivity in regard to the weights is hidden. Especially for sensible political decisions, such a proceeding seems highly critical because of its missing transparency. In short, these optimizations hide the political decision necessary.

To make a decision more transparent, the decision maker must have the opportunity to play with the weights for the criteria easily and he should have the chance to see the influence of technical criteria on the optimization result separated from the influence of the (mostly) reluctant environmental criteria. In this situation, this chapter intends:

(a) To demonstrate the situation by giving a typical but generally understandable example: The conflict between the criteria that has been explained generally comes up characteristically in connection of the personal decision "Which car shall I buy?." The technical criteria are acceleration, maximum speed, number of passengers, load capacity, … Typical ecological criteria are fuel consumption, CO_2 emissions. It is obvious that these two types of criteria effect in opposite directions concerning the objective function, e.g., a high velocity will cause high emissions and one will not get a car that is first ranked in both.

(b) To show how a software platform might support the decision. To make just this decision situation as transparent as possible for the decision maker, a user interface for a decision support software tool is presented in Chap. 2.

(c) To provide different methods for comparison to demonstrate the effects of the method chosen.
(d) To sensitize the users for the interdependencies between the comparison method and the resulting ranking.

So the main objective of the software proposed is to provide flexibility to formulate and to parameterize the decision problem to learn about the significance of the results, i.e., the resulting ranking. This is the content of the first section of this chapter.

In the second section the method of partial ordered set shall be applied to this problem with the intention to give the users some deeper understanding about the weights and their influence on the ranking results. On the other hand, the analysis gained by partial order methods could be used to interpret the ranking and to find critical thresholds for the weighting parameters.

Thus, we do not have the intention to make a decision automatically, but we try to elaborate some findings that help to understand an already existing ranking and to examine it concerning sensitivity and stability.

7.2 The Example: The Comparison of Cars

The example will deal with the decision to find a new car according to individually scalable ratios. General data on different cars mainly are in conflict with the ratio of CO_2 expressing the environmental aspects of the cars to select.

The chapter proposes a software platform that allows to deal with these conflicting parameters by individually weighting and a flexible interface for comparison.

7.2.1 Specification of the Software Tool

The software tool was programmed using the Microsoft .net Framework which assists the developer in the efficient software development. The programming language used is C#. The basic functionalities can be grouped into three categories: administration, look up and compare, and finally search. Each functionality in these categories is presented to the user through a unique form.

7.2.1.1 Administration

Everything that has do to with the information base can be altered through this module. There is one form showing the directory of the program and the name and connection properties of the database. The latter can be altered here.

Another form is dedicated to the information itself. The user can add and edit existing brands and car models as well as define properties of a car. These properties

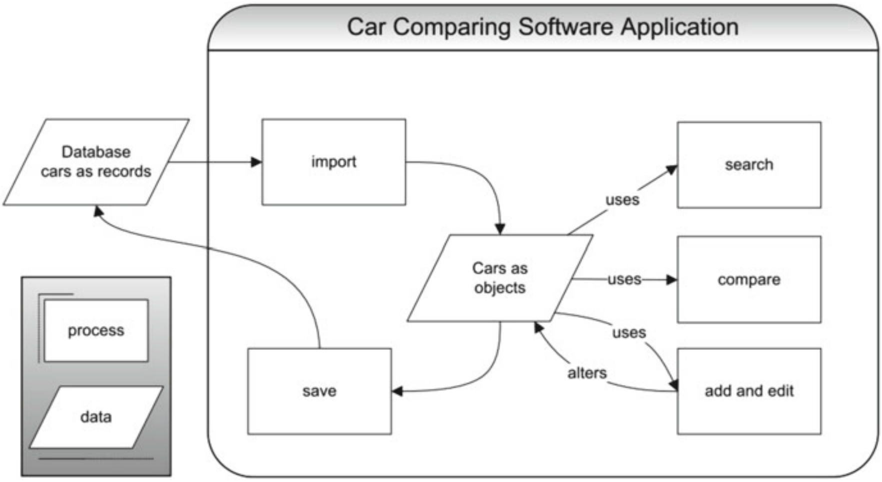


Fig. 7.1 The structure of the software tool

are: brand, model name, type, CO_2 emissions (g/km), fuel consumption (l/100 km), fuel type, power (kW), emission standard, Ncap rating, price (€), taxes p.a. (€), and luggage capacity (l). Since cars have to share the same attributes in order to be comparable changes made in this section affect all cars.

Adding or editing a specific model, the user has to select certain attribute values from drop-down lists. This is done due to the need for a consistent information base since a little typing error would result in a car that cannot be found or be compared with correctly. The additions or changes made will be saved to the database on leaving the form.

The following chart (Fig. 7.1) shows how information is handled and stored.

7.2.1.2 Compare and Search

There are three methods available for the user to comparing the cars: compare by chosen attributes, compare by priority, and a direct comparison of selected cars. The differences and routines of the methods are explained to detail in Sect. 7.2.2 “User specific weighting and method results.”

7.2.1.3 Look Up

On one hand, the user can browse a list containing all cars and all their respective attributes stored in the database. On the other hand, a custom list can be created containing only the attributes the user is interested in. The program uses the database to fill the drop-down lists that are used for attribute selection.

7.2.2 User Specific Weighting and Method Results

Special interest lies on the individual weighting algorithm. The weighting itself is not a mathematically new approach, but the user-friendly integration into the software application especially with respect to the great number of single attributes of a car that potentially have to be weighted implies some methodical and software-technological deliberations.

Different methods for comparing cars have been implemented up front due to the fact that multiple scenarios are possible. For example, a user might not want to compare all cars available but has made up his mind and needs assistance in comparing two or three predetermined models.

Two of the methods are based around comparing all cars that are stored in the database. Starting point for these methods is the selection of attributes and definition of their priority for the weighting process by the user. The implemented algorithms then compare all cars based on the information provided. This process is divided into several steps:

(a) The range or scale of each criterion is being determined through finding the highest and lowest value.
(b) Every attribute of every car is given a value representing its position on the scale.
(c) The values of each criterion (normalized) are multiplied by the corresponding factor of the priorities (i.e., weights) the user has chosen.
(d) The final values for each car are added up and a ranked list is prompted to the user.

However, there are some differences distinguishing the two methods.

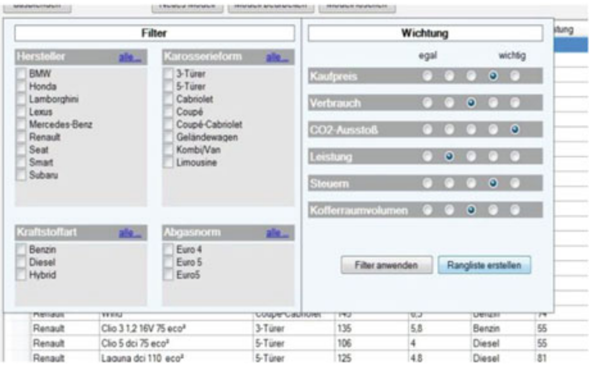
7.2.2.1 Comparison on Behalf of Ranked Criteria

This method is based on how relevant each attribute is to the user's decision-making process. To determine this, the user is asked to rank the attributes by selecting their importance to him (Fig. 7.2). All cars are then compared to each other. The main difference to the comparison on behalf of selected attributes is the normalization that takes place during the comparison whereby the original ranking is translated into a percentage "scale of meeting the requirements" assigning a certain percentage to each criteria. This method ensures that attributes with the same priority do get the same percentage assigned to them.

Due to the mass of data that needs to be shown to keep the result as transparent as possible, the output is text based which makes it difficult for the user to get an instant overview, unfortunately.

7.2.2.2 Comparison on Behalf of Selected Attributes

For running this method, the user selects three different attributes that he wants to base his decision on. Then the cars' attributes are compared as described above.



Filter

Hersteller alle...

- □ BMW
- □ Honda
- □ Lamborghini
- □ Lexus
- □ Mercedes-Benz
- □ Renault
- □ Seat
- □ Smart
- □ Subaru

Karosserieform alle...

- □ 3-Türer
- □ 5-Türer
- □ Cabriolet
- □ Coupé
- □ Coupé-Cabriolet
- □ Geländewagen
- □ Kombi/Van
- □ Limousine

Kraftstoffart alle...

- □ Benzin
- □ Diesel
- □ Hybrid

Abgasnorm alle...

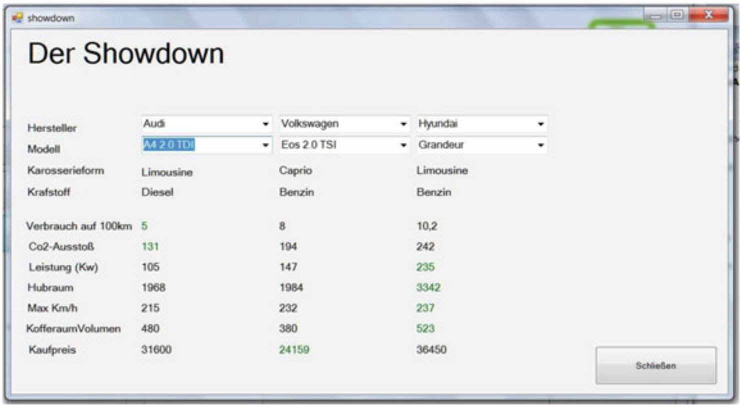
- □ Euro 4
- □ Euro 5
- □ Euro5

Wichtung

	egal				wichtig
Kaufpreis	○	○	○	●	○
Verbrauch	○	○	●	○	○
CO2-Ausstoß	○	○	○	○	●
Leistung	○	●	○	○	○
Steuern	○	○	○	●	○
Kofferraumvolumen	○	○	●	○	○

Renault	Clio 3 1.2 16V 75 eco²	3-Türer	135	5,8	Benzin	55
Renault	Clio 5 dci 75 eco²	5-Türer	106	4	Diesel	55
Renault	Laguna dci 110 eco²	5-Türer	125	4.8	Diesel	81

Fig. 7.2 Weighting



showdown

Der Showdown

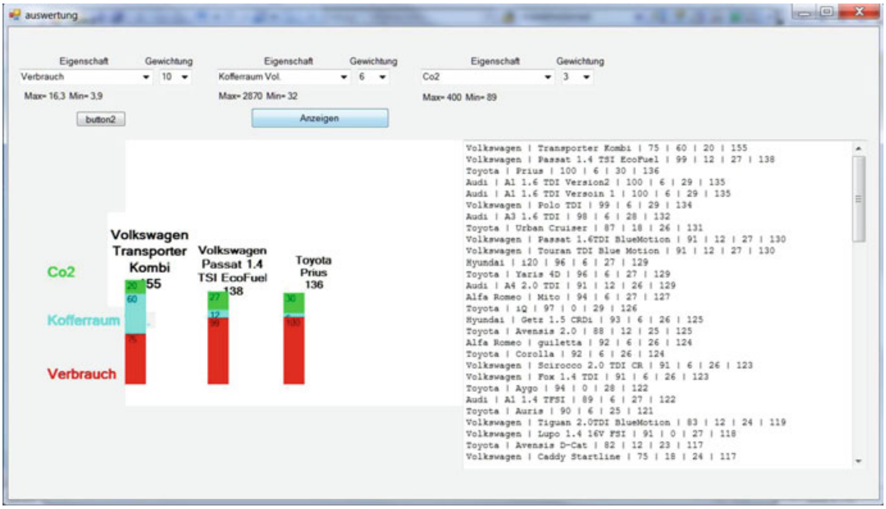
Hersteller	Audi	Volkswagen	Hyundai
Modell	A4 2.0 TDI	Eos 2.0 TSI	Grandeur
Karosserieform	Limousine	Caprio	Limousine
Krafstoff	Diesel	Benzin	Benzin
Verbrauch auf 100km	5	8	10,2
Co2-Ausstoß	131	194	242
Leistung (Kw)	105	147	235
Hubraum	1968	1984	3342
Max Km/h	215	232	237
KofferaumVolumen	480	380	523
Kaufpreis	31600	24159	36450

Fig. 7.3 Screenshot of the method "shutdown"

Each attribute is assigned a value representing its relative position on the scale for that attribute. Combined with the prioritization factor, the attributes can reach a value between 1 and 100. When the algorithm finishes assigning the values, the attributes of the best three cars are combined by stacking each car's attributes in a bar chart. Each vertical bar represents one car. Hereby, the overall position of one car is clearly visible (see also Sect. 7.2.3 for a screenshot).

7.2.2.3 Comparison of Predetermined Models

The third method is called Showdown (see screenshot in Fig. 7.3). It provides a form in which the user can choose up to three different cars, and the program displays all properties of them. The program compares the properties and displays the best result of each category in a light green font.



Eigenschaft: Verbrauch Gewichtung: 10
Max= 16.3 Min= 3.9

Eigenschaft: Kofferraum Vol. Gewichtung: 6
Max= 2870 Min= 32

Eigenschaft: Co2 Gewichtung: 3
Max= 400 Min= 89

```
Volkswagen | Transporter Kombi | 75 | 60 | 20 | 155
Volkswagen | Passat 1.4 TSI EcoFuel | 99 | 12 | 27 | 138
Toyota | Prius | 100 | 6 | 30 | 136
Audi | A1 1.6 TDI Version2 | 100 | 6 | 29 | 135
Audi | A1 1.6 TDI Versoin 1 | 100 | 6 | 29 | 135
Volkswagen | Polo TDI | 99 | 6 | 29 | 134
Audi | A3 1.6 TDI | 98 | 6 | 28 | 132
Toyota | Urban Cruiser | 87 | 18 | 26 | 131
Volkswagen | Passat 1.6TDI BlueMotion | 91 | 12 | 27 | 130
Volkswagen | Touran TDI Blue Motion | 91 | 12 | 27 | 130
Hyundai | i20 | 96 | 6 | 27 | 129
Toyota | Yaris 4D | 96 | 6 | 27 | 129
Audi | A4 2.0 TDI | 91 | 12 | 26 | 129
Alfa Romeo | Mito | 94 | 6 | 27 | 127
Toyota | iQ | 97 | 0 | 29 | 126
Hyundai | Getz 1.5 CRDi | 93 | 6 | 26 | 125
Toyota | Avensis 2.0 | 88 | 12 | 25 | 125
Alfa Romeo | guiletta | 92 | 6 | 26 | 124
Toyota | Corolla | 92 | 6 | 26 | 124
Volkswagen | Scirocco 2.0 TDI CR | 91 | 6 | 26 | 123
Volkswagen | Fox 1.4 TDI | 91 | 6 | 26 | 123
Toyota | Aygo | 94 | 0 | 28 | 122
Audi | A1 1.4 TFSI | 89 | 6 | 27 | 122
Toyota | Auris | 90 | 6 | 25 | 121
Volkswagen | Tiguan 2.0TDI BlueMotion | 83 | 12 | 24 | 119
Volkswagen | Lupo 1.4 16V FSI | 91 | 0 | 27 | 118
Toyota | Avensis D-Cat | 82 | 12 | 23 | 117
Volkswagen | Caddy Startline | 75 | 18 | 24 | 117
```

Fig. 7.4 Screenshot of the method "chart"

7.2.3 Some Results

The results will be sketched by an example with the corresponding screenshots of the software package. These results show the conflict for the decision between environmental relevant attributes and the classical ones for rating a car. Therefore, the approach of partially ordered set is applied to the decision problem in the following.

To find the best three cars with a user-defined weighting of the properties, the user can use the form "chart" (see screenshot in Fig. 7.4). Therein, he chooses the three properties which are most important to him. For every property he also can set a factor of importance. The number *zero* means least important attribute; *ten* means most important one.

The program looks for the best and the worst value for each property. With this information it creates a value range. Now the program looks at the value of the property from that special cars that are on that range and assesses it with points from null to ten. This mark will be multiplied with the users weighting factor. This number creates the high of the bars. At the end the sum of all property numbers became calculated. On the left side the properties are shown in the same color like the associated bars. Over the bars the names of the producer and the model are displayed.

7.3 Partially Ordered Sets as Decision Support

So far, the user support given by the three methods cannot give any insight into the interdependencies of the attributes:

- Comparison by weighted attributes
- Comparison by single selected attributes
- Comparison by predetermined models

7.3.1 Introduction to Partial Order Methods

Due to this problem for the user to get an overview on the set of cars under observation together with the conflicting criteria, a nonstandard method has been tested for the decision process. It is the idea of partially ordered set theory (see for instance Brüggemann and Patil 2010, 2011). In the following, we give a short introduction into this:

Suppose C is a set of cars. A comparison between elements of this set shall be pairwise and fulfill the following restrictions:

- Reflexivity: Each car shall be comparable with itself.
- Antisymmetry: If an Audi type x is better than an Audi type y, and y shall be better than x simultaneously, then it must follow $x = y$, i.e., x and y are equal. In our application we are relaxing to: x and y are equivalent (in sign: $x \cong y$).
- Transitivity: If a BMW type x is better than a BMW type y, and BMW type y is better than a BMW type z, then it shall apply: BMW type x is better than BMW type z.

A binary operation between two objects (in this case the two car types) that is reflexive, anti-symmetric, and transitive is a partial order. For our application the partial order shall be defined as follows:

With $q_i(x)$ the value of the i-th attribute of an object x follows: $x < y : q_i(x) \leq q_i(y)$ for $i = 1, 2 \ldots m$, with at least one attribute q_{i^*}, with $q_{i^*}(x) < q_{i^*}(y)$.

The m attributes of the objects have to be considered simultaneously. From a theoretical point of view, by the relaxation to equivalence, instead of equality we are dealing with a quasi-order instead of a partial order.

Furthermore the concept of chains is important: Chains are subsets of C such that each element is comparable with each other.

If there is a partial order for a given set of n objects, the set is called a partially ordered set (or poset). A graphical representation of partially ordered sets is the so-called Hasse diagram. Hasse diagrams are good visualizations for sets with a small number of elements n but tend to be not clearly arranged if the number of elements increases (see Brüggemann and Patil 2010, 2011; Myers and Patil 2008; Newlin and Patil 2010).

The analysis of posets on the base of the partial order defined above is called Hasse diagram Technique (HDT) (Brüggemann et al. 2001; Brüggemann and Voigt 2008). The role of HDT in decision support is recently discussed by Brüggemann and Carlsen (2012). The HDT provides a huge collection of methods that are each quite simple but tend to be sumptuous if the number of elements in the poset increases. Therefore, special software packages have been developed to support the HDT, such as WHASSE (Brüggemann et al. 1999), DART (Manganaro et al. 2008), PRORANK (Pudenz 2005; Voigt et al. 2006), and recently PyHasse (Brüggemann and Voigt 2009; Voigt et al. 2010a, b, see also Voigt et al. 2014 and Brüggemann et al. 2014).

7.3.2 The Hasse-Diagram for the Comparison of Cars

We demonstrate the partial order application 19 types of BMW and by three [0,1]-normalized attributes, namely,

BV: fuel consumption
CA: CO_2 emission
Kp: purchase price

In Fig. 7.5 the Hasse diagram based on the 19×3 matrix of the below table is shown.

Data matrix BMW:

Objects

BC1, BC2, BL1, BL2, BK, Bca, BL3, BL4, BL5, BL6_1, BK2, BCa2, BC3, BL6_2, BL7, BL8, BS1, BS2, BS3

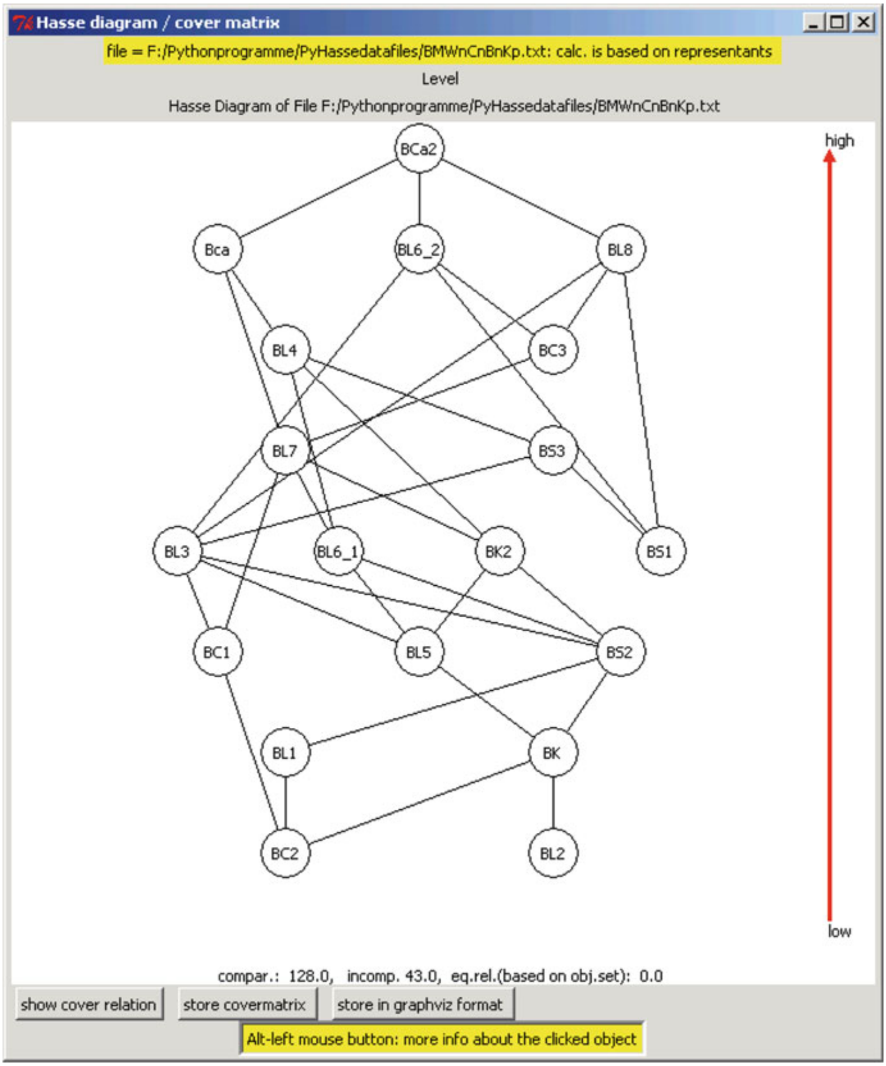
Properties

nCA, nBV, nKp

Data matrix

	nCA	nBV	nKp
BC1	0.434659	0.466667	0.007315
BC2	0.338068	0.266667	0.001045
BL1	0.414773	0.4	0.012539
BL2	0.335227	0.266667	0.020899
BK	0.355114	0.333333	0.068966
Bca	0.84375	0.866667	0.503135
BL3	0.502841	0.533333	0.15047
BL4	0.690341	0.666667	0.449321
BL5	0.366477	0.333333	0.130094
BL6_1	0.460227	0.4	0.301985
BK2	0.491477	0.4	0.289446
BCa2	1.0	1.0	1.0
BC3	0.519886	0.466667	0.514629
BL6_2	0.659091	0.666667	0.541275
BL7	0.505682	0.466667	0.478579
BL8	0.622159	0.6	0.819227
BS1	0.542614	0.533333	0.0
BS2	0.423295	0.4	0.121212
BS3	0.670455	0.666667	0.286311

We see that:



file = F:/Pythonprogramme/PyHassedatafiles/BMWnCnBnKp.txt: calc. is based on representants

Level

Hasse Diagram of File F:/Pythonprogramme/PyHassedatafiles/BMWnCnBnKp.txt

compar.: 128.0, incomp. 43.0, eq.rel.(based on obj.set): 0.0

Fig. 7.5 Hasse diagram using PyHasse software of 19 BMW car types characterized by three attributes (high: car affects the environment severely; low: car does not affect the environment severely)

- BCa2 is a greatest element. BCa2 has a highest purchase price, highest CO_2 emission, and highest fuel consumption. At least with respect to these three attributes, this type of car cannot be recommended.
- BC2, BL2, and BS1 are minimal elements. Any other car type has simultaneously disadvantages with respect to the three attributes.
- By the vertical arrangement of the vertices, we can identify car types which form antichains, for example: BL3, BL6_1, BK2, BS1. Clearly, there are not only incomparability relations among the members of one selected level, but

additionally also between elements belonging to different ones. See in this context Carlsen and Brüggemann (2014). Here, for example, BK2 is incomparable with all the other elements of the level to which BK2 belongs but also with BS3 and BC1.
- We also can find chains, where the three attributes are weakly simultaneously increasing. For example, BC2, BL1, BS2, BK2, BL4, BCa, BCa2 is such a chain. As it is tedious to identify chains by optical inspection, an own PyHasse module is focusing on chains and their properties.
- {BL4, BCa} and {BL6_2,BL8,BC3} are separated subsets, because there are no <–relations among the members of the two sets. Such separated subsets are of high interest, because each subset must have a property profile, which is characteristically different from that of the other one, so that no order relation can be found.

Here, however, our focus is on the interplay between weights of a composite indicator and the partial order, so we cannot deepen the partial order analysis concerning chains, antichains, and separability [see for details Brüggemann and Patil (2011)].

7.3.3 The Relation Between Rank and Weights

The decision problem so far has been discussed under two perspectives: Firstly, finding weights by more or less "intelligent" software support, by different approaches to visualize the consequences of a set of weights, and by facilitating to see the implications a certain set of weights would have. All these types of approaches try to bring as much transparency into the decision task as possible and to give some feeling about the sensitivity of the decision in correspondence to the set of weights chosen.

Our second approach applied partial order techniques and thereby generated a ranking giving by the Hasse diagram depicted in the section before. For this approach no weights are necessary so far.

In this situation, the idea may come up: how the rank gained by partial order conceptually following the approach of Winkler (1982) and the rank gained by the given set of weights correspond to each other. To work on this idea, a distance between the both rankings is calculated, first. In a further step this distance is minimized by variation of the set of weights. Figure 7.6 depicts the corresponding relations in overview.

7.3.4 Some Findings

Considerable attempts were performed to derive from a partially ordered set a linear or weak order. Most of these attempts are based on the results of Winkler (1982) where the heights of the elements of the partially ordered sets of each linear extensions are averaged in order to get averaged ranks. The high interest in this computational difficult problem is caused by the expectation that this ranking does not need any weighting of the indicators and could therefore be a mean to judge

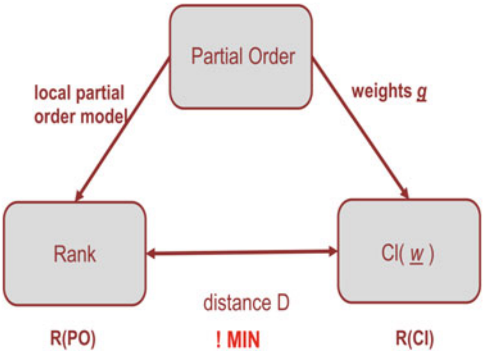


Fig. 7.6 Which weights would lead to a composite indicator whose ranks are in minimal distance to the ranks based on partial order theory

empirical rankings based on Composite Indicators or on any other procedure which provides a one-dimensional index suitable to get a ranking. For a recent development in calculating the average ranks, see Brüggemann and Carlsen (2011) and De Loof et al. (2006, 2011).

Here, we try an opposite approach: We do not judge a CI (composite indicator) by the ranking, derived from partial order theory, but we are asking us how does the vector of weights, g = (consumption/CO$_2$/prize), looks like, when the demand is to get CIs, whose ranks R(CI(g)) have a minimal distance to the partial order derived ranks R(PO). As a distance measure $D = D(R(\text{CI}), R(\text{PO}))$ we select the squared Euclidian distance and consider the ranks, identified by the associated elements as components of vectors. Let g^* be the set of weights which minimizes $D(R(\text{CI}), R(\text{PO}))$. Then we can compare g_{emp} (empirical weights) and arrive at an evaluation of g_{emp}. When we go along this sketched strategy we are faced with the following problems:

1. There may be several $g_1{}^* \neq g_2{}^*, \neq \ldots$ which minimize D.
2. The natural lowest value for D will be 0. However, often we will not arrive at $D = 0$ because the partial order may have order symmetries. Order symmetries will lead to averaged ranks for objects x, y, … which are equal. Because of the metric structure of the construction of the CI(g), the ranks of $R(\text{CI},x) \neq R(\text{CI},y)$. Then necessarily D is unequal to zero.
3. Instead of comparing g^* and g_{emp} we can also consider $D(R(\text{CI}), R(\text{PO}))$ as a function of the weights ("D-map"). In the case of three weights, and under the demand of normalized weights, the projection to g_i, g_j yields sufficient insights, where weights may lead to CI, near to R(PO) or not.

7.3.4.1 Weight Vectors, Minimizing *D*

The minimal value for D was found by Monte Carlo Simulation varying the weight vector $g = (g_{\text{fuel consumption}}, g_{\text{CO2}}, g_{\text{purchase prize}})$ with $D = 0.6316$. There are several weight vectors having minimal distance (minimal weight vectors).

The minimal weight vector that has the lowest value in fuel consumption is (0.004, 0.3, 0.696). Because of the normalization of weights, a small value in one component of the triple of weights needs high contributions of the remaining two. Therefore, the main contribution comes from the third weight. That is, when we accept that we give the purchase price a high weight, then CI(*g*) will be close to *R*(PO).

Maximal weight for fuel consumption is 0.226. Then the weight combination minimizing *D* is 0.226, 0.098, 0.676. We see a trade-off in fuel consumption and CO_2 emission; here, the weights can vary; obviously, the main factor is the purchase price, the weight of which is nearly the same as in the first case.

Checking the set of weights for that triple having the lowest weight for purchase price (g_3), we arrive at 0.675. There are several triples with g_3=0.675. How different could be g_1 and g_2? The lowest value of g_1 is 0.179, the highest one: 0.24 ($D=0.676$).

Hence, we find that any empirical weight, which does not give a high weight to the purchase price ($g_{3emp} \geq 0.675$) will automatically deviate from *i*(PO). Indeed, it is of high interest that also the sensitivity analysis rendered the purchase price as most important.

7.3.4.2 Order Symmetries

There is order symmetry, when two labels can be interchanged without changing the order relations.

We find two pairs of objects that are ordered theoretically symmetric:

(a) BL6_2, BL8
(b) BL6_1, BK2

Now, demanding a distance less than 0.8 leads to the scatter plot, shown in Fig. 7.7b.

We can verify this symmetry by constructing the local Hasse diagrams for these four elements, or more easily but approximately by checking the number of elements in the up- and down sets of these elements. Because of these two pairs of objects that are in an order symmetry, it is improbable to get a distance $D=0$. Indeed, we performed 5,000 Monte Carlo simulations and performed a systematic search within the cube of the three weights, requiring that sum = 1, and did not find a distance <0.6316.

7.3.4.3 D-maps

In Fig. 7.7 we see all points g_1, g_3 are shown, with distances $D \leq 3$ and $D \leq 0.8$, respectively.

We see from these two scatter plots:

- The high nonlinearity of the problem, minimizing the distance, because there is no unique solution.
- The set of acceptable points is topologically not connected.
- By an additional check (requiring distance $D \leq 0.7$) we see that then also the point sets with $g_3 < 0.6$ disappear, confirming the results reported above.

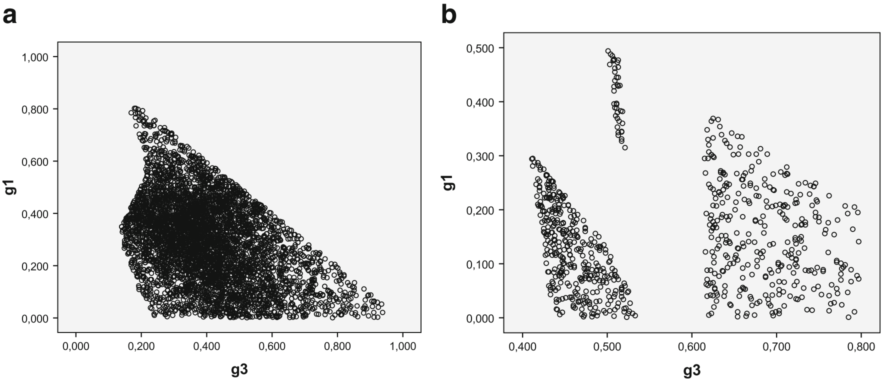


Fig. 7.7 Scatter plots (**a**) with $D \leq 3.0$, (**b**) with $D \leq 0.8$

So we can show that any weight system, with a lower weight for the purchase price than 0.6 will automatically lead to deviations in the rank based on the corresponding CI and the ranks based on partial order.

When, for example, a weight triple is selected with minimal D, then the partially ordered set based on CI and the averaged ranks has 4 incomparable pairs, whereas when weights are selected with a distance near 3, then 12 incomparable pairs appear, i.e., three times more conflicts between R(PO) and R(CI).

Another way to look upon the influence of weights is to apply the stability analysis, as provided by the PyHasse software package. If fuel consumption and purchase price are considered alone (the CO_2 emission is correlated with the fuel consumption), then we can see where weights g_{BV} and g_{Kp} (BV: from german Benzinverbrauch, Kp from Kaufpreis) will have influence on the order due to varying g_{BV}.

In Fig. 7.8 the horizontal axis is g_{BV}, and the bars indicate positions of g_{BV} where a ranking is changing.

The stability histogram refers to only two properties, here nBV and nKp, with

$$\mathrm{CI}(g) = g \times \mathrm{nBV} + (1 - g) \times \mathrm{nKp}$$

Any stability field in Fig. 7.8 (top) means that here the order due to a given g does not vary until g crosses the position of a bar. For more information, compare Brüggemann et al. (2008) and Restrepo et al. (2008a, b).

One striking question clearly is the role of a real empirical CI, made of **all three** attributes.

Assume, for example, the empirical CI is constructed as follows:

Empirical weights: $g_1 = 0.25$, $g_2 = 0.1$, $g_3 = 0.65$. This means that the weights are already selected in that way that R (CI(g_1,g_2,g_3)) is near to the R(PO) following the findings of the paragraphs above. Now: Which weight combination in the **two-dimensional** nBV,nKp-system would lead to a CI whose ranks have a minimum distance to R (CI(g_1,g_2,g_3))? The distance D(R(CI(g_1,g_2,g_3)), R(CI(g)) as a function of g is shown in Fig. 7.8 (bottom). By application of the tools of the PyHasse module

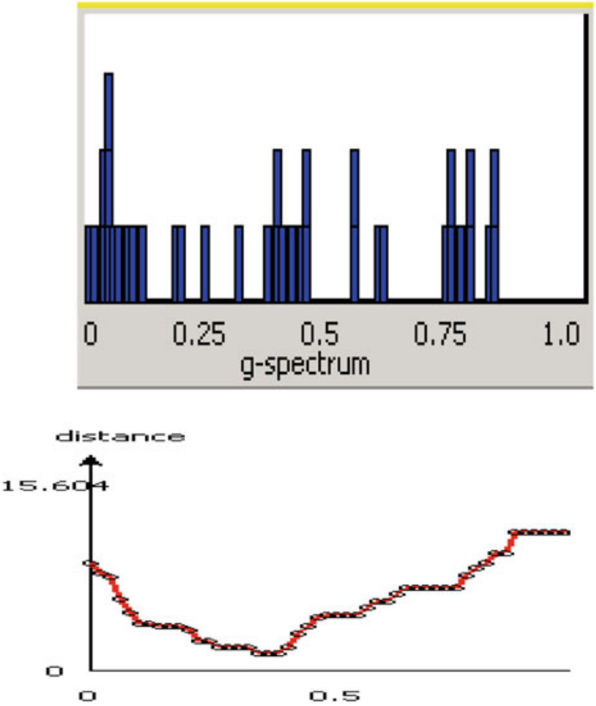


Fig. 7.8 *Top*: Stability histogram. (The *height of bars* indicates the number of interchanging ranks. The stability fields are the ranges between two bars and indicate robustness of the orders versus slight variations of *g*.) *Bottom*: distance of ranks due to a given empirical CI to ranks, due to the linear orders in the stability fields (In steps of 0.05 increments for *g*)

stability9.py, it turns out that $g = 0.35$ leads to the minimum distance. In Table 7.1 the orders due to the empirical CI and to $g = 0.365$ are shown.

Any other *g*-value will lead to more inversions and hence increases the distance.

Summarizing: The concept of stability fields may be a tool to reduce the dimensionality of the CI by searching for orders in the stability fields minimizing the distance to the orders due to the empirical CI. Unfortunately, however, the stability fields are pretty small. Hence, the determination of weights needs a careful analysis.

7.4 Conclusion and Outlook

This chapter proposes a decision support tool exemplarily for the selection of car types. It emphasizes the basic problem that the criteria for a decision often are in conflict to each other, and a decision support should not be restricted to a simple

Table 7.1 Order due to the empirical CI with weights: $g_1 = 0.25$, $g_2 = 0.1$, $g_3 = 0.65$ and to the CI in the nBV, nKp-system with $g = 0.35$

$i=0$: BC2	$i=0$: BC2
$i=1$: BL2	$i=1$: BL2
$i=2$: BL1	$i=2$: BL1
$i=3$: BK	$i=3$: BK
$i=4$: BC1	$i=4$: BC1
$i=5$: BL5	$i=5$: BL5
$i=6$: BS1	$i=6$: BS1
$i=7$: BS2	$i=7$: BS2
$i=8$: BL3	$i=8$: BL3
i=9: BK2	**i=9: BL6_1**
i=10: BL6_1	**i=10: BK2**
$i=11$: BS3	$i=11$: BS3
$i=12$: BL7	$i=12$: BL7
$i=13$: BC3	$i=13$: BC3
$i=14$: BL4	$i=14$: BL4
$i=15$: BL6_2	$i=15$: BL6_2
$i=16$: Bca	$i=16$: Bca
$i=17$: BL8	$i=17$: BL8
$i=18$: BCa2	$i=18$: BCa2

The only one rank inversion is marked with bold literals

ranking without any possibility for the user to get a sensibility for the effects changes in his priorities and/or weightings will have. Therefore this chapter offers four methods for comparison and decision making:

1. Classical ranking according to a weighted ratio: The criteria are weighted and merged. The result is a ranked list. The advantage of this method is its clear final result, its disadvantage the intransparency, how the resulting ranking is determined and especially how sensible the ranking would be in respect to changes in the weights for the different criteria.
2. Weighted ranking without explicit merging of the relevant criteria: This method tries to bring some more transparency concerning the influence of the decision criteria by showing the detailed valuation for each of them and thus giving more information about the final position of each car type in the ranking list. However, such a method is restricted to a quite small number of criteria. In the example, the implementation concentrated on the three most interesting ones.
3. "Showdown" with the complete information for preselected car types: To heal the deficit of the second method, the so-called showdown has been introduced: for a small number of preselected car types, the complete data set is shown and the "winner" for each criterion is highlighted. If method 2 shows a restricted number of criteria for all cars, then method 3 shows for a restricted number of car all their criteria.
4. Partially ordered set approach: A much more complex approach to explore relations between elements of a given set is proposed by the theory of partially ordered sets. It helps to understand exactly the crucial criteria, but it needs a set of methods that are not easy to apply for an inexperienced user.

5. Results of the partial order approach are not only the list of minimal and maximal objects but also the possibility to identify the next best object. The fact that there are several optimal cars shows that the potential user can still select according to his preferences within the set of attributes.
6. Relations between partial order ranking and thes composite index: An analysis of these relations gives new insights into the decision space by making the interdependencies between the weights more transparent and by making a visualization of the situation possible.
7. To find optimal weights is a highly nonlinear problem. If, for example, a set of weights is to be determined so that the composite indicator models appropriately the ranks due to the partial order techniques, one can see that there is no unique solution. Several sets of weights minimize the distance.
8. With two attributes the range of weights, where the induced order may change dramatically, can be visualized. Now one may search for weights referring to three attributes which map best into the two-attribute system. The applied technique may be the basis for a methodology to systematically reduce the dimensionality of the weighting problem.

For the example discussed in this chapter, the HDT-software (PyHasse) has not been included into the software package but has been used as a standby method, only. However, the experiences made with the very simple car selection example show that the user has to be supported to learn and to understand a resulting ranking and its sensibility by just “playing” with the criteria under observation and/or the weightings for each of them.

To understand and to emphasize a decision as a result of a decision-making process, this chapter proposes a set of quite simple methods and tries a first proposition to integrate them within a new software tool. In the first step of the project, the software design trusts in the users’ curiosity to navigate the decision set under various points of view supported by simple methods offered. With our last step to integrate the partial order methods, we reach the limits of this concept and in the further work a sophisticated user guidance has to be developed.

References

Brans JP, Vincke PH (1985) A preference ranking organisation method (The PROMETHEE Method for Multiple Criteria Decision-Making). Manag Sci 31:647–656

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partially orders. Match Commun Math Comput Chem 65:383–414

Brüggemann R, Carlsen L (2012) Multi-criteria decision analyses. Viewing MCDA in terms of both process and aggregation methods: some thoughts, motivated by the paper of Huang, Keisler and Linkov. Sci Total Environ 425:293–295

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems - introduction to partial order applications. Springer, New York

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Brüggemann R, Voigt K (2009) Analysis of partial orders in environmental systems applying the new software PyHasse. In: Wittmann J, Flechsig M (eds) Simulation in Umwelt- und Geowissenschaften- Workshop Potsdam 2009. Shaker-Verlag, Aachen, pp 43–55

Brüggemann R, Bücherl C, Pudenz S, Steinberg C (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim Hydrobiol 27:170–178

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (2001) Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. J Chem Inf Comp Sci 41:918–925

Brüggemann R, Voigt K, Restrepo G, Simon U (2008) The concept of stability fields and hot spots in ranking of environmental chemicals. Environ Model Softw 23:1000–1012

Brüggemann R, Carlsen L, Voigt K, Wieland R (2014) PyHasse software for partial order analysis: Scientific background and description of selected modules. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 19. Springer, New York

Carlsen L, Brüggemann R (2014) Indicator analyses. What is important – and for what? In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 18. Springer, New York

De Loof K, De Meyer H, De Baets B (2006) Exploiting the lattice of ideals representation of a poset. Fundam Inform 71:309–321

De Loof K, De Baets B, De Meyer H (2011) Approximation of average ranks in posets. Match Commun Math Comput Chem 66:219–229

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA

Manganaro A, Ballabio D, Consonni V, Mauri A, Pavan M, Todeschini R (2008) The DART (Decision Analysis by Ranking Techniques) software. In: Pavan M, Todeschini R (eds) Scientific data ranking methods: theory and applications. Elsevier, Amsterdam, pp 193–207

Munda G (2008) Social multi-criteria evaluation for a sustainable economy. Springer, Berlin

Myers WL, Patil GP (2008) Semi-subordination sequences in multi-measure prioritization problems. In: Todeschini R, Pavan M (eds) Data handling in science and technology, vol 27. Elsevier, New York, pp 161–170

Newlin J, Patil GP (2010) Application of partial order to stream channel assessment at bridge infrastructure for mitigation management. Environ Ecol Stat 17:437–454

Pudenz S (2005) ProRank - software for partial order ranking. Match Commun Math Comput Chem 54:611–622

Restrepo G, Weckert M, Brüggemann R, Gerstmann S, Frank H (2008a) Ranking of refrigerants. Environ Sci Technol 42:2925–2930

Restrepo G, Brüggemann R, Weckert M, Gerstmann S, Frank H (2008b) Ranking patterns, an application to refrigerants. Match Commun Math Comput Chem 59:555–584

Schwefel H-P (1995) Evolution and optimum seeking. Wiley, New York, 1995

Voigt K, Brüggemann R, Pudenz S (2006) A multi-criteria evaluation of environmental databases using the Hasse diagram technique (ProRank) software. Environ Model Softw 21:1587–1597

Voigt K, Brüggemann R, Scherb H, Shen H, Schramm K-H (2010a) Evaluating the relationship between chemical exposure and cryptorchidism by discrete mathematical method using PyHasse software. Environ Model Softw 25:1801–1812

Voigt K, Brüggemann R, Kirchner M, Schramm K-W (2010b) Influence of altitude concerning the contamination of humus soils in the German Alps: a data evaluation approach using PyHasse. Environ Sci Pollut Res 17:429–440

Voigt K, Brüggemann R, Scherb H, Cok I, Mazmanci B, Mazmanci MA, Turgut C, Schramm K-W (2014) PyHasse software features applied on the evaluation of chemicals in human breast milk samples in Turkey. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 17. Springer, New York

Winkler P (1982) Average height in a partially ordered set. Discrete Math 39:337–341

Part III
New Trends in Partial Order

Chapter 8
Coordination of Contrariety and Ambiguity in Comparative Compositional Contexts: Balance of Normalized Definitive Status in Multi-indicator Systems

Wayne L. Myers and Ganapati P. Patil

Abstract We address oppositional aspects of comparative compositional contexts for some particular purpose. Compositional components of land cover in localities provide our context, with the exemplifying purpose being cooperative conservation. A subset of cover components is considered definitely propitious (pro) for the purpose, with another subset being definitely contraindicative (con), and the rest as ambiguous "other." Plotting percent pro on the ordinate and percent con on the abscissa gives a "definitive domain display" for visualization. A "Balance Of Normalized Definitive Status" (BONDS) is used for scalar sequencing. Using concepts of "down-set" and "up-set" from theory of partially ordered sets (posets), this is extended to obtain an intrinsically compositional context of pro and con that applies objectively to any suite of (monotonic) indicators. Indicators are eliminated in a systematic manner to resolve ties in the extended version by lexicographic suborder. Computations are specified in terms of **R** software.

8.1 Introduction

We focus on multi-indicator systems for practical prioritization (Brüggemann and Patil 2010, 2011). We begin by using innovative approaches to spatial synthesis (Myers and Patil 2011, 2012a) for configuring a context of land cover composition

W.L. Myers (✉)
Penn State Institutes of Energy and Environment, The Pennsylvania State University,
5 Land & Water Research Building, University Park, PA 16802, USA
e-mail: wlm@psu.edu

G.P. Patil
Center for Statistical Ecology and Environmental Statistics, Department of Statistics,
The Pennsylvania State University, 421 Thomas Building, University Park, PA 16802, USA
e-mail: gpp@stat.psu.edu

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_8, © Springer Science+Business Media New York 2014

in a spatial setting as a basis for explication. In this context, we consider a prospective prioritization purpose of selection, designation, recognition, intervention, etc. Relative to the prospect, we partition the cover components (kinds of cover as available indicators) in a tripartite manner by aggregation as affirmatively promotive of the prospect (pro), contraindicative to the prospect (con), or indefinite (ambiguous) with regard to the prospect (other). We produce a definitive domain display for visualization of the comparative context (Myers and Patil 2010, 2012b) and use a "Balance Of Normalized Definitive Status" (BONDS) for scalar sequencing. The partitioning may be heuristic in such a compositional context, with visualization and scalar index serving to contrast competing views in terms of consequences for ordering and thus raises the level of debate.

We then invert the contrary indicators and draw upon partial order theory to obtain an intrinsically compositional context that applies objectively to any suite of (monotonic) indicators. The (noninclusive) "down-set" under product–order relation (Brüggemann and Voigt 2008; Brüggemann and Patil 2011) supplies the "pro" aspect, while the "up-set" supplies the "con" aspect with incomparable as ambiguous "other." A corresponding scalar index follows in a straightforward manner, but is subject to appreciable occurrence of ties. We extend the ranking strategy to obtain a lexicographic protocol for breaking ties among instances that are not identical in terms of indicators. This entails a progressive search for distinctiveness by successive deletion of indicators having lesser collectivity; which is followed, if necessary, by considering indicators individually. Such segregation within tied sets helps to reveal the particularities of influence among the indicators. The extended lexicographic indexing becomes a suitable candidate for motivating nonparametric rank-sum tests of hypothesis.

Computational facilities are shown in **R** software (Allerhand 2011; **R** Development Core Team 2008; Short 2009; Venables et al. 2005) which provides a convenient platform for progressive exploration of data. **R** commands are given in the text. When an **R** command generates a data frame, the header and one or two lines are given to show structure of the data frame. Output that elucidates details of the approach is presented in tables. **R** function facilities are presented as appendices.

8.2 Integrated Vicinity Indicators for Spatial Settings

A localization paradigm for integrative vicinity indicators (IVIs) is presented by Myers and Patil (2011, 2012a) for spatial settings. This begins with a locality layer as a grid of numbered point positions. Octagonal integrating vicinities (OCTIVs) are established around the locality points as a localizing layer, and indicators of interest are compiled within the vicinities and referenced to the central point positions. Octagonal vicinities are chosen as being parsimonious approximations to circular zones, requiring orders of magnitude fewer vertices than would be needed for circular buffer zones in a geographic information system (GIS). We apply this paradigm to obtain a context of land cover composition for use in our explorations of indicators.

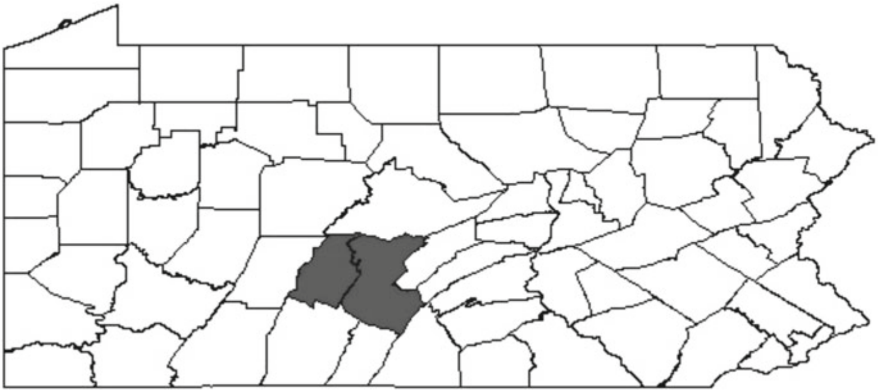


Fig. 8.1 Blair County (*left*) and Huntingdon County (*right*) in Pennsylvania, USA

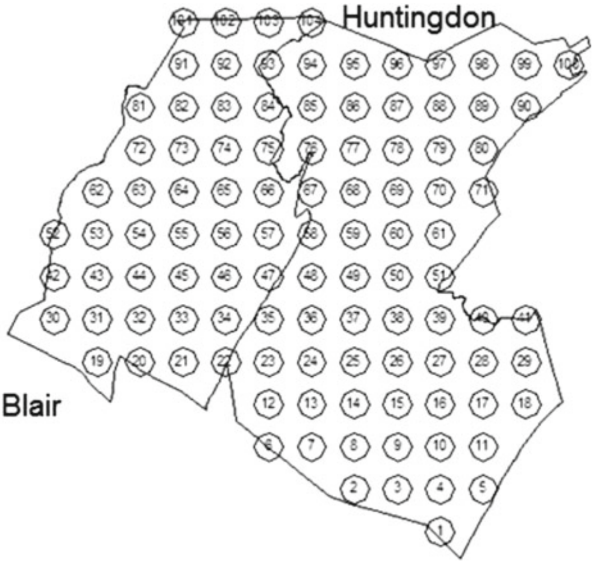


Fig. 8.2 Blair and Huntingdon Counties with perimeters of numbered octagonal vicinities (OCTIVs) having size determined by 2-km circumscribing circle. Centers of OCTIVs are spaced on a 6-km grid

We choose two counties (Blair and Huntingdon) having south-central location in the State of Pennsylvania, USA as shown in Fig. 8.1. These counties offer contrasts in terms of land cover, with Blair County hosting the urban area of Altoona, whereas Huntingdon is more rural and contains a large reservoir called Raystown Lake.

Figure 8.2 shows octagonal perimeters of the numbered vicinities (OCTIVs) with central locality points being suppressed. These OCTIVs have a circumscribing circle with 2-km radius.

Table 8.1 Component codes for 2001 (generalized) National Land Cover Database (NLCD)

Code	Cover type
11	Water
20	Development
31	Barren: rock, sand, clay
32	Quarries, strip mines, gravel pits
33	Transitional
40	Forest
81	Hay/pasture
82	Row crops
85	Urban/recreational grasses
91	Woody wetlands
92	Emergent herbaceous wetlands

Land cover composition is determined from a generalization of the 2001 National Land Cover Database (NLCD) produced by the Multi-Resolution Land Characteristics (MRLC) Consortium (Homer et al. 2004; Chander et al. 2009). Coding of components (kinds of land cover) for the 2001 generalized dataset is shown in Table 8.1. Each 30-m pixel is assigned one of these codes.

The composition (%) of each land cover component in the OCTIVs was compiled with a GIS as an integrated vicinity indicator (IVI). Since composition is in terms of percentage, there are two options for OCTIVs that overlap county boundaries. One is to clip the OCTIVs, and the other is to provide land cover data extending beyond county boundaries. The latter has been used for present purposes. These IVIs are transferred to **R** software as a “data frame,” with header and first six lines (head) as shown for subsequent reference in Table 8.2.

8.3 Prospect of Cooperative Conservation

We posit an assessment of prospect for cooperative conservation and consideration of candidacy for solicitation. For this prospect, we initially consider water (Pct11) and forest (Pct40) IVIs as being promotive. Development IVI (Pct20) along with quarries, strip mines, and gravel pits IVI (Pct32) are considered as contraindicative. These are referenced to their column positions in the data frame (Table 8.2) for work in **R** as follows:

```
> IDs
[1] 1
> ThePros
[1] 2 6
> TheCons
[1] 3 4
```

Since the identifiers for the OCTIVs (Fig. 8.2) occupy the first column, the first IVI indicator is in the second column position. The remaining six IVI indicators are initially seen as “other”; that is, neither definitively “pro” nor definitively “con.”

Table 8.2 Head of **R** data frame for available indicators as columns of percent composition for different kinds (components) of land cover

```
> head(BHcvr)
  ID  Pct11    Pct20    Pct32    Pct33    Pct40    Pct81     Pct82    Pct85 Pct91    Pct92
1 1   0.071644 0.660722 0.000000 0.000000 69.59083 24.120362 3.319535 0     2.236904 0.000000
2 2   0.111500 0.055750 0.000000 0.000000 69.69576 24.163746 5.853775 0     0.000000 0.119464
3 3   2.348352 0.055723 0.000000 0.007960 56.11368 31.985352 9.409329 0     0.000000 0.079605
4 4   0.039802 0.000000 0.000000 1.432892 87.74080 7.944594  2.818022 0     0.007960 0.015921
5 5   0.103486 0.127368 0.000000 0.000000 70.45852 22.138194 6.249004 0     0.772170 0.151249
6 6   0.071667 0.844083 6.051919 0.899824 89.05877 2.986144  0.000000 0     0.087593 0.000000
```

The ID column gives OCTIV number as shown in Fig. 8.2. In column headers Pct denotes percent and number is code (Table 8.1)

This constitutes a tripartite partition recognizing primary oppositional aspects of the prospect. The **R** function named Procon in Appendix 1 is used to extract the "pro" and "con" composites into a secondary data frame named ProAndCon as follows:

```
> ProAndCon <- Procon(BHcvr,IDs,ThePros,TheCons)
> head(ProAndCon)
Ids Pro Con
[1,] 1 69.66247 0.660722
```

8.4 Definitive Domain Display

The secondary data frame (ProAndCon) provides the basis for configuring a definitive domain display in which the definitive Pro and Con aspects are on the axes. Such a plot will be considered more informally here as a "pro-pensity plot" or simply a "propensity plot." The **R** function ProconPlot given in Appendix 2 produces the graphic in Fig. 8.3 when invoked as follows:

```
> ProconPlot(ProAndCon)
> identify(ProAndCon[,3],ProAndCon[,2])
[1] 21 54 63
```

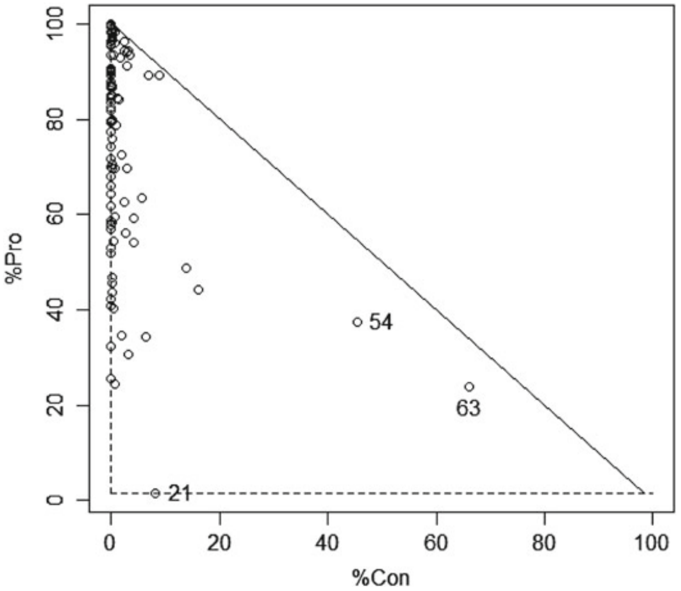


Fig. 8.3 Definitive domain display (or propensity plot) for water and forest as "Pros" with development and quarries/strip mines/gravel pits as "Cons." Selected points are labeled by row position (which is also OCTIV ID) in the data frame

The greater the Pro and the less the Con, the more propitious is the positioning of an instance. The diagonal line in Fig. 8.3 is a "limiting line" representing 100 % composition. Thus, the farther a point is situated (horizontally) to the left of this line, the greater the contribution of "other" and the less definitive the indications. Ties would be indicated by + through the circle if any were present.

Three points are selectively labeled in Fig. 8.3 by row position (also ID for OCTIV) in the data frame. The full compositional structure of these instances is given in Table 8.3. Instance (OCTIV) 21 is seen not to be very definitive, being composed 90 % of codes 81 and 82 (hay/pasture and row crops). In other words, this locality is strongly agricultural, which was not declared as either pro or con in the original prospect of cooperative conservation. This is signified by being low on both axes and thus far to the left of the limiting line. In contrast, locality 63 quite clearly has little to recommend it, being two-thirds developed. Locality 54 is somewhat less than half developed with the remainder being more rural as forest and farmland. This ability to probe the details interactively shows a major strength of **R** for those acquainted with its protocols.

The more pronounced candidacies are situated in the upper (affirmative to the prospect) apex, which is densely populated in this context. Thus, it is more a question here of which localities not to consider further.

The promotive apex can be explored further as shown in Fig. 8.4 by a single **R** plotting command which specifies the ranges of values on the axes along with a particular plotting character and appropriate labels for the axes. Note that specification of what is to appear on the abscissa comes immediately after the opening parenthesis as follows:

```
> plot(ProAndCon[,3],ProAndCon[,2],xlab="%Con",ylab="%Pro",
+ xlim=c(0,20),ylim=c(80,100),pch="x")
```

The expansion in Fig. 8.4 suggests that further pursuit of priorities would entail listing all instances with a "Pro" value of perhaps 95 %. This would be easily done in **R** by taking advantage of the protocols for incorporating conditionals in subscripting for retrieval with the following command.

```
> ProAndCon[ProAndCon[,2]>95,]
```

8.5 Balance of Normalized Definitive Status

We proceed to undertake capturing aspects of information in a definitive domain display by a scalar sequence that reflects both Pro and Con as well as being responsive to degree of definitiveness. Such scaling helps to address questions that would entail comparing definitive domain displays, which is not readily done directly.

Capability for comparison is needed for visualizing the implications of altering the oppositional aspects of indication, such as whether something should be definitive and in what manner. An obvious question then concerns which of the instances (OCTIVs) are comparatively affected and how. Plotting scalars for alternatives against each other

Table 8.3 Full composition for OCTIVs 21, 54 and 63

ID	Pct11	Pct20	Pct32	Pct33	Pct40	Pct81	Pct82	Pct85	Pct91	Pct92
21	0.262822	8.211213	0	0.000000	1.330041	51.951258	38.093341	0.000000	0	0.151322
54	0.549538	45.388658	0	1.385791	37.026122	12.942019	0.302644	2.405224	0	0.000000
63	0.000000	66.063889	0	0.023898	23.898669	9.256751	0.756791	0.000000	0	0.000000

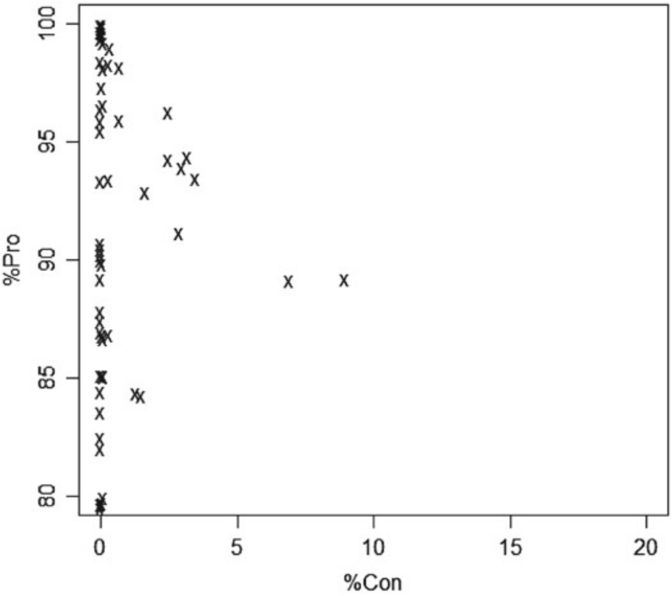


Fig. 8.4 Probing the affirmative apex of Fig. 8.3 with an **R** plotting command

facilitates comparative analysis. To this end, we calculate a Balance Of Normalized Definitive Status (BONDS) for each instance, and then rank across instances:

BONDS = (Pro – Con) × (Pro + Con)/100.0.

In this formulation, the (Pro + Con)/100.0 factor has the effect of lending emphasis to differences according to the degree of definitiveness. Positive values reflect more Pro than Con, whereas negative values reflect more Con than Pro. More definitive differences carry more weight. The **R** BONDings function in Appendix 3 serves to calculate and rank the BONDS values according to pro and con tuples as follows:

```
> ProAndConBONDS <- BONDings(ProAndCon)
> head(ProAndConBONDS)
      Ids      Pro      Con BONDSrank
[1,] 1 69.66247 0.660722 37
```

There is a lexicographic component to the BONDSranks. The factor (Pro + Con)/100.0 has no effect for 0.0 values of BONDS. However, such neutral values of BONDS are sub-ranked according to increasing definitiveness. Thus, definitiveness is incorporated in all BONDSranks.

We can illustrate BONDS-based exploration of sensitivity in our land-cover context by suggesting that the wetland components (Pct91 and Pct92 in Table 8.2) should be included in the promotive partition. We obtain Pro and Con for inclusion of wetlands as follows, with plotting as shown in Fig. 8.5:

```
> ThePros2 <- c(2,6,10,11)
> ProVsConBONDS <- BONDings(ProVsCon)
> plot(ProAndConBONDS[,4],ProVsConBONDS[,4],
+ xlab="BONDSrank1",ylab="BONDSrank2")
```

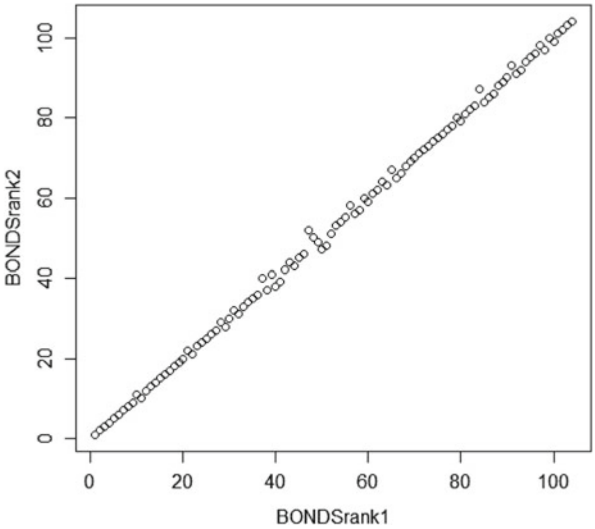


Fig. 8.5 Cross-plot treating wetlands as promotive (BONDSrank2) versus wetlands as not being definitive (BONDSrank1), showing relative insensitivity to the shift

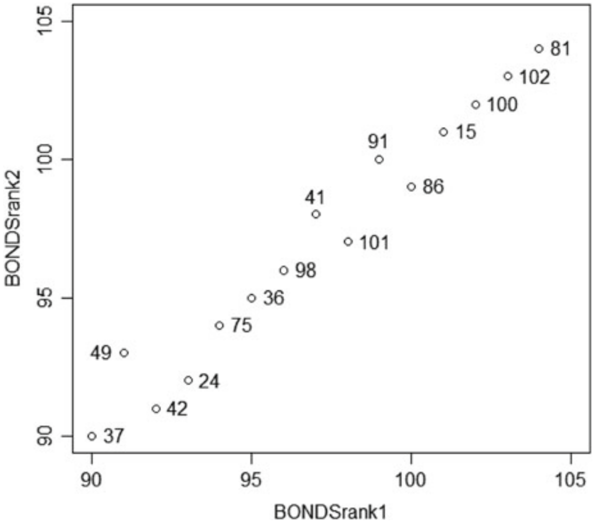


Fig. 8.6 Expansion of *upper-right corner* of Fig. 8.5 and labeling of OCTIVs that are propitious from both initial perspectives

The cross-plot in Fig. 8.5 shows relative insensitivity to the shift and any OCTIV instances that show some discrepancy could be identified by labeling as shown in previous plots.

We expand the upper-right corner of the cross-plot and label some of the propitious prospects in Fig. 8.6.

8.6 Comparative Compositional Context Using Partial Order Theory

We next draw upon partial order theory (Patil and Taillie 2004; Brüggemann and Patil 2011) to provide inputs for visualization and coordination that are less heuristic. This gives an intrinsically compositional context involving contrariety and ambiguity while being objectively applicable to any suite of (monotonic) indicators. Opposition aspects do not enter directly at the level of indicators, since we invert/reverse/negate those that are contraindicative to lend all indicators the same sense. Oppositional aspects emerge secondarily as a consequence of comparative compilation.

For explication of the approach, we choose a suite of indicators from among the compositional components of land cover (Tables 8.1 and 8.2). It bears emphasis, however, that the indicators need not entail composition directly. The **R** function named Pickind given in Appendix 4 provides for selecting and orienting indicators from a data frame such as BHcvr (Table 8.2) above. In addition to the name of the data frame containing the raw indicators, it requires a specification of Id column number and a vector of column numbers selected as indicators. If the column number is made negative, the data in the column will be negated to change the sense of direction. This should be done with all indicators that are seen as being inherently contraindicative. We choose water (Pct11) and forest (Pct40) as affirmative indicators, with three contraindicative indicators being development (Pct20), quarries/strip mines/gravel pits (Pct 32), and row crops (Pct 82). Row crops were not used explicitly earlier, but Fig. 8.3 and Table 8.3 suggested that row crops can constitute a major land cover and they entail intensive human disturbance on a continuing basis. This suite of indicators is configured from BHcvr as follows:

```
> IDs
[1] 1
> Apick <- c(2,-3,-4,6,-8)
> Apicking <- Pickind(BHcvr,IDs,Apick)
> head(Apicking)
  IDs    Pct11     Pct20     Pct32    Pct40     Pct82
1   1 0.071644 -0.660722  0.000000 69.59083 -3.319535
```

We proceed to use product–order relation (Brüggemann and Patil 2011) to determine the percentage of other instances which each instance dominates, along with the percentage by which it is dominated. One instance dominates another in the product–order sense if it is at least as high on all indicators and higher on at least one indicator. Conversely, an instance is dominated by another if it is at least as low on all indicators and lower on at least one indicator. Paired comparisons that do not exhibit domination are considered as incomparable and not definitive. The percentage that an instance dominates is a “down-set” in the parlance of partial order theory (De Loof et al. 2008), whereas the percentage by which an instance is dominated constitutes an “up-set.” The percentage as “down-set” supplies the propitious aspect in the sense described earlier PRO=Pro(down)=|$O(x)$|, whereas the percentage as “up-set” supplies the contraindicative aspect CON=Con(up)=|$F(x)$|. The remaining

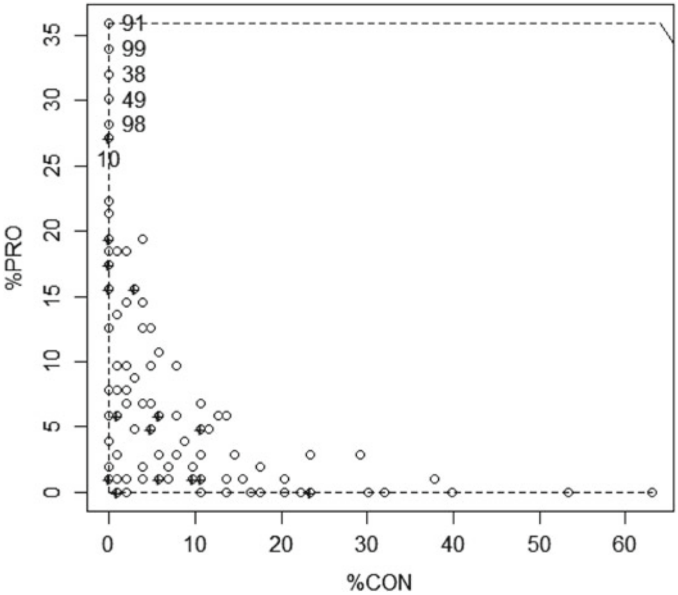


Fig. 8.7 Definitive domain display for product–order tabulation of five indicators, with occurrence of ties indicated by a *cross* in the *circle*. Note intentional truncation of limiting line (*upper-right corner*) to give more detail for plotted points

percentage (if any) as neither “PRO” or “CON” is comprised of “incomparable” instances that are not definitive. Note that we are using “PRO” and “CON” for the poset-based version to distinguish it from “Pro” and “Con” of the earlier version. Thus, we obtain an analytically derived compositional context with definitive oppositional aspects and ambiguity in the incomparable “Other.”

The product–order compilation of PRO and CON is performed the **R** POprocon function in Appendix 5 and then plotted as a definitive domain display for visualization in Fig. 8.7 as follows:

```
> PickProCon <- POprocon(Apicking)
> PikProCon <- BONDings(PickProCon)
> ProconPlot(PikProCon,2)
> identify(PikProCon[,3],PikProCon[,2])
```

A noteworthy aspect of Fig. 8.7 is the relatively low maximum on the PRO axis which is further reflected in only a small section of the limiting line being included at the right of the plot. As shown earlier in Fig. 8.4, several of the OCTIVs have a composition of more than 95 % forest and water, with most of that being forest. This similarity will tend to induce a high degree of conflict and that is accentuated here by the designation of row crops as a contrary indicator among instances that are not so heavily forested. Conflict and ties can thus encompass considerable information on interaction of indicators among some subsets of instances, and each set of ties may entail some particular aspect of such interaction. It can thus be interesting and instructive to investigate breaking of ties in an extended lexicographic manner by systematically suppressing some indicators for a given subset of instances that is tied in terms of BONDSranks in order to disambiguate incomparability. Toward this end, we note that ties enter as averaged (BONDS)ranks and that compilation of dominance can be conducted for individual (OCTIV) instances (Brüggemann et al. 2004, 2005).

8.7 Collectivity of Indicators

BONDSrank is a collective expression of the indicators, and this is a consideration for disambiguation by systematic suppression of indicators. A strategic goal in finding a subset of indicators for breaking ties is to favor indicators that contribute to a collective consensus. A random vector would be completely lacking in regard to collective consensus since it would only introduce non-indicative "static." Likewise, indicators that speak to uncorrelated aspects of variability are essentially lacking in regard to collective consensus. We use correlation of an indicator with BONDSrank as a working criterion of collective consensus (collectivity). Accordingly, we compute correlations of indicators with BONDSrank and arrange the indicators in order of decreasing magnitude of correlation as follows:

```
> PickProCor <- cor(cbind(Apicking[,-1],PicsProCon[,5]))
> PickProCor <- PickProCor[,6]
> PickProCor <- PickProCor[1:5]
> PickProCor
Pct11   Pct20   Pct32   Pct40   Pct82
0.1757874 0.2314213 0.1284951 0.6088157 0.5143591
> Pickor <- c(4,5,2,1,3)
> Pickor
[1] 4 5 2 1 3
```

The first of the foregoing commands gives the correlation matrix for the indicators after removing the ID column and appending the BONDSrank as an additional column. The second command extracts the last column of the correlation matrix pertaining to the BONDSrank, and the third command omits the unit correlation for BONDSrank with itself. The vector of correlations is then listed, from which an order vector of indicator numbers is prepared according to priority for retention with lower magnitude of correlation (less collectivity) given less priority. Forest and row crops exhibit the greatest collectivity followed by development. The indicator with lowest collectivity is that for very localized quarries, strip mines, and gravel pits.

8.8 Tie Tracking

The prevalence of ties can be seen more fully by plotting BONDSranks against a simple step structure (sss). This is shown in Fig. 8.8, with a tie set occurring wherever there is a horizontal subsequence of points. Labeling of points can be used to identify

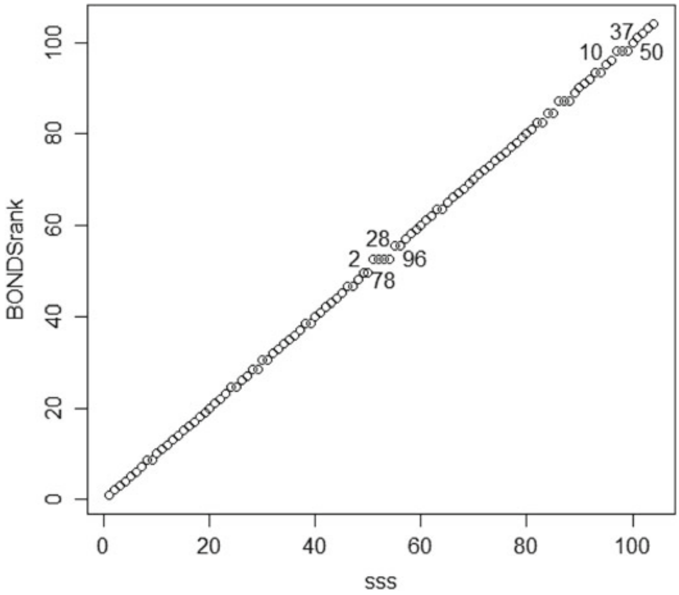


Fig. 8.8 Tie tracking with simple step structure (sss) as a supplement to Fig. 8.7. A tie set occurs wherever there is a horizontal subsequence of points, examples being (10, 37, 50) and (2, 28, 78, 96)

ties; so the highest tie set in Fig. 8.8 consists of rows (OCTIV instances) 10, 37, and 50. The tie set with the largest membership includes instances 2, 28, 78, and 96.

```
> sss <- rank(PikProCon[,4],ties.method="first")
> plot(sss,PikProCon[,4],ylab="BONDSrank")
> identify(sss,PikProCon[,4])
```

A compendium of ties can be obtained by applying the **R** TieSpecs function (Appendix 6) to augmented output of the BONDings function as follows to produce Table 8.4.

```
> PicProCon <- BONDings(PickProCon,2)
> PicsProCon <- TieSpecs(PicProCon)
> PicsProCon[PicsProCon[,6]>0,]
```

Since BONDSrank has a lexicographic element for BONDS values of 0.0, TieSpecs tabulates sets as identical PRO, CON pairs which are also assured to have identical BONDSrank. Thus, TieSet 1 and TieSet 14 have different BONDSrank but both have 0.0 as a BONDS value. The bracketed numbers on the left of Table 8.4 are line numbers in the table. The CaseIDs are ID numbers of the OCTIVs and also line numbers in the PicsProCon data frame from which the lines with ties are extracted. The TieSets column in Table 8.4 shows the set number of the tie set for each instance involved in a tie. The TieLink column effectively contains a linked list for membership in each of the tied sets. It contains the ID of another member in the same set or (for the last member of the set) the negative of the rank above which the ties in the set start. Table 8.4 shows 15 sets of ties, with the largest being one of two sets

Table 8.4 Compendium of ties among instances with respect to BONDSrank for five indicators

	CaseIDs	PRO	CON	BONDS	BONDSrank	TieSets	TieLink
[1,]	2	5.83	5.83	0.000000	52.5	1	28
[2,]	6	0.00	0.97	-0.009409	46.5	2	12
[3,]	7	0.97	9.71	-0.933432	28.5	3	74
[4,]	10	27.18	0.00	7.387524	98.0	4	37
[5,]	12	0.00	0.97	-0.009409	46.5	2	-45
[6,]	14	0.97	10.68	-1.131215	24.5	5	76
[7,]	15	19.42	0.00	3.771364	93.5	6	102
[8,]	16	0.97	5.83	-0.330480	38.5	7	57
[9,]	18	4.85	10.68	-0.905399	30.5	8	26
[10,]	21	0.00	23.30	-5.428900	8.5	9	85
[11,]	22	15.53	2.91	2.327128	82.5	10	56
[12,]	23	15.53	0.00	2.411809	84.5	11	79
[13,]	26	4.85	10.68	-0.905399	30.5	8	-29
[14,]	28	5.83	5.83	0.000000	52.5	1	78
[15,]	33	17.48	0.00	3.055504	87.0	12	41
[16,]	37	27.18	0.00	7.387524	98.0	4	50
[17,]	39	5.83	0.97	0.330480	63.5	13	40
[18,]	40	5.83	0.97	0.330480	63.5	13	-62
[19,]	41	17.48	0.00	3.055504	87.0	12	58
[20,]	43	4.85	4.85	0.000000	49.5	14	67
[21,]	50	27.18	0.00	7.387524	98.0	4	-96
[22,]	51	0.97	0.00	0.009409	55.5	15	62
[23,]	56	15.53	2.91	2.327128	82.5	10	-81
[24,]	57	0.97	5.83	-0.330480	38.5	7	-37
[25,]	58	17.48	0.00	3.055504	87.0	12	-85
[26,]	62	0.97	0.00	0.009409	55.5	15	-54
[27,]	67	4.85	4.85	0.000000	49.5	14	-48
[28,]	74	0.97	9.71	-0.933432	28.5	3	-27
[29,]	76	0.97	10.68	-1.131215	24.5	5	-23
[30,]	78	5.83	5.83	0.000000	52.5	1	96
[31,]	79	15.53	0.00	2.411809	84.5	11	-83
[32,]	85	0.00	23.30	-5.428900	8.5	9	-7
[33,]	96	5.83	5.83	0.000000	52.5	1	-50
[34,]	102	19.42	0.00	3.771364	93.5	6	-92

comprised of instances having equal Pro and Con. The last entry in TieLink for set number 1 is −50, meaning that the lowest rank in the set is 51. This set is the one labeled in the middle of Fig. 8.8 consisting of OCTIVs 2, 28, 78, and 96.

8.9 Trial Indicator Eliminations Based on Collectivity

Each set of tied instances (OCTIVs) is addressed in a three-phase process involving a systematic progression of Trial Indicator Eliminations (TIE) leading to the Tie Resolving Indicator Modification (TRIM) that is least disruptive of collectivity

Table 8.5 Output of TIEphasC function for TieSet 1 (Untie=1)

	TieIds	TRIMrank	Instep
[1,]	2	51	-2
[2,]	28	54	-2
[3,]	78	52	-2
[4,]	96	53	-2

unless instances are identical with respect to all indicators. The comparators generated in the first two phases are TIEBONDS, and the third phase uses the TIEBONDS from the first two phases to produce TRIMranks as lexicographic replacements for the tied ranks.

The A-phase is that of progressively dropping indicators according to least collectivity-based priority for retention. This phase is performed by the **R** TIEphasA function in Appendix 7. To facilitate checking, the first pass of this function uses all indicators and is labeled as Step 0. The next step drops the indicator having lowest collectivity-based priority for retention and is labeled as Step -1. For subsequent steps, the negative step number corresponds to the number of indicators dropped.

The B-phase consists of using the indicators one-by-one starting with the one having the highest collectivity-based priority for retention by virtue of greatest magnitude of correlation with the all-indicator BONDSranks. This phase is performed by the **R** TIEphasB function in Appendix 8.

The C-phase concatenates the outputs of the A-phase and B-phase and then finds lexicographic ranks for the least disruptive Tie Resolving Indicator Modification (TRIM). This phase is conducted by the **R** TIEphasC function in Appendix 9.

Here we choose to break apart the first (and largest) tied set, which is the set of four instances numbered as TieSet 1 in Table 8.4 comprised of OCTIVs 2, 28, 78, and 96. The **R** commands are as follows with output shown in Table 8.5. The third column of Table 8.5 shows that these ties were broken by dropping the quarries/strip mines/gravel pits Pct32 indicator and the Pct11 water indicator.

```
> Untie <- 1
> UntieSet1A <- TIEphasA(Apicking,PicsProCon,Pickor,Untie)
> UntieSet1B <- TIEphasB(Apicking,PicsProCon,Pickor,Untie)
> UntieSet1C <- TIEphasC(PicsProCon,UntieSet1A,UntieSet1B,
Untie)
```

This approach allows lexicographic latitude to resolve ties for different sets with different suppressions of indicators. If desired, it would be simple to replace the tied ranks of a set with TRIM ranks using a negative sign to flag the substitutions. Of course, one could also bundle the tie-resolving operations in a programmed manner that would address resolving of all ties in a single run. Ties are intrinsic for instances having identical values of all indicators.

8.10 Synopsis

We return now to the spatial setting and show in Fig. 8.9 the localities that are labeled at the top of Figs. 8.7 and 8.8 as being propitiously positioned. It should not be surprising that spatial affinities appear in Fig. 8.9. Seven of these eight vicinities are located in Huntingdon County, and half are adjacent in central Huntingdon County with two together in northeastern Huntingdon County.

There is evident merit in considering the spatial aspects beyond treating the localities comparatively as individuals (Myers and Patil 2012a), and localization with vicinity variates helps facilitate such further consideration. One such strategy is dual clustering as CLANs (clustered localities agglomerated nonspatially) and CLUMPs (clustered localities using map positions).

We have developed and demonstrated the dually definitive and implicitly ambiguous oppositional display of pros and cons for a prospect in compositional contexts, along with BONDSranks (Balance Of Normalized Definitive Status) to compare aspects of alternatives.

Generalization of the definitive domain display and BONDSranks to any multi-indicator system was obtained through partial order theory in terms of product–order relation and frequencies of domination as "down-sets" and "up-sets." The overall (all indicators) BONDSranks are subject to appreciable ties but provide a basis for quantifying collectivity among the indicators. An explicitly numerical expression of collectivity serves to sequence subsets of indicators for consideration

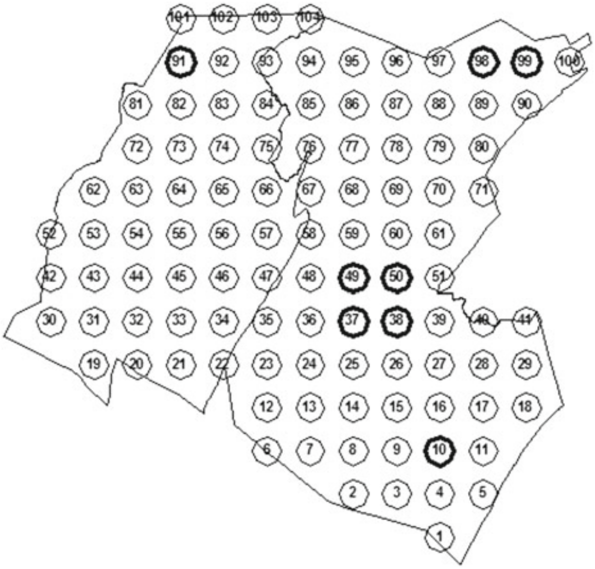


Fig. 8.9 Location of vicinities having propitious positioning in Figs. 8.7 and 8.8

in lexicographic breaking of ties with TRIMranks (Tie Resolving Indicator Modification) while adding insight regarding roles of individual indicators.

In addition to serving purposes of prioritization, BONDSranks and TRIMranks can provide a basis for constructing rank-sum tests of hypothesis regarding subsets of instances. In the current context, this might address the question of differences between the two counties.

Appendix 1: R Procon Function to Partition Compositional Components

```
Procon <- function(PctTabl,Id,Pros,Cons)
# This function is for composition indicators as %.
# Id is column number of case ID.
# Pros is vector of column numbers as Pros.
# Cons is vector of column numbers as Cons.
{Ids <- PctTabl[,Id]
 Npro <- length(Pros)
 Ncon <- length(Cons)
 Itms <-length(Ids)
 Pro <- rep(0.0,Itms)
 Con <- rep(0.0,Itms)
 for(I in 1:Itms)
  {for(J in 1:Npro)
    {K <- Pros[J]
     Pro[I] <- Pro[I] + PctTabl[I,K]
    }
   for(J in 1:Ncon)
    {K <- Cons[J]
     Con[I] <- Con[I] + PctTabl[I,K]
    }
  }
 ProCon <- cbind(Ids,Pro,Con)
 ProCon
}
```

Appendix 2: R Proconplot Function for Definitive Domain Display

```
ProconPlot <- function(PpCc,Capital=1)
  # Input is pro and con data frame.
  # Idz is column number of CaseIDs.
  # Pp is column number of Pro part.
  # Cc is column number of Con component.
  {Cases <- length(PpCc[,1])
   Idz <- 1
   Pp <- 2
   Cc <- 3
   Ymax <- max(PpCc[,Pp])
   Ymin <- min(PpCc[,Pp])
   Xmax <- max(PpCc[,Cc])
   Xmin <- min(PpCc[,Cc])
   Xright <- Ymax
   if(Xmax>Ymax) Xright <- Xmax
   if(Capital==1) plot(PpCc[,Cc],PpCc[,Pp],ylab="%Pro",
   xlab="%Con",xlim=c(0,Xright))
   if(Capital>1) plot(PpCc[,Cc],PpCc[,Pp],ylab="%PRO",
   xlab="%CON",xlim=c(0,Xright))
   YY <- c(Ymax,Ymin)
   XX <- c(100-Ymax,100-Ymin)
   lines(XX,YY,lty=1)
   XX <- c(Xmin,Xmin)
   YY <- c(Ymin,Ymax)
   lines(XX,YY,lty=2)
   XX <- c(Xmin,Ymax)
   if(Xmax>Ymax) XX <- c(Xmin,Xmax)
   YY <- c(Ymin,Ymin)
   lines(XX,YY,lty=2)
   XX <- c(Xmin,100-Ymax)
   YY <- c(Ymax,Ymax)
   lines(XX,YY,lty=2)
   for(I in 1:Cases)
    {IPp <- PpCc[I,2]
     ICc <- PpCc[I,3]
     Frq <- 0
     for(J in 1:Cases)
      {JPp <- PpCc[J,2]
       JCc <- PpCc[J,3]
       if(IPp==JPp & ICc==JCc) Frq <- Frq + 1
      }
     if(Frq>1) points(ICc,IPp,pch="+")
    }
  }
```

Appendix 3: R BONDings Function for Calculating BONDS Values and BONDSranks

```
BONDings <- function(ProAnCon,Items=1)
  # This function takes output of Procon or POprocon.
  # Appends Balance Of Normalized Definite Status (BONDS)
and BONDSrank.
  # Items=2 gives both BONDS values and BONDSranks.
  {ID <- ProAnCon[,1]
   Pro <- ProAnCon[,2]
   Con <- ProAnCon[,3]
   Ncase <- length(ID)
   BONDS <- rep(0,Ncase)
   for(I in 1:Ncase)
   BONDS[I] <- (Pro[I] - Con[I]) * (Pro[I] + Con[I])/100.0
   BONDSrank <- rank(BONDS,ties.method="average")
   ProconBOND <- cbind(ProAnCon,BONDS,BONDSrank)
   ProconBOND <- ProconBOND[order(ProconBOND[,4]),]
   Lo <- -1
   Hi <- -1
   for(I in 1:Ncase)
    {if(ProconBOND[I,2]==ProconBOND[I,3] & Lo<0) Lo <- I
     if(ProconBOND[I,2]==ProconBOND[I,3]) Hi <- I
    }
   Zros <- 0
   if(Lo>0) Zros <- Hi - Lo + 1
   if(Zros>0)
    {Zrows <- rep(0,Zros)
     for(I in 1:Zros)
      {J <- Lo + I - 1
       Zrows[I] <- ProconBOND[J,2]
      }
     Zrows <- rank(Zrows,ties.method="average")
     for(I in 1:Zros)
      {J <- Lo + I - 1
       ProconBOND[J,5] <- Zrows[I] + Lo - 1
      }
    }
   ProconBOND <- ProconBOND[order(ProconBOND[,1]),]
   if(Items<2) ProconBOND <- ProconBOND[,-4]
   ProconBOND
  }
```

Appendix 4: R Pickind Function for Selecting and Orienting Indicators

```
Pickind <- function(Aframe,IDitem,Pickings)
  # This function takes data frame of general indicators.
  # IDitem is column number of case ID.
  # Pickings is vector of column numbers of indicators.
  # Negative column number negates column as cons.
  {Items <- length(Aframe)
   IDs <- Aframe[,IDitem]
   INcas <- length(IDs)
   Picks <- length(Pickings)
   Kindups <- Aframe
   Xitems <- 0
   for(I in 1:Picks)
    {J <- Pickings[I]
     if(J<0)
      {J <- -1 * J
       Kindups[,J] <- -1 * Kindups[,J]
      }
    }
   for(I in 1:Items)
    {K <- Items - I + 1
     Xitem <- 1
     for(J in 1:Picks)
      {KK <- Pickings[J]
       if(KK<0) KK <- -1 * KK
       if(KK == K) Xitem <- 0
      }
     Xitems <- Xitems + Xitem
     if(Xitem>0) Kindups <- Kindups[,-K]
    }
   Kindups <- cbind(IDs,Kindups)
   Kindups
  }
```

Appendix 5: R POprocon Function for Product-Order Comparative Compilation

```
POprocon <- function(Rating)
  # Function takes output of Pickind as input.
  {CaseIDs <- Rating[,1]
   Ratings <- Rating[,-1]
   Ncase <- length(CaseIDs)
   Ncol <- length(Ratings)
   DD <- Ncase - 1
   Status1 <- rep(-1,Ncase)
   Status2 <- Status1
   for(I in 1:Ncase)
    {Nosub <- 0; Levl <- 0
     for(J in 1:Ncase)
      {if(I<J | I>J)
        {MatchA <- 0; MatchB <- 0; Undom <- 1
         VecA <- Ratings[I,] - Ratings[J,]
         if(max(VecA) > 0) MatchA <- 1
         if(min(VecA) < 0) MatchB <- 1
         if(MatchA==1 & MatchB==0) Nosub <- Nosub + 1
         if(MatchA==0 & MatchB==1) Undom <- 0
         Levl <- Levl + Undom
        }
      }
     Status1[I] <- Nosub
     Status2[I] <- DD - Levl
    }
   Pct <- 100.0/DD
   PRO <- round((Status1 * Pct),digits=2)
   CON <- round((Status2 * Pct),digits=2)
   POpropensity <- cbind(CaseIDs,PRO,CON)
   POpropensity
  }
```

Appendix 6: R TieSpecs Program for Membership of Instances in Tied Sets

```
TieSpecs <- function(Rating)
  # Function takes extended output of BONDings as input.
  {Ncase <- length(Rating[,1])
   Status1 <- Rating[,2]
   Status2 <- Rating[,3]
   TieSets <- rep(0,Ncase)
   TieLink <- TieSets
   TopTie <- 1
   for(I in 1:Ncase)
    {Ties <- 0
```

```
     for(J in 1:Ncase)
      {if(I<J | I>J)
     {if(Status1[I]==Status1[J] & Status2[I]==Status2[J]
& TieSets[J]<1)
        {Ties <- Ties+1
         TieSets[J] <- TopTie
        }
       }
      }
     if(Ties>0)
      {TieSets[I] <- TopTie
       TopTie <- TopTie + 1
      }
    }
   TopTie <- TopTie - 1
   if(TopTie > 0)
    {for(I in 1:Ncase)
     {TieTo <- 0
      if(TieSets[I]>0 & I<Ncase)
       {TieSet <- TieSets[I]
        II <- I + 1
        for(J in II:Ncase)
         if(TieSets[J]==TieSet & TieTo==0) TieTo <- J
       }
      TieLink[I] <- TieTo
     }
    }
   LexIndx <- Rating[,5]
   if(TopTie>0)
    {ccc <- rank(LexIndx,ties.method="first")
     for(I in 1:TopTie)
      {TieLo <- Ncase
       for(J in 1:Ncase)
              {if(TieSets[J]==I  &  ccc[J]<TieLo)  TieLo
<- ccc[J]
        if(TieSets[J]==I & TieLink[J]==0) TieLink[J] <-
-1 * (TieLo-1)
        }
      }
    }
   POpropensity <- cbind(Rating,TieSets,TieLink)
   POpropensity
  }
```

Appendix 7: R TIEphasA Function for Dropping Indicators to Break Ties

```
TIEphasA <- function(Ratings,Lexings,KeepOrdr,SepraSet)
   # Ratings is output of Pickind.
   # Lexings is output of TieSpecs.
   # KeepOrdr is retention priority order for ratings.
   # SepraSet is TieSets number in Lexings.
   {Ncase <- length(Lexings[,1])
    DD <- Ncase - 1
    Inset <- 0
    for(I in 1:Ncase)
     if(Lexings[I,6]==SepraSet) Inset <- Inset + 1
    TieIds <- rep(0,Inset)
    TieDex <- 1
    for(I in 1:Ncase)
     if(Lexings[I,6]==SepraSet)
      {TieIds[TieDex] <- Lexings[I,1]
       TieDex <- TieDex + 1
      }
    Keeps <- length(KeepOrdr)
    Rated <- length(Ratings) - 1
    Outings <- Keeps * Inset
    Status1 <- rep(-1,Outings)
    Status2 <- rep(-1,Outings)
    IDti <- rep(0,Outings)
    Step <- rep(0,Outings)
    Outdex <- 0
   # Drop cycle
   Keeping <- Keeps
   Instep <- 0
   for(M in 1:Keeps)
    {VecA <- rep(0,Keeping)
     VecB <- rep(0,Keeping)
     for(II in 1:Inset)
      {I <- TieIds[II]
       Nosub <- 0; Levl <- 0
       for(K in 1:Keeping)
        {KK <- KeepOrdr[K] + 1
         VecA[K] <- Ratings[I,KK]
        }
       for(J in 1:Ncase)
        {for(K in 1:Keeping)
          {KK <- KeepOrdr[K]+1
```

```
           VecB[K] <- Ratings[J,KK]
          }
         if(I<J | I>J)
          {MatchA <- 0; MatchB <- 0; Undom <- 1
           VecB <- VecA - VecB
           if(max(VecB) > 0) MatchA <- 1
           if(min(VecB) < 0) MatchB <- 1
           if(MatchA==1 & MatchB==0) Nosub <- Nosub + 1
           if(MatchA==0 & MatchB==1) Undom <- 0
           Levl <- Levl + Undom
          }
        }
       Outdex <- Outdex + 1
       Status1[Outdex] <- Nosub
       Status2[Outdex] <- DD - Levl
       Step[Outdex] <- Instep
       IDti[Outdex] <- I
      }
     Keeping <- Keeping - 1
     Instep <- Instep - 1
    }
   Pct <- 100.0/DD
   Pro <- round((Status1 * Pct),digits=2)
   Con <- round((Status2 * Pct),digits=2)
   TIEBONDS <- rep(0,Outings)
   for(I in 1:Outings)
      TIEBONDS[I]  <-  (Pro[I]  -  Con[I])  *  (Pro[I]  +
Con[I])/100.0
   Untidrop <- cbind(Step,IDti,Pro,Con,TIEBONDS)
   Untidrop
  }
```

Appendix 8: R TIEphasB Function for Breaking Ties with Individual Indicators

```
TIEphasB <- function(Ratings,Lexings,KeepOrdr,SepraSet)
   # Ratings is output of Pickind.
   # Lexings is output of TieSpecs.
   # KeepOrdr is priority order for ratings.
   # SepraSet is TieSets number in Lexings.
   {Ncase <- length(Lexings[,1])
    DD <- Ncase - 1
```

```
 Inset <- 0
 for(I in 1:Ncase)
  if(Lexings[I,6]==SepraSet) Inset <- Inset + 1
 TieIds <- rep(0,Inset)
 TieDex <- 1
 for(I in 1:Ncase)
  if(Lexings[I,6]==SepraSet)
   {TieIds[TieDex] <- Lexings[I,1]
    TieDex <- TieDex + 1
   }
 Keeps <- length(KeepOrdr)
 Rated <- length(Ratings) - 1
 Outings <- Keeps * Inset
 Status1 <- rep(-1,Outings)
 Status2 <- rep(-1,Outings)
 IDti <- rep(0,Outings)
 Step <- rep(0,Outings)
 Outdex <- 0
# Singular cycle
Keeping <- 1
Instep <- 1
for(M in 1:Keeps)
 {for(II in 1:Inset)
   {I <- TieIds[II]
    Nosub <- 0; Levl <- 0
    K <- Keeping
    KK <- KeepOrdr[K] + 1
    VecA <- Ratings[I,KK]
    for(J in 1:Ncase)
     {K <- Keeping
      KK <- KeepOrdr[K]+1
      VecB <- Ratings[J,KK]
      if(I<J | I>J)
       {MatchA <- 0; MatchB <- 0; Undom <- 1
        VecB <- VecA - VecB
        if(VecB > 0) MatchA <- 1
        if(VecB < 0) MatchB <- 1
        if(MatchA==1 & MatchB==0) Nosub <- Nosub + 1
        if(MatchA==0 & MatchB==1) Undom <- 0
        Levl <- Levl + Undom
       }
     }
    Outdex <- Outdex + 1
    Status1[Outdex] <- Nosub
    Status2[Outdex] <- DD - Levl
```

```
       Step[Outdex] <- Instep
       IDti[Outdex] <- I
      }
     Keeping <- Keeping + 1
     Instep <- Instep + 1
    }
   Pct <- 100.0/DD
   Pro <- round((Status1 * Pct),digits=2)
   Con <- round((Status2 * Pct),digits=2)
   TIEBONDS <- rep(0,Outings)
   for(I in 1:Outings)
     TIEBONDS[I] <- (Pro[I] - Con[I]) * (Pro[I] +
Con[I])/100.0
   Unti1x1 <- cbind(Step,IDti,Pro,Con,TIEBONDS)
   Unti1x1
  }
```

Appendix 9: R TIEphasC Function for Assigning Ranks Among Ties

```
TIEphasC <- function(Lexings,Untidrop,Unti1x1,SepraSet)
# Tie Resolving Indicator Modification ranks
# Lexings is output of TieSpecs
# Untidrop is output of TIEphasA
# Unti1x1 is output of TIEphasB
# SepraSet is TieSets number in Lexings
{Ncase <- length(Lexings[,1])
 Inset <- 0
 SubTie <- 0
 for(I in 1:Ncase)
  {if(Lexings[I,6]==SepraSet) Inset <- Inset + 1
   if(Lexings[I,6]==SepraSet & Lexings[I,7]<0) SubTie <-
-1 * Lexings[I,7]
    }
   TieIds <- rep(0,Inset)
   TieDex <- 1
   for(I in 1:Ncase)
    if(Lexings[I,6]==SepraSet)
     {TieIds[TieDex] <- Lexings[I,1]
      TieDex <- TieDex + 1
     }
   Unties <- rbind(Untidrop,Unti1x1)
```

```
Outings <- length(Unties[,1])
TRIMrank <- rep(0,Inset)
LexRnk <- rep(0,Inset)
for(I in 1:Inset) LexRnk[I] <- SubTie + I
Instep <- rep(0,Inset)
Idone <- 0
Ilo <- 1
Ihi <- Inset
Nleft <- Inset
while(Idone < 1)
 {Lexleft <- rep(0,Nleft)
  StepLex <- rep(0,Nleft)
  StepRank <- rep(0,Nleft)
  J <- 1
  for(I in 1:Inset)
   if(TRIMrank[I]==0) {Lexleft[J] <- I;J <- J + 1}
  J <- 1
  K <- Ilo
  for(I in 1:Inset)
   if(TRIMrank[I]==0)
    {StepLex[J] <- Unties[K,5]
     J <- J + 1
     K <- K + 1
    }
  StepRank <- rank(StepLex,ties.method="average")
  for(I in 1:Nleft)
   {Tied <- 0
    for(J in 1:Nleft)
     if(I != J & StepRank[I]==StepRank[J]) Tied <- 1
    if(Tied==0)
     {K <- Lexleft[I]
      KK <- StepRank[I]
      TRIMrank[K] <- LexRnk[KK]
      LexRnk[KK] <- 0
      Instep[K] <- Unties[Ilo,1]
     }
   }
  Nleft <- 0
  KK <- 0
  for(K in 1:Inset)
   if(LexRnk[K]>0) Nleft <- Nleft + 1
  if(Nleft>0)
   {for(K in 1:Inset)
     {if(LexRnk[K]>0)
       {KK <- KK +1
```

```
         LexRnk[KK] <- LexRnk[K]
        }
      }
     if(KK<Inset)
      {KK <- KK + 1
       for(K in KK:Inset) LexRnk[K] <- 0
      }
    }
   if(Nleft==0) Idone <- 1
   Ilo <- Ilo + Inset
   Ihi <- Ihi + Inset
   if(Ihi>Outings) Idone <- 1
  }
 LexdSet <- cbind(TieIds,TRIMrank,Instep)
 LexdSet
}
```

References

Allerhand M (2011) A tiny handbook of **R**. Springer, New York, NY

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17(4):383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, New York, NY

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Brüggemann R, Sorensen P, 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, Simon U, Mey S (2005) Estimation of averaged ranks by extended local partial order models. Match Commun Math Comput Chem 54:489–518

Chander G, Huang C, Yang L, Homer C, Larson C (2009) Developing consistent Landsat data sets for large area applications – the MRLC protocol. IEEE Geosci Remote Sens Lett 6(4):777–781

De Loof K, De Baets B, De Meyer H, Brüggemann R (2008) Hitchhiker's guide to poset ranking. Comb Chem High Throughput Screen 11:734–744

Homer C, Huang C, Yang L, Wylie B, Coan M (2004) Development of a 2001 National Landcover Database for the United States. Photogramm Eng Remote Sensing 70(7):829–840

Myers W, Patil GP (2010) Preliminary prioritization based on partial order theory and **R** software for compositional complexes in landscape ecology, with applications to restoration, remediation, and enhancement. Environ Ecol Stat 17:411–436

Myers, W, Patil GP (2011) Geoinformatics for human environment interface. In: Proceedings of the joint statistical meetings (JSM) 2011, July 31, 2011, Miami Beach, FL, session 206322, presentation 300319, http://www.amstat.org on-line archives

Myers W, Patil GP (2012a) Statistical geoinformatics for human environment interface. Chapman & Hall/CRC, Boca Raton, FL

Myers W, Patil GP (2012b) Multivariate methods of representing relations in **R** for prioritization purposes: selective scaling, comparative clustering, collective criteria and sequenced set. Springer, New York, NY

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

R Development Core Team (2008) **R**: A language and environment for statistical computing. **R** Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R--project.org/

Short T (2009) **R** reference guide. Revolution Computing, New Haven, CT

Venables WN, Smith DM, the R Development Core Team (2005) An introduction to **R**. Network Theory LTD, Bristol

Chapter 9
Partial Orders in Socio-economics: A Practical Challenge for Poset Theorists or a Cultural Challenge for Social Scientists?

Marco Fattore and Filomena Maggino

Abstract In this “position paper” we discuss the potential role of partial order theory in socio-economic statistics and social indicators construction. We maintain that the use of concepts and tools from poset theory is needed and urgent to improve currently adopted methodologies, which often prove ineffective for exploiting ordinal data. We also point out that the difficulties in spreading partial order tools are cultural in nature, and that some open-mindedness is needed among social scientists. We address these issues introducing some examples of open questions in socio-economic data analysis: (i) the problem of multidimensional poverty evaluation, (ii) the problem of assessing inequality and societal polarization, and (iii) the problem of clustering in multidimensional ordinal datasets.

9.1 Introduction

During a workshop held in Italy in 2010, a full professor in Statistics, presenting an evaluation study pertaining to service quality and based on ordinal data, made a statement like: “…here we’re dealing with ordinal data, so there is no room for mathematics and statistics.” The speaker was certainly aware of the number of methodologies in the statistical literature for dealing with ordinal variables. Yet the statement reveals something true and, somehow, interesting. Still today, when social

M. Fattore (✉)
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
Milano, Italy
e-mail: marco.fattore@unimib.it

F. Maggino
Department of Statistics, Computer science, Applications “G. Parenti”,
University of Florence, Florence, Italy
e-mail: filomena.maggino@unifi.it

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_9, © Springer Science+Business Media New York 2014

scientists address multidimensional complex[1] problems involving qualitative data (like service quality evaluation studies) they feel basically uncomfortable and consider such problems, in a sense, “ill-posed.” Unfortunately, this perception conflicts with the following evidences: (i) more and more crucial socio-economic issues may be meaningfully described only by involving ordinal information (e.g., material deprivation and multidimensional poverty, subjective well-being or quality-of-life, customer satisfaction and service quality perception, to mention a few) and (ii) de facto, more and more socio-economic datasets offer ordinal data to scholars and researchers. The question is therefore whether this means that we cannot adequately describe and understand socio-economic facts, or whether this is still possible, but requires some change of paradigms and tools. In this position paper, we address this issue. We try to identify the logical roots of the problem and to reveal the interconnections between them and the need for new statistical tools, stressing the role of partial order theory and of a “partial order culture,” to overcome limitations of current statistical practice. We will pursue this, identifying some questions that are still unsolved in applied socio-economic research and that could be (and are being) fruitfully addressed through partial order theory. The proposed open issues are in no way intended to be an exhaustive list. They have been chosen based on our research interests and experience; still, they are really urgent problems and are useful to explain our position. Given the aims of the paper, we will not go into technical details, which would lead us to a long and complicated exposition. Rather, we will focus on the essentials which better clarify the issues we are addressing.

9.2 Old Paradigms and Open Questions

Among statisticians and social scientists, there is a widespread feeling that sound scientific knowledge may be achieved only when precise measurements may be attained. This idea comes from natural sciences, physics in particular, and was embodied in social sciences from the late nineteenth century to the first half of twentieth century. This is one of the root problems, often preventing ordinal data to be considered valuable. Clearly, we are not arguing against the relevance of measuring socio-economic facts precisely, whenever possible. The question one should answer is in fact different: “How can we obtain faithful representations of socio-economic facts?” Sometimes, faithful representations may be built using precise measurement models, as in physics, sometimes not. We may, at least ideally, measure with great precision prices and quantities of different goods to account for

[1] In this paper, we often use terms like “complex” or “complexity.” We use them informally, to mean problems or systems that cannot easily be solved, reduced or described, as they are made of many linked elements or facets. Perhaps the formal definition of complexity which is closest in spirit to the way we employ the term is that of “Kolmogorov complexity,” used in algorithmic information theory. However, we stress that this is only an analogy.

inflation,[2] but we cannot claim to faithfully represent in numerical terms ordinal issues like the democracy level of a country or subjective well-being or quality of services. Representing a physical, economic, or social phenomenon means identifying its essential features and the interrelations among its components and sketching them in a suitable formalism, so that by performing formal computations one may get insights into it. Often, social scientists simply code qualitative information into numerical scores and proceed to analytical computations. Sometimes, prior to computations they employ complex algorithms to turn qualitative variables into numerical scales. In any case, the basic questions one should answer are: "What do these numbers represent? May we assume the results are effective to understand what we are interested in? Do our computations convey valid information, reflecting elements of reality?" We admit that sometimes the answer may be positive, so one may fruitfully proceed this way. But it is important to pose the question.

The problem, in fact, is not just epistemological, but very practical. Today there is a great amount of ordinal information available to social scientists, and focus is shifting toward qualitative socio-economic issues, like assessing well-being, quality-of-life, or multidimensional poverty. Still, social scientists approach these topics with "numerical" paradigms, using methodologies and computational procedures designed to deal with quantitative information. So the problem often becomes how to fit ordinal data into well-established and routinely used procedures, rather than how to build new appropriate methodologies. When ordinal data is used in this way, the risk is getting questionable and biased results, which affect our understanding of social facts. This matter of fact is also due to the implicit assumption that ordinal data cannot be handled in a consistent and effective way, since no formal tools are available; so even those who are aware of the problem cannot easily see any way out. The use of partial order theory and other related tools from discrete mathematics and relational calculus have not spread into the "methodological imagination" of social scientists' community yet, and there is little awareness of the possibilities that they may open.

A paradigmatic example comes from the problem of extending classical socio-economic indicators (e.g., inequality indices) to multidimensional ordinal datasets. This is one of the core issues in current research, since any attempt to represent modern societies and their complexity requires taking into account many different aspects jointly. Historically, there has been a great deal of research on giving sound mathematical and axiomatic foundations to the theory of statistical indices (consider, e.g., the theories of price indices, poverty indices, or concentration indices). It is much more difficult to achieve multidimensional extensions of these axiomatic systems and, usually, results are less neat and general. In the case of ordinal variables the situation is even worse, since systematic theories of this kind are still lacking, even if some attempts are being made. Unfortunately, the use of statistical indicators is the basis of many socio-economic studies and, in many cases, the

[2] Measuring inflation should involve also measuring services and it is quite debatable how to precisely define the concept of quantity in this case.

absence of effective tools for ordinal data forces social scientists to fall back on numerical representations.

On the whole, epistemological difficulties, lack of awareness about possible alternatives, and unsuitable statistical tools are major obstacles for the development of statistical methodologies capable of exploiting ordinal data and answering the information needs of researchers, policy-makers, and citizens. In the following paragraphs, we illustrate these issues and give some ideas about the role of partial order theory, by means of three examples of open questions in applied socio-economic statistics: the evaluation of multidimensional poverty and well-being, the measurement of inequality and polarization in a multidimensional ordinal setting, and the development of procedures to perform cluster analysis on ordered structures.

9.2.1 Evaluating Multidimensional Poverty: A Matter of Multidimensional Comparison?

One of the most relevant examples in socio-economic analysis where the issue of multidimensional ordinal data is crucial is the wide field of evaluation studies pertaining to quality-of-life, well-being, and multidimensional poverty. Following the Commission on the Measurement of Economic Performance and Social Progress (the "Stiglitz-Sen-Fitoussi Commission"), several attempts to assess well-being and to go beyond GDP (Gross Domestic Product) as a measure of societal wealth are being pursued in many countries. The authors of this paper are involved in the Italian project for the construction of official well-being indicators,[3] promoted by CNEL (National Council of Economy and Work) and ISTAT (National Institute of Statistics). Twelve well-being dimensions have been identified (e.g., health, education, work, social relations, and environment), each comprising different indicators, both numerical and ordinal (e.g., those pertaining to subjective well-being). One of the main issues under discussion is whether to produce synthetic indicators and how; perhaps computing composite indicators? The drawbacks of aggregative and compensative procedures of this kind are well known (see, e.g., Fattore et al. 2012), but often no alternatives are pursued. Given the impact of official well-being statistics on public opinion and policy-makers, it is clear that any choice about how to produce final indicators requires great care. In social evaluation studies, the aggregation problem is at least twofold:

1. There is a technical issue when ordinal data are at hand, since in that case usual procedures designed for numerical variables break down and no aggregation can be performed directly. To overcome this problem, various procedures are often implemented to transform ordinal data into numerical figures,[4] so as to apply

[3] http://www.misuredelbenessere.it.

[4] These so-called *scaling tools* range from simply coding and using ordinal scores as integers, to running complex numerical algorithms, like in the Gifi homogeneity analysis (Michailides and de Leeuw 1998), or using various regression or model-based approaches.

aggregation procedures. Unfortunately, the existence of latent numerical scales behind ordinal data may often be questioned. Moreover, scaling procedures often generate numerical figures minimizing some loss function, so in practice introducing into the analysis an optimization criterion that need not be intrinsic to the data, albeit mathematically appealing. Therefore, one may legitimately ask whether the final figures produced that way actually give a faithful representation of the underlying socio-economic facts or are just the output of numerical algorithms with a limited capability of enlightening data.

2. There is also a general conceptual problem. The basic assumption behind the development of aggregated indicators is the existence of one main latent dimension accounting for most data variability, so that by exploiting variable interdependencies one may hope to reduce data complexity. As a matter of fact, evaluation dimensions are often weakly interdependent and, even conceptually, one cannot accomplish any satisfactory synthesis, drawing on the principle of explaining joint variability. We remark that this problem does not depend upon the nature (cardinal or ordinal) of the variables to handle. It is intrinsic to the true multidimensionality of the concepts related to quality-of-life, which often prevents the aggregative procedure from getting meaningful results. What makes social evaluation studies challenging is precisely this feature; the evaluation problem is not reducible to aggregation.

In practice, and more and more often, the two problems combine together, making the development of synthetic indicators more demanding for statisticians, who must find new technical tools to build them, and more urgent for policy-makers, who need them to interpret even more complex societies. The current debate on these problems is quite heated. An interesting issue of the Journal of Economic Inequality published in 2011, hosting a forum on the topic, is particularly enlightening of the state-of-the-art. The main debate is polarized around two different positions: that of Alkire and Foster, who propose their aggregative counting approach to the measurement of multidimensional poverty (see Alkire and Foster 2011a and Alkire and Foster 2011b), and that of Ravaillon Ravaillon [2011], who suggests avoiding any synthetic procedures, in favor of using dashboards (panels of indicators). The Alkire–Foster procedure is perhaps the most consistent framework to assess multidimensional poverty based on both ordinal and cardinal indicators, and its use is spreading. The structure of the procedure is very simple and can be described as follows.

Let $T_{n \times k}$ be the data matrix, comprising the scores of n statistical units on k evaluation dimensions $v_1, \ldots, v_k$ (i.e., each row of the data matrix contains the profile of the corresponding statistical unit). Then the following steps are implemented:

1. a set of $c_1, \ldots, c_k$ cutoffs is exogenously chosen, one for each evaluation dimension;
2. each individual is assessed against the cutoffs and is declared deprived on v_i if his/her score on that dimension is less than c_i;

3. matrix $T_{n\times k}$ is transformed into a binary matrix $G_{n\times k}$, where $G_{ij}=1$ if individual i is deprived on dimension j and $G_{ij}=0$ otherwise;
4. the rows of G are then summed up, possibly weighting each column with weights expressing the relative importance of being deprived on the various dimensions;
5. finally, an individual is declared definitely deprived if his/her overall score is equal to or greater than an overall cutoff c, exogenously chosen.

In practice, this procedure leads to the definition of an identification function which classifies individuals as deprived or not in a binary way. Once individuals have been classified, several poverty indicators may be computed (for a complete discussion see Alkire and Foster 2011a). Notice that irrespective of dealing with cardinal or ordinal variables, the Alkire–Foster procedure turns the original data matrix into binary matrix G and applies a weighted aggregation function (i.e., a weighted sum) to its rows. If all of the weights are set to 1, the methodology simply counts individual deprivations.

The debated point is essentially on the meaningfulness and utility of aggregating indices using weights. It is instructive to quote the final comment of the Forum Editor (Lustig 2011):

> At the bottom of the discussion is a fundamental disagreement on the "legitimacy" of the weights used to aggregate dimensions of wellbeing […] Ravallion and those who agree with him consider that the alternative weights used in the MPI (or similar indices) are not a good solution as they may imply unappealing trade-offs and that these aggregate poverty measures are generally not consistent with consumer welfare theory.[…] Thus, given this problem and the fact that for policy purposes disaggregation will be required, Ravallion asks: what is the advantage of using composite indices […] instead of a "dashboard" of multiple indices? One key unresolved issue in the "dashboard approach", however, is that if we agree that welfare depends on a series of dimensions, how do we address the fact that the marginal effect of increasing an individual's access to one of the dimensions (e.g., health services) depends not only on that individual's access to the dimension in question, but also on the individual's level of all the other indicators of welfare?
>
> Future research will need to focus on how to identify weights in ways that are consistent (1) with welfare economics and (2) with theories of justice. Will we have to choose between the two?

From the last sentence we see that the weighting problem is considered as the central issue. But weighting is a consistent operation only in a numerical setting, so what about ordinal data? Basically, we are left with two alternatives: (i) scaling ordinal scores to cardinals and proceeding to usual computations, getting arguable results or (ii) sticking to the Alkire–Foster procedure, turning ordinal scores into binary scores and counting, possibly with weights, losing a great deal of information on the degree of individual and societal poverty (Fattore et al. 2011b). Both cases seem to be driven by the (presumed) impossibility of exploiting ordinal data on their own. The debate goes on trapped within the "weight and aggregate" framework, the problem being to search for more sophisticated weighting procedures or, as a radical alternative, to abandon synthetic indicators.

In our view, the way out of this trap is via some change of paradigm:

1. no longer considering “synthetic indicator” and “aggregated indicator” as equivalent concepts;
2. considering that evaluation processes could be better addressed as problems of multidimensional comparisons against suitable benchmarks, rather than problems of aggregation;
3. realizing that ordinal data may be consistently handled, with appropriate mathematical tools.

In the wider literature about evaluation, the terms “synthetic” and “aggregated” are used interchangeably and it is taken for granted that to get synthetic information, some aggregation procedure is needed. Since aggregating basically requires summing up scores, we are inevitably led back to the problem of incompatibility between analytical tools and ordinal data. The problem would be solved if we could get synthetic indicators from ordinal data without aggregating variables. This is indeed possible, provided we reconsider the evaluation process as a multidimensional comparison problem and employ partial order theory to address it. In assessing multidimensional poverty and similar issues, no natural measuring scale exists and, implicitly or explicitly, assessments are often based on the selection of some “reference points” or of some “prototypes,” to be used as benchmarks.[5] In the unidimensional case, e.g., monetary thresholds are adopted and individuals’ income is compared against them. In a multidimensional setting the concept of benchmark is more complex. First, multidimensional poverty may assume different shapes, i.e., there are “several ways” to be poor and so more benchmarks are needed; second, being poor or not depends upon the individual’s scores on all the dimensions of concern (i.e., his/her profile), so benchmarks should be identified in terms of prototypical score configurations. Assessing the poverty state of an individual therefore means comparing his score configuration to those constituting the benchmarks. Being a problem of multidimensional comparison, partial ordered theory naturally comes into play.

This idea is currently being pursued by the authors and other colleagues, e.g., in Fattore et al. [2011a,b], and Fattore et al. [2012]. Without entering into technical details, the basic idea is quite simple. Let $v_1,\ldots,v_k$ be k ordinal evaluation dimensions. To each individual in the population, a profile $\mathbf{p}=(p_1,\ldots,p_k)$ is associated, whose components are the scores of the statistical unit on the evaluation dimensions. The set P of profiles is turned into a partially ordered set $(P,\trianglelefteq)$ defining

$$\mathbf{p} \trianglelefteq \mathbf{q} \Leftrightarrow p_i \leq q_i \quad \forall i=1,\ldots,k. \tag{9.1}$$

In this framework, a multidimensional poverty threshold τ is a minimal set of profiles[6] such that any profile below[7] one of its elements is classified as poor. Given

[5] Whenever cutoffs or thresholds are involved in the assessment, benchmarks are de facto introduced.

[6] That is, the smallest set of profiles with the cited property.

[7] Here, we assume that the lower the scores, the worse off the individual.

the threshold, any other profile may be assessed in terms of poverty, based on its position with respect to τ, in the Hasse diagram of the profile poset. The multidimensional comparisons involving the profiles and the threshold are performed counting over linear extensions of $(P, \lhd)$ how frequently a profile is classified below an element of the threshold (see Fattore et al. 2011a for details). The resulting *evaluation function* assigns to each profile (and thus to any individual sharing it) a score in $[0, 1]$, representing the degree of poverty, given τ. In practice, the procedure quantifies the degree of ambiguity in the classification of a profile into the set of poor profiles and may be better interpreted as a way to compute a fuzzy membership function. What is relevant here, is that such a quantification does not involve any ordinal variable scaling; the focus is on profiles and information is extracted out of the mathematical structure representing the basic relation existing among them, i.e., the partial order relation. The resulting evaluation procedure, even if heavier from a computational point of view, is more effective and general than the Alkire–Foster counting approach, which in fact may be seen as a special case of the former (for a complete comparison, see Fattore et al. 2011b). Apart from these technicalities, the interesting feature of poset-based evaluation procedures is that they show how to exploit ordinal data, implementing the same logical structure of classical unidimensional evaluation studies.[8]

To provide some insights into the poset approach to evaluation, we briefly outline the example reported in Fattore et al. [2011a]. Five deprivation variables have been considered, from the EU-SILC survey pertaining to Italy, for year 2004:

1. HS040—Capacity to afford paying for one week annual holiday away from home;
2. HS050—Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day;
3. HS070—Owning a phone (or mobile phone);
4. HS080—Owning a color TV;
5. HS100—Owning a washing machine.

All of the variables are expressed on a yes/no scale, so that deprivation profiles are sequences of five 0/1 digits (1: non-deprivation; 0: deprivation). Clearly, there are $2^5 = 32$ different profiles. The threshold has been chosen as $\tau = (01011, 00111)$, i.e., deprivation on HS040 and HS070, or deprivation on HS040 and HS050.[9] Figure 9.1 reports the Hasse diagram of the profile poset, with the threshold elements in black. The top element corresponds to profile 11111, the bottom to profile 00000. Profiles with the same number of 1s have the same distance from the bottom element (the distance is measured as the number of edges in a downward path from the profile to the bottom). All of the elements of the threshold and all of the profiles below one of them are scored 1 (i.e. unambiguously deprived) by the evaluation

[8] In Fattore et al. [2012], e.g., it is also suggested how the classical notion of "weighting" variables may be translated into purely poset terms.

[9] This threshold has been chosen for exemplification purposes only.

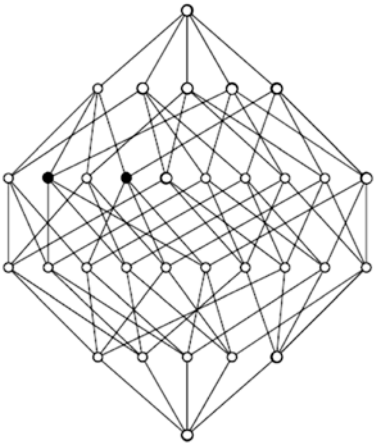


Fig. 9.1 Profile poset on a set of five binary variables (threshold elements in *black*)

Table 9.1 Evaluation function for the elements of the profile poset depicted in Fig. 9.1, given the threshold $\boldsymbol{\tau} = (01011, 00111)$

Profile	11111	11110	11101	11100	11011	11010	11001	11000
Evaluation	0.00	0.11	0.11	0.65	0.06	0.66	0.66	0.98
Profile	10111	10110	10101	10100	10011	10010	10001	10000
Evaluation	0.6	0.66	0.66	0.98	0.67	0.98	0.98	1.00
Profile	01111	01110	01101	01100	01011	01010	01001	01000
Evaluation	0.00	0.67	0.67	0.98	1.00	1.00	1.00	1.00
Profile	00111	00110	00101	00100	00011	00010	00001	00000
Evaluation	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

function, since they represent profiles that are deprived as much as or more than threshold profiles. Profiles above both threshold elements are instead scored 0 by the evaluation function (a direct inspection of the Hasse diagram shows that there are only two such elements, namely 01111 and 11111). All of the other elements of the poset are scored in $[0, 1]$ by the evaluation function. The choice of the threshold breaks the symmetry of the profile poset, so that profiles with the same number of 1s may be scored differently. For the sake of completeness, Table 9.1 reports the evaluation function, estimated extracting a sample of 10^8 linear extensions out of the profile poset. For more comments on the results, see Fattore et al. [2011a].

Remark. We conclude this paragraph with a brief discussion on the possible use of partial order theory in evaluation problems involving continuous variables. When truly numerical variables are at hand, there is apparently no need for partial order theory to be employed. At least in principle, composite indicators might be computed and metric information preserved. However, a closer look at the problem shows that things are not so neat. Composite indicators often mix up variables expressed on different scales, producing almost uninterpretable results. Even the trick of scaling variables to unit variance does not solve the problem and could be justified only by assuming a latent variable model behind observed data. Unfortunately, involving latent constructs leads to other subtle issues, such as the indeterminacy of factor scores (Vittadini 1989, 2007), which raise doubts on the validity of the approach. Moreover, as previously noticed, interdependencies among poverty variables are often not very strong, reducing the effectiveness of correlation-based procedures. In addition to these technical problems, other (perhaps more fundamental) considerations may be given in favor of partial order theory. As already discussed, evaluation is primarily a problem of comparison against benchmarks, rather than against an absolute scale. In the composite indicator approach (with or without latent variables), benchmarking is performed through variable aggregation, i.e., achieving unidimensionality to eliminate incomparabilities. Aggregation introduces compensations and trade-offs between evaluation dimensions, which are often debatable, but usually accepted as the only way to get full comparability among statistical units. However, if one addresses multidimensional evaluation through partial order theory, i.e., as a problem of "comparability quantification," the existence of incomparabilities stops being a problem and aggregation is no longer needed, even conceptually. Up to now, this point of view has been pursued only for ordinal data, but one may hold it also about continuous variables. There are indeed some technical problems to face. For example, the trick of considering linear extensions may not be directly applied to continuous partial orders and its implementation must be reconsidered (we are currently working on that and a possible solution has been already identified). Apart from these technical issues, however, we see that poset theory is a general tool for multidimensional evaluation problems, since its conceptual and formal structure is fully consistent with the very nature of evaluation processes. This is one of the main reasons why poset theory should be part of the "encyclopedia" of social scientists.

9.2.2 Inequality and Polarization in Multidimensional Ordinal Datasets: How to Assess Them?

A major concern in current socio-economic research is assessing inequality patterns among individuals and polarization within societies. Historically, the measure of inequality is one of the most studied and developed research fields in socio-economic statistics and the amount of literature about it is huge. Inequality measurement focused primarily on income distribution and monetary well-being. This led to classical axiomatic systems for inequality and concentration indices. As social scientists' and policy-makers' focus is shifting from a monetary analysis to well-being, questions about inequality are moving toward a multidimensional setting, often comprising ordinal information. A similar process is occurring with respect to another crucial phenomenon affecting modern societies and which is attracting more and more interest by social scientists: social polarization. The first aspect of polarization, technically referred to as *bi-polarization*, pertains primarily to the well-known phenomenon of the *disappearing middle class*. The existence of the middle class is one of the most relevant consequences of the development of modern societies as an effect of the diffusion of well-being, both in monetary and non-monetary forms. However, in the last two decades, in many countries there are evidences that "the rich get richer and the poor get poorer." That is, societies are becoming polarized and the middle class is partly disappearing. A broader concept of polarization, closely linked to existence of different social and ethnic groups in modern democracies, is related to the "alienation that individuals and groups feel from one another [. . .] fuelled by notions of within-group identity." Here, the interest lies primarily on "the correlates of organized, large-scale social unrest-strikes, demonstrations, processions, widespread violence, and revolt or rebellion. Such phenomena thrive on differences, to be sure. But they cannot exist without notions of group identity either." (see Duclos et al. 2004, pp. 1737–1738). From a conceptual, and then statistical, point of view, it is interesting to notice that inequality and polarization (both in the bi-polarization and in the broader sense) are two distinct concepts. A first evidence of this dates back to the paper of Wolfson [1994], where it is unequivocally shown how a sequence of income distributions may be built with decreasing inequality and increasing bi-polarization. Inequality does not capture either the notion of identification-alienation polarization, as discussed in Duclos et al. [2004]. The interest in polarization led to many statistical studies devoted to measuring it and to developing related axiomatic systems, primarily in the unidimensional case (Duclos et al. 2004, Fusco and Silber 2011, Permanyer 20 12, Zhang and Kanbur 2001). While the concept of polarization is being carefully analyzed and theoretical and empirical differences or interconnections with inequality are being investigated, a new issue is emerging as relevant and urgent. Inequality and polarization do not only concern the monetary perspective; instead they involve the whole well-being concept, comprising health, work, education, culture, environment, material deprivation, and so on. Some interesting studies address the issue of labor market segmentation and the link between job polarization and wage polarization (Ercolani and Jenkins 1998, Gregg and Wadsworth 2004). A great deal of research is also being done on how to measure inequality and polarization in health services and, in particular, in the subjective assessment of health responsiveness (Apouey 2007, Lones 2010). Involving well-being raises two issues in the theory of inequality/polarization measurement: (i) building multidimensional indices and (ii) defining formulas suitable to ordinal data. Multidimensional inequality measures have been already extensively studied (see, e.g. Maasoumi 1999, Tsui 1986), while multidimensional extensions of polarization measures are still at an initial stage (see, e.g. Gigliarano and Mosler 2009). The problem of ordinal data is instead urgent for both inequality and polarization measurements. There are indeed several formulas to treat ordinal information (consider, e.g., Abul Naga and Yalcin 2008, Allison and Foster 2004), but at the same time, and this is the point of interest for our aims, their use still meets some "resistance." For example, in Doorslaer van and Jones [2003] the issues of using ordinal data are clearly addressed and overcome in favor of transforming ordinal scores into numerical figures. Quoting from the Introduction:

> One of the challenges in investigating inequalities in health is that, very often, health information is only available at an ordinal level. One of the most commonly used indicators of overall individual health in general population surveys is the simple question, “how is your health in general?”, with response categories ranging from “very good” or “excellent” to “poor” or “very poor”. This categorical variable has been shown to be a very good predictor variable of other outcomes, such as subsequent use of medical care or of mortality [...]. However, it does not provide a cardinal health (utility) scale that can be used, for instance, for quality adjustments of life expectancy. Categorical measures of health create a problem for the measurement of inequalities in health. The health concentration index, and the related slope index of inequality, require information on health in the form of either a continuous variable or a dichotomous variable. (Doorslaer van and Jones 2003, pp. 61–62).

Ordinal data are thus a problem since the computational and interpretative processes are designed for numerical figures. To be clear, we are not lessening the relevance of jointly considering life expectancy and health status. We simply remark that, according to the quoted text, the whole conceptual framework is not compatible with ordinal data. Whether this is a problem of the data generation process or a limit of the conceptual framework, we leave to the reader. The way out from this incompatibility is usually the application of scaling tools to transform ordinal data into numerical figures, by means of latent variable models, probit models, or other form of regressions, together with all the burden of hypotheses and assumptions that they carry (Doorslaer van and Jones 2003), affecting in some way the final computations. An interesting example showing possible effects of scaling is provided by the study reported in Madden [2010], which concerns health status in Ireland for years 2003–2006. Inequality in self-reported health status is analyzed and compared using the Abul Naga and Yalcin indices (designed for ordinal data) and, after transforming original ordinal data into cardinal figures by means of interval regression, through the Generalized Entropy indices. Before commenting on the results of the study, it is interesting to quote its motivations:

> As the vast majority of summary inequality indices are mean-based they require a cardinal measure of the outcome variable in question. While there are some health measures that are cardinal (e.g. body mass index) they are typically not comprehensive. More general health measures are nearly always categorical and ordinal rather than cardinal. Thus, to obtain a summary measure of inequality it is necessary to either (a) employ an inequality measure that is specifically designed to deal with ordinal data or (b) to transform the ordinal measure into a cardinal measure and then employ a standard inequality index. [...] It could be argued that since inequality measures specifically designed to deal with ordinal data are now available, analysts should always use such indices. However, it also seems fair to suggest that such measures are less well developed than their cardinal counterparts.

Table 9.2 reproduces a part of the results reported in Madden [2010]; it compares the inequality measures obtained by the Abul Naga and Yalcin index and those obtained by the Generalized Entropy index for each year.[10]

[10] The Abul Naga and Yalcin index is computed setting $\alpha = \beta = 1$, while the Generalized Entropy is computed setting $a = 0$.

Table 9.2 Comparison between inequality measures computed using the Abul Naga and Yalcin (AN–Y) index and the Generalized Entropy (GE) index, extracted from Table 2 of Madden [2010]

Year	AN–Y	GE
2003	0,3563	0.0039
2004	0.3455	0.0043
2005	0.3427	0.0045
2006	0.3330	0.0040

In Madden [2010], it is emphasized that an absolute comparison between the two indices is not meaningful and so the author focuses on the orderings induced by them on the four years considered. As it can be directly checked, the orderings are very different: according to the Abul Naga and Yalcin Index we get 2006 < 2005 < 2004 < 2003, while according to Generalized Entropy we have 2003 < 2006 < 2004 < 2005. It is particularly noticeable that year 2003 is ranked as the most unequal year by the first ordering and as the least by the second. Thus we see that the judgment on the temporal evolution of self-reported health status would be almost reversed, if one chooses the ordinal or the cardinal way of measuring inequality.

This example is quite instructive and shows the consequences of assuming latent cardinal variables behind ordinal data (not to mention the problems concerning the numerical results of the scaling procedure: are the differences in the Generalized Entropy measures really significant?).

When the issues of multidimensionality and of using ordinal data combine together, the situation becomes much more complex and challenging. The problem is to build a multidimensional index of inequality or polarization for ordinal data and this seems to be an almost completely unexplored field. As far as we know, the only attempt in this direction is Kobus [2011]. In general terms, the extension of unidimensional indices to multidimensional settings is pursued building axiomatic systems that generalize unidimensional axioms to sets of many variables. The problem with this approach is that inequality, polarization, or other similar issues in a multidimensional framework may assume so many different forms and may show such a great number of shapes that it is often extremely difficult to identify neatly natural properties to be turned into axioms. The result is quite complicated and debatable axiomatic systems. Without entering into technical details, it seems to us that one of the problems is that axiomatization attempts tend to focus directly on the ordinal variables at hand, instead of focusing on the partial order induced by them on the (equivalence classes of) statistical units. Basically, the approach is to define a partial order on the set of joint ordinal frequency distributions that should reflect a partial ordering of (say) inequalities and then to impose inequality indices to be consistent with it. The idea is in itself quite appealing and resembles classical approaches to multidimensional inequality and concentration indices, but putting it into practice leads to quite complicated axiomatics and non-neat results, also depending on some arbitrary choices that lessen the generality of the arguments. It is our personal feeling that in applied statistics, axiomatic systems should be kept as simple and natural as possible. In this effort, one should be driven by a clear idea of what is to be measured, by suitable formal representations of the problem and by appropriate mathematical tools. While the concepts of inequality and polarization are quite clear, the formal tools usually employed, borrowed from classical mathematical analysis, are not so effective. A possible alternative is to cast the whole problem in partial order terms, representing (equivalence classes of) statistical units through Hasse diagrams, linking inequality/polarization axioms to the structure of the partial order and to the distribution of the population on it. Focusing on statistical units instead of variables (i.e., considering data matrix rows instead of columns) has many advantages: it (i) makes the structure of the data explicit, (ii) helps identify alternative properties that indices may fulfill and that may be turned into axioms, (iii) is completely consistent with the ordinal nature of the data, and (iv) generalizes also to posets not explicitly built upon a set of variables. Basically, the idea could be to build a system of axioms which is more algebraic in nature than analytical, requiring the indices to be "well behaved" both when the frequency distributions change and when the partial order structure changes. We are currently working on this, with promising results.

9.2.3 Searching for Patterns: Clustering over Posets and Lattices?

It is not unusual that social surveys collect ordinal information by asking respondents to score their judgments using scales with up to ten degrees. The resulting set of profiles (i.e., sequence of scores on the investigated dimensions) may comprise even thousands of different elements. The need to reduce complexity and identify groups and clusters of respondents naturally arises. Similarly, many economic studies comparing countries' features lead to systems of multidimensional comparisons on ordinal data; also in such cases one may be interested to group similar statistical units. The study of clustering techniques is one of the most developed branches of data analysis. Many different methodologies are available, ranging from simple hierarchical procedures (Rencher 2002), to neural networks algorithms (Kohonen 2001, Ripley 2005) or to model-based techniques (Vermunt and Magidson 2002). Most of the clustering tools are designed to work with cardinal variables, but there are also procedures for ordinal data. In practice, however, the partial ordinal nature of the data is seldom taken into account explicitly. Recently, an interesting book about clustering on ordinal data came out, where also generalizations of dissimilarity measures taking values in posets are considered (Janowitz 1978, 2010), but most of the techniques applied in daily research are of a classical kind and treat ordinal scores as cardinal (e.g., considering scores directly as numbers). The question we pose is whether it is possible to develop clustering methods that take full account of the partial order structure of the dataset, that is to build procedures which extract clustering information not only from some metric structure, but also from the underlying partial order. To be more explicit, let us consider the following simple problem. Suppose we record data on k ordinal variables, each on two-degree scales (e.g., pertaining to the ownership of k different goods). The set of 2^k k-dimensional profiles is naturally turned into a lattice L, under the product order. Suppose also that a frequency distribution is assigned on L. We now want to cluster individuals, i.e., profiles using some hierarchical procedure, based on the choice of a metric. If we perform this task in the usual way, at each step of the procedure we do obtain groups, but we lose information on the underlying lattice structure. In other words, we merge profiles into groups, but we do not know how to partially order them and we cannot embed them into a lattice structure. This is a critical problem, since the lattice structure of the profiles conveys a lot of information on the data. This information should be used when clustering and should be preserved, as much as possible, at each step of the procedure. A possible way to achieve this relies on the concept of lattice congruence. Any partition of the elements of L defines an equivalence relation on the lattice, which, however, may or may not be compatible with the join and meet operations defined on it, that is which may or may not be a congruence. Forcing, at any step of the clustering procedure, the obtained partition to be a congruence, the clustering process would naturally partially order the clusters, in a way compatible with the original lattice. A procedure might be designed where, at any step, (i) some profiles are merged (i.e., they are declared as equivalent) based on some metric (or dissimilarity) criterion and (ii) the smallest congruence comprising that equivalence is computed. In this way, the information comprised in the relational structure of L would be employed in the clustering process, making the local metric information spread across the lattice, through the congruence constraint. A trivial example is given in Fig. 2 (see Davey and Priestley 2002). When the selected elements of the lattice are merged in cluster "a," other clusters must be formed ("b," "c," "d," "e") for the partition to be consistent with the order relation. The resulting set of clusters is again a lattice, whose Hasse diagram is depicted in Fig. 9.2, with black nodes representing groups. The equivalence relation induced by the final partition is the smallest congruence comprising the equivalence class "a." Similar approaches might be also followed when dealing with posets which are not lattices, possibly drawing upon poset generalizations of the notion of a lattice congruence. We notice that also the choice of the metric might be made taking into account that profiles are partially ordered (Monjardet 1981) and this could improve the coherence and the effectiveness of clustering algorithms on partially ordered structures. Clustering procedures are not our own research field, so we limit ourselves to the above hints. However, we invite lattice experts to address this problem, the solution to which would be very useful to social scientists.

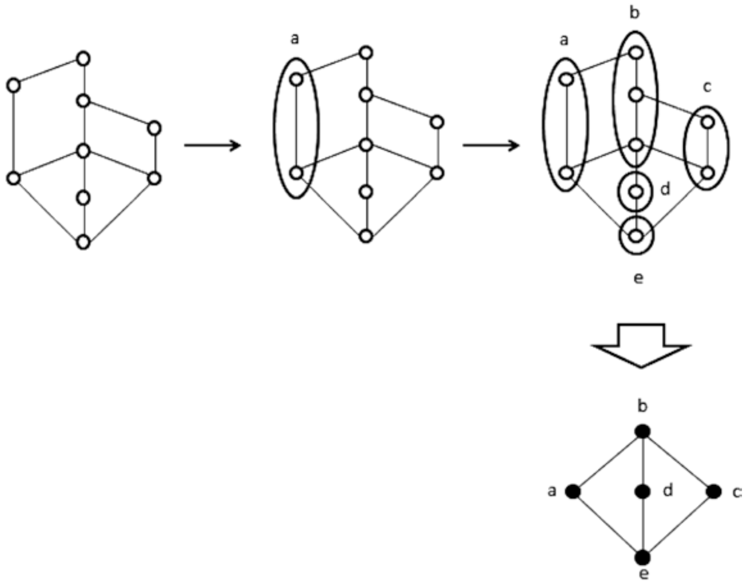


Fig. 9.2 An example of clustering on a small lattice

9.3 Conclusion

In this paper we have commented on the possible role of partial order theory in socio-economic analysis, from the (certainly narrow) point of view of our main research interests. Here we simply want to stress some of the key concepts addressed. The study of social facts is asking for new tools and new languages, more oriented to complexity and more capable of reproducing the reality, in modern societies, of "patterns," "shapes," and "nuances" which are relevant for policy-making. The issue of multidimensionality combined with the increasing availability of ordinal data is particularly challenging for socio-economic scientists, who need new tools to address social issues, but also tend to stick to "old" paradigms. So the problem is both technical, in that new statistical procedures must be developed, and cultural, in that some open-mindedness is necessary for scholars to modify, at least partly, their methodological habits. Partial order theory may play a key role in this challenge, as we have suggested through examples pertaining to real issues, crucial for our comprehension of societal dynamics and for policy definition. Proving that ordinal data may be effectively and consistently treated and exploited, partial order theory opens new possibilities to socio-economic statistics. Certainly, the technical and the cultural sides of the challenge go together. As concrete applications of partial order tools begin to prove their usefulness to social science, it will become easier for the wider scientific community to accept and employ them successfully. This challenge involves both partial order theorists and social scientists, since only by joining different points of view and complementary competencies it is possible to advance in this research field. We hope that this paper, raising questions and soliciting answers, may contribute to fruitful developments.

References

Alkire S, Foster J (2011) Counting and multidimensional poverty measurement. J Publ Econ 96(7–8):476–487

Alkire S, Foster J (2011) Understandings and misunderstandings of multidimensional poverty measurement. J Econ Inequal 9(2):289–314

Abul Naga R, Yalcin T (2008) Inequality measurement for ordered response health data. J Health Econ 27:1614–1625

Allison RA, Foster J (2004) Measuring health inequality using qualitative data. J Health Econ 23:505–524

Apouey B (2007) Measuring health polarization with self-assessed health data. Health Econ 16:875–894

Davey BA, Priestley HA (2002) Introduction to lattices and order. Cambridge University Press, New York

Doorslaer van E, Jones AM (2003) Inequalities in self-reported health: validation of a new approach to measurement. J Health Econ 22:61–87

Duclos J, Esteban J, Debraj R (2004) Polarization: concepts, measurement, estimation. Econometrica 72:1737–1772

Ercolani MG, Jenkins SP (1998) The polarisation of work and the distribution of income in Britain, Institute for Labour Research and ESRC Research Centre on Micro-Social Change University of Essex, UK

Fattore M, Brüggemann R, Owsiński J (2011) Using poset theory to compare fuzzy multidimensional material deprivation across regions. In: Ingrassia S, Rocci R, Vichi M (eds) New perspectives in statistical modeling and data analysis, Springer, Berlin, Heidelberg

Fattore M, Maggino F, Greselin F (2011) Socio-economic evaluation with ordinal variables: integrating counting and poset approaches. Statistica & Applicazioni, Special Issue, 31–42

Fattore M, Maggino F, Colombo E (2012) From composite indicators to partial order: evaluating socio-economic phenomena through ordinal data. In: Maggino F, Nuvolati G (eds) Quality of life in Italy: research and reflections. Social indicators research series, vol 48. Springer, the Netherlands

Fusco A, Silber J (2011) Ordinal Variables and the Measurement of Polarization. CEPS/INSTEAD Working Paper 2011–2033

Gigliarano C, Mosler K (2009) Constructing indices of multivariate polarization. J Econ Inequal 7:435–460

Gregg P, Wadsworth J (2004) Two sides to every story: measuring the polarisation of work. CEP discussion paper No 632, London School of Economics and Political Science

Janowitz MF (1978) An order theoretic model for cluster analysis. SIAM J Appl Math 34(1):55–72

Janowitz MF (2010) Ordinal and relational clustering. World Scientific, Singapore

Kobus M (2011) Multidimensional inequality indices for ordinal data. Warsaw University, Faculty of Economic Sciences (available at http://coin.wne.uw.edu.pl/mkobus/) accessed August 22, 2013

Kohonen T (2001) Self-organizing maps. Springer, Berlin

Lustig N (2011) Multidimensional indices of achievements and poverty: what do we gain and what do we lose? An introduction to JOEI Forum on multidimensional poverty. J Econ Inequal 9(2):227–234

Maasoumi E (1999) The measurement and decomposition of multi-dimensional inequality. Econometrica 54:771–779

Madden D (2010) Ordinal and cardinal measures of health inequality: an empirical comparison. Health Econ 19:243–250

Michailides G, de Leeuw J (1998) The Gifi System for Nonlinear Multivariate Analysis, Department of Statistics Papers, UCLA. Available at http://escholarship.org/uc/item/0789f7d3 accessed August 22, 2013

Monjardet B (1981) Metrics on partially ordered sets—A survey. Discrete Math 35(1–3): 173–184

Lones AM, Rice N, Robone S, Dias PS (2010) Inequality and polarisation in health systems' responsiveness: a cross-country analysis. HEDG Working Paper 10/27

Permanyer I (2012) The conceptualization and measurement of social polarization. J Econ Inequal 10:45–74

Ravaillon M (2011) On multidimensional indices of poverty. J Econ Inequal 9(2):235–248

Rencher AC (2002) Methods of multivariate analysis. Wiley, New York

Ripley DP (2005) Pattern recognition and neural networks. Cambridge University Press, Cambridge

Tsui K (1986) Multidimensional inequality and multidimensional generalized entropy measures: an axiomatic derivation. Soc Choice Welfare 16:145–157

Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Applied latent class analysis. Cambridge University Press, Cambridge, pp 89–106

Vittadini G (1989) Indeterminacy problems in the Lisrel model. Multivariate Behav Res 24(4):397–414

Vittadini G, Minotti SC, Fattore M, Lovaglio PG (2007) On the relationships among latent variables and residuals in PLS path modeling: the formative-reflective scheme. Comput Stat Data Anal 51:5828–5846

Wolfson MC (1994) When inequalities diverge. Am Econ Rev 84(5):353–358

Zhang X, Kanbur R (2001) What difference do polarisation measures make? An application to China. J Dev Stud 37(3):85–98

Part IV
Applications

Chapter 10
Ranking Hazardous Chemicals with a Heuristic Approach to Reduce Isolated Objects in Hasse Diagrams

Ghanima Al-Sharrah

Abstract This work identifies hazardous chemicals that cause chemical accidents in plants using simple and available properties. Particular attention is given to reactive chemicals and the relation between corrosion and accidents frequencies. The identification of hazardous chemicals is done by categorized ranking using the Hasse diagram technique. Hasse diagrams are the most promising method due to simplicity, wide use, and nonparametric advantage. To achieve our goal, the large number of isolated objects in Hasse diagrams was reduced/eliminated using a heuristic approach that presents strategies to speed up the ranking task. The basis is to collect suitable indicators and to define the suitable ranking objective. The ranking is presented using a case study of 22 chemicals with 8 simple hazardous indicators. Results show that the reduction of isolated objects is essential before evaluating the hazardous results. Also, simple and readily available data were used successfully as indicators for identifying chemicals causing accidents.

10.1 Introduction

Hazardous chemicals present health threats to workers in industrial and residential areas. Identifying, then controlling these chemicals is high up the agenda for all industries, but particularly for the chemical industry. The consumers, employees, stakeholders, legislators, and the communities for which the industry operates are all becoming increasingly aware of hazardous chemicals issues and demand ever-higher standards. Over the last few decades, the chemical industry has reduced its harmful emissions significantly, by using environmental and technological developments together with an increased awareness of the safety aspects of plant

G. Al-Sharrah (✉)
Department of Chemical Engineering, Kuwait University,
P.O. Box 5969, 13060 Safat, Kuwait
e-mail: g.sharrah@ku.edu.kw

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_10, © Springer Science+Business Media New York 2014

operation. The task of identifying and ranking chemicals is not easy to apply due to its huge tasks and components, number of chemicals, and hazardous indicators. However, identifying the chemicals that need special attention is essential to protect the people and the environment as far as possible from the dangers that can arise from an industrial plant.

The primary hazard in the chemical industry resides in the material, because materials are a hazard even if only in storage, with no processing or other activity being performed. The raw material, the intermediate, and the finished products present the primary independent hazard element (Ward 2002). The hazardous effect of chemicals comes through three ways: fire, explosion, and toxicity. Many flammable, explosive, and toxic properties were used as hazard indices to rank chemicals according to their hazardous effect on humans and the environment, i.e., accidents consequences or severity. Another side of hazardous effect of chemicals is their ability to cause accidents; highly reactive and corrosive chemicals are included in this side.

In a process, not only the substances but also the equipment or unit operations play an important role. Equipments and units represent an inherent hazard, secondary to material, because they act on the materials and cannot be the cause of problems without the materials and operation. Accident consequences due to equipment failure alone are mainly an economic loss and an in-plant problem, while the equipment failure accompanied by a chemical release is an off-plant problem, mainly for human health and environmental damage. This means that the existence of a chemical within the equipment increases the consequences of accidents.

10.2 Accident-Causing Chemicals

Chemical accidents have been a big concern for facilities that process, handle, transport, or store chemicals. The chemical and petrochemical industry is one of these industries that suffer from the continuous occurrence of these accidents. Even with the large number of studies, these incidents continue to occur. The ability to learn from previous incidents has long been regarded as an essential aspect of any program designed to reduce the frequency and severity of future incidents. Many major events, which capture media attention, continue to implicate “failure to learn from previous losses.” If obvious similarities are apparent between an existing operation and one that experienced loss, follow-up action is more likely to be pursued and future loss may be avoided. This indicates the importance of identifying “accident-causing chemicals” from their properties. From the chemical side, some studies give alerts to reactivity of chemicals when changes in chemical structures have the potential to generate heat, energy, and gaseous byproducts that cannot be safely absorbed by the immediate surrounding (Bretherick and Urben 1999).

Wei et al. (2004) and Liu et al. (2006) analyzed 167 reactive chemical incidents and concluded that information needed to identify hazards in order to prevent incidents already exist in the literature. A useful source of reactive chemical properties is the Chemical Engineering Thermodynamic and Hazard Evaluation CHETAH program developed by ASTM (2005). CHETAH calculates the value of six hazard evaluation criteria, as follows:

- Maximum heat of decomposition
- Fuel value minus heat of decomposition
- Oxygen balance
- CHETAH energy release evaluation criterion
- Overall energy release potential
- Net plosive density

Wei et al. (2004) used CHETAH to analyze reactive incidents reported in a study conducted by the U.S. Chemicals Safety and Hazard Investigation Board (2002).

Another simple and well-known hazard indices used when ranking involves chemicals are the National Fire Protection Association (NFPA) (1994) indices. NFPA has developed a system for indicating health, flammability, and reactivity hazards of chemicals. The system is based on giving a number (from 0 to 4) to a chemical indicating its effect. Al-Sharrah et al. (2001) used these NFPA health ratings as an index for an environmental objective in petrochemical planning.

Corrosion is another factor that supports the occurrence of equipment failure and thereafter, accidents and the release of chemicals. Corrosion may be defined as the process of unwanted attack on a metal by its environment. In practice, corrosion is an insidious process which is often difficult to recognize until deterioration is well advanced. Corrosion is one of the largest causes of plant and equipment breakdown in the process industries (Chandrasekaran 2010). For most applications it is possible to select materials of construction that are completely resistant to the attack by the process fluids, but the cost of such an approach is often prohibitive. In practice it is usual to select materials that corrode slowly at a known rate and to make an allowance for this in specifying the material thickness. However, a significant proportion of corrosion failures occur due to some form of localized corrosion, which results in failure in a much shorter time than would be expected from uniform wastage. Additionally, it is important to take into account that external atmospheric corrosion leads to many instances of loss of containment and tends to be a greater problem than internal corrosion. All these aspects of corrosive behavior need to be addressed both at plant design time and during the life of the plant.

Thomas (1981) analyzed data from pipe and pressure vessel failure and found that corrosion and erosion contribute to 24.6 % of total leaks. Balasubramanian and Louvar (2002) analyzed data from refinery accidents for a 40-year period from 1960 to 2000 and indicated that 38.1 % of these accidents are due to corrosion and stress. Therefore, corrosion must be considered in evaluating the safety aspect of industrial process. Its inclusion into the evaluation of hazard and safety indices reflects the fact that there exists a strong relation between accident frequency and a chemical's compatibility with the construction material. The Dow Fire & Explosion Index (1994) and Heikkila (1999) indices considered corrosion in evaluating safety indices. Dow gave a penalty for corrosion and erosion in the range of 0.1–0.75, and Heikkila gave a score of 0, 1, and 2 for the construction material carbon steel, stainless steel, and better material, respectively. Corrosion is considered as a part of reactivity in a reactive index proposed Gubta and Babu (1999). Also, Al-Sharrah et al. (2007) studied the corrosion resistance of some chemicals and found that the accident frequency decrease with increasing corrosion resistance.

The molecular structure, chemical, or physiochemical properties are major factors influencing the corrosion caused by chemicals on metal surfaces. The acceleration or inhibition of corrosion process results from an interaction of the metal with the chemical. Special attention has been given to molecular mass, conductivity, density, viscosity, dielectric constant, pH, adsorption–solvation effects, and dipolar moment, with a strong belief that none of the properties can serve as a sole parameter determining the corrosion of metals in all media (Ekilik and Grigor'ev 2002; Dewan et al. 2002). The literature on such approach applied to chemical properties are lacking unlike the opposite of this, many theoretical considerations are published for correlations between the molecular structure or properties of chemical compounds and their ability to inhibit corrosion [see for example Popova et al. (2007)].

10.3 Ranking and Heuristics for Isolated Objects

Chemical ranking and sorting is a tool for assessing chemicals by considering health, environmental or other hazards, and exposure. During the last decade there have been vast improvements in the methods used to rank chemicals and to interpret their data within a risk assessment framework. Risk-base ranking and scoring methods can be used to focus attention and resources on the largest potential hazard. One useful method for ranking chemicals is the Hasse diagram; it is a visual representation of partially ordered sets. It ranks objects given a number of indicators that present their performance. Details of this method are presented elsewhere (see Halfon and Reggiani 1986) and its result is to rank chemicals (objects) into categories with different importance levels. The Hasse diagram has been used in many chemical ranking needs; examples are ranking high production volume chemicals (Lerche et al. 2002), pesticides (Halfon et al. 1996), and chemicals for environmental hazard (Halfon and Reggiani 1986).

The advantage of the Hasse diagram is the simple mathematical background, being nonparametric and flexible enough to be adapted for most purposes which encourage its continuous use for chemical ranking. However, one of its disadvantages is the existence of isolated object; isolated objects are common in Hasse diagrams. This isolation indicates an extreme level of contradiction in object's properties when they are compared with other objects, so they appear as isolated circles in the Hasse diagram and can freely be located from maximal to minimal position. The existence of isolated objects in the Hasse categorized ranking diagram is sometimes misleading or it can direct the attention and resources into the wrong direction, i.e., nonhazardous chemicals. The existence of isolated objects is not considered as a problem if their number is low or if the decision makers can definably decide whether these isolated objects be located conventionally as maximal or

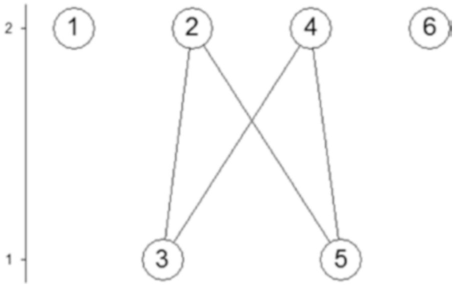


Fig. 10.1 Hasse diagram for the data in Table 10.1

Table 10.1 Sample data for chemical ranking, six chemicals with their four hazard indicators

	Hazard indicator			
Chemical no.	1	2	3	4
1	6	1	1	0
2	4	2	5	1
3	3	2	0	1
4	5	3	4	2
5	4	1	4	1
6	5	1	2	5

minimal elements. Figure 10.1 shows an example of a Hasse diagram for the data in Table 10.1; chemicals 1 and 6 are isolated objects. The rank indicated from the Hasse diagram is that chemicals 1, 2, 4, and 6 are in the top rank and chemicals 3 and 5 in the bottom rank. Objects 1 and 6 can be drawn at the bottom of the diagram resulting into other ranking results. Top rank means most hazardous chemicals and bottom rank means low hazardous chemicals. Note that, actual Hasse diagrams may have multiple levels.

The problem of the existence of isolation has been addressed, in general, in the form of indicator reduction for improving comparability (Voigt et al. 2010) and lattice theory (Brüggemann et al. 1997). However, no clear or simple solution for this problem has been studied although it appears in many Hasse diagrams. It is important to note now that the isolation exists only in Hasse diagrams and if it is desired to overcome this problem in a strict manner, one can use any total ranking method which can give an exact rank to each object (chemical); examples of these methods are Hasse average rank (Brüggemann et al. 2004), Copeland method (Al-Sharrah 2010), and Simple Additive Ranking (SAR). The last method is simply ranking the objects with respect to each indicator separately then subsequent aggregation of the weighted ranks by arithmetic mean and, finally, the normalization of the obtained values.

A simple question then arises: “If there is a feeling that isolated objects will appear in the Hasse diagram, why not use any total ranking method that may give practical results from the beginning?” The author hopes that this chapter gives a positive answer to the question; using the Hasse diagram and solving the problem of isolated objects if any occur will achieve the following:

- Satisfy the requirement of categorized ranking which is more suitable for chemicals. Usually it is desired to classify chemicals as high hazardous, medium hazardous, and so on.
- Not exclude the possibility of using any total ranking methods that can strengthen the results.
- Avoid the problem of combining the indicators at the beginning of the ranking task which usually needs decision maker's input. Combining indicators by total raking may be used in a limited manner at the final stages of ranking.
- Solving the problem of isolated objects, using heuristics, will give insight to any weakness in the indicators used or ranking objective.

Isolation is contradiction; therefore it should be reduced, if possible, by gaining knowledge on objects and indicators. We propose the following heuristics to aid in the knowledge and thereafter reducing the number of isolated objects and any confusion they can cause:

- Rank according to hazardous not safe properties.
- Select suitable indicators, effective and evenly comparable.
- Divide your data in a suitable manner for a specific rank objective.
- Use total ranking to relocate the isolated objects only (you don't want to re-rank).
- Compare ranking results with existing knowledge about the objects to see if any contradiction exists especially for the relocated isolated objects.

The above heuristics are explained below.

Isolated objects can be seen as maximal and minimal elements at once but according to the caution principle they are located in the diagram within those objects that require priority attention (most often they are located at the top of the diagram). Therefore, it is important to adjust indicators to represent the hazardous, or in general undesired, effect. In this case, if any isolated objects exist they will be located with those objects that need attention. If the opposite was done, i.e., the indicators were adjusted to reflect safe features, isolated objects would be located with safe objects (at the top of the diagram), and the results would have some negligence.

The number of indicators also affects the existence of isolated objects. The higher the number of indicators, the higher the possibility of contradiction. The number of indicators can be reduced by excluding weak indicators. This includes very old, very general, or indicators with limited applicability. Therefore, it is advantageous to make all indicators comparable in strength to represent the desired rank objective.

Reducing the number of indicators can be done also by specifying and limiting the rank objective. For example, the term hazardous is very wide when considered as a rank objective and sometimes vague. In the chemical side, the term hazardous has many representations such as toxicological behavior, chronic health, exposure potential, etc. A total hazard value is always an ideal objective for assessing chemicals; however, this can increase the number of indicators and hence isolated objects. Total hazard is best handled using total ranking methods that have aggregation such as CHEMSI (Swanson et al. 1997) and Indiana Relative Chemical Hazard Score (IRCHS). IRCHS indicates how a chemical compares with others in terms of its capacity to impact human health, ecosystems, or environmental health generally. Hazard values have been assigned to over one thousand chemicals. The hazard values are on Clean Manufacturing Technology Institute CMTI's Web site, http://www.ecn.purdue.edu/CMTI.

Although, the proposed heuristics seem to be very basic and obvious, it is found that they help to explain the existence of isolated objects in some Hasse diagrams found in literature. For example, Patil and Taillie (2004) ranked ten Latin American countries using human–environmental indicators with values between 0 and 1; large values representing "better conditions." The resulted Hasse diagram showed Peru, Uruguay, and Venezuela as isolated objects on the top level; this may have indicated good environmental conditions for inexpert readers. However, other studies indicate that the capital of Peru has a Green City Index of well below average and the capital of Uruguay of below average (Economist Intelligence Unit 2010). An example on the problem of not selecting comparable indicators is the data used by Brüggemann and Carlsen (2011); their work was to rank 13 chemicals using 7 health effects indicators; the resulted Hasse diagram has 5 isolated objects. A close look at the indicators reveals that six indicators were organ specific and one was an Ames test (a test to estimate the carcinogenic potential of a chemical). If the Ames test indicator was removed, the isolated objects will be reduced to 2. Selecting the rank objective is also an important issue in reducing the number of isolated objects. Voigt et al. (2000) ranked environmental online and CD-ROM databases and used a rank objective of "quality." Due to the wide definition and components of the rank objective, the resulted Hasse diagram has a high level of isolation.

By using the above-mentioned five heuristics, it is expected that the number of isolated objects will be reduced and any existing isolated object will be in the higher level of attention. If the number of isolated objects is still high, then further reduction is required. One way is to apply one of the total ranking methods as follows: (a) calculate a total rank for all objects, (b) compare the rank of each isolated object with the average rank of the different levels, and (c) relocate the isolated objects. Applying the above to the examples in Table 10.1 and Fig. 10.1 will start by calculating a total rank using any desired method. SAR is suitable for the relocation and details of this method are shown in Table 10.2 and Eqs. (10.1) and (10.2). As mentioned earlier, SAR method is based on the ranking of the objects with respect to each criterion separately, and the subsequent aggregation of the weighted ranks by arithmetic mean. To explain in an equation form, if all the rankings r_{ij} of the i-th object for the j-th indicator are calculated, the SAR rank is calculated as

$$\overline{SAR_i} = \frac{\sum_{j=1}^{P} w_j \cdot r_{ij}}{n}, \qquad (10.1)$$

Table 10.2 SAR for the data in Table 10.1

	Rank with respect to indicator				
Chemical no.	1	2	3	4	SAR
1	6	1	2	1	0.300
2	2.5	4.5	6	3	0.600
3	1	4.5	1	3	0.275
4	4.5	6	4.5	5	0.800
5	2.5	1	4.5	3	0.350
6	4.5	1	3	6	0.525

where n is the total number of objects, P is the total number of indicators, and w_i is the weight for indicator. In our case, indicators have the same importance, therefore, equal weights of 0.25 were used. Then the rank is normalized as:

$$SAR_i = \frac{\overline{SAR_i} - 1/n}{1 - 1/n}. \tag{10.2}$$

Normalization is not necessary to obtain ranks of objects, but it gives a more concise rank range.

The ranks using SAR are 0.3, 0.6, 0.275, 0.8, 0.35, and 0.525 for chemicals 1–6, respectively. Based on SAR the average total rank for level two chemicals in the Hasse diagram (counting from the bottom, excluding isolated object) AV_2 is 0.7 and level one AV_1 is 0.313. Chemical 6 has an absolute relative difference with AV_2:

$$DeltaAV_2 = \left|\frac{0.7 - 0.525}{0.525}\right| \times 100 = 33.3\%$$

and with AV_1:

$$DeltaAV_1 = \left|\frac{0.313 - 0.525}{0.525}\right| \times 100 = 40.4\%.$$

Therefore, chemical 6 can be kept at the top level 2. For chemical 1, the differences were 133.3 % with AV_2 and 4.3 % with AV_1 and therefore for this chemical it is more appropriate to transfer it to level one. This relocation of isolated objects is not strict and is optional but it helps to reduce the number of isolated objects in the higher level, i.e., limit the number of objects that need more attention and strict regulations. The last confirmation for the relocation is using existing knowledge, if available, about the nature and effect of these chemicals (objects) so relocation can be accepted.

10.4 Data and Results

Our aim is to rank chemicals for their ability to cause accidents using simple physiochemical properties and reactivity indicators. The ranking is done using the Hasse diagram which is plotted using DART (2009); DART is a package developed by Talete srl and it is made freely available as a service to scientific researchers from the Institute of Health and Consumer Protection Web site. Also, ranking will be accompanied with heuristics to speed up the process of finding a satisfactory solution and to tackle the problem of isolated objects in Hasse diagrams.

Identification of reactive hazards is an important part of any chemical plant to reduce the risk of chemical accidents. Reactive hazards are complex because they may involve external process conditions, however, the most important and more easily evaluated is the thermophysical/physical properties of these chemicals that cause reactivity and hence hazards. The same can be said for corrosive hazard chemicals. From the discussion earlier, the following properties will be selected to rank a number of chemicals for their reactivity and ability to cause accidents:

1. Maximum heat of decomposition
2. Overall energy release potential
3. Oxygen balance
4. NFPA reactivity index
5. Density
6. Viscosity
7. Molecular weight (MW)
8. Compatibility with process materials

The maximum heat of decomposition, overall energy release potential, and oxygen demand were taken from CHETAH computer program (ASTM 2005); the NFPA (1994) reactivity index is an integer index from 4 to 0 to represent the reactivity of a chemical ranging from explosive material at room temperature to nonreactive. The density, viscosity, and molecular weight were taken from the Hyprotech-HYSYS 3.1 software. Compatibility with process material is a number estimated from the compatibility of the chemical with the mostly used process material, i.e., carbon steel, 304-stainless steel, and 316-stainless steel. Compatibility tables usually indicate compatibility as excellent, good, fair, or severe. If these qualitative compatibilities were transferred to quantitative values as 4, 3, 2, and 1, respectively, this will give an overall number for compatibility. For example, sulfur dioxide has a compatibility of good for both carbon steel and 304-stainless steel and severe with 316-stainless steel; therefore, its overall compatibility will be 7. A good source for compatibility is FMC technologies compatibility manual (FCM 1996).

The relation between hazardous effect and the above properties is inversely proportional except for viscosity. The data should be adjusted to be able to present high value for high hazard and this can be done either by multiplying by -1 or taking the reciprocal; both ways give the same result but the former method is preferred if some indicators have zero values. The data used for this case study are presented in

Table 10.3 with 22 chemical and 8 indicators. Ranking will go through stages until all the isolation problems in the Hasse diagram are resolved. Note that the chemicals will be the objects and the chemical properties will be indicators.

Stage 1: At this stage, all indicators were taken and the Hasse diagram for this case is shown in Fig. 10.2. The figure clearly shows that 21 chemicals are isolated objects with no ranking information detected from the Hasse diagram; only objects 7 and 6 are comparable. The indicators have to be rechecked to delete weak indicators.

In the CHETAH manual, the oxygen balance indicator is attributed to Lothrop and Handrick (1949), the work of which reported that oxygen balance was recognized as a standard parameter in explosive design. Lothrop and Handrick (1949) demonstrated a strong correlation between effective oxygen balance and explosive performance of carbon, hydrogen, nitrogen, and oxygen (CHNO) explosives. They neither recommended nor mentioned the use of oxygen balance as a hazard evaluation criterion. The ranges of oxygen balance referred to in CHETAH are not taken from Lothrop and Handrick. Therefore, oxygen balance as calculated in CHETAH is an unsatisfactory measure of reactivity hazard. Consequently, the indicator of oxygen balance will be removed.

U.S. Chemical Safety and Hazard Investigation Board (2002) analyzed 167 serious incidents in the USA involving uncontrolled chemical reactivity from January 1980 to June 2001. Approximately 60 % of the 167 incidents involved chemicals that either not rated by NFPA or have “no special hazard” (NFPA reactivity index of 0). Only 10 % of the 167 incidents involved chemicals with NFPA published ratings of 3 or 4. Moreover, NFPA Standard covers only 325 chemical substances, a very small percentage of the chemicals used in industry and ratings were established by a system that relies, in part, on subjective criteria and judgment. Therefore, NFPA reactivity index is insufficient as a basis for determining reactive hazard.

Stage 2: At this stage, all indicators were used except oxygen balance and the NFPA reactivity index. The Hasse diagram is shown in Fig. 10.3 with the number of isolated objects reduced from 21 to 1 (which is chemical 16). This indicates the importance of carefully selecting indicators for ranking. Now the diagram has three level and most of the objects are on the high level with one isolated object.

Although the number of isolated objects is acceptable, applying more rank heuristics will improve the results and clarify more. Now, the rank objective needs to be checked, i.e., to define the exact meaning of hazardous chemicals. If the objective is the accident side of chemicals, then we have to see how well do people process information related to the number of accidents and the level of risk? For example, a decision maker might be told how many accidents occurred in a given chemical plant; when determining the accident frequency, the number of accidents will be the numerator and the size of plant will be the denominator. Do people combine this information in a reliable manner in making judgment about the recklessness of particular activities? Our hypothesis is that people are much more responsive to information about the numerator, or the total number of accidents than they are to the denominator, which is the measure of the total level of the particulate

Table 10.3 Data used for ranking chemicals

No	Chemical	Max. heat of decomposition (kJ/g)	Energy release potential (kJ/g)	O_2 balance (g O_2/g)	NFPA reactivity	Density (kg/m^3)	Viscosity (cp)	MW	Compatibility
1	Ammonia	0	0.197	−140.918	0	0.68699	0.008348	17.03	12
2	Chlorine	0	0.032	0	0	2.8604	0.013699	70.91	8
3	Hydrogen fluoride	0	0.032	0	1	0.80705	0.0051614	20.01	3
4	Chlorine dioxide	−0.371	−0.754	0	4	2.757	0.0118	67.45	9
5	Propane	−0.242	0.563	−362.827	0	1.7967	0.008175	44.1	12
6	Sulfur dioxide	0	0.032	0	0	2.6102	0.011821	64.06	7
7	Hydrogen chloride	0	0.032	0	1	1.4708	0.013591	36.46	3
8	Hydrogen	0	0.821	−793.668	0	0.081326	0.0087357	2.02	6
9	Methane	0	0.523	−398.919	0	0.64717	0.011274	16.04	11
10	Butane	−0.251	0.561	−357.846	0	2.3447	0.0074459	58.12	11
11	Ethylene oxide	−1.229	−0.234	−181.592	3	1.7949	0.008495	44.04	11
12	Hydrogen sulfide	0	0.228	−140.832	0	1.3746	0.011965	34.08	9
13	Formaldehyde	−1.002	−0.404	−106.569	0	1.2112	0.0077287	30.03	10
14	Isobutane	−0.216	0.548	−357.846	0	2.3447	0.0074459	58.12	12
15	Pentane	−0.564	0.649	−342.187	0	620.71	0.22043	72.15	9
16	Titanium tetrachloride	0	0.064	−16.869	2	1985.7	1.861	189.71	3
17	Phosgene	−0.063	0.079	−16.175	1	3.9903	0.011284	98.92	6
18	Nitric acid	0	0.185	63.477	0	23.855	0.8296	63.01	9
19	Ethane	−0.229	0.569	−372.454	0	1.231	0.0094066	30.07	12
20	Ethylene	−1.084	−0.086	−342.187	2	1.1317	0.010178	28.05	12
21	Vinyl chloride	−0.605	−0.515	−127.998	2	2.5209	0.0094694	62.49	12
22	Trichlorosilane	−0.015	0.06	−23.624	2	1350.9	0.33788	135.45	12

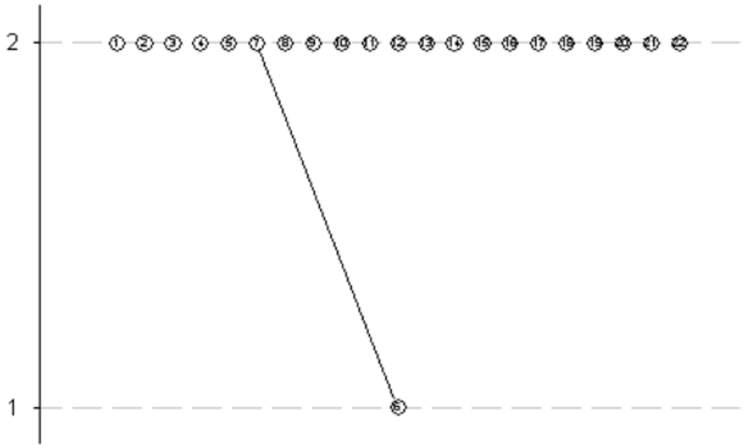


Fig. 10.2 Hasse diagram with all indicators

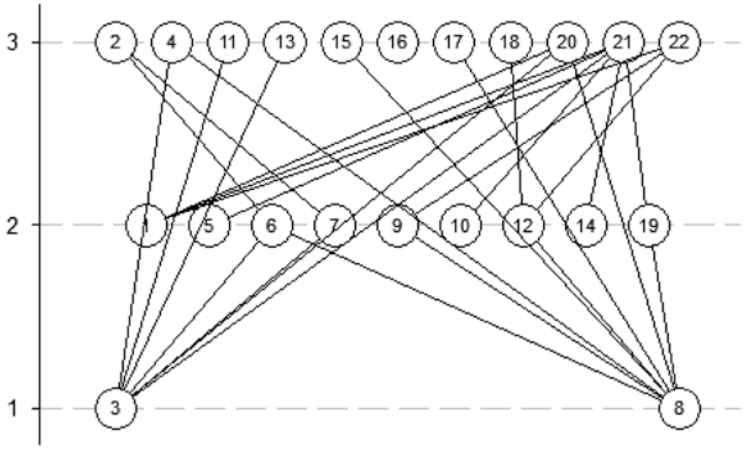


Fig. 10.3 Hasse diagram without oxygen demand and NFPA reactivity

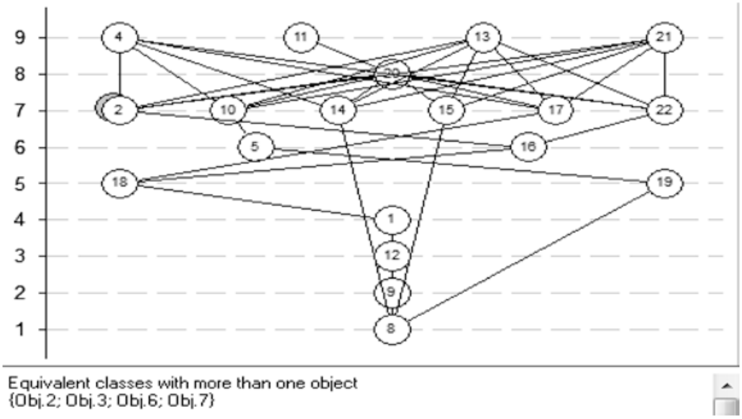
economic activity. If so, then large-scale plants would be at a disadvantage in term of the public perception of their safety performance when compared to smaller scale plants. The denominator blindness bias does not appear to have been fully explored in the literature. However, it has been identified by Viscusi and Zeckhauser (2004) and presented in the accidents analysis of Belke (2000).

Incidents rates are commonly normalized by dividing the number of incidents by some measure of the number of opportunities for an accident to occur. For example the U.S. Department of labour calculated occupational injury and illness rates by dividing the number of occupational injuries at a facility by the total number of person–hour worked at the facility over a given period. This allows large and small facilities to be fairly compared, assuming that all other safety issues being equal. The overall number of occupational injuries at a workplace over a given time period will generally be directly proportional to the number of employees working there. Likewise, this study should use normalization. However, since hazardous chemicals facilities vary so great in size, number of processes, chemical quantities stored and produced, operating schedule, and other characteristics, it is difficult to say which single divisor best represents the number of accidents opportunities over the full spectrum of facilities. This study uses the number of processes and aggregate chemical quantity as normalization factors. In choosing these factors, the assumption implied is that, a chemical contained in a large number of processes or in large quantities has more opportunities to be accidently released than does a chemical contained in fewer processes or smaller quantities. While these divisors are certainly not perfect, they appear to be reasonable (Belke 2000).

Our objective is to rank the chemicals in Table 10.3 for their ability to cause accidents, i.e., accidents frequency. Accidents frequency is different depending on the denominator, and as discussed previously a possible choice was, the number of processes and aggregate chemical quantity. When the frequency of chemical accidents are presented in number of accidents per number of processes, this means that the basic bulk properties of chemicals affect these accidents such as density, viscosity, molecular weight, and compatibility with process materials. The principle is that the chemical environment within the equipment controls the severity of corrosion and hence increases the failure rate of a process. Other properties, i.e., maximum heat of decomposition and overall energy release potential are related to the reactivity of the chemical and the existence of high quantities in a chemical plant. High quantities will release higher energy and then violent self-reaction and explosion. This classification will direct us to the next stage.

Stage 3: The data will be divided for the two objectives in ranking; the first dataset is 22 chemicals with the first two indicators in Table 10.3, namely, Maximum Heat of Decomposition and Energy Release Potential. The objective is high hazard presented by high accidents per mass of chemical existing in the process which we will call "*extrinsic frequency*." The second dataset is 22 chemicals with the last four indicators namely, compatibility, molecular weight, viscosity, and density, with the objective of high accidents per process containing these chemicals which we will call "*intrinsic frequency.*" Results of the Hasse diagrams are shown in Figs. 10.4 and 10.5. The ranking for extrinsic frequency has no isolated objects and the chemicals are ranked into nine levels. For the intrinsic frequency, the Hasse diagram has five levels and three isolated objects; the isolated objects can be relocated as shown in stage 4.

Rank of some chemicals can significantly change when changing the rank objective; see, for example, objects 8 and 21 which where hydrogen and vinyl chloride, respectively. In Fig. 10.4, vinyl chloride is at the top and hydrogen at the bottom while in Fig. 10.5, hydrogen is at the top and vinyl chloride is with the lower objects. This exactly corresponds to the properties of these chemicals; hydrogen is a nonreactive chemical, however, it is highly corrosive. While vinyl chloride is a reactive chemical that forms explosive polymeric peroxides, however, it is noncorrosive unless moisture is present.



Equivalent classes with more than one object
{Obj.2; Obj.3; Obj.6; Obj.7}

Fig. 10.4 Hasse diagram for extrinsic accident frequency

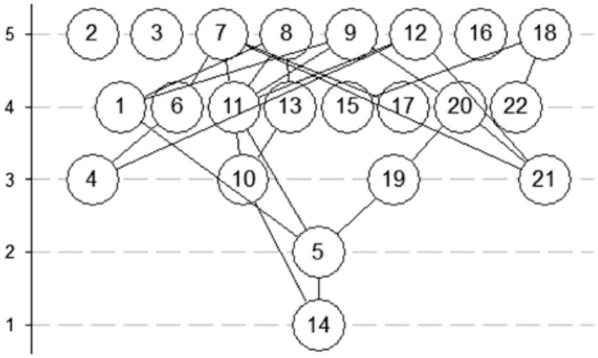


Fig. 10.5 Hasse diagram for intrinsic accident frequency

Stage 4: Isolated objects in Fig. 10.5 can be relocated into the level that has the smallest difference from their indicators. First, any simple total ranking method can be used like SAR or Copeland method (Al-Sharrah 2010) to obtain total ranking of the objects. Ranks obtained for the isolated object and ranks of the other objects in different levels are compared for the relocation. Using the SAR method, the total ranks shown in Table 10.4 are obtained.

The ranks for the isolated object are underlined and the average rank for the chemicals in levels 5, 4, 3, 2, and 1 are $AV_5=0.686$, $AV_4=0.463$, $AV_3=0.387$, $AV_2=0.345$, and $AV_1=0.333$, respectively. Note that the average rank of levels might not always be in a descending order as above; this is because the SAR ranks objects in a different way than the Hasse diagram. Table 10.5 shows absolute rank difference between the isolated objects and the different levels DeltaAV.

Looking at Table 10.4, it is clear that isolated objects have different total ranks and hence can be relocated to the level with the lowest absolute difference from

Table 10.4 SAR ranking; chemical numbers are referred to Table 10.3

Chemical no.	1	2	3	4	5	6	7	8	9	10	11
SAR rank	0.571	0.488	0.679	0.417	0.345	0.512	0.738	0.786	0.714	0.321	0.440
Chemical no.	12	13	14	15	16	17	18	19	20	21	22
SAR rank	0.667	0.536	0.333	0.417	0.476	0.429	0.524	0.464	0.548	0.345	0.250

Table 10.5 Comparison between isolated objects ranks and the levels average rank

Isolated object	Rank	DeltaAV$_5$	DeltaAV$_4$	DeltaAV$_3$	DeltaAV$_2$	DeltaAV$_1$	New level	Intrinsic accident frequency
2	0.488	40.574	5.123	20.697	29.303	31.762	4	0.022
3	0.679	1.031	31.811	43.004	49.190	50.957	5	0.064
16	0.476	44.118	2.731	18.697	27.521	30.042	4	0.056

their ranks as shown in Table 10.5. Note that if an isolated object has the same minimum difference with two levels, then the highest level should be selected to be more conservative. It is important to be aware that some total ranking methods are not suitable for this stage of relocation like the Desirability (Harrington 1965) and the Hasse average ranking method (Brüggemann et al. 2004). The approach of the desirability function is to define a function for each indicator in order to transform values *y* of the indicators to some scale:

$$d_j = f_j(y) \quad 0 \le d_j \le 1 \quad j = 1,2,\cdots,P.$$

The overall desirability is calculated by combining all the desirability through a geometrical mean. If an object is poor (very low) with respect to one indicator, its overall desirability will be poor. If any desirability d_j is equal to zero, then the overall desirability will be zero. This property makes it very probable that isolated objects will have zero rank with desirability due to the high contradiction in their indicators, i.e., some are very high and some are very low. On the other hand, the Hasse average rank gives that same rank to all isolated objects since it depends on the total number of objects, the number of objects ranked below the isolated object, and the number of objects incomparable with the isolated object.

The final and the most important stage is to check the appropriateness of the partial ranking presented by the Hasse diagrams and the relocation of the isolated objects; this is done in stage 5.

Stage 5: The Hasse diagrams presented previously show hazardous chemicals in the field of chemical accidents. Belke (2000) wrote a very useful analysis from the RMP*info database; it was a preliminary characterization of the database. Federal law requires industrial facilities that use large amounts of extremely hazardous substances to file a Risk Management Plan (RMP) with the Environmental Protection Agency (EPA). EPA does not release to the public all the data in the RMP*info database; these data can only be accessed through a federal reading room.

Belke (2000) used the number of processes in the facility and the aggregate chemical quantity as normalization factors for accidents rate. The data in Belke (2000) can be used to check the appropriateness of the ranking from the Hasse diagrams shown in Figs. 10.4 and 10.5 and the relocation of the isolated objects. The most important issue is to check the ability of the Hasse diagram created from the proposed set of indicators and the heuristics to identify hazardous chemicals, i.e., objects in the highest level should be the chemicals that have high accident frequency. Starting with "*extrinsic*" accidents frequency; the chemicals which cause more accidents related to their quantity in processing are the top objects (take the top two levels). In Fig. 10.4, the top chemicals are chlorine dioxide, ethylene oxide, formaldehyde, ethylene, and vinyl chloride.

Chlorine dioxide is a reactive, unstable gas; at above 10 % in air it can decompose spontaneously with a corresponding pressure pulse that may be more violent and explosive at higher partial pressure. For more than 25 % in air, explosion relief may be inadequate and rupture of the vessel may occur. These explosions can ignite combustible material. When chlorine dioxide gas is heated to decomposition, chlorine gas is produced. Chlorine gas creates hydrochloric acid when mixed with water or steam. The actual safety performance of this gas in the industry is poor with high accidents frequency 1.97 accident/Mlbs stored per year, which is more than 100 times the median. Chlorine dioxide is produced at typically 10–40 tons/day and with storage of 1 month of production; this will result in 0.66 Mlbs stored in the plant, therefore, a minimum of 1.3 accidents/year can occur. The thermophysical properties coincide with the accidents performance of these chemicals in the industry indicating that it is inherently hazardous chemicals and the industry process safety regulations need to be modified to better cover the reactive hazards of this chemical.

For ethylene oxide the C–C bond is short and the bond angles strained, therefore, it can be opened easily. The highly reactive nature of ethylene oxide helps to make it a versatile and commercially important chemical intermediate. However, serious incidents have occurred during its processing, storage, and transportation. Moreover, ethylene oxide has a tendency to polymerize, a number of sources have observed thermally driven polymerization initiated at different temperatures. The usual catalysts for ethylene oxide reactions, such as strong alkali, iron oxide (rust), and other metal oxides accelerate the reaction (American Chemistry Council 2007). When catalyzed by rust at ambient temperature, polymerization can create blockages or plugs in operations, such as plug of lines, relief valve inlets, and instrumentation taps. The actual accidents performance in the industry is moderate hazard of 0.045 accident/Mlbs stored per year, twice the median. This indicates a good application of safety guards and protection when operating with this chemical. Similar performance is observed with the ethylene, formaldehyde, and vinyl chloride that easily polymerize.

For the "*intrinsic*" frequency, the most hazardous chemicals were hydrogen fluoride, hydrogen chloride, hydrogen, methane, hydrogen sulfide, and nitric acid. All of these chemicals, except methane, are corrosive and as mentioned earlier, corrosion is the largest single cause of plant and equipment breakdown in the process industries. And since its control is a huge task, corrosive chemical accidents will occur in a high rate. For methane, the physical properties of liquefied natural gas (LNG) such as its cryogenic temperature, flammability, and vapor dispersion characteristics, add additional concerns of potential safety issues. Therefore, continuous monitoring and implementation of appropriate actions are essential to prevent, control, and mitigate unfavorable consequences of LNG production and use (Rathnayaka et al. 2012). This shows that simple indicators can identify other hazardous chemicals other than corrosive chemicals.

Next, several chemicals have notably high normalized accident rates relative to other chemicals listed in Belke (2000). The most obvious example is phosgene, which has an extrinsic frequency 138 times the median and it was not on the top level of the Hasse diagram of Fig. 10.4, i.e., not identified as very hazardous chemical. The reason for this result is that phosgene itself is highly toxic but not very reactive chemical; however, its ability to react violently with water makes it very susceptible to hazard conditions since water contamination is highly probable in plants. This introduces water reactivity as another reactivity index for future work.

To study the appropriateness of the relocation of isolated objects in Table 10.5, we need to compare the results with existing industry performance. The isolated objects were chlorine, hydrogen fluoride, and titanium tetrachloride. Hydrogen fluoride was kept at the top level and not relocated, while the others were lowered to the fourth level. This coincides with their accidents frequencies found in Belke (2000) where hydrogen fluoride (objects 3) has a higher accident frequency than both chlorine and titanium tetrachloride as listed in Table 10.5.

10.5 Conclusion and Outlook

Careful reporting and classification of data concerning chemical accidents, will greatly improve the quality of retrospective studies of chemical accidents. Therefore, our capability of analyzing data presenting properties of the chemicals with the help of Hasse diagram technique was successful of identifying hazardous chemicals and will transmit learning lessons from past accidents to predict future ones.

Obtaining simple and available indicators in chemical ranking will facilitate hazard evaluation in all level of planning stages and for all personal levels. The data used as indicators for ranking can be used as criteria for raw material selection during process screening and process concept development. Selection of safer, stable, and compatible raw material can eliminate or reduce the overall risk of the chemical process plant operation. Added to that is the advantage of Hasse diagram being a nonparametric method that requires basic mathematical knowledge and zero decision maker input.

The heuristics used were able to reduce the number and/or ambiguity of isolated objects in ranking chemicals. These heuristics or strategies are applied according to the needs of decision makers to speed up the process of finding satisfactory results. The results show that reactive and corrosive chemicals have to be identified in chemical plants as soon as possible and taken into consideration in design and material selection. Also, any hazardous chemicals not identified by Hasse diagrams directed us to an additional indicator that should be included.

Future research direction include the application of the heuristics on a wider range of datasets and to study their efficiency in reducing the number of isolated objects relative to the data size, i.e., are these heuristics useful for systems with large number of objects and/or large number of indicator? The results should also be compared with the sensitivity analysis provided by PyHasse program which specifies the indicators having the highest impact on the order relations (for PyHasse software, see several other chapters in this book). Another interesting field to extent this work is to study the best total ordering method for the last stage of relocating the isolated objects; SAR is criticized from methodological points of view and other methods may be more appropriate.

References

Al-Sharrah G (2010) Ranking using the Copeland score: a comparison with the Hasse diagram. J Chem Inf Model 50:785–791

Al-Sharrah G, Alatiqi I, Elkamel A, Alper E (2001) Planning an integrated petrochemical industry with an environmental objective. Ind Eng Chem Res 40:2103–2111

Al-Sharrah G, Edwards D, Hankinson G (2007) A new safety risk index for use in petrochemical planning. Trans IChemE Part B Proc Saf Environ Prot 85(B6):533–540

American Chemistry Council's Ethylene oxide/Ethylene glycols Panel (2007) The ethylene oxide product stewardship guidance manual, 3rd edn. http://www.sunocochemicals.com/hes/tech_manuals_EthyleneOxide.pdf. Accessed 22 Jan 2012

ASTM (2005) Stock # DS51E. ASTM computer program for chemical thermodynamic and energy release evaluation- CHETAH. Version 8.0, ASTM, Philadelphia, PA. ISBN: 0 8031 3366 9

Balasubramanian SG, Louvar JF (2002) Study of major accidents and lessons learned. Proc Saf Prog 21(3):237–244

Belke JC (2000) Chemical accident risks in U.S. industry – a preliminary analysis of accident risk data from U.S. hazardous chemical facilities. United States Environmental Protection Agency. Chemical emergency preparedness and prevention Office. http://www.epa.gov/ceppo/pubs/stockholmpaper.pdf. Accessed 20 Jan 2012

Bretherick L, Urben PG (1999) Bretherick's handbook of reactive chemical hazards, 6th edn. Butterworth-Heinemann, Jordan Hill, Oxford

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partial orders. Match Comm Math Comput Chem 65:383–414

Brüggemann R, Voigt K, Steinberg C (1997) Application of formal concept analysis to evaluate environmental databases. Chemosphere 35:479–486

Brüggemann R, Sørensen P, Lerche D, Carlsen L (2004) Estimation of average ranks by a local partial order model. J Chem Inf Comput Sci 44:618–625

Chandrasekaran V (2010) Rubber as a construction material for corrosion protection: a comprehensive guide for process equipment designers. Wiley, Hoboken, NJ, p 46

DART (2009) Version 2.05. Institute of Health and Consumer Protection, European commission joint research center. http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=DART. Accessed 11 Apr 2012

Dewan AK, Valenzuela D, Dubey S, Dewan A (2002) Corrosion at metal interfaces – a study of corrosion rate and solution properties, including electrical conductance, viscosity and density. Ind Eng Chem Res 41:914–921

Dow Chemical Company, Fire & Explosion Index (1994) Hazard classification guide, 7th edn. American Institute of Chemical Engineers, New York, NY

Economist Intelligence Unit (2010) Latin American green city index: assessing the environmental performance of Latin America's major cities. Siemens, Munich

Ekilik VV, Grigor'ev VP (2002) Metal corrosion in organic and aqueous-organic media. Prot Met 38(2):124–131

FMC technologies compatibility manual (1996) http://info.smithmeter.com/literature/docs/ab0a002.pdf. Accessed 20 Feb 2012

Gubta JP, Babu BS (1999) A new hazardous waste index. J Hazard Mater 67(1):1–7

Halfon E, Reggiani M (1986) On ranking chemicals for environmental hazard. Environ Sci Technol 20:1173–1179

Halfon E, Galassi S, Brüggemann R, Provini A (1996) Selection of priority properties to assess environmental hazard of pesticides. Chemosphere 33:1543–1562

Harrington EC (1965) The desirability function. Ind Qual Cont 21:494–498

Heikkila A-M (1999) Inherent safety in process plant design: an index-based approach. Dissertation, Technical Research Centre of Finland, VTT Publication 384, Espoo

Lerche D, Brüggemann R, Sørensen P, Carlsen L, Nielsen O (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

Liu Y, Rogers W, Sam Mannan M (2006) Screening reactive chemical hazards. CEP Mag, May:41–47

Lothrop WC, Handrick GR (1949) The relationship between performance and constitution of pure organic explosive compounds. Chem Rev 44:419–445

National Fire Protection Association (NFPA) (1994) Standard 49. Hazardous chemical data, Quincy

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

Popova A, Christov M, Zwetanova A (2007) Effect of the molecular structure on the inhibitor properties of azoles on mild steel corrosion in 1 M hydrochloric acid. Corros Sci 49: 2131–2143

Rathnayaka S, Khan F, Amyotte P (2012) Accident modeling approach for safety assessment in an LNG processing facility. J Loss Prev Process Ind 25(2):414–423

Swanson M, Davis G, Kincaid L, Schultz T, Bartmess J, Jones S, George E (1997) A screening method for ranking and scoring chemicals by potential human health and environmental impact. Environ Toxicol Chem 16:372–383

Thomas HM (1981) Pipe and vessel failure probability. Reliab Eng 2(2):83–124

U.S. Chemical Safety and Hazard Investigation Board (2002) Hazard investigation: improving reactive hazard management. Report No. 2001-10-H

Viscusi WK, Zeckhauser RJ (2004) The denominator blindness effect: accidents frequencies and the misjudgement of recklessness. Am Law Econ Rev 6(1):72–94

Voigt K, Gasteiger J, Brüggemann R (2000) Comparative evaluation of chemical and environmental online and CD-ROM database. J Chem Inf Comput Sci 40:44–49

Voigt K, Brüggemann R, Kirchner M, Schramm K (2010) Influence of altitude concerning the contamination of humus soil in the German Alps: a data evaluation approach using PyHasse. Environ Sci Pollut Res 17:429–440

Ward PB (2002) Analyzing the past, planning the future, for the hazard of management. Trans IChemE Part B Proc Saf Environ Prot 80(1):47–54

Wei C, Rogers W, Sam Mannan M (2004) Application of screening tools in the prevention of reactive chemical incidents. J Loss Prev Proc Ind 17:261–269

Chapter 11
Hasse Diagram Technique Can Further Improve the Interpretation of Results in Multielemental Large-Scale Biomonitoring Studies of Atmospheric Metal Pollution

Stergios Pirintsos, Michael Bariotakis, Vaios Kalogrias, Stella Katsogianni, and Rainer Brüggemann

Abstract Lichens and mosses have extensively been used in multielemental large-scale biomonitoring studies of atmospheric metal pollution. Despite its high importance in the assessment of cumulative risk and the communication with risk managers, the presentation and interpretation of biomonitoring results have only been partially the center of interest for a standardized methodology and for the harmonization of the techniques. Here we attempt to expand and improve the up-to-date formal presentation of biomonitoring results, combining the Hasse diagram technique with GIS techniques. The implementation using real data has demonstrated that such an expansion and improvement, in the direction of cumulative risk assessment and management, is feasible and it is suggested for incorporation in biomonitoring studies.

11.1 Biomonitors and Biomonitoring

Lichens and mosses may be considered as the most commonly used organisms in biomonitoring studies of metal pollution in the atmospheric environment (Kularatne and Freitas 2013; Ares et al. 2012).

Lichens are by definition symbiotic organisms, usually composed of a fungal partner, the mycobiont, and one or more photosynthetic partners, the photobiont, which is most often either a green alga or cyanobacterium (Nash 2008), while mosses are photosynthetic non-vascular and non-woody plants grouped in the

S. Pirintsos (✉) • M. Bariotakis • V. Kalogrias • S. Katsogianni
Department of Biology, University of Crete, P.O.B. 2208, 71409 Heraklion, Greece
e-mail: pirintsos@biology.uoc.gr

R. Brüggemann
Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Mueggelseedamm 310, 12587 Berlin, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*, DOI 10.1007/978-1-4614-8223-9_11, © Springer Science+Business Media New York 2014

division of Bryophyta, differing among others from vascular plants in lacking water-bearing xylem tracheids or vessels (Goffinet et al. 2009).

Their use in biomonitoring studies is largely based on their lack of any roots comparable with higher plants, which implies that both obtain their mineral supplies from aerial sources and not from the substratum (Little and Martin 1974; Steinnes et al. 1994; Tyler 2008; Salo et al. 2012). However, uptake or other contributions from the substrate have also been reported and studied to detect its relative share (Garty 2001; Wolterbeek 2002).

In the literature two types of biomonitoring are clearly differentiated: (1) passive biomonitoring, using lichens and mosses that grow naturally in a particular area and (2) active biomonitoring, by transplanting lichens and mosses from other locations. While the use of native lichens and mosses is more appropriate for extensive studies in large areas (i.e., regional or national studies), active biomonitoring is more useful for intensive studies in smaller areas (i.e., urban or industrial areas) (Garty 2001; Ares et al. 2012).

For active biomonitoring, lichen and moss samples are collected from relatively unpolluted habitats; they are then cleaned, selected, and pretreated before being exposed in a different environment. Several papers have appeared stimulating methodological issues concerning a standardized implementation of active biomonitoring (Szczepaniak and Biziuk 2003; Szczepaniak et al. 2007; Moreira et al. 2005; Godinho et al. 2008; Aničić et al. 2009; Giordano et al. 2009 among others). Selection and preparation of proper species, pre-exposure treatment, and vital state of the species, preparation of the transplants or the bags, method and duration of exposure of the samples, post-exposure treatments, analytical techniques and methods, and quality assurance are basic concepts in such an effort for a standardized methodology and for the harmonization of techniques (Ares et al. 2012; Giordano et al. 2013).

Despite its high importance, the presentation and interpretation of biomonitoring results have only been partially the center of interest for the standardization of the biomonitoring methodology. For spatial interpolation, ordinary kriging method (Haining 2003) has been widely employed, based on the assumption that there is a definite correlation between the distance between points and the degree of their similarity, expressed through the semi-variogram (Dmuchowski et al. 2011).

An important contribution among others in mapping biomonitoring results is that of Nimis and Bargagli (1999). It classifies the results according to a suggested seven class scale of environmental alteration based on element concentration in lichen samples. Interpretation is based on the naturality/alteration scale developed for element concentrations (Al, Ba, As, Cd, Cr, Cu, Fe, Hg, Mn, Ni, Pb, V, and Zn) in foliose epiphytic lichens of Italy. This seven class scale (1: very high naturality, 2: high naturality, 3: middle naturality, 4: low naturality/alteration, 5: middle alteration, 6: high alteration, and 7: very high alteration) expresses the degrees of deviation (environmental alteration) from background conditions (naturality) based on the percentile distributions of element concentrations in lichens (Nimis et al. 2000).

The presentation and interpretation of biomonitoring results nowadays become also very important for two extra reasons: (1) the assessment of cumulative risk and (2) the communication with risk managers (U.S.EPA 2007).

Cumulative risk is defined as the combined risks from aggregate exposures to multiple agents or stressors, while cumulative risk assessment is the analysis, characterization, and possible quantification of the combined risks to human health or the environment from multiple agents or stressors (U.S.EPA 2003).

Cumulative risk assessment of pollution in relation to human health is a fast emerging research field of high social priority (Callahan and Sexton 2007; Skubisz et al. 2009). Interpretation of cumulative risk assessment results is depending on the target group of interest, such as stakeholders, risk managers, professional groups, citizen groups, etc. (Vaughan 1995), and it is part of the "risk communication process" (Leiss 1996), which is defined as an interactive process of exchange of information and opinions between individuals, groups and institutions, involving discussions of types and levels of risk and measures for dealing with risks.

The interpretation of pollution risk has been promoted by the development of GIS technology and high throughput analytical instruments. Therefore, in local or regional scales, numerous pollution maps are published in the literature (Briggs et al. 1997; Vu et al. 2013) and numerous pollution maps are produced/used in the daily practice from the stakeholders (RoTAP 2012). Nevertheless, cumulative risk maps (Lahr et al. 2010; Wang et al. 2012) constitute a small part of the total available pollution maps, as there is no formal way for a holistic and cumulative interpretation of the results.

This is also the case for maps presenting the metal accumulation in lichens and mosses (Scerbo et al. 1999; Carreras and Pignata 2002) within the framework of biomonitoring metal pollution of the atmospheric environment as an indicator of potential exposure of citizens to pollutants.

In this chapter, our analysis rests on the conception that Hasse diagram technique can further expand and improve in the direction of cumulative risk the up-to-date formal presentation and interpretation of results in biomonitoring studies of metal atmospheric pollution. The idea will be tested with real data, using the dataset published in Demiray et al. (2012).

11.2 Partial Order Theory and Hasse Diagram Technique

11.2.1 Partial Order

As mentioned in Sect. 11.1, stations can be characterized by a set of metals and their concentrations in biological targets or the metals can be characterized with respect to their concentrations in different stations. In both cases a multi-indicator system is to be analyzed (Brüggemann and Patil 2011). A ranking (of stations) due to their metal contamination or of metals due to the pollution of stations is far from being trivial. The increasing field of multicriteria decision support systems, MCDS (see for instance Munda 2008; Huang et al. 2011; Brüggemann and Carlsen 2012) demonstrates that there is no unique and simple solution. Partial order theory can be helpful, however, it should be rather seen as an analysis tool than as a pure decision support system.

Given a finite set of objects it is possible to define partial order relations among them in a multitude of ways, whence partial order theory became a powerful technique in many applied sciences (compare in that context Wolski 2008). Even, when a data matrix is at hand and the data matrix can be considered as suitable for a ranking (see this book, Chapter 11) there are many different possibilities to define a partial order relation.

11.2.2 Hasse Diagram Technique

Let X be a finite set of objects (often also called "elements") and IB the set of indicators q_j ($j = 1,\ldots,m$), then one of the most simple ways to define a partial order is Eq. (11.1a):

$$x, y \in X, x \leq y :\Leftrightarrow q_j(x) \leq q_j(y) \quad \mathit{for\,all}\, q_j \in IB. \tag{11.1a}$$

When

$$q_j(x) = q_j(y) \quad \mathit{for\,all}\, q_j \in IB\, \mathit{then}\, x \cong y \tag{11.2}$$

with ≅ indicating an equivalence.

In most practical applications based on (11.1a) and (11.2), only a representative element of an equivalence class is considered and (11.1a) is specified as follows:

$$x, y \in X, x < y :\Leftrightarrow q_j(x) \leq q_j(y) \quad \mathit{for\,all}\, q_j \in IB \tag{11.1b}$$

with at least one q_{j^*} for which a strict inequality holds.

Equations (11.1a), (11.1b), and (11.2) are a very specific realization of a partial order, whence in the literature the resulting analysis technique is called Hasse diagram technique (HDT). Elements which obey (11.1a) are called comparable. If (11.1a) does not hold, then x and y are incomparable, in sign: $x \parallel y$.

Equations (11.1a) is already applied for example on monitoring studies, concerning the pollution of pesticides (Sørensen et al. 2003); in that context also a new index, based on (11.1a), was suggested (Sørensen et al. 2010).

11.2.3 Cover Relation and Drawing a Hasse Diagram

Equation (11.1a) fulfills the axioms of partial order, especially the transitivity:

$$x, y, z \in X, \quad \mathit{if}\ x \leq y\ \mathit{and}\ \ y \leq z,\ \mathit{then} : x \leq z. \tag{11.3}$$

The situation, where $x \leq z$ and there is no element y for which (11.3) holds is of special importance, because x and z can be considered as immediate neighbors.

Table 11.1 Cover-relation as the basis for Fig. 11.1

Stations	Is covered by	Stations	Is covered by
1	3	11	6
2	4	12	6, 8, 9
3		13	6
4	3	14	2
5	3	15	1, 11, 12 , 13
6	5, 7, 14	16	12
7	3	17	6, 8
8	5, 7, 10, 19	18	3
9	2	19	18
10	14	20	6, 10, 19

A pair $x < y$ where no third element of X is between x and y is called a cover-relation and designed as x<: y. Cover-relations are the basis to draw a Hasse diagram. Anticipating Sect. 11.3, Table 11.1 shows the cover-relations:

(a) x <: y is drawn in a plane such that y is located vertically above x.
(b) The location of the elements of X (to be more exact: of the representative elements) is done as symmetric as possible.
(c) The vertical arrangement of the objects should be done with the least number of different heights (taking from the bottom). Hence subsets of objects are arranged in the same vertical level.
(d) In cases objects can be assigned to different levels, they will be located in the highest one. This rule is not justified by any order theoretical argument but by conservatism: Most often high values in indicators are associated with high risk. Hence risky objects should be at the top of the diagram.

A view on Fig. 11.1 may be helpful:

17 <: 6 is a cover relation, 20 < 14 is not a cover relation, because there is element 10 in between.

The objects of X are positioned symmetrical with respect to a thought central axis; elements 3 and 5 are located at this thought line, elements 2 and 19 have the same distance to the central line.

The elements are organized in levels. The first level is constituted by 15 and 16, the third, e.g., by 8, 11, 13, and 20.

Element 17 could also be located at the first level, however, rule (d) is applied (other examples are elements 1 and 20).

11.2.4 Chains and Levels

A chain is a subset $X' \subseteq X$, where all elements are mutually comparable. If $X' = X$, then the complete object set can be linearly ordered. In that case the partial order provides a complete ranking.

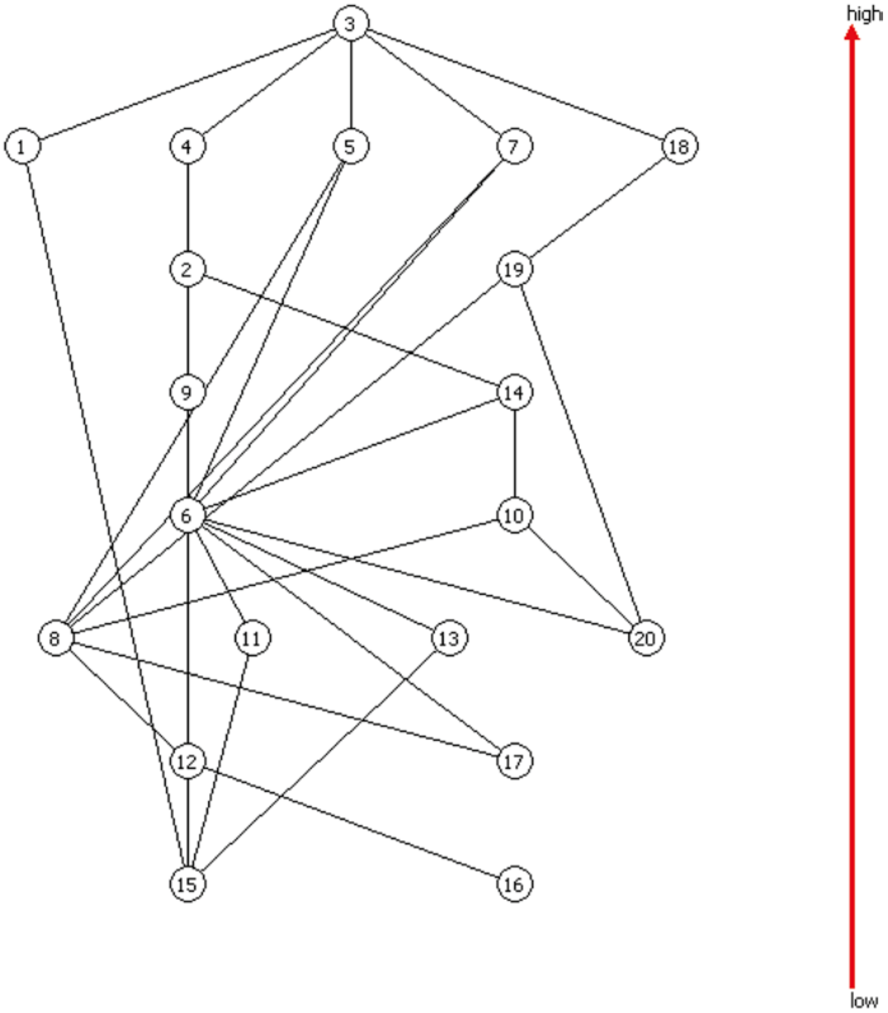


Fig. 11.1 Hasse diagram of the twenty sampling stations of Demiray et al. (2012) dataset modified according to the naturality/alteration seven class scale of Nimis and Bargagli (1999)

If X' is a proper subset of X, then nevertheless X' may be a chain. Interpreted on the basis of (11.1a), the presence of chains indicates that the indicator values are not countercurrent. Therefore a "partial ranking" is possible. At least for the elements of a chain a ranking can be performed, without applying knowledge beyond the data matrix or the need of sophisticated MCDSs.

Now consider a limit α as follows: If the indicator value of a metal ≤α, then the metal may be considered as not present at the corresponding site.

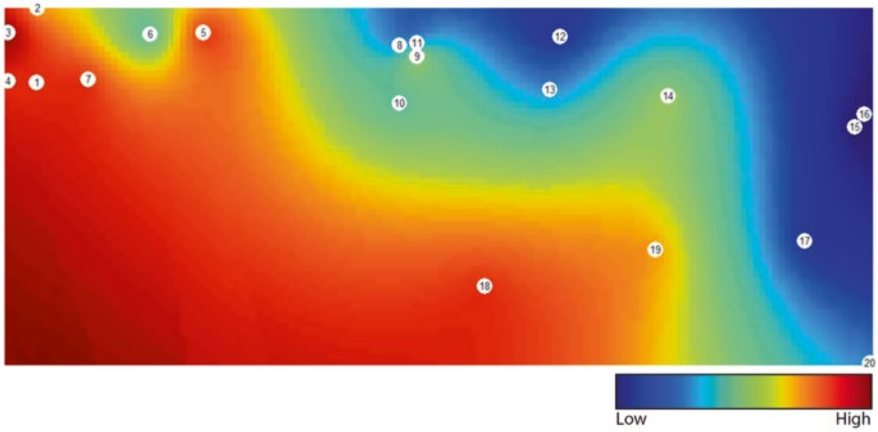


Fig. 11.2 The spatial pattern of alteration/naturality classes based on the output of sampling stations. HDT, kriging interpolation, and SAGA-GIS software (available at http://www.saga-gis.org/) for geospatial data have been used for the mapping

When sites are ordered due to their indicator values of metals, then there are two cases:

(a) $x<:y$ all indicator values I of metal j of x ($I_j(x)$) are less than those of y, and $I_j(x)> \alpha$.
(b) $x<:y$ all $I_j(x)) < I_j(y)$, and for $j \in \{1,\ldots,m\}$ $I_j(x)\leq \alpha$ and $I_j(y) > \alpha$ holds.

Case (a) is denoted a quantitative step and case (b) a qualitative one, because in site x less metals are considered as present than in y. When the direction in a Hasse diagram is taken downwards, then contextually the steps are "improvements".

Chains can have different lengths, i.e., different number of elements. Once again in Fig. 11.2 the sequence 15 < 1 < 3 is a chain and its length is 3.

Chains of maximal length are of special interest (because of the interpretation due to the indicator values). So 15 < 1 <3 is a maximal chain, because there is no element which can be inserted keeping the mutual comparability of {1, 3, 15}. A maximal chain which has a maximum length defines the number of possible levels:

$$\textit{Number of levels} = \textit{Length of maximum of maximal chain.} \tag{11.4}$$

Taking once again Fig. 11.1 as an example, a maximum of maximal chains is

$$15 < 12 < 8 < 6 < 14 < 2 < 4 < 3.$$

Accordingly the Hasse diagram, shown in Fig. 11.1 has eight levels.

Partial order theory shows that levels constitute an order, i.e., level1 < level 2 < level 3 <…< level 8 (see Fig. 11.1). The order relation between any two levels I and k is based on the order relation among the elements of the level i and level k. Two levels are neighbored if order relations between the elements of both levels are cover relations.

Hence, the system of increasing levels indicates increasing indicator values, because order relations exist among the elements of any two levels. More formally, a weak order exists among the elements of X as follows:

$$x \cong y \text{ if } x,y \text{ belong the same level}$$

$$x < y \quad \text{if } x \text{ belongs to levels } i \text{ and } y \text{ belongs to level } k \text{ with } i < k. \tag{11.5}$$

11.2.5 Transposition of the Data Matrix

A multi-indicator system is quantified by a data matrix, where the rows are defined by the objects and the columns by the indicators. Application of (11.1a) or (11.1b) defines an order among the row defining elements by examining the values of the column-defining indicators. As mentioned in Sect. 11.2.1 often the interpretation of a data matrix allows both: (X, IB) or (IB, X). In the first case the elements of X ("the objects", here the sites) are ordered due to the metal concentrations; in the second case the metals are ordered due to their site-specific concentrations. Let the data matrix be denoted as $D(i,j)$, subscript i counts the rows, and subscript j counts the columns. Then $D(i,j)$ is associated with (X,IB), whereas $D(j,i)$, the transposed matrix, is associated with (IB,X). In the first case a Hasse diagram of stations, in the second case a Hasse diagram of metals is obtained. It may be noted that partial order theory allows a unifying view on both (X,IB) and (IB,X), namely, by the Formal Concept Analysis (see Ganter and Wille 1996; Annoni and Brüggemann 2008), which, however, is outside of the scope of this chapter.

11.3 Expanding the Interpretation of Results in Biomonitoring

The published lichen biomonitoring dataset in Demiray et al. (2012) has been used for the implementation of Hasse diagram technique, modified according to the naturality/alteration seven class scale of Nimis and Bargagli (1999) (Table 11.2). In this article, airborne metal deposition in the major urban and industrial districts of Kocaeli was monitored using *Xanthoria parietina* lichen specimen as a biomonitoring organism. Samples were collected from 20 sampling stations (1–20) distributed around the major industrialized, urban–suburban areas of Kocaeli and at distances of at least 100 m from highways and main roads and any settlement area. Lichen samples were analyzed for Al, As, Co, Cd, Cu, Fe, Hg, Mn, Ni, Pb, Ti, Tl, V, and Zn contents to determine the relationship between the potential pollutant sources in the region and the degree of airborne metal deposition. Results showed (Demiray et al. 2012) that airborne metal deposition in the Kocaeli province was widespread and environmental alteration was serious near the industrial facilities.

Table 11.2 The lichen biomonitoring dataset of Demiray et al. (2012), modified according to the naturality/alteration seven class scale of Nimis and Bargagli (1999)

Station	Al	As	Cd	Cu	Fe	Hg	Mn	Ni	Pb	V	Zn
1	4	2	5	5	7	4	7	7	7	7	7
2	5	5	7	5	7	3	7	7	7	7	7
3	7	7	7	7	7	5	7	7	7	7	7
4	5	6	7	7	7	3	7	7	7	7	7
5	5	7	4	4	7	3	6	6	5	7	7
6	4	5	3	3	7	3	6	5	4	6	7
7	4	5	4	5	7	4	7	7	6	7	7
8	4	4	2	3	7	2	4	4	4	7	6
9	3	4	2	3	6	2	5	4	7	4	7
10	5	4	3	4	7	2	6	7	4	7	7
11	3	4	2	3	5	3	5	4	3	4	6
12	3	3	2	3	6	2	4	4	3	4	5
13	4	3	1	3	7	3	5	4	3	5	6
14	5	5	3	4	7	3	6	7	4	7	7
15	3	2	1	3	5	2	4	3	3	4	3
16	3	3	2	3	6	2	4	4	2	4	4
17	4	3	2	2	7	2	4	4	2	5	5
18	7	4	3	4	7	2	6	7	4	7	6
19	6	4	3	4	7	2	5	7	4	7	6
20	4	3	3	2	5	2	4	3	2	4	5

The results of applying the Hasse diagram technique (HDT) for the 20 sampling stations of Demiray et al. (2012) dataset are presented in Fig. 11.1. Figure 11.1 is based on (*X*,IB). *X* the set of stations and IB the set indicators with values modified according to the naturality/alteration seven class scale of Nimis and Bargagli (1999).

Eight levels of sampling stations were resulted: Level 1: 15,16; Level 2: 12,17; Level 3: 8,11,13,20; Level 4: 6,10; Level 5: 9,14; Level 6: 2,19; Level 7: 1,4,5,7,18; and Level 8: 3 ranging from low to high pollution or from very high naturality class to very high alteration class.

The spatial pattern of alteration/naturality classes based on the output of HDT, which reflects the spatial pattern of cumulative risk in the area considering all metals, is presented in Fig. 11.2. Blue corresponds to very high naturality and red to very high alteration class reflecting low and high cumulative risk correspondingly.

Comparing this output with the spatial pattern of naturality and environmental alteration for each metal (Demiray et al. 2012), it becomes clear that using HDT we can expand the presentation and interpretation of the results adding the spatial pattern of cumulative risk.

For example (see also Table 11.2), in station 3, the cumulative risk is the maximum, as all the recorded metals, with the exception of Hg, appeared with the maximum value of alteration (7 = very high alteration), while in station 15, the cumulative risk compared to station 3 is minimum, as only one metal (Fe) appeared with the middle value of alteration (5 = middle alteration).

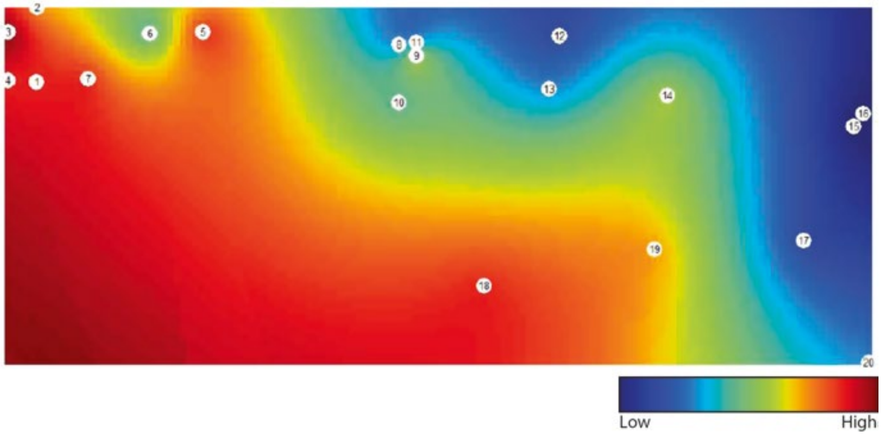


Fig. 11.3 The spatial pattern of alteration/naturality classes based on the output of sampling stations. Hasse diagram for toxic metals category (Al, As, Cd, Hg, Pb, Ni, and V). For mapping details see Fig. 11.1

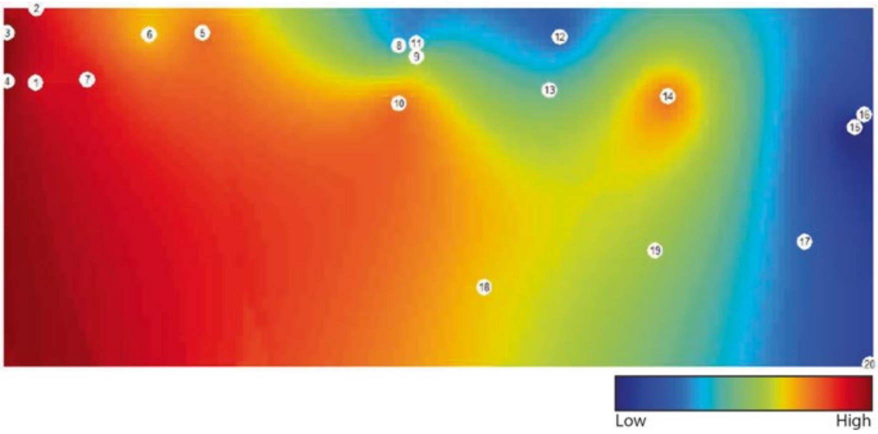


Fig. 11.4 The spatial pattern of alteration/naturality classes based on the output of sampling stations. Hasse diagram for nontoxic metals category (Cu, Fe, Mn, and Zn). For mapping details see Fig. 11.1

Moreover, applying the HDT, we can map cumulative risk of specific metal categories. Metals can be classified in different categories using different classification criteria, such as the origin of the metals, the toxicity of the metals, the health effects on specific organs (e.g., metals which affect lung, metals which affect liver), etc.

For the same published dataset, we have conventionally attempted a classification of the metals in two categories, toxic (Al, As, Cd, Hg, Pb, Ni, and V) vs. nontoxic (Cu, Fe, Mn, and Zn), based on official databases about their health effects. The spatial pattern of alteration/naturality classes based on the output of HDT, considering separately each metal category are presented in Figs. 11.3 and 11.4.

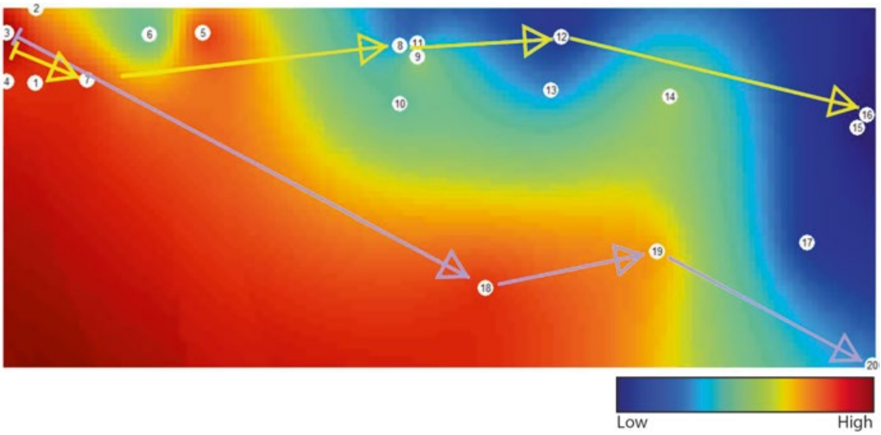


Fig. 11.5 Nested pollution gradients of sampling stations based on the output of sampling stations Hasse diagram

It seems that the spatial pattern of cumulative risk, based on the naturality and environmental alteration scale of sampling stations, can result from applying the HDT for specific metal categories, expanding further the potentiality of presentation and interpretation of the biomonitoring results.

Additionally, another advantage from the implementation of HDT in the presentation of biomonitoring results is the fact that nested pollution gradients of the sampling stations can be revealed. From the Hasse diagram of Fig. 11.1, several nested pollution gradients can be selected for a mapping presentation. Nested is used in the sense of a chain where each member contains the next. For example in Fig. 11.5, two nested pollution gradients are presented. Analyzing the chains of the partial order, whose graphical representation is shown in Fig. 11.1, the gradient 3→7→8→12→16 (yellow arrows) and the gradient 3→18→19→20 (purple arrows). So, it seems that through HDT we can also further expand the potential of presenting the biomonitoring results, presenting nested pollution gradients of sampling stations.

Nested pollution gradients may reflect gradients of qualitative and/or quantitative improvement. In qualitative improvement, at the transition from one station to another along the gradient, environmental alteration is presented only in a subset of metals, while in quantitative improvement it is only a matter of changing class without removal of metals. This information could also expand the potential of presenting the biomonitoring results. In Figs. 11.6 and 11.7, two nested pollution gradients are presented, a qualitative and a qualitative/quantitative, respectively.

Likewise, we can consider (IB,*X*), i.e., the transposed data matrix and apply HDT for the metals, according to their naturality/alteration scale in the 20 sampling stations. Such a Hasse diagram is presented in Fig. 11.8. Metals at the bottom of the diagram have low contribution, while metals at the top of the diagram have high contribution in the alteration of the environment in the study area.

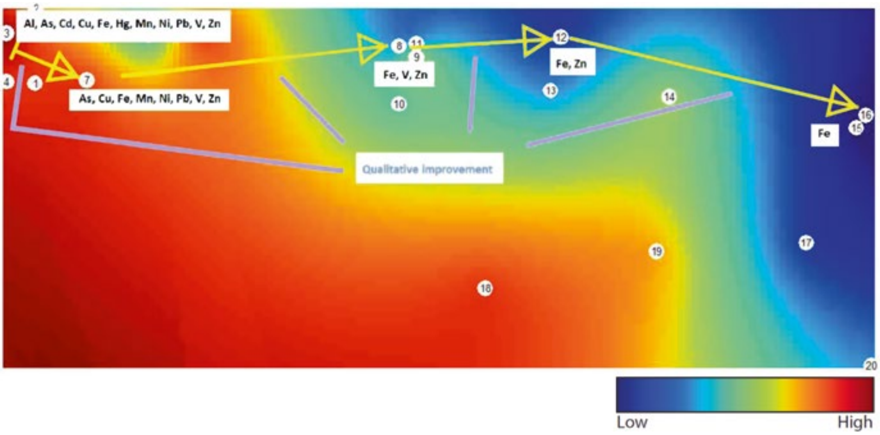


Fig. 11.6 As Fig. 11.5, however now qualitative improvements (see Sect. 11.2.3) is recorded

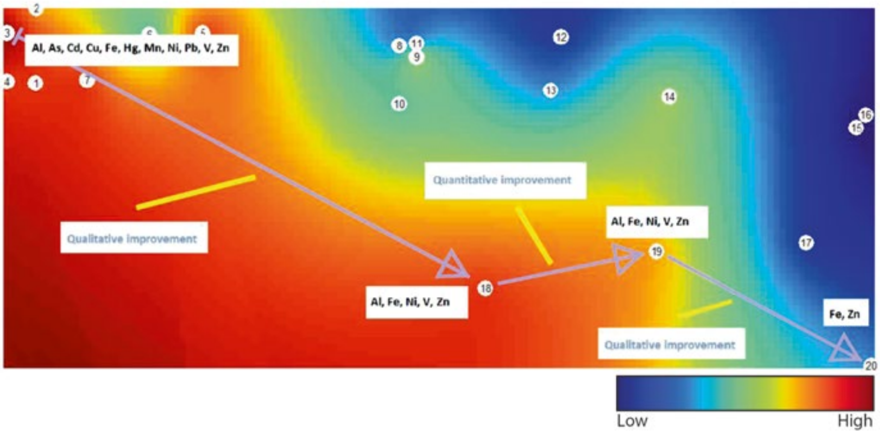


Fig. 11.7 A nested pollution gradient where both qualitative and quantitative improvements are recorded

In this diagram Cadmium (Cd) is connected to three lines (a) Iron (Fe)—Vanadium (V)—Nickel (Ni)—Cadmium (Cd), (b) Zinc (Zn)—Lead (Pb)—Cadmium (Cd), and (c) Iron (Fe)—Manganese (Mn)—Cadmium (Cd), which possibly reflect different nested pollution origin, the line (a) possibly reflects sources of burning coal or oil, the line (b) possibly reflects car pollution, while the line (c) possibly reflects other sources such as alloys industry.

So, the Hasse diagram of metals can extend the interpretation of the results contributing to the perception of the discrimination of possible pollution sources.

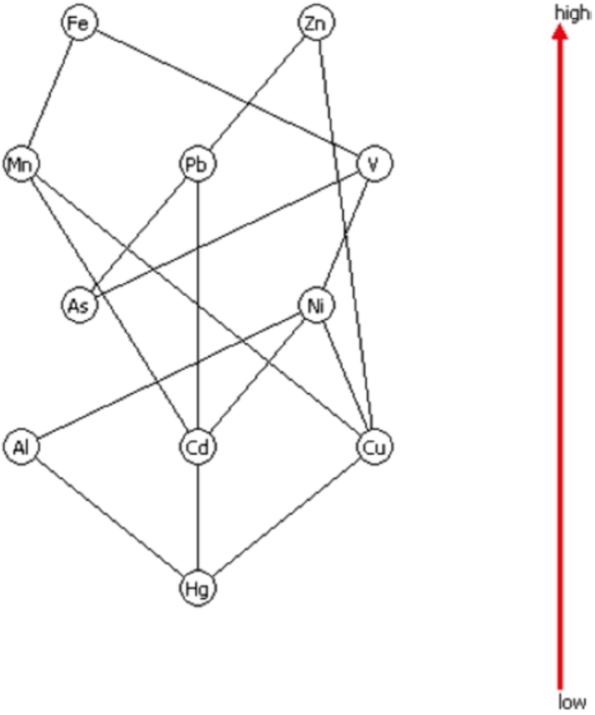


Fig. 11.8 Hasse diagram of the metals, according to their naturality/alteration scale in the 20 sampling stations

11.4 Conclusions and Outlook

In conclusion, in biomonitoring studies of metal pollution in the atmospheric environment, using HD technique

1. We can map cumulative risk.
2. We can map risk of specific metal categories.
3. We can draw in detail the pollution gradients.
4. We can extract quantitative vs. qualitative alterations.
5. We can contribute to the perception of the discrimination of possible pollution sources.

We will further expand the presentation and interpretation of biomonitoring results, improving also the risk communication processes. Especially it will be of high interest to study temporal developments. Furthermore the intrinsic relations between indicators, indicator values, and the ranking of stations will be studied applying Formal Concept Analysis (Ganter and Wille 1996; Annoni and Brüggemann 2008).

References

Aničić M, Tomašević M, Tasić M, Rajšić S, Popović A, Frontasyeva MV, Lierhagen S, Steinnes E (2009) Monitoring of trace element atmospheric deposition using dry and wet moss bags: accumulation capacity versus exposure time. J Hazard Mater 171:182–188

Annoni P, Brüggemann R (2008) The dualistic approach of FCA: a further insight into Ontario Lake sediments. Chemosphere 70:2025–2031

Ares A, Aboal JR, Carballeira A, Giordano S, Adamo P, Fernández JA (2012) Moss bag biomonitoring: a methodological review. Sci Total Environ 432:143–158

Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, Pryl K, van Reeuwijk H, Smallbone K, van der Veen A (1997) Mapping urban air pollution using GIS: a regression-based approach. Int J Geogr Inf Sci 11:699–718

Brüggemann R, Carlsen L (2012) Multi-criteria decision analyses. Viewing MCDA in terms of both process and aggregation methods: some thoughts, motivated by the paper of Huang, Keisler and Linkov. Sci Total Environ 425:293–295

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems – introduction to partial order applications. Springer, New York, NY

Callahan MA, Sexton K (2007) If cumulative risk assessment is the answer, what is the question? Environ Health Perspect 115:799–806

Carreras HA, Pignata ML (2002) Biomonitoring of heavy metals and air quality in Cordoba City, Argentina, using transplanted lichens. Environ Pollut 117:77–87

Demiray AD, Yolcubal I, Akyol NH, Çobanoğlu G (2012) Biomonitoring of airborne metals using the Lichen *Xanthoria parietina* in Kocaeli Province, Turkey. Ecol Indic 18:632–643

Dmuchowski W, Gozdowski D, Baczewska A (2011) Comparison of four bioindication methods for assessing the degree of environmental lead and cadmium pollution. J Hazard Mater 197: 109–118

Ganter B, Wille R (1996) Formale Begriffsanalyse Mathematische Grundlagen. Springer, Berlin

Garty J (2001) Biomonitoring atmospheric heavy metals with lichens: theory and application. Crit Rev Plant Sci 20:309–371

Giordano S, Adamo P, Monaci F, Pittao E, Tretiach M, Bargagli R (2009) Bags with oven-dried moss for the active monitoring of airborne trace elements in urban areas. Environ Pollut 157: 2798–2805

Giordano S, Adamo P, Spagnuolo V, Tretiach M, Bargagli R (2013) Accumulation of airborne trace elements in mosses, lichens and synthetic materials exposed at urban monitoring stations: towards a harmonization of the moss-bag technique. Chemosphere 90:292–299

Godinho RM, Wolterbeek HT, Verburg T, Freitas MC (2008) Bioaccumulation behaviour of transplants of the lichen *Flavoparmelia caperata* in relation to total deposition at a polluted location in Portugal. Environ Pollut 151:318–325

Goffinet B, Buck WR, Shaw J (2009) Morphology, anatomy, and classification of the Bryophyta. In: Goffinet B, Shaw J (eds) Bryophyte biology. Cambridge University Press, Cambridge

Haining R (2003) Spatial data analysis. Theory and Practice. Cambridge University Press, Cambridge

Huang IB, Keisler J, Linkov I (2011) Multi-criteria decision analysis in environmental sciences: ten years of applications and trends. Sci Total Environ 409:3578–3594

Kularatne KIA, Freitas CR (2013) Epiphytic lichens as biomonitors of airborne heavy metal pollution. Environ Exp Bot 88:24–32

Lahr J, Münier B, De Lange HJ, Faber JF, Sørensen PB (2010) Wildlife vulnerability and risk maps for combined pollutants. Sci Total Environ 408:3891–3898

Leiss W (1996) Three phases in the evolution of risk communication practice. Ann Am Acad Pol Soc Sci 545:85–94

Little P, Martin MH (1974) Biological monitoring of heavy metal pollution. Environ Pollut 6:1–19

Moreira FR, Borges RM, Oliveira RM (2005) Comparison of two digestion procedures for the determination of lead in lichens by electrothermal atomic absorption spectrometry. Spectrochim Acta Part B At Spectrosc 60:755–758

Munda G (2008) Social multi-criteria evaluation for a sustainable economy. Springer, Berlin

Nash III TH (2008) Introduction. In: Nash III TH (ed) Lichen biology. Cambridge University Press, Cambridge

Nimis PL, Bargagli R (1999) Linee-guida per l'utilizzo dei licheni epifiti come bioaccumulatori di metalli in traccia. In: Proceedings of Workshop Biomonitoraggio della qualità dell'aria sul territorio nazionale, Roma, 26–27 Giugno 1998. ANPA-Serie Atti, pp 279–287

Nimis PL, Lazzarin G, Lazzarin A, Skert N (2000) Biomonitoring of trace elements with lichens in Veneto (NE Italy). Sci Total Environ 255:97–111

RoTAP (2012) Review of transboundary air pollution: acidification, eutrophication, ground level ozone and heavy metals in the UK. Contract Report to the Department for Environment, Food and Rural Affairs. Centre for Ecology & Hydrology

Salo H, Bućko MS, Vaahtovuo E, Limo J, Mäkinen J, Pesonen LJ (2012) Biomonitoring of air pollution in SW Finland by magnetic and chemical measurements of moss bags and lichens. J Geochem Explor 115:69–81

Scerbo R, Possenti L, Lampugnani L, Ristori T, Barale R, Barghigiani C (1999) Lichen (*Xanthoria parietina*) biomonitoring of trace element contamination and air quality assessment in Livorno Province (Tuscany, Italy). Sci Total Environ 241:91–106

Skubisz C, Reimer T, Hoffrage U (2009) Communicating quantitative risk information. In: Beck CS (ed) Communication yearbook, vol 33. Routledge, New York, NY, pp 177–211

Sørensen PB, Brüggemann R, Carlsen L, Mogensen BB, Kreuger J, Pudenz S (2003) Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory. Environ Toxicol Chem 22:661–670

Sørensen PB, Giralt F, Rallo R, Espinosa G, Münier B, Gyldenkærne S, Thomsen M (2010) Conscious worst case definition for risk assessment. Part II: a methodological case study for pesticide risk assessment. Sci Total Environ 408:3860–3870

Steinnes E, Hanssen JE, Rambæk JP, Vogt NB (1994) Atmospheric deposition of trace elements in Norway: temporal and spatial trends studied by moss analysis. Water Air Soil Pollut 74: 121–140

Szczepaniak K, Biziuk M (2003) Aspects of the biomonitoring studies using mosses and lichens as indicators of metal pollution. Environ Res 93:221–230

Szczepaniak K, Astel A, Simeonov V, Tsakovski S, Biziuk M, Bode P, Przyjazny A (2007) Comparison of dry and living *Sphagnum palustre* moss samples in determining their biocumulative capability as biomonitoring tools. J Environ Sci Health Part A 42:1101–1115

Tyler G (2008) Bryophytes and heavy metals: a literature review. Bot J Linn Soc 104:231–253

U.S.EPA (2003) Framework for cumulative risk assessment. U.S.EPA/ORD/NCEA, Washington, DC. EPA/600/P-02/001F

U.S.EPA (2007) Concepts, methods and data sources for cumulative health risk assessment of multiple chemicals, exposures and effects: a resource document. ORD, NCEA, Cincinnati, OH. EPA/600/R-06/013F

Vaughan E (1995) The significance of socioeconomic and ethnic diversity for the risk communication process. Risk Anal 15:169–180

Vu V-H, Le X-Q, Pham N-H, Hens L (2013) Application of GIS and modelling in health risk assessment for urban road mobility. Environ Sci Pollut Res 20:5138–5149

Wang M, Bai Y, Chen W, Markert B, Peng C, Ouyang Z (2012) A GIS technology based potential eco-risk assessment of metals in urban soils in Beijing, China. Environ Pollut 161:236–242

Wolski M (2008) Information quanta and approximation operators: once more around the track. In: Peters JF, Skowron A (eds) Transactions in rough sets VIII, vol 5084, Lecture notes in computer science. Springer, Berlin, pp 237–250

Wolterbeek B (2002) Biomonitoring of trace element air pollution: principles, possibilities and perspectives. Environ Pollut 120:11–21

Chapter 12
Application of Partial Orders and Hasse Matrices in Ranking Contaminated Sites

Ron J. Thiessen and Gopal Achari

Abstract Contaminated site cleanup is a multibillion dollar issue in Canada. Practical decision support tools are needed to help prioritise contaminated site cleanup funding across a portfolio of many sites. This chapter illustrates the concept of prioritising contaminated site management decisions using environmental risk, societal perception and environmental liability as the measures. Using previously published information on 20 contaminated sites located in Canada, these three aspects are combined by applying partial order concepts to prioritise sites. The authors show how societal perception and environmental liability influence site prioritisation compared to prioritisation based on environmental risk alone. They recommended additional research in quantifying the societal perception of contaminated sites prior to practical application of the concepts explained in the chapter.

12.1 Introduction and Background

Contaminated site cleanup is a multibillion dollar industry in Canada. In 2012, the Canadian federal government was responsible for managing 13,000 contaminated sites having a combined environmental liability of CAN$7.7 billion (OAG 2012). Canada's environment minister has developed a multiyear plan to remediate high priority contaminated sites (EC 2012).

This challenge is not unique to the federal government and extends to provincial governments, municipalities and industries responsible for contaminated sites. Site remediation is commonly viewed as a business issue having a low return on investment which must compete with opportunities that have a positive and tangible effect

R.J. Thiessen (✉) • G. Achari
Schulich School of Engineering, University of Calgary,
ENF 262, 2500 University Drive NW, Calgary, AB, Canada T2N 1N4
e-mail: ron.thiessen@shaw.ca; gachari@ucalgary.ca

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_12, © Springer Science+Business Media New York 2014

on an organisation's financial health. Environmental managers are often required to justify contaminated site assessment and remediation budgets to business managers and need to communicate with them in ways that make business sense. Given limited funds, tools are needed to prioritise contaminated site management.

Current decision-support tools used in contaminated site management do not adequately quantify environmental, societal and economic aspects in a practical way. Carlon et al. (2009) reviewed many existing tools used in Europe and North America and concluded that, "[they] have found poor application in the real world. It may be partly due to differences between the decision making process proposed by the [tools] and that occurring in practice". Goosen et al. (2007) were more candid when they concluded that "too often, systems are developed from a technological developers' push (supply driven) rather than through a demand driven process…"

12.2 Objective and Scope

The objective is to develop a practical tool to prioritise contaminated site management funding across a portfolio of many sites. As a step towards this objective, the scope of this chapter is to illustrate the concept of prioritising contaminated site management using environmental risk, societal perception and environmental liability as the measures. Environmental risk is calculated using human health and ecological risk assessment methods accepted by North American environmental regulators. Societal perception is approximated by applying current research in risk perception; however, further research is required before perception estimation is used in practical applications. Environmental liability is quantified according to accounting practices acceptable in Canada.

12.3 Literature Review

12.3.1 Environmental Risk

Human health and ecological risk assessments frameworks in North America have been developed and refined over the past three decades. The United States Environmental Protection Agency (USEPA) conducted its first human health risk assessment, on vinyl chloride, in 1975 and issued its first health risk assessment guidance document in 1976 (USEPA 1976, 2012).

The US National Research Council (NRC 1983) subsequently established a standard process for conducting human health risk assessments. Similar guidance documents were prepared by the USEPA (1992) and NRC (1993) for ecological risk assessment.

In Canada, the Canadian Council of Ministers of the Environment (CCME 1996, 1997) defined the structure of ecological risk assessments based on guidance from NRC (1983) and the USEPA (1992). Similarly, Health Canada (HC 2004a, b)

Table 12.1 Human risk perception characteristics and latent risk factors (Slovic et al. 1980)

Risk perception characteristics	Latent risk factors
1. *Severity uncontrolled*: If a mishap occurs, can the damage be controlled?	Dread
2. *Dreaded risk*: Is this a risk that people have learned to live with and can think about reasonably calmly, or is it one that people have great dread?	
3. *Global catastrophic*: Does the hazard threaten global catastrophe?	
4. *Mishap unpreventable*: Can mishaps be prevented?	
5. *Fatal consequence*: If exposed to the risk, to what extent can you avoid death while engaging in the activity?	
6. *Inequitable risks and benefits*: Are the benefits equitably distributed among those at risk?	
7. *Catastrophic*: Is this a risk that kills people one at a time or kills large numbers of people at once?	
8. *Future generations threatened*: Does the hazard threaten future generations?	
9. *Risk not easily reduced*: Can the risk be reduced easily?	
10. *Risk increasing*: Is the risk increasing or decreasing?	
11. *Risk involuntary*: Do people get into these risk situations voluntarily?	
12. *Personal affect*: Are you personally at risk from this hazard?	
13. *Not observable*: Are the damage-producing processes observable as they occur?	Familiarity
14. *Unknown to exposed*: To what extent are the risks known precisely by the persons who are exposed to the risk?	
15. *Effect immediacy*: To what extent is the risk of death immediate?	
16. *Unfamiliar*: Are these risks new, novel ones or old, familiar ones?	
17. *Unknown to science*: To what extent are the risks known to science?	
18. *Many people exposed*: How many people are exposed to this hazard?	Number exposed

established the structure of human health risk assessments related to contaminated sites under the purview of the Canadian federal government. Both Health Canada (2004a) and CCME (1996, 1997) identified the major elements of risk assessments as problem formulation (also known as a conceptual site model development), exposure assessment, hazard (or toxicity) assessment and risk characterisation.

The structure of environmental risk assessments is well established and the reader is encouraged to review the referenced, publicly available documents for further background. Details on the risk assessments conducted as part of the chapter's topic are provided in Sect. 12.4.2.

12.3.2 Risk Perception

In the field of risk perception, the articles written by Paul Slovic and his colleagues are frequently referenced given the volume of research they have completed in the last several decades. Slovic et al. (1980) investigated 90 hazards to humans that spanned many technologies, activities and substances. Approximately 7 % of those hazards were environmental contaminants. Survey respondents were asked to rate each hazard in terms of the 18 risk perception characteristics summarised in Table 12.1. Respondents were also requested to estimate how the risk of each hazard should be adjusted to make it acceptable. If one thought a hazard was “too risky”, the respondent was asked to provide a numerical entry, X, to the statement, “to be acceptable [the hazard] would have to be X times safer”. A similar statement was provided if a respondent thought the hazard was safe and could be riskier.

A common factor analysis was performed on the responses to derive three latent risk factors describing the underlying dimensions of the data. Referring again to Table 12.1, Slovic et al. (1980) observed the first 12 questions were intercorrelated and described the concept of risk dread; the following five questions characterised risk unfamiliarity; and the last question was distinct from the others, describing the latent factor of number of persons exposed. Hazards having large, positive factor scores were ones that respondents deemed highly dreaded, very unfamiliar, or affecting many people, depending on the latent factor in question.

Dread and unfamiliarity factor scores for the six identified environmental contaminants were clustered together, having positive values in the low to mid-ranges (Slovic et al. 1980). The positive sign and magnitude of factor scores suggested respondents were not neutral regarding these environmental contaminants but had moderate levels of concern about their use.

McDaniels et al. (1995) and Lazo et al. (2000) expanded the work of Slovic et al. (1980) to investigate perceptions of ecological or ecosystem risks. Lazo et al. (2000) examined perceptions of 25 ecosystem risks with approximately half of them related to global climate change. Only one of those risks, pesticides, related to contaminated sites. Respondents rated each of the 25 ecosystem risks considering the 27 risk perception characteristics summarised in Table 12.2. Lazo et al. (2000) conducted a common factor analysis of the responses and derived four latent factors that influence ecological risk perceptions: impact, control, acceptance and understanding. Referring to the same table, the risk perception characteristics were grouped according to these factors shown in the second column.

Kasperson et al. (2003) describes the ripple effect of a growing perception issue outward from directly affected individuals to the local community, professional organisations, other stakeholder groups and eventually society as a whole. Ordered by those directly affected to society as a whole, Sandman (1993) divided the public into nine groups: employees and retirees, neighbours and occupants, concerned citizens, subject matter experts, industry peers, elected officials, regulatory agencies, social issue activists and the media. He noted that each group has different perceptions and motivations for those perceptions.

Kasperson et al. (2003) also described the results of adverse public perceptions. Examples that were cited were financial losses incurred by the offending party (i.e. risk owner), regulatory action or litigation, community concern and loss of trust.

12.3.3 Environmental Liability

The Canadian federal government defines environmental liability as: “*…the estimated costs related to the management and remediation of environmentally contaminated sites. Based on management's best estimates, a liability is accrued and an expense*

Table 12.2 Ecological risk perception characteristics and latent risk factors (Lazo et al. 2000)

Risk perception characteristics	Latent risk factors
1. *People affected*: How many people could be affected by the risk?	Impact
2. *Human health threat*: To what extent does the risk threaten human health?	
3. *Human suffering*: How much human suffering will result from the risk?	
4. *Life relevance*: How relevant is the risk to your life?	
5. *Impact scope*: How big an area will the risk affect?	
6. *Emotional response*: What degree of negative emotions do you feel when thinking about the risk?	
7. *Impact duration*: How long will the risk's effects on the ecosystem last?	
8. *Species loss*: What is the potential for animal or plant species loss?	
9. *Rights infringement*: To what degree does the risk infringe on nonhuman species' rights?	
10. *Destructive potential*: How destructive is the risk to the ecosystem?	
11. *Animal or plant suffering*: How much animal or plant suffering will result from the risk?	
12. *Media attention*: How much media attention does the risk receive?	
13. *Effect certainty*: What is the degree of certainty that the risk will affect the ecosystem?	
14. *Impact control*: How controllable is the risk's impact on the ecosystem?	Control
15. *Risk regulation*: To what extent can the risk be regulated by governments?	
16. *Impact avoidance*: To what extent can society avoid the risk's occurrence?	
17. *Alternative availability*: Are there reasonable alternatives to the risk or actions that lead to the risk?	
18. *Risk goodness*: How good or bad is the risk in terms of its impact on the ecosystem?	Acceptance
19. *Societal benefit*: How much benefit to society may result from the risk's effect on the ecosystem?	
20. *Risk acceptance*: How acceptable is the risk's effect on the ecosystem?	
21. *Ecosystem adaptation*: How well will the ecosystem adapt to the risk?	
22. *Impact ethics*: How ethical is the risk's impact on the ecosystem?	
23. *Effects observed*: How observable are the risk's effect on the ecosystem?	Understanding
24. *Effect prediction*: To what extent can scientists predict the effects on the risk on ecosystems?	
25. *Impact recognition*: How quickly do experts recognise the risk's effects on ecosystems?	
26. *Effect immediacy*: How quickly are the risk's effects observed on the ecosystem?	
27. *Effect understanding*: To what degree is the effect of the risk understood?	

recorded when the contamination occurs or when the [government] *becomes aware of the contamination and is obligated, or is likely to be obligated to incur such costs*" (TBCS 2012).

A legal entity is accountable for an environmental liability when it has an obligation to restore a contaminated site to meet applicable environmental standards. This concept is the polluter pays principle (CCME 2006b). Often, the obligation is motivated to comply with regulations that are protective of human and ecological health.

12.3.4 Ranking Methods

There are several methods to rank objects given a set of several attributes to which scores for each object have been assigned. A review of the various methods is beyond the scope of this chapter since many reviews and descriptions have already been published. Examples include books by Pavan and Todeschini (2008) and Brüggemann and Patil (2011). A key distinction across methods is how they rank a pair of objects when the attribute scores of one object do not dominate all the corresponding scores of the other object. No single method is best in all applications and practitioners should be aware of the advantages and limitations of considered methods before selection.

12.4 Methods

12.4.1 Contaminated Site Selection (Excerpt from Thiessen and Achari 2011)

Real world contaminated sites were used in the prioritisation. Contaminated sites were chosen from the Federal Contaminated Site Inventory (FCSI), which is a web-accessible database containing information on contaminated sites under the responsibility of the Government of Canada (TBCS 2011). Four criteria were used to narrow the search in the FCSI for a group of candidate sites:

- The group of sites should represent Canada's varied climate and geography.
- At minimum, a Phase 2 ESA[1] should have been completed at each site.
- The group of sites should represent the more frequently encountered contaminant types.
- The group of sites should present risk to a range of human and ecological receptors.

Sites were chosen from those in the Northwest Territories, Nunavut and Yukon to represent Canada's North; British Columbia and the Alberta Foothills to represent the West Coast and Rocky Mountains; Saskatchewan and Manitoba for the Prairies; Ontario for Central Canada and the Atlantic Provinces for the Canadian Maritimes. Second, sites meeting at least *Step 5–Detailed Testing Programme* of the Federal 10-Step Process were included to meet the Phase 2 ESA requirement. The proportions of contaminant type reported in the FCSI were as follows: 36 % of sites had metals contamination; 36 % had petroleum hydrocarbon (PHC) contamination; 11 % were contaminated by polycyclic aromatic hydrocarbons (PAHs); 6 % were contaminated

[1] A Phase 2 Environmental Site Assessment (ESA) is a contaminated site investigation where soil, sediment, soil vapour, surface water or groundwater samples are obtained and analysed for contaminants.

by benzene, toluene, ethylbenzene and xylene (BTEX) and the balance impacted by a collection of other contaminant types (TBCS 2011). Since almost 90 % of sites were impacted by metals, PHCs, PAHs and BTEX, the sites selected had primarily these contaminants to meet the third criterion. To meet the fourth criterion, the FCSI's method of categorising sites into high, medium, low and negligible risk was referenced. This method is the National Classification System for Contaminated Sites (NCSCS) and is explained by CCME (2008). An approximately equal number of sites from each of the four classes were obtained.

Thus, a total of 20 sites were selected from the FCSI. ESA information on the selected sites was requested via the Canadian Access to Information Act. Table 12.3 briefly summarises information on the selected sites.

12.4.2 Environmental Risk Estimation (Excerpt from Thiessen and Achari 2012)

Preliminary quantitative risk assessments (PQRAs) were conducted on each of the 20 sites. Health Canada (2004a) describes a PQRA, in the context of human receptors, as a screening assessment that uses "prescribed methods and assumptions that ensure that exposures and risks are not underestimated". CCME (1996) describes a screening ecological risk assessment as "based primarily on data from literature, previous or preliminary studies of the contaminated site, monitoring studies, historical data of the site, and a reconnaissance visit to evaluate the receptors, exposure, hazards, and risk at the site". For the purposes of this chapter, both screening human health and ecological risk assessments are referred to as a PQRA. The four major elements of risk assessments identified in Sect. 12.3.1 above were considered.

12.4.2.1 Problem Formulation

In problem formulation, the following biological receptors identified by CCME (2006a) and elaborated by Thiessen and Achari (2011) were considered:

- Humans (toddler and adult)
- Terrestrial plants and soil invertebrates
- Soil microorganisms responsible for nutrient and energy cycling
- Agricultural livestock (dairy cow)
- Primary consumer (meadow vole)
- Secondary consumer (masked shrew)
- Tertiary consumers (American kestrel)
- Aquatic life including fish, aquatic plants and benthic organisms

The environmental data presented in the ESA reports were spatially referenced according to the polar coordinate system shown in Fig. 12.1. The grey ellipse represents a contaminant source zone with a plume and the black dots represent sample locations.

Table 12.3 Selected contaminated sites

Site Id	Contaminant source	Province or territory[a]	Ecoregion[b]	Contaminant types	Contaminated soil volume (m^3)
1	Waste soil landfill	BC	Eastern Vancouver Island	Metals, PCBs[c], PAHs, PHCs	16,000
2	Aboveground storage tank	BC	Western Vancouver Island	PAHs, PHCs	10
3	Weathered paint	ON	Manitoulin-Lake Simcoe	Metals	250
4	Mechanical repair area	BC	Coastal Gap	Metals, PAHs, PHCs	2,200
5	Aboveground storage tank	YT	Ruby Ranges	PAHs, PHCs	290
6	Waste dump	BC	Eastern Vancouver Island	Metals, PAHs, PCBs[c], PHCs	3,700
7	Soak away pit	ON	St. Laurent Lowlands	CHCs [d]	70
8	Salt storage area	AB	Northern Continental Divide	Salts	760
9	Underground storage tank	PE	Prince Edward Island	Metals, PAHs, PHCs	1,100
10	Spilled fuel	NB	Maritime Lowlands	Metals, PAHs, PHCs	690
11	Chemical dump	ON	St. Laurent Lowlands	CHCs[d], Metals, PHCs	6,100
12	Underground storage tank	ON	Thunder Bay-Quetico	PHCs	240
13	Aboveground storage tank	MB	Aspen Parkland	PHCs	350
14	Underground storage tank	SK	Aspen Parkland	PHCs	3,200
15	Wastewater lagoon	SK	Aspen Parkland	PHCs	1
16	Pesticide dump	SK	Aspen Parkland	Metals, Phenols	40
17	Equipment dump	NU	Eureka Hills	Metals	1,200
18	Aboveground storage tank	NT	Tazin Lake Upland	PHCs	6
19	Waste dump	AB	Fescue Grassland	DDT[e], Metals, PAHs	600
20	Aboveground storage tank nest	NT	Tuktoyaktuk Coastal Plain	PHCs	400

[a] *BC* British Columbia, *ON* Ontario, *YT* Yukon, *AB* Alberta, *PE* Prince Edward Island, *NB* New Brunswick, *MB* Manitoba, *SK* Saskatchewan, *NU* Nunavut, *NT* Northwest Territories

[b] Refer to NRCan (2007)

[c] Polychlorinated biphenyls

[d] Chlorinated hydrocarbons

[e] Dichlorodiphenyltrichloroethane

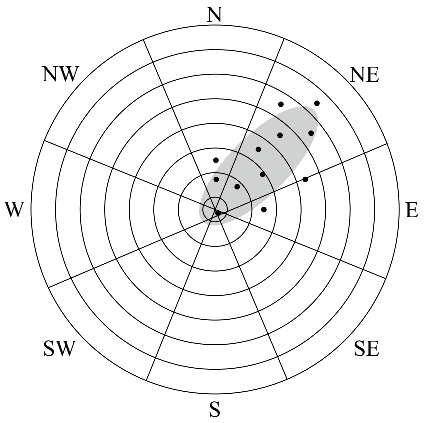


Fig. 12.1 Polar coordinate system

The origin was positioned within the contaminant source zone where the majority of contaminants were at their maximum concentrations. A site, *s*, was then divided into cells, *i*, each identified by its direction and radius from the origin. At each cell, soil was divided vertically into three strata where topsoil was defined as the top 0.3 m, surface soil as between 0.3 m and 1.5 m below ground surface (mbgs) and subsoil below 1.5 mbgs. The depth definitions of surface soil and subsoil were consistent with definitions proposed by CCME (2006a).

Each cell was assigned one of the following seven land uses to define the relevant receptors and exposure parameters (CCME 2006a; AENV 2010): natural area, agricultural, residential, parkland, commercial, industrial and water body. The selection of land uses for a site was based on information in the associated ESA reports, aerial imagery and land use zoning maps, where available.

12.4.2.2 Exposure Assessment and Hazard Assessment

The details of how exposure and hazard assessments were conducted were presented by Thiessen and Achari (2011) and are beyond the scope of this chapter. Briefly, regulatory documents published by AENV, CCME, Health Canada, ORNL and USEPA were referenced.

12.4.2.3 Risk Characterisation

At each cell, *i*, the hazard posed by each contaminant, *j*, to each receptor, *k*, via each exposure route, *m*, was expressed as a dimensionless hazard quotient, $HQ_{i,j,k,m,s}$, which is the measured or predicted dose divided by the tolerable dose (HC 2004a; CCME 1996). For the identified human and terrestrial animal receptors, this hazard quotient was calculated using (12.1) below. Equation (12.2) was applied for terrestrial plants, soil invertebrates and microorganisms. For aquatic plants and animals, Eqs. (12.3) and (12.4) were used:

$$HQ_{i,j,k,m,s} = \frac{PDI}{TDI}, \tag{12.1}$$

$$HQ_{i,j,k,m,s} = \frac{PSoC}{TSoC}, \tag{12.2}$$

$$HQ_{i,j,k,m,s} = \frac{PWC}{TWC}, \tag{12.3}$$

$$HQ_{i,j,k,m,s} = \frac{PSdC}{TSdC}. \tag{12.4}$$

The variables PDI and TDI are predicted daily intake and tolerable daily intake, typically expressed in contaminant mass per receptor mass per day. PSoC and TSoC are the predicted and tolerable soil concentrations, respectively, and quantified in contaminant mass per soil mass. Predicted sediment concentration, PSdC, and tolerable sediment concentration, TSdC, are similarly quantified. Predicted water concentration, PWC, and tolerable water concentration, TWC, have units of contaminant mass per water volume.

A hazard quotient does not communicate a probability of adverse effect but is a measure of potential adverse effect where a higher value means greater potential without quantifying that potential in an absolute sense. Typically, hazard quotients are used to characterise the impact of non-carcinogens to human receptors and risk values (e.g. 10^{-4} probability) used to characterise carcinogenic contaminants (HC 2004a). However, the health risk of a carcinogenic contaminant was also expressed as a hazard quotient solely for consistency in data comparison when discussing the results. The acceptable risk level of 10^{-5} (HC 2004a) was used as the denominator in carcinogenic hazard quotients.

Hazard quotients were summed across exposure routes for a receptor to obtain a hazard quotient at each cell for each contaminant, $HQ_{i,j,k,s}$ (USEPA 2005):

$$HQ_{i,j,k,s} = \sum_{m} HQ_{i,j,k,m,s}. \tag{12.5}$$

The maximum hazard quotients across all cells for each contaminant and receptor at a site, $\max HQ_{j,k,s}$, were used to characterise receptor risk:

$$max\, HQ_{j,k,s} = \max_i \left(HQ_{i,j,k,s} \right). \tag{12.6}$$

At each site, a set of these values was calculated for each of the eight receptor groups previously defined. The values in each set were Pareto ordered by largest to smallest and tabulated by receptor. The reason for ordering these values was to allow comparison of the largest value at one site to the largest value at another, the second largest at one to the second largest at another and so on. However, Pareto-ordering is not a requirement of partial order ranking described in Sect. 12.4.5 below. Hazard quotients less than 1×10^{-2} were deemed negligible and were set to zero.

12.4.3 Risk Perception Estimation

Risk perception research has largely focused in areas other than the risks associated with contaminated sites, as alluded in Sect. 12.3.2. Ideally, risk perception estimation should be derived from contaminated site perception data gathered via a focused survey. This survey should be designed to accommodate multivariate analysis via multidimensional scaling (MDS). This method is similar to common factor analysis because it seeks to find the underlying structure or dimensions to the data (Hair et al. 2010). However, its purpose is to find similarities or preferences across objects, which differs from common factor analysis. When used to clarify similarities, MDS orders objects along scales according to how alike they are to each other. In contrast, MDS can also be used to define preferences where objects are ranked according to several variables. In context of the current topic, the objects could be contaminated sites and the variables could be data dimensions similar to the latent factors identified by Slovic et al. (1980) and Lazo et al. (2000).

Unfortunately, relevant research articles focused on contaminated site risk perception are not currently available. In lieu and for illustrative purposes only, the authors assumed the role of the most directly affected stakeholders, which are contaminated site occupants and neighbours as per Sandman (1993), and arbitrarily rated each site against six dimensions (i.e. $D=6$). Those dimensions were denoted Dim1–Dim6 and ratings were integers between 1 and 5, inclusive. A rating of 1 indicated very low risk perception and 5, very high risk perception.

12.4.4 Environmental Liability Estimation

Environmental liability estimation adds one extra indicator and was limited to the cost of remediating the selected contaminated sites. For illustrative simplicity, the selected remediation method was contaminated soil excavation, disposal offsite at a nearby landfill and backfilling with non-contaminated soil. Contaminated soil volumes were determined from the environmental and soils information in the ESA reports and summarised in the last column of Table 12.3. Groundwater remediation was not included. A non-commercial version of AECOM's (2011) RACER™ software was applied to derive cost estimates using the built-in, default 2013 system cost database containing Canadian average costs. No site-specific cost estimating modifications were made.

12.4.5 Site Ranking

Partial order ranking (POR) was selected to rank sites according to environmental risk, societal perception and environmental liability. POR minimises bias by avoiding the need to choose attribute weights as is required in commonly used multicriteria

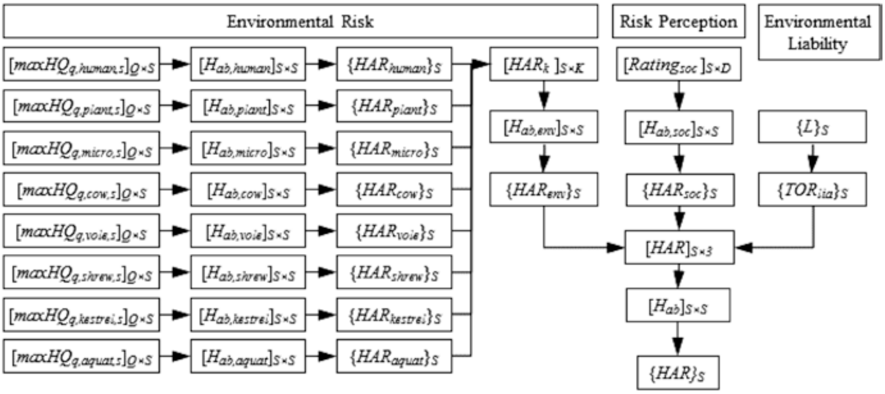


Fig. 12.2 HPOR procedure using environmental risk, maxHQ$_{q,k,s}$; perceptions and liability, L_s

analysis (MCA) methods. A limitation of POR is objects are ranked on an ordinal scale meaning the degree of difference between ranked objects cannot be determined as can be done with MCA.

HPOR as described by Carlsen (2008) and Brüggemann and Patil (2011) was first applied to environmental risk results because of the large number of attributes used to rank the sites: 48 Pareto-ordered, contaminant-specific hazard quotients for human receptors, 22 for plants and soil invertebrates, 6 for soil microorganisms, 15 for cows, 15 for meadow voles, 24 for masked shrews, 9 for American kestrels and 41 for aquatic life. The HPOR procedure for environmental risk results is summarised in Fig. 12.2. In the first step, a table of Pareto-ordered hazard quotients from across all sites for each receptor was prepared, [maxHQ$_{q,k,s}$]$_{Q\times S}$. Eight tables were prepared, where q is a contaminants position in the order and Q is the largest number of hazard quotients at any site (rows) and S is the number of sites evaluated (columns). For example, the table for human receptors had 48 ordered hazard quotients for 20 sites. Q varied depending on the receptor but S was constant (20 sites). Table 12.4 for human receptors is an example of the first step's output. Similar tables were prepared for the remaining seven receptors.

Second, a Hasse matrix summarising the rank relationship of every site to every other site, [$H_{ab,k}$]$_{S\times S}$, based on maxHQ$_{q,k,s}$ values was prepared for each of the eight receptors. Each entry to the matrix was determined using the following logic as defined by Mauri and Ballabio (2008):

$$H_{ab,k}(a)=\begin{Bmatrix} 1, & if\, maxHQ_{q,k}(a) \geq maxHQ_{q,k}(b), \forall q \in [1,Q] \\ -1, & if\, maxHQ_{q,k}(a) < maxHQ_{q,k}(b), \forall q \in [1,Q] \\ 0, & otherwise. \end{Bmatrix} \tag{12.7}$$

Table 12.4 Hazard quotients for humans, $[\mathrm{maxHQ}_{q,\mathrm{human},s}]_{Q\times S}$ (Thiessen and Achari 2012)

	Site & Pareto-ordered hazard quotients, $\mathrm{maxHQ}_{q,\mathrm{human},s}$																		
q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
48	3×10^{3}	1×10^{1}	2×10^{3}	3×10^{3}	1×10^{1}	2×10^{4}	9×10^{1}	3×10^{0}	3×10^{2}	2×10^{2}	2×10^{5}	6×10^{2}	2×10^{2}	4×10^{3}	2×10^{1}	3×10^{3}	9×10^{-2}	2×10^{0}	5×10^{2}
47	3×10^{2}	7×10^{0}	1×10^{3}	7×10^{2}	7×10^{0}	7×10^{2}	3×10^{1}	2×10^{0}	3×10^{1}	9×10^{1}	4×10^{3}	2×10^{1}	7×10^{1}	2×10^{1}	1×10^{0}	4×10^{1}	–	6×10^{-1}	4×10^{1}
46	2×10^{2}	4×10^{0}	2×10^{1}	9×10^{1}	5×10^{0}	5×10^{2}	2×10^{0}	–	7×10^{0}	3×10^{1}	2×10^{2}	1×10^{1}	6×10^{1}	2×10^{1}	3×10^{-1}	3×10^{1}	–	4×10^{-1}	2×10^{1}
45	2×10^{2}	1×10^{0}	3×10^{0}	7×10^{1}	3×10^{0}	5×10^{2}	–	–	3×10^{0}	2×10^{1}	1×10^{2}	9×10^{0}	4×10^{1}	7×10^{0}	–	2×10^{0}	–	1×10^{-1}	2×10^{1}
44	9×10^{1}	7×10^{-1}	2×10^{0}	3×10^{1}	3×10^{0}	5×10^{2}	–	–	1×10^{0}	9×10^{0}	9×10^{1}	3×10^{0}	3×10^{1}	6×10^{0}	–	9×10^{-1}	–	7×10^{-2}	1×10^{1}
43	9×10^{1}	2×10^{-1}	8×10^{-1}	3×10^{1}	1×10^{0}	1×10^{2}	–	–	9×10^{-1}	9×10^{0}	8×10^{1}	4×10^{-1}	2×10^{1}	2×10^{0}	–	2×10^{-1}	–	–	1×10^{1}
42	9×10^{1}	3×10^{-2}	–	2×10^{1}	1×10^{-0}	8×10^{1}	–	–	7×10^{-1}	7×10^{0}	1×10^{1}	2×10^{-1}	1×10^{0}	2×10^{0}	–	2×10^{-1}	–	–	7×10^{0}
41	7×10^{1}	1×10^{-2}	–	1×10^{1}	8×10^{-1}	8×10^{1}	–	–	7×10^{-1}	3×10^{0}	5×10^{0}	8×10^{-2}	8×10^{-1}	2×10^{0}	–	4×10^{-2}	–	–	3×10^{0}
40	5×10^{1}	–	–	9×10^{0}	6×10^{-1}	6×10^{1}	–	–	6×10^{-1}	2×10^{0}	8×10^{-1}	8×10^{-2}	6×10^{-1}	8×10^{-1}	–	–	–	–	2×10^{0}
39	5×10^{1}	–	–	8×10^{0}	2×10^{-1}	4×10^{1}	–	–	4×10^{-1}	1×10^{0}	2×10^{-2}	1×10^{-2}	5×10^{-2}	4×10^{-1}	–	–	–	–	2×10^{0}
38	4×10^{1}	–	–	4×10^{0}	1×10^{-1}	3×10^{1}	–	–	3×10^{-1}	9×10^{-1}	–	–	4×10^{-2}	2×10^{-1}	–	–	–	–	3×10^{-1}
37	4×10^{1}	–	–	3×10^{0}	5×10^{-2}	2×10^{1}	–	–	2×10^{-1}	5×10^{-1}	–	–	–	–	–	–	–	–	1×10^{-1}
36	3×10^{1}	–	–	3×10^{0}	4×10^{-2}	2×10^{1}	–	–	1×10^{-1}	2×10^{-1}	–	–	–	–	–	–	–	–	4×10^{-2}
35	2×10^{1}	–	–	2×10^{0}	1×10^{-2}	2×10^{1}	–	–	1×10^{-1}	4×10^{-2}	–	–	–	–	–	–	–	–	2×10^{-2}
34	2×10^{1}	–	–	7×10^{-1}	–	1×10^{1}	–	–	5×10^{-2}	1×10^{-2}	–	–	–	–	–	–	–	–	–
33	8×10^{0}	–	–	5×10^{-1}	–	1×10^{1}	–	–	3×10^{-2}	–	–	–	–	–	–	–	–	–	–
32	6×10^{0}	–	–	5×10^{-1}	–	1×10^{1}	–	–	1×10^{-2}	–	–	–	–	–	–	–	–	–	–
31	4×10^{0}	–	–	3×10^{-1}	–	1×10^{1}	–	–	1×10^{-2}	–	–	–	–	–	–	–	–	–	–
30	4×10^{0}	–	–	3×10^{-1}	–	9×10^{0}	–	–	–	–	–	–	–	–	–	–	–	–	–
29	2×10^{0}	–	–	2×10^{-1}	–	6×10^{0}	–	–	–	–	–	–	–	–	–	–	–	–	–

Sites having all hazard quotients less than 1×10^{-2} are not shown. Human receptors were not present at site 20 and thus hazard quotients for this site are zero

$H_{ab,k}$ (a) is the rank relationship of site a relative to site b based solely on risks to receptor k. A $H_{ab,k}$ entry of 1 resulted if the Pareto-ordered, maximum hazard quotients for all contaminants at site a were greater than or equal to the corresponding Pareto-ordered, maximum hazard quotients for all contaminants at site b. A value of −1 meant the Pareto-ordered maxHQ$_{q,k}$ values for site b dominated the corresponding values for site a. Zero was the result when neither of these conditions applied; sites a and b were incomparable. The resulting Hasse matrix described the partial order site ranks. The ranking software DART developed by TALETE (2007) was used to generate the Hasse matrices, although PyHasse is a more recently developed and flexible alternative.

Third, an estimated total order site ranking based on each receptor, $\{\mathrm{HAR}_k\}_S$, was estimated using the following Hasse average rank equation (Carlsen 2008):

$$HAR_k\left(a\right)=\left(S+1\right)\text{-}\left(sub\left(a\right)+1\right)\quad\left(S+1\right)/\left(S+1-Inc\left(a\right)\right). \tag{12.8}$$

HAR_k (a) is the Hasse average rank of site a and S is the number of sites being ranked, as defined previously. Sub(a) is the number of comparable sites, where site a is subordinate (i.e. the number of $H_{ab,k}$ values equal to −1) and Inc(a) the number of sites incomparable with site a (i.e. the number of $H_{ab,k}$ values equal to zero). Carlsen (2008) identified the epistemic uncertainty in using average ranks and acknowledged that a range of possible ranks is associated with each average rank. The lower and upper rank boundaries were defined by Brüggemann et al. (2004) in the following equations to calculate minimum and maximum ranks, $\mathrm{R}_{\mathrm{min},k}(a)$ and $\mathrm{R}_{\mathrm{max},k}(a)$, respectively:

$$R_{min,k}\left(a\right)=Sup\left(a\right)+1, \tag{12.9}$$

$$R_{max,k}\left(a\right)=sup\left(a\right)+Inc\left(a\right)+1. \tag{12.10}$$

Sup(a) is the number of comparable sites, where site a is superior (i.e. the number of $H_{ab,k}$ values equal to 1).

Fourth, the eight receptor-specific, Hasse average site ranks were combined into a table, $[\mathrm{HAR}_k]_{S\times K}$, where K is the number of receptors (i.e. eight). A single Hasse matrix, $[H_{ab,\mathrm{env}}]_{S\times S}$, was then prepared from this table to summarise the overall relationship of every site to every other site in terms of environmental risk. The following relationship was applied, which is similar to (12.7) above, but where $H_{ab,\mathrm{env}}(a)$ is the overall PQRA rank relationship of site a relative to site b:

$$H_{ab,env}\left(a\right)=\begin{cases}1, & if\,HAR_k\left(a\right)\geq HAR_k(b),\forall k\in[1,K]\\ -1, & if\,HAR_k\left(a\right)<HAR_k(b),\forall k\in[1,K]\\ & 0,\quad otherwise\end{cases} \tag{12.11}$$

Minimum, maximum and Hasse average ranks, $\{\mathrm{HAR}_{\mathrm{env}}\}_S$, for all sites in terms of environmental risk were then calculated in a manner similar to (12.2), (12.3) and (12.4).

Referring again to Fig. 12.2, a similar but abbreviated procedure was completed to determine average site ranks based on perception, $\{\mathrm{HAR}_{\mathrm{soc}}\}_S$, beginning with table of perception ratings, $[\text{Rating}_{\text{soc}}]_{S \times D}$. The subscript D is the number of risk perception characteristics (i.e. six). Total order site ranks, $\{\text{TOR}_{\text{lia}}\}_S$, were determined directly from the set of liability estimates, $\{L\}_S$.

Finally, the three sets of site ranks were combined, $[\text{HAR}]_{S \times 3}$; the corresponding Hasse matrix was determined, $[H_{ab}]_{S \times S}$ and overall average site rankings determined, $\{\text{HAR}\}_S$.

12.4.6 HPOR Limitations

A limitation of the HPOR procedure defined by Carlsen (2008) is that rank uncertainty propagation is not calculated. This is because only a measure of central tendency in site ranks is passed from one step to the next. Quantifying uncertainty propagation would require Monte Carlo simulation and a return to the computational challenges discussed by Lerche and Sorensen (2003) and Lerche et al. (2003).

12.5 Results and Discussion

As mentioned previously, Table 12.4 summarises the Pareto-ordered hazard quotients obtained from the PQRAs for human receptors. Similar tables were prepared for the other seven receptors. The results are presented with only one significant figure because of the large uncertainties in conducting environmental risk assessment. The PQRA protocol yields results with significant epistemic uncertainty, also called ignorance or subjective uncertainty, since it is a screening tool that uses sparse environmental data and approximate calculations to simplify reality. Aleatoric uncertainty, which is data variability and often called objective uncertainty, has secondary significance. The PQRA protocol is conservative because it tends to overestimate risk by applying reasonable worst case assumptions in contaminant transport and receptor exposure. Furthermore, the human and ecological receptor toxicological reference values applied in the PQRAs are generally very low to compensate for the epistemic uncertainty in their derivations. The intended bias is to consistently overestimate risk to compensate for significant epistemic uncertainty.

Table 12.5 describes stakeholder perceptions of the issues at each of the 20 contaminated sites. To reiterate, the ratings were fabricated by the authors to illustrate the concept of ranking site by combining environmental risk, societal perception and environmental liability.

Table 12.6 summarises estimated costs to remediate the sites using excavation and landfill disposal as the selected remediation option. Costs ranged over three orders of magnitude from CAN$2,000 to CAN$1,000,000. Similar to the PQRA results, costs are presented to one significant figure because of the large uncertainties in the soil excavation volumes listed in Table 12.3.

Table 12.5 Fabricated stakeholder perception ratings

	Site and stakeholder ratings[b]																			
Dimension[a]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dim1	3	1	3	3	1	4	3	3	2	3	3	2	1	1	1	3	3	1	4	2
Dim2	5	4	3	4	3	5	4	5	3	5	5	3	3	3	3	5	5	3	5	5
Dim3	4	1	2	2	1	4	4	3	1	5	5	1	1	1	2	2	2	1	3	3
Dim4	4	1	2	1	1	4	3	3	2	4	5	1	1	1	1	2	1	1	3	2
Dim5	1	1	1	1	1	1	2	1	1	1	4	1	1	1	1	1	1	3	1	1
Dim6	1	1	1	1	1	1	3	1	1	1	4	1	1	1	2	2	1	3	1	1

[a] Dimension are arbitrary for illustration only

[b] Five point scale; 1 = very low risk perception; 5 = very high risk perception

Table 12.6 Remediation excavation cost estimates, L_s

Site	Estimate ($), L_s
1	1×10^{6}
2	1×10^{4}
3	4×10^{4}
4	2×10^{5}
5	3×10^{4}
6	5×10^{5}
7	4×10^{4}
8	6×10^{4}
9	2×10^{5}
10	1×10^{5}
11	9×10^{5}
12	2×10^{4}
13	3×10^{4}
14	4×10^{5}
15	2×10^{3}
16	1×10^{4}
17	1×10^{5}
18	8×10^{3}
19	5×10^{4}
20	3×10^{4}

Tables 12.7 and 12.8 are the Hasse matrices summarising the pair-wise comparison of sites and site rank statistics based on the environment risk alone and then environmental risk, stakeholder perception plus environmental liability. In both tables, there is a noticeable uncertainty in the Hasse average ranks as shown by the wide ranges described by the minimum and maximum ranks. This uncertainty is due to the large number of incomparable sites, expressed in the matrices as a zero. This highlights a central concept in contaminated site prioritisation and management: decisions are made using sparse, imprecise data and must be made with a clear understanding of the associated uncertainties.

In Fig. 12.3, site ranks using environmental risk, perception and liability combined, *HAR*, were compared to ranks according to environmental risks alone, HAR_{env}. The number adjacent to each data point is the site identifier. The diagonal, dashed

Table 12.7 Hasse matrix, $[H_{ab,\text{env}}]_{S\times S}$, and rank statistics using environmental risks, $\text{maxHQ}_{q,k,s}$ (Thiessen and Achari 2012)

Site *a*	…Site *b*																				Rank statistics		
compared to…	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	$R_{\text{env,min}}$	$R_{\text{env,max}}$	HAR_{env}
1	–	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	1	0	1	14	20	19.6
2	−1	–	−1	−1	−1	−1	0	0	−1	−1	−1	0	−1	0	1	−1	0	1	−1	0	3	9	4.2
3	−1	1	–	0	1	−1	1	0	1	0	0	1	0	0	1	0	1	1	0	1	10	18	16.2
4	−1	1	0	–	0	−1	1	0	1	0	0	1	0	0	1	0	1	1	0	0	8	18	15.3
5	−1	1	−1	0	–	−1	1	0	0	0	0	0	−1	0	1	−1	0	1	−1	0	5	14	8.8
6	0	1	1	1	1	–	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	20	20.0
7	−1	0	−1	−1	−1	−1	–	0	−1	−1	−1	−1	−1	−1	1	−1	0	1	−1	0	3	7	3.7
8	−1	0	0	0	0	−1	0	–	0	0	0	0	0	0	0	0	0	0	0	0	1	18	5.3
9	−1	1	−1	−1	0	−1	1	0	–	0	−1	1	0	0	1	0	0	1	−1	0	6	14	9.7
10	−1	1	0	0	0	−1	1	0	0	–	0	1	0	0	1	0	0	1	0	0	6	18	14.0
11	0	1	0	0	0	−1	1	0	1	0	–	1	0	1	1	0	1	1	0	1	10	19	17.5
12	−1	0	−1	−1	0	−1	1	0	−1	−1	−1	–	−1	0	1	−1	0	1	−1	0	4	10	5.6
13	0	1	0	0	1	−1	1	0	0	0	0	1	–	0	1	0	0	1	0	1	8	19	16.8
14	0	0	0	0	0	−1	1	0	0	0	−1	0	0	–	1	0	0	1	0	0	4	18	12.0
15	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	−1	–	−1	0	1	−1	0	2	5	2.3
16	0	1	0	0	1	−1	1	0	0	0	0	1	0	0	1	–	0	1	0	0	7	19	16.3
17	−1	0	−1	−1	0	−1	0	0	0	0	−1	0	0	0	0	0	–	0	−1	0	1	14	2.6
18	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	−1	−1	−1	0	–	−1	0	1	4	1.2
19	0	1	0	0	1	−1	1	0	1	0	0	1	0	0	1	0	1	1	–	1	10	19	17.5
20	−1	0	−1	0	0	−1	0	0	0	0	−1	0	−1	0	0	0	0	0	−1	–	1	14	2.6

Table 12.8 Hasse matrix, $[H_{ab}]_{S \times S}$, and rank statistics using environmental risks, perceptions and liabilities

Site *a* compared	…Site *b*																				Rank statistics		
to…	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	$R_{\min}$	$R_{\max}$	HAR
1	–	1	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	0	1	14	20	19.6
2	−1	–	−1	−1	0	−1	0	−1	−1	−1	−1	0	0	0	0	−1	0	0	−1	0	1	10	1.8
3	−1	1	–	0	1	−1	0	0	0	0	−1	1	0	0	0	0	0	0	−1	0	4	16	9.3
4	−1	1	0	–	1	−1	0	0	1	0	−1	1	0	0	0	0	0	0	0	0	5	17	11.7
5	−1	0	−1	−1	–	−1	0	0	−1	−1	−1	0	−1	−1	0	0	0	0	−1	0	1	10	1.8
6	0	1	1	1	1	–	1	1	1	1	0	1	1	1	1	1	1	1	1	1	18	20	19.9
7	0	0	0	0	0	−1	–	0	0	−1	−1	0	0	0	1	0	0	1	−1	1	4	16	9.3
8	−1	1	0	0	0	−1	0	–	0	−1	−1	0	0	0	1	0	0	0	0	1	4	16	9.3
9	−1	1	0	−1	1	−1	0	0	–	0	−1	1	0	0	0	0	0	0	0	0	4	16	9.3
10	0	1	0	0	1	−1	1	1	0	–	−1	1	0	0	1	0	1	1	0	1	10	18	16.2
11	0	1	1	1	1	0	1	1	1	1	–	1	1	1	1	1	1	1	1	1	18	20	19.9
12	−1	0	−1	−1	0	−1	0	0	−1	−1	−1	–	0	0	0	0	0	0	−1	0	1	12	2.1
13	−1	0	0	0	1	−1	0	0	0	0	−1	0	–	0	0	0	0	0	−1	0	2	16	6.0
14	−1	0	0	0	1	−1	0	0	0	0	−1	0	0	–	0	0	0	0	0	0	2	17	7.0
15	−1	0	0	0	0	−1	−1	−1	0	−1	−1	0	0	0	–	−1	0	0	−1	−1	1	11	1.9
16	0	1	0	0	0	−1	0	0	0	0	−1	0	0	0	1	–	0	1	−1	0	4	17	10.5
17	−1	0	0	0	0	−1	0	0	0	−1	−1	0	0	0	0	0	–	0	0	0	1	16	3.5
18	−1	0	0	0	0	−1	−1	0	0	−1	−1	0	0	0	0	−1	0	–	−1	0	1	13	2.3
19	0	1	1	0	1	−1	1	0	0	0	−1	1	1	0	1	1	0	1	–	1	11	18	16.5
20	−1	0	0	0	0	−1	−1	−1	0	−1	−1	0	0	0	1	0	0	0	−1	–	2	13	4.2

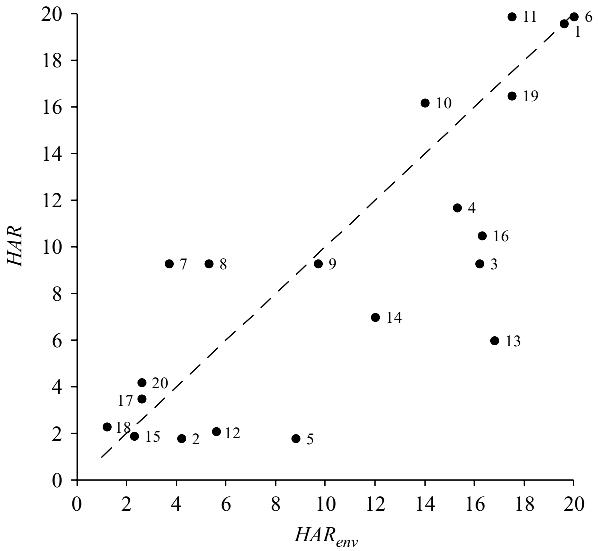


Fig. 12.3 Comparison of overall site ranks, HAR, to environmental risk ranks, $\mathrm{HAR}_{\mathrm{env}}$

line shows where the points would be if there was perfect correlation. The significant data scatter shows how including perception and environmental liability considerations can influence site prioritisation.

12.6 Conclusions and Recommendations for Future Work

A proposed contaminated site prioritisation method that considers environmental risk, public perception and financial liability was presented in this chapter. The results illustrate how perception and liability influence the prioritisation compared to a prioritisation based on environmental risk alone. Furthermore, the results highlight the degree of uncertainty often involved when making decisions about contaminated site management.

Refinement and further research are needed to develop the proposed method into a practical tool that will be valuable to site occupants, property owners, regulatory agencies, financial lenders and insurers and contaminated site professionals. First, environmental risk estimation could be simplified and refined by summing receptor-specific hazard quotients for chemicals causing similar adverse effects (e.g. carcinogens). This would reduce the number of environmental risk attributes per site and thus reduce the potential number of incomparable sites in the Hasse matrices. A result would be less uncertainty in site rankings. Second, the environmental risk estimation method applied here could be broadened to allow estimates to be made when no intrusive investigation has been completed but only desktop studies and site interviews (e.g. Phase 1 ESA). Contaminated site portfolio owners often need to prioritise funds on intrusive investigations to determine whether environmental risks are present or to understand the magnitude of those risks once discovered. Third, research in understanding stakeholders' perceptions of contaminated sites and contaminants needs to be advanced with the objective of satisfying the needs of practical applications. The concept of quantifying perceptions as a multiplier of environmental risk results should be investigated. Fourth, epistemic and aleatoric uncertainty in data and modelling should be quantified explicitly so that decision makers understand the true uncertainty in the output. Both probabilistic and possibilistic approaches could be applied as supported by the data. These recommendations will be a challenge to achieve, and overall, model advancement should be parsimonious with the users' specific requirements as primary research objectives.

References

AECOM (AECOM Technology Corporation) (2011) Remedial Action Cost Engineering and Requirements (RACER) System, Version 11.1.12.0. Retrieved from http://www.fecpractice.com/?p=RACER

AENV (Alberta Environment) (2010) Alberta Tier 1 Soil and Groundwater Remediation Guidelines. Edmonton, AB. Retrieved from http://environment.gov.ab.ca/info/library/7751.pdf

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, New York, NY

Brüggemann R, Sorensen PB, Lerche D, Carlsen L (2004) Estimation of averaged ranks by a local partial order model. J Chem Inf Comput Sci 44(2):618–625. doi:10.1021/ci034214m

Carlon C, Hope B, Quercia F (2009) Contaminated land: a multi-dimensional problem. In: Marcomini A, Suter GW II, Critto A (eds) Decision support systems for risk-based management of contaminated sites. Springer, New York, NY, pp 113–135

Carlsen L (2008) Hierarchical partial order ranking. Environ Pollut 155:247–253. doi:10.1016/j.envpol.2007.11.023

CCME (Canadian Council of Ministers of the Environment) (1996) A framework for ecological risk assessment: general guidance. Winnipeg, MB. Retrieved from http://www.ccme.ca/assets/pdf/pn_1195_e.pdf

CCME (Canadian Council of Ministers of the Environment) (1997) A framework for ecological risk assessment: technical appendices. Winnipeg, MB. Retrieved from http://www.ccme.ca/assets/pdf/pn_1274_e.pdf

CCME (Canadian Council of Ministers of the Environment) (2006a) A protocol for the derivation of environmental and human health soil quality guidelines. Winnipeg, MB. Retrieved from http://www.ccme.ca/assets/pdf/sg_protocol_1332_e.pdf

CCME (Canadian Council of Ministers of the Environment) (2006b) Recommended principles on contaminated site liability. Winnipeg, MB. Retrieved from http://www.ccme.ca/assets/pdf/csl_14_principles_e.pdf

CCME (Canadian Council of Ministers of the Environment) (2008) National classification system for contaminated sites: guidance document. Winnipeg, MB. Retrieved from http://www.ccme.ca/assets/pdf/pn_1403_ncscs_guidance_e.pdf

EC (Environment Canada) (2012, October 4) Harper government launches Phase II of contaminated sites clean-up plan, 4 Oct 2012. Retrieved from Environment Canada. http://www.ec.gc.ca

Goosen H, Janssen R, Vermaat JE (2007) Decision support for participatory wetland decision-making. Ecol Eng 30(2):187–199. doi:10.1016/j.ecoleng.2006.11.004

Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis, 7th edn. Prentice Hall, Upper Saddle River, NJ

HC (Health Canada) (2004a) Federal contaminated site risk assessment in Canada, part I: guidance on human health preliminary quantitative risk assessment (PQRA). Ottawa, ON

HC (Health Canada) (2004b) Federal contaminated site risk assessment in Canada, part III: guidance on peer review of human health risk assessments for federal contaminated sites in Canada. Ottawa, ON

Kasperson JX, Kasperson RE, Pidgeon N, Slovic P (2003) The social amplification of risk: assessing fifteen years of research and theory. In: Pidgeon N, Kasperson RE, Slovic P (eds) The social amplification of risk. Cambridge University Press, Cambridge, pp 13–46

Lazo JK, Kinnell JC, Fisher A (2000) Expert and layperson perceptions of ecosystem risk. Risk Anal 20(2):179–193

Lerche D, Sorensen PB (2003) Evaluation of the ranking probabilities for partial orders based on random linear extensions. Chemosphere 53:981–992. doi:10.1016/S0045-6535(03)00558-7

Lerche D, Sorensen PB, Brüggemann R (2003) Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability. J Chem Inf Comput Sci 43:1471–1480. doi:10.1021/ci0300036

Mauri A, Ballabio D (2008) Similarity/diversity measure for sequential data based on Hasse matrices; theory and applications. In: Pavan M, Todeschini R (eds) Scientific data ranking methods: theory and applications. Elsevier, Oxford, UK

McDaniels T, Axelrod LJ, Slovic P (1995) Characterizing perception of ecological risk. Risk Anal 15(5):575–588

NRC (National Research Council) (1983) Risk assessment in the Federal government: managing the process. National Academy Press, Washington, DC. Retrieved from http://www.nap.edu/openbook.php?isbn=0309033497

NRC (National Research Council) (1993) Issues in risk assessment: a paradigm for ecological risk assessment. National Academy Press, Washington, DC. Retrieved from http://www.nap.edu/openbook.php?record_id=2078&page=241

NRCan (Natural Resources Canada) (2007) The Atlas of Canada-Ecological Framework. Retrieved from http://atlas.nrcan.gc.ca

OAG (Office of the Auditor General of Canada) (2012) Commissioner of the environment and sustainable development's opening statement, 2012 Spring report press conference, 8 May 2012. Retrieved from Office of the Auditor General of Canada. http://www.oag-bvg.gc.ca

Pavan M, Todeschini R (eds) (2008) Scientific data ranking methods: theory and applications. Elsevier, Oxford, UK

Sandman PM (1993) Responding to community outrage. American Industrial Hygiene Association, Falls Church, VA. Retrieved from http://www.psandman.com/media/RespondingtoCommunityOutrage.pdf

Slovic P, Fischhoff B, Lichtenstein S (1980) Perceived Risk. In: Schwing RC, Albers WA (eds) Societal risk assessment: how safe is safe enough? Plenum, New York, NY

TALETE (2007) DART – decision analysis by ranking techniques, Version 2.0.5. Retrieved from http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/DART

TBCS (Treasury Board of Canada Secretariat) (2011) Federal contaminated sites inventory, 24 Mar. Retrieved from Treasury Board of Canada Secretariat. http://www.tbs-sct.gc.ca/fcsi-rscf/home-accueil-eng.aspx

TBCS (Treasury Board of Canada Secretariat) (2012) Treasury Board Accounting Standard 1.2 – Departmental and Agency Financial Statements, 4 Oct. Retrieved from Treasury Board of Canada Secretariat. http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=18830&section=text

Thiessen RJ, Achari G (2011) A comparison of 2008 National Classification System for Contaminated Sites scores to preliminary quantitative risk assessment hazard quotients. Can J Civil Eng 38(7):719–728. doi:10.1139/l11-039

Thiessen RJ, Achari G (2012) Can the National Classification System for Contaminated Sites be used to rank sites? Can J Civil Eng 39(4):415–431. doi:10.1139/l2012-015

USEPA (United States Environmental Protection Agency) (1976) Interim procedures and guidelines for health risk and economic impact assessments of suspected carcinogens. Washington, DC

USEPA (United States Environmental Protection Agency) (1992) Framework for ecological risk assessment. Washington, DC. Retrieved from http://rais.ornl.gov/documents/FRMWRK_ERA.PDF

USEPA (United States Environmental Protection Agency) (2005) Human health risk assessment protocol for hazardous waste combustion facilities. Washington, DC. Retrieved from http://www.epa.gov/osw/hazard/tsd/td/combust/risk.htm

USEPA (United States Environmental Protection Agency) (2012) The history of risk at EPA, 31 July. Retrieved from USEPA. http://epa.gov/risk/history.htm

Chapter 13
Evaluating Ranking Robustness in Multi-indicator Uncertain Matrices: An Application Based on Simulation and Global Sensitivity Analysis

Claudio M. Rocco and Stefano Tarantola

Abstract Multi-indicator matrices represent a set of objects or alternatives characterized simultaneously by several criteria or attributes. In many situations, a decision-maker is interested in assessing each object, by considering simultaneously all criteria, and defining a ranking able to synthesize the global characteristic of each object, for example, from best to worst. However the assessment could be influenced by uncertain factors. For example, the cost of a project could be affected by variations in the interest rate. The effects of such variations could affect the initial or base rank. In this chapter, the robustness of the base rank is analyzed. The first part analyzes how the uncertainty in the numerical value of the criteria associated with each objects affects its rank. Additionally, it proposes some ideas for assessing the rank robustness. The second part proposes the use of global sensitivity analysis to assess the importance of each uncertain factor on, for example, the base rank. An example related to a real portfolio management, using three techniques that do not require additional preference parameters, is presented.

13.1 Introduction

In many situations objects are characterized by several criteria or attributes. For example, engineering projects could be represented by cost, availability, and environmental impact, among others. Each criterion is quantified via performance values (PV), which can be either numerical or categorical, but in any case ordinal.

C.M. Rocco (✉)
Facultad de Ingeniería, Universidad Central de Venezuela, Caracas, Venezuela
e-mail: croccoucv@gmail.com

S. Tarantola
Institute of the Protection and Security of the Citizen,
JRC, European Commission, Ispra, Italy

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_13, © Springer Science+Business Media New York 2014

This information is typically structured in a matrix, where rows represent objects and columns are associated with several criteria.

A typical problem faced by a decision-maker (DM) is to define a ranking of such objects, able to synthesize the global characteristic of each one. The idea of ranking objects or alternatives is based on one of the four discrete decision-making problems defined as "*Problematique* γ" in Roy (1985), that is, ranking the alternatives from the best to the worst ones.

In general, such assessment is performed as follows:

1. Define n objects and m criteria (or attributes)
2. Define the multi-indicator matrix Q, based on each PV_{ij} (for each alternative $i=1, \ldots, n$ and for each criterion $j=1, \ldots, m$)
3. Select a ranking technique (RT)
4. Produce a rank of alternatives

However when objects are characterized by several criteria, their evaluations could lead to partial order, since an object "*a*" is not always better than an object "*b*," considering the set of given attributes. In other words, it is very difficult to obtain a complete order.

Ranking techniques to generate the desired rank are classified as parametric and nonparametric. The first group requires information about decision-maker preferences (e.g., criterion weights), while nonparametric techniques do not use such information. Widely used parametric techniques include (Dorini et al. 2011):

- ELECTRE (I, II, III, IV, and TRI) series methods (Roy 1968)
- PROMETHEE—Preference Ranking Organization METHod for Enrichment Evaluations (Brans and Vincke 1985)
- The Analytical Hierarchy Process (Saaty 1980)
- TOPSIS—Technique for Order Preference by Similarity to Ideal Solution (Hwang and Yoon 1981)

Nonparametric ranking techniques include, among others, partial order ranking (Brüggemann et al. 1999), Hasse diagram technique (Brüggemann et al. 1995), and Copeland Scores (Al-Sharrah 2010).

The use of these ranking techniques allows ranking the alternatives from best to worst. This ranking is defined as the base rank (BR). However, no matter which assessment technique is selected, each PV_{ij} could be influenced by uncertain factors. For example, the cost of a project could be affected by variations in the interest rate. If these variations are represented as a probability distribution function then the rank of each alternative could be considered as a random variable.

In the literature, there are several methods proposed for assessing the impact of uncertainties. Some techniques consider only uncertainties in the decision-maker preferences [e.g., by varying the weight associated with each criterion (Rios Insua and French 1991; Wolters and Mareschal 1995; Yu et al. 2012)]. Other approaches are available for quantifying the impact of PV_{ij} on the ranking of alternatives (Triantaphyllou and Sanchez 1997; Yu et al. 2012). These approaches are in general limited since they focus on the variation of one parameter at a time. Applications are usually illustrated using parametric multi-criteria methods.

The first part of this chapter (*P1*) extends previous works by Hyde et al. (2004), Hyde and Maier (2006), and Yu et al. (2012) for analyzing the uncertainty problem, i.e., how the uncertainty in the PV_{ij} (the input) is propagated or affects the object ranks (the output). The approach, based on Monte Carlo simulation, allows answering several questions regarding ranking robustness. For example, the decision-maker could be interested in knowing whether the rank obtained for a specific alternative "a" in the base case is "maintained" no matter how the PV_{ij} values are affected by uncertainties. Alternatively, the decision-maker could be interested in knowing how the rank of a specific object varies when uncertainties are considered.

The second part (*P2*) proposes the use of a global sensitivity analysis technique, the Morris method (Morris 1991; Saltelli et al. 2000), to assess the importance of uncertainties in the PV on, for example, the base ranking.

To the best of our knowledge, these types of assessments have not been reported in the literature, at least under multi-indicator matrices.

The contribution of the approaches *P1* and *P2* is illustrated using three ranking techniques: two techniques related to the evaluation of partial order sets (or *posets*) and the third known as the Copeland Scores (CS) technique. All the ranking techniques selected are considered as nonparametric techniques since object ranking is produced by considering the data matrix Q alone, that is, with no additional information (e.g., the decision-maker preferences). It is important to mention that for these specific ranking techniques, there are sensitivity-based studies that will be referenced in the following sections.

The rest of the chapter is organized as follows: Section 13.2 describes the uncertainty analysis, while Sect. 13.3 presents an overview of global sensitivity analysis. The nonparametric evaluation techniques are presented in Sect. 13.4 while Sect. 13.5 describes a case study. Finally, Sect. 13.6 shows the conclusions and future work.

13.2 Uncertainty Analysis

As defined by Morgan and Henrion (1992), uncertainty analysis is an approach for assessing how the uncertainty in the inputs is propagated to the outputs. In our case, performance values PV_{ij} are considered as inputs whose uncertainties are modeled as random variables properly characterized through known probability distribution functions (pdf). The output is the ranking position of each alternative.

For this analysis, a Monte Carlo simulation approach is proposed:

1. Random deviates are generated for each PV_{ij}, according to its known probability distribution function. These set of values define a multi-indicator matrix Q sample.
2. A ranking of alternatives is obtained using Q and the ranking technique selected.

Steps 1 and 2 are repeated NSAMPLE times. The result of the Monte Carlo approach provides an approximated probability distribution function of the rank position of each alternative.

This analysis allows answer for a given alternative "a," for example:

(a) What is the probability that the base rank position is maintained?
(b) Which is the rank position with the highest probability?
(c) What are the possible rank positions and their corresponding probability?

13.3 Sensitivity Analysis

13.3.1 Introduction

Sensitivity analysis (SA) is "the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input factors" (Saltelli et al. 2000). In our case, the input factors $\mathbf{X}=\{x_1, x_2, \ldots, x_k)$ are the performance values stored in the multi-indicator matrix Q (i.e., $k=n\times m$) and the output is the ranking of the alternatives obtained by a specific ranking technique.

Sensitivity analysis methods can be classified as local and global. Local methods estimate partial derivatives of the model and determine the effects of the uncertainty in one factor on the model output while the other factors are fixed to a specific value (also referred to as one factor at a time (OAT) methods). These methods do not address the possibility of interactions among factors.

Global methods evaluate the effect of a factor while all the others are varying as well, thus allowing the exploration of the multidimensional input space. Global SA methods also include screening methods [e.g., the Morris technique (Morris 1991)], which produce a "list of factors ranked in order of decreasing importance" (Saltelli et al. 2004). On the other hand, techniques such as FAST, extended FAST, and Sobol's (Saltelli et al. 2004) are able to quantify the importance of the factors through variance decomposition techniques, though requiring more model evaluations than the screening techniques.

Variance decomposition techniques are capable of computing the main effects of a factor and the so-called total sensitivity indices, which jointly capture the single effect and the interaction terms (of any order) involving that factor. In the chapter, the screening method of Morris is selected as the technique for analyzing the case study.

It is important to realize that in the context defined by the multi-indicator matrix, two different sensitivity analyses can be performed. The first assesses the effects of uncertainty in Q or in parameters associated with the criteria (e.g., weights, in the case of parametric ranking techniques) and is known as indicator value-related sensitivity (Annoni et al. 2011). As mentioned in the introduction, previous works on sensitivity analysis are considered as local sensitivity methods, since they focus on the variation of one parameter at a time. However, Annoni et al. (2011) present a global sensitivity analysis for a specific ranking technique, described in Sect. 13.4.1. The second type of sensitivity analysis evaluates the effects of not considering a single attribute in the matrix Q. The analysis performed in this chapter is related to the first type of sensitivity analysis.

13.3.2 The Method of Morris

The method of Morris is able to "screen" a subset of a few important inputs among the many often contained in models (Morris 1991). Input factors can assume specific values (called levels) within their range of variation. The Morris method is useful for detecting factors having negligible effect on the output, having linear effect, or interacting with other factors.

The Morris method calculates elementary effects in the form of incremental ratios computed at different points in the space of the input and averaged over the same space. Since the exploration of the input space is performed in several regions, the method is considered as global (Saltelli et al. 2004).

The *Elementary Effect* for the *i*th input in a point $\mathbf{X}^0$ is calculated as

$$EE_i(x_1^0, \quad x_2^0) = \frac{y(x_1^0, x_2^0, \quad , x_{i-1}^0, x_1^0 + \Delta, x_{i+1}^0, \quad , x_k^0) - y(x_1^0, \quad x_k^0)}{1} \qquad (13.1)$$

where y is the model output in k factors and Δ is a given increment. In this chapter, the model output is evaluated through several ranking techniques, which are then considered as *black-box* models.

The EE_i is an OAT measure of sensitivity for the *i*th factor as it varies only one coordinate at a time. The method computes EE_i at r different points $\mathbf{X}^1, \ldots, \mathbf{X}^r$ and then takes the average as an indicator of importance for the *i*th factor. A modification of the method (Campolongo et al. 2007) estimates $\mu^*(x_i)$, as the average of the absolute value of the EE_i $\mu^*(x_i)$ is useful to identify inputs that are not responsible for output variation, while the standard deviation σ of the EE_i is a measure of the overall interactions of factor x_i with other inputs. As suggested by Saltelli et al. (2004), $\mu^*(x_i)$ is useful when the goal of the analysis is "factor prioritization." Recently, the method has been ameliorated by generalizing the sampling design to estimate the $\mu^*(x_i)$ (Campolongo et al. 2011). The design proposed coincides with the classic design used for estimating Sobol' indices.

The total number of model evaluations required by the Morris technique is $r(k+1)$ (r, the number of elementary effects computed per factor, is in the range 4–10 (Saltelli et al. 2000); k is the number of factors). Even if the method does not estimate the main effects of factors, it is a quick and reliable method capable of detecting negligible factors.

13.4 Ranking Techniques Considered

13.4.1 Average Rank Derived from Partial Order Sets

Let P define a set of n objects (for example, alternatives) to be analyzed and let the descriptors, $q_1, q_2, \ldots, q_m$ define m different attributes or criteria selected to assess the objects in P (for example, cost, availability, environmental impact). It is important that attributes are defined to reflect, for example, that a low value indicates low rankings, while a high value indicates high ranking (Restrepo et al. 2008).

If only one descriptor is used to rank the objects, then it is possible to define a total order in P. In general, given $x,\ y \in P$, if $q_i(x) \leq q_i(y)\ \forall\, i$, then x and y are said to be comparable. However, if two descriptors are used simultaneously, the following could happen: $q_1(x) \leq q_1(y)$ and $q_2(x) > q_2(y)$. In such case x and y are said to be incomparable (denoted by $x \| y$). If several objects are mutually incomparable, set P is called a partially ordered set or *poset*. Note that since comparisons are made for each criterion, no normalization is required.

The objects in a poset can be represented by a directed acyclic graph whose vertices are the objects $\in P$ and there is an edge between objects x and y if they are comparable. Such graph is termed a Hasse diagram (HD) (Brüggemann et al. 1995). All Hasse diagram plots presented in this chapter are obtained using the software DART 2.05.

[http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/qsar_tools/DART].

Hasse diagram is then a nonparametric ranking technique and can perform ranking decisions from the information available without using any aggregation criterion. However, in general Hasse diagram could not provide a total order of objects but gives an interesting overall picture of the relations among objects.

A useful approach to produce a ranking is based on the concept of the average rank of each object in the set of linear extensions of a poset (De Loof et al. 2011). Since the algorithms suggested for calculating such average ranks are exponential in nature (De Loof et al. 2011), special approximations have been developed, such as the local partial order model (LPOM) (Brüggemann et al. 2004), the extended LPOM (LPOMext) (Brüggemann and Carlsen 2011), or the approximation suggested in (De Loof et al. 2011). In this chapter, the LPOM and LPOMext average ranks will be used.

From the Hasse diagram, several sets could be derived. If $x \in P$ (Brüggemann and Carlsen 2011):

1. $U(x)$: The set of objects incomparable with x: $U(x) := \{y \in P : x \| y\}$
2. $O(x)$, the down set: $O(x) := \{y \in P : y \leq x\}$
3. $S(x)$, the successor set: $S(x) := O(x) - \{x\}$
4. $F(x)$, the up set: $F(x) := \{y \in P : x \leq y\}$

Then, the following average rank indexes are defined:

(a) $$LPOM(x) = \left(|S(x)| + 1\right) \quad (n+1) / \left(n + 1 - |U(x)|\right)$$

(b) $$LPOMext(x) = |O(x)| + \sum_{y \in U(x)} \frac{p_y^<}{p_y^< + p_y^>}$$

where n is the number of objects,
$|V|$ defines the cardinality of a generic set V

$$p_y^< = |O(x) \cap U(y)|, p_y^> = |F(x) \cap U(y)|, y \in U(x)$$

At this point it is important to mention that in Sørensen et al. (2000) the authors analyze the influence of data uncertainty on the partial order ranking and define a robustness parameter. Brüggemann and Patil (2011) assess the impact of removing a criterion by analyzing the structure of the Hasse diagram. Also, Annoni et al. (2011) analyze the effects on the structure of the Hasse diagram, due to uncertainty in the data matrix Q, using a global sensitivity analysis.

13.4.2 Copeland Score

The Copeland Score is a simple nonparametric ranking technique that has been applied outside of its usual political environment (voting) to rank objects in the sciences. Al-Sharrah (2010) shows that this method facilitates the analysis of large partially ordered sets, since it avoids the disadvantages of the Hasse diagram or the linear extension approach usually employed to resolve this issue. The method selects the alternative with the largest Copeland Score, which is the number of times an alternative is better than other alternatives minus the number of times that the alternative is worse than other alternatives, when they are compared pair-wise for each criterion. As in the previous technique, comparisons are made for each criterion and no normalization is required. Copeland Scores also assume that each criterion has equal importance.

The method builds a comparison matrix C. Each position $C(i,k)$ represents the count of comparison between alternative i and alternative k, considering each criterion q_j. If $q_j(i) \geq q_j(k)$, then $C(i,k) = C(i,k) + 1$. If $q_j(i) \leq q_j(k)$, then $C(i,k) = C(i,k) - 1$. Summing up $C(i,k)$ over all objects $(1 \leq k \leq n)$ yields the CS(i) of alternative i. Objects are then ranked using the corresponding CS(i). As mentioned in Al-Sharrah (2010), Copeland Scores could be also used as a categorized ranking tool, that is, to cluster objects.

In Al-Sharrah (2010), the author assesses the robustness of Copeland Scores ranking using several multi-indicator matrices. In Al-Sharrah (2011), the author compares Copeland Scores with other ranking techniques.

13.5 Case Study

The case study analyzed here corresponds to a real situation in portfolio management (Hernandez 2006) faced by policy makers, central and developing banks, and financial and economics ministries. The example is related to a Venezuelan case (Badillo 2010), where a set of projects (public investments) are used by decision-maker and analysts to address the growth and development of the economy, like refineries, bridges, and petroleum exploration, among others. The decision-maker and their representatives collected the data and estimated the impacts by using their own prospective model, which includes financial, economic, social, and environmental aspects.

Table 13.1 Performance values for the projects considered (base case)

PR	REO	EE	PEO	IA	PR	REO	EE	PEO	IA
1	521,855	504,692	106,619	–0.80	11	539,168	519,548	107,252	–0.40
2	407,903	394,287	84,373	–0.80	12	836,229	795,506	161,048	–0.80
3	680,613	654,939	130,118	–0.60	13	597,126	593,121	121,416	–0.80
4	533,198	509,258	111,742	–0.40	14	728,751	771,938	134,190	–0.40
5	3,508,350	3,935,177	582,314	–0.60	15	5,058,332	5,714,542	827,703	–0.20
6	581,041	561,471	119,371	–0.40	16	550,751	522,426	114,819	–0.40
7	817,653	778,248	158,007	–0.60	17	644,306	614,408	126,006	–0.20
8	902,646	888,821	168,828	–0.20	18	1,084,835	1,167,140	194,516	–0.40
9	2,429,447	2,275,772	421,882	–0.20	19	1,340,322	1,462,931	234,238	–0.60
10	2,667,175	2,983,373	445,109	–0.60	20	2,332,584	2,185,915	406,265	–0.80

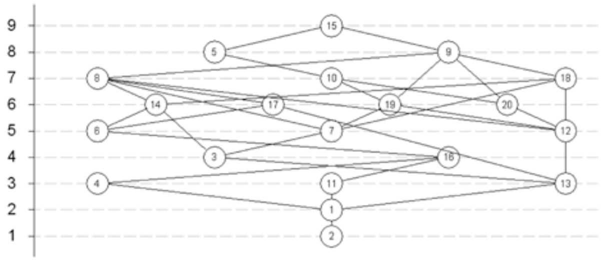


Fig. 13.1 Hasse diagram for the multi-indicator matrix in Table 13.1

Projects were evaluated according to their performance on four criteria ($m=4$) previously selected by the decision unit. The first three criteria are related to macroeconomic variables: Compensation of employees (REO) (an economic criterion that represents the aggregate income of employees that a project could produce (MMBs): a high value is better); profitability (EE) (a financial criterion related to each project which represents not only the net present value but also the aggregate benefit for an economy (MMBs): a high value is better); and employment (PO) (the capability to hire people (number of employees): a high value is better). The last criterion considered is the impact of the project on the environment, defined as environmental impact assessment (IA) (a low value is better).

Table 13.1 lists the set of 20 projects ($n=20$) under analysis. Each column represents the performance value $\mathrm{PV}_{i,j}$ for each project ($i=1, \ldots, 20$) and for each attribute ($j=1, \ldots, 4$) and defines the base case to be assessed.

Figure 13.1 shows the Hasse diagram for this example. There are nine levels. The maximal object located at level 9 corresponds to project 15 (i.e., project 15 is the most convenient project for the decision-maker) while the minimal object corresponds to project 2 (level 1). For all of the projects in level 8 (5 and 9), for at least one of the criteria, there is no clear decision about which project is ranked as more convenient. Thus, these two projects are incomparable. The same behavior is noted at other levels.

Table 13.2 Projects ranking base case

Project	Copeland	LPOM	LPOMext	Project	Copeland	LPOM	LPOMext
1	19	19	19	11	16	16	16
2	20	20	20	12	12	14	14
3	13	15	15	13	17	18	18
4	18	16	16	14	9	8	8
5	3	3	3	15	1	1	1
6	14	11	10	16	15	12	12
7	11	12	13	17	10	7	7
8	5	4	4	18	6	5	5
9	2	2	2	19	7	9	9
10	4	6	6	20	8	10	11

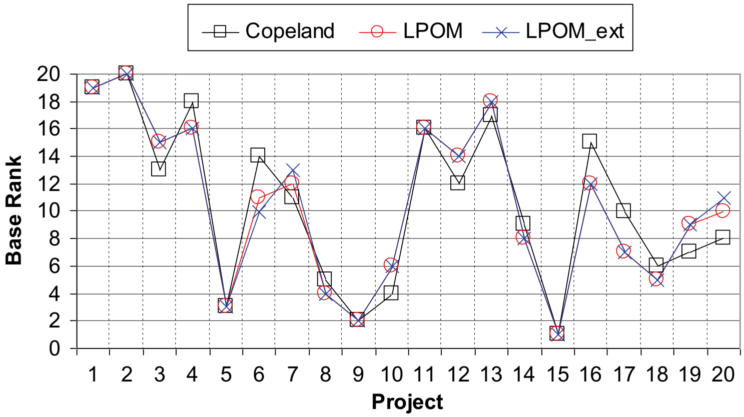


Fig. 13.2 Base rank comparison among ranking techniques, for each project

It is interesting to note that the Hasse diagram is able to define particular rankings. For example, consider the right chain in Fig. 13.1: 15–9–18–12–13–1–2. This means that project 15 is more "convenient" than project 9; that project 4 is more "convenient" than project 18, and so on. That is, 15 > 9 > 18 > 12 > 13 > 1 > 2.

In summary the Hasse diagram technique reveals that excluding projects 15 and 2, it is very difficult to rank the rest of the projects. So the decision to select a "second" best among these projects is not evident, and additional procedures are required to achieve a "complete" rank.

13.5.1 Base Rank

Table 13.2 shows the base rank obtained using the ranking technique described in Sect. 13.4. All ranking techniques rank project 15 as the first followed by projects 9 in the second position and then project 5 in the third position. Figure 13.2 shows the

Table 13.3 Spearman correlation coefficient among rankings for selected RT

SCC	Copeland	LPOM	LPOMext
Copeland	1	0.957	0.946
LPOM	0.957	1	0.997
LPOMext	0.946	0.997	1

BR provided by each ranking technique for every project. It seems that all the ranking techniques produce almost the same ranking.

Table 13.3 shows the Spearman correlation coefficient (SCC) among the selected ranking techniques. An SCC = 1 means that the rankings obtained by the two ranking techniques are equal, while smaller SCC values indicate that the rankings are quite different. Copeland ranking is almost concordant with LPOM or LPOMext rankings, while rankings by LPOM and LPOMext are practically equals.

13.5.2 Uncertainty Propagation

Badillo (2010) assumes that the data from Table 13.1 are uncertain and could be modeled by random variables with known probability distribution function. Random variables (RV) are numbered column-wise: RV_1 corresponds to the performance value $PV_{1,1}$ of the first project for the first criterion. RV_2 corresponds to the performance value $PV_{2,1}$ of the second project and the first criterion, and so on. In total, there are 80 RV (or factors).

In this example a triangular distribution for each RV is assumed. Table 13.4 shows the lower and upper bound associated with each RV. It is interesting to note that some PVs for specific projects are considered by Badillo (2010) as deterministic values.

Figure 13.3 shows the corresponding Hasse diagram when considering lower and upper bound. Note that in both cases project 15 is still the maximal object (as in the base case), but minimal objects vary. This fast evaluation suggests that project ranking could vary due to the uncertainty in the data.

As previously mentioned, the procedure is as follows (1) a random deviate is generated for each factor and a sample of matrix Q is built, and (2) the three ranking techniques are used to rank the projects. In this evaluation 10,000 samples are selected. It is important to mention that the same sample matrix Q is used by each ranking technique. As a result, an approximation of the probability distribution function of the rank position of each alternative is obtained.

Figure 13.4 shows the effects of the uncertainty in the data on each ranking technique. Left panels show the variations associated with CS, LPOM, and LPOMext, while right panels show the best and the worst ranks for each project, along with the their base ranks.

Let us consider the results for the Copeland ranking technique (Fig. 13.4a). From the left panel, it is clear that the range of the Copeland Scores for each project overlap.

Table 13.4 Lower and upper limit considered

	REO		EE		PEO		IA	
PR	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
1	332,090	711,621	321,167	688,216	67,848	145,390	–1.00	–0.60
2	89,403	726,403	86,419	702,156	18,493	150,254	–1.00	–0.60
3	291,691	1,069,535	280,688	1,029,190	55,765	204,472	–0.80	–0.40
4	266,599	799,797	254,629	763,887	55,871	167,613	–0.60	–0.20
5	3,157,515	3,859,185	3,541,659	4,328,694	524,083	640,545	–0.80	–0.40
6	581,041	581,041	561,471	561,471	119,371	119,371	–0.60	–0.20
7	490,591	1,144,714	466,948	1,089,547	94,804	221,210	–0.80	–0.40
8	309,478	1,495,814	304,738	1,472,904	57,884	279,773	–0.40	0.00
9	2,429,447	2,429,447	2,275,772	2,275,772	421,882	421,882	–0.40	0.00
10	1,568,926	3,765,423	1,754,925	4,211,821	261,829	628,390	–0.80	–0.40
11	215,667	862,669	207,819	831,277	42,901	171,604	–0.60	–0.20
12	334,491	1,337,966	318,202	1,272,809	64,419	257,677	–1.00	–0.60
13	597,126	597,126	593,121	593,121	121,416	121,416	–1.00	–0.60
14	291,500	1,166,002	308,775	1,235,101	53,676	214,705	–0.60	–0.20
15	5,058,331	5,058,331	5,714,542	5,714,542	827,703	827,703	–0.40	0.00
16	550,750	550,751	522,426	522,426	114,819	114,819	–0.60	–0.20
17	106,121	1,182,492	101,196	1,127,619	20,754	231,258	–0.40	0.00
18	376,554	1,793,117	405,123	1,929,158	67,518	321,514	–0.60	–0.20
19	490,362	2,190,283	535,218	2,390,644	85,697	382,780	–0.80	–0.40
20	2,078,541	2,586,628	1,947,845	2,423,985	362,019	450,512	–1.00	–0.60

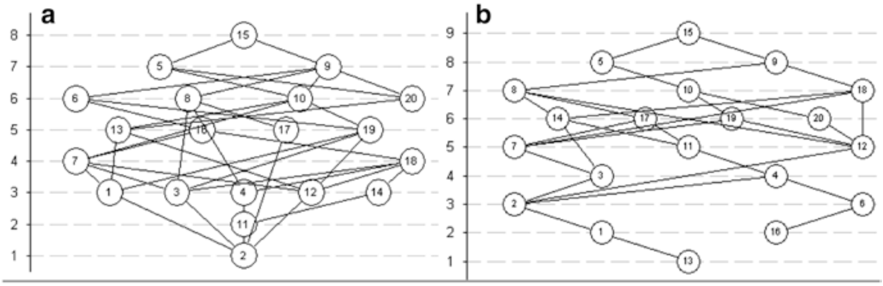


Fig. 13.3 Hasse diagram considering extreme bounds: (**a**) Lower bounds; (**b**) Upper bounds

This means that, for a possible set of performance values, a project could be ranked better that another.

For example, the range of CS for project 15 overlaps the range of CS for project 9. Therefore project 9 could be ranked as the first and project 15 as the second. However this rank-reversal situation is not possible between projects 15 and 5, since their CS ranges do not overlap. In conclusion, project 15 could be ranked as first or second. Note that a rank reversal is possible between projects 9 and 5 as well as other projects.

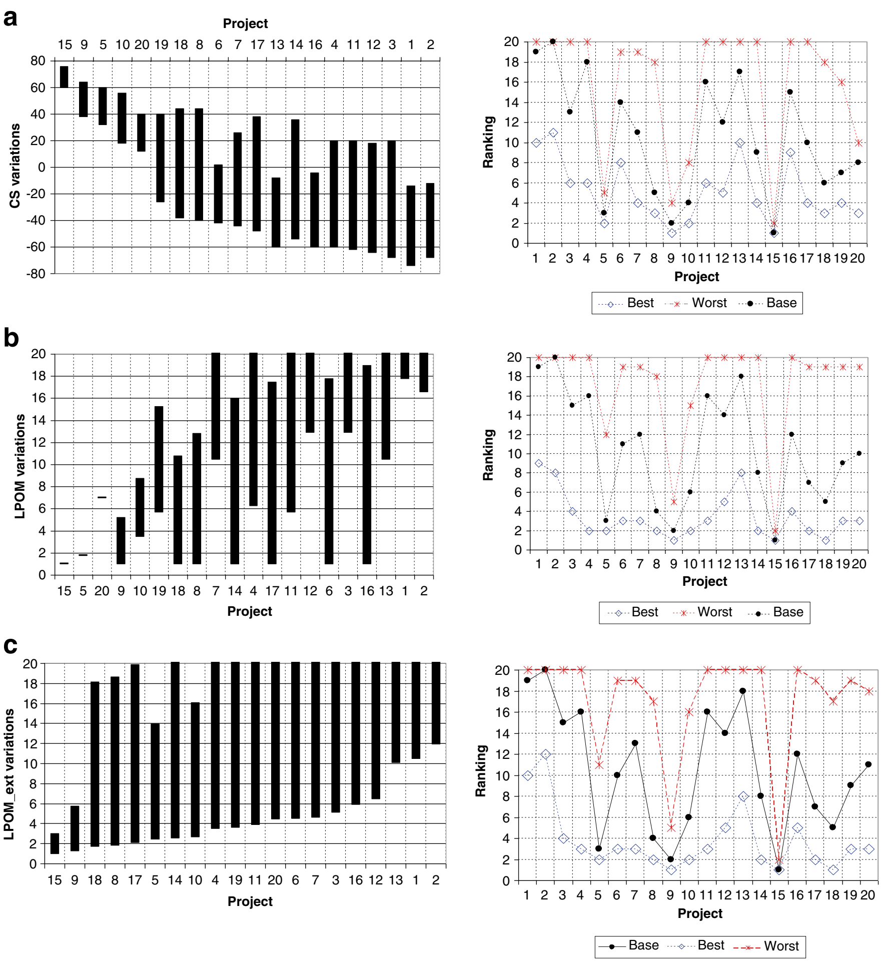


Fig. 13.4 Effects of uncertainty on different RT: (**a**) Copeland, (**b**) LPOM, and (**c**) LPOMext. *Left panel* shows the corresponding intervals, while *right panel* shows the ranking variations for each project

The best and the worst rank that a project can occupy (i.e., the highest or lowest position) are presented in the right panel. For example, project 1 is ranked in the 19th position in the base case, but the best and the worst position could vary between 10th and 20th. Similar results are obtained with the other ranking techniques.

From Fig. 13.4, however, it is not possible to assess if a project could occupy the entire set of rankings. To clarify this situation, Fig. 13.5 shows the approximated probability distribution function for each project ranking. That means that each possible ranking position has a corresponding probability. Figure 13.5 is developed using the LPOM technique. Similar results are obtained with the other ranking techniques.

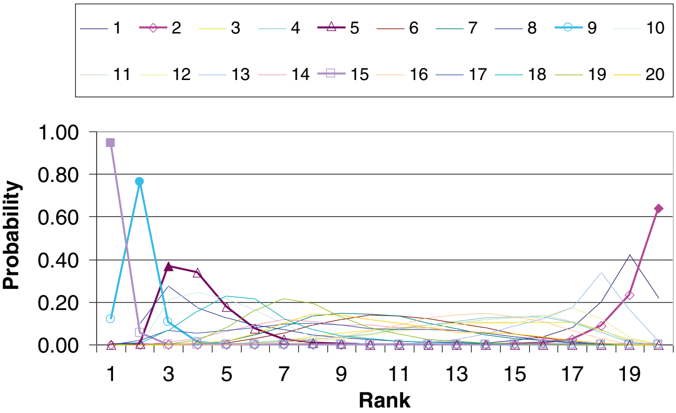


Fig. 13.5 Approximate probability distribution functions for each project (LPOM case)

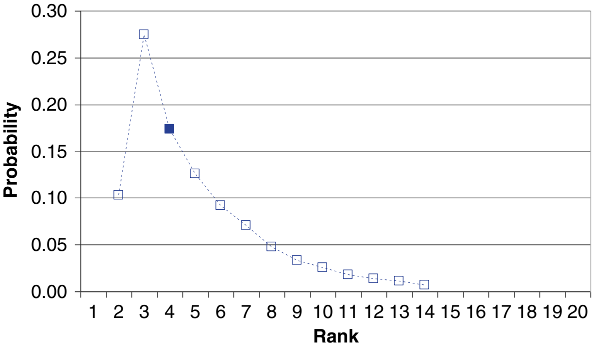


Fig. 13.6 Approximate probability distribution functions for project 8 (LPOM case)

Figure 13.4c shows that, for example, project 9 can be ranked between the 1st and the 5th position, but no probabilities are associated with each position.

From Fig. 13.5 (light-blue circled curve) it is clear that the most probable position for project 9 is the 2nd (which corresponds to the position occupied in the base case), followed by positions 1st, 3rd, and 5th.

Figure 13.6 shows the approximated rank distribution for project 8. In this case, it is easy to detect that the most probable position is the 3rd while the base case position is the 4th.

Table 13.5 shows the details for selected positions (LPOM case).

Several questions can be answered from Table 13.5:

1. What is the probability that a specific alternative, ranked in position *i* in the base case, is still ranked in the same position when uncertainty is considered?
2. What is the rank that could be occupied with the highest probability?
3. What are the probabilities for the best and the worst ranking?

Table 13.5 Uncertainty analysis: Probability for several rank conditions (LPOM case)

Project	Base rank	Probability for base rank	Rank with maximum probability	Maximum probability	Best rank	Probability for best rank	Worst rank	Probability for worst rank
1	19	0.4257	19	0.4257	9	0.0002	20	0.2176
2	20	0.6409	20	0.6409	8	0.0001	20	0.6409
3	15	0.1050	16	0.1062	4	0.0005	20	0.0027
4	16	0.1306	15	0.1309	2	0.0002	20	0.0012
5	3	0.3692	3	0.3692	2	0.0042	12	0.0001
6	11	0.1370	10	0.1378	3	0.0011	19	0.0004
7	12	0.1012	9	0.1488	3	0.0002	19	0.0006
8	4	0.1730	3	0.2744	2	0.1027	18	0.0003
9	2	0.7610	2	0.7610	1	0.1198	5	0.0008
10	6	0.1660	4	0.2466	2	0.001	15	0.0001
11	16	0.1390	16	0.139	3	0.0005	20	0.001
12	14	0.1008	17	0.1725	5	0.0001	20	0.0025
13	18	0.3378	18	0.3378	8	0.0003	20	0.0182
14	8	0.1122	7	0.1184	2	0.0017	20	0.0001
15	1	0.9451	1	0.9451	1	0.9451	2	0.0549
16	12	0.1254	14	0.1498	4	0.0007	20	0.0002
17	7	0.0967	8	0.1036	2	0.0192	19	0.0002
18	5	0.2307	5	0.2307	1	0.0003	19	0.0001
19	9	0.1330	7	0.2165	3	0.0037	19	0.0001
20	10	0.1207	8	0.1447	3	0.0018	19	0.0008

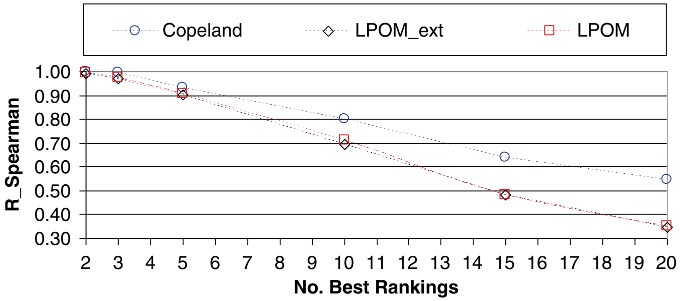


Fig. 13.7 Effects of uncertainty up to specific positions

At this point, the decision-maker knows that, for example, project 15 has a high probability (0.9451) of being the best project and a low probability (0.0549) to occupy the second position (using Copeland, these probabilities are 0.9999/0.001 respectively; using LPOMext, these probabilities are 0.9295/0.0705).

Figure 13.7 shows the SCC between the base rank and the samples generated, up to selected best positions. Under uncertainty, for example, the decision-maker, could be interested in assessing what happens with the projects that were ranked in the base case as the first and the second ones (i.e., up to the second best position). In this case, the SCC is almost 1, indicating that projects 15 and 9 maintain their

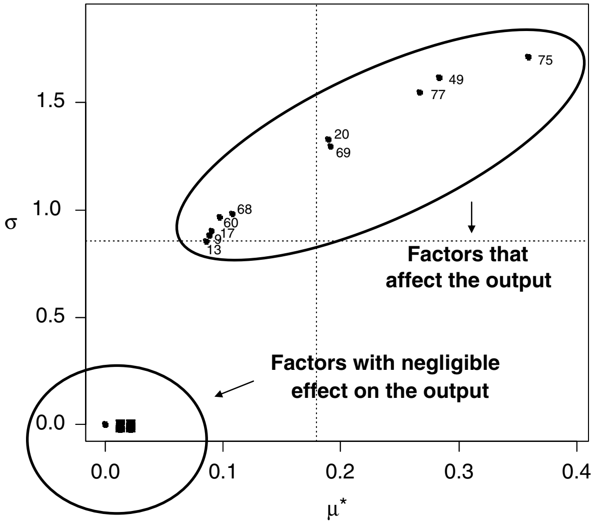


Fig. 13.8 Morris screening results, for the first two positions of the BR (LPOM case)

base positions, during the simulation. This conclusion is valid no matter which ranking technique is used.

However, as the number of best positions being considered increases, the SCC decreases, indicating that the rank of projects in the BR tends to be different under uncertainty. While the SCCs evaluated by LPOM or LPOMext are almost equals, the SCC obtained by Copeland is higher, suggesting that Copeland rankings are less sensitive to data uncertainties.

13.5.3 Global Sensitivity Analysis

The variable of interest (i.e., the output) for the global sensitivity analysis is the SCC between the base rank obtained using the nominal performance values, and the rank obtained by any of the 10,000 random samples. In this chapter, Morris evaluations are performed using the R procedure “sensitivity” (Pujols 2009).

The idea of the global sensitivity analysis is to know which factors are more responsible for the variation of a selected output. For example, Fig. 13.8 shows the results of the Morris screening analysis ($r = 10$) when the selected variable is the SCC between the first two positions of the base ranking and the first two positions of the ranking generated in each sample (using LPOM as the evaluating technique). As mentioned in Sect. 13.3.2, $\mu^*(x_i)$ is useful when the goal of the analysis is to detect those factors that contribute the most to the variation of the output. Figure 13.8 shows two set of factors (a) the set of factors with negligible effects ($\mu^*(x_i) \approx 0$) and (b) the set of factors that affect the output. For example, factors 20, 49, 69, 75, and

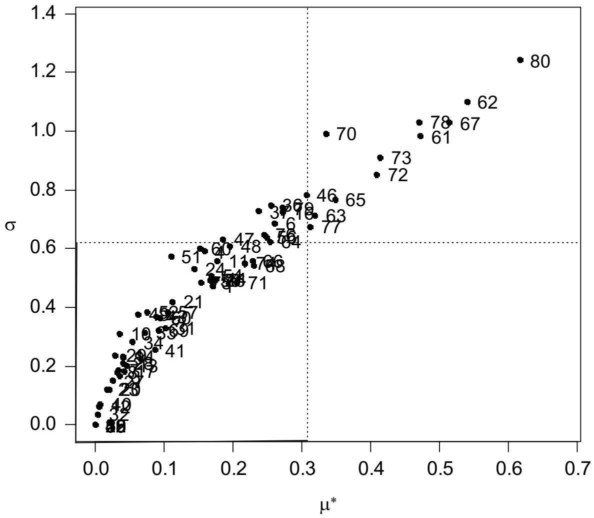


Fig. 13.9 Morris screening results, for the complete BR (LPOM case)

77 (located in the upper right position) correspond to the attribute REO of project 20 ($PV_{20,1}$), attributes PEO and IA of project 9 ($PV_{9,3}$ and $PV_{9,4}$), and attribute IA of projects 15 ($PV_{15,4}$) and 17 ($PV_{17,4}$). Note that, from Table 13.5, the best position for these projects is the 2nd.

Figure 13.9 shows the results considering the SCC between the full BR and the full ranking generated in each sample. For example, the performance values and the uncertainties associated with factors 46, 61, 62, 63, 65, 67, 70, 72, 73, 77, 78, and 80 (upper right position) have to be analyzed carefully to understand how it is possible to reduce the volatility of the rankings.

However, no matter the selected output, the analyst could perform a Monte Carlo filtering analysis to determine the key factors that produce a “defined behavior for the model” (Saltelli et al. 2004). The analyst could also be interested to knowing what PV sets are responsible of specific rank conditions (e.g., projects 15 and 9 ranked always as the best two projects). In this case, the use of classification techniques can provide an answer, based on a set of **If–Then** rules (Rocco 2012).

13.6 Conclusions

In this chapter uncertainty and sensitivity analysis are used as additional tools for quantifying the effects of the variations in the coefficients of a multi-indicator matrix, on the ranking of objects. Uncertainty analysis can be useful to study the robustness of a specific ranking, for example, by evaluating the possible positions that an object can occupy or its most probable rank. Sensitivity analysis allows analysts to assess the importance of uncertainties in performance values with a view to possibly reduce the volatility of the rankings.

The use of these tools is illustrated on a real example along with three specific nonparametric ranking techniques that consider only the information on the multi-indicator matrix. Two of them are based on the partial order theory, while the third is based on the Copeland approach. However, the analysis could be also performed using any multi-criteria technique. The results show that the approach is able to produce additional information that could be useful in a decision-making process.

References

Al-Sharrah G (2010) Ranking using the copeland score: a comparison with the Hasse diagram. J Chem Inf Model 50:785–791

Al-Sharrah G (2011) The copeland score as a relative and categorized ranking tool. Statistica Applicazioni, Spl Iss: 81–95

Annoni P, Brüggemann R, Saltelli A (2011) Partial order investigation of multiple indicator systems using variance based sensitivity analysis. Environ Model Softw 26(2011):950–958

Badillo R (2010) Project ranking under uncertainty. M.Sc. Thesis in Operational Research, Universidad Central de Venezuela (In Spanish)

Brans JP, Vincke PH (1985) A preference ranking organization method: the PROMETHEE method for multiple criteria decision making. Manag Sci 31(6):647–656

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partial orders. Match Commun Math Comput Chem 65:383–414

Brüggemann R, Patil G (2011) Ranking and prioritization for multi-indicator systems. Springer, Dordrecht

Brüggemann R, Schwaiger J, Negele RD (1995) Applying Hasse diagram technique for the evaluation of toxicological fish tests. Chemosphere 30:1767–1780

Brüggemann R, Bücherl C, Pudenz S, Steinberg C (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim Hydrobiol 27:170–178

Brüggemann R, Sorensen P, Lerche D, Carlsen L (2004) Estimation of averaged ranks by a local partial order model. J Chem Inf Comp Sci 44:618–625

Campolongo F, Cariboni J, Saltelli A (2007) An effective screening design for sensitivity analysis of large models. Environ Model Softw 22:1509–1518

Campolongo F, Saltelli A, Cariboni J (2011) From screening to quantitative sensitivity analysis. A unified approach. Comput Phys Commun 182:978–988

De Loof K, De Baets B, De Meyer H (2011) Approximation of average ranks in posets. Match Commun Math Comput Chem 66:219–229

Dorini G, Kapelan Z, Azapagic A (2011) Managing uncertainty in multiple-criteria decision making related to sustainability assessment. Clean Techn Environ Policy 13:133–139

Hernandez E (2006) A multicriteria methodology to rank investment projects. M.Sc. Thesis in Operational Research, Universidad Central de Venezuela (In Spanish)

Hwang CL, Yoon KS (1981) Multiple attribute decision making: methods and applications. Springer, Berlin

Hyde KM, Maier HR (2006) Distance-based and stochastic uncertainty analysis for multi-criteria decision analysis in Excel using Visual Basic for applications. Environ Model Softw 21:1695–1710

Hyde KM, Maier HR, Colby CB (2004) Reliability-based approach to multicriteria decision analysis for water resources. J Water Resour Plan Manage 130(6):429–438

Morgan MG, Henrion M (1992) A guide to dealing with uncertainty in quantitative risk and policy analysis. Wiley, Chichester

Morris MD (1991) Factorial sampling plans for preliminary computational experiments. Technometrics 33(2):161–174

Pujol G (2009) Sensitivity analysis R package. http://cran.r-project.org/web/packages/sensitivity

Restrepo G, Brüggemann R, Weckert M, Gerstmann S, Frank H (2008) Ranking patterns, an application to refrigerants. Match Commun Math Comput Chem 59:555–584

Rios Insua D, French S (1991) A framework for sensitivity analysis in discrete multi-objective decision-making. Eur J Oper Res 54(2):176–190

Rocco CM (2012) Classification algorithms for extended factor mapping setting in sensitivity analysis. Internal Report UCV-DIOC, Universidad Central de Venezuela (In Spanish)

Roy B (1968) Classement et choix en presence de points devue multiples (la methode ELECTRE). Revue d'Informatique et de recherché opérationelle 6(8):57–75

Roy B (1985) Méthodologie multicritère d'aide à la décision. Economica, París

Saaty TL (1980) The analytic hierarchy process. McGraw Hill, New York

Saltelli A, Chan K, Scout EM (2000) Sensitivity analysis. Wiley, Chichester

Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) Sensitivity analysis in practice. A guide to assessing scientific models. Probability and statistics series. Wiley, Chichester

Sørensen PB, Mogensen BB, Carlsen L, Thomsen M (2000) The influence on partial order ranking from input parameter uncertainty. Definition of a robustness parameter. Chemosphere 41: 595–601

Triantaphyllou E, Sanchez A (1997) A sensitivity analysis approach for some deterministic multi-criteria decision-making methods. Decis Sci 28(1):151–194

Wolters WTM, Mareschal B (1995) Novel types of sensitivity analysis for additive MCDM methods. Eur J Oper Res 81(2):281–290

Yu O-Y, Guikema SD, Briaud J-L, Burnett D (2012) Sensitivity analysis for multi-attribute system selection problems in onshore environmentally friendly drilling (EFD). Syst Eng 15(2):153–171

Chapter 14
Hasse Diagram Technique Contributions to Environmental Risk Assessment

Stefan Tsakovski and Vasil Simeonov

Abstract This chapter deals with the successive application of self-organizing map (SOM) classification and Hasse diagram technique (HDT) as chemometric tools for assessment of river water and sediment quality. Both studies are carried out by using long-term water quality monitoring data from the Struma River catchment, Bulgaria and lake sediment samples from Mar Menor lagoon in Spain. The advantages of the SOM algorithm for advanced visualization and classification of large datasets are used for proper selection of chemical parameters being most effective in quality assessment combined with some state directives for surface water quality parameters in the river water study and as preprocessing procedure of the initial sediment data matrix. The simultaneous application of the SOM methodology or legislation norms with Hasse diagram technique allows to visualize the spatial and temporal evolution of water quality parameters or to reveal specific sediment pollution patterns.

14.1 Introduction

Environmental risk assessment (ERA) is one of the main tools for environmental impact assessment. Usually ERA framework starts with (1) problem formulation as a critical first step (which includes hazard identification); (2) hazard characterization, which examines potential hazards and their magnitude; (3) exposure characterization, which covers levels and likelihood of exposure; (4) risk characterization, in which the magnitude of consequences and the likelihood of occurrence are integrated; and finish with (5) risk evaluation. The main goal of assessment procedure is to determine a predicted environmental concentration (PEC) and a predicted no

S. Tsakovski (✉) • V. Simeonov
Group of Chemometrics and Environmetrics, Faculty of Chemistry
and Pharmacy, University of Sofia, Sofia, Bulgaria
e-mail: STsakovski@chem.uni-sofia.bg

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_14, © Springer Science+Business Media New York 2014

effect concentration (PNEC) for each environmental compartment using the relevant data. If the PEC exceeds the PNEC, it is considered to be a risk of environmental damage in proportion to the ratio of PEC to PNEC. The estimation of PEC and PNEC in complex environmental system is quite difficult task (European Commission 2003). The well-established practice of comparison of environmental indicators with empirically defined limits could not be used for confident ERA. A much more reliable approach proves to be the chemometric strategy for classification, modeling, and data interpretation of environmental monitoring results, since they consider the environmental system as multivariate and treat it correspondingly (Einax et al. 1997).

Normally, the monitoring datasets include many objects of interest (e.g., sampling locations observed in a shorter or extended time interval) described by many variables (e.g., chemical pollutant concentrations, physicochemical parameters, or ecotoxicity test values). In order to interpret correctly the multiparametric datasets, one often needs preselection of significant variables (in order to reduce the dataset dimensionality) followed by some ordering or ranking procedure (in order to determine reliably the impact of the selected parameters on the environmental system).

The successive application of self-organizing maps (SOM) and introducing of legislation norms like preprocessing procedures and Hasse diagram technique (HDT) as ranking approach will be presented as a new strategy in ERA studies. The major objective of the present chapter is to demonstrate this strategy in two important water quality assessment studies.

14.2 Basic Features of Hasse Diagram Technique and Self-Organizing Maps

14.2.1 Hasse Diagram Technique

Hasse diagrams visualize partial order relations between objects described by a certain number of variables. HDT is well described (see for example Brüggemann et al. 2001) and here only a brief description concerning the present studies will be given.

In HDT the ranking of objects (sampling points) is done with respect to a set of variables (e.g., water quality parameters, heavy metal distributions, or ecotoxicity), which is called the "information basis" (IB). The set of objects is called *E*. The processed data matrix Q (N × R) contains *N* objects and *R* variables. The entry q_{ir} is the numerical value of the *r*th variable of the *i*th object. The q_r is a variable by which the objects will be ranked. The two objects *s* and *t* are comparable if

$$s,t \in E; s \le t \Leftrightarrow q(s) \le q(t),$$

$$q(s) \le q(t) \Leftrightarrow q_r(s) \le q_r(t) \quad for\ all\ q_r \in IB.$$

If there are some q_r for which $q_r(s) > q_r(t)$ *and some others for which* $q_r(s)<q_r(t)$, then the objects s and t are incomparable ($s||t$). Two objects are equivalent if they have one and the same data with respect to a given set of variables ($s = t$). A partial order set can be easily developed by the cover relation matrix, which collects relations between each pair of objects. The "Hasse matrix" (see for instance Mauri and Ballabio 2008) is a (N × N) antisymmetric matrix where for each pair of elements s and t the entry h_{st} is given:

$$h_{st} = \begin{cases} 1 \quad \mathit{if}\, q_r(s) \geq q_r(t) \quad \mathit{for\,all}\; q_r \in IB \\ -1 \quad \mathit{if}\, q_r(s) < q_r(t) \quad \mathit{for\,all}\; q_r \in IB \\ \qquad\quad 0 \quad \mathit{otherwise.} \end{cases}$$

The order relations stored in a the matrix are visualized by a Hasse diagram, where the objects are drawn as small circles together with an appropriate identifier grouped in levels. Hasse diagrams are oriented acyclic graphs and usually when object t is "greater" than object s, object t is located above s in the plane. In the present studies, the objects near to the upper part of the Hasse diagram indicate objects that are the "polluted" or "toxic" ones according to the criteria used for their ranking. The objects not "covered" by other objects are called *maximal* objects. These objects, which do not cover other objects, are called *minimal* objects. In some diagrams there also exist *isolated* objects which can be considered to be maximal and minimal objects at the same time. A chain is a set of comparable objects; levels can be defined according to the longest chain within the diagram. An anti-chain is a set of mutually incomparable objects. The height of the diagram is the longest chain and longest anti-chain is its width (Brüggemann and Patil 2011).

The sensitivity analysis of Hasse diagram toward variables describing objects could be done by the dissimilarity matrix (W matrix). The W matrix represents the influence of the variables on metric distance between partially ordered sets (posets), based on different subsets of IB [$R-1$ variables, see for details, e.g., Brüggemann and Patil (2011)].

The similarity analysis of two Hasse diagrams derived from different sets of variables gives an opportunity to calculate the similarity between the two resulting (posets). In the first study of this chapter, two different sets of variables: heavy metal distributions (HM) and ecotoxicity (TOX) will be compared. Let N be the set of sampling sites, TOX the columns of ecotoxicity values, and HM the set of heavy metal distributions (heavy metal concentrations in different lake compartments). Thus, two partial order sets—(N, TOX) and (N, HM)—can be represented by two Hasse diagrams. Similarity analysis between the resulting Hasse diagrams can be performed by comparing the order relations between each two objects in both posets. The similarity between both Hasse diagrams can be presented by a tuple (#isotone, #antitone, # weak isotone, #indifferent, and #equivalences), where # is the number of respective relations. The total number of relations is equal to $n \times (n-1)$, with n, the number of sampling locations. The degree of each partial order combination can be derived. For example, the degree of isotones is equal to #isotones/($n \times (n-1)$). A detailed similarity analysis description can be found in Brüggemann and Patil (2011).

Moreover, the similarity between the above-mentioned two partial orders can be optimized. This can be done by finding an appropriate subset of variables from HM which gives an order preserving map (*N*, TOX) → (*N*, HM). Selection of the appropriate subset among 2^R-1 subsets can be performed by finding the maximum of the constructed objective function (Voigt et al. 2010). The objective function *O* is the weighted sum of two other functions: O_1 and O_2. O_1 selects the most similar TOX column out of HM but accepts inverse relations ("antitones"). O_2 is a function avoiding inverse relations and obtaining an order preserving map by the fraction of "indifferent".

All calculations concerning HDT were performed using the free available software package PyHasse, written by the first author in Brüggemann and Patil (2010, 2011). DART software freely available on the ECB Computational toxicology site (Manganaro et al. 2008) and WHASSE written by the first author (Brüggemann et al. 2001) were used for visualization.

14.2.2 Self-Organizing Map

SOM is an algorithm used to visualize and interpret large high-dimensional datasets (Kohonen 2001). SOM is an unsupervised pattern cognition method similar to cluster analysis. The main advantage of SOM is the simultaneous clustering of variables and objects (sampling locations). Typical applications are visualization of process states or observation results by representing the central dependences within the data on the map. Usually the map consists of a regular grid of processing units called neurons (n) whose number is determined using the formula $n=5\sqrt{\text{number of samples}}$.

An input vector of some multidimensional observation, eventually a vector consisting of features (variables, attributes), is associated with each unit. The map attempts to represent all available observations with optimal accuracy using a restricted set of output vectors. At the same time, the input vectors become ordered on the grid so that similar input vectors are close to each other and dissimilar input vectors are far from each other. Fitting of the input vectors is usually carried out by a sequential regression process, where $t=1,2,\ldots$ is the step index: For each sample $x(t)$, first the winner index c (best matching unit—BMU) is identified by the condition:

$$\forall i, x(t) - m_c(t) \le x(t) - m_i(t).$$

After finding the BMU, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space. Further, all output vectors or a subset of them that belong to nodes centered around node $c=c(x)$ are updated as $m_i(t+1)=m_i(t)+h_{c(x),i}(x(t)-m_i(t))$.Here, $h_{c(x),i}$ is the "neighborhood function," a decreasing function of the distance between the ith and cth nodes on the map grid. This regression is usually reiterated over the available objects. The quality of SOM is evaluated by quantization error (QE) and topographic error (TE). QE is the average distance between each input vector and its BMU, while TE is a measure for topological preservation of all input vectors. The network is trained different numbers of map units and optimum size is chosen based on minimum QE and TE.

The trained map consisting of nodes linked with respect to their output vectors could be graphically presented by 2D planes for each variable indicating variable distribution values on the different map regions by different colors. These "component" planes could be compared in order to detect similarities between variables. Close placed planes are indication for similar behavior and correlation between respected variables.

SOM gives also the U-matrix plane, where distances between nodes are visualized. Additionally, the node vectors of the respective objects could be used for further statistical data treatment. All calculations concerning SOM were performed by a free SOM Toolbox 2.0 (Vesanto 1999).

14.3 Case Studies

14.3.1 Mar Menor Lagoon

The heavy metal pollution of aquatic ecosystems has reached serious proportions as a result of their toxicity and accumulative behavior. During their transport, heavy metals released into an aquatic system from natural or anthropogenic sources are unevenly distributed among the different compartments of the aquatic ecosystem: water, sediment, and biota (Moore and Ramamoorthy 1984; Felipe-Sotelo et al. 2007). Heavy metals do not remain dissolved in water; they are mainly adsorbed on bottom sediments, where they accumulate. Thus, the sediments can serve both as reservoirs and as potential sources of contaminants to the water column and can adversely affect sediment-dwelling organisms, aquatic-dependent wildlife, and ultimately human health (USA, EPA 2005). Therefore, bottom sediments can be used in investigations of sources of long-term pollution impact.

It is obvious that the environmental fate of heavy metals depends on many processes in the aquatic system governing their mobility and redistribution. In (Einax et al. 1999; Aulinger et al. 2004) the relationship between the distributions of heavy metals in different environmental compartments was found and their concentration in river sediments predicted with the use of partial least squares (PLS) multi-way models. Another group of studies, using N-way modeling, was performed to interpret heavy metal specific distribution determined by sequential extraction procedures (Pardo et al. 2004; Singh et al. 2007). Recently, an interpretation of heavy metal fractionation in suspended particulate matter and sediment fractions due to sedimentation processes was presented (Tsakovski et al. 2009). Two of the abovementioned studies (Pardo et al. 2004; Tsakovski et al. 2009) provide an estimation of the heavy metal environmental pollution impact derived from N-way models performed in these studies.

The aim of the present study is to estimate the ecotoxicity of five different heavy metals (Zn, Cu, Mn, Pb, and Cd) distributions (dissolved in water, suspended on two

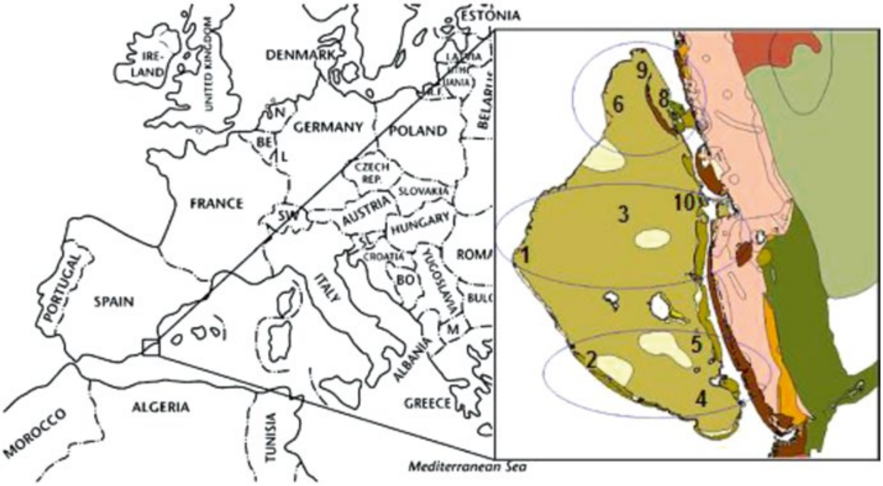


Fig. 14.1 Sampling region of Mar Menor Lagoon

particulate matter fractions and on two sediment fractions) in the Mar Menor coastal lagoon, south-eastern Spain. The environmental hazard was assessed by comparing the different heavy metal distributions with two ecotoxicological tests (Microtox$^{®}$ and Ostracodtoxkit FTM) using the Hasse diagram technique.

14.3.1.1 Sampling Area, Sampling, and Sample Measurement

The area of sediment monitoring was the Mar Menor lagoon (Spain) (see Fig. 14.1). The Campo de Cartagena is located in the south-east of the Murcia Region. It is a wide plain of very soft topography, with a gentle south-eastward slope towards the Mar Menor. This is a terrestrial geo-ecosystem of the neogen-quaternary basin of the same name. It is a lagoon that occupies a natural surface of around 180 km^2. The lagoon ecosystem is being affected to a degree that depends on the phase in which the metals are carried, since the bioavailability and toxicity of the metals are influenced by the distribution of trace metals in particulate, sediments, or dissolved phases.

The sampling was conducted during May 2007 and included tracing points of interest concerning or affecting pollution conditions in the area (from farmlands or from a mining area).

The aim here was to estimate concentration levels of metals in the water column (total, dissolved and suspended) and in the sediments. The water samples were passed through millipore nitrocellulose filters. Dissolved metals were preconcentrated with Chelex 100 resin and measured with FAAS and GFAAS. Suspended metals were determined by GFAAS and FAAS after digestion of the filters with conc. nitric acid (Tsakovski et al. 2009). The sediment samples were wet sieved and dried. Non-lattice held metals were extracted with dilute hydrochloric acid and total metals

Table 14.1 Average values of heavy metal distributions (all in μg g^{-1} except DISS values which are presented in μg L^{-1})

	Zn	Cu	Mn	Pb	Cd
DISS	11.5	0.50	3.42	2.21	0.05
PM1	9.30	9.56	28.3	17.7	0.17
PM2	402	36.2	694	287	1.22
SED1	877	25.0	509	782	3178
SED2	1.10	20.5	476	58.0	3438

with concentrated nitric and hydrofluoric acids. The acid extracts were analyzed for trace metals with FAAS (Kudłak et al. 2011).

The Ostracodtoxkit was used for the chronic toxicity (MORT variable) determinations and for acute toxicity "81.9 % Basic Test" of Microtox® (EC50 variable) were chosen (Tsakovski et al. 2012).

14.3.1.2 Results and Discussion

For statistical data treatment HDT using optimized similarity analysis between posets is chosen. The SOM is applied for comparison and explanation of obtained results in respect to relationships between heavy metal distributions and ecotoxicity parameters.

1. The dataset used for exploration consists of ten sampling stations, as each one is described by: five heavy metal concentrations (Zn, Cu, Mn, Pb, Cd) in water (coded as D), two suspended particulate matter fractions (PM0.45, PM8, coded as PM1 and PM2, respectively) and two sediment fractions (SED<63 μm, SED>63 μm, coded as SED1 and SED2, respectively), sediment acute (EC50), and chronic toxicity (MORT).
2. The mean concentration values of heavy metal distributions are presented in Table 14.1. More detailed information on heavy metals determination studies are given in (Marin-Guirao et al. 2007).

The Zn distributions in the PM1 and SED2 fractions were excluded from further data treatment. Since HDT is rather sensitive to even very small differences in parameter values uniform distribution as in the case with Zn variable could lead to unwanted ranking results that apparently would be without any physical meaning. So the dataset consists of 10 sampling stations characterized by 23 heavy metal distributions and 2 ecotoxicities (EC50 and MORT). The resulting dataset was normalized across the variable in order to be used for SOM calculations.

Figure 14.2 shows the U-matrix, all variable planes, and the station hits diagram for the input dataset. With the aid of a color scale, the distribution of each variable on the SOM map and the distances between the nodes in the U-matrix plane are easily found. For example, the stations with high Cu and Cd distributions in the PM2 fraction are located in the lower right-hand part of the SOM plane, while the stations with high Mn distributions in the SED1 and SED2 fractions are placed in the lower left-hand part of the plane. Comparison of acute toxicity (low EC50 values indicate

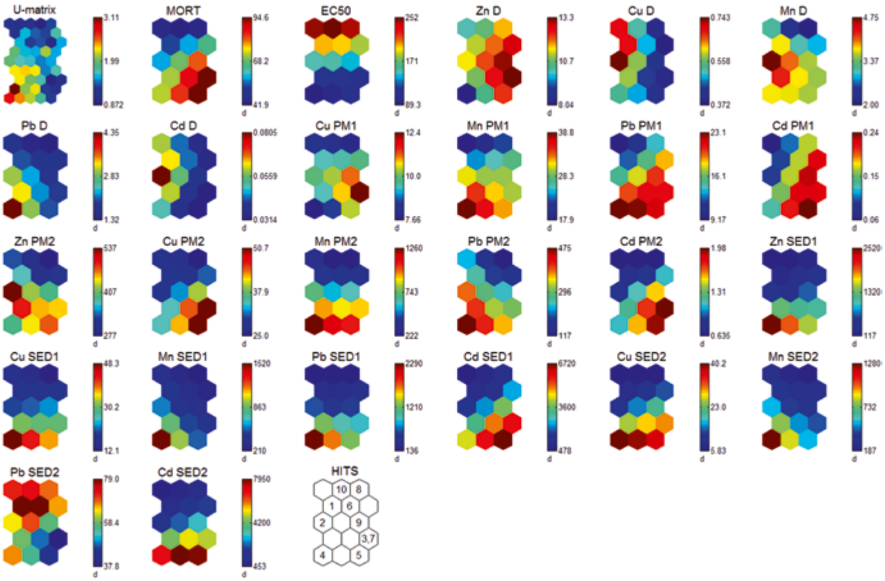


Fig. 14.2 SOM clustering of samples and sampling station hit diagrams

higher toxicity) and chronic toxicity (MORT) with the station hits diagram reveals that stations (3, 4, 5, 7) with elevated toxicity values are located in the lower part of the SOM map, while less toxic stations like 1,6, 8, and 10 are placed in the upper part of the map. This result corresponds well with the potential environmental hazard based on N-way analysis of heavy metal distributions in the investigated area (Tsakovski et al. 2009) where the potential environmental impact of sampling locations (1,2,3,…) increases in the order $4 > 5 > 3 \approx 7 > 9 \approx 2 > 6 > 1 > 10 > 8$.

The ordering of component planes (Fig. 14.3) allows to detect relationships between ecotoxicity parameters and heavy metal concentrations in different fractions, based on their position and color scale. It was found that chronic toxicity is strongly positively correlated with Cd PM2, Cu PM2, Cd SED1, and Pb SED2. The position of the EC50 plane is an indication of the similarity between acute toxicity and Mn PM2, Cd SED2, Cu SED2, Mn PM1, and Pb PM1. Visual inspection of ordering does not detect any special grouping by a certain heavy metal concentration in different fractions or grouping by a certain fraction.

The above-mentioned similarities are based on all the heavy metal concentrations in the different fractions and toxicity values apart from those being excluded. In order to detect heavy metal distributions having the highest impact on acute and chronic ecotoxicity, Hasse diagram optimized similarity analysis is carried out. The dataset is organized in three partial ordered sets. The first poset includes the heavy metal distributions at all the sampling locations (*N*, HM) and the other two include acute (*N*, EC50) and chronic toxicity (*N*, MORT) also for all sampling sites. EC50 values are multiplied by −1 in order that high values correspond to high acute toxicity levels. The optimized similarity analysis was performed by finding the most similar Hasse

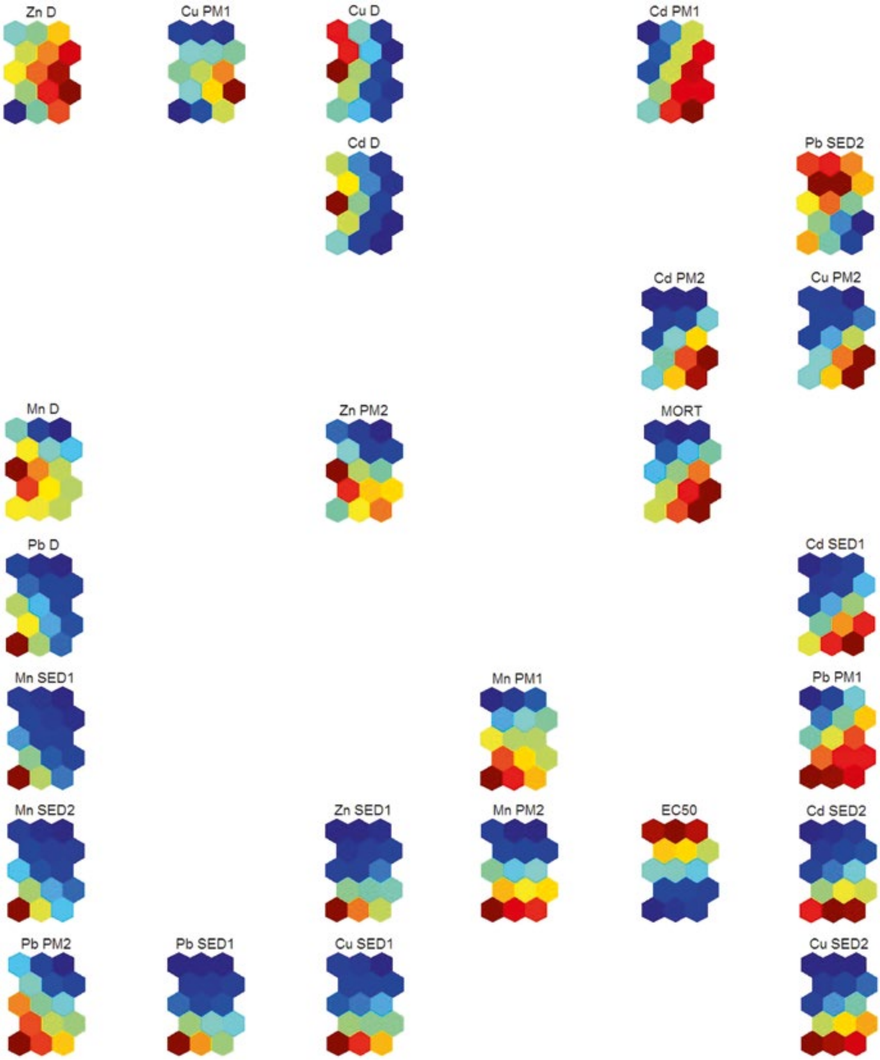


Fig. 14.3 Classification of heavy metal distributions and ecotoxicity variables

diagram based on a selected subset of heavy metal distributions for each one of the Hasse diagrams obtained from ecotoxicity parameters.

Optimized similarity analysis between (*N*, MORT) and (*N*, HM) partial ordered sets reveals that the optimal subset of heavy metal distributions includes Cu PM2, Cd PM2, and Cd SED1. This means that chronic toxicity is connected predominantly with Cu concentrations in the coarse particulate fraction and with Cd concentrations in coarse particulate matter and in the fine sediment fraction. This result is not surprising considering the similarities between chronic toxicity and the heavy metal distributions presented in Fig. 14.3. This connection also corresponds very well to

Table 14.2 Results of similarity analysis of two datasets (chronic toxicity and Cu PM2, Cd PM2, and Cd SED1 distributions)

Order relations	Number
Isotone (>> or <<)	50
Antitone (>< or <>)	0
Weak isotone (=ǁ, =>, =<)	2
Indifferences (ǁǁ, ǁ>, ǁ<)	38
Equivalences (==)	0

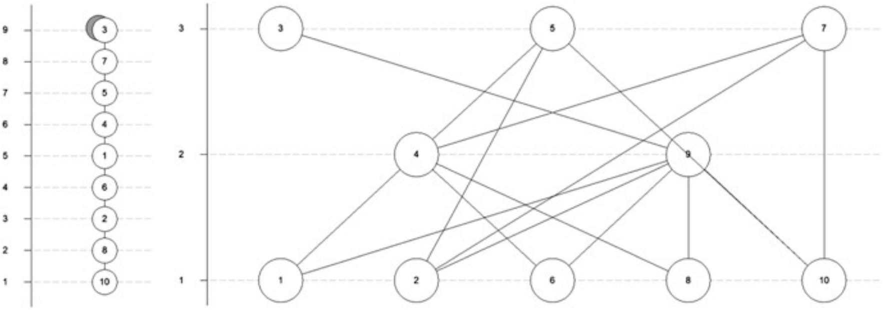


Fig. 14.4 Hasse diagram of chronic toxicity (lhs) and Hasse diagram of selected heavy metal distribution subset (rhs). The sampling locations 3 and 9 are equivalent in lhs

the recently reported order of metal toxicity to *H. incongruens*: Cd Cu > Mn > Zn > Pb (Kudłak et al. 2011). An explanation of the influence of Cu and Cd concentrations in the coarse particulate matter fraction to chronic toxicity can be given due to the increased release of heavy metals present in PM in the gastrointestinal tract of the Ostracods.

The results presenting the comparison of Hasse diagrams induced by both datasets of chronic toxicity and a selected subset of heavy metal distributions are listed in Table 14.2.

The degree of indifferences (d.ind.) is 0.42, which means that the difference between the two datasets has quite a significant impact on the structure of the Hasse diagrams. As stated in the theoretical part, the indifferent relations indicate combinations of comparabilities of the one Hasse diagram with incomparabilities of the other. The examination of both resulting Hasse diagrams (Fig. 14.4) shows that sampling location 9 makes the biggest contribution to the indifferent relations. The different position of the location in both diagrams can be explained by agricultural activity, which affects the northern part of the lagoon. Since organic pollutants are not measured in this study, the underestimation of chronic toxicity of sampling location 9 is not surprising.

Optimized similarity analysis between (N, EC50) and (N, HM) partial ordered sets reveals another optimal subset of heavy metal distributions, which includes Cd SED2, Cu SED2, and Pb SED1. This means that acute toxicity is related mainly to Cd and Cu concentrations in the coarse sediment fraction and Pb concentrations in the fine sediment fraction. The connection between Cd and Cu concentrations in the coarse sediment fraction with acute toxicity is easily explained by their similarities

Table 14.3 Results of similarity analysis of two datasets (acute toxicity and Cd SED2, Cu SED2, and Pb SED1 distributions)

Order relations	Number
Isotone (>> or <<)	72
Antitone (>< or <>)	0
Weak isotone (=ll, =>, =<)	0
Indifferences (llll, ll>, ll<)	18
Equivalences (==)	0

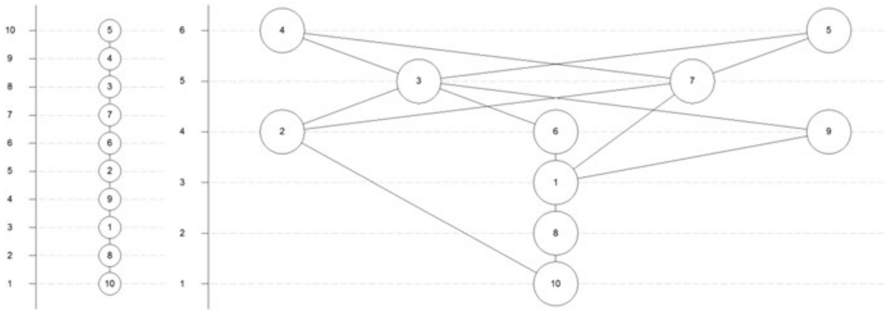


Fig. 14.5 Hasse diagram of acute toxicity (lhs) and Hasse diagram of selected heavy metal distribution subset (rhs)

(see Fig. 14.3) and the reported metal toxicity order to *Vibrio fischeri*: Cu > Cd Mn > Zn > Pb (Kudłak et al. 2011). The only explanation for the influence of Pb concentration in the fine sediment fraction on the acute toxicity could be the easier leaching of lead from the fine sediment fraction during acute toxicity determination. It is worth mentioning that mining activities in the region might be the reason for the Pb and Zn accumulation in the sediment fraction below 63 μm (Tsakovski et al. 2009).

The results presenting the comparison between both datasets of acute toxicity and the selected subset of heavy metal distributions are listed in Table 14.3.

In this case d.ind is 0.20 and the impact of the indifference relations is quite small. The reason for the better similarity between the two resulting Hasse diagrams (Fig. 14.5) can be found in the similar position of sampling location 9 in both Hasse diagrams. We suggest that the acute *Vibrio fischeri* test is less sensitive to the expected organic pollutants, not measured in this study, than the chronic *Heterocypris incongruens* test.

14.3.2 Struma River Catchment

The assessment of surface water quality is an extremely important environmental issue. The traditional approach of quality assessment by comparison of monitored and standardized by legislation threshold values lacks completeness. The significance and reliability of the chemometrics approaches using multivariate statistics is already proven by many studies (Astel et al. 2006; Mattikalli 1996; Cun and Vilagines 1997; Wunderlin et al. 2001; Goetz et al. 1998; Lu and Lo 2002; Simeonov et al. 2002; Mendiguchia et al. 2004; Stefanov et al. 1999; Kowalkowski et al. 2006; Tsakovski et al. 1999; Marengo et al. 1995; Brodnjak-Voncina et al. 2002). In this selection the trials were undertaken by means of traditional chemometric approaches like cluster analysis, factor analysis, principal component regression, discriminant analysis, etc.

Efforts have been made to involve more sophisticated approaches such as self-organizing maps (SOM) (Astel et al. 2007; Yan et al. 2001; Giraudel and Lek 2001; Mingoti and Lima 2006; Mangiameli et al. 1996) or Hasse diagram technique (HDT) (Tsakovski and Simeonov 2008) in classification and modeling studies with surface water data sets or to compare SOM classification with more traditional multivariate statistical classification methods. It has to be mentioned that the application of SOM, for example, makes it possible to reach a specific "resolution" of the classification scheme offered by Astel et al. (2007) as compared to more traditional methods such as cluster analysis and, thus, (a) to reveal specific features of the sampling sites within the monitoring net along a big river catchment and (b) to detect additional hidden sources of pollution along the same catchment.

On the other hand, the application of HDT in a lake sediment study (Tsakovski and Simeonov 2008) indicated that it is possible to offer a specific expertise of chronic and acute toxicity in the lake environment by ensuring additional information to the traditional classification patterns (e.g., by SOM) about pollution priority of chemical species as well as relations between sampling locations.

The aim of the present study is to assess the river water quality by combining preliminary sampling site and sampling parameters' classification (SOM) with water quality norm data and a decision support system (HDT) expertise. Thus, more specific features of the sampling locations and of the water quality parameters could be revealed and practically used.

The Struma River is located in the southern part of Bulgaria. It flows from north to south and has a length of 290 km as far as the Greek border. From that point to the Aegean Sea, the river is about 110 km long. Its total watershed in Bulgaria is nearly 10,250 km^2 and covers the Vitosha Mountains and the Rila, Pirin, and surrounding mountains (Fig. 14.6). Being a cross-border river, the Struma basin is of substantial importance to both Bulgaria and Greece. That is why careful monitoring of water quality in the long or short term at different sampling sites is not only an ecological but also a political issue.

The dataset used for the chemometric exploration consists of more than 15,000 measurements on the Struma River. The sites chosen almost completely cover the length of the river from its source to the Greek border. Water samples were collected between 1989 and 1998. The coding of the sampling location included digits indicating the number of the site within the National Monitoring Net (three digits for each site, e.g., 123, 124, etc., as indicated in Fig. 14.6) and the year of sampling (two digits for each site, e.g., 89, 90, 91, etc.). In the dataset, the objects of interest were coded as 123_89, 123_90, 124_98, etc. Annual averages for the quality parameters were used.

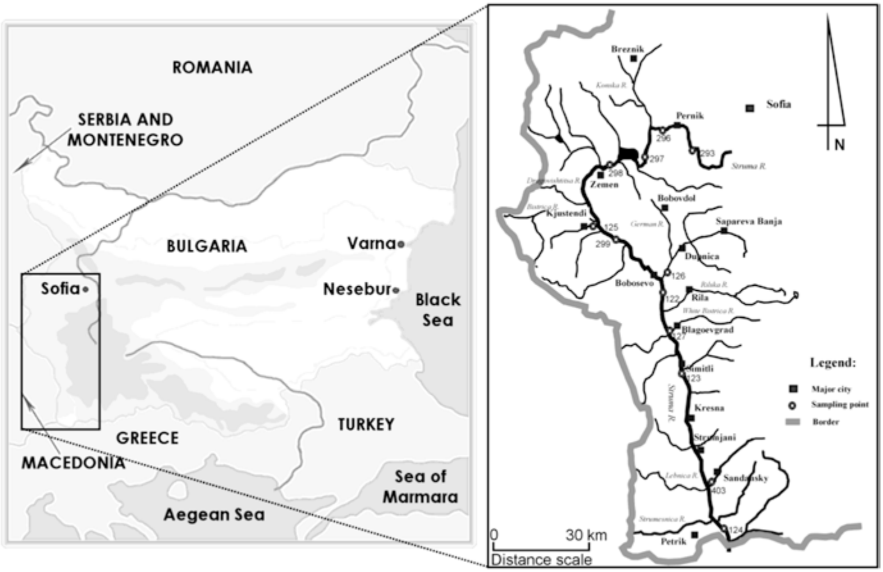


Fig. 14.6 Struma river catchment

The chemical indicators involved were pH, dissolved oxygen (O_2) [mgO_2 L^{-1}], oxidation ability (OXIS) [mgO_2 L^{-1}], biological oxygen demand (BOD) [mgO_2 L^{-1}], chemical oxygen demand (COD) [mgO_2 L^{-1}], dissolved matter (DISS) [mg L^{-1}], nondissolved matter (N-DISS) [mg L^{-1}], chloride (Cl^-) [mg Cl L^{-1}], sulfate (SO_4^{2-}) [mg S-SO_4 L^{-1}], ammonium (NH_4^+) [mg N-NH_4 L^{-1}], nitrate (NO_3^-) [mg N-NO_3 L^{-1}], and iron (Fe^{2+}) [mg Fe L^{-1}]. Sample preparation and sample measurements are described in detail elsewhere (Bulgarian State Standards 1985).

The surface water quality norms were extracted from Directive 7 issued by the Bulgarian Ministry of Environment and Water (Bulgarian Ministry of Environment and Water, Directive 1997). Water quality norm I refers to drinking water, norm II to recreation usage, and norm III to irrigation and industrial usage. They are presented for comparison and interpretation along with the basic statistics of the input data in Table 14.4.

14.3.2.1 Results and Discussion

The dataset used for exploration consists of 68 objects as each one is described by 12 variables derived on annual basis (Table 14.4). The arrangement of the variable planes (Fig. 14.7) shows four well-defined groups of correlated variables and some variables with specific location. The first group includes the water quality parameters: nitrates and DISS. This fact is an indication for the similar information value of the two parameters. The second well-defined group reveals the connection between NH_4^+ and N-DISS, which could be easily related to common discharge

Table 14.4 Basic statistics ($n = 68$) and surface water quality norms (in mg L^{-1} except O_2)

					Surface water quality norms		
Parameter	Min	Max	Mean	St. dev.	I	II	III
pH	7.05	8.82	7.90	0.36	6.5–8.5	6.0–8.5	6.0–9.0
O_2 (in %)	3.83	10.68	7.85	1.54	75	40	20
BOD	2.33	46.44	9.75	7.95	5	15	25
OXIS	3.16	44.29	15.00	9.64	10	30	40
COD	7.81	266.67	40.19	45.53	25	70	100
DISS	111.00	679.33	347.78	145.23	700	1,000	1,500
N-DISS	4.17	180.06	57.19	33.82	30	50	100
Cl^-	11.67	44.99	26.82	7.87	200	300	400
SO_4^{2-}	24.67	174.00	64.20	23.88	200	300	400
NH_4^+	0.08	6.78	1.65	1.51	0.1	2	5
NO_3^-	0.04	17.58	4.87	3.40	5	10	20
Fe	0.08	2.09	0.47	0.42	1	2	5

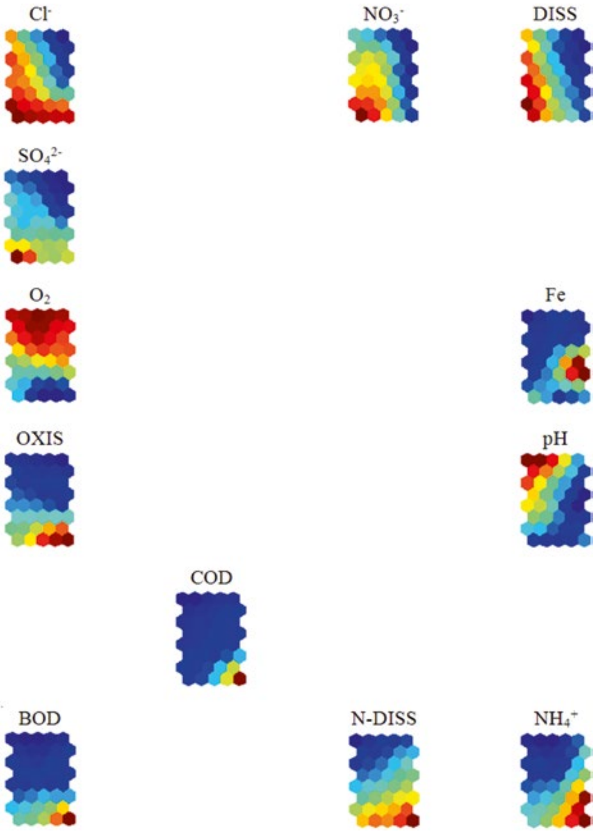


Fig. 14.7 Classification of water quality parameters

sources (Simeonova et al. 2003). The next group that includes the parameters Cl^-, SO_4^{2-}, O_2, and OXIS, with a positive correlation between Cl^- and SO_4^{2-} and negative correlation between O_2 and OXIS parameters. The fourth group is formed by Fe and pH, which are also negatively correlated. The specific positions of BOD and COD could be explained with their complex information ability to describe various, often uncontrolled pollutants and their transformations (Simeonova et al. 2003).

Using this classification scheme and data about the surface water quality norms in Bulgaria, a proper selection of surface water quality parameters could be done. Each well-defined group could be obviously presented only by one member of each group. Thus, NO_3^- and NH_4^+ water quality parameters were selected as representatives of the first two groups. These parameters were preferred to the other two (DISS and N-DISS), since they undergo a more reliable and accurate analytical determination and are directly attributed to specific anthropogenic influences along the river catchment (Simeonova et al. 2003). The OXIS was preferred as representative of the group also including Cl^-, SO_4^{2-}, and O_2. Since the majority of the objects monitored possess chloride and sulfate concentrations below the surface water quality norms (Bulgarian Ministry of Environment and Water, Directive 1997), so OXIS parameter describes the water quality more completely than the dissolved oxygen parameter. Because of their specific positions on the component plane and the presence of a pharmaceutical factory in the river catchment (EPA Memorandum 1988), the parameters BOD and COD were also included in the new water quality parameter set. The fourth group is not presented in the selected water quality parameter set, since Fe and pH do not discriminate well polluted (located in the bottom part of SOM) from unpolluted objects. The presented selection of five water quality parameters, NO_3^-, NH_4^+, OXIS, BOD, and COD, is in good agreement with previously obtained results where these selected parameters discriminate well polluted from less polluted objects (Astel et al. 2007).

The 68 sampling situations (number of sampling sites multiplied by the number of periods of sampling) described by five selected parameters were projected on a 2D map with dimensionality 5×8. The stations are grouped in 40 plane units (nodes). Each populated node could be used as an object, which includes the matching sampling situations (Table 14.5).

For parameter values, the map-trained vector of corresponding node will be bined in regular intervals using surface water quality norms from Table 14.1. Values below norm I will be set as 0, values between quality norms I and II will be set as 1, values between II and III will be set as 2, and values higher than III will be set as 3 (Table 14.5). This operation makes the dataset more homogeneous leading to reduction in the number of incomparable objects. Thus, a more useful and interpretable Hasse diagram analysis could be performed. This preprocessing of variables (bins partition) leads to formation of equivalence classes, where objects have one and the same numerical values for all variables. Hence, in Fig. 14.8, each one of the five equivalence classes is presented only by one object. The contents of equivalence classes is as follows: 3 = [3,11]; 20 = [20,22, 28, 29, 30, 35, 36, 38]; 23 = [23,31]; 25 = [25,33]; and 32 = [32,40].

Table 14.5 Objects (SOM nodes) representing the sampling stations with their coded surface water quality indexes

Objects	Belonging sampling stations	BOD	OXIS	COD	NH_4^+	NO_3^-
1	127_93, 127_94, 127_95	3	2	3	2	0
3	125_98	1	1	1	2	0
5	123_97, 123_98, 124_93, 124_94, 124_95, 124_97, 124_98	0	0	1	2	0
7	123_93, 403_93	1	0	1	1	0
8	122_98, 123_95, 126_98, 127_97, 293_98, 299_98, 403_97	1	0	0	1	0
11	403_94, 403_95	1	1	1	2	0
16	121_98, 293_97, 298_97	0	0	0	1	0
17	125_97, 127_90	2	2	2	2	1
20	122_93, 297_93	1	1	1	1	1
22	297_95, 298_97, 299_93	1	1	1	1	1
23	297_98	1	1	0	1	1
24	121_97, 122_97, 296_98, 299_97, 403_98	1	0	0	1	1
25	125_93, 127_89	2	2	1	2	1
27	125_92, 125_95, 126_93	1	1	1	2	1
28	126_94	1	1	1	1	1
29	122_94, 122_95, 126_95, 296_95, 297_94	1	1	1	1	1
30	296_97	1	1	1	1	1
31	124_92, 126_97, 297_97	1	1	0	1	1
32	123_92, 123_94	1	0	1	1	1
33	125_94	2	2	1	2	1
34	296_93, 296_94	2	1	1	2	1
35	126_92	1	1	1	1	1
36	120_91, 127_98	1	1	1	1	1
38	122_92, 299_94	1	1	1	1	1
40	121_91, 121_92, 123_89, 124_89, 299_95	1	0	1	1	1

The Hasse diagram presented for 14 representative objects (SOM nodes representing belonging sampling situations) and five variables (surface water quality parameters) is shown in Fig. 14.8. The order of the parameters in the diagram is as follows: BOD, OXIS, COD, NH_4^+, and NO_3.

The Hasse diagram shows nine levels, two maximal objects (1 and 17), and one minimal object (16) (in order terminology: a least element). It could be concluded that the two maximal objects determine two directed subgraphs in the diagram. The objects related to object 17 (e.g., 17-25-34-27-20-(23 or 32)-24) form a directed subgraph that could be conditionally named "nitrate." The sampling situations belonging to these objects are impacted by high nitrate loads. The other directed subgraph includes objects related to object 1 (1-3-5 and 1-3-7-8), where the rest of the surface water quality parameters (BOD, OXIS, COD, NH_4^+) exhibit elevated values. This separation makes it easy to distinguish patterns of sampling stations related to the influence of surface water quality parameters. Since the Hasse diagram

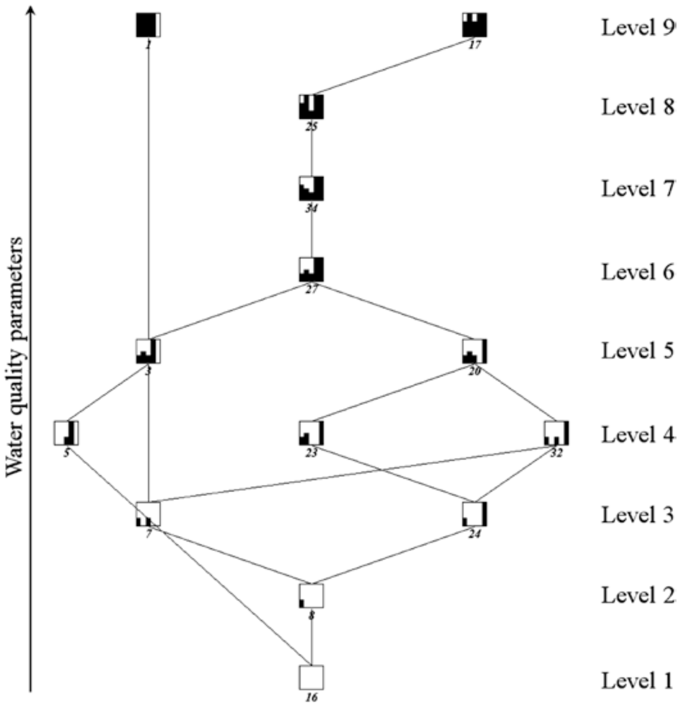


Fig. 14.8 Hasse diagram for the 25×5 selected dataset. The order (from *left* to *right*) of water quality parameters: BOD, OXIS, COD, NH_4^+, and NO_3^-. Data are normalized and represented as *bars* (software WHASSE)

Table 14.6 Location of sampling situations in the equivalence classes of Hasse diagram for sampling stations belonging to nitrate-directed subgraph

	Year						
Station	91	92	93	94	95	97	98
121	32	32				24	16
122		20	20	20	20	24	8
296			34	34	20	20	24
297			20	20	20	23	23

shows relations between objects, time series evaluation for each sampling station could also be performed. For instance, for sampling station 121 data for 4 years are available (Table 14.6). Sampling situations during 91 and 92 belong to object 32, data from 97 belongs to object 24, and data from 98 is linked to the less polluted object 16. It could be concluded that sampling station 121 belongs to the nitrate-directed subgraph and the nitrate loads are decreasing with time. The same pattern is found for the other three sampling stations (122, 296, and 297) that are affected by local steel and food industry and coal mining. Obviously, the high nitrate loads are due to the industrial activity. The time trend is not surprising, since in the last 15 years most of the factories were closed.

Table 14.7 Location of sampling situations in the equivalence classes of Hasse diagram for sampling stations "migrating" from nitrate to the directed subgraph controlled by the other water quality parameters

	Year						
Station	89	92	93	94	95	97	98
123	32	32	7	32	8	5	5
124	32	23	5	5	5	5	5
125		27	25	25	27	17	3
126		20	27	20	20	23	8
299			20	20	32	24	8

Table 14.8 Sensitivity analysis results

Variable	Sensitivity
BOD	4
OXIS	24
COD	28
NH_4^+	35
NO_3^-	50

For another group of sampling stations (123, 124, 125, 126, and 299), the decreasing trend of nitrate loads leads to their "transfer" from the "nitrate"-directed subgraph to the directed subgraph controlled by the other parameters (Table 14.7).

This situation occurs at the end of sampling period. Thus, the nitrate load "overlaps" the other anthropogenic impacts such as pharmaceutical wastes (site 126) and agricultural and farm stock activities (sites 123, 124, 125). The specificity of sampling station 127 should be mentioned, since it belongs to both maximal objects. High nitrate loads at the beginning of sampling period (89–90) located this site into the more impacted objects of the "nitrate"-directed subgraph, namely 17 and 25. During the next period (93–95), the site is shifted to the other directed subgraph (object 1) seriously affected by other surface water quality parameters. At the end of the sampling period, this particular site already belongs to the less polluted levels indicated by the Hasse diagram. Two other sampling stations (293, 298) are also located in the less polluted levels but for the whole sampling period. Sampling station 293 confirms its position as a background site in the monitoring net with sampling situations belonging to less affected equivalence classes 8 and 16. Sampling station 298 is not defined as background site but the close resemblance to site 293 could be sought in the sedimentation processes taking place in Pchelin reservoir placed just before the station. These processes obviously contribute to the water self-purification.

The sensitivity analysis of Hasse diagram is performed by calculation of dissimilarity W matrix for different combinations of $R-1$ variables and the calculated sensitivities are presented in Table 14.8. Therefore, NO_3^- is the most important parameter within the selected surface water quality parameter set for Hasse diagram obtained. The other three parameters, NH_4^+, COD, and OXIS, could be also treated as important quality parameters for decision-making. The low sensitivity of BOD is an indication that for environmental risk assessment COD seems to be a more appropriate surface water quality parameter.

14.4 Conclusions

The presented case studies indicate that a HDT could contribute to ERA of a certain marine or river environment by revealing of important environmental information like:

- The parallel monitoring of the content of heavy metals in various sediment samples and ecotoxicity measurements of the same samples; the routine approach uses separately the heavy metal indicator data and ecotoxicity data in attempts to detect some correlation between metal concentrations and levels of chronic or acute toxicity (e.g., mortality or EC50 values).
- The reliable interpretation of the two sets of data (metal concentrations and ecotoxicity tests) by the Hasse diagram technique; thus, the relationships between these two types of indicators of sediment quality is obtained and understood; additionally, a ranking of the pollution impact of the different sediment compartments could be achieved.
- The successive performance of SOM classification and HDT, which made it possible to determine specific patterns between the sampling locations within the monitoring net along the Struma River catchment for a long time period.
- The determination of limited number of water quality parameters (nitrates, ammonia, OXIS, COD, and BOD) being enough to reach a ranking in the set of objects (sampling situations) showing two specific pollution loads in the river system—nitrate load (due to industrial activity), decreasing with time, and a complex load of the other parameters related mainly to the other anthropogenic impacts.
- The classification scheme allowing to study the time series for the sites and to observe specific "shift" of sites from one pattern to another in time.

References

Astel A, Glosinska G, Sobczynski T, Boszke L, Simeonov V, Siepak J (2006) Chemometrics in assessment of sustainable development rule implementation. Cent Eur J Chem 4:543–664

Astel A, Tsakovski S, Barbieri P, Simeonov V (2007) Comparison of self organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Res 41:4566–4578

Aulinger A, Einax JW, Prange A (2004) Setup and optimization of a PLS regression model for predicting element contents in river sediments. Chemom Intell Lab Syst 72:35–41

Brodnjak-Voncina D, Dobcnik D, Novic M, Zupan J (2002) Chemometrics characterization of the quality of river water. Anal Chim Acta 462:87–100

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, Berlin

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg CEW (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

Bulgarian State Standards (1985) EN and ISO, Sofia

BulgarianMinistry of Environment and Water (1997) State directive 7 (Surface water quality norms), Sofia

Cun C, Vilagines R (1997) Time series analyses on chlorides, nitrates, ammonium and dissolved oxygen concentration in the Seine river near Paris. Sci Total Environ 208:59–569

Einax J, Zwanziger H, Geiss S (1997) Chemometrics in environmental analysis. Wiley, Weinheim

Einax JW, Aulinger A, Tümplng WV, Prange A (1999) Quantitative description of element concentrations in longitudinal river profiles by multiway PLS models. Fresenius J Anal Chem 363:655–661

EPA Memorandum (1988) The COD of pharmaceutical wastewaters

European Commission (2003) Technical guidance document in risk assessment. Part II. Joint Research Center, Ispra

Felipe-Sotelo M, Andrade JM, Carlosena A, Tauler R (2007) Temporal characterization of river waters in urban and semi-urban areas using physico-chemical parameters and chemometric methods. Anal Chim Acta 583:128–137

Giraudel J, Lek S (2001) A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol Model 146:329–339

Goetz R, Steiner B, Friesel P, Klaus R, Walkow F, Maass V, Reincke H, Stachel B (1998) Dioxin (PCDD/F) in the river Elbe—investigations of their origin by multivariate statistical methods. Chemosphere 37:1987–2002

Kohonen T (2001) Self-organizing maps, 3rd edn. Springer, Berlin

Kowalkowski T, Zbytniewski R, Szpenja J, Buszewski B (2006) Application of chemometrics in river water classification. Water Res 40:744–752

Kudłak B, Wolska L, Namiesnik J (2011) Determination of EC $_{50}$ toxicity data of selected heavy metals toward Heterocypris incongruens and their comparison to “direct- contact” and microbiotests. Environ Monit Assess 174:509–516

Lu R, Lo S (2002) Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. Water Res 40:2265–2274

Manganaro A, Ballabio D, Consonni V, Mauri A, Pavan M, Todeschini R (2008) The DART (decision analysis by ranking techniques) software. In: Pavan M, Todeschini R (eds) Scientific data ranking methods: theory and applications. Elsevier, Amsterdam, pp 193–207. Software is freely available on http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/doc/DART2_05Setup.zip. 10 Dec 2012

Mangiameli P, Chen S, West D (1996) A comparison of SOM neural network and hierarchical clustering methods. Eur J Oper Res 93:402–417

Marengo E, Gennaro M, Giacosa D, Abrigo C, Saini G, Avignone M (1995) How chemometrics can helpfully assist in evaluating environmental data from lagoon water. Anal Chim Acta 317:53–63

Marin-Guirao L, Lloret J, Marin A, Garcia G, Garcia Fernandez AJ (2007) Pulse-discharges of mining wastes into a coastal lagoon: water chemistry and toxicity. Chem Ecol 23:217–231

Mattikalli N (1996) Time series analysis of historical surface water quality data of the River Glen Catchment, UK. J Environ Manage 46:149–172

Mauri A, Ballabio D (2008) Similarity/diversity measure for sequential data based on Hasse matrices: theory and applications. In: Pavan M, Todeschini R (eds) Scientific data ranking methods: theory and applications. Elsevier, Amsterdam, pp 111–138

Mendiguchia C, Moreno C, Galindo-Riano D, Garcia-Vargas M (2004) Using chemometric tools to assess anthropogenic effects in river water. A case study: Guadalquivir (Spain). Anal Chim Acta 515:143–149

Mingoti S, Lima J (2006) Comparing SOM neural network with fuzzy c-means, K-means and traditional hierarchical clustering algorithms. Eur J Oper Res 174:1742–1759

Moore J, Ramamoorthy S (1984) Heavy metals in natural waters. Applied monitoring and impact assessment. Springer, New York, NY

Pardo R, Helena BA, Cazurro C, Guerra C, Deban L, Guerra CM, Vega M (2004) Application of two- and three-way principal component analysis to the interpretation of chemical fractionation results obtained by the use of the B.C.R. procedure. Anal Chim Acta 523:125–132

Simeonov V, Einax J, Stanimirova I, Kraft J (2002) Environmetric modelling and interpretation of river water monitoring data. Anal Bioanal Chem 374:898–905

Simeonova P, Simeonov V, Andreev G (2003) Water quality study of the Struma River basin, Bulgaria. Cent Eur J Chem 2:121–136

Singh KP, Malik A, Basant N, Singh VK, Basant A (2007) Multi-way data modeling of heavy metal fractionation in sediments from Gomti River (India). Chemom Intell Lab Syst 87:185–193

Stefanov S, Simeonov V, Tsakovski S (1999) Chemometrical analysis of waste water monitoring data from Yantra river basin, Bulgaria. Toxicol Environ Chem 70:473–482

Tsakovski S, Simeonov V (2008) Hasse diagrams as exploratory tool in environmental data mining: a case study. In: Owsinski J, Brüggemann R (eds) Multicriteria ordering and ranking: partial orders, ambiguities and applied issues. SRI, PAS, Warsaw

Tsakovski S, Simeonov V, Stefanov S (1999) Time-series analysis of long-term water quality records from Yantra river basin, Bulgaria. Fresenius Environ Bull 8:28–36

Tsakovski S, Kudłak B, Simeonov V, Wolska L, Garcia G, Dassenakis M, Namiesnik J (2009) N-way modelling of sediment monitoring data from Mar Menor lagoon, Spain. Talanta 80:935–941

Tsakovski S, Kudlak B, Simeonov V, Wolska L, Garcia G, Namiesnik J (2012) Relationship between heavy metal distribution in sediment samples and their ecotoxicity by the use of the Hasse diagram technique. Anal Chim Acta 719:16–23

U.S. EPA (Environmental Protection Agency) (2005) Predicting toxicity to amphipods from sediment chemistry. National Center for Environmental Assessment, Washington, DC, EPA/600/R-04/030. Available from National Technical Information Service, Springfield, VA, and online at http://www.epa.gov/ncea

Vesanto J (1999) SOM-based data visualization methods. Intell Data Anal 3:111–126

Voigt K, Brüggemann R, Scherb H, Shen HQ, Schramm KW (2010) Evaluating the relationship between chemical exposure and cryptorchidism. Environ Model Softw 25:1801–1812

Wunderlin D, Diaz M, Ame M, Pesce S, Hued A (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Siquia River Basin (Cordoba—Argentina). Water Res 35:2881–2894

Yan X, Chen D, Chen Y, Hu S (2001) SOM integrated with CCA for the feature map and classification of complex chemical patterns. Comput Chem 25:597–605

Part V
Software Aspects

Chapter 15
PARSEC: An R Package for Poset-Based Evaluation of Multidimensional Poverty

Marco Fattore and Alberto Arcagni

Abstract The paper introduces PARSEC, a new software package implementing basic partial order tools for multidimensional poverty evaluation with ordinal variables. The package has been developed in the R environment and is freely available from the authors. Its main goal is to provide socio-economic scholars with an integrated set of elementary functions for multidimensional poverty evaluation, based on ordinal information. The package is organized in four main parts. The first two comprise functions for data management and basic partial order analysis; the third and the fourth are devoted to evaluation and implement both the poset-based approach and a more classical counting procedure. The paper briefly sketches the two evaluation methodologies, illustrates the structure and the main functionalities of PARSEC, and provides some examples of its use.

15.1 Introduction

PARSEC[1] is a new R (R Core Team 2012) package implementing basic partial order tools for multidimensional poverty evaluation with ordinal variables. Poset theory use overcomes the drawbacks of classical evaluation procedures, which prove scarcely effective and often inconsistent for handling ordinal data (Fattore et al. 2012). The poset-based approach has been primarily developed for poverty and material deprivation assessment (Fattore et al. 2011a,b), but it may be virtually applied to any kind of evaluation problem with ordinal variables, like assessing quality-of-life, well-being, or customer satisfaction. For the sake of completeness,

[1] PARtial orders in Socio-EConomics.

M. Fattore (✉) • A. Arcagni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy
e-mail: marco.fattore@unimib.it

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_15, © Springer Science+Business Media New York 2014

PARSEC implements also the counting approach to multidimensional poverty evaluation, developed by the Oxford Poverty & Human Development Initiative (OPHI) group (Alkire and Foster 2011a). This procedure is gaining relevance at international level and may be used as a benchmark for the poset approach. The paper is organized as follows: Sect. 15.2 gives a brief sketch of the poset-based and OPHI evaluation procedures; Sect. 15.3 illustrates the main functionalities of PARSEC; Sect. 15.4 provides some scripts, showing PARSEC in action and giving some ideas of its performances; Sect. 15.5 discusses the improvements to be implemented in the next release of the package; Sect. 15.6 concludes and the Appendix provides a list of the functions currently available in PARSEC.

15.2 Poset-Based and Counting Evaluation Methodologies

Multidimensional poverty evaluation increasingly involves ordinal variables. This poses some critical methodological problems: (1) classical evaluation procedures based on variable aggregation are not directly applicable to ordinal data and (2) data are often truly multidimensional and variable interdependencies are too weak to achieve any dimensionality reduction, even conceptually. The scientific community is currently debating these issues and while some scholars stress the relevance of getting synthetic indicators anyway (e.g., for policy-making purposes), others argue their consistency and suggest relying on multidimensional dashboards (Ravaillon 2011). These difficulties are partly unavoidable, due to the complexity of the problems at hand; but they are also amplified by the use of unsuitable statistical tools, borrowed from classical multivariate analysis and based on linear algebra. Linear algebra tools break down when addressing ordinal variables and produce inconsistencies that may be mistaken for an intrinsic impossibility to get well-founded results. The use of partial order theory clarifies that this is not the case, it shows that the computation of synthetic indicators need not require variable aggregation and paves the way to alternative and consistent evaluation procedures. An example of the possibilities offered by partial order theory is provided by the poset-based methodology implemented in PARSEC and briefly introduced in the following. The methodology provides a consistent framework for ordinal evaluation problems, preserving the logic of classical procedures, but using partial order theory for data representation and information extraction. In the following, we limit ourselves to a very essential introduction to the methodology. More complete presentations can be found in Fattore et al. [2011a,b, 2012].

Let $v_1, \ldots, v_k$ be k ordinal variables representing poverty dimensions (we assume that lower degrees of $v_1, \ldots, v_k$ represent higher deprivations). Each variable is recorded on a different scale, possibly with a different number of degrees $m_1, \ldots, m_k$. Each statistical unit in the population is scored against the k variables. The vector $\mathbf{p} = (p_1, \ldots, p_k)$ of k scores associated to a statistical unit is called a *profile*. The set of possible profiles P has cardinality $|P| = m_1 \cdot m_2 \cdot \ldots \cdot m_k$, even if some profiles may not be observed within the population. The set P is naturally turned into a partially ordered set $(P, \lhd)$ putting

$$\mathbf{p} \lhd \mathbf{q} \Leftrightarrow p_i \leq q_i \ \ \forall i \in 1, \ldots, k \tag{15.1}$$

where $\mathbf{p}$ and $\mathbf{q}$ are elements of P. An *evaluation function* $E\ v\ a\ l(\cdot)$ is defined to assign a degree of poverty in $[0, 1]$ to each profile. In fact, the existence of incomparabilities among profiles leads quite naturally to describing poverty on a continuous scale (in a fuzzy spirit). Poset $(P, \lhd)$ is just a mathematical structure; as such, it conveys no information on the degree of poverty of its elements. To transform P in a tool for poverty evaluation, exogenous information is embedded into it choosing a *threshold* τ, that is, by selecting a minimal set of profiles[2] considered as poor and scored 1 by the evaluation function. The choice of the threshold allows the evaluation function to be extended to all of the profiles in P, according to the following procedure:

1. consider the set $L\ E$ of linear extensions of P;
2. for any linear extension $l \in L\ E$, assign poverty score 1 to all the profiles that are below an element of τ in l;
3. assign to profiles of P a final poverty score averaging the scores they get on the elements of $L\ E$.

In practice, given a profile $\mathbf{p}$, $E\ v\ a\ l(\mathbf{p})$ is computed as the relative frequency of linear extensions where $\mathbf{p}$ is below an element of the threshold. By construction, $E\ v\ a\ l(\cdot)$ scores to 1 all the profiles in τ or in the downset of τ and to 0 all the profiles in the intersection of the upsets of the elements of τ. All of the other profiles in P are assigned scores in $]0, 1[$. Finally, each statistical unit in the population is assigned the poverty degree of the profile it shares. Once the population has been assessed in this way, classical overall indicators may be computed, particularly the Head Count Ratio, here defined as the average degree of poverty in the population.

The poset-based methodology provides a radical alternative to classical aggregative procedures. Among the latter ones, the counting approach developed by the OPHI group (Alkire and Foster 2011a) is gaining more and more relevance and its application is spreading. One of its merits is to provide a general and unified framework for multidimensional poverty assessment, even if it suffers from drawbacks typical of aggregative methodologies (Fattore et al. 2012). Due to its importance, the OPHI procedure is implemented in PARSEC, also to provide a benchmark for the poset-based approach.[3] The OPHI procedure is conceptually quite simple. Let $v_1, \ldots, v_k$ be k poverty dimensions, as before. A set $c_1, \ldots, c_k$ of k cutoffs is exoge-

[2] In a multidimensional setting, the threshold need not be composed of just one profile, but may comprise several profiles, since the shapes of poverty can be different and incomparable. It may be proved that a threshold can be always chosen as an antichain (Fattore et al. 2011a).

[3] The OPHI approach can be applied also when cardinal variables are of concern, but here we limit the discussion to the ordinal case.

nously defined, identifying a different poverty threshold for each evaluation dimension. A statistical unit scoring a degree d_i lower than c_i is considered as deprived on dimension v_i. Statistical units are classified as definitely poor if the number of dimensions they are deprived on equals or exceeds an overall cutoff c, also to be defined exogenously. In practice, the OPHI approach defines a yes-or-no evaluation function (more precisely, it defines an *identification* function) and classifies statistical units in just two classes, the poor and the non-poor. The final output of the OPHI procedure comprises the Head Count Ratio, that is, the fraction of statistical units scored as poor, and the Adjusted Head Count Ratio, which is the product of the Head Count Ratio and the average number of deprivations suffered by poor statistical units. This last indicator is of interest since it helps to realize the severity of deprivation in a given population. A complete description and discussion of the methodology can be found in Alkire and Foster [2011a] and Alkire and Foster [2011b]. It is interesting to note that the OPHI approach can be cast in poset terms and can be seen as a special case of the poset-based methodology (Fattore et al. 2011b).

15.3 The Structure of PARSEC and Its Main Functionalities

PARSEC is organized in four main sections, each comprising a set of functions for specific tasks:

1. Data management.
2. Basic poset analysis.
3. Poset-based evaluation.
4. OPHI counting approach.

In the following we give a brief account of each section, referring to the Appendix for a complete list of the functions currently available in the package.

This set of functions is used to build partial orders, possibly out of original data. Function `var2prof` allows the user to specify an arbitrary number of ordinal variables, each coded with a different scale, and produces the list of all the profiles built on them. It is very useful for building posets from scratch. It is also possible to assign a weight to each profile (usually the number of units sharing the profile). Function `pop2prof` extracts all the unique profiles out of a population of statistical units assessed against a set of ordinal variables. It also assigns to each observed profile the correspondent absolute frequency. Once the set of profiles is obtained, it can be turned into a partial order according to expression (15.1). The square binary matrix (usually labeled Z) representing the partial order (i.e., the incidence matrix of the corresponding Hasse diagram) is obtained through function `getzeta`.

The functions of this section manage posets and allow the investigation of their basic features. PARSEC represents posets in matrix terms, so many of its functions rely on matrix calculus. Through functions like `binary`, `reflexivity`, `antisymmetry`, and `transitivity`, one may check whether the input square matrix is binary and represents a reflexive, antisymmetric, or transitive relation. Checking these properties jointly, functions is.preorder and is.partialorder verify whether the input matrix represents a preorder or a partial order. Often, it is useful to handle directly the cover relation generating the partial order. The cover relation matrix may be obtained from the partial order matrix using incidence2cover, while cover2incidence performs the opposite (which is useful, since often it is easier to specify a partial order through the cover relation). Maximal and minimal elements of a poset are directly obtained invoking maximal and minimal. The heights of poset elements are obtained through heights, and similar functions exist to compute the depths and the levels of poset elements (see Patil and Taillie 2004 for appropriate definitions). The poset-based evaluation methodology draws upon the concepts of antichain, downset and upset. PARSEC thus provides functions to get the set of incomparable elements of a given poset element (incomp), to check whether a list of elements forms a downset (is.downset) or an upset (is.upset), to return the downset or the upset generated by a given set of elements (downset and upset, respectively) and to identify the antichain generating a downset (gen.downset) or an upset (gen.upset).

PARSEC implements the poset-based approach to evaluation through function evaluation. Given the partial order matrix and the selected threshold, evaluation returns the evaluation function, the rank distribution of the profiles and the frequency distribution of the distances (rank differences) between a profile and the threshold.[4] Given the number of statistical units sharing each profile, the Head Count Ratio can then be easily obtained. Function evaluation is based on a precompiled C implementation of the Bubley–Dyer algorithm for (almost) uniform sampling of linear extensions (Bubley and Dyer 1999). Even in medium size posets (e.g., 40–50 elements), the computation of the evaluation function requires sampling several of hundred million linear extensions, a task that an interpreted scripting language like R could only accomplish in a very long time. The pre-compiled C routine decreases dramatically the computation time and allows the evaluation methodology to be applied to larger posets, composed of some hundreds elements. The next section provides some examples of the use of evaluation combined with other PARSEC functions; some tests are also presented to give an idea of the package performances.

Function count implements[5] in a single call the computations involved in the OPHI counting approach (Alkire and Foster 2011a). Passing to count the profiles of the statistical units, the vector of cutoffs on the evaluation dimensions and the overall cutoff, a complete output is returned comprising the Head Count Ratio and the Adjusted Head Count Ratio. As already mentioned, it is easily seen that the OPHI approach can be considered as a special case of the poset-based methodology. The link between the two methodologies is given by function count2threshold

[4] Precisely, for any linear extension, the differences between the rank of the higher ranked element of the threshold and the ranks of the other profiles are computed.

[5] The OPHI approach can be applied also when cardinal variables are of concern. PARSEC implements the methodology for ordinal variables only.

which returns the threshold (in profile terms) generating the set of poor profiles identified by `count`. The returned threshold can be directly used in `evaluation`, to compare the results of the two methodologies.

15.4 Some Examples

In this section, we provide some application examples of PARSEC. First, we give a simulated example of multidimensional poverty evaluation using both the OPHI and the poset-based approaches, comparing the results. Then we illustrate the computation of the evaluation function through `evaluation` and show how increasing the number of sampled linear extensions leads to more accurate results. Finally, we test the performances of `evaluation` applying it to a sequence of posets of increasing complexity, comparing the computation times.

All of the computations described in the following were done on an Intel®; Core$^{\text{TM}}$2 Duo CPU E8400 @ 3.00GHz ×2, RAM 1.9 GB, equipped with Linux 32 bit operative system.

We simulate a multidimensional poverty evaluation process, on a poset of 40 profiles built out of three ordinal variables recorded on scales with number of degrees 4, 5, and 2, respectively. The simulation compares the OPHI and the poset-based approaches and is organized in three steps:

1. Data definition.
2. Implementation of the OPHI procedure.
3. Implementation of the poset-based procedure and comparison of the results.

Step 1: Data definition. To build the poset, we first define a vector `vs` whose length is the number of variables and whose components are the number of degrees of each variable

```
> vs <- c(4, 5, 2)
```

Vector `vs` is passed to function `var2prof` which generates all 4 ×5 ×2=40 poset profiles, computing all the combinations of variable degrees. A vector `freq` of 40 randomized integer numbers is also passed to `var2prof` to simulate a distribution of frequencies on the profiles

```
> prof <- var2prof(vs, freq = rbinom(prod(vs), 100, .5))
```

Step 2: Implementation of the OPHI procedure. To apply the OPHI procedure to the simulated data, the vector `var_cut` of cutoffs and the overall cutoff `over_cut` are first defined

```
> var_cut <- c(1, 1, 1)
> over_cut <- 2
```

According to `var_cut`, a statistical unit is deprived on a variable if the corresponding degree is lower than 1, that is, if it scores 0; according to `over_cut`, a statistical unit is considered as definitely deprived if it scores 0 on at least two variables out of three. The OPHI methodology is now applied to data, calling function `count`

```
> countres <- count(prof, var_cut, k=over_cut)
```

Typing `countres$H` returns the Head Count Ratio

```
> countres$H
[1] 0.2190476
```

Step 3: Implementation of the poset-based procedure. To apply the poset-based evaluation procedure, partial order matrix `Z` is first built out of the profiles

```
> Z <- getzeta(prof)
```

Next a threshold must be defined. To compare the results of the two methodologies, the threshold is chosen to be that implicitly defined in the OPHI approach. This is achieved by typing

```
> threshold <- count2threshold(countres, prof, Z)
```

It turns out that the threshold is composed of profiles 300, 040, 001

```
 > prof$profiles[threshold,]
    Var1 Var2 Var3
P04    3    0    0
P17    0    4    0
P21    0    0    1
```

The evaluation function is computed through `evaluation`, passing to it `Z`, together with `threshold` and other two parameters, namely, an arbitrary linear extension `lin_ext` to initialize the Bubley–Dyer algorithm and the number of iterations `nit` to achieve an almost uniform sampling (the number of iterations depends upon the parameter `error`, i.e., the acceptable maximum total variation distance from a uniform distribution). The initializing linear extension is computed by typing

```
> lin_ext <- lingen(Z)
```

Selecting a maximum total variation distance from uniformity of 10^{-5}

```
> error <- 10^(-5)
```

`nit` is estimated using a formula of Karzanov and Khachiyan (see Bubley and Dyer 1999)

```
> nit <- floor(n^5*log(n)+n^4*log(error^(-1)))
> nit
[1] 407214345
```

where `n` is the number of profiles (here, 40). Finally calling

```
> eval <- evaluation(Z, lin_ext, nit, threshold)
```

the evaluation function is obtained (computation takes about 2 min).

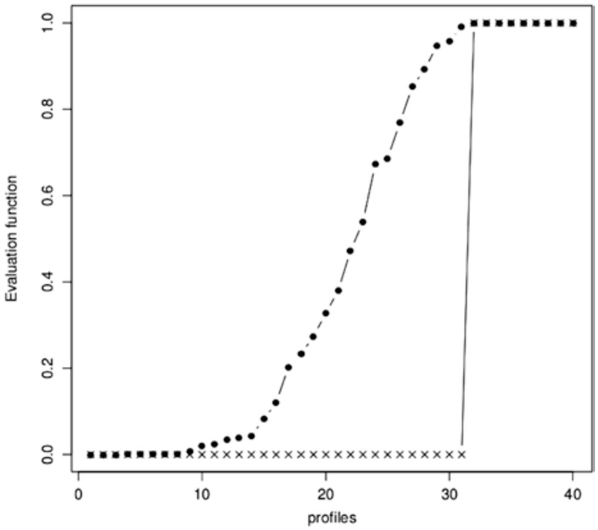


Fig. 15.1 Poset-based (*circles*) and OPHI (*crosses*) evaluation functions

The plot of the evaluation functions obtained from `count` and `evaluation` is depicted in Fig. 15.1 and is obtained by

```
 > ord <- order(eval$poorfreq)
> plot(eval$poorfreq[ord], type="b", pch=16,
    ylab="Evaluation function", xlab="profiles")
> lines(1:40, countres$Z_k[ord], type="b", pch=4)
```

The two evaluation functions coincide on profiles scored to 1 or 0 by the poset-based approach, but differ on all of the other profiles. This difference has a great impact on the Head Count Ratio (`hc`), which turns out to be much greater than that computed by `count`

```
>hc<-as.vector(eval$poorfreq%*%prof$freq/sum(prof$freq))
> hc
[1] 0.4623521
```

We now show how increasing the number of sampled linear extensions reflects on the estimation of the evaluation function. We consider the same poset and the same threshold introduced in the previous paragraph, running `evaluation` seven times, with increasing sample sizes of 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, and 10^9 linear extensions. The computation times range from 0.001 s (sample of 10^3) to about 6 min (sample of 10^9) and increases almost linearly with sample size. The results are displayed in Fig. 15.2. In addition, Table 15.1 reports the distances (measured as the maximum point-wise absolute differences) of the first six evaluation functions from the last, computed on 10^9 linear extensions.

As can be noticed, the evaluation function estimated with 10^4 iterations is worse than that estimated with 10^3. This is due to randomness and to the fact the 10^3 and 10^4 are far below the number of linear extensions needed to approach a uniform distribution.

To check the computational performances of `evaluation` as poset complexity increases, we have run it on a sequence of nine posets, built as the product of

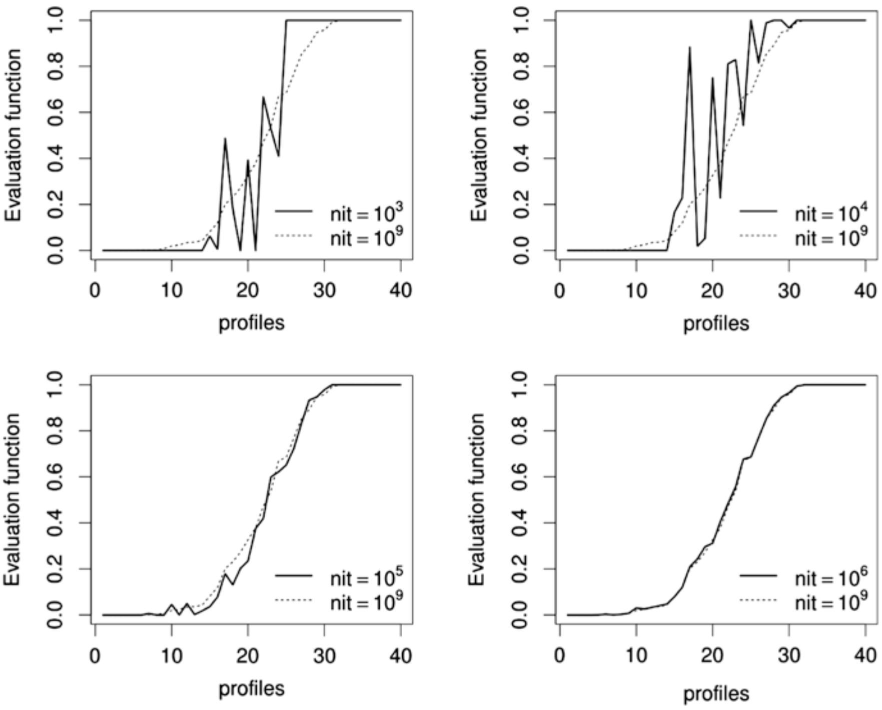


Fig. 15.2 Convergence of the evaluation function as number of iterations (nit) increases

Table 15.1 Distances from the evaluation function computed on a sample of 10^9 linear extensions

Sample size	10^3	10^4	10^5	10^6	10^7	10^8
Distance	0.378	0.682	0.100	0.026	0.017	0.005

Table 15.2 Computation time (in minutes) to run `evaluation` sampling 10^9 linear extensions, as poset complexity increases

Poset	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
Minutes	2	2	3	5	6	10	21	75	183

$2, 3, \ldots, 10$ two-element chains. The posets are labeled $2^2, 2^3, \ldots, 2^{10}$. In each case, we have extracted 10^9 linear extensions, with a fixed threshold composed of a single poset element. Before illustrating the results, it must be considered that `evaluation` computes various statistics on the input poset, some of which require the iterative execution of many `if-then` statements, in addition to the computational burden due to the core Bubley–Dyer algorithm. The computation time of these additional calculations depends critically, among other things, upon the cardinality and the structure of the poset, explaining the figures reported in Table 15.2. As can be seen, the computation time rapidly increases as the number of poset elements grows, reaching 183 min for poset 2^{10}, which is composed of 1,024 elements.

15.5 Planned Developments and Improvements

PARSEC is currently in its beta version. Some improvements are in order and will be implemented in the next release. They pertain to usability, to the addition of functionalities to handle large posets, and to the introduction of graphical capabilities.

Currently, PARSEC provides all the basic functions needed to implement the poset-based evaluation methodology introduced in the paper. However, to perform concrete computations some scripting is still needed by users, who must combine together different functions to obtain the required outputs. Although this allows for great flexibility, it may cause some problems to non-programmers. For this reason, some "macro" functions are being implemented to obtain the desired results in a single call, similarly to the `count` function. The aim is to reduce to a minimum the need for coding; similarly, a graphical user interface will be considered (see, for example, the R package Rattle Williams 2009), to assist non-expert users.

PARSEC represents poset by means of matrices and employs simple matrix computations (Patil and Taillie 2004) to address poset analysis. This makes the package quite easy to develop and to maintain and sufficiently effective for most real applications, but also makes PARSEC scarcely scalable. This could be a problem, when handling posets with several hundreds or thousands of elements. In this case, more sophisticated programming techniques will be needed, to effectively manage large sparse matrices. However, the main issue pertains to the way the evaluation function is computed. At present, the computation of the evaluation function is implemented by sampling linear extensions through the Bubley–Dyer algorithm. Although this is the fastest algorithm currently available, the number of linear extensions to be sampled and the computation time increase steeply with the complexity of the poset and in practice huge posets cannot be handled this way. In socio-economic applications, one may work with posets comprising many hundreds of elements, since statistical units are often assessed against ordinal variables coded in up to 10°. Evaluation function computation in this kind of partial orders is better addressed by using analytical formulas which provide approximations to mutual ranking probabilities. Different approximation formulas can be found in literature that can be used for our purposes, see, for example, De Loof et al. [2008], De Loof [2010]. Actually, available formulas are designed to approximate mutual ranking probabilities of two elements at a time, while the evaluation methodology requires computing the mutual ranking probability of an element with respect to an antichain. Thus, some adaptation is required before implementation.

Visualizing data is one of the most effective ways to ease user experience. In the near future, a set of functionalities will be implemented to give standard graphical representations to PARSEC outputs and to draw Hasse diagrams, projecting on their nodes various kinds of information, such as the value of the evaluation function and the corresponding number of statistical units, or inserting pictorial representations of the profiles [e.g., in the spirit of graphical representations available in the Kohonen package (Werhens and Buydens 2007)].

15.6 Conclusion

It is progressively clear to scholars and to decision-makers that addressing socio-economic issues in modern societies, and evaluation issues in particular, requires a change of paradigm. It is no longer the time to "aggregate and average," to produce macro-indicators that would fail to represent real societies and their structural dynamics. Instead, it is the time to represent and make explicit "shapes" and "patterns," "similarities" and "dissimilarities," "structural affinities" and "structural differences," and "nuances" and "complexities." Partial order theory surely plays an important role in this challenge and PARSEC can spread its use across the community of socio-economic scholars. PARSEC surely needs to be improved and extended in many directions, and we hope to get suggestions from users to fix possible bugs and to add new functionalities. In fact, it is our intention to transform PARSEC into an official and publicly available R package, publishing it on the CRAN web site. This task will be accomplished in the near future, after completing the test of the beta version. At present, PARSEC is freely available from the authors together with the technical documentation, for both Windows and Linux operating systems.

15.7 Appendix: Function List

Here, we list the functions currently available in PARSEC. The list is not for technical reference, but to give an idea of the scope and the capabilities of the package. The list is organized according to the four PARSEC sections.

15.7.1 Data Management

`var2prof` Generates all possible profiles out of *k* ordinal variables. A vector of frequencies may also be passed, for subsequent use.

`pop2prof` Reads a dataframe comprising statistical unit scores on *k* variables and extracts all unique profiles together with the corresponding frequencies.

`getzeta` Generates the partial order matrix (i.e., the incidence matrix of the corresponding Hasse diagram) according to (1), from the profile list.

`popelem` Associates each observed statistical unit with the index (i.e., the row or column of the partial order matrix) of the corresponding profile.

15.7.2 Basic Poset Analysis

`binary` Checks whether a matrix is binary.

`reflexivity` Checks whether a binary relation is reflexive.

`antisymmetry`	Checks whether a binary relation is antisymmetric.
`transitivity`	Checks whether a binary relation is transitive.
`is.preorder`	Checks whether a binary relation is a preorder.
`is.partialorder`	Checks whether a binary relation is a partial order.
`validate.partialorder`	Checks whether an input binary matrix defines a partial order and validates it as the incidence matrix of the corresponding Hasse diagram. If the input matrix does not represent a poset, the function returns which poset properties are not fulfilled.
`incidence2cover`	Builds a cover matrix from the partial order matrix.
`cover2incidence`	Builds a partial order matrix from the cover matrix.
`transitiveClosure`	Computes the transitive closure of a binary relation.
`upset`	Returns the upset of a set of elements.
`downset`	Returns the downset of a set of elements.
`is.upset`	Checks whether a set of elements of a poset is an upset.
`is.downset`	Checks whether a set of elements of a poset is a downset.
`gen.upset`	Returns the antichain generating the input upset.
`gen.downset`	Returns the antichain generating the input downset.
`incomp`	Returns the set of elements incomparable with a selected poset element.
`minimal`	Returns the minimal elements of a poset.
`maximal`	Returns the maximal elements of a poset.
`heights`	Returns the heights of the elements of the poset in the corresponding Hasse diagram.
`depth`	Returns the depths of the elements of the poset in the corresponding Hasse diagram.
`levels`	Returns the levels of the elements of the poset in the corresponding Hasse diagram.
`colevels`	Returns the colevels of the elements of the poset in the corresponding Hasse diagram.
`height.poset`	Returns the height of a poset.
`synopsis`	Gives a summary of the input poset features.

15.7.3 Poset-Based Evaluation

`lingen` Returns a linear extension extracted by the input poset.

`linzeta` Returns the (partial) order matrix of a linear extension, given the profile indexes of the original poset.

`evaluation` From (1) the partial order matrix of a poset, (2) a linear extension of the poset, (3) the number of linear extensions to be sampled and (4) a threshold (as an antichain), the function returns (a) the estimated evaluation function, (b) the rank frequency distribution of each element of the poset, (c) the frequency distribution of the rank differences between each element of the poset and the higher ranked element of the threshold, and (d) the last linear extension sampled (that may be used to initialize other executions of the function).

15.7.4 OPHI Counting Approach

`count` Given a population, the single variable cutoffs and the overall cutoff, the function implements the OPHI procedure and returns, among other results, (a) the indexes and the number of statistical units classified as poor, (b) the number of deprivations suffered by each statistical unit or by each profile, (c) the deprivation map (which observations or profiles are deprived on which dimensions), (d) the Head Count Ratio, (e) the Average Deprivation Share, and (f) the Adjusted Head Count Ratio.

`count2threshold` Returns the profile threshold determining the poor profiles in the OPHI procedure.

References

Alkire S, Foster J (2011) Counting and multidimensional poverty measurement. J Public Econ 96(7–8):476–487

Alkire S, Foster J (2011) Understandings and misunderstandings of multidimensional poverty measurement. J Econ Inequal 9(2):289–314

Bubley R, Dyer M (1999) Faster random generation of linear extensions. Discrete Math 201: 81–88

De Loof K (2010) Efficient computation of rank probabilities in posets. Ph.D. dissertation. https://biblio.ugent.be/publication/874495

De Loof K, De Baets B, De Meyer H (2008) Properties of mutual rank probabilities in partially ordered sets. In: Owsinski JW, Brueggemann R (eds) Multicriteria ordering and ranking: partial orders, ambiguities and applied issues. Polish Academy of Sciences, Warsaw

Fattore M, Brueggemann R, Owsiński J (2011) Using poset theory to compare fuzzy multidimensional material deprivation across regions. In: Ingrassia S, Rocci R, Vichi M (eds) New perspectives in statistical modeling and data analysis. Springer, Berlin

Fattore M, Maggino F, Greselin F (2011) Socio-economic evaluation with ordinal variables: integrating counting and poset approaches. Stat Appl, Special Issue, 31–42

Fattore M, Maggino F, Colombo E (2012) From composite indicators to partial order: evaluating socio-economic phenomena through ordinal data. In: Maggino F, Nuvolati G (eds) Quality of life in Italy: research and reflections. Social indicators research series, vol. 48. Springer, Berlin

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

R Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org/

Ravaillon M (2011) On multidimensional indices of poverty. J Econ Inequal 9(2):235–248

Werhens R, Buydens LMC (2007) Self- and super-organizing maps in R: the kohonen package. J Stat Softw 21(5)

Williams GJ (2009) Rattle: a data mining GUI for R. R J 1/2:45–55

Chapter 16
Higher Order Indicator with Rank-Related Clustering in Multi-indicator Systems

Wayne L. Myers and Ganapati P. Patil

Abstract We extend exploration and application of a compound ranking regime for multi-indicator systems based on partial order theory that coordinates principal down-set, up-set, and comparability. We use this "balance of normalized definitive status (BONDS)" ranking in conjunction with augmented hierarchical clustering for comparing clusters as graded groups to obtain information relevant to number of clusters, group-wise ordering, and interaction of indicators. A case study is conducted with localities as instances (objects) and percentages for kinds of land cover as indicators. Hierarchical clustering is accomplished with the **hclust** facility of **R** software in conjunction with customized computational support in **R**.

16.1 Introduction

Four primary posetic aspects that emerge from basic analysis of multi-indicator systems under partial order theory are Hasse diagram, principal down-set, principal up-set, and (in)comparability (Brüggemann and Patil 2010, 2011; Brüggemann and Voigt 2008; Patil and Taillie 2004). While the first of these is most informative for small numbers of objects (or instances), the complexity of the Hasse diagram is compounded rapidly with increasing numbers of instances to the point of essentially losing interpretability. The other three aspects as (principal) down-set=|$O(x)$|, up-set=|$F(x)$|, and comparability retain their interpretability although requiring

W.L. Myers (✉)
Penn State Institutes of Energy and Environment, The Pennsylvania State University,
5 Land & Water Research Building, University Park, PA 16802, USA
e-mail: wlm@psu.edu

G.P. Patil
Center for Statistical Ecology and Environmental Statistics, Department of Statistics,
The Pennsylvania State University, 421 Thomas Building, University Park, PA 16802, USA
e-mail: gpp@stat.psu.edu

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_16, © Springer Science+Business Media New York 2014

additional computational time. Myers and Patil (2013b) formulate a "balance of normalized definitive status (BONDS)" approach to compound ranking of instances that coordinates these latter three posetic aspects. BONDS-based ranks effectively serve as a single higher order scalar indicator that captures prominent implications of the multi-indicator system.

Additional insight to accompany the BONDSranks is desirable regarding interaction of the actual multiple indicators as it relates to the compound ranking. Here we address this need by hierarchical clustering conducted on an augmented matrix of multiple indicators, whereby the BONDS ranking augments the actual indicators and thereby conditions the clustering in a rank-related manner. The imbedded ranks are used to direct displays that facilitate choosing an appropriate number of clusters, aggregate ordering of the clusters, and compositional comparisons of the clusters. Hierarchical clustering is accomplished with the **hclust** facility of **R** software (Allerhand 2011; **R** Development Core Team 2008; Short 2009; Venables et al. 2005) in conjunction with customized computational support in **R**. **R** commands are shown for clustering and displays. Customized **R** facilities for obtaining BONDSranks are given by Myers and Patil (2013b).

16.2 Vicinity Variates for Octagonal Area Objects

For consistency of comparison, we pursue a geographic multi-indicator context used by Myers and Patil (2013b) for elucidating the BONDS approach. The context is a two-county region in south-central Pennsylvania, USA. The counties involved are Blair and Huntingdon (Fig. 16.1). For a particular biogeomorphic region, different kinds of land cover have intrinsic implications as indicators for landscape ecology. This region has forest (Pct40) as a natural land cover if ecological succession is allowed to proceed without intervention. Natural disturbance dynamics have transitional cover (Pct33) where disturbed patches are again reverting to forest. Humans are a major (non-natural) disturbance agent in this region. The most extreme and enduring scars of humanization in these landscapes are quarries, strip mines, and gravel pits (Pct32). Development (Pct20) along with urban/recreational grasses (Pct85) constitutes a prevalent aspect of humanization as long-term disturbance. Agriculture comprises a second prominent aspect of humanization comprised of crops (Pct82) and hay/pasture (Pct81) with crops involving annual exposure of the soil surface to erosion in this relative steep terrain. Loss of natural wetlands to humanization has led to restrictions on further conversion of woody wetlands (Pct91) and emergent herbaceous wetlands (Pct92) which together currently comprise a small fraction of total landscape area. Water (Pct11) naturally occurs primarily as rivers and streams, but Huntingdon County hosts a large impoundment called Raystown Lake. All indicator data start as area percentages for designated vicinities as per the ensuing description, with percentages being given a negative sign for contrary indicators.

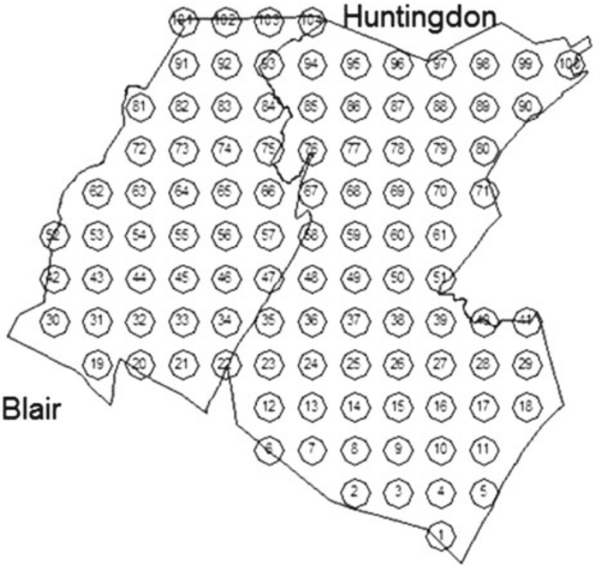


Fig. 16.1 Octagonal vicinities (OCTIVs) with centers having 6-km spacing showing identifying numbers in Blair and Huntingdon Counties of south-central Pennsylvania, USA

The contextual goal is to find areas having relatively natural (naturalistic) character, and five types of cover are taken as indicators of naturalistic character or human influence (humanistic character). Water (Pct11) and forest (Pct40) are taken as indicative of naturalistic character. Excavations (Pct32), development (Pct20), and crops (Pct82) are taken as indicative of intensive human influence (humanistic character). To give all indicators the same sense of sign relative to naturalness, the negatives of the human influence indicators are utilized as indicators.

Center points of localities are spaced on a 6-km grid in a geographic information system (GIS), and OCTagonal Integrating Vicinities (OCTIVs) are delineated for integrative computation of percentages for kinds of cover. This is a geostatistical setting of sampling as addressed by Myers and Patil (2013a) with aspects of so-called support (as sensitivity to scope of integration) to be considered. Vicinities would usually be addressed as circular buffers for points in GIS. The octagons have a 2-km circumscribing circular radius and are much more parsimonious for representation in GIS than circular zones. The percentages constitute Integrative Vicinity Indicators (IVIs). The serial numbering of OCTIVs in Fig. 16.1 provides identification numbers for these areas as objects (or instances). This sampling approach offers a regional overview of localities at much more modest computational cost than would a contiguous square lattice, and the diagonal deviance of squares makes for less coherent localities.

16.3 Balance of Normalized Definitive Status

The BONDS approach to higher order indexing by coordinated compound ranking uses domination under product–order relation (Brüggemann and Patil 2011) as a point of departure. Object *A* dominates object *B* in this sense if all indicators of *A* are at least as good (great) as those of *B* and at least one is better. We take the fraction of objects that *A* dominates as a measure (**A**↓) of its "down-set." Likewise, we take the fraction of objects by which *A* is dominated as a measure (**A**↑) of its "up-set." Then we define the BONDS value of the object *A* as follows:

$$BONDS(A) = 100 \times \left(\mathbf{A}\downarrow - \mathbf{A}\uparrow\right) \times \left(\mathbf{A}\downarrow + \mathbf{A}\uparrow\right).$$

The BONDS values are then ranked, whereupon we proceed to use the rank numbers. For a BONDS-based ranking, we first rank the BONDS values in the usual manner, and then sub-rank the objects having a zero value of BONDS according to the magnitude of (**A**↓ + **A**↑). A map of BONDS ranks for the study setting is given in Fig. 16.2.

We begin with **R** software having a data frame of indicators that we call Apicking and the BONDSranks in a vector that we call PikBONDS. A boxplot of the indicators is shown in Fig. 16.3.

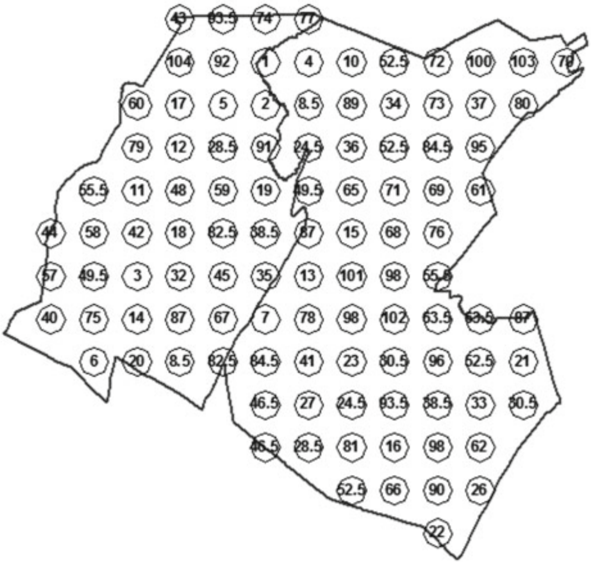


Fig. 16.2 Map of BONDSrank values for OCTIVs in Blair and Huntingdon Counties of south-central Pennsylvania, USA as shown with identification numbers in Fig. 16.1

16.4 Rank-Related Hierarchical Clustering

We proceed to augment the data frame of indicators with the BONDSranks as an additional column, with the first couple lines of the augmented data frame being shown as follows:

```
> Apickings <- cbind(Apicking,PikBONDS)
> head(Apickings)
  IDs  Pct11    Pct20     Pct32    Pct40    Pct82   PikBONDS
1  1 0.071644 -0.660722  0.000000 69.59083 -3.319535 22.0
2  2 0.111500 -0.055750  0.000000 69.69576 -5.853775 52.5
```

We use hierarchical clustering with Euclidean distance as a metric and Ward linkage. Scaling of the indicators is a consideration with this method of clustering, whereby the major concern is to have the BONDSranks influence but not dominate the clustering. If one of the actual indicators has a range similar to the BONDSranks, then this concern is nullified. Here the Pct40 (forest) indicator has a range similar to the ranks, so the concern is not present. If all actual indicators had ranges smaller (or larger) than the ranks, then a range rescaling multiplier would be used to equilibrate the indicator having largest range with the range of ranks. This range rescaling factor would be applied to all actual indicators, but only to the point where objects have been assigned cluster numbers. Thereafter, actual values of indicators would be used. A constant multiplicative rescaling does not alter groupings for

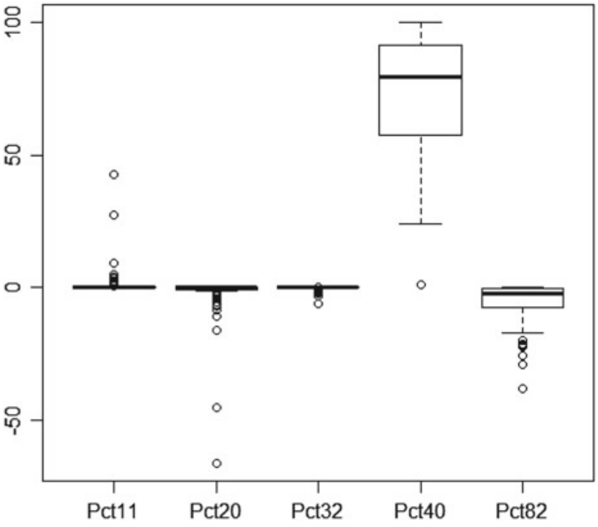


Fig. 16.3 Boxplot of five land-cover composition indicators; *Y*-scale is based on percent area, with percent being positive for propitious indicators and negative for contrary indicators

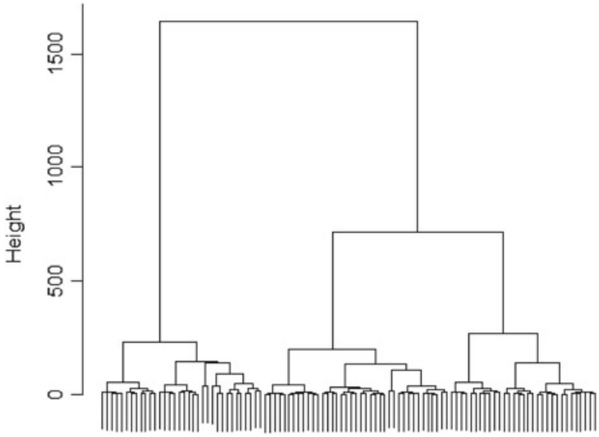


Fig. 16.4 Dendrogram for hierarchical clustering of augmented indicator data

agglomerative hierarchical clustering based on Euclidean distance. Since this analysis is of an exploratory nature, questions regarding sensitivity to scaling should be resolved by comparing groups from different scalings.

The clustering is accomplished as follows, with dendrogram shown in Fig. 16.4:

```
> DistMat <- dist(Apickings[,-1],method="euclidean")
> ApicksClus <- hclust(DistMat,method="ward")
> plot(ApicksClus,labels=F)
```

Having the dendrogram as a roadmap of mergers from individuals to a group of the whole, we can investigate where to truncate the process of progressively merging subgroups. This is best considered as top-down work in the dendrogram, which is the opposite of the actual merger process. Object assignments to a specific number of clusters are obtained with the **R** cutree command, which can be viewed as a horizontal cutting line across the dendrogram that transects the given number of verticals. With our present approach, the appropriateness of a partition can be addressed by plotting rank against cluster number for each object. We begin by extracting six clusters and making the plot shown in Fig. 16.5:

```
> ApicksClus6 <- cutree(ApicksClus,k=6)
> plot(ApicksClus6,Apickings[,7],ylab="BONDSrank")
```

The efficacy of partition can be viewed in terms of vertical proximity of points in each cluster and presence or absence of overlap for the vertical spans in the clusters. In Fig. 16.5 there is evident need for further partition due to extensive overlap of clusters 4 and 6 with the first cluster also contained entirely within the range of cluster 6. Accordingly, Fig. 16.6 shows the result of progressively increasing the number of clusters to 10 which will be considered a stopping point for present purposes.

```
> ApicksClus10 <- cutree(ApicksClus,k=10)
> plot(ApicksClus10,Apickings[,7],ylab="BONDSrank")
> identify(ApicksClus10,Apickings[,7])
```

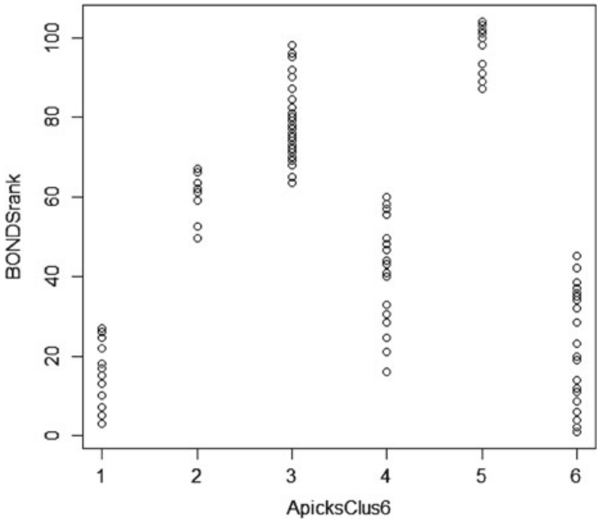


Fig. 16.5 Plot of BONDSrank against cluster number for six clusters

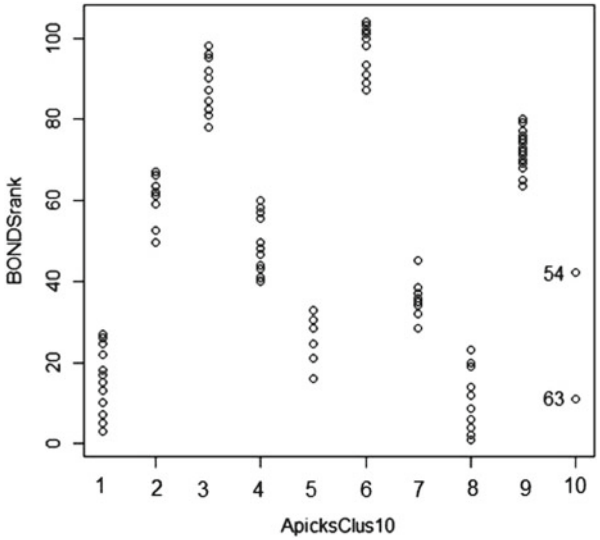


Fig. 16.6 Plot of BONDSrank against cluster number for ten clusters with labeling for members of cluster 10

Nine of the ten clusters in Fig. 16.6 show substantial vertical coherence, with cluster 10 being an exception consisting of only two instances. The two instances in cluster 10 are extracted and listed as follows:

```
> Apickings[c(54,63),]
   IDs    Pct11    Pct20  Pct32   Pct40    Pct82    PikBONDS
54  54 0.549538 -45.38866    0 37.02612 -0.302644       42
63  63 0.000000 -66.06389    0 23.89867 -0.756791       11
```

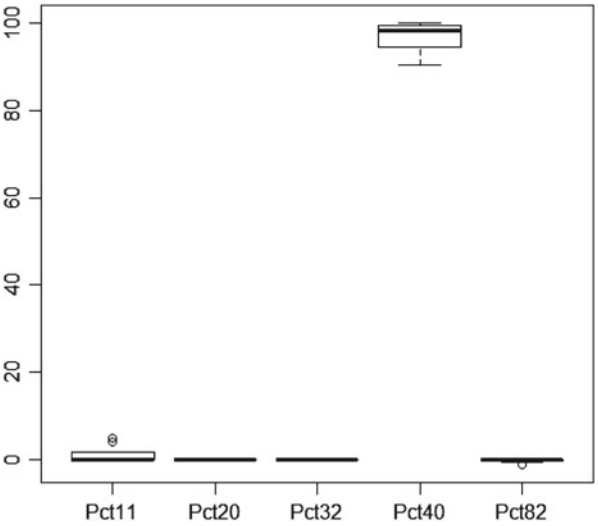


Fig. 16.7 Boxplots for members of cluster 6 out of 10 clusters, with *Y*-axis being BONDSrank

This anomalous binary cluster is seen to contain instances consisting of opposing kinds of land cover as development in forests (e.g., housing in areas with tree cover), with more development than forest. The balance for OCTIV (instance) 63, is considerably more toward development with less forest. It is well to have these anomalous instances of the same general sort segregated in a separate cluster.

With respect to group-wise aggregate ordering, there remains extensive vertical overlap among several subsets of clusters. This is not pathology in the clustering, but shows instead that a range of BONDSranks can arise by several combinatorial interactions of indicators. Clusters could be ordered according to median BONDSrank, but the graphic depiction of Fig. 16.6 is more informative. Clusters 3, 6, and 9 obviously contain premier instances arising by different combinatorial conditions. The combinatorial conditions can be revealed by making separate boxplots for each cluster, starting with cluster 6 as shown in Fig. 16.7. It is evident from Fig. 16.7 that the predominant feature of most instances in cluster 6 is a high component of forest cover with little else.

Cluster 3 of 10 is next in order of interest with regard to BONDSrank, and its boxplots are shown in Fig. 16.8. Forests again predominate, but with a somewhat lesser percentage of cover as seen from the vertical axis.

Cluster 9 is the third of the three upper clusters. Its boxplots are omitted here for brevity but also reveal the primary character as high forest cover. A map highlighting cluster 6 locations in shown in Fig. 16.9, and a map highlighting the combined locations of clusters 3, 6, and 9 is shown in Fig. 16.10. Figure 16.10

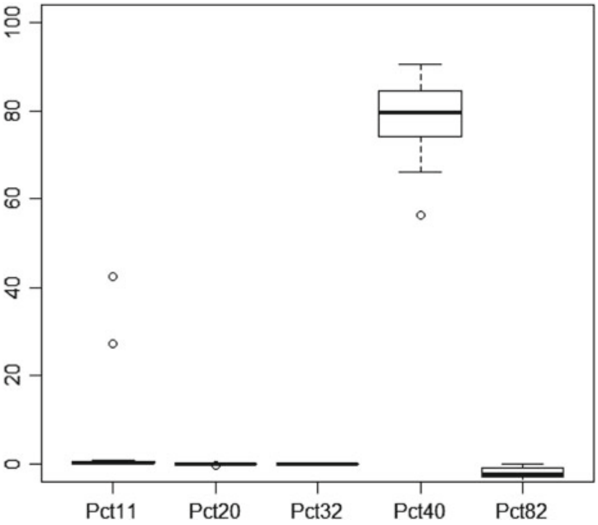


Fig. 16.8 Boxplots for members of cluster 3 out of 10 clusters

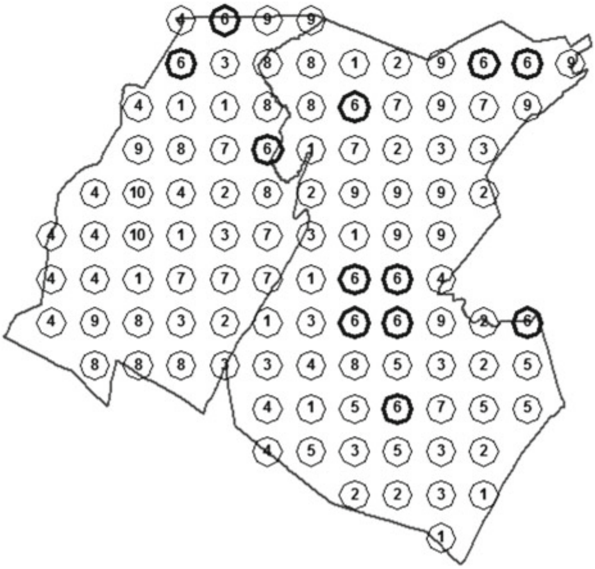


Fig. 16.9 Map of ten clusters with locations of cluster 6 highlighted

reflects the more heavily forested character of Huntingdon County including state forest and experimental forest of Pennsylvania State University. Careful reference back to Fig. 16.1 shows that the two anomalous localities of cluster 10 have north–south adjacency in west-central Blair County.

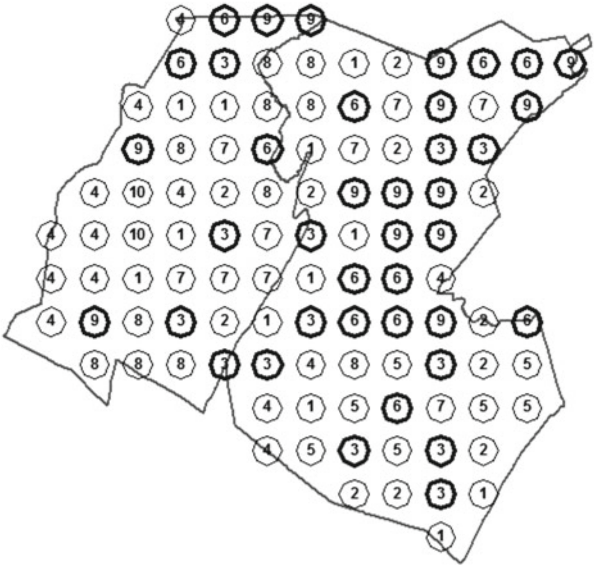


Fig. 16.10 Map of ten clusters with locations of clusters 3, 6, and 9 highlighted

16.5 Summary

We have developed and demonstrated a rank-related clustering approach for multi-indicator systems that enhances the informational value of provisional rankings for the objects (instances) in question. The approach helps to elucidate collectivity of interaction among the indicators in relation to the ranking.

Hierarchical clustering is conducted on an augmented matrix of indicators with the ranks included in the manner of an additional indicator. Multiplicative rescaling of indicators may be appropriate to obtain rank influence on grouping without having rank domination of grouping (see text). Need for such rescaling can be determined by comparing the largest range among the indicators to the range of ranks. There are two kinds of diagnostic displays in the approach. The display that is innovative plots rank values against cluster number to show cluster coherence and allow aggregate ordering of the clusters. The time-tested display shows cluster-specific parallel boxplots.

The clustering approach is applied here with ranks based on “Balance Of Normalized Definitive Status (BONDS)” that coordinate three posetic aspects from partial order theory as principal down-set, up-set, and comparability. The context of the multi-indicator system has different kinds of land covers as landscape ecological indicators in a temperate hardwood moderately mountainous biogeographic setting of central Pennsylvania, USA.

References

Allerhand M (2011) A tiny handbook of **R**. Springer, New York, NY

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17(4):383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, New York, NY

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Myers W, Patil GP (2013a) Statistical geoinformatics for human environment interface. Chapman & Hall/CRC, Boca Raton, FL

Myers W, Patil GP (2013b) Coordination of contrariety and ambiguity in comparative compositional contexts: balance of normalized definitive status in multi-indicator systems. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 8. Springer, New York, NY

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multicriterion ranking and prioritization. Environ Ecol Stat 11:199–228

R Development Core Team (2008) **R**: A language and environment for statistical computing. **R** Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R--project.org/

Short T (2009) **R** reference guide. Revolution Computing, New Haven, CT

Venables WN, Smith DM, the R Development Core Team (2005) An introduction to **R**. Network Theory LTD, Bristol

Chapter 17
PyHasse Software Features Applied on the Evaluation of Chemicals in Human Breast Milk Samples in Turkey

Kristina Voigt, Rainer Brüggemann, Hagen Scherb, Ismet Cok, Birgül Mazmanci, M. Ali Mazmanci, Cafer Turgut, and Karl-Werner Schramm

Abstract In this chapter we evaluate the data of 18 Organochlorine pesticides (OCPs) found in breast milk samples from 44 mothers in the Taurus Mountains in Turkey. In this approach the association of concentration levels in breast milk samples with the two confounding factors: smoking habit and habit of taking medication is the goal. For all data evaluation approaches, we applied the Hasse diagram technique and its software package, namely the PyHasse software. Special emphasis was laid on the software features "similarity" and "Local Partial Order Model" to

K. Voigt (✉) • H. Scherb
Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Institute of Computational Biology, Ingolstaedter Landstr.1, 85764 Neuherberg, Germany
e-mail: kvoigt@helmholtz-muenchen.de

R. Brüggemann
Department of Ecohydrology, Leibniz-Institute of Fresh Water Ecology and Inland Fisheries, Mueggelseedamm 310, Berlin, Germany
e-mail: brg_home@web.de

I. Cok
Department of Toxicology, Faculty of Pharmacy, Gazi University, 06330 Ankara, Turkey

B. Mazmanci
Department of Biology, Faculty of Sciences and Letters, University of Mersin, Mersin 33363, Turkey

M.A. Mazmanci
Department of Environmental Engineering, Faculty of Engineering, University of Mersin, Mersin, 33363, Turkey

C. Turgut
Faculty of Agriculture, Adnan Menderes University, 09100 Aydin, Turkey

K.-W. Schramm
Helmholtz Zentrum München, German Research Center for Environmental Health, Molecular EXposomics (MEX), Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany

Department fuer Biowissenschaften, Wissenschaftszentrum Weihenstephan fuer Ernaehrung und Landnutzung, TUM, Weihenstephaner Steig 23, 85350 Freising, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_17, © Springer Science+Business Media New York 2014

draw further conclusions out of the data. The data analyses resulted in differences between the smoking women and those who did not smoke as well as between the medication and non-medication breast milk samples. Little differences were found comparing hormone taking mothers and mothers taking other medication.

17.1 Introduction

Environmental pollutants such as Organochlorine Pesticides (OCPs) have been reported in human breast milk samples for several decades. These OCPs are widespread used chemicals in agriculture and industry for different purposes all over the world. Epidemiological evidence and theoretical considerations imply that these compounds are potentially hazardous to human and wildlife reproductive health.

In a recent approach 18 OCPs in 44 breast milk samples, measured in 5 different regions in the Taurus Mountains in Turkey, were evaluated (Voigt et al. 2013). Furthermore, we analyzed the influence of the fish eating habits on the ordering of these 18 chemicals. In this chapter now we will take a closer look at the impact of other co-variables, like, e.g., the smoking habit and the use of pharmaceuticals. Concerning the smoking habit we can only distinguish between nonsmokers and smokers (Voigt et al. 2012). With respect to the co-variable using medicine, we can distinguish among no medication, hormone treatment, and other pharmaceuticals.

For all data evaluation approaches, we applied the Hasse diagram technique (HDT) and its software package, namely the PyHasse software. This software is written in the free available interpreter language Python by the second author, and it is under constant development. PyHasse can be obtained from the second author on request. It comprises many modules, some of which are of great support also in the data evaluation of environmental health data. In this presentation we will apply the main Hasse Diagram Technique Module (mainHD20), the Similarity Analysis (similarity9) for the comparison of two data matrices and the Local Partial Order Model (LPOM).

17.2 PyHasse Software and Its Features

17.2.1 PyHasse Software

The data analysis is performed by the free available software package PyHasse, written by the second author (Brüggemann and Patil 2010, 2011). Apart from the calculation of Hasse diagrams and some basic characteristics (mainHD20) many other features are provided, such as for example the similarity analysis (similarity9) and the Local Partial order Model (LPOMext2) applied in this chapter.

PyHasse is written in the modern interpreter language Python (vs 2.7). Python, used as ‘rapid prototyping’ programming language, is considered as a ‘Very High Level Language’ (VHLL) and allows a high level of abstraction (Muller and Schwarzer 2007). For example a combinatorial analysis such as rendering all combinations with repetitions needs only one line in Python programming.

Python can freely be downloaded from http://www.python.org/download, is platform independent and it runs on many operating systems. It also has access to many packages either in algebra, in statistics, or in graphics or even in preparing sound effects, game developments, and image processing, allows object-oriented programming, and is itself modularized to a very high degree (http://www.pythonology.com/home). In contrast to the older but professionally programmed WHASSE, PyHasse is considered as “experimental” software, which bridges the gap between professional software and software, exclusively developed in laboratories and in general only applicable for the programming scientist. To obtain good graphics is not in the focus of PyHasse, as professional software can be used for that. Recently, a new module of PyHasse became available which has an interface to the professionally written graph drawing program graphviz (Gansner et al. 1993). In graphviz the drawing of digraphs is optimized in order to get clear visualizations by minimizing the number of line crossings.

17.2.2 Similarity Analysis

In complex datasets it is often necessary to compare different sets of criteria which are quantified by the attributes or (synonym) the indicators. In the similarity analysis we intend to calculate the proximity of different posets (partially ordered sets, denoted generally as $(G, \leq)$, with G the ground set (of objects)) based on the same ground set. This similarity analysis is an important feature of the newly developed software package called PyHasse and is briefly explained as follows:

The outcome of a partial order for two objects a, b may be $a < b$, $a > b$, $a \parallel b$, $a \cong b$. When two partial orders $(G, \leq_1)$ and $(G, \leq_2)$ are to be compared, then the combinations are counted such as: $a <_1 b$, $a <_2 b$, or $a <_1 b$, $a >_2 b$, or $a \parallel_1 b$, $a <_2 b$ for which we use the shorthand notation >>, >< , ||, etc.

Most important are the entries like >> or <<, which are counting the “isotone” character of both partial orders (ISO) and the entries like ><, <> which contribute to the “antitone” character, i.e., to the conflicts between the two partial orders (ANTI). There are still more combinations to look upon: <||, >||, ||>,||< , =||, ||=, or || || are considered as indifferent (IND); combinations like > =, < =, = <, = > are called weak isotone (WISO). Finally, the entry of type = = contributes to equivalence relations (IDE). A detailed description can be found in Brüggemann and Patil (2011).

Recent examples of the application of the similarity on an environmental health data set are given by Voigt et al. (2010, 2011, 2013).

17.2.3 Local Partial Order Model

The basic idea is to construct from partial order a linear order. This task can approximately be performed using the Local Partial Order Module of the PyHasse software package. Some recent publications by Carlsen and Brüggemann (2008) and Brüggemann and Carlsen (2011) give some further insight into this feature.

Only some basic ideas in order to understand the example in this chapter are presented now.

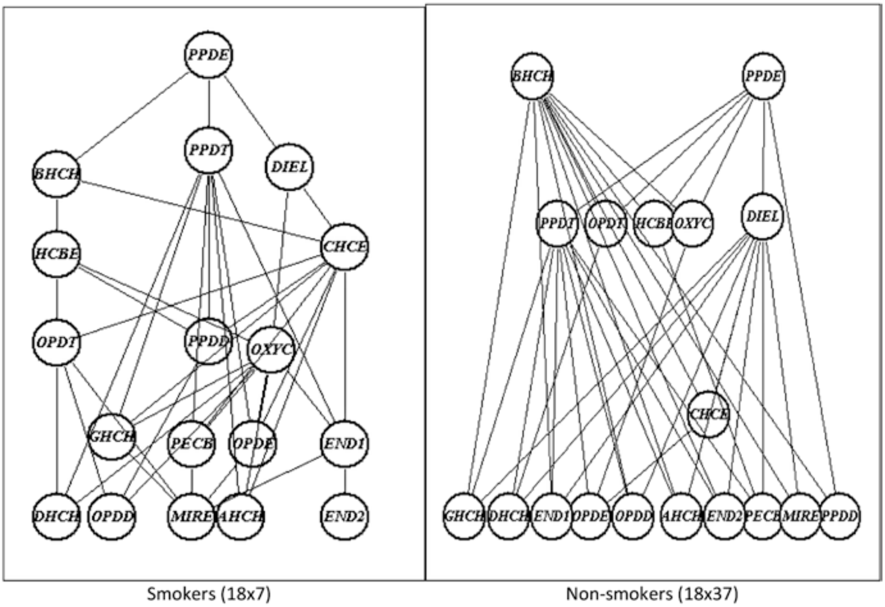
The number of incomparable objects in the partial ordering constitutes a limitation in the attempt to rank, e.g., a series of chemical substances based on their potential environmental or human health hazard. To a certain extent this problem can be remedied through the application of the so-called linear extensions of the partial order ranking (Fishburn 1974; Graham 1982). A linear extension is a total order, where all comparabilities of the partial order are reproduced (Brüggemann et al. 2001). Due to the incomparabilities in the partial order ranking, a number of possible linear extensions correspond to one partial order. If all possible linear extensions are found, a ranking probability can be calculated. Hence, based on the linear extensions, the probability that a certain object, *Q*, has a certain rank can be derived. Further, it is possible to calculate the average ranks of the single objects in a partially ordered set (Winkler 1982, 1983). Whereas theoretically the calculation of average ranks is straightforward, the computational realization is very difficult, hence approximations are needed.

The generation of the average rank of the single object in the Hasse diagram can be obtained through deriving a large number of randomly generated linear extensions (Bubley and Dyer 1998; Sørensen et al. 2001; Lerche et al. 2002, 2003). Recently, an improved version of the random linear extension approach has been suggested by Sørensen et al. (2007), taking into account that not all descriptors may be evenly important. Alternatively, an approximate generation of the average rank of the single objects in the Hasse diagram is obtained applying the simple relation recently reported by Brüggemann et al. (2004). Within the simplest version of the Local Partial Order Model, the average rank (seen as average height) of a specific object, *Q*, can be obtained by

$$Rkav = \frac{(Suc + 1) x (N + 1)}{(N + 1 - U)},$$

where *N* is the number of objects in the diagram, Suc the number of successors, i.e., comparable object located below *Q* and *U*, the number of objects being incomparable to *Q* (Brüggemann et al. 2004). For further developments, especially the more sophisticated Local Partial Order Model (LPOMext) (see Brüggemann et al. 2005; Brüggemann and Carlsen 2011). Its basic equation is:

$$Rkav(x) \approx | O(x) | + \sum_{y \in U(x)} \frac{p_y^<}{p_y^< + p_y^>}.$$



Smokers (18x7)

Non-smokers (18x37)

Fig. 17.1 Hasse diagrams of smokers versus non-smokers

Wherein

$|O(x)|$ is the number of elements in the down set of *x*. For example *O* (OPDT) = {OPDT, DHCH, OPDD, MIRE}, see Fig. 17.1, lhs. Hence $|O(\text{OPDT})| = 4$.

U(*x*) is the set of elements incomparable to *x*, for example, *U* (OPDT) = {PPDT, DIEL, PPDD, OXYC, GHCH, PECB, OPDE, END1, END2, AHCH}, see Fig. 17.1 lhs.

F(*x*) is the up set of *x* i.e., all elements comparable with *x* upwards.

$$p_y^{<} := | U(y) \cap O(x) |$$

$$p_y^{>} := | U(y) \cap F(x) |$$

17.3 Chemicals in Human Breast Milk Samples

The occurrence of POPs in the Taurus Mountains in Turkey was studied in 2010 (Turgut et al. 2011). Taurus Mountains were suggested for this study because of their potential to act as a sink for organic pollutants by cold condensation and can reflect the atmospheric pollution in Turkey as well as neighboring countries e.g. Arabia, Africa, and Russia. Referring to the role of mountains as sinks in mathematical modeling, the long-range transport of chemicals we recommend the publications of Scheringer et al. (2000, 2001).

Table 17.1 List of 18 POPs detected in breast milk samples in Turkey

Nr.	Acronym	Name	CAS-Number
01	AHCH	alpha-Hexachlorcyclohexane	319-84-6
02	BHCH	beta-Hexachlorcyclohexane	319-85-7
03	GHCH	gamma-Hexachlorcyclohexane	58-89-9
04	PECB	Pentachlorobenzene	608-93-5
05	HCBE	Hexachlorobenzene	118-74-1
06	PPDT	p, p′-Dichlordiphenyltrichlorethane	50-29-3
07	OPDT	o, p′-Dichlordiphenyltrichlorethane	789-02-6
08	PPDD	p, p′-Dichlordiphenyldichlorethane	72-54-8
09	OPDD	o, p’-Dichlordiphenyldichlorethane	53-19-0
10	PPDE	p, p′-Dichlordiphenyldichlorethene	72-55-9
11	OPDE	o, p′-Dichlordiphenyldichlorethene	3424-82-6
12	OXYC	Oxychlordane	27304-13-8
13	CHCE	cis-Heptachloroepoxide	1024-57-3
14	DIEL	Dieldrin	60-57-1
15	END1	Endosulfan-1	959-98-8
16	END2	Endosulfan-2	33213-65-9
17	MECH	Methoxychlor	72-43-5
18	MIRE	Mirex	2385-85-5

The Taurus Mountains are a complex in southern Turkey, from which Euphrates and Tigris descend into Iraq. It divides the Mediterranean coastal region of southern Turkey from the Central Anatolian Plateaus. Five locations were selected in Taurus Mountains. The monitoring of pesticides in Turkey is widely described in a recent paper (Cok et al. 2012). In a newly performed Turkish-Germany collaboration study in the Taurus mountain area in Turkey, breast milk samples were analyzed for POPs (Turgut et al. 2011).

Now the occurrence of POPs in breast milk samples and the analysis of the role of co-variables is the aim of the current study.

In this chapter we will evaluate the data of 18 OCPs (Table 17.1) found in breast milk samples from 44 mothers. The study did not aim at finding the association between the concentration of POPs and the occurrence of diseases and/or malfunctions. In this approach the association of concentration levels in breast milk samples with the two confounding factors, smoking habit and habit of taking medication, is the goal.

17.3.1 Smoking Habits

Several studies have been conducted to examine the influence of smoking habits of the pregnant mothers on the newborn children.

Table 17.2 Evaluation results of Hasse diagrams of different co-variables

	Smoking (SM)	No smoking (NSM)	Medication (ME)	No medication (NME)	Hormone medication (HME)	Other medication (OME)
Data set	18×7	18×37	18×11	18×33	18×6	18×5
Maximal objects	1: PPDE	2: BHCH, PPDE	1: PPDE	2: BHCH, PPDE	1: PPDE	1: PPDE
Minimal objects	5: AHCH, DHCH, END2, MIRE, OPDD	10: AHCH, GHCH, DHCH, PECB, PPDD, OPDD, OPDE, END1, END2, MIRE	5: AHCH, DHCH, OPDD, END2, MIRE	10: AHCH, GHCH, DHCH, PECB, PPDD, OPDD, OPDE, END1, END2, MIRE	4: DHCH, OPDD, END2, MIRE	5: AHCH, DHCH, OPDD, END2, MIRE
Levels	6	4	6	4	7	6
Comparabilities	95	52	90	52	112	100
Incomparabilities	58	101	63	101	41	53

It is known that nicotine accumulates in human breast milk. Maternal smoking during the period of lactation will expose the infant to nicotine through the breast milk (Dahlstrom et al. 2004). The documented knowledge about the effects of smoking and nicotine exposure on the fetus during pregnancy is substantial (Nieburg et al. 1985; Law et al. 2003; Rogers and Abbott 2003).

There exist further evidence on the occurrence of increased blood pressure in neonates and infants whose mothers smoke during pregnancy (Beratis et al. 1996).

Only few studies have been considering the effect of smoking on OCP levels in breast milk and results of these studies were inconsistent (Harris et al. 2001). In a study performed more than 20 years ago, Skaare and Polder (1990) found higher levels of p,p′-DDE and HCB in smoking Norwegian women. In a recently published study Polder et al. (2009) found out that smoking was associated with higher levels of PCBs, p,p′-DDE, and beta-HCH. The levels of Hexachlorobenzene in breast milk in relation to birth weight in a Norwegian cohort were also correlated with the smoking habit of the mothers (Eggesbo et al. 2009).

In our Turkish breast milk study out of 44 women 37 did not smoke before, during, and after their pregnancy and only 7 admitted to be smokers. So we encounter two data sets: 18×7 smokers and 18×37 nonsmokers and constructed the corresponding two Hasse diagrams (see Fig. 17.1).

Differences can be found in the maximal and minimal objects as well as in the number of levels. More information is found in Table 17.2. Furthermore, we can see that, e.g., a sequence can be found such as DHCH < OPDT < HCBE < BHCH < PP DE for smokers, which is reduced to DHCH < PPDT < PPDE for the nonsmokers, described by a data matrix with around five times more columns.

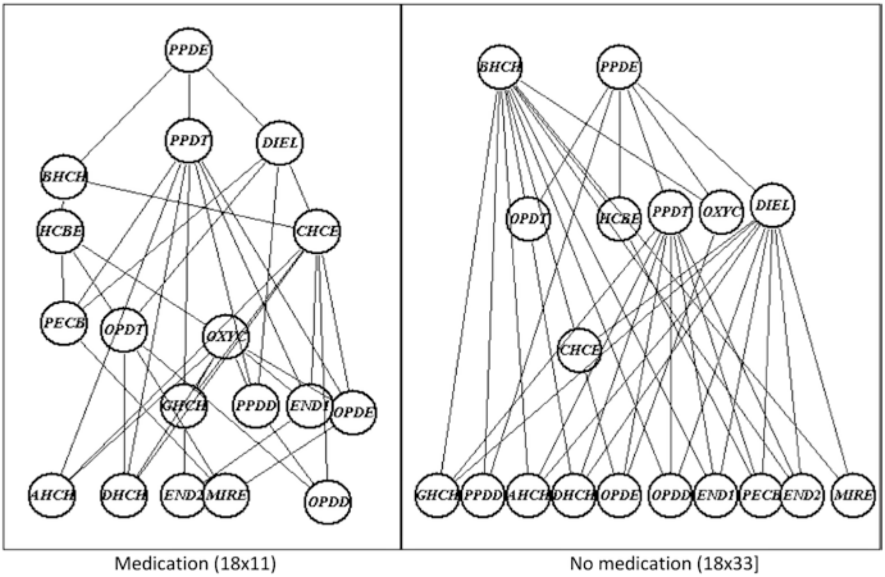


Fig. 17.2 Hasse diagrams of mothers taking medication versus no medication

17.3.2 Medication Habits

The Committee on Drugs published lists of drugs with their reported sign or symptoms in infants or effect on lactation (Committee on Drugs 2001). Also the food and environmental agents with their reported effects on breast feeding are listed.

First we start with the evaluation of those breast milk studies where mothers did not take any medication versus the mothers who took different kinds of medication.

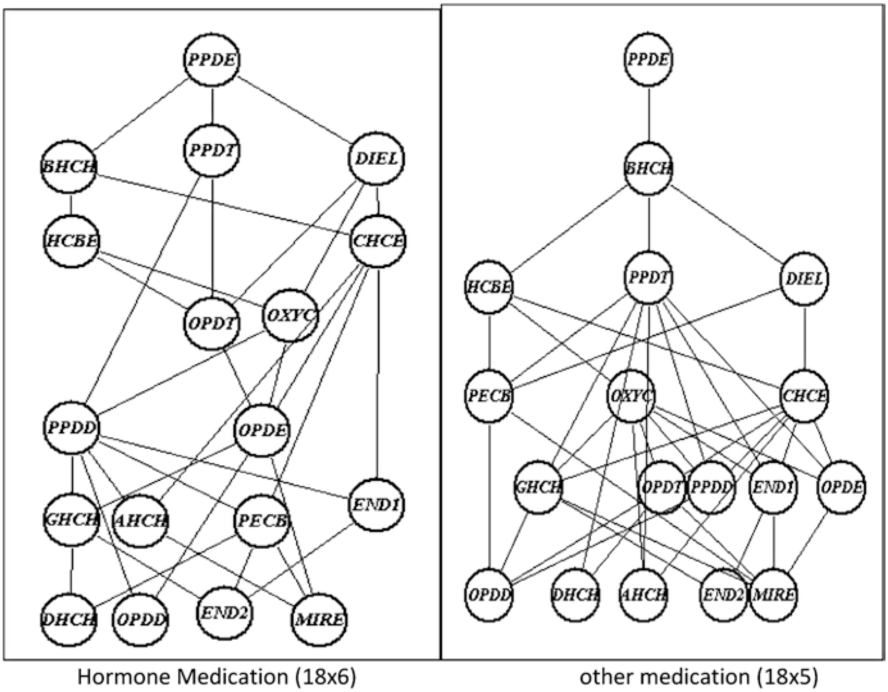
The Hasse diagrams are displayed in Fig. 17.2.

18 × 11 medication comprising six hormones and five other pharmaceuticals versus 18 × 33 no medication.

The differences are visible in the maximal/minimal positions of objects as well as in the number of levels (see also Table 17.2).

The medication can be divided into hormones (18 × 6) and other medication (18 × 5). The Hasse diagrams are displayed in Fig. 17.3.

The visibility of the differences of the two diagrams is much less than in Figs. 17.1 and 17.2. It is clear that the partitioning of medication cases into hormone medication and other medications must preserve the order relations of medication. Therefore, it cannot be expected that much changed Hasse diagrams can be obtained (see also Table 17.2).



Hormone Medication (18x6)

other medication (18x5)

Fig. 17.3 Hasse diagrams of mothers taking hormone medication versus other medication

17.4 Similarity Analysis

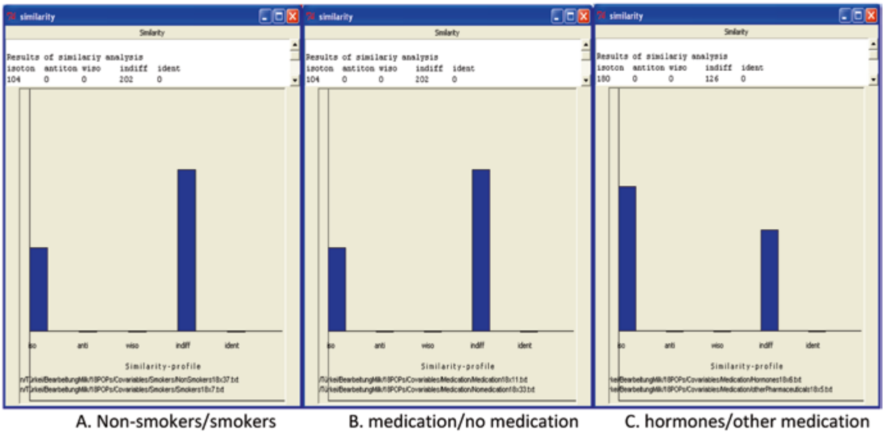
The similarity analysis should quantify the relations between smoker versus nonsmoker on the one side and medication versus non-medication on the other side, as well as hormone versus other medication.

The data sets to be examined by the similarity analysis tool are the following:

A	18×7	Smokers	18×37	Nonsmokers	iso: 104, ind: 202
B	18×11	Medication	18×33	No medication	iso: 104, ind: 202
C	18×6	Hormone medication	18×5	Other medication	iso: 180, ind: 126

In Fig. 17.4 the three similarity graphs are displayed.

The analysis reveals only isotone and indifferent relations. Whereas isotone relations demonstrate a high degree of similarities, indifferent relations reveal all relations with incomparabilities. In the similarity approaches concerning the nonsmokers versus smokers (Fig. 17.4 lhs) and medication versus no medication (Fig. 17.4 middle), the relation isotone/indifferent is approximately 2:1. This means that the two data sets in question as well as the two Hasse diagrams are very different.



```
Results of similariy analysis
isoton  antiton viso    indiff  ident
104     0       0       202     0
```

n/TürkeiBearbeitungMilk/18POPs/Covariables/Smokers/NonSmokers18x37.txt
n/TürkeiBearbeitungMilk/18POPs/Covariables/Smokers/Smokers18x7.txt

A. Non-smokers/smokers

```
Results of similariy analysis
isoton  antiton viso    indiff  ident
104     0       0       202     0
```

/TürkeiBearbeitungMilk/18POPs/Covariables/Medication/Medication18x11.txt
/TürkeiBearbeitungMilk/18POPs/Covariables/Medication/Nomedication18x33.txt

B. medication/no medication

```
Results of similariy analysis
isoton  antiton viso    indiff  ident
180     0       0       126     0
```

keiBearbeitungMilk/18POPs/Covariables/Medication/Hormones18x6.txt
keiBearbeitungMilk/18POPs/Covariables/Medication/otherPharmaceuticals18x5.txt

C. hormones/other medication

Fig. 17.4 (A) Nonsmokers/smokers, (B) medication/no medication, and (C) hormones/other medication

The last analysis the one of the hormones versus other medication shows the opposite. In this calculation more isotone than indifferent relations are calculated. This means that the two data matrices are very similar.

We can conclude from this that a separation into hormones and other pharmaceuticals is not necessary. The smoking habits as well as the medication taking habits seem to have an impact on the concentrations of OCPs in human breast milk. A decisive answer cannot yet be given, because just the effect of different numbers of samples (attributes) changes the number of incomparabilities. Hence, a comparison of a partial order of n objects and m_1 attributes with another one having the same number of objects but m_2 ($>m_1$) attributes is hampered just by the different number of attributes, which not only bear contextual information but which also induce more noise.

In a recently performed study we took a closer look at the fish eating habits of the mothers. The similarity analysis of the two data sets of fish eating habit versus non-fish eating habit indicated a strong influence of fish on the concentration of OCPs in human breast milk (Voigt et al. 2012).

17.5 Application of the Local Partial Order Model

As mentioned above just the change from m1 to m2 attributes will in general cause changes of the partial order. Therefore, we eliminate this effect by discussing the average rank. We apply, for example, the extended local partial order model (LPOMext) to estimate a weak or linear order based on average ranks of the studied objects (chemicals) originally being partially ordered. In the concept of weak order tied ranks are not excluded.

Table 17.3 Average rank values of all data sets

Chemical	RkavNSM	RkavSM	RkavNME	RkavME	RkavHME	RkavOME
AHCH:	6,167	5,583	6,167	5,283	6,367	5,583
BHCH:	17,208	16,5	17,208	16,583	16,467	17
GHCH:	6,167	6,917	6,167	7,067	5,65	8,667
DHCH:	4,433	4,25	4,433	4,117	2,417	4,2
PECB:	4,433	6,917	4,433	8	7,5	8,667
HCBE:	10,667	14,6	10,667	14,933	14,633	15,467
PPDT:	15,189	14,556	15,189	14,806	14,917	14,222
OPDT:	10,667	9,85	10,667	10,317	10,764	9,083
PPDD:	8,167	7,917	8,167	7,617	10,667	7,083
OPDD:	6,167	3,167	6,167	3,083	3,75	2,233
PPDE:	17,75	18	17,75	18	18	18
OPDE:	3,45	6,917	3,45	6,95	8,279	6,917
OXYC:	10,5	12,308	10,5	12,775	12,533	13,233
CHCE:	9,417	13,483	9,417	12,552	12,55	12,633
DIEL:	15,727	15,967	15,727	15,717	15,917	15
END1:	6,167	8,25	6,167	7,067	6,2	8,067
END2:	6,167	3,95	6,167	2,76	1,926	3,2
MIRE:	7,083	1,768	7,083	2,467	2,333	1,902

17.5.1 LPOM Analysis on Smoking Habits and Medication

The procedure of the calculation of the extended local partial order model (LPOMext) will be exemplified at the data sets of the nonsmoking habits (18 × 37) versus smoking, medication versus non-medication, medication hormones versus other medications.

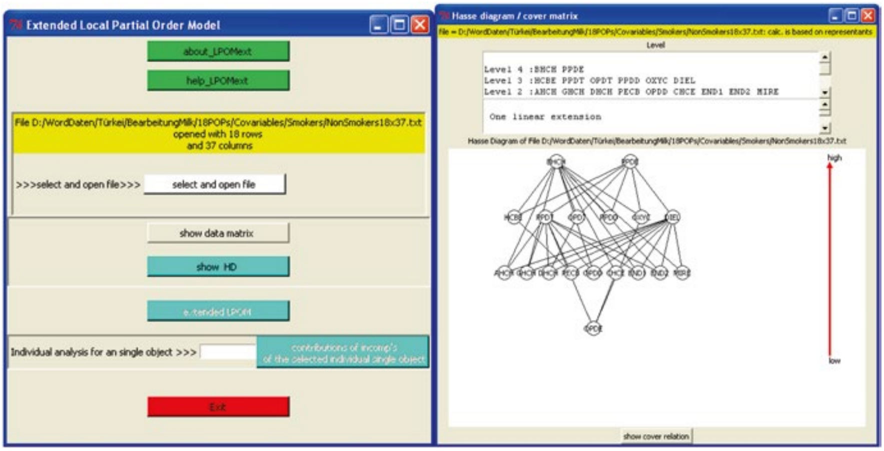
This procedure will be performed for all six data sets:

1	NSM	Non-smoking	18 × 37
2	SM	Smoking	18 × 7
3	NME	No medication	18 × 33
4	ME	Medication	18 × 11
5	HME	Hormone medication	18 × 6
6	OME	Other medication	18 × 5

The abbreviations are used in Table 17.3.

In Fig. 17.5 the typical graphical user interface of LPOMext module and a resulting Hasse diagram are shown.

The results of the calculation of the extended partial order model, namely the Rkav values (average ranks) are given in Table 17.3.


Extended Local Partial Order Model
about_LPOMext
help_LPOMext
>>>select and open file>>>
select and open file
show data matrix
show HD
extended LPOM
Individual analysis for an single object >>>
contributions of incomp's
of the selected individual single object
Exit
Hasse diagram / cover matrix
Level
BHCH
PPDE
HCBE
PPDT
OPDT
PPDD
OXYC
DIEL
AHCH
GHCH
DHCH
PECB
OPDD
CHCE
END1
END2
MIRE
OPDE
high
low
show cover relation

File D:/WordDaten/Türkei/BearbeitungMilk/18POPs/Covariables/Smokers/NonSmokers18x37.txt
opened with 18 rows
and 37 columns

file = D:/WordDaten/Türkei/BearbeitungMilk/18POPs/Covariables/Smokers/NonSmokers18x37.txt: calc. is based on representants

```
Level 4 :BHCH PPDE
Level 3 :HCBE PPDT OPDT PPDD OXYC DIEL
Level 2 :AHCH GHCH DHCH PECB OPDD CHCE END1 END2 MIRE
```

```
One linear extension
```

Hasse Diagram of File D:/WordDaten/Türkei/BearbeitungMilk/18POPs/Covariables/Smokers/NonSmokers18x37.txt

Fig. 17.5 LPOMext nonsmoking habits (18 × 37)

17.5.2 Calculation of the Correlation Coefficients

We calculated the Pearson correlation coefficient for the following three data sets derived from the average ranks determination described above.

Smokers/nonsmokers	0.88
No medication/medication	0.88
Hormones/other medication	0.95

The results show relatively good correlation in all three cases. The highest Pearson correlation coefficient value is obtained for the comparison of the hormones/other medication data sets. In this case merely no difference can be detected by the bivariate statistical approach of correlation.

A somewhat lower correlation coefficient is found for the two other pairs of co-variables indicating that concentration levels may indeed be affected by the cases represented by the co-variables.

In contrast the similarity analysis of two partial ordered sets displays structured differences by the number of indifferent relations. Hence, it provides more precise similarity information.

17.6 Results and Outlook

A clear difference can be detected between the contamination profiles of breast milk samples from mother with smoking habits to those mothers who did not smoke. This can be easily read from the initial Hasse diagrams. The same applies to the co-variable medication habits. To quantify these results the similarity feature of the PyHasse software is applied and reveals that a great similarity is calculated between hormone and other medication. This means that no distinction between hormone medication and other medication is necessary for our data analysis approach.

There is no degree of antitone, as can be expected because the data sets are all derived from one base set: A comparability in a data set with many attributes will remain, even if some attributes are eliminated from the data matrix. This means for the similarity analysis discussed here: when a concentration profile of one chemical is larger than that of another chemical with respect to the one case (one subset of attributes), then there will be no contradiction in the other case (the other subset of attributes), although still a greater relation in one case may meet an incomparability (||) in the other case.

In the current study the similarity analysis does not distinguish between the mere number of attribute effects and the contextual effects. It is, however, of great importance to get an idea to which degree the similarity is merely based on the number of attributes and the noise they are implying on order relations. We are working on the separation of these two effects in the PyHasse software package. Then we will evaluate the three co-variables smoking/no smoking, medication/no medication, and fish/no fish in this respect.

It is evident that there is increasing pressure to intensify the research and to more efficiently evaluate the data on persistent and bioaccumulative chemicals in the environment as well as in human bodies. In this respect mathematical ranking methods and their corresponding software are essential. It would be advisable to professionalize this software in order to find more applications and applicants in environment and health research in the future.

References

Beratis NG, Panagoulias D, Varvarigou A (1996) Increased blood pressure in neonates and infants whose mothers smoked during pregnancy. J Pediatr 128:806–812

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partial orders. Match Commun Math Comput Chem 65:383–414

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems. Springer, Berlin

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg CEW (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, Sørensen PB, 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, Simon U, Mey S (2005) Estimation of averaged ranks by extended local partial order models. Match Commun Math Comput Chem 54:489–518

Bubley R, Dyer ME (1998) Faster random generation of linear extensions. In: Proceedings of the 9th annual ACM-SIAM symposium on discrete algorithm, San Francisco, CA, pp 350–354

Carlsen L, Brüggemann R (2008) Accumulating partial order ranking. Environ Model Softw 23:986–993

Cok I, Mazmanci B, Mazmanci MA, Turgut C, Henkelmann B, Schramm K-W (2012) Analysis of human milk to assess exposure to PAHs, PCBs and organochlorine pesticides in the vicinity Mediterranean city Mersin, Turkey. Environ Int 40:63–69

Committee on Drugs (2001) Transfer of drugs and other chemicals into human milk. Pediatrics 108:776–789

Dahlstrom A, Ebersjo C, Lundell B (2004) Nicotine exposure in breastfed infants. Acta Paediatr 93:810–816

Eggesbo M, Stigum H, Longnecker MP, Polder A, Aldrin M, Basso O, Thomsen C, Skaare JU, Becher G, Magnus P (2009) Levels of hexachlorobenzene (HCB) in breast milk in relation to birth weight in a Norwegian cohort. Environ Res 109:559–566

Fishburn PC (1974) On the family of linear extensions of a partial order. J Comb Theory 17:240–243

Gansner ER, Koutsofios E, North C, Vo K-P (1993) A technique for drawing directed graphs. IEEE Trans Softw Eng 19:214–230

Graham RL (ed) (1982) Linear extensions of a partial order and the FKG inequality. Reidel, Dodrecht

Harris CA, Woolridge MW, Hay AW (2001) Factors affecting the transfer of organochlorine pesticide residues to breastmilk. Chemosphere 43:243–256

Law KL, Stroud LR, LaGasse LL, Niaura R, Liu J, Lester BM (2003) Smoking during pregnancy and newborn neurobehavior. Pediatrics 111:1318–1323

Lerche D, Brüggemann R, Sorensen P, Carlsen L, Nielsen OJ (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, Sorensen PB, Brüggemann R (2003) Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability. J Chem Inf Comput Sci 43:1471–1480

Muller M, Schwarzer S (eds) (2007) Python im deutschsprachigen Raum. Lehmanns Media, Berlin

Nieburg P, Marks JS, McLaren NM, Remington PL (1985) The fetal tobacco syndrome. JAMA 253:2998–2999

Polder A, Skaare JU, Skjerve E, Loken KB, Eggesbo M (2009) Levels of chlorinated pesticides and polychlorinated biphenyls in Norwegian breast milk (2002–2006), and factors that may predict the level of contamination. Sci Total Environ 407:4584–4590

Rogers JM, Abbott BD (2003) Screening for developmental toxicity of tobacco smoke constituents. Toxicol Sci 75:227–228

Scheringer M, Wegmann F, Fenner K, Hungerbuehler K (2000) Investigation of the cold condensation of persistent organic pollutants with a global multimedia fate model. Environ Sci Technol 34(9):1842–1850

Scheringer M, Hungerbuehler K, Matthies M (2001) The spatial scale of organic chemicals in multimedia fate modeling. Recent developments and significance for chemical assessment. Environ Sci Pollut Res Int 8(3):150–155

Skaare JU, Polder A (1990) Polychlorinated biphenyls and organochlorine pesticides in milk of Norwegian women during lactation. Arch Environ Contam Toxicol 19:640–645

Sørensen PB, Lerche D, Carlsen L, Brüggemann R (2001) Statistically approach for estimating the total set of linear orders. In: Brüggemann R, Pudenz S, Luhr H-P (eds) Order theoretical tools in environmental science and decision systems. Berichte des IBB, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, pp 87–97

Sørensen PB, Thomsen M, Fauser P, Muenier B (2007) The usefulness of a stochastic approach for multi-criteria selection. In: Hryniewicz O, Studzinski J, Szediw A (eds) EnviroInfo 2007, environmental informatics and systems research, vol 2. Shaker-Verlag, Aachen, pp 187–194

Turgut C, Atatanir L, Mazmanci B, Mazmanci MA, Henkelmann B, Schramm KW (2011) The occurrence and environmental effect of persistent organic pollutants (POPs) in Taurus Mountains soils. Environ Sci Pollut Res Int 19:325–334

Voigt K, Brüggemann R, Scherb H, Shen HQ, Schramm KW (2010) Evaluating the relationship between chemical exposure and cryptorchidism. Environ Model Softw 25:1801–1812

Voigt K, Brüggemann R, Scherb H, Cok I, Mazmnci B, Mazmanci A, Turgut C, Schramm KW (2011) Similarities of environmental health data of persistent organic pollutants in three countries analyzed by the PyHasse software. In: Pillmann W, Schade S, Smits P (eds) EnviroInfo Ispra 2011, 25th international conference environmental informatics, innovations in sharing environmental observation and information. Shaker-Verlag, Ispra, pp 137–144

Voigt K, Brüggemann R, Scherb H, Cok I, Mazmanci B, Mazmanci A, Turgut C, Schramm K-W (2012) Features of the PyHasse software used for the evaluation of chemicals in human breast milk samples in Turkey. In: Wittmann J, Page B (eds) Workshop Simulation in den Umwelt- und Geowissenschaften. Shaker-Verlag, Hamburg, pp 169–180

Voigt K, Brüggemann R, Scherb H, Cok I, Mazmanci B, Mazmanci A, Turgut C, Schramm K-W (2013) Evaluation of organochlorine pesticides in breast milk samples in Turkey applying features of the partial order technique. Int J Environ Health Res 23(3):226–246

Winkler PM (1982) Average height in a partially ordered set. Discrete Math 39:337–341

Winkler PM (1983) Correlation among partial orders. SIAM J Algebraic Discrete Methods 4:1–7

Chapter 18
Indicator Analyses: What Is Important—and for What?

Lars Carlsen and Rainer Brüggemann

Abstract Simple elements of partial order theory appear helpful for a causal analysis in the context of ranking. The Hasse diagrams may seem as a confusing system of lines and a high number of incomparabilities. Thus, they indicate that metric information may be lost, but, on the other side partial order tools offer a wide variety of additional information about the interplay between the objects of interest and indicators. In this chapter a series of tools are presented to reveal such information.

As an illustrative example the so-called Failed State Index (*FSI*) is used. *FSI* is a composite indicator based on 12 individual indicators by simply summarizing the single values. The *FSI* comprises 177 states, which are the objects of our study.

A selection of appropriate partial order tools are applied to reveal specific information about the interplay between the states and the 12 indicators, such as A: sensitivity analysis, where the indicators are ordered relatively to their impact on the structure of the partially ordered set, B: a "vertical," i.e., chain analysis that is directed towards the comparabilities within a Hasse diagram, and C: a "horizontal," i.e., antichain analysis focusing on incomparabilities, including also the use of tripartite graphs as well as a derivation of an ordinary graph.

Partial order does not necessarily constitute as a Multicriteria Method solving all inherent problems. However, this chapter discloses that a detailed analysis by partial order tools prior to a possible derivation of a ranking index apparently is highly attractive.

L. Carlsen (✉)
Awareness Center, Linkøpingvej 35, Trekroner, 4000 Roskilde, Denmark

Center of Physical Chemical Methods of Research and Analysis, al-Farabi Kazakh,
National University, 96A Tole Bi street, 050012 Almaty, Kazakhstan
e-mail: LC@AwarenessCenter.dk

R. Brüggemann
Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries,
Müggelseedamm 310, 12587, Berlin, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_18, © Springer Science+Business Media New York 2014

18.1 Introduction

To obtain a ranking among a series of objects of the study, in the present study 177 states, based on multiple criteria/indicators is a priori not an easy task. The problem is that on the one hand the various indicators may be conflicting, i.e., they do not all vary in the same way for the investigated objects, nevertheless aiming at the same goal of ranking. On the other hand, if a perfect correlation between all indicators prevailed, only one single descriptor would be necessary as this selected indicator would mimic the influence of all indicators on the ranking. This apparent problem is often overcome simply by generating one single composite indicator, e.g., by addition of the single – possibly weighted and [0,1]-normalized – indicator. Although the field of multi-criteria decision support systems (DSS) offers a variety of approaches, this use of a simplistic version of utility theory (Fishburn 1970; Schneeweiss 1991) is often preferred. A promising approach, overcoming some of the disadvantages of conventional weighted sums, was developed by Yager (1988; 1993) where the weights are based on fuzzy "and" or "or" decisions. A critical recent discussion of DSS can be found by Munda (2008).

As an illustrative example of a Multi Indicator System (MIS), it serves the so-called Failed States Index (*FSI*) (FSI 2011) generated by the Fund for Peace (FFP 2011a). The 2011 index comprises 177 states that are ranked based on a composite index, the *FSI* index, which is composed by simple addition of 12 individual indicators (FFP 2011b), which comprises **Social indicators** (1: Mounting Demographic Pressures, 2: Massive Movement of Refugees or Internally Displaced Persons, 3: Legacy of Vengeance-Seeking Group Grievance or Group Paranoia, 4: Chronic and Sustained Human Flight), **Economic indicators** (5: Uneven Economic Development Along Group Lines, 6: Sharp and/or Severe Economic Decline), and **Political/Military Indicators** (7: Criminalization and/or Delegitimization of the State, 8: Progressive Deterioration of Public Services, 9: Suspension of the Rule of Law and Widespread Violation of Human Rights, 10: Security Apparatus Operates as a "State within a State," 11: Rise of Factionalized Elites, 12: Intervention of Other States or External Political Actors) (Baker 2006; FFP 2006, 2011b). For a more elaborate description of the single indicators, see Appendix 1. For the data, see the Appendix 2.

Obviously, having the simple scale based on the additive index construction, ranking is easy and straightforward. The price for such a simplification is in the best case that a significant amount of valuable information is lost. However, more crucial seems that such simple addition of the indicator values may lead to quite erroneous conclusions as high score(s) in certain indicator(s) may be leveled off by low scores in other indicator(s), without taking into account that these indicators point towards quite different topics. This compensation among indicator values is the main disadvantage in any construction of a one-dimensional ranking index from a MIS (Munda 2008).

It is immediately intelligible that the variation in the single indicator values may not be identical for all 12 indicators, since the indicators are mutually not or only slightly correlated (cf. Fig. 18.1, depicting the correlation between indicators

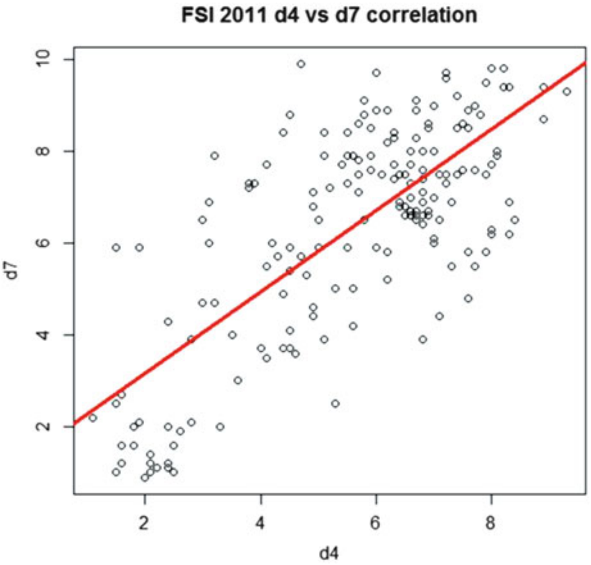


Fig. 18.1 Correlation between indicators *d*4 and *d*7 ($r = 0.734$, number of nations (cases) 177)

*d*4 and *d*7 as an illustrative example), the correlation coefficients, *r*, ranging from 0.61 to 0.95. Thus, conflicts among the states are unavoidable. In fact, there is no single natural measure to quantify the extent how a state fails. In a recent paper we discussed this apparent problem and suggested the partial order methodology to deal with these conflicts (Brüggemann and Carlsen 2012).

It is important to note that in contrast to a ranking based on a composite indicator, which obviously lead to a linear order, the partial order methodology will display a number of incomparabilities among the investigated elements due to conflicting indicator values. We have in recent papers studied the Failed States applying partial order tools (Carlsen and Brüggemann 2012; Carlsen and Brüggemann 2013). In Fig. 18.2 the rather complicated Hasse diagram, visualizing the partial order ranking, is displayed.

It should be noted that partial order methodology provides a (weak (i.e., tied ranks are not excluded)) linear order by calculating the average rank of the single objects. Both exact methods, based on lattice and polytope theory (Wienand 2006; De Loof et al. 2006) and approximate methods based on the local partial order model (abbr.: LPOM) (Brüggemann et al. 2004; Brüggemann and Carlsen 2011) are available. The comparison between ranks due to a composite indicator and the weak order based on average ranks was previously discussed in detail using the Failed States data (Carlsen and Brüggemann 2013), and it shall not be further discussed in the present context.

At a first glance, the partial order methodology leading to a picture as displayed in Fig. 18.2 may look rather confusing due to the significant number of incomparable

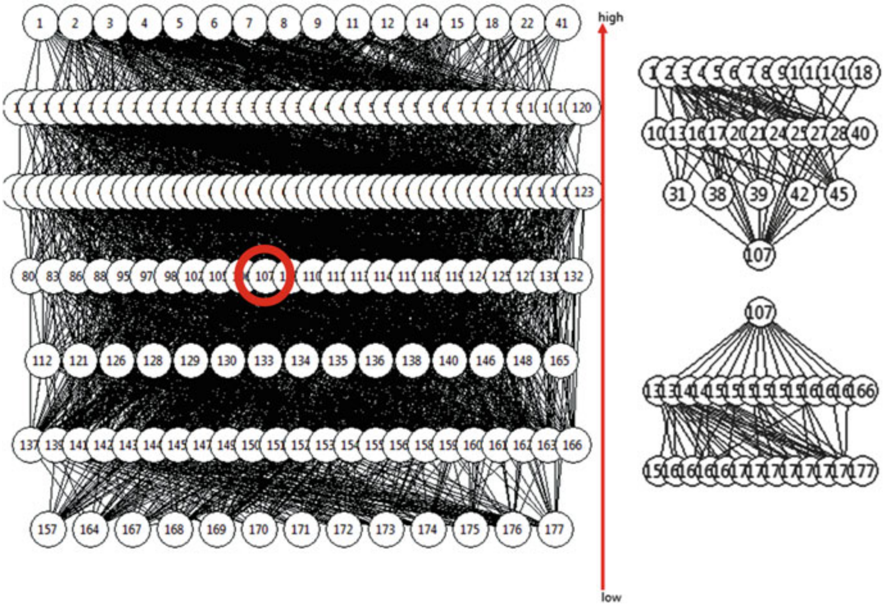


Fig. 18.2 Hasse diagram based on the *FSI* data with 177 states and 12 indicators. The ID's for the single states are kept as used in the *FSI* Orientation: High ID values indicating stable states at the *bottom*; Low ID values, indicating unstable states at the *top*. The *insets* show as an example the connections upwards and downwards for Kazakhstan (#107) as an example

as well as comparable objects (the lines in Fig. 18.2). Following the usual drawing rules (Brüggemann and Patil 2011), we find the elements (states) studied being positioned in levels, each level being a so-called anti-chain where the single elements by construction are incomparable. Although we will focus on those antichains, which are rendered by the levels, it is worthwhile to mention that a variety of other incomparabilities are found in the partial order (see for instance Fig. 18.9), which in principle can be treated analogously to what will be described in the following.

This chapter focuses on the analyses of indicators. What general information can we actually get from analyzing the indicators and what information can be retrieved more specifically in the cases where we are dealing with comparable elements and especially when we are dealing with incomparable elements. From a statistical explorative point of view, clearly other methods such as principal component analysis and cluster analysis may be applicable too. However, in the present context the aspect of ranking and evaluation is the focus.

The organization of this chapter follows the idea to outline some important modules of the software PyHasse.

1. An overview about the module *mainHD20_5.py* is given focusing on those tools, which will be applied within the context of Failed Nations. It is assumed that the reader is familiar with the basic concepts of partial order (see for instance Brüggemann and Voigt 2008), where the main module *mainHD20* is described

and the special application of the "level" structure is rendered. The "levels" being subsets of nations are equivalence classes under the equivalence relation: "Same length of maximal chains upwards." Thus, for example, nation 177 is in the same level as nation 80, because both have chains of maximum length, including four vertices in the Hasse diagram (Fig. 18.1). By the level construction a weak order is found for the set of objects. This weak order is highly degenerated, as the equivalence classes are large. Later, weak order due to average ranks will be introduced which remediate the situation.

2. The next important module is the sensitivity-module *sensitivty18_3.py*. This module is designed to find an order among indicators with respect to how far they affect the graph-theoretical structure of the partially ordered set and thus their relative importance in relation to the ranking. It will be shown that the 12 indicators have a very different impact on the structure of the Hasse diagram.
3. The module *chain7.py* is helpful when the graphical presentation of the partially ordered set is very complex, as in the case studied here (see Fig. 18.2). Thus, by visual inspection it is difficult to find out sequences of objects which are mutually comparable. Any chain, once identified, can be seen as a subset of objects with mutually high correlation between any pair of indicators. The point is, however, to find these subsets. This is obviously a graph-theoretical task and is solved with tools taken from algorithmic graph theory.
4. The most typical and striking aspect of an application of partial order theory on data matrices is the appearance of incomparabilities. As in hierarchical cluster analysis, where clusters are not only linearly sequenced, partial order theory lead to graph-theoretical structures, which beside the vertical component also have a horizontal component. The horizontal development is a consequence of incomparabilities, and it is a main task to investigate the reasons for incomparabilities in terms of the indicators used in the data matrix. Two modules of PyHasse are devoted for this purpose: *antichain20.py* and *sepanal16.py*. As the name may indicate, *antichain20* analyzes sets of objects which are mutually incomparable, whereas the module *sepanal16* is to analyze why two subsets of objects are not or only loosely connected within a Hasse diagram. Here, concepts of *sepanal16* are applied on trivial object subsets, namely on singletons out of an antichain. The main tool is the analysis of tripartite graphs, which will be explained in more detail in the appropriate section.
5. Note in the following we omit the extension *py*, which indicates that the programming language is Python.

18.2 Methods

In the present study we use the data provided by the Fund for Peace (FFP 2011a) for the 12 indicators applied for generation of the 2011 *FSI* (FSI 2011; Baker 2006; FFP 2006, 2011b), which we have earlier treated using partial order methodologies (Carlsen and Brüggemann 2012, 2013).

The central concept in partial order is the "concept of comparison." Studying objects applying partial order methodology almost unambiguously leads to a series of incomparabilities between some of the elements included (cf. Fig. 18.1). The above shown partial order based on the *FSI* comprises 6,307 comparabilities and 9,269 incomparabilities with no equivalent elements, e.g., elements with all indicators being pair-wise identical.

Calculations are made using the PyHasse software, which has been developed (and currently being extended and improved) by the second author of this chapter. In the present context we shall not describe the principles of partial order theory as this can be found in numerous publications previously published (Al-Sharrah 2010; Annoni and Brüggemann 2008; Bick et al. 2011; Brüggemann and Voigt 2008; Brüggemann and Patil 2010, 2011; Brüggemann and Carlsen 2006; Brüggemann 2011; Duchowicz et al. 2008; Freier et al. 2011; Kardaetz et al. 2008; Newlin and Patil 2010; Restrepo et al. 2008; Tsakovski and Simeonov 2011; Voigt and Brüggemann 2008).

In the following a short description of the modules of the PyHasse software package that have been applied in the present work is given.

18.2.1 PyHasse

The calculations are performed using the PyHasse software. PyHasse is programmed using the interpreter language Python (version 2.6), which is freely downloadable from the Internet. PyHasse is available on request from the second author and should be considered as experimental, nonprofessional software. Beside DART (Talente 2007), the software package PyHasse is the actual available one for applying ordinal analyses on data matrices. When a ranking is intended or an ordinal evaluation, then this software may be appropriate.

When interfaces and more technical tools are counted too, PyHasse consists actually of close to 100 modules. These are specific programs, such as canonweight9, which analyzes the role of weights in terms of partial order theory or similarity9, which provides tools to compare two partial orders of the same set of objects. There are even modules, which are simplified versions of classical MultiCriteriaDecisionAnalysis (MCDA) method, such as PROMETHEE or ORESTE, or TOPSIS.

Only a limited number of the available modules is applied and described in this chapter.

18.2.2 Main Module (mainHD20_5)

The main entrance to partial order ranking studies applying the PyHasse software is the module *mainHD20_2* that offers information on, e.g., a chain statistics rendering the number and lengths of chains (comparable objects) and the level structure (objects in the same vertical position in a Hasse diagram, comprising number of objects in each level). The concept of levels constitutes the simplest approach to obtain a weak order from the objects. Beyond this, the number of successors of each element (downward comparable elements), of predecessors (upward comparable objects), and incomparable elements is rendered. Furthermore, a "local" Hasse diagram is constructed, where a specific object can be selected and its order relations are displayed.

In addition to drawing of the Hasse diagram (HD) based on the data matrix, a two-dimensional representation of the objects by a FOU-plot is offered: Each object x is characterized by the coordinates $U(x)$ (number of elements incomparable with x) and $|F(x)| - |O(x)|$. The quantity $|F(x)| - |O(x)|$ is the difference of the contents of the principal upset of x and downset of x.

The module *mainHD20_2* also includes tools for an order theoretical navigation. It renders information on three important order theoretical subsets namely

- Principal upsets, F(x) (synonym: principal order filters), i.e., objects being upwards comparable to a given object studied
- Principal downsets, O(x) (synonym: principal order ideals), i.e., objects being downward comparable to a given object studied
- Order interval graphs, i.e., objects order theoretically between two given objects studied

(For further details see Brüggemann and Patil 2011.)

Finally, the main module provides tools to get weak orders, somewhat more sophisticated than that, by the levels. For example, one can select the method of Bubley and Dyer (1999) or a simple LPOM approach (Brüggemann et al. 2004). Whereas the Bubley Dyer method offers a rather good approximation based on a Markov chain Monte Carlo simulation, the LPOM method in *mainHD20_5* is thought of as a screening method for first considerations.

18.2.3 Sensitivity (Sensitivity18_3)

Working with a MIS it is obviously of significant interest to retrieve information about the relative importance of the single indicators as it potentially will provide crucial input for possible actions, e.g., to be taken by the authorities or regulators.

The sensitivity expresses how important any single indicator is for the structure of the partial order (i.e., the system of levels, chains, and antichains). The PyHasse module *sensitivity18_3* offers an estimation of the global sensitivity (for all objects), the local sensitivity for any single object as well as an approach to decompose the set of indicators into one of important and one of "fine-tuning" indicators (Brüggemann and Patil 2011). The outcome of the sensitivity analyses is presented both in tabular as well as in graphical form (for details see: Brüggemann et al. 2001).

The leading idea is to find distances among partially ordered sets (posets), where one poset is the original one with all indicators and the others are those posets where one indicator is left out. The most important indicator for the structure of the MIS is obviously that one, whose elimination from the data matrix results in the maximal distance to the original one. Recently, an elegant and mathematically deep approach was provided by Annoni et al. (2011), where a variance-based sensitivity analysis is suggested. The variance-based sensitivity analysis provides two important advantages over the approaches presently available in PyHasse (1) The role of indicator values and that of changing the indicator set (attribute value sensitivity and attribute-related sensitivity, respectively (Brüggemann and Patil 2011), are that they are simultaneously analyzed and (2) the mutual influences of indicators on the graph-theoretical structure of the partially ordered set can be quantified. As shown in Annoni et al. (2011) the influence of the uncertainty in indicator values of one indicator depends on those of all others. The "total order sensitivity coefficient is indicative for interactions among the indicators". For the present, this approach is implemented in Matlab only and not yet available in PyHasse.

18.2.4 Chain Analyses (Chain7)

Whereas only a crude statistical view concerning the chains is taken in *mainHD20_5* about the frequency of chain lengths, the module *chain7* provides a detailed analysis: an analysis of chains, e.g., series of comparable elements starting from the top (or a higher level) of a Hasse diagram and walking through it "vertically," i.e., following the comparability indicating edges downwards until an element of the bottom, or lower, level of the diagram is reached. Such a chain is of interest because we have series of comparable objects where all indicators monotonously change in decrease from the start (source) to the end (sink). Thus, based on the choice of source and sink, the PyHasse module *chain7* provides information on

- All possible chains i.e., the number of chains with number of objects ("height") above a given threshold
- The height of the single chains
- The average number of elements in the chains
- Information on how far within one chain indicator values are increasing, relatively to the full range of the indicator of interest
- Information on the individual chains such as the included objects
- The decomposition of a Hasse diagram into disjoint parts rendering similar data profiles (A data profile is, for example, $d1(x) > d2(x)$, d_i being normalized indicators. Then it is of interest which other objects would have the same relation in $d1$ and $d2$.)

18.2.5 Antichain Analyses (Antichain20)

Objects found at the same level in the Hasse diagram are by definition incomparable and constitute a so-called antichain. In studies on the comparison of objects, it is obviously of importance not only to study actually comparable objects but also to elucidate the reasons that cause incomparabilities between objects ("horizontal analysis").

The module calculates the number of times the indicators exceed given threshold values for the objects of an antichain. The threshold values are set individually for the single indicators by the user.

To illustrate this consider three objects characterized by four indicators within an antichain with the data profiles (1,2,3,4), (2,2,2,2), and (4,4,3,1). Now select a threshold value equal to 2. Then the first object exceeds this value with respect to two of 4 indicators, the second object for no indicator (nevertheless, it is incomparable with the two others), the third object with respect to 3 indicators.

Obviously, the resulting values for the single objects may vary from 0 to m, m being the number of indicators. The background for this tool is that in practical applications, two objects may be incomparable because of many conflicting attribute pairs. However, the numerical differences may be only slight for every conflict. Hence, the number of indicators exceeding a certain limit gives initial information. The evaluation of object pairs therefore needs two types of information:

1. The number of conflicting indicator pairs
2. What is the largest numerical discrepancy

Many tools of the PyHasse that module *Antichain20* are based on the antichain matrix, where the rows are defined by the pairs of objects, being incomparable, whereas the columns are defined by the pairs of indicators. An entry of this antichain matrix can be 0 if the indicator pair (d_i,d_j) does not contribute to the incomparability of the object pair (x,y) or 1 if it contributes (Brüggemann and Voigt 2012). By the antichain matrix information is available about which object pair is most often separated by the indicators and which indicator pair is most often participating in the separation of the object pairs. It is convenient to introduce row- and column sums and by normalizing them to get "densities." An object pair whose normalized row sum equals 1 is incomparable with respect to all available indicator pairs. An indicator pair with density 1 is causing the incomparability for all object pairs taken from the set of objects of the selected antichain.

The *Antichain20* module additionally offers graphical elucidation of which indicator pair(s) is/are involved in the incomparability of a certain pair of objects, which graphically is displayed by the so-called tripartite graph (Brüggemann and Voigt 2011).

As mentioned above, normalizing the number of cases where a given indicator pair is causing incomparabilities leads to the density of indicator pairs, which is a number in the interval [0,1]. Choosing a limiting value, limit, we can draw a graph, where the two vertices d_i and d_j are the indicators and a connecting line is drawn, when the density ≥ limit (most often the limit is selected as median or as third quartile of all densities related with antichain being studied). The valence, which we define as the number of incident edges, tells us how important an indicator is in causing an antichain. Details can be found in Brüggemann and Voigt (2012).

Table 18.1 Level population of the Hasse diagram based on the *FSI* data with 177 states and 12 indicators (Fig. 18.2) (Level 1: bottom, level 7: top)

Level	Population	Range of FSI-rank
7	1 2 3 4 5 6 7 8 9 11 12 14 15 18 22 41	1–41
6	10 13 16 17 19 20 21 23 24 25 26 27 28 29 30 32 33 35 36 37 40 43 44 52 53 54 55 57 59 61 71 72 75 77 90 100 103 109 120	10–120
5	31 34 38 39 42 45 46 47 48 49 50 51 56 58 60 62 63 64 65 66 67 68 69 70 73 74 76 78 79 81 82 84 85 87 89 91 92 93 94 96 99 101 104 116 117 122 123	31–123
4	80 83 86 88 95 97 98 102 105 106 107 108 110 111 113 114 115 118 119 124 125 127 131 132	80–132
3	112 121 126 128 129 130 133 134 135 136 138 140 146 148 165	112–165
2	137 139 141 142 143 144 145 147 149 150 151 152 153 154 155 156 158 159 160 161 162 163 166	137–166
1	157 164 167 168 169 170 171 172 173 174 175 176 177	157–177

18.3 Results and Discussion

From the Hasse diagram (Fig. 18.2) it is seen that the partial order ranking leads to a structure with seven discrete levels where the lowest level (level 1) comprises the most stable states whereas the top level (level 7) includes the most unstable states (Table 18.1).

The last column of Table 18.1 shows that there is some increasing tendency of the range of *FSI* ranks from the bottom to the top level. This shows that if a state is good with respect to one indicator, there is a tendency to be also good in most of the other indicators. The pretty large population of each level shows the high degree of degeneracy when the level-concept is applied to obtain a weak order.

18.3.1 Indicator Importance

First of all the question arises: what indicator(s) influence(s) the partial order ranking the most, i.e., which indicator(s) being of highest importance (attribute-related sensitivity)? This is obviously of major interest if some intervention is necessary, e.g., to improve the overall ranking of an object. Looking at the single indicators, all given equally weight for the *FSI*, it is hard to believe that they actually should be equally important for rating the countries according to their stability or most at-risk of collapse and violence, respectively.

In the present case a sensitivity analysis (Carlsen and Brüggemann 2012) clearly demonstrated that the different indicators have different importance in relation to the overall ranking of the 177 states, which is illustrated in Fig. 18.2 that visualizes the relative importance of the 12 indicators.

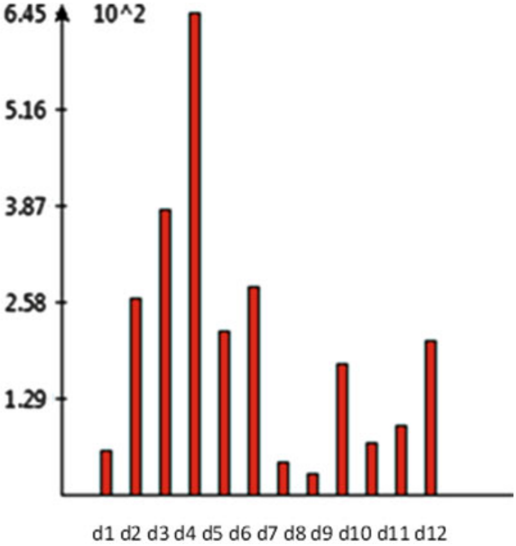


Fig. 18.3 Relative importance of the 12 FS indicators (*d*1, *d*2, …, *d*12) as disclosed by partial order methodology

It is immediately clear (Fig. 18.3) that on the global scale (i.e., taking all 177 states into account), the absolute top rating among the indicators is indicator *d*4, which describes the "Chronic and Sustained Human Flight," i.e., brain drain, followed by indicators *d*3 ("Legacy of Vengeance-Seeking Group Grievance or Group Paranoia") and *d*6 ["Sharp and/or Severe Economic Decline" including "Increase in levels of corruption and illicit transactions among the general populace" (Baker 2006)]. This may not be surprising. However, this is obviously not in any way elucidated by the original *FSI*.

If we look at the relative importance of the 12 FS indicators on a local scale, i.e., referring to single states, the picture is more differentiated. On the one hand one finds a sensitivity pattern similar to that of Fig. 18.3, on the other hand the differences among the relative sensitivity values seem to smooth out when moving from the most vulnerable states to the more stable. Hence, for the most vulnerable states, the picture is similar to that shown in Fig. 18.3. Thus, for Somalia (#1) we find $d4 \gg d3$, whereas for the most stable states like Finland, (#177) the single indicators appear to be virtually of equal importance. For a state placed somewhere in the middle of the *FSI*, e.g., Kazakhstan (#107), we find some intermediary values, however, brain drain (indicator *d*4) still being the most important. To fully visualize this, Table 18.2 summarizes the relative importance of the single indicators, both on a global scale as well as on the local scales using #1 (Somalia), #107 (Kazakhstan), and #177 (Finland) as representative for the top and bottom of the *FSI* ranking (Somalia and Finland, respectively) and a country being intermediary ranked (e.g., Kazakhstan).

Table 18.2 Relative importance of the single indicators on a global scale as well on local scales for #1 (Somalia), #107 (Kazakhstan), and #177(Finland)

Indicator	Global	Somalia #1	Kazakhstan #107	Finland #177
*d*4	0.261	0.429	0.225	0.087
*d*3	0.155	0.095	0.143	0.086
*d*6	0.113	0.048	0.085	0.086
*d*2	0.107	0.048	0.083	0.082
*d*5	0.089	0.048	0.082	0.082
*d*12	0.084	0.048	0.080	0.082
*d*9	0.071	0.048	0.057	0.082
*d*11	0.038	0.048	0.050	0.082
*d*10	0.028	0.048	0.049	0.082
*d*1	0.024	0.048	0.049	0.082
*d*7	0.018	0.048	0.049	0.082
*d*8	0.012	0.048	0.049	0.082

Based on these calculations it may be concluded that, especially in the most vulnerable states, such as Somalia, focus should be on initiatives that will secure not only the generation of human capital but of equally high priority that the human capital generated in the single countries is retained there for future benefit of the state. Further focus areas appear to be possible decreasing group paranoia and reducing or even stopping economic decline.

In the case of economic indicators, it is of interest to note that indicator *d*6 that includes "Increase in levels of corruption and illicit transactions among the general populace" has a significant higher impact on ranking (sensitivity) than indicator 7 that includes "Massive and endemic corruption or profiteering by ruling elites" (Baker 2006). However, these data are in accordance with the widespread corruption that prevails in countries like, e.g., Somalia and Kazakhstan (Transparency 2011). Thus, on a scale from 0 (highly corrupt) to 10 (very clean) Transparency International (Transparency 2011) has ranked 183 countries (including the actual considered 177 states included in the *FSI*) with Finland as one of the least corrupted countries almost in the top on place 2 with a score of 9.4 and Somalia as the most corrupted country in the bottom on place 183 with a score of 1.0. Kazakhstan is also here found on rather low intermediary place 120 with a score of 2.7 (Transparency 2011).

18.3.2 Chain Analyses

As mentioned in the introduction, the *FSI* is a composite indicator and as such the *FSI* leads to a linear order where all states are mutually comparable. Thus, an obvious question will be to what extent are the single states comparable in the partial order approach? A good starting point for such a discussion could be to select states from the bottom and top level. Here for example we selected state Finland (#177) and Somalia (#1). Finland is located in the bottom level and based on the *FSI* the most stable state, whereas Somalia is found at the top level as the most unstable state in the study of "failed states." Hence, the question is: are Finland and Somalia

Table 18.3 Ten arbitrarily chosen chains between #177 (Finland) and #1 (Somalia)[a]

Chain number	Count of states	
526	7	177, 139, 112, 80, 31, 10, 1
1034	7	177, 143, 126, 80, 31, 10, 1
1042	7	177, 143, 126, 105, 31, 10, 1
1672	7	177, 147, 112, 80, 31, 10, 1
1924	7	177, 149, 126, 80, 31, 10, 1
1932	7	177, 149, 126, 105, 31, 10, 1
2059	7	177, 149, 135, 80, 31, 10, 1
2349	7	177, 152, 130, 97, 38, 10, 1
2433	7	177, 152, 133, 105, 31, 10, 1
2705	7	177, 153, 121, 86, 31, 10, 1

[a] For legends to the single states the Appendix 2 should be consulted

Table 18.4 Individual indicators for the 7 states in the #1 (Somalia)–#177 (Finland) chain 2059

ID	State	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12	FSI
1	Somalia	9.7	10	9.5	8.2	8.4	9.3	9.8	9.4	9.7	10	9.8	9.7	113.5
10	Cote d’Ivoire	8.1	8.5	8.7	7.9	8	7.7	9.5	8.4	8.6	8.6	9.1	9.7	102.8
31	Kyrgyzstan	7.6	6.5	8.3	7	7.6	7.6	9	6	8	8	8.3	7.9	91.8
80	Venezuela	6	4.8	7	6.4	7.3	6.1	7.5	5.8	7.4	7	7.3	5.5	78.1
135	Lativa	4.2	3.9	4.9	4.8	5.7	5.8	5.3	3.9	3.6	3.3	4.3	4.4	54.1
149	Lithuania	4.1	3.2	3.7	4.6	5.7	5.3	3.6	2.9	3.1	2.5	2.8	3.8	45.3
177	Finland	2	2.1	1.7	2.5	1.3	2.8	1	1.5	1.1	1	1.2	1.5	19.7

comparable? One could get immediately the answer if the Hasse diagram would be a clear graph. However, in the present case the system of lines is too messy to obtain an answer simply by visual inspection. This general problem motivated to the development of the chain module. Hence, by chain analysis it is seen that in fact the two states are comparable as a chain analysis disclosed 52 chains starting with Finland (#177) and ending at Somalia (#1)! In Table 18.3 ten arbitrarily chosen chains between Finland (#177) and Somalia (#1) are rendered.

As can be seen, the starting and ending point must be by construction the same. The sequence 31, 10, 1 is 9 times realized, more differences among the chains are found in the lower part. As only an arbitrary number of chains are selected out of a total of 52, a firm conclusion should not be drawn. This has been left for future research.

Obviously, all 12 indicators for the 52 chains will monotonously decrease from the low values found for #177 to high values found for #1. In Table 18.4 the indicator values for chain No. 2059 can be seen.

A closer look at the data (Table 18.3) displays obvious differences between top and bottom. As illustrative examples can be used the indicators *d*7: Criminalization and/or Delegitimization of the State and *d*10: Security Apparatus Operates as a “State within a State,” range from 1 for #177 to 9.8 and 10, respectively, for #1 thus realizing the full scale (1–10). Further Table 18.4 clearly points to the differences in the stable Finland vs. the highly unstable Somalia.

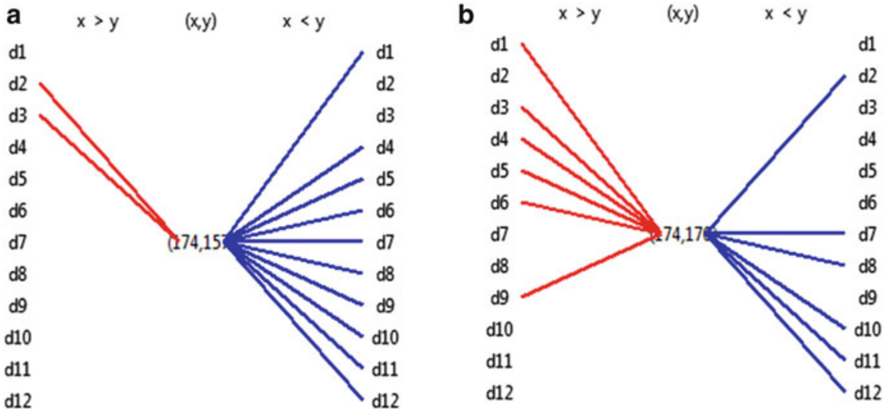


Fig. 18.4 Comparisons between #174 (Switzerland) and (**a**) #157 (Singapore) and (**b**) #170 (Luxembourg), respectively

18.3.3 Antichain Analyses

Antichain analyses in relation to the "Failed States" have initially been presented by Carlsen and Brüggemann (2013). Further and more elaborate analyses are rendered here based on the PyHasse module *Antichain20*.

The discussion is twofold:

1. What are the reasons for incomparability between two specific states?
2. To what extent do the single indicators contribute to the formation of a specific antichain?

It is initially important to note that the answers to the above questions are not necessarily related to the impact of the single indicators as the outcome of the sensitivity analysis; see, however, the final discussion of this subsection. The focus is to search for differences in the indicator values between incomparable states (here of incomparabilities arising among states of a specified level). Hence, in the present context we concentrate on the top (7) and bottom (1) level, respectively.

The reasons for incomparabilities between two specific elements can be illustrated through the so-called tripartite graphs (Brüggemann and Voigt 2011). Let x and y be incomparable objects. The idea behind tripartite graph is to make evident which indicator implies $x>y$ and which $x<y$. Therefore, an arrangement into three vertical parts appears useful. To the left and to the right a list of indicators are displayed and in the middle part the set of pairs of objects taken from the antichain under investigation is given. If an indicator on the right side implies $x>y$, then a line connects this indicator with the pair (x,y). If another indicator implies $y>x$ (as it must be, when $x||y$) then a line of the indicator from the left side is drawn to the pair (x,y).

Clearly this representation has its limits when the number of pairs of objects or of indicators is large. Therefore, there are many tools for a local analysis. That is,

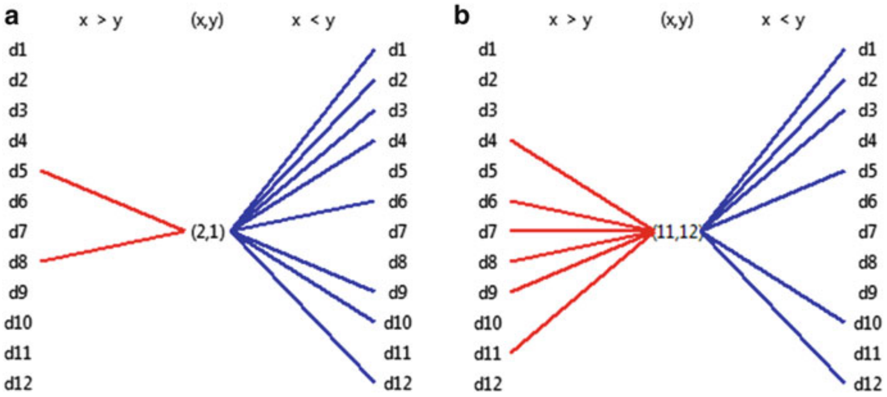


Fig. 18.5 Comparisons between (**a**) #2 (Chad) and #1 (Somalia) and (**b**) #11 (Guinea) and #12 (Pakistan), respectively

starting either from a selected object (state), or a selected pair of objects (states), or from an indicator. This local analysis is applied here. A representation by a bar diagram (also available in sepanal16) indicates that how often an indicator implies a $>-$or $<-$relations among the pairs of objects is useful too; however, it does not indicate which indicator pair actually is contributing to $x||y$.

Looking at level 1, the most stable states, we exemplify the procedure by economically affluent states like Switzerland, Singapore, and Luxembourg. In Fig. 18.4 the tripartite graphs for the pairs #174 (Switzerland)/#157 (Singapore) (Fig. 18.3a) and #174 (Switzerland)/#170 (Luxembourg) (Fig. 18.3b) are shown.

A detailed analysis and explanation of the single indicators is outside the scope of the present chapter. It can just overall be concluded that in the cases shown in Fig. 18.4a, wide range of conflicting indicators is in play. However, it can further be noted that in both the two studied cases, all 12 indicators are involved in generating incomparabilities. Note the unbalanced nature in Fig. 18.4a (states 174, 157), where 10 indicators are causing a higher instability of Singapore in contrast to only 2 ($d2$: Massive Movement of Refugees or Internally Displaced Persons and $d3$: Legacy of Vengeance-Seeking Group Grievance or Group Paranoia) of Switzerland. In Fig. 18.4b (states 174, 170) the contributions of the indicators are balanced: 6 indicators favor Switzerland over Luxembourg and 6 others favor Luxembourg over Switzerland.

The same is virtually the case looking at states originating at level 1 as, e.g., #2 (Chad) and #1 (Somalia) and #11 (Guinea) and #12 (Pakistan), respectively, Fig. 18.5. Here, again all indicators are involved in the comparison between #11 (Guinea) and #12 (Pakistan), whereas for the study on #2 (Chad) and #1 (Somalia) indicators 7 and 11 apparently are not involved in the incomparability.

Even an antichain or level has its degree of differentiation. Some pairs of states obviously are incomparable because almost the same number of indicators favor one state over the other, whereas in some cases, such as for pair (174, 157) or pair (2,1)

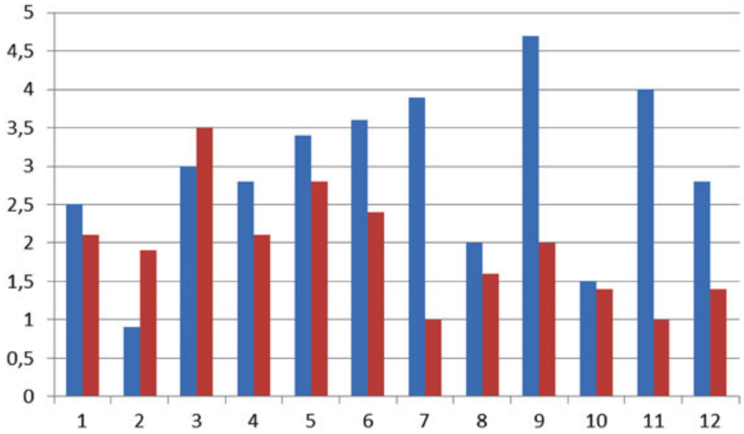


Fig. 18.6 Indicator values for state #174 (Switzerland) (*red*) and #157 (Singapore) (*blue*)

only few indicators contradict the tendency realized by the majority of indicators, namely 157 > 174 and 1 > 2, respectively, with respect to 10 indicators.

A somewhat more detailed analysis can be obtained by looking at a simple bar diagram displaying the single indicator values for the two elements under discussion. Thus, in Fig. 18.6 the comparison between the indicator values for #174 (Switzerland) and #157 (Singapore) is depicted.

It is here immediately clear that the two states #174 and #157 as such are incomparable. Thus, it is seen that #157 is worse than #174 with respect to indicator 1 and 4–12, whereas the reverse is true for indicator 2 and 3, i.e., #174 is worse than #157 in agreement with Fig. 18.4a. It is further noted that from a numerical point of view, the difference in terms of values favoring Switzerland are small in comparison to those favoring Singapore.

Turning to the overall assessment of the influence of the single indicators in the incomparabilities at the level state, we have to look at the density distribution of indicator pairs (cf. Methods' section). From a more philosophical point of view, this analysis reflects the fact that entities (here such as the set of the 7 levels) are not a homogenous system, ordered just by the level number but can be highly heterogeneous. Indeed, it is very important to provide tools to exploit "local" features. Here, the role of indicator pairs can differ from one to the next level.

The number of possible element pairs at level 1 (13 element, cf. Table 18.1) is 78, whereas the number of pairs of indicators is 66, as all 12 indicators are in play in this analysis. The third quartile of the density distribution is 0.478. In Fig. 18.7 the graph of indicators is drawn corresponding to density values ≥ 3rd quartile (0.478).

Figure 18.7 clearly shows that the indicators 2 and 4 (valences 7 and 6, respectively) play dominant roles. Thus, these indicators are in different combination with other indicators most often responsible for the incomparabilities between the single states of level 1 (cf. Fig. 18.2 and Table 18.1). The valences indicate in a

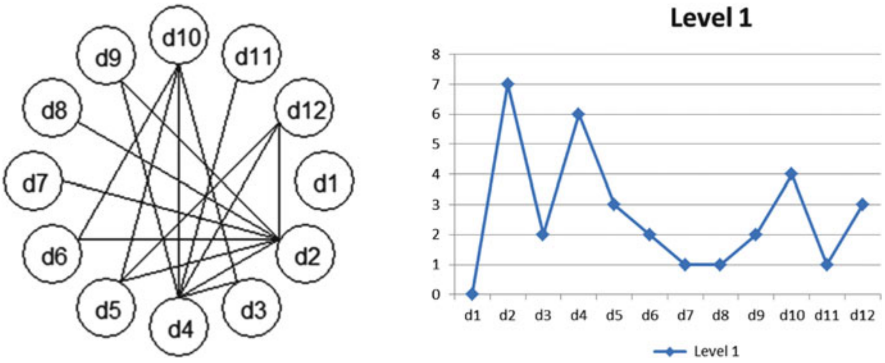


Fig. 18.7 Graph of indicators for level 1 for density values ≥ 3rd quartile (0.478) and a plot of the valences of the single indicators

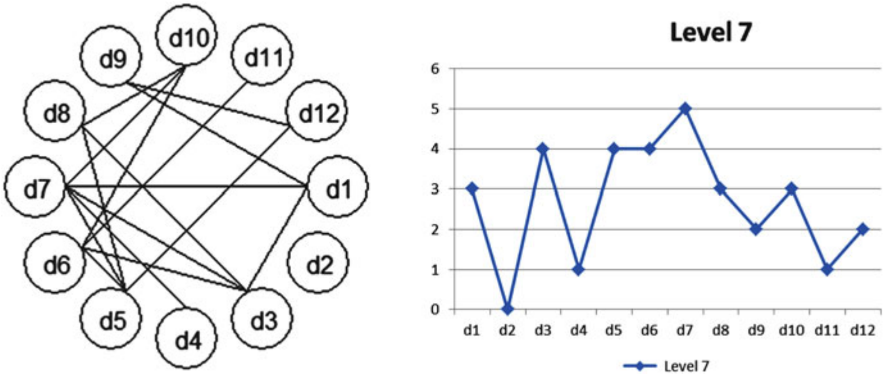


Fig. 18.8 Graph of indicators for level 7 for density values ≥ 3rd quartile (0.469) and a plot of the valences of the single indicators

quantitative way the importance of the indicators for causing incomparabilities in antichains, like in level 1.

A similar analysis of level 7, i.e., the most unstable states, gives a different picture (Fig. 18.8). The number of object pairs (16 objects) is 120, whereas clearly the number of indicator pairs remains 66. The 3rd quartile of the density distribution is in this case 0.469. In Fig. 18.8 the corresponding graph of indicators are depicted. It can here be seen that the most important indicator that in combination with others are responsible for incomparisons is indicator 7 (valence 5) followed by indicators 3, 5, and 6 (all with valence 4).

Looking at the two graphs depicted in Figs. 18.7 and 18.8, respectively, a clear difference is seen. As we here looked at the top (level 7) and bottom (level 1) levels, it is a tempting thought that once again a “smooth,” i.e., monotonous change of the valence distributions throughout the seven levels could be observed and thus further

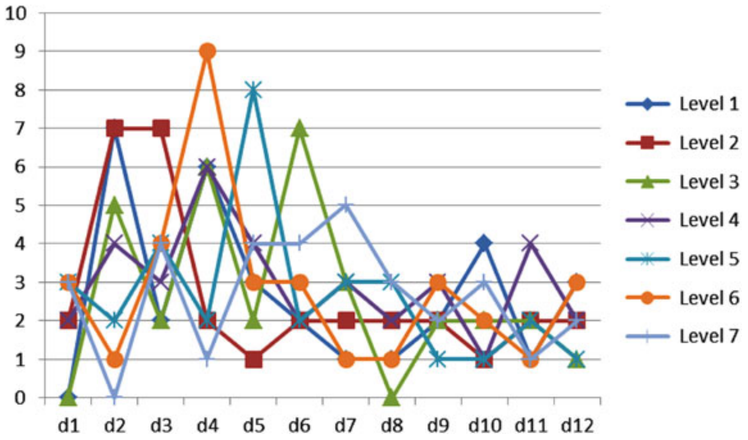


Fig. 18.9 Valence distribution for the 12 indicators for all 7 levels

contribute with info about the background for incomparabilities as function of, in the present case, state stability. A study along the lines above for the levels 1 and 7, however, disclosed that this apparently is not reality. In Fig. 18.9 a summary of the valence distributions for all 7 levels are depicted. It is immediately clear that no specific trend can be deducted, i.e., each indicator contributes in different combinations with others to the incomparabilities causing the antichain in that specific level.

Figure 18.8 shows qualitatively that indicator *d*4 has large values for many levels, followed by indicator *d*5.

A large value of valence in a level means that the corresponding indicator is most often responsible for the incomparabilities, which prevail in that antichain. Taking level 6 as an illustrative example it is seen (Fig. 18.8) that the indicator *d*4 seems to be rather important for the incomparabilities of this level as *d*4 9 times is contributing to all of the incomparable pairs generated from the objects of this level. Hence, in order to obtain a quantitative measure we calculated first for each level a relative number drel($d_{i,j}$), where *i* stands for the *i*th indicator d_i and *j* for the *j*th level:

drel($d_{i,j}$) = valence($d_{i,j}$)/Σ valences ($d_{s,j}$), where *s* varies over all indicators, i.e. 1 to 12.

In the case of indicator *d*4 we have the valence of *d*4 in level 6 is 9 (Fig. 18.8) and the sum of all valences of indicator *d*4 over all levels is 32. Therefore, drel(*d*4, 6) = 9/32 = 0.26, and similarly drel(*d*4, 1) = 6/32 = 0.19, etc.

The contribution of *d*4 over all levels is therefore dreltotal(*d*4) = Σ drel(4, level *t*) (*t* = 1, …, 7), which amounts to 0.96.

The quantity dreltotal informs about the overall responsibility of an indicator for the incomparabilities, appearing within all levels. In Table 18.5 the values of dreltotal(*i*) are shown, together with the results of the sensitivity study as retrieved from Table 18.2.

Table 18.5 Relevance of indicators *d*1, …, *d*12, and their global sensitivity values [taken from subsection "indicator importance" (rounded to two positions)]

	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12
dreltotal	0.39	0.80	0.80	0.96	0.76	0.68	0.55	0.37	0.45	0.43	0.39	0.43
Global importance[a]	0.024	0.107	0.155	0.261	0.089	0.113	0.018	0.012	0.071	0.028	0.038	0.084

[a] Imported from Table 18.2

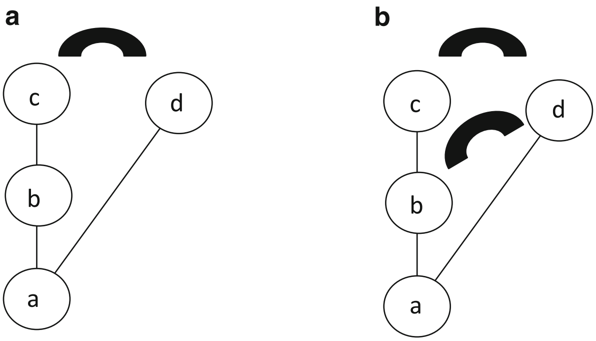


Fig. 18.10 The incomparability within the top level (**a**) is considered which is the basis for drel(*i*). All incomparabilities (**b**), especially also those which appear between any two levels, are the basis for the global sensitivity

By means of the dreltotal-values we found $d4 \gg d2 = d3 > d5 > d6 > d7 > d9 > d10 = d12 > d1 = d11 > d8$, which once again shows the importance of the indicator *d*4. There is a good coincidence to the result of the sensitivity analysis, although some discrepancies are noted. Thus, the sensitivity analysis leads to the order $d4 > d3 > d6 > d2 > d5 > d12 > d9 > d11 > d10 > d1 > d7 > d8$. The discrepancies arise, because the focus here is on the incomparabilities caused by the objects located in the same level. However, there are many further incomparabilities between objects of one and objects of another level that should be taken into account in a full analysis, which is done by the sensitivity analysis. Figure 18.10 shows the difference schematically.

18.4 Conclusions and Outlook

In this chapter we have shown how far simple elements of partial order theory can be helpful for a causal analysis in the context of ranking. The Hasse diagram in Fig. 18.1 shows a confusing system of lines and a pretty high number of incomparabilities. This high degree of incomparabilities indicate that by construction of a one-dimensional ranking index often the generally unwanted compensation among indicators appears: One good value may level out certain bad values and *vice versa*. So, the Hasse diagram indicates that we loose on the one side metric information when a composite indicator is generated, on the other side the analysis by partial order tools reveals many additional information about the interplay between the states and their indicator values. How do we get this additional information? `

First of all we perform a sensitivity analysis, where now the indicators themselves are ordered relatively to their impact on the structure of the partially ordered set. The next logical steps are the "vertical" and "horizontal" analysis. Whereas the vertical analysis is directed towards the comparabilities and the resulting chains within a Hasse diagram, the horizontal analysis is focusing on incomparabilities. The most striking visual effect of incomparability is that the Hasse diagram gets horizontal dimensions, best seen by the extent and number of levels.

The appearance of incomparabilities (causing levels with more than one element in them) is often seen as main disadvantage of partial order theory. This fact thus kicks partial order approaches out from the list of seriously considered decision support systems. However, the horizontal analysis as described in the present study is of main interest. Hence, we concentrated on the antichain analysis. The most important result is that even incomparable states information can be found. Thus, the appearance of antichains makes the decision about priority difficult, but at the same moment, we can find out the reasons why and to which extent certain incomparabilities appear. One of the potential techniques is the use of tripartite graphs.

Another tool which is useful in the context of a horizontal analysis is the derivation of graphs as shown in Figs. 18.7 and 18.8. In the top level, the most unstable states, the indicators *d*2: "Massive Movement of Refugees or Internally Displaced Persons" and *d*4: "Chronic and Sustained Human Flight" are most often causing the states to be incomparable. In the bottom level, the most stable states, the indicator *d*7: Criminalization and/or Delegitimization of the State is striking, followed by indicators *d*3, *d*5, and *d*6 ("Legacy of Vengeance-Seeking Group Grievance or Group Paranoia," "Uneven Economic Development Along Group Lines" and "Sharp and/or Severe Economic Decline," respectively).

We do not claim that the partial order is the Multicriteria Method solving all inherent problems in that field, but we propose to expand the use of analysis by partial order tools before the final step in derivation of a ranking index (by several methods) is performed. The simplicistic use of just a weighted sum of (normalized) indicators has one main advantage over many other sophisticated multicriteria methods: It is so simple that it is transparent and hence has rarely an acceptance problem. So it is a logical step in our eyes to continue the simplicity, i.e., in the simple rules of the Hasse diagram technique (see Brüggemann and Carlsen 2012) by the simple aggregation method of weighted sums. A major part of future work will deal with the relations between the sets of weights, needed to obtain a composite indicator, and the partial order. A first step is published by Brüggemann et al. (2013).

Appendix 1: Indicators of State Collapse and Internal Conflict (Baker 2006)

Social Indicators

1. Mounting Demographic Pressures

   - Pressures deriving from high population density relative to food supply and other life-sustaining resources.—Pressures deriving from group settlement patterns that affect the freedom to participate in common forms of human and physical activity, including economic productivity, travel, social interaction, religious worship, etc.
   - Pressures deriving from group settlement patterns and physical settings, including border disputes, ownership or occupancy of land, access to transportation outlets, control of religious or historical sites, and proximity to environmental hazards.
   - Pressures from skewed population distributions, such as a "youth or age bulge," or sharply divergent rates of population growth among competing communal groups.

2. Massive Movement of Refugees or Internally Displaced Persons

   - Forced uprooting of large communities as a result of random or targeted violence and/or repression, causing food shortages, disease, lack of clean water, land competition, and turmoil that can spiral into larger humanitarian and security problems, both within and between countries.

3. Legacy of Vengeance-Seeking Group Grievance or Group Paranoia

   - History of aggrieved communal groups citing injustices of the past, sometimes going back centuries.
   - Pattern of atrocities committed with impunity against communal groups.
   - Specific groups singled out by state authorities, or by dominant groups, for persecution or repression.
   - Institutionalized political exclusion.
   - Public scapegoating of groups believed to have acquired wealth, status, or power as evidenced in the emergence of "hate" radio, pamphleteering, and stereotypical or nationalistic political rhetoric.

4. *Chronic and Sustained Human Flight*—"Brain drain" of professionals, intellectuals, and political dissidents fearing persecution or repression.

   - Voluntary emigration of "the middle class," particularly economically productive segments of the population, such as entrepreneurs, businesspeople, artisans, and traders, due to economic deterioration.
   - Growth of exile communities.

Economic Indicators

5. Uneven Economic Development Along Group Lines

    - Group-based inequality, or perceived inequality, in education and economic status.
    - Group-based impoverishment as measured by poverty levels, infant mortality rates, educational levels, etc.
    - Rise of communal nationalism based on real or perceived group inequalities.

6. Sharp and/or Severe Economic Decline

    - A pattern of progressive economic decline of the society as a whole as measured by per capita income, GNP, debt, child mortality rates, poverty levels, business failures, etc.
    - Sudden drop in commodity prices, trade revenue, or foreign investment.
    - Collapse or devaluation of the national currency.
    - Extreme social hardship imposed by economic austerity programs.
    - Growth of hidden economies, including the drug trade, smuggling, and capital flight.
    - Increase in levels of corruption and illicit transactions among the general populace.

Political/Military Indicators

7. Criminalization and/or Delegitimization of the State

    - Massive and endemic corruption or profiteering by ruling elites.
    - Resistance of ruling elites to transparency, accountability, and political representation.
    - Widespread loss of popular confidence in state institutions and processes, e.g., widely boycotted or contested elections, mass public demonstrations, sustained civil disobedience, inability of the state to collect taxes, resistance to military conscription, rise of armed insurgencies.
    - Growth of crime syndicates linked to ruling elites.

8. Progressive Deterioration of Public Services

    - Disappearance of basic state functions that serve the people, including failure to protect citizens from terrorism and violence and to provide essential services, such as health, education, sanitation, public transportation, etc.
    - State apparatus narrows to those agencies that serve the ruling elites, such as security agencies, presidential staff, the central bank, the diplomatic service, and customs and collection agencies.

9. *Suspension of the Rule of Law and Widespread Violation of Human Rights*—Emergence of authoritarian, dictatorial, or military rule in which constitutional and democratic institutions and processes are suspended or manipulated.

    - Outbreak of politically inspired (as opposed to criminal) violence against innocent civilians.—Rising number of political prisoners or dissidents who are denied due process consistent with international norms and practices.

- Widespread abuse of legal, political, and social rights, including those of individuals, groups, and institutions (e.g., harassment of the press, politicization of the judiciary, internal use of military for political ends, public repression of political opponents).

10. Security Apparatus Operates as a "State within a State"

- Emergence of elite or praetorian guards that operate with impunity.
- Emergence of state-sponsored or state-supported "private militias" that terrorize political opponents, suspected "enemies," or civilians seen to be sympathetic to the opposition.
- Emergence of an "army within an army" that serves the interests of the dominant military or political clique.

11. Rise of Factionalized Elites

- Fragmentation of ruling elites and state institutions along ethnic, class, clan, racial, or religious lines.
- Use of nationalistic political rhetoric by ruling elites, often in terms of communal irredentism (e.g., a "greater Serbia") or of communal solidarity (e.g., ethnic "cleansing" or defending "the faith").

12. Intervention of Other States or External Political Actors

- Military or paramilitary engagement in the internal affairs of the state at risk by outside armies, states, identity groups, or entities that affect the internal balance of power or resolution of the conflict.

Appendix 2: Original *FSI* Data (FSI 2011; reproduced with permission from The Fund for Peace)

ID	State	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12	FSI
1	Somalia	9.7	10	9.5	8.2	8.4	9.3	9.8	9.4	9.7	10	9.8	9.7	113.5
2	Chad	9.2	9.5	9.4	8	8.9	8.5	9.8	9.6	9.3	9.2	9.8	9.1	110.3
3	Sudan	8.5	9.6	9.9	8.2	9.1	6.4	9.4	9	9.7	9.6	9.9	9.5	108.8
4	Congo (D. R.)	9.7	9.6	8.3	7.7	9.2	8.7	9	8.9	9.2	9.6	8.8	9.5	108.2
5	Haiti	10	9.2	7.3	8.9	8.8	9.2	9.4	10	8	8.4	8.8	10	108
6	Zimbabwe	9.3	8.2	9	9.3	9.2	9	9.3	9	9.2	9	9.6	7.8	107.9
7	Afghanistan	9.1	9.3	9.3	7.2	8.4	8	9.7	8.5	8.8	9.8	9.4	10	107.5
8	Central African Rep.	8.9	9.6	8.6	5.8	8.9	8.1	9.1	9	8.6	9.7	9.1	9.6	105
9	Iraq	8.3	9	9	8.9	9	7	8.7	8	8.6	9.5	9.6	9.3	104.9
10	Cote d'Ivoire	8.1	8.5	8.7	7.9	8	7.7	9.5	8.4	8.6	8.6	9.1	9.7	102.8
11	Guinea	8.2	7.7	7.9	8.3	8.4	8.6	9.4	8.7	9.2	9.3	9.2	7.6	102.5
12	Pakistan	8.8	9.2	9.3	7.5	8.5	6.6	8.6	7.3	8.7	9.4	9.1	9.3	102.3
13	Yemen	8.7	8.4	8.6	6.9	8.3	7.7	8.6	8.7	7.7	9.3	9.3	8.2	100.4
14	Nigeria	8.3	6	9.6	7.7	9	7.3	9	9	8.6	9.1	9.5	6.9	100
15	Niger	9.8	6.6	7.8	6.2	7.9	8.9	8.9	9.5	8.2	8	8.6	8.7	99.1
16	Kenya	8.8	8.5	8.7	7.6	8.5	7	8.9	7.8	7.7	7.9	8.8	8.5	98.7
17	Burundi	9.1	8.7	8.2	6.2	8.1	8.5	8.2	8.8	8	7.7	8.2	9	98.7
18	Myanmar	8.2	8	8.7	6	9	7.9	9.7	8.3	9	8.5	8.3	6.7	98.3
19	Guinea Bissau	8.7	7.2	5.4	7.4	8.1	8.7	9.2	8.4	7.8	9.3	9.2	8.8	98.2
20	Ethiopia	9.1	8.2	8.4	7.2	8.2	7.7	7.5	8.4	8.5	7.9	9	8.1	98.2
21	Uganda	8.8	8	8	6.6	8.4	7.5	7.7	8.3	7.5	8.6	8.6	8.2	96.2
22	North Korea	8.2	5.3	6.9	4.7	8.5	9.2	9.9	9.3	9.5	8.1	7.4	8.6	95.6
23	Timor-Leste	8.5	8	7.1	5.8	7.3	7.9	8.8	8.7	6.8	8.3	8.3	9.3	94.8
24	Cameroon	8	7.3	7.8	7.8	8.4	7	8.8	8.3	8.1	7.8	8.5	6.8	94.6
25	Bangladesh	8.3	6.5	9.2	8.1	8.4	7.7	8	8	7.1	7.9	8.9	6.2	94.3
26	Liberia	8.3	8.6	6.8	7	8	8.4	7	8.8	6.3	7.3	8.1	9.3	93.9
27	Nepal	7.8	7.4	9	5.9	8.7	7.9	7.9	7.7	8.5	7.8	8	7.1	93.7
28	Eritrea	8.3	6.8	6.1	7.4	6.5	8.3	8.5	8.4	8.9	7.7	8.1	8.5	93.5
29	Sri Lanka	7	8.6	9.4	6.9	8.4	5.3	8.5	6.1	8.6	8	9.5	6.8	93.1
30	Sierra Leone	8.9	7.5	6.5	8	8.5	8	7.7	8.8	6.7	6	7.9	7.6	92.1
31	Kyrgyzstan	7.6	6.5	8.3	7	7.6	7.6	9	6	8	8	8.3	7.9	91.8
32	Congo (Republic)	8.5	7.7	6	6.7	8.2	7.3	8.9	8.3	7.5	7.3	6.7	8.2	91.3
33	Malawi	9.1	6.5	6	8.1	8	8.8	7.9	8.2	7	5.2	7.6	8.7	91.1
34	Rwanda	8.9	7.3	8.2	6.8	7.4	7	7.1	7.8	8.2	5.8	8.4	8	90.9
35	Iran	6.1	7.9	8.5	6.7	7	5.4	9.1	5.6	9	8.6	9.2	7	90.1
36	Togo	8.1	6.5	5.4	7	7.9	8	8	8.5	7.7	7.3	7.8	7.1	89.3
37	Burkina Faso	8.9	6.2	5.5	6.3	8.5	8	7.7	8.7	6.4	7	7.3	8	88.5
38	Cambodia	7.7	5.6	7.2	7.6	6.8	7.2	8.5	8.4	8	6.2	8	7.4	88.6
39	Tajikistan	7.7	5.9	7.2	6	6.8	7.4	8.9	6.9	8.5	7.4	8.6	7	88.3
40	Uzbekistan	7.3	5.7	7.4	6.3	8.2	6.8	8.4	6	9	8.5	8.7	6	88.3
41	Equatorial Guinea	8.5	2.7	6.6	7.2	9.1	4.5	9.6	8.1	9.4	8.1	8.2	6	88
42	Mauritania	8.2	6.8	7.8	5.5	6.5	7.3	7.3	7.9	7	7.9	7.9	7.9	88

(continued)

ID	State	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12	FSI
43	Lebanon	6.5	8.5	8.7	6.6	6.8	5.7	7	5.8	6.6	8.7	8.8	8	87.7
44	Colombia	6.7	8.7	7.5	7.9	8.6	4.1	7.5	5.6	7.2	7.5	8	7.7	87
45	Egypt	7.1	6.4	8.3	5.7	7.4	6.5	8.6	5.9	8.3	6.8	8	7.8	86.8
46	Laos	7.6	5.8	6.5	6.8	5.7	7.2	8	7.7	8.5	7.1	8.6	7.2	86.7
47	Georgia	5.8	7.5	8	5.5	6.9	6	8.4	6	6.9	7.9	9	8.5	86.4
48	Syria	5.6	8.5	8.7	6.3	7.4	5.8	8.3	5.8	8.6	7.5	7.9	5.5	85.9
49	Solomon Islands	7.9	4.5	6.8	5.1	8	7.6	7.9	8.1	6.5	6.7	8	8.8	85.9
50	Bhutan	6.6	6.9	7.8	6.8	8.2	6.9	6.6	6.9	7.6	6.2	7.5	7	85
51	Philippines	7.3	6.5	7.2	6.7	7.1	5.6	8.3	6.1	7.3	8.3	8.5	6.1	85
52	Angola	8.6	6.6	6.2	5.9	8.8	4.5	8.5	8.2	7.5	6.2	7	6.7	84.7
53	Israel/West Bank	6.8	7.6	9.6	3.8	7.8	4.3	7.3	6.5	7.9	7	8.1	7.8	84.5
54	Papua New Guinea	7.4	4.5	6.9	7.4	9.1	6.4	7.5	8.7	6.3	6.6	7.1	6.4	84.3
55	Zambia	8.9	7.6	5.7	6.8	7.3	7.7	7.6	7.8	6.1	5.3	5.8	7.3	83.9
56	Comoros	7.5	4	5.3	6.6	5.8	7.6	8	8.2	6.6	7.5	8	8.7	83.8
57	Mozambique	9	4	4.6	7.7	7.4	8.2	7.6	8.6	7	7.1	5.6	6.7	83.5
58	Madagascar	8.3	4.6	5.2	4.9	7.8	7.6	7.1	8.6	6	6.8	8	8.3	83.2
59	Bolivia	7.2	4.6	7.7	6.4	8.9	6.5	6.8	7.1	6.3	6.5	8	6.9	82.9
60	Dijbouti	7.8	7.2	6.2	5.2	6.8	6	7.2	7.2	7	6.2	7.5	8.3	82.6
61	Swaziland	9.2	4.6	3.9	5.9	6.5	7.8	8.5	7.5	8.2	6.6	7	6.9	82.6
62	Ecuador	5.9	6.4	6.9	7.1	7.7	6.3	7.5	7.2	5.7	7	8.2	6.3	82.2
63	Azerbaijan	5.8	7.9	7.5	5.4	6.9	5.5	7.7	5.7	7.2	7	7.8	7.5	81.9
64	Indonesia	7.4	6.6	6.6	6.9	7.5	6.4	6.7	6.5	6.3	7.1	7	6.5	81.5
65	Tanzania	8.1	7.4	6.1	5.8	6.3	7.4	6.5	8.6	6.2	5.5	6	7.4	81.3
66	Moldova	6.1	4.4	6.6	7.5	6.5	6.7	7.6	6.3	6.5	7.8	8	7.2	81.2
67	Nicaragua	6.9	4.9	6	7.2	8.2	7.3	7.3	7.3	6	6.2	6.8	7.1	81.2
68	Fiji	5.9	3.9	7.6	6.9	7.7	7	8.6	5.5	6.5	7	7.9	6.6	81.1
69	Gambia	7.9	6.4	4	6.5	6.6	7.1	7.5	7	7.5	6.1	6.8	7.5	80.9
70	Bosnia and Herzegovina	5	6.8	8.4	5.9	6.8	5.2	7.6	5	6.1	7	9.2	8	81
71	Lesotho	9	4.6	5	6.8	6.1	8.1	6.9	8.2	6	5.5	7	7.2	80.4
72	China	8.2	6.2	7.9	5.6	8.6	4.4	7.9	6.6	8.8	5.7	6.9	3.3	80.1
73	Guatemala	7.3	5.6	6.9	6.5	7.7	6.5	6.8	6.9	6.9	7.6	6	5.3	80
74	Benin	8.1	7.1	3.9	6.6	7.2	7.9	6.7	8.5	5.7	6	5	7.3	80
75	Turkmenistan	6.5	4.2	6.6	5.1	7.1	6	8.4	6.7	8.7	7.5	7.7	5.2	79.7
76	India	8	5	8.2	6.2	8.5	5.4	5.8	7.2	5.9	7.8	6.8	4.5	79.3
77	Mali	8.8	5.3	6	7.3	6.7	7.8	5.5	8.2	4.9	7.1	4.5	7.2	79.3
78	Honduras	7.6	3.9	5.3	6.6	8.1	7	7.3	6.6	6.3	6.5	6.3	6.9	78.4
79	Thailand	6.4	6.6	8	4.4	7.2	4	8.4	5	7.3	7.6	8.5	4.9	78.3
80	Venezuela	6	4.8	7	6.4	7.3	6.1	7.5	5.8	7.4	7	7.3	5.5	78.1
81	Algeria	6.4	6.1	7.8	5.7	6.8	5.2	7.1	6.1	7.5	7.2	6.8	5.3	78
82	Russia	6.3	5.1	7.6	5.7	7.6	4.6	7.8	5.3	8.1	7.2	7.8	4.6	77.7
83	Belarus	6.3	3.6	6.8	4.5	6.3	6.2	8.8	5.8	8	6.3	8	7	77.6
84	Dominican Republic	6.5	5.5	6.1	7.9	7.5	5.6	5.8	6.8	6.3	5.8	6.8	6.2	76.8
85	Senegal	7.6	6.4	6.3	6	7.2	6.5	5.9	7.8	6.2	6.3	4.5	6.1	76.8
86	Cuba	6.3	5.4	5.1	6.9	6.3	6	6.6	5.3	7.4	6.9	6.9	7.5	76.6

(continued)

ID	State	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12	FSI
87	Morocco	6.4	6.5	6.4	6.4	7.5	6	6.9	6.6	6.4	5.9	6.3	4.9	76.2
88	Vietnam	6.7	5	5.7	5.7	6.2	6.1	7.5	6.4	7.7	6	6.9	6.1	76
89	El Salvador	7.6	5.3	5.8	7.1	7.6	6.3	6.5	6.9	6.7	7	4.3	4.9	76
90	Cape Verde	7.3	4.3	4.2	8.3	6.3	6.3	6.9	6.9	5.7	5.7	5.7	8.2	75.8
91	Maldives	6	5.9	4.9	6.8	5	6.7	7.4	6.9	7	5.7	7.6	5.8	75.7
92	Gabon	6.8	6.2	3.3	6.1	7.9	5.5	7.5	6.7	6.7	5.7	7.1	5.8	75.3
93	Saudi Arabia	6	5.8	7.5	3.2	7	3.4	7.9	4.2	8.9	7.5	7.9	5.9	75.2
94	Mexico	6.5	4.2	6.1	6.5	7.7	6	6.6	5.8	5.9	7.9	5.2	6.7	75.1
95	Turkey	5.9	6	8.3	4.5	7.4	5.5	5.9	5.7	5.2	7.4	7.5	5.6	74.9
96	Jordan	6.4	7.6	6.7	4.7	6.9	5.8	5.7	4.9	6.8	6	6.3	6.8	74.6
97	Sao Tome	7.1	4.3	4.8	7.3	6.2	6.9	6.9	7	4.9	5.8	6.3	6.9	74.4
98	Serbia	5.3	6.4	7.5	5	6.5	5.7	6.5	4.9	5.3	6.5	8	6.8	74.4
99	Peru	6.1	4.1	6.8	6.7	8	5.1	6.6	6.1	5.2	7.2	6.6	5.1	73.6
100	Guyana	6.4	3.6	5.9	8.4	7.4	6.4	6.5	5.5	5	6.3	5.1	6	72.5
101	Paraguay	5.9	1.9	6.5	5.5	8.3	5.9	7.9	5.5	6.4	6.4	7.7	4.5	72.4
102	Armenia	5.5	6.6	6	6.6	6.2	5.3	6.6	5	6.5	5.2	7	5.8	72.3
103	Micronesia	7.1	3.5	4.2	8	7.2	6.7	6.3	6.9	2.5	5.4	5.6	8.5	71.9
104	Namibia	7.2	5.6	5.3	7.1	8.5	6.3	4.4	6.7	5.5	5.5	3.5	6.2	71.8
105	Suriname	6	3.5	6.1	7	7.5	6.1	6.1	4.9	5.6	5.8	5.8	6.7	71.1
106	Macedonia	4.5	4.6	7.4	6.7	6.8	6.2	6.7	4.2	5	6	6.7	6.2	71
107	Kazakhstan	5.5	3.8	6	3.8	5.9	6.2	7.2	5.1	6.9	6.2	7.7	5.9	70.2
108	Tunisia	5.5	3.4	5.6	5.2	6.6	5	7.2	5.3	7.7	7	6.8	4.8	70.1
109	Samoa	7	2.7	4.8	8.3	6.6	5.9	6.2	4.7	4.2	5.5	5.1	8.6	69.6
110	Ukraine	5.3	3.1	6.5	6.3	5.9	6	7.4	4.1	5.5	4	8	6.8	68.9
111	Libya	5.5	4.6	6	3.9	6.9	4.6	7.3	4.3	8.3	5.9	7	4.4	68.7
112	Malaysia	6	4.8	6.7	4.2	6.7	4.9	6	5.1	6.9	6	6.4	5	68.7
113	Botswana	8.9	6.4	4.5	5.6	7.4	6.3	5	6	5	4.1	3.3	5.4	67.9
114	Belize	6.7	5.4	4.4	7	6.8	5.7	6	5.8	3.8	5.5	4.3	6.3	67.7
115	Ghana	6.8	5.5	5.5	7.6	6.3	6.1	4.8	7.7	4.5	3	4.2	5.6	67.6
116	Cyprus	4.4	4.4	7.6	5.3	7.3	5	5	3.3	3.3	5.3	7.9	8.8	67.6
117	South Africa	8.4	6.7	5.9	4.1	8.2	5.3	5.5	5.5	4.6	4.5	5.9	3	67.6
118	Jamaica	6.2	3.4	4.3	6.7	6.2	6.3	6.5	5.9	5.3	6.3	3.7	6.3	67.1
119	Seychelles	5.8	3.9	4.8	4.9	6.6	5.4	6.8	4.1	5.8	6.1	5.7	7.1	67
120	Grenada	5.8	3.2	3.9	8	6.5	5.7	6.2	4.2	4.3	5.3	5.6	7.7	66.4
121	Albania	5.5	3.1	5.1	6.8	5.4	5.9	6.4	5	5	5.4	6.3	6.3	66.2
122	Brunei	5.1	3.9	6.2	4.1	7.8	3.4	7.7	3.2	6.7	5.6	7.4	4.7	65.8
123	Brazil	6.1	3.5	6.5	4.5	8.5	3.9	5.9	5.8	5.1	6.5	4.9	3.9	65.1
124	Trinidad	5.3	3.2	4.7	7.7	6.9	4.5	5.5	4.9	5.1	5.5	5.6	4.8	63.7
125	Antigua and Barbuda	5.2	3	4.1	7.6	5.9	5.1	5.8	4.3	4.5	4.9	3.7	5.8	59.9
126	Romania	5.1	3.2	6	5	5.8	5.8	5.9	4.5	4	4.1	5.2	5.2	59.8
127	Mongolia	5.5	1.6	4	1.9	6.2	5.3	5.9	5.6	6	5	5.5	7.1	59.6
128	Kuwait	5.1	3.8	4.9	4.3	5.9	4	5.7	2.9	6.2	4.5	7.2	5	59.5
129	Bahrain	4.5	2.9	6.8	3.1	6	3.4	6.9	2.7	5.9	4.8	6.6	5.3	58.9
130	Bulgaria	4.1	3.6	4.3	5.5	5.7	5.3	5.9	4.6	4.3	4.9	5.3	5.5	59
131	Panama	6	3.9	4.6	4.9	7.4	4.9	4.6	5.2	4.5	5.7	2.5	3.6	57.8
132	Croatia	4.3	5.5	5.5	4.9	5	5.9	4.4	3.4	4.3	4.4	4.7	5	57.3

(continued)

ID	State	*d*1	*d*2	*d*3	*d*4	*d*5	*d*6	*d*7	*d*8	*d*9	*d*10	*d*11	*d*12	FSI
133	Bahamas	5.8	2.8	4.4	6.2	6.2	4.8	5.2	4.2	3.2	4.3	4.5	4.9	56.5
134	Montenegro	4.5	4.5	6.4	2.4	4.1	5.2	4.3	3.6	5	4.8	6.2	5.3	56.3
135	Lativa	4.2	3.9	4.9	4.8	5.7	5.8	5.3	3.9	3.6	3.3	4.3	4.4	54.1
136	Barbados	4.3	2.9	4.4	6.8	6.3	5	3.9	2.9	2.5	4.2	4.2	5.4	52.8
137	Costa Rica	5.1	4.3	4.1	4.1	6.5	4.9	3.5	4.2	3	2.5	3.5	4.9	50.6
138	United Arab Emirates	4.1	2.8	4.6	3	5.4	4.2	6.5	3.3	5.7	3	3.6	4.1	50.3
139	Qatar	4.2	2.7	4.9	3.1	5	3.7	6	2.3	5	3	5	4.6	49.5
140	Estonia	4.1	3.9	5.4	4.5	4.9	4.3	4.1	2.9	3	2.9	5.5	3.9	49.4
141	Oman	5.1	1.5	3	1.5	3	3.8	5.9	4.4	6.9	5.3	6.3	2.4	49.1
142	Hungary	3.1	3.1	3.5	4.5	5.5	5.4	5.4	3.7	3	2.5	4.7	4.3	48.7
143	Greece	4.1	2.6	4.5	4.4	4.3	5.1	4.9	3.8	3.1	3.8	2.5	4.3	47.4
144	Slovakia	3.8	2.3	5	5.1	5.2	4.6	3.9	3.6	3.6	2.3	3.7	3.9	47
145	Argentina	4.4	2.6	4.9	3.5	6	4.4	4	3.5	4	2.7	3	3.8	46.8
146	Poland	4.3	3.5	3.5	5.6	4.7	4.3	4.2	3.3	3.5	2.5	3.6	3.9	46.9
147	Italy	3.6	3.5	5.3	3.2	4.1	4.2	4.7	2.8	3.1	4.9	4.4	2	45.8
148	Malta	3.4	5.4	4	4.4	4.1	4.1	3.7	2.9	3.4	3.7	2	4.4	45.5
149	Lithuania	4.1	3.2	3.7	4.6	5.7	5.3	3.6	2.9	3.1	2.5	2.8	3.8	45.3
150	Mauritius	3.3	1.6	3.5	3	5.4	4.5	4.7	3.9	3.5	3.6	3.2	4	44.2
151	Spain	3.3	2.9	6	1.9	4.7	4.5	2.1	2.4	2.6	4.9	5.6	2.2	43.1
152	Czech Republic	3	2.8	3.8	4	3.8	4.6	3.7	3.9	3	2.1	3.8	3.8	42.3
153	Chile	5	3	3.5	2.8	5	4.6	2.1	4.3	3.3	2.5	1.4	3.3	40.8
154	Uruguay	3.9	1.7	2.4	5.3	4.7	3.8	2.5	3.3	2.5	3.7	2.7	3.9	40.4
155	South Korea	3.3	3	3.7	4.5	2.3	2.2	3.7	2.2	2.6	1.7	3.6	6	38.8
156	Slovenia	3.1	1.7	3.1	3.6	4.7	3.7	3	2.8	2.8	3	1.1	2.9	35.5
157	Singapore	2.5	0.9	3	2.8	3.4	3.6	3.9	2	4.7	1.5	4	2.8	35.1
158	United States	3.4	2.9	3.6	1.1	5.4	3.7	2.2	2.7	3.3	1.6	3.6	1.3	34.8
159	United Kingdom	2.9	3.3	4.4	2.1	4.2	3.3	1.4	2.2	2	2.7	3.6	1.9	34
160	Belgium	2.5	2.1	4.4	1.6	4.4	3.6	2.7	2.5	1.6	2	4	2.6	34
161	France	3.3	2.8	5.9	1.8	4.9	3.5	1.6	1.9	2.5	1.9	1.9	2	34
162	Germany	2.9	4.2	4.7	2.6	4.4	2.9	1.9	2	2	2.2	2.1	2	33.9
163	Portugal	3.3	2	2.5	2.5	3.6	4.8	1.6	3.3	3.3	1.6	1.4	2.5	32.4
164	Japan	3.6	1.1	3.9	1.8	2.3	3.5	2	1.7	3	2	2.6	3.5	31
165	Iceland	1.6	1.5	1	3.3	2.2	6.2	2	1.9	1.6	1	1.8	6	30.1
166	Netherlands	3	3	4.4	2.2	2.9	3.2	1.1	1.7	1	1.4	2.4	2.1	28.4
167	Australia	3.3	2.8	3.6	1.6	3.9	2.9	1.6	1.8	1.9	1.7	1.6	1.4	28.1
168	Canada	2.9	2.5	3.3	2.4	4.1	2.4	1.2	1.9	1.6	1.5	2.5	1.4	27.7
169	Austria	2.6	2.6	3.8	1.6	4.4	2.3	1.2	1.6	1.5	1.1	2.4	2.2	27.3
170	Luxembourg	1.7	2.1	2.8	1.5	2	2.3	2.5	1.9	1	2.3	3.4	2.6	26.1
171	Ireland	2.3	2	1.3	2.4	2.6	3.9	2	2.2	1.2	1.6	1.4	2.4	25.3
172	New Zealand	2	1.7	3.5	2.4	4	3.8	1.1	1.9	1.2	1.1	1.1	1.1	24.9
173	Denmark	2.9	2.1	3.3	2.1	1.7	2.5	1.2	1.6	1.3	1.5	1	2.6	23.8
174	Switzerland	2.1	1.9	3.5	2.1	2.8	2.4	1	1.6	2	1.4	1	1.4	23.2
175	Sweden	2.8	2.9	1.3	2	2.2	1.9	0.9	1.5	1.6	2.3	1.8	1.6	22.8
176	Norway	2	2	1.3	1.5	2.1	2.9	1	1.4	1.9	1.2	1.2	1.9	20.4
177	Finland	2	2.1	1.7	2.5	1.3	2.8	1	1.5	1.1	1	1.2	1.5	19.7

References

Al-Sharrah G (2010) Ranking using the copeland score: a comparison with the Hasse diagram. J Chem Inf Model 50:785–791

Annoni P, Brüggemann R (2008) The dualistic approach of FCA: a further insight into Ontario Lake sediments. Chemosphere 70:2025–2031

Annoni P, Brüggemann R, Saltelli A (2011) Partial order investigation of multiple indicator systems using variance-based sensitivity analysis. Environ Model Softw 26:950–958

Baker PH (2006) Conflict Assessment System Tool (CAST). An analytical model for early warning and risk assessment of weak and failing states. Fund for Peace, Washington, DC

Bick A, Brüggemann R, Oron G (2011) Assessment of the intake and the pretreatment design in existing seawater reverse osmosis (SWRO) plants by Hasse diagram technique. Clean 39:933–940

Brüggemann R (2011) Special Issue: Partially ordered sets (editorial). Stat Appl Spl Iss:3–6

Brüggemann R, Carlsen L (2006) Partial order in environmental sciences and chemistry. Springer, Berlin

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partially orders. Match Commun Math Comput Chem 65:383–414

Brüggemann R, Carlsen L (2012) Multicriteria decision analyses. Viewing MCDA in terms of both process and aggregation methods: some thoughts, motivated by the paper of Huang, Keisler and Linkov. Sci Total Environ 425:293–295

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems - introduction to partial order applications. Springer, New York

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Brüggemann R, Voigt K (2011) A new tool to analyze partially ordered sets. application: ranking of polychlorinated biphenyls and alkanes/alkenes in river Main, Germany. Match Commun Math Comput Chem 66:231–251

Brüggemann R, Voigt K (2012) Antichains in partial order, example: pollution in a German region by lead, cadmium, zinc and sulfur in the herb layer. Match Commun Math Comput Chem 67:731–744

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (2001) Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. J Chem Inf Comp Sci 41:918–925

Brüggemann R, Sørensen PB, Lerche D, Carlsen L (2004) Estimation of averaged ranks by a local partial order model. J Chem Inf Comp Sci 44:618–625

Brüggemann R, Restrepo G, Voigt K, Annoni P (2013) Weighting intervals and ranking, exemplified by leaching potential of pesticides. Match Commun Math Comput Chem 69(2):413–432

Bubley R, Dyer M (1999) Faster random generation of linear extensions. Discrete Math 201:81–88

Carlsen L, Brüggemann R (2012) The 'failed state index' offers more than just a simple ranking. Soc Indic Res. doi:10.1007/s11205-012-9999-6

Carlsen L, Brüggemann R (2013) An elaborate analysis of the 'failed states index' by partial order methodology JoSS -J.Soc.Struct. 14:(in press)

De Loof K, De Meyer H, De Baets B (2006) Exploiting the lattice of ideals representation of a poset. Fundam Inform 71:309–321

Duchowicz PR, Castro EA, Fernandez FM (2008) Application of a novel ranking approach in QSPR- QSAR. J Math Chem 43:620–636

FFP (2006) Methodology behind the index, http://ffp.statesindex.org/methodology. Accessed Aug 2013

FFP (2011a) The fund for peace. http://www.fundforpeace.org/global/. Accessed May 2012

FFP (2011b) Conflict assessment indicators. The fund for peace state analysis indicators and their measures. The Fund for Peace Publication CR-10-97-CA (11-05C). http://www.fundforpeace.org/global/library/cr-10-97-ca-conflictassessmentindicators-1105c.pdf. Accessed May 2012

Fishburn PC (1970) Utility theory for decision making. Wiley, New York

Freier KP, Brüggemann R, Scheffran J, Finckh M, Schneider UA (2011) Assessing the predictability of future livelihood strategies of pastoralists in semi-arid Morocco under climate change. Technol Forecast Soc Change 79:371–382

FSI (2011) Failed states index. http://www.foreignpolicy.com/failedstates. Accessed May 2012

Kardaetz S, Strube T, Brüggemann R, Nützmann G (2008) Ecological scenarios analyzed and evaluated by a shallow lake model. J Environ Manag 88:120–135

Munda G (2008) Social multi-criteria evaluation for a sustainable economy. Springer, Berlin

Newlin J, Patil GP (2010) Application of partial order to stream channel assessment at bridge infrastructure for mitigation management. Environ Ecol Stat 17:437–454

Restrepo G, Weckert M, Brüggemann R, Gerstmann S, Frank H (2008) Ranking of refrigerants. Environ Sci Technol 42:2925–2930

Schneeweiss C (1991) Planung 1 - Systemanalytische und entscheidungstheoretische Grundlagen. Springer, Berlin

Talente (2007) DART – Decision Analysis by Ranking Techniques (ver. 2.05), http://www.talete.mi.it/products/dart_description.htm. Accessed May 2012

Transparency (2011) Corruption perception index 2011. Transparency International, Berlin

Tsakovski S, Simeonov V (2011) Hasse diagram technique as exploratory tool in sediment pollution assessment. J Chemometrics. doi:10.1002/cem.1381:1-8

Voigt K, Brüggemann R (2008) Ranking of pharmaceuticals detected in the environment: aggregation and weighting procedures. Comb Chem High Throughput Screen 11:770–782

Wienand O (2006) lcell. http://bio.math.berkeley.edu/ranktests/lcell/. Accessed May 2012

Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybern 18:183–190

Yager RR (1993) Families of OWA operators. Fuzzy Sets Syst 59:125–148

Chapter 19
PyHasse Software for Partial Order Analysis: Scientific Background and Description of Selected Modules

Rainer Brüggemann, Lars Carlsen, Kristina Voigt, and Ralf Wieland

Abstract The software PyHasse is an elaborated "experimental" software for ordinal analysis of data matrices. PyHasse is based on the interpreter programming language Python. A brief introduction to the programming language Python is given and the general principles behind PyHasse are outlined. An actual overview about PyHasse (status, April 2013) is provided. Today PyHasse comprises 91 modules covering 9 different categories, such as basic Partial Order Analysis, i.e., the drawing Hasse diagrams and the calculation of some important quantities. A selection of newer or rarely used modules are discussed in detail in order to explain some principles of PyHasse. As a leading example the pollution by Lead, Cadmium, and Zinc of regions of south-western Germany is discussed.

An outlook is given, where future projects are discussed. Such projects comprise among others, Internet access to some of the more important modules, inclusion of the Formal Concept Analysis tools, and of tools derived from POSAC and the variance-based sensitivity.

R. Brüggemann (✉)
Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Mueggelseedamm 310, Berlin, Germany
e-mail: brg_home@web.de

L. Carlsen
Awareness Center, Linkøpingvej 35, Trekroner, DK-4000 Roskilde, Denmark

Center of Physical Chemical Methods of Research and Analysis, al-Farabi Kazakh, National University, 96A Tole Bi street, 050012 Almaty, Kazakhstan

K. Voigt
Institute of Computational Biology, Helmholtz Center, Munich, Neuherberg, Germany

R. Wieland
Leibniz Centre for Agricultural Landscape Research (ZALF), Institute of Landscape Systems Analysis, Muencheberg, Germany

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*,
DOI 10.1007/978-1-4614-8223-9_19, © Springer Science+Business Media New York 2014

19.1 Introduction

The analysis of multi-indicator systems (Brüggemann and Patil 2010, 2011), aiming at a ranking of multiple characterized objects is of increasing interest in many scientific fields, e.g., environmental health (Voigt et al. 2011, 2012) or sociology and politics (Annoni 2007; Carlsen and Brüggemann 2013a, b). In this context, partial order methodology appears to be increasingly applied (see Brüggemann and Carlsen 2012 in their response to Huang et al. 2011). The tools of partial order are not as ancient as those of general decision-making methods which started with the scientific work of Condorcet, Borda, at the end of the eighteenth century (cf. Munda 2008). Partial order as a mathematical discipline seems to go back to the late nineteenth century, where Dedekind was exploring the Diedergroups. Strong impacts on the theory of partially ordered sets can be related to Hasse (1927, 1952) and Birkhoff (1984), two mathematicians who, as Dedekind, were mainly interested in algebraic aspects. Within the context of data matrices, i.e., within a statistical point of view, main contributions can be traced back to Patil on the one side (within the context of biological diversity, see Patil and Taillie 1976), and, without knowing each other, to the team Halfon and Reggiani (Halfon and Reggiani 1986), on the other side. The work of Halfon and his coauthors gave the basis for the computerized Hasse diagram technique (HDT), which is specifically related to partial order and their application to the ranking of objects simultaneously described by several indicators, i.e., by data matrices. A third line of development of the analysis of data matrices can be identified, which is the field of Formal Concept Analysis (FCA), developed in the 1980s (Ganter and Wille 1996), which also finds increasing interest (see for instance Bartel and Brüggemann 1998; Davey 2004; Carlsen 2009; Brüggemann and Patil 2011).

The application of many of the tools of partial order theory on data matrices is a priori extremely simple, however, tedious if performed manually. Therefore, it is understandable that together with development of computer programming, an increasing and more and more detailed support in the ordinal analysis of data matrices is ongoing.

In this book some chapters describe the application of selected modules of the PyHasse package, whereas in Brüggemann and Patil (2011), a state-of-the-art overview (by 2010) of the software packages Rapid and PyHasse is given.

This chapter explains some background material on Python, the programming language on which PyHasse is based, and renders some more and general information about PyHasse.

19.2 HDT Software

An overview about software, a status by 2006, was given by Halfon (2006).

A complete overview about theory and applications of partial order on multi-indicator systems is outside the scope of this chapter, which instead aims at a description of PyHasse. For introductory texts we refer to papers by Brüggemann et al. (2001) and Brüggemann and Voigt (2008).

Table 19.1 Software aiming at partial order analysis of data matrices

Software	Authors	Remark	Reference
Hasse	Halfon	Drawing Hasse diagrams	Halfon et al. (1986)
WHASSE	Brüggemann	Drawing Hasse diagrams and first attempts to introduce tools beyond the drawing	Brüggemann et al. (1999)
conimp4	Burmeister	Analyzing data matrices on the basis of Formal Concepts	Burmeister, CONIMP4, Programm zur Formalen Begriffsanalyse (1997)
conexp	Yevtushenko	Analyzing data matrices on the basis of Formal Concepts	Yevtushenko (2003) http://www.comp.dit.ie/pbrowne/compfund2/UserGuide.pdf (assessed 7 Nov 2012) Download, see for instance: http://sourceforge.net/projects/conexp/files/conexp/1.3/ (accessed Aug 2013)
DART	Talente (Manganaro et al. 2008)	Drawing Hasse diagrams, utility functions	Manganaro et al. (2008)
ProRank	Pudenz	Drawing Hasse diagrams with emphasis on simple data management	Pudenz (2005)
Rapid	Joshi, Brüggemann and Patil	Drawing Hasse diagrams, some analysis tools	Brüggemann and Patil (2011)
PyHasse	Brüggemann and Patil	Analyzing partially ordered sets, derived from data matrices	Brüggemann and Patil (2011)
Parsec	Fattore	Analysis on the basis of R; see also Myers and Patil (2010, 2014)	Fattore and Arcagni, Chap. 16

In Table 19.1 a—certainly not complete—overview of software is given. The newest (so far the authors are aware), the PyHasse, will be explained in more detail in this chapter.

19.3 Python as Programming Language for Contemporary Software Generation

Clearly the first question often stated may be: Why Python, why not JAVA, PERL, or traditional languages such as C++, Fortran, or VisualBasic? The most honest answer is, simply because Python fulfills to a wide degree the personal taste of the programmer, in this case of Rainer Brüggemann. This very personal way to find a decision about the suitable programming language may be unsatisfactory for many readers. Hence, we discuss some objective points that in the view of the programmer favor Python (however, without arguing that other programming languages do not have these features).

19.3.1 General Remarks

In the present context the arguments are following those of Lutz and Ascher (2003) closely, although there is a book available, where specifically Python for scientific uses is explained and which is recommended for further reading (Langtangen 2009).

When Python was developed by Guido von Rossum (cf. Venners 2003) it was developed in one step. Hence, Python had a very homogenuous structure from the very beginning. Clearly, Python has been further developed and will be further developed in the future. Actually, currently Python is delivered in version 3, whereas PyHasse is developed on the basis of Python 2.6, available since around 2007.

Python is—in contrast to C languages—comfortably readable and coherent. Python supports consequently the object-oriented programming style.

It is of further interest that typically Python codes are "1/3 to 1/5 the size of equivalent C++ or Java code" (Lutz and Ascher, page 3, 2003).

Briefly speaking there is a Python slogan which says that "In the Python way of thinking, explicit is better than implicit, and simple is better than complex" (Lutz and Ascher 2003, page 5).

Python is an interpreter language. That means, there is no need to compile and link the software before it is applied. In the Web site python.org, we find "Python is a programming language that lets you work more quickly and integrate your systems more effectively." Clearly, it is to be expected that the linear reading of the programming code may be time consuming. However, the personal experience is that even the combinatorial algorithms, which are typical for the application field of partially ordered sets do not need much time, i.e., even the impatient programmer may await the result, sitting before his machine!

Technically spoken, Python belongs to the Very High-Level Languages (VHLL) (Müller and Schwarzer 2007). For the PyHasse author, the fact that developing new modules and testing them does not need to first compile parts of the program makes Python a very efficient and quick programming tool.

19.3.2 Portability

The portability of Python programs is high. For example, PyHasse programs run without problems on different Windows operating systems as well as on different UNIX or Linux machines. Although there is not much experience with Macintosh operating systems, examples are known that PyHasse can be ported to the Macintosh without major difficulties.

19.3.3 Libraries

As most other modern programming languages Python provides many freely downloadable libraries. All possible applications of “modern life” can be handled by such libraries:

- NumPy and Matplotlib: powerful libraries for numerical calculations and visualization which can replace MATLAB in many applications.
- Statistical libraries, drawing libraries are available, as well as libraries designed to handle databases (MySQL, Oracle, …).
- The Internet programming libraries are of increasing importance and many web frameworks provide support for a quick construction of Web sites for example Plone (with parts of ZOPE) or Django.

Also more exotic applications are supported like:

- PIL: (PhotoImageLibrary) a library for manipulating electronically photos
- PyGame: a library facilitating game programming

19.3.4 Programming Support

One important point is that Python supports the development of own written libraries, which are specifically designed for the scientific purpose of any software package.

Another important point is that Python supports the development process by a set of tools: For example, Cython expands Python adding type information to a Python program to make Python modules faster. Cython may further be used to include C/C++ Code in a Python library. Alternatively SIP or SWIG can wrap existing C/C++ code to use it as a Python module. There are modules available that include the QT library which allows to implement modern graphical user interfaces or another module which includes the gnu scientific library (gsl).

For some applications like Monte-Carlo Simulations, Python as interpreter language is too slow. Techniques, discussed above, may be used to accelerate Python modules. It should further be noted that Python supports parallel programming. Thus, modules like PyPar connects Python programs to the powerful Message Passing Interface (MPI) and allow parallel processing even on a personal computer with four or eight cores. This underlines that Python is not restricted to fast prototyping but Python is a modern programming toolbox, which can be used for challenging projects.

Some useful links are:
python: http://www.python.org/
numpy, scipy: http://numpy.scipy.org/
matplotlib: http://matplotlib.org/

networkX: http://networkx.lanl.gov/
Cython: http://cython.org/
SWIG: http://www.swig.org/
QT: http://qt.digia.com/
PyQT: http://www.riverbankcomputing.co.uk/software/pyqt/intro
gsl: http://www.gnu.org/software/gsl/
PyPy: http://pypy.org/
Pypar: http://code.google.com/p/pypar/

19.4 A Practical Ranking Problem, i.e., a Test Set for Explaining PyHasse

In order to have an illustrative example at hand we look at nine regions in Germany. In order to measure the air quality, the concentrations of deposited Pb, Cd, and Zn are monitored in epiphytic mosses (Brüggemann et al. 1998). The question arises: Can we rank the regions simultaneously taking into account the concentrations of all three metals? In Table 19.2 the data matrix is shown.

Partial order theory provides an answer as displayed as a Hasse diagram, i.e., a transitively reduced acyclic, trianglefree digraph of order relations (as explained in several chapters of this volume) (Fig. 19.1). The first observation is that the

Table 19.2 Data matrix, nine regions, three metals, concentrations in mg/kg dry weight (rounded)

Region	Pb (Lead)	Cd (Cadmium)	Zn (Zinc)
6	11	0.2	31
8	20	0.4	55
7	14	0.3	41
17	13	0.3	63
9	17	0.3	45
16	13	0.4	51
14	12	0.6	41
5	14	0.4	45
29	9	0.4	29

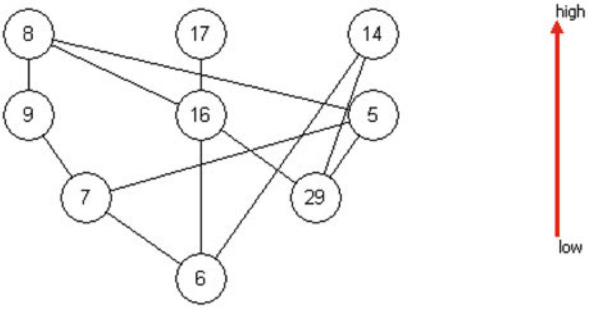


Fig. 19.1 Hasse diagram of nine regions with three attributes, namely the (rounded) metal concentrations in epiphytic mosses of Pb, Cd, and Zn

Hasse diagram is not slim, i.e., it obviously deviates remarkably from a linear order (Fig. 19.1).

Figure 19.1 shows the main characteristic of HDT: There are regions (generally "objects") which cannot be compared, as, e.g., region 8 and 17. The reason is that region 8, with respect to one metal, has a higher concentration than region 17, whereas region 17 on the other hand, has a higher concentration than region 8 with respect to another metal. Thus, the two regions are "in conflict with each other." Technically we describe an incomparability by the symbol ||, i.e., here as 8 || 17. A set of objects which are mutually incomparable is called an antichain, in contrast to a set of objects, which are mutually comparable, i.e., a chain. Hence, the set {9, 16, 5} is an antichain, whereas the set {6, 7, 9, 8} is a chain.

19.5 The PyHasse Software

19.5.1 Intention Behind the Software

19.5.1.1 Modules

PyHasse is a software consisting of a series of mutually independent programs. These programs are called "modules." When programming tools, as well as interfaces and all the partial order analysis tools are counted, the complete number of modules of PyHasse software is 91 (April 2013). However, this number is continuously changing, as new modules may replace a couple of older modules or new ideas to analyze partial orders derived from the ordinal analysis of data matrices eventually result in new modules.

In total, PyHasse is a software package with more than 50,000 lines of programming code (including comment lines, empty lines, which help to get a clear program code). Obviously, parts of program codes often appear several times, due to the intention that the modules should be mutually independent.

19.5.1.2 PyHasse as Experimental Software

PyHasse is intended to help solving daily problems applying partial order concepts on data matrices. It does not intend to provide either perfect statistical or graphical tools, especially it does not intend to include the vast number of applications, which more or less routinely are performed by applying spreadsheet software, such as Microsoft Excel®. The same kind of philosophy holds when a drawing of Hasse diagrams is considered. Thus, virtually all PyHasse modules offer the drawing of Hasse diagrams following the drawing convention, which has its origin in the work of Halfon (Halfon and Reggiani 1986). Nevertheless, these PyHasse-generated graphs are far from being perfect drawings. Hence in this context PyHasse cannot compete with the powerful freely downloadable program Graphviz, see Gansner and North (1999), which visualizes partially ordered sets in an almost perfect manner (see Sect. 19.5.4.3).

In sum PyHasse tries to fill the gap between highly specialized programs often developed in laboratories but not generally applicable and professionally written software, which usually may not reflect the state of the art of the theoretical development, even though updates are made available from time to time.

19.5.2 Basic Structure

19.5.2.1 Contextual Categories

PyHasse is structured in two ways: Contextually and from the programming point of view. In Table 19.3, nine contextual categories are explained.

In Fig. 19.2 a bar diagram displays the distribution of the PyHasse modules over the nine categories described in Table 19.3.

19.5.2.2 Programming Structure

The 91 modules are supported by four libraries (Table 19.4).

These four libraries are delivered together with the PyHasse modules (and some additional files) and the user has to put them into the folder, where Python is localized.

In order to facilitate the installation of PyHasse software, the programmer, Brüggemann, did not extensively use other comfortable libraries, such as MatplotLib or NumPy.

Together with the utility functions, the programming structure can be characterized by a scheme, as shown in Fig. 19.3.

19.5.2.3 Graphical User Interface

Most of the modules have similar graphical user interfaces (GUIs). In Python GUIs can be programmed, applying the standard library Tkinter, which is derived from Tcl/Tk. Thus, all user interfaces in the PyHasse package are built using Tkinter. The location of the typos: buttons (which govern the user activity) are vertically arranged following the most typical logical sequence of steps. A few modules are menu oriented, such as pyhassemenue8_3.py, DAHP.py, and modelHD9.py. In almost every user interface an "about" function is found, which informs briefly about the aim of the module and the programmer and (sometimes) about the leading idea out of the literature.

Table 19.3 Alphabetically sorted contextual categories of PyHasse, references are found in the Appendix

Group	Explanation	Modules	Described or applied in this book
AC	Antichain analysis in the broadest sense Why incomparabilities of objects appear in antichains, but also, why different sets of vertices in a Hasse diagram are not or only loosely connected In hdgt6.py some graph–theoretical concepts are programmed to explore distances in the Hasse graph	antag2.py antagscattplot2.py antichain20_4.py checksepset5.py findsomesepset4.py hdgt6.py recocgnizesepset3.py sepanal15_3.py sepanal16_coloured.py	Chap. 18 Chap. 18 Chap. 18
Comp-Ind	The intention is near to a concept having its origin in Patil’s intention of “comparative knowledge discovery” (Patil and Joshi 2014), namely to explore the role of weights when composite indicators are known or are to be constructed In owa4.py the fuzzy concept after Yager (Yager 1988, 1993) is applied CompInd is in pyhassemenue8 listed under MCDA	canonweight9_2.py canonweight_3D-grid2.py conflict7.py gevol2.py HDCI6.py linagg10.py owa4.py stability9.py weightbased modeling2.py	Chaps. 6 and 7 Chap. 2
Inter-face	Partial order may be expressed in different ways. When zeta or cover matrices are used, then some of the interface modules can read these matrices Different PyHasse modules can communicate with each other applying a special format. Files with this internal format have the extension, i.e., *.pdt Finally an interface to apply graphviz is available	covi_interactive3.py covreader4.py graphviz2.py txtpdt_interface1.py zetareader1.py	

(continued)

Table 19.3 (continued)

Group	Explanation	Modules	Described or applied in this book
LinExt	Beside the directed acyclic graph displaying partially ordered sets (Hasse diagram), partially ordered sets can be presented by a set of linear extensions	avrank5.py avrkmut3.py	Chap. 6 Chap. 6
	From linear extensions average ranks can be derived, as well as probabilities as prob($x > y$) or prob(x: rank=Rk) Winkler 1982; Brüggemann and Carlsen 2011; Wienand 2012	genlinext1.py linext_play2.py LPOMext4_2.py	
	This group contains modules, which are most crucial with respect to memory limitations, especially avrank5.py	LPOMstruct1.py mutprobavrk.py BubleyDyer8.py CRF1.py	
MCDA	PyHasse contains some "classical" MCDA methods such as PROMETHEE (Brans and Vincke 1985), the concordance–discordance analysis of ELECTRE (Peters and Zelewski 2007; Opperhuizen and Hutzinger 1982), AHP (Saaty 1994), as well as TOPSIS and DEA (http://mat.gsia.cmu.edu/classes/QUANT/NOTES/chap12.pdf). However, typically these methods are included in extremely simplified versions. Beyond this, there are several variants available, where an outranking algorithm, for example, the derivation of the Copeland index is applied (Al-Sharrah 2010, 2011)	condor2_2.py copel4.py dahp4 deaMC1.py Discordance3.py genoutrk7.py oreste6.py outrkHD6.py prom6_2.py topsissimpl1.py	
METEOR	Method of evaluation by order theory. The basic idea is to analyze the role of weightings of the single descriptors by a stepwise procedure. METEOR could also be seen as a part of CompInd	cap7.py HDweightMC2.py meteorHD4.py meteorparallel2.py	
Model	In the broadest sense, proximity (similarity) analysis as well as exploring partially ordered sets being most similar to a given one can be seen as a modeling technique. Furthermore, the influence of graph–theoretical structures in Hasse diagrams on average ranks is of interest and can be analyzed by "model posets"	combsimilarity7.py concord2_2.py modelHD9.py similarity10_1.py similarity_search7_2.py	Chap. 17

POT	Drawing Hasse diagrams in different configurations of inputs or of the graph, with different additional information, such as navigation tools or even a simple approach to estimate average ranks	mainHD20_5.py	Chap. 18
		mainHD20_5.py (without “psyco”)	Chaps. 11 and 18 Chap. 10
	A fuzzy concept of partial order after Kosko is included, as well as a MC simulation, fuzzydds7.py, after Wieland and Brüggemann (2013). A nonnumerical aggregation based on concepts developed by Carlsen (2008) is realized in hpor3.py. The module probranksThAC3.py follows an idea, published by Thiessen and Achari (2012)	**mHDCl2_7**.py	Chap. 2
		chain7_1.py	
		dds12.py	
		ExcelHD1.py	
		fuzzyHD13_2.py	
		fuzzydds7.py	
		graphvizHD1.py	
		HDalter6_1.py	
		HDedit3.py	
		HDscratch2.py	
		HDsimpl1.py	
		HD_msptree1.py	
		hpor3.py	
		interval8.py	
		levelheuristic2	
		majoriz1.py	
		palg4_2.py	
		pooc6.py	
		POTanalysis2.py	
		sensitivity18_3.py	
		SingleObjects_analysis4.py	
		pycluster1_2.py	
		orientation1.py	
		POcdn1.py	
		probranksThAC3.py	

(continued)

Table 19.3 (continued)

Group	Explanation	Modules	Described or applied in this book
Utility	This group of models is least relevant for the users. They are mainly helpful in documentation and programming. Some of those modules are internally called by pyhassemenue8.py. Another, pir3.py, can be used to get a complete tutorial text by means of all the help texts available for the PyHasse modules	discretiz3.py instruction1.py pir3.py pyhasseinfo1.py pyhassemenue8_4.py pyHasse_progr1.py randomdm2.py tool_libraryreader1.py utility_libraryreader1.py utility_PyHassecount1.py utility_PyHasseInfo1.py utility_PyHasse_newsince1.py (abbr. UPN) utility_PyHasse_interval1.py	

Modules described in this chapter are written in italic and bold characters

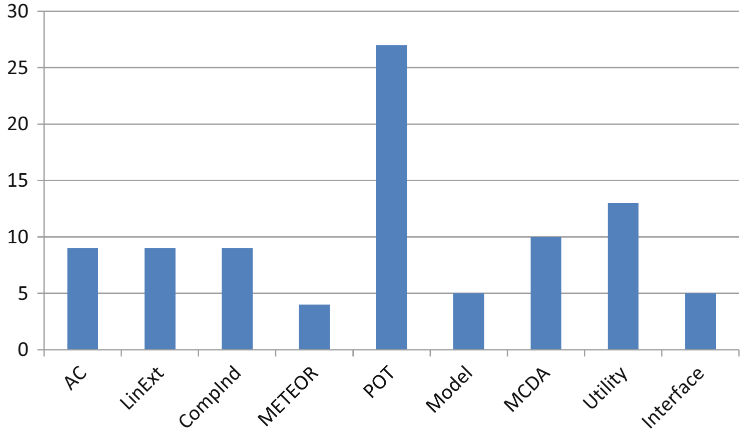


Fig. 19.2 Distribution of the 91 PyHasse modules within the nine contextual categories given in Table 19.3

Table 19.4 Libraries, supporting the PyHasse modules

Name of library	Description	Remark
raioop2.py	A library of classes, i.e., on procedures based on object oriented programming Mainly: user interfaces and graphics	Written by Brüggemann 4,500 lines of programming code
rmod2.py	Library of procedures, mainly of combinatorial character and manipulating matrices	Written by Brüggemann More than 6,800 lines of programming code
pstat.py	Statistics	Free downloadable from Internet. However, routinely delivered together with the other two libraries above as part of the PyHasse package
stats.py	Statistics	Free downloadable from Internet: however routinely delivered together with the other two libraries above as part of the PyHasse package

In addition, there is a “help” function, which has the following structure:

- Aim
- Prerequisites
- Usage or steps
- Results (not in all cases)
- Difficulties
- Literature
- Example data files

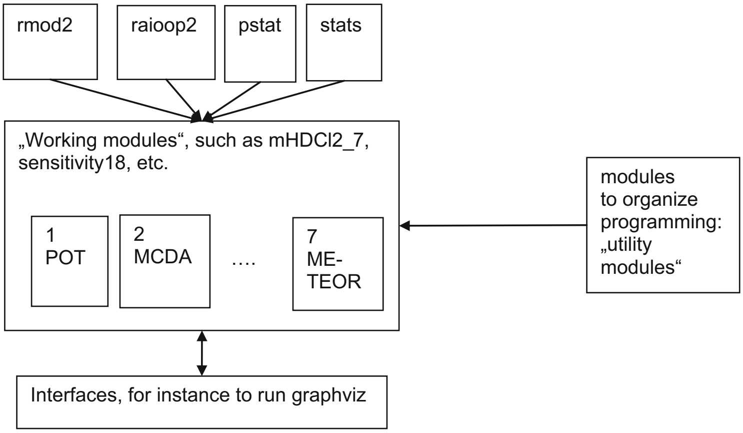

rmod2
raioop2
pstat
stats
„Working modules", such as mHDCl2_7, sensitivity18, etc.
1 POT
2 MCDA
….
7 ME-TEOR
modules to organize programming: „utility modules"
Interfaces, for instance to run graphviz

Fig. 19.3 Programming structure of PyHasse

19.5.2.4 PyHasse Data Flow (Example: Windows® as Operating System)

Within the Windows® environment the majority of potential users will apply Microsoft Excel®.

In order to fulfill the input requirement for the PyHasse module, it is important that the rows as well as the columns have a short label (optimal are labels with up to three characters) and that the (0,0) position of the data matrix (in Excel the A,1) is not empty. Furthermore, none of the PyHasse modules accept data gaps. Hence, it is in the responsibility of the users to provide a data sheet with all labels and no data gaps. In contrast, software packages such as DART (see Manganaro et al. 2008) and WHASSE (Brüggemann et al. 1999) provide some facilities to handle missing data.

Typically the PyHasse modules require the Excel sheet stored as a tab-separated txt file. Only the module EXCELHD1.py can directly apply the data by copying the appropriate field in the Excel sheet. Once the data matrix is read in, one may perform calculations and results can be stored in the internal format pdt. Some more important modules therefore offer to read these intermediate results as *.pdt files.

19.5.3 Overview

19.5.3.1 Most Often Used Modules

The application of the following modules is well described (cf. Table 19.1 and the appendix at the end. Further, specific references are available within the single modules).

- mainHD20_5.py and mHDCl2.py, resp.: Beside the Hasse diagram, these module provide navigation tools and much structural information, as well a variety of other facilities. As "basic" modules these are the most important

- chain7_1.py: Search and analysis of chains
- dds12.py: Dominance and separability of disjoint subsets of objects on the basis of the order relations among their elements
- LPOMext4_2.py: Average ranks calculated after two different approximations based on the “local partial order concept”
- fuzzyHD13.py: Instead of analyzing the “<” relation directly a subsethood is defined (Kosko measure, cf. Van de Walle et al. 1995) and a fuzzy partial order defined
- sensitivity19_1.py: A partially ordered set has a structure. This structure is characterizable by chains and antichains. What is the impact of any single matrix column (representing the indicator values for all the objects)? i.e., what is the impact of any single indicator on the structure of a poset?
- similarity10_1.py: The same set of objects may be described by different multi-indicator systems. What is the proximity between the two resulting posets?

19.5.4 Description of Some Modules of PyHasse Software

19.5.4.1 Module mHDCl2_7: The “New Main”

This module is one of the newest and is completely written in an object oriented programming style. The reason, why mHDCl2_7.py was developed, was threefold:

1. The similar module mainHD20_5.py runs into memory error when the data matrices are too large
2. The GUI and the logical organization were no more adequat
3. After some years of practical applications some adaptions appeared appropriate

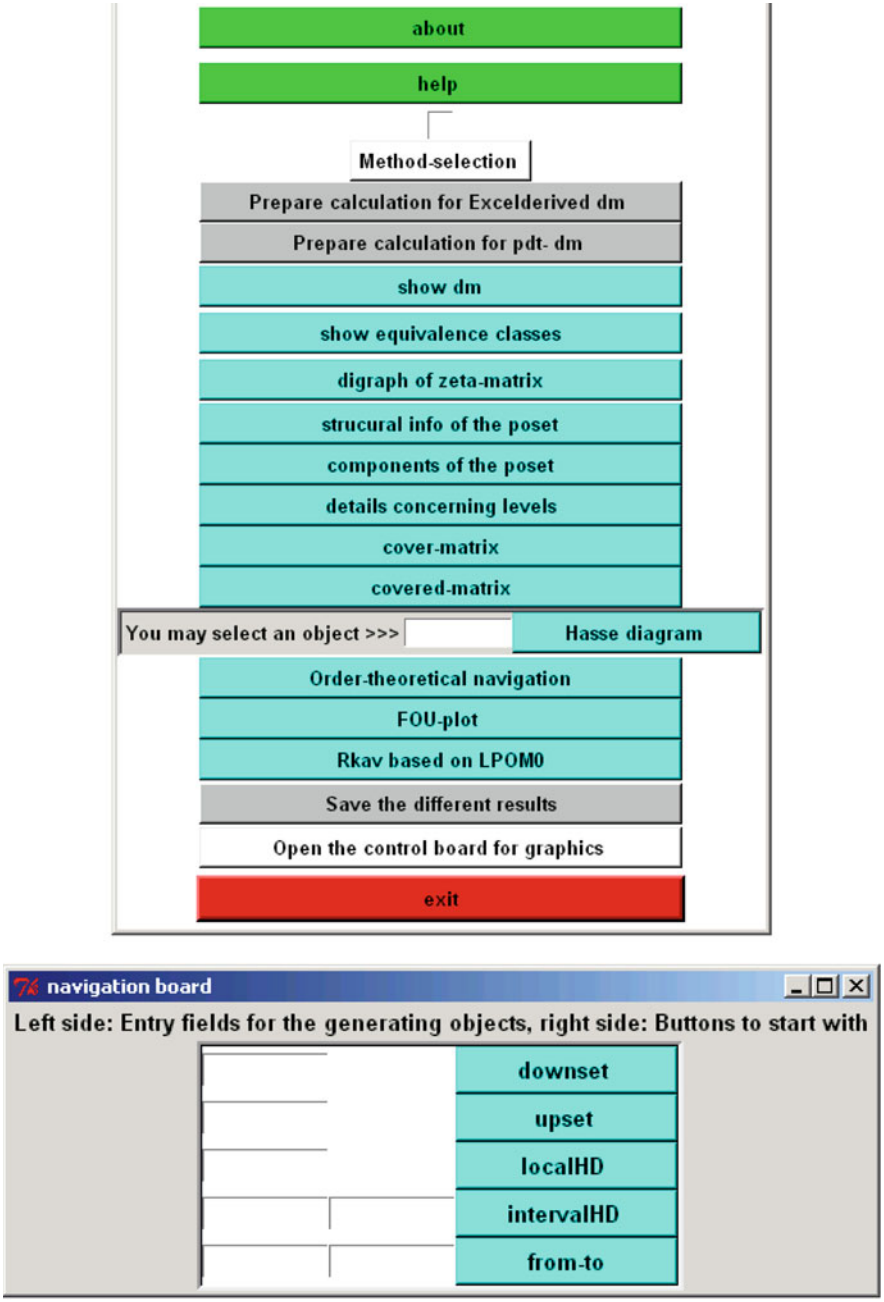
The purpose is, as with mainHD20_5.py, to provide a complete basical analysis of a partially ordered set as derived from a data matrix. This includes as results:

- Level structure
- Information of each object about its successors, predecessors, and incomparable objects in the Hasse diagram, in tabular form
- Hasse diagram
- Navigation tools: principal down- and upsets, interval graphs, local Hasse diagrams, the most simple approximation of average rank by the local partial order (LPOM0) (Brüggemann et al. 2004)

The GUI and its subsequent windows are shown in Figs. 19.4 and 19.5.

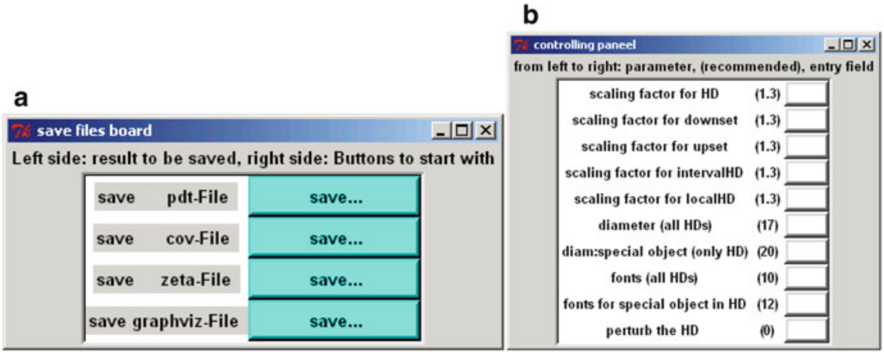
In the following we describe each button given in Fig. 19.5, starting from the top in Table 19.5.

In mHDCl2.py there are three other tools to overcome the difficulties of drawing Hasse diagrams: (a) by rendering information in a tabular form (Table 19.6) and (b) by the FOU plot, which is a realization of the concept of posetic coordinates (see Chap. 8), see Fig. 19.6.


about
help
Method-selection
Prepare calculation for Excelderived dm
Prepare calculation for pdt- dm
show dm
show equivalence classes
digraph of zeta-matrix
strucural info of the poset
components of the poset
details concerning levels
cover-matrix
covered-matrix
You may select an object >>>
Hasse diagram
Order-theoretical navigation
FOU-plot
Rkav based on LPOM0
Save the different results
Open the control board for graphics
exit
navigation board
downset
upset
localHD
intervalHD
from-to

Left side: Entry fields for the generating objects, right side: Buttons to start with

Fig. 19.4 GUI of mHDCl2_7.py and the window opening after pressing "Order theoretical navigation." Note that the first three navigation buttons need the input of one single object, whereas the buttons "intervalHD" and "from–to" need two objects as input



save files board

Left side: result to be saved, right side: Buttons to start with

save pdt-File

save cov-File

save zeta-File

save graphviz-File

controlling paneel

from left to right: parameter, (recommended), entry field

scaling factor for HD (1.3)

scaling factor for downset (1.3)

scaling factor for upset (1.3)

scaling factor for intervalHD (1.3)

scaling factor for localHD (1.3)

diameter (all HDs) (17)

diam:special object (only HD) (20)

fonts (all HDs) (10)

fonts for special object in HD (12)

perturb the HD (0)

Fig. 19.5 Windows popping up after pressing “Save the different results” (**a**) and “Open the control board for graphics” (**b**)

In Fig. 19.6, a FOU plot is shown. Myers and Patil (2014) are focusing on possibilities to represent partially ordered sets by scatter plots in order to avoid too complex Hasse diagrams. Here our aim is similar. The basic idea is to describe partially ordered sets by “posetic coordinates,” i.e., by numbers which are derived from partial order theory, e.g., the contents of principal down- and upsets and of *U*(*x*). When equivalence relations are possible, the number of equivalent elements could be used too to obtain posetic coordinates. Here we characterize the poset by two order theoretical coordinates for each object *x*, i.e., by the difference of the contents of down (*O*(*x*)) and upsets (*F*(*x*)), *OF* and the content of the set of elements incomparable with *x: U*.

$$OF := \left(\left|O\left(x\right)\right| - \mid F\left(x\right) \mid\right) \text{ and } U := \mid U(x) \mid . \tag{19.1}$$

- In contrast to the coordinates, the original data matrix may render (Pb, Cd, Zn) now posetic coordinates, namely OF and *U* are used to characterize the objects.
- In contrast to the triangle coordinate representation (Brüggemann and Patil 2011), which is more detailled, the scatter plot, based on OF amd *U* is simple to be interpreted.

Generally, it is a promising new task in partial order theory to find best “posetic coordinates” allowing presentations of partial orders not so much depending on the clarity of the relational graph, such as the Hasse diagram.

Figure 19.6 shows that

- There are two regions selected (namely 8 and 14) being maximal elements, however, they differ in their values of their posetic coordinates.
- There is one region being at most incomparable |*U*(*x*)|= 7. object, this is region 17, which also is a maximal element.

Table 19.5 Explanations of the buttons of the GUI of mHDCl2_7.py

Button	Explanation	Remark
Method selection	0 as input A method to perform the transitive reduction is performed, which eliminates step by step the transivities 1 as input A method is used, following Simon (1992), which, however, leads to memory errors when the adjacency matrix is to be calculated and the number of objects is too large (>200) and at the same time the number of comparabilities is high	
Prepare calculation for Excel-derived dm	When this button is pressed, the module expects a data matrix following the principles explained in section "*PyHasse data flow (example Windows as operating system)*"	Internally all calculations are performed to get the Hasse diagrams and other combinatorial results
Prepare calculation for data matrices in the pdt Format	The module expects data matrices in the internal format pdt. This facility is not as often used as the Excel-derived dm	Internally all calculations are performed to get the Hasse diagrams and other combinatorial results
Show dm	Attributes (indicators) as well as the labels of the objects are shown. Furthermore, the complete data matrix is displayed	When the button "Hasse diagram" is pressed, the exact label of the object is needed
Show equivalence classes	If two rows, i.e., two objects have identical values for all indicators then the two objects are considered as equivalent, the alphabetically first object is retained A graphical as well as a tabular presentation of equivalence classes can be obtained	
Digraph of zeta matrix	The zeta matrix describes the order relations among the objects. In contrast to the representation in the Hasse diagram, which is based on the cover relations, the relations corresponding to transitivity of the order relation are shown too	
Structural info of the poset	Being aware that Hasse diagrams can be a complex system of lines (see Carlsen and Brüggemann 2013a) all needed information are provided in tabular form	See Table 19.6
Components of the poset	Graph theoretically the acyclic-directed graph may have vertices which are not connected (in former publications also called "hierarchies"). Here an information about the number of components and the distribution of objects over these components is available	

(continued)

Table 19.5 (continued)

Button	Explanation	Remark
Details concerning levels	Levels are an important structure and a mean to get the set of objects weakly ordered (Brüggemann and Patil 2011) Therefore there is a multitude of more detailed information available	See below
Cover matrix	Even if the Hasse diagram is drawn, one may want to get a list of cover relations Here “Object *x* is covering …” is given	
Covered matrix	Similar as above. However, here “Object *x* is covered by…” is given	
Hasse diagram	A Hasse diagram is drawn. When the entry field at the left side is filled with the correct label of an object, this object will be marked in the graphic	See also Fig. 19.1
Order-theoretical navigation	It pops up an extra window, where it can be specified which navigation is wanted	See for instance Fig. 19.10
FOU plot	In order to analyze large data matrices, the Hasse diagram is often not suitable because of its complexity. Then other representations must be selected, as is pointed out in Myers et al., in several papers and in this book (Myers and Patil 2008; Myers et al. 2006) Here new coordinates are introduced for each object *x*: Abscissa: Difference of objects in downset and upsets of *x*: $(\|O(x)\|\text{-}\|F(x)\|)$ Ordinate, number of objects incomparable with *x*, $\|U(x)\|$. Because $F(x)$ is used as symbol for upset(x), $O(x)$ as symbol for downset *x*, and $U(x)$ for the set of incomparable objects with *x*, the name FOU plot was used	See also Fig. 19.11
Rkav based on LPOM0	The local partial order model LPOM0 will be applied to get an approximation the average ranks	In LPOMext4.py a more sophisticated approximation is available, due the extended LPOM In avrank5.py the exact average rank is available when certain conditions are fulfilled
Save the different results	See also Fig. 19.5	
Open the control board for graphics	See also Fig. 19.5 to get an impression about the multitude how graphics can be manipulated. In parentheses the default values are shown	
Exit	It is important to exit the program in order to avoid damages	

Table 19.6 Structural information of the data matrix of Table 19.2, related with the Hasse diagram of Fig. 19.1

One linear extension

6 < 29 < 7 < 5 < 16 < 9 < 14 < 17 < 8

Maximal elements

8, 14, 17

Minimal elements

29, 6

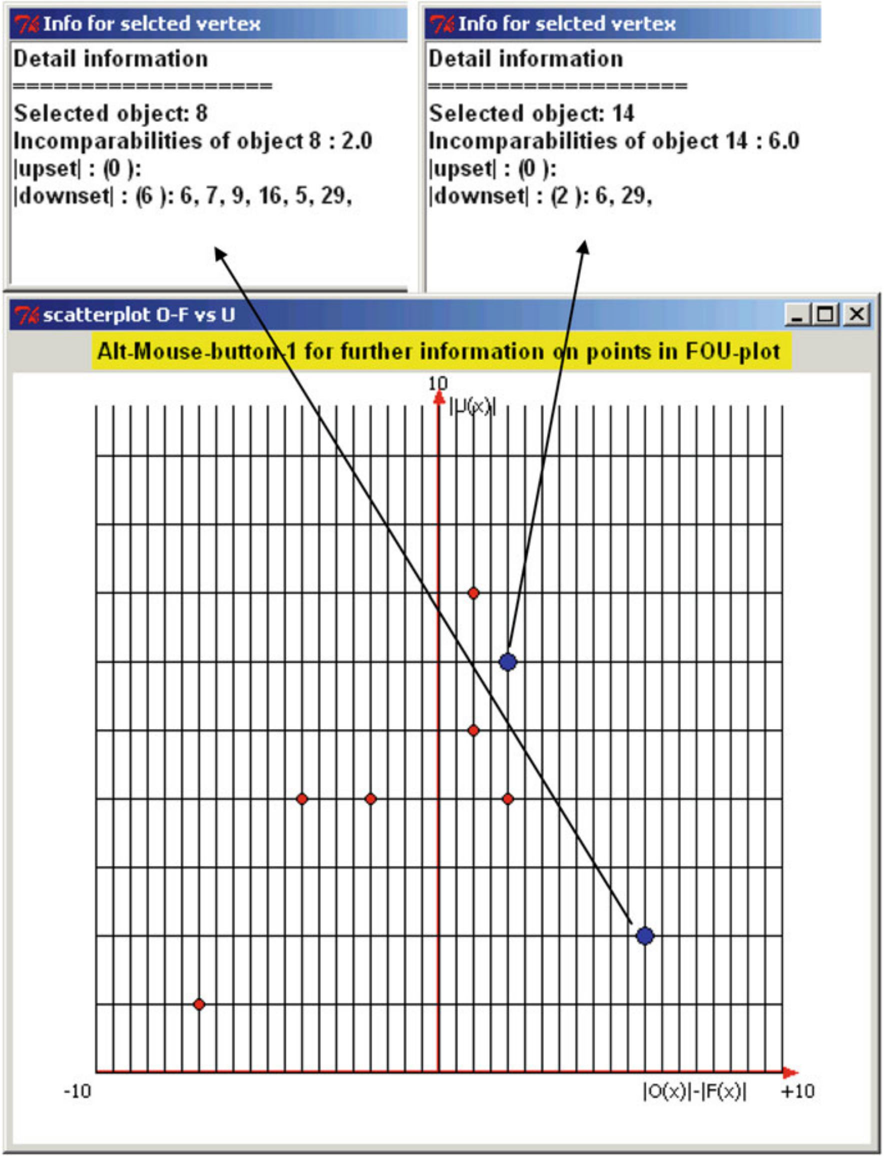
Isolated elements

Individual info:

First the object, then in parentheses: count of, then the list of elements
Sets of incomparable elements
6: (1): 29
8: (2): 14, 17
7: (4): 29, 14, 17, 16
17: (7): 14, 16, 29, 5, 7, 9, 8
9: (5): 29, 5, 14, 17, 16
16: (5): 9, 5, 17, 7, 14
14: (6): 17, 16, 5, 7, 9, 8
5: (4): 9, 14, 17, 16
29: (4): 9, 17, 7, 6
Downsets
6: (1): 6
8: (7): 16, 29, 5, 7, 6, 9, 8
7: (2): 7, 6
17: (2): 17, 6
9: (3): 9, 7, 6
16: (3): 16, 29, 6
14: (3): 14, 29, 6
5: (4): 5, 29, 7, 6
29: (1): 29
Upsets
6: (8): 14, 17, 16, 5, 7, 6, 9, 8
8: (1): 8
7: (4): 9, 8, 5, 7
17: (1): 17
9: (2): 9, 8
16: (2): 8, 16
14: (1): 14
5: (2): 8, 5
29: (5): 8, 5, 14, 29, 16

Checking the data matrix one can see that indeed regions 8 and 14 are pretty different with respect to their data profile (for the sake of clarity, the min, and max values over all regions for each of the three attributes are additionally given):

	Pb	Cd	Zn
Max:	20	0.6	63
8:	20	0.4	55
14:	12	0.6	41
Min:	9	0.2	29



```
Detail information
===================
Selected object: 8
Incomparabilities of object 8 : 2.0
|upset| : (0 ):
|downset| : (6 ): 6, 7, 9, 16, 5, 29,
```

```
Detail information
===================
Selected object: 14
Incomparabilities of object 14 : 6.0
|upset| : (0 ):
|downset| : (2 ): 6, 29,
```

Fig. 19.6 FOU plot (see text) based on Table 19.2, using posetic coordinates. The greater *blue circles* are obtained by clicking with the mouse on them. The abscissa counts from −10 to +10 with steps of 0.5, the ordinate, however, counts from 0 to 10 with steps of 1

Region 8 is dominantly polluted by Lead and Zinc, whereas the main contribution of pollution of region 14 is Cadmium. The maximal and minimal values of Pb, Zn, and Cd taken over all objects of the data matrix (Table 19.2) are added to facilitate the interpretation.

The FOU plot is mainly useful for an interactive analysis and can be further explored using the mouse. So the FOU plot fulfills similar tasks as those, explained by Myers in this book (Myers and Patil 2014) There is an abscissa which describes the relative position on a bad–good axis and the ordinate which quantifies the conflicts associated with each object.

Clicking with the left mouse button, pessing "ALT" a window pops up with more information (Fig. 19.6, top, left side, and right side). Basically, depending on the ranking aim, the points near the lines given by (19.2a) and (19.2b)

$$|U(x)| = n+1-OF,\ OF = (|O(x)|-|F(x)|), |F(x)| = 0 \tag{19.2a}$$

and

$$|U(x)| = n+1+OF,\ OF = (|O(x)|-|F(x)|), |O(x)| = 0 \tag{19.2b}$$

are of most interest, as they are the extremal points.

In contrast to mainHD20_5.py, the module mHDCl2.py does no more contain the Bubley–Dyer algorithm (Bubley and Dyer 1999) to get average ranks (see Patil and Joshi 2014) and the statistics concerning chain length. The BubleyDyer algorithm is now the central part of the module BubleyDyer8.py where also the algorithm, proposed by Patil and Taillie (2004), the Cumulatice Rank Frequency (CRF) iterative method is provided. The CRF algorithm can be applied to enrich the poset until a weak order is obtained. See for details Chap. 6.

19.5.4.2 The Module to Check the Role of Single Indicator Values: POOC6.py

As mentioned by Annoni et al. (2011, 2012) and explained in more detail by Brüggemann and Patil (2011), there are two types of sensitivity analysis:

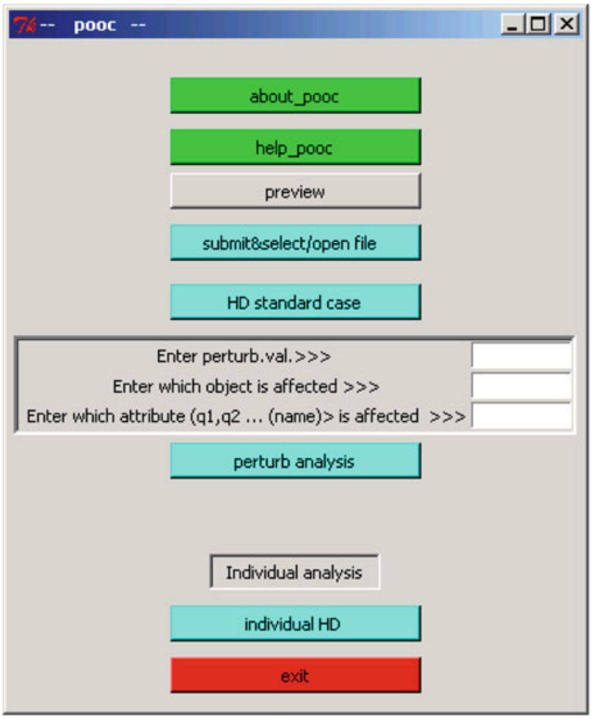
- Variation of the set of indicators, e.g., to elucidate the effect if one indicator is eliminated from the data matrix
- Variation of the values of indicators

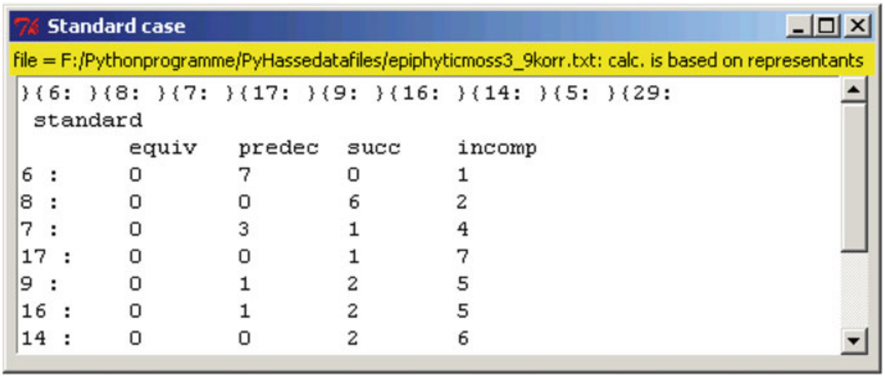
The first is referred to as attribute-related sensitivity (ARS), the second as attribute value-related sensitivity (AVRS). The ARS is the task of sensitivity18_3.py and is well described in the literature. Attribute value-related sensitivity is the task of POOC6.py (perturbation on order characteristics). With the new concept of variance-based sensitivity (Annoni et al. 2011, 2012; see also Chap. 13), the development concerning POOC6.py was slowed down. Nevertheless, this module appears mandatory, as long as the variance-based sensitivity is not programmed within PyHasse.

The GUI of POOC6.py is shown in Fig. 19.7.

After selecting the same data matrix as for Fig. 19.1, a posetic overview over the data matrix (Fig. 19.8) is first obtained, whereby now four coordinates are used.

Fig. 19.7 GUI of POOC6.py





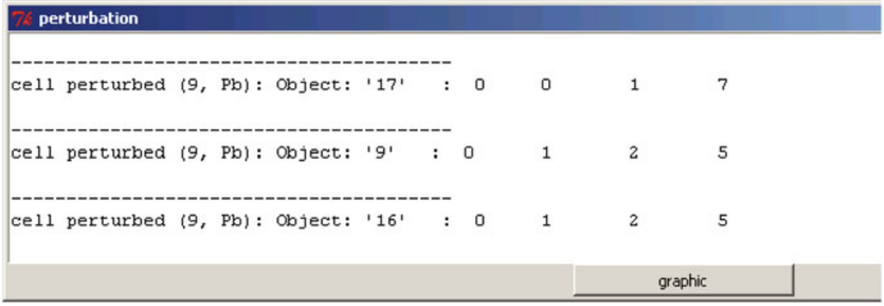
file = F:/Pythonprogramme/PyHassedatafiles/epiphyticmoss3_9korr.txt: calc. is based on representants

```
}{6: }{8: }{7: }{17: }{9: }{16: }{14: }{5: }{29:
 standard
          equiv    predec   succ     incomp
6  :      0        7        0        1
8  :      0        0        6        2
7  :      0        3        1        4
17  :     0        0        1        7
9  :      0        1        2        5
16  :     0        1        2        5
14  :     0        0        2        6
```

Fig. 19.8 Posetic "coordinates" (equiv, predec, succ, and incomp) of the data matrix, describing metal pollution of epiphytic mosses, in south-west of Germany (cf. Table 19.2)

The coordinates are:

- *equiv* (*eq*): number of equivalent elements with *x*
- *predec* (*pred*): number of elements above *x*, *pred* =|*O*(*x*)-{*x*}|
- *succ*: number of elements below *x*, *succ* = |*F*(*x*)-{*x*}|
- *incomp* (*ic*): number of elements incomparable with *x*, *ic*= |*U*(*x*)|



```
----------------------------------------
cell perturbed (9, Pb): Object: '17'   :  0     0       1       7

----------------------------------------
cell perturbed (9, Pb): Object: '9'   :  0      1       2       5

----------------------------------------
cell perturbed (9, Pb): Object: '16'   :  0     1       2       5
```

Fig. 19.9 Posetic coordinates, after perturbing the value of attribute Pb of region 9 by changing the attribute by adding the value of 1

Thus, as an example, select as element of interest region 9, one sees that there is no element, equivalent with region 9, there is one element above 9, two elements below region 9 and 5 regions which are not comparable to region 9.

The role of pooc6.py is now to check how a change in an attribute value will change the set of posetic coordinates.

We enter a perturbing value 1, select object 9 and attribute "Pb," i.e., pollution of lead in the epiphytic moss. Note that clearly a perturbing value for one of the columns of the data matrix is specific, for instance, 1 for Pb concentration is a small value, 1 for Cd pollution would be larger than the whole span of Cd values! Here we perturb Pb by 1/20 of the maximal value.

Technically a perturbation by 1 means that we change the original entry *q1(9)* by adding 1, i.e., changing the value from 17 to 18, and thus observe the possible effects (Fig. 19.9).

To the most left side: the site of perturbation and the perturbed indicator are explained, then information is given how the different elements of the poset are reacting.

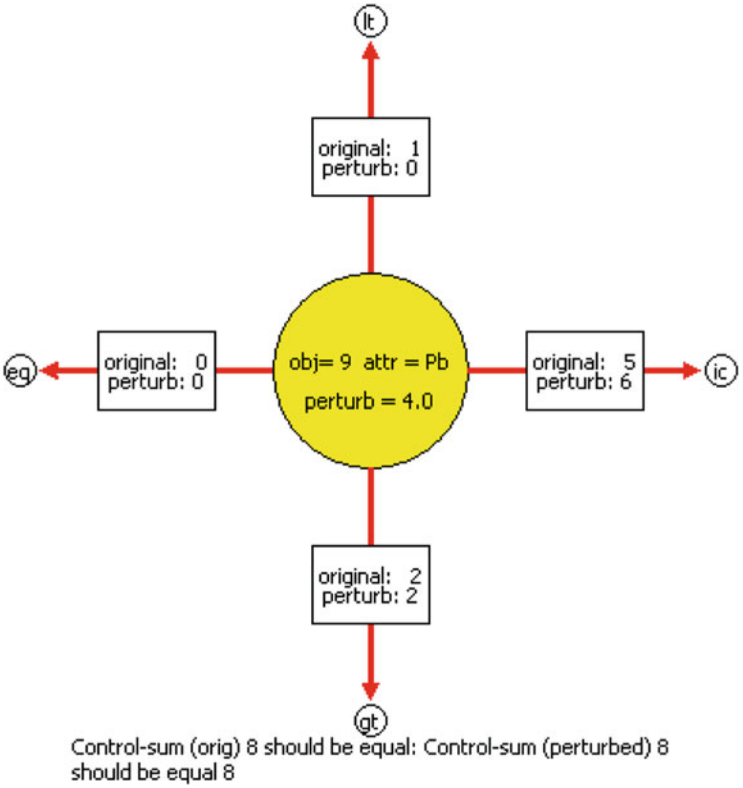
A perturbation by adding 4 to the original value, i.e., 20 % of the maximum of lead concentrations with the regions considered, changes the coordinates.

	eq	predec	succ	incomp
(Perturbed) (9, Pb): Obj: 9:	0	0	2	6

The number of predecessors of region 9 would in this case be reduced and the number of incomparable elements with region 9 increased.

We could conclude that the posetic information concerning region 9 is rather stable with respect to increasing the value of Pb. Clearly this procedure can be repeated for every element of interest, every attribute, and with every perturbing value.

In Fig. 19.10 a graphical display on what happens after perturbing the value of Pb for region 9 by 4 is given (in terms of region 9 less than (lt), greater then (gt), incomparable with (ic), and equivalent with (eq)).



Control-sum (orig) 8 should be equal: Control-sum (perturbed) 8 should be equal 8

Fig. 19.10 Schematic overview about the four posetic coordinates: eq: number of equivalences, ic: number of incomparabilities, gt: number of predecessors (greater), and lt: number of successors (less than) after changing the attribute Pb of region 9 by four units

Figure 19.10 deserves some further explanations: The large yellow circle informs about the perturbation itself.

The affected object (obj) is region 9, the attribute (attr) perturbed is Pb, and the amount of perturbation (perturb) is 4.0. In the little white circle the four posetic coordinates are indicated and in the rectangular box an information is given, what happens with respect to this specific coordinate: For example, the value of "lt" does not changed, i.e., it is not perturbed (perturbed value of lt:=0). In contrast, the coordinate "ic" changes by perturbation. The original value is 5, after perturbation this value changed to 6. (perturb (of ic)): 6.

Figures 19.8, 19.9, and 19.10 are the result of POOC6.py, which in turn is designed to help to find answers concerning the Hasse diagram in Sect. 19.4, namely the effect of data uncertainty. Additionally, however not shown, any perturbed data matrix can be visualized by a Hasse diagram.

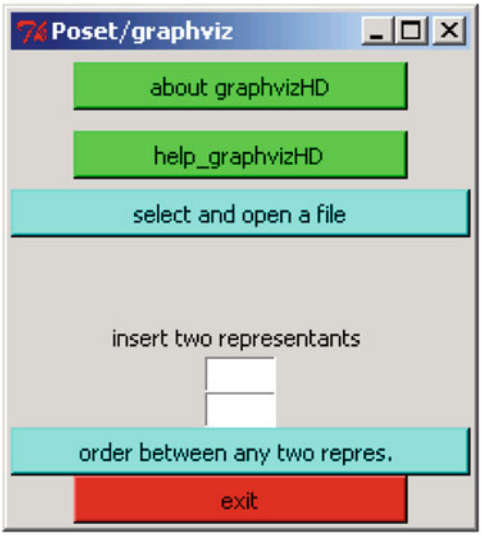


Fig. 19.11 GUI of graphvizHD1

19.5.4.3 Module graphvizHD1.py

Introduction

This module serves as a nice example for the general philosophy in the context of PyHasse, i.e., not to compete with professional software, if available. In the module graphvizHD1.py, some information on the partially ordered set is given. However, the graph drawing is a matter of the well-known graph–theoretical program Graphviz (Gansner and North 1999), which is explained below. In Fig. 19.11 the GUI is shown.

As for other modules, about and help functions are found. Behind the button "select and open a file," the facilities of the Tkinter library are applied.

After the selection of the data file, a window pops up with more information (see Table 19.7).

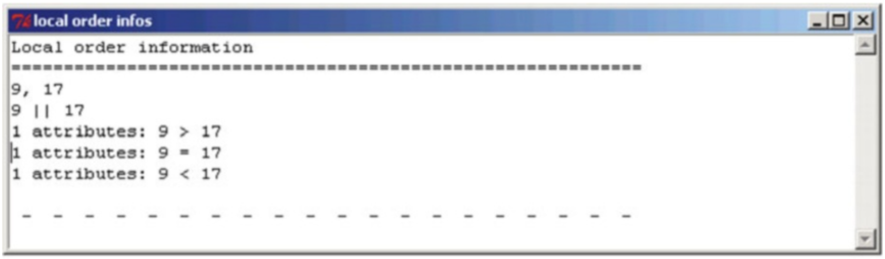
It is seen that local information (i.e., information not related to the complete object set, but to a user selected pair of objects) is available by inserting objects into the two open entry fields. Inserting for example "9" and "17," two objects of the data matrix of the selected example file, an information is obtained: a) comparable or not and b) in which orientation the two regions are comparable. Here it is found: 9 || 17, see also Fig. 19.12.

For a deeper analysis procedure, the concept of "distance due to incomparability" (Bartel and Mucha, Chap. 3) may be applied.

Further, a window is opened to select name and site of the file, subsequently to be analyzed by graphviz.

Table 19.7 Content of the window, popping up after selecting and opening the file (containing the data matrix about pollution in epiphytic mosses) (and doing subsequently all needed calculations to obtain the partial order)

Info about poset of F:/Pythonprogramme/PyHassedatafiles/epiphyticmoss3_9korr.txt
1. General info
Objects (representants)
6, 8, 7, 17, 9, 16, 14, 5, 29
Properties (indicators, attributes)
Pb, Cd, Zn
2. Posetic info
Number of levels (= length of maxim. chains) = 4
Number of elements in largest level = 3
Comparabilities = 17
Incomparabilities = 19.0
Count of maximal elements = 3
Maximal elements
8, 17, 14
Count of minimal elements = 2
Minimal elements
6, 29
Count of isolated objects = 0
Isolated objects



```
Local order information
==============================================================
9, 17
9 || 17
1 attributes: 9 > 17
1 attributes: 9 = 17
1 attributes: 9 < 17

 -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
```

Fig. 19.12 Local information about a pair of objects, here about a pair of regions

Graphviz

Graphviz is a professional program to draw graphs, i.e., visualize binary relations on a ground set. Graphviz draws binary relations of the object set, or more exactly, of the set of representatives. The software Graphviz is freely downloadable from the Internet and is described by Gansner et al. (1993) and Gansner and North (1999).

The version used here is 2.26.3, and among the programs available in Graphviz, the program Gvedit, v: 1.01 is used.

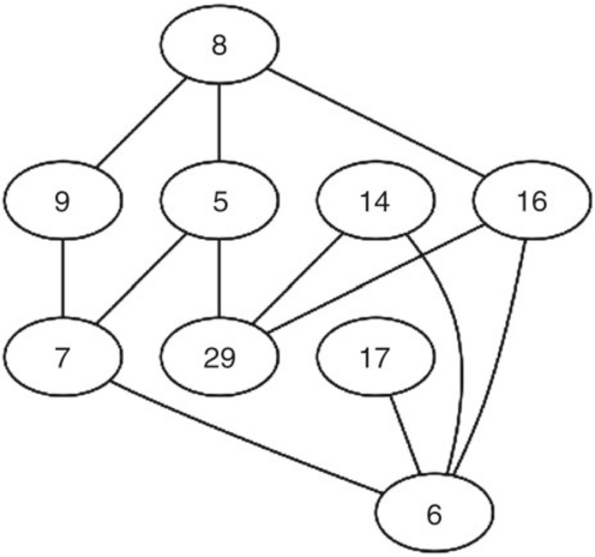


Fig. 19.13 Result after running Gvedit: a gif File

By successful running Gvedit, for instance, a gif File is obtained (see Fig. 19.13), many other formats are available too. It represents the same order relation as in Fig. 19.1. However, the drawing rules in Graphviz are dominated by minimizing the crossings of lines. Gvedit allows many controlling interactions by the user: However, for most purposes those specifications are not needed. A more detailed description of the graphviz option is outside the scope of the present chapter.

19.6 Summary and Conclusions

19.6.1 Summary

When a data matrix is to be analyzed with respect to some ranking or evaluation, then usually one has to select a software. Whereas the construction of a composite indicator is simple and can be done with spreadsheet facilities, like MS Excel, the analysis within partial order methodology can in general not be done by spreadsheet software.

With the example of a small real-life data matrix, where the regional pollution is measured in the special target of epiphytic the technical performance by some modules of PyHasse are demonstrated. There are PyHasse modules available, which unequivocallly are important and often used, for example, mainHD20_5.py, mHDCL2.py, similarity10_1, or sensitivity18_3, LPOM4ext.py, and dds12.py. Some others are described and follow crudely the logical line:

1. What information is provided by the Hasse diagram? (mHDCl2_7.py, Sect. 19.5.4.1)?
2. What happens when data entries are changed? (POOC6.py, Sect. 19.5.4.2)
3. Graphical display in the form of Hasse diagrams is appealing. However, the display of partial orders allows many freedoms. In PyHasse firstly a conservative point of view is taken, i.e., to locate the objects in the highest position, which is order theoretically possible. Secondly the objects are arranged in levels. These two principles often lead, in the graphical presentation of the results (the Hasse diagram), to crossing of lines, which may be rather confusing. Thus, an alternative is discussed, and the use of the freely downloadable software Graphviz is suggested (Sect. 19.5.4.3).

19.6.2 Conclusions

PyHasse is today applied by many teams around the world. It is clear that correspondingly many ideas are expressed how PyHasse can be improved

- In its technical handling
- Contextually, in its tools to ordinally analyze data matrices

PyHasse is not claimed as a user-friendly software with a good guidance of the users. However, it should be clear that PyHasse does not want to (and cannot) compete with, for instance, DART (Manganaro et al. 2008), which provides a very convenient tool to get Hasse diagrams as well as some basic information derived from a data matrix—even with missing data. The application of PyHasse needs some preparatory steps in data handling before it can be run. It also most often needs an a posteriori activity by the user. This is the consequence of the conception behind PyHasse, to help specifically in studies of partial ordering, i.e., in all consequences which arise from the ordinal analysis of multi indicator systems. PyHasse provides copy-and-paste texts to support the documentation of results.

When graphical representations are available (bar diagrams, Hasse diagrams, scatter plots, etc.), their purpose is to give the user a first impression, and when the user wants a professional graphic software, such as Excel, then some few steps of data handling are necessary.

PyHasse is rapidly developing as ideas from users as well as concepts from the literature relatively easily may be programmed leading to new modules. The price is that the total absence of bugs cannot be guaranteed, albeit most modules are tested rather carefully. Futher the user interfaces may not always be as comfortable as possibly desirable and philosophies how to guide the user are only rudimentarily realized. PyHasse is an "experimental" software under constant development and suggestions, comments, and wishes from users are always welcome and appreciated.

19.7 Outlook

For the time being, eight major objectives are still on the agenda. However, obviously time is required and the development further constantly compete with other more rapidly realizable ideas. The eight future objective can be summarized as:

1. PyHasse being made available in an Internet version. Some preliminary attempts have been made. However, the cooperation with web designers, etc., appears crucial.
2. Although the powerful conexp3, written in Java is available for analysis of Formal Concepts (Yevtushenko 2003; Ganter and Wille 1996, Burmeister 2003) a Formal-Concept-Analysis-module within PyHasse would facilitate many applications.
3. POSAC is a program performing a reduction of the attributes of the data matrix to two coordinates. The underlying idea is to maintain the typical outcome of partial order theory, i.e., the appearance of incomparabilities but at the same time simplifiying the analysis. POSAC is an approximation. Nevertheless a, possibly simplified version in PyHasse would be helpful (see Brüggemann and Patil 2011 and references therein).
4. When a reduction to two new attributes as in POSAC is intended, a calculation of the poset dimension would be useful. However, the calculation of the dimension of a poset is computationally extremely difficult. Nevertheless, it is important to get ideas about the dimension of posets.
5. The variance-based sensitivity analysis is most urgently needed as an implementation in PyHasse. So far, the needed calculations are performed using Matlab. Consequently, the extensive numerical part should be programmed in C++ or at least by including the library NumPy.
6. In multivariate statistics, cluster analysis plays an important role. A straightforward application of cluster analysis is suitable in order to get clear Hasse diagrams by reducing the number of vertices. This reduction can be done in the form of deriving a poset on cluster centers instead on the single objects. This reduction is the main feature of the PyHasse module pycluster1_2.py. However, an order theoretical approach would be helpful too: Instead of defining equivalence relations such as "belonging to the same cluster," one could appropriately define equivalence relations among the elements of a poset, also called "blocks" (Davey and Priestley 1990) and analyze the resulting posets based on the representative elements, which clearly is simpler than the original poset.
7. Finally, a project aiming at extending PyHasse by an additional fuzzy-poset analysis is in progress. A first variant is provided in fuzzydds7.py. Now an intensive testing phase is needed.
8. The further analysis of the two approximations of average ranks based on local partial order model is a task for the future. It is hoped to give improved statements about the accuracy of the LPOM model.

19.8 (a) List of Abbreviations (Alphabetically Sorted)

Abbreviation	Meaning
AC, ac	Antichain
ACM	Antichain matrix
ARS	Attribute-related sensitivity
AVRS	Attribute value-related sensitivity
CompInd	Composite indicators
DART	Decision analysis by ranking techniques
dm	Data matrix
equiv(eq)	Number of equivalent elements of a certain object
FCA	Formal concept analysis
FOU	Plot derived from $\lvert F(x)\rvert$, $\lvert O(x)\rvert$, $\lvert U(x)\rvert$ (contents of upset of x, downset of x, incomparables with x)
GUI	Graphical user interface
HD, hd	Hasse diagram
HDT	Hasse diagram technique
IB	Information base (set of attributes of a certain ranking study)
incomp(ic)	Number of incomparable elements of an object x
LinExt	Linear extensions
LPOM	Local partial order model
LPOM0	LPOM, based on the simplest approximation
LPOMext	LPOM, based on an extended method
MAC	Macintosh
MCDA	Multicriteria decision analysis
METEOR	Method of evaluation by order theory
OS	Operating system
Parsec	Partial orders in Socioeconomics
POSAC	Partial order scalogram with coordinates
poset	Partially ordered set
POT	Partial order theory
predec(pred)	Number of predecessors of a certain object
Rapid	Ranking and prioritization information delivery
Rkav	Average height, commonly called average rank
succ	Number of successors of a certain object
VHLL	Very high-level language

19.9 (b) Further Recommended References Within the Context of PyHasse and HDT

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

Brüggemann R, Voigt K, Restrepo G, Simon U (2008) The concept of stability fields and hot spots in ranking of environmental chemicals. Environ Model Softw 23:1000–1012

Brüggemann R, Kerber A, Restrepo G (2011) Ranking objects using fuzzy orders, with an application to refrigerants. Match Commun Math Comput Chem 66(2):581–603

Carlsen L, Brüggemann R (2009) Partial order ranking as a tool in environmental impact assessment. In: Halley GT, Fridian YT (eds) PAH and PCB pollution of the River Main as an illustrative example. . environmental impact assessment. Nova Science Publishers, pp 335–354

Carlsen L, Brüggemann R (2011) Risk assessment of chemicals in the River Main (Germany): application of selected partial order ranking tools. Statistica and Applicazioni, special issue 125–140

De Loof K, De Meyer H, De Baets B (2006) Exploiting the lattice of ideals representation of a poset. Fundamenta Informaticae 71:309–321

De Loof K, De Baets B, De Meyer H, Brüggemann R (2008) A Hitchhiker's guide to poset ranking. Comb Chem High Throughput Screen 11:734–744 (3-3)

De Loof K, De Baets B, De Meyer H (2011) Approximation of average ranks in posets. Match Commun Math Comput Chem 66:219–229

Restrepo G, Brüggemann R (2008) Dominance and separability in posets, their application to isoelectronic species with equal total charge. J Math Chem 44: 577–602

Restrepo G, Weckert M, Brüggemann R, Gerstmann S, Frank H (2008) Ranking of refrigerants. Environ Sci Technol 42:2925–2930 (3-8)

Sailaukhanuly Y, Zhakupbekova A, Amutova F, Carlsen L (2013) On the ranking of chemicals based on their PBT characteristics: comparison of different ranking methodologies using selected POPs as an illustrative example. Chemosphere 90:112–117

Simon U, Brüggemann R, Mey S, Pudenz S (2005) METEOR – application of a decision support tool based on discrete mathematics. Match Commun Math Comput Chem 54:623–642

Simon U, Brüggemann R, Behrendt H, Shulenberger 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 Hydrochim Hydrobiol 34:126–136

Tsakovski S, Simeonov V (2011) Hasse diagram technique as exploratory tool in sediment pollution assessment. J Chemometrics. doi:10.1002/cem.1381:1-8

Tsakovski S, Kudlak B, Simeonov V, Wolska L, Garcia G, Namiesnik J (2012) Relationship between heavy metal distribution in sediment samples and their ecotoxicity by the use of the Hasse diagram technique. Analytica Chimica Acta. doi: http:///10.1016/j.aca.2011.12.052

Voigt K, Brüggemann R, Scherb H, Shen H, Schramm K-H (2010a) Evaluating the relationship between chemical exposure and cryptorchidism by discrete mathematical method using PyHasse software. Environ Model Softw 25:1801–1812

Voigt K, Brüggemann R, Kirchner M, Schramm K-W (2010b) Influence of altitude concerning the contamination of humus soils in the German Alps: a data evaluation approach using PyHasse. Environ Sci Pollut Res 17:429–440

References

Al-Sharrah G (2010) Ranking using the Copeland score: a comparsion with the Hasse diagram. J Chem Inf Model 50:785–791

Al-Sharrah G (2011) The Copeland method as a relative and categorized ranking tool. Stastidtica et Applicazioni, special issue, 81–95

Annoni P (2007) Different ranking methods: potentialities and pitfalls for the case of European opinion poll. Environ Ecol Stat 14:453–471

Annoni P, Brüggemann R, Saltelli A (2011) Partial order investigation of multiple indicator systems using variance – based sensitivity analysis. Environ Model Softw 26:950–958

Annoni P, Brüggemann R, Saltelli A (2012) Random and quasi-random designs in variance-based sensitivity analysis for partially ordered sets. Reliab Eng Syst Saf 107:184–189

Bartel H-G, Brüggemann R (1998) Application of formal concept analysis to structure-activity relationships. Fresenius J Anal Chem 361:23–28

Birkhoff G (1984) Lattice theory, vol XXV. American Mathematical Society, Providence, RI

Brans JP, Vincke PH (1985) A preference ranking organisation method (The PROMETHEE method for multiple criteria decision – making). Manag Sci 31:647–656

Brüggemann R, Carlsen L (2011) An improved estimation of averaged ranks of partial orders. Match Commun Math Comput Chem 65(2):383–414

Brüggemann R, Carlsen L (2012) Multi-criteria decision analyses. Viewing MCDA in terms of both process and aggregation methods: Some thoughts, motivated by the paper of Huang, Keisler and Linkov. Sci Total Environ 425:293–295

Brüggemann R, Patil GP (2010) Multicriteria prioritization and partial order in environmental sciences. Environ Ecol Stat 17:383–410

Brüggemann R, Patil GP (2011) Ranking and prioritization for multi-indicator systems – introduction to partial order applications. Springer, New York, NY

Brüggemann R, Voigt K (2008) Basic principles of Hasse diagram technique in chemistry. Comb Chem High Throughput Screen 11:756–769

Brüggemann R, Sørensen PB, 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, Kaune A, Pudenz S, Komossa D, Friedrich J (1998) Vergleichende ökologische Bewertung von Regionen in Baden- Württemberg GSF-Bericht 20/98. GSF, Neuherberg

Brüggemann R, Bücherl C, Pudenz S, Steinberg C (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim Hydrobiol 27:170–178

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (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

Bubley R, Dyer M (1999) Faster random generation of linear extensions. Discrete Math 201: 81–88

Burmeister P, CONIMP4, Programm zur Formalen Begriffsanalyse (1997) Technische Hochschule Darmstadt, Arbeitsgruppe 1, Fachbereich 4 (Mathematik) WWW-Adresse: http://www.mathematik.tu-darmstadt.de/~burmeister/ConImpIntro.pdf (last access August, 2013)

Carlsen L (2008) Hierarchical partial order ranking. Environ Pollut 155:247–253

Carlsen L (2009) The interplay between QSAR/QSPR studies and partial order ranking and formal concept analyses. Int J Mol Sci 10:1628–1657

Carlsen L, Brüggemann R (2013a) Indicator analyses, what is important: and for what? In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order, Chap 18. Springer, New York, NY

Carlsen L, Brüggemann R (2013b) The 'Failed Nations Index' offers more than just a simple Ranking. Soc Indic Res. doi:10.1007/s11205-012-9999-6

Davey BA (2004) Formal concept analysis. In: Eklund P (ed) ICFCA 2004, LNAI 2961. Springer, Berlin, pp 55–56

Davey BA, Priestley HA (1990) Introduction to lattices and order. Cambridge University Press, Cambridge

DEA. http://mat.gsia.cmu.edu/classes/QUANT/NOTES/chap12.pdf. Assessed 16 Oct 2012

Gansner ER, North SC (1999) An open graph visualization system and its applications to software engineering. Softw Pract Exp 30(11):1203–1233

Gansner ER, Koutsofios E, North SC, Vo K-P (1993) A technique for drawing directed graphs. IEEE Trans Softw Eng 19:214–230

Ganter B, Wille R (1996) Formale Begriffsanalyse Mathematische Grundlagen. Springer, Berlin

Halfon E (2006) Hasse diagrams and software development. In: Brüggemann R, Carlsen L (eds) Partial order in environmental sciences and chemistry. Springer, Berlin, pp 385–392

Halfon E, Reggiani MG (1986) On ranking chemicals for environmental hazard. Environ Sci Technol 20:1173–1179

Halfon E, Hodson J, Miles K (1986) An algorithm to plot Hasse diagrams on microcomputers and Calcomp plotters. Ecol Model 47:189–197

Hans Petter Langtangen (2009) A primer on scientific programming with Python (Texts in Computational Science and Engineering). Springer, Auflage: 1 (4 Aug 2009)

Hasse H (1927) Höhere Algebra II Gleichungen höheren Grades. Walter De Gruyter, vormals G.J. Göschen'sche Vetrlagshandlung, Berlin und Leipzig

Hasse H (1952) Über die Klassenzahl abelscher Zahlkörper. Akademie Verlag, Berlin

Huang IB, Keisler J, Linkov I (2011) Multi-criteria decision analysis in environmental sciences: ten years of applications and trends. Sci Total Environ 409:3578–3594

Lutz M, Ascher D (2003) Learning Python. O'Reilly, Beijing

Manganaro A, Ballabio D, Consonni V, Mauri A, Pavan M, Todeschini R (2008) The DART (Decision Analysis by Ranking Techniques) software. In: Pavan M, Todeschini R (eds) Scientific data ranking methods: theory and applications. Elsevier, Amsterdam, pp 193–207

Müller M, St. Schwarzer (2007) Python im deutschsprachigen Raum, Tagungsband zum Workshop am 8 September 2006 in Leipzig. Lehmanns Media, Berlin

Munda G (2008) Social multi-criteria evaluation for a sustainable economy. Springer, Berlin

Myers WL, Patil GP (2008) Semi-subordination sequences in multi-measure prioritization problems. In: Todeschini R, Pavan M (eds) Data handling in science and technology, vol 27. Elsevier, New York, NY, pp 161–170

Myers WL, Patil GP (2010) Preliminary prioritization based on partial order theory and R software for compositional complexes in landscape ecology, with applications to restoration, remediation, and enhancement. Environ Ecol Stat 17:411–436

Myers WL, Patil GP (2014) Higher-order indicator with rank-related clustering in multi-indicator systems. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order. Springer, New York, NY

Myers WL, Patil GP, Cai Y (2006) Exploring patterns of habitat diversity across landscapes using partial ordering. In: Brüggemann R, Carlsen L (eds) Partial order in environmental sciences and chemistry. Springer, Berlin, pp 309–325

Opperhuizen A, Hutzinger O (1982) Multi-criteria analysis and risk assessment. Chemosphere 11:675–678

Patil GP, Joshi S (2014) Comparative knowledge discovery with partial order and composite indicator. In: Brüggemann R, Carlsen L, Wittmann J (eds) Multi-indicator systems and modelling in partial order. Springer, New York, NY

Patil GP, Taillie C (1976) Ecological diversity: concepts, indices and applications. In: The Biometric Society (ed) Proceedings of the 9th biometric conference, vol II, Boston, The Biometric Society, Boston, MA, pp 383–411, 22–27 Aug 1976

Patil GP, Taillie C (2004) Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization. Environ Ecol Stat 11:199–228

Peters ML, Zelewski S (2007) TOPSIS als Technik zur Effizienzanalyse. WiSt January 2007:9–15

Pudenz S (2005) ProRank – software for partial order ranking. Match Commun Math Comput Chem 54:611–622

Saaty TL (1994) How to make a decision: the analytical hierarchy process. Interfaces 24:19–43

Thiessen RJ, Achari G (2012) Can the national classification system for contaminated sites be used to rank sites. Can J Civ Eng 39:415–431

Van de Walle B, De Baets B, Kersebaum KC (1995) Fuzzy multi-criteria analysis of cutting techniques in a nuclear dismantling project. Fuzzy Set Syst 74:115–126

Venners B (2003) The making of, Python, a conversation with Guido van Rossum, Part I. http://www.artima.com/intv/python.html. Assessed 20 Sept 2012

Voigt K, Scherb H, Brüggemann R, Schramm K-W (2011) Application of the PyHasse program features: sensitivity, similarity, and separability for environmental health data. Statistica and Applicazioni, special issue 155–168

Voigt K, Brüggemann R, Scherb H, Cok I, Mazmanci B, Mazmanci MA, Turgut C, Schramm K-W (2012) Evaluation of organochlorine pesticides in breast milk samples in Turkey applying features of the partial order technique. Int J Environ Health Res. doi:10.1080/09603123.2012.71915

Wieland R, Brüggemann R (2013) Hasse Diagram technique and Monte Carlo simulations. Match Commun Math Comput Chem 70:45–59

Wienand O (2012) http://bio.math.berkeley.edu/ranktests/lcell/. Assessed 6 Nov 6

Winkler P (1982) Average height in a partially ordered set. Discrete Math 39:337–341

Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybern 18:183–190

Yager RR (1993) Families of OWA operators. Fuzzy Set Syst 59:125–148

Yevtushenko SA (2003) Concept explorer the user guide: system of data analysis. Concept Explorer 127–134

Index

R. Brüggemann et al. (eds.), *Multi-indicator Systems and Modelling in Partial Order*, DOI 10.1007/978-1-4614-8223-9, © Springer Science+Business Media New York 2014