Assessment of the environmental acceptability of refrigerants by discrete mathematics: cluster analysis and Hasse diagram technique

A dissertation submitted to the
Faculty of Biology, Chemistry and Geosciences
University of Bayreuth
Germany

to attain the academic degree of
Dr. rer. nat.

presented by
Guillermo Restrepo
M. Sc.

born August 12, 1976
in Bogotá, Colombia

Supervisors:
1. Prof. Dr. Hartmut Frank
2. Dr. Rainer Brüggemann

Bayreuth, February 14, 2008

Assessment of the environmental acceptability of refrigerants by discrete mathematics: cluster analysis and Hasse diagram technique

By
Guillermo Restrepo
Environmental Chemistry and Ecotoxicology
University of Bayreuth
Bayreuth
Germany

This doctoral thesis was funded by

COLCIENCIAS - Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología “Francisco José de Caldas” (Colombia)

and the

Universidad de Pamplona
(Colombia)

Vollständiger Abdruck der von der Fakultät für Biologie, Chemie und Geowissenschaften der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.).

Gefördert durch COLCIENCIAS – Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología "Francisco José de Caldas" (Kolumbien) und die Universidad de Pamplona (Kolumbien).

Promotionsgesuch eingereicht am: 24. Oktober 2007

Tag des wissenschaftlichen Kolloquiums: 14. Februar 2008

Prüfungsausschuss:

Prof. Dr. Hartmut Frank (Erstgutachter)
Prof. Dr. Adalbert Kerber (Zweigutachter)

Prof. Dr. Stefan Peiffer (Vorsitzender)
Prof. Dr. Klaus Bitzer
Prof. Dr. Dieter Brüggemann

Die vorliegende Arbeit entstand im Zeitraum von Dezember 2005 bis Oktober 2007 und wurde am Lehrstuhl für Lehrstuhl für Umweltchemie & Ökotoxikologie der Universität Bayreuth unter der Anleitung von Prof. Dr. Hartmut Frank angefertigt.

Contents

Summary

The recognition of the adverse environmental impact of chlorofluorocarbons (CFCs), mainly used as refrigerants, has lead to look for environmentally acceptable CFC replacements. Main environmental concern CFCs face is their ability to deplete the stratospheric ozone layer, quantified by the ozone depletion potential (ODP). Some of the first replacements mooted were hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs), which contribute to the global warming, quantified by the global warming potential (GWP). ODP and GWP are related to the atmospheric lifetime (ALT), a third indicator. Hence, the environmental impact of a refrigerant may be characterised by a triple of ODP, GWP and ALT values. In this respect, an acceptable refrigerant is a chemical with low ODP, GWP and ALT values.

One of the first steps to assess the environmental acceptability of refrigerants is to classify them in order to find classes of substances sharing common features. In this respect, a supervised and unsupervised classification was performed over 40 refrigerants used in the past, presently used and some proposed substitutes. First one was a classification based upon elemental composition and functional groups present in refrigerant molecules which leads to different substance families: CFCs, HCFCs, HFCs, hydrocarbons (HCs), hydrofluoro ethers (HFEs), chloromethanes (CMs) and single refrigerants like carbon dioxide, trifluoroiodomethane, dimethyl ether and ammonia. The unsupervised classification was performed using hierarchical cluster analysis. In this case, refrigerants were characterised according to three kinds of descriptors: Environmental properties (ODP, GWP, ALT), thermodynamic features related to their refrigeration performance and molecular descriptors derived from their molecular structure. Eight clustering methodologies were applied to each kind of refrigerant descriptors. In order to assess the stability of these classifications, the cluster index, a method for quantifying the resemblance between pairs of classifications was developed and further applied to refrigerant classifications. Results showed that the environmental descriptors are the only case in which refrigerant classes formed are stable when varying the classification method.

The chemotopological procedure, a method for studying similarity relationships, was applied to the environmental classification of refrigerants. It was found that CFCs are similar to themselves and also to 1,1,1,3,3,3-hexafluoropropane, a HFC. The most similar substances to all CFCs considered were trichlorofluoromethane and 1,1,2-trichloro-1,2,2-trifluoroethane. The other refrigerant families were found to be similar to many other substances, therefore there is no clear affiliation of refrigerants of one family to one certain class.

It was found a disagreement between the supervised classification leading to refrigerant families and the three unsupervised classifications (environmental, thermodynamic and molecular ones). Therefore, refrigerant classification into families does not imply same classification based upon environmental properties, thermodynamic features and molecular descriptors of the refrigerants considered.

A different refrigerant classification was performed, namely the one based upon order relationships of refrigerant environmental properties. In this case the Hasse diagram technique, a method based on partial order theory, was applied to the 40 refrigerants characterised by environmental properties. A parameter free procedure for ordering classes based upon order relationships of their elements was developed. For that purpose, the dominance and separability degrees were introduced, first one indicates the extent to which members of one class hold higher descriptor values than the members of another class; while separability degree quantifies the lack of order relationships between two classes. Dominance and separability degrees were related by a theorem. By the application of dominance and separability degrees to refrigerant families three main classes were detected: problematic substances, gathering CFCs, octafluorocyclobutane and bromochlorodifluoromethane; least problematic ones, collecting HCs, CMs, carbon dioxide, trifluoroiodomethane, dimethyl ether and ammonia; and moderately problematic refrigerants, made from HCFCs, HFCs and HFEs. It was found that some HFEs are not dominated by CFCs, which raises the question on the applicability of these substances as environmentally acceptable replacements.

METEOR (Method of evaluation by order theory), a procedure for prioritising descriptors and studying its effect on the order relationships of the objects considered was discussed. When applied to the refrigerants, the effect of prioritising ODP, GWP and ALT in the order relationships of these substances was studied. It was found that pentafluorodimethyl ether, a HFE, is one of the most problematic refrigerants under a large range of priorities of the environmental properties considered.

Due to the mathematical generality of the methods here introduced, they are not restricted to the analysis of refrigerants but can be used to the study of different sets whose elements are characterised by various attributes.

Zusammenfassung

Die wissenschaftliche Aufdeckung der umweltschädlichen Einflüsse der Fluorchlorkohlenwasserstoffe (CFC*), die vornehmlich als Kältemittel Verwendung fanden, führte dazu dass nach umweltverträglichen CFC-Ersatzstoffen gesucht wurde. Das größte Umweltproblem der CFC ist deren Fähigkeit, die stratosphärische Ozonschicht zu zerstören. Die Stärke eines Stoffes zur Ozonschichtschädigung wird durch das Ozonzerstörungspotential (ODP) quantifiziert. Einige der ersten Ersatzstoffe waren die teilhalogenierten Fluorchlorkohlenwasserstoffe (HCFC) und die teilfluorierten Kohlenwasserstoffe (HFC), die jedoch zum Treibhauseffekt beisteuern. Der Beitrag zum Treibhauseffekt wird durch das Treibhauspotential (GWP) beschrieben. ODP und GWP sind mit der atmosphärischen Lebensdauer (ALT), einem dritten Indikator, verbunden. Das Umweltverhalten eines Kältemittels kann durch die drei Indikatoren ODP, GWP und ALT charakterisiert werden. Ein umweltfreundliches Kältemittel ist ein Stoff mit niedrigen ODP, GWP und ALT Werten.

Einer der ersten Schritte, um die Umweltverträglichkeit von Kältemitteln abzuschätzen, ist deren Klassifizierung, um Klassen von Substanzen zu finden, die gemeinsame Merkmale aufweisen. Diesbezüglich wurde eine überwachte und nicht überwachte Klassifizierung an 40 Kältemitteln durchgeführt. Die Gruppe von 40 Kältemitteln besteht aus Kältemitteln, die in der Vergangenheit eingesetzt wurden, die derzeit verwendet werden und solche, die als Ersatzstoffe vorgeschlagen werden. Die erste Klassifizierung war eine Klassifizierung, die auf der elementaren Zusammensetzung und den funktionellen Gruppen innerhalb der molekularen Struktur der Stoffe beruht und zu unterschiedlichen Familien von Stoffen führt: CFC, HCFC, HFC, Kohlenwasserstoffe (HC), teilfluorierte Ether (HFE), Chlormethane (CM) und einzelne Kältemittel wie Kohlenstoffdioxid, Trifluorjodmethan, Dimethylether und Ammoniak. Bei der nicht überwachten Klassifizierung wurde die hierarchische Clusteranalyse eingesetzt. Hierbei wurden die Kältemittel anhand von drei Kategorien von Deskriptoren charakterisiert: Umwelteigenschaften (ODP, GWP, ALT), thermodynamischen Eigenschaften bezüglich ihres Kühlverhaltens und molekulare Deskriptoren, die sich aus ihrer Molekülstruktur ergeben. Acht Cluster-Methoden wurden auf jede Gruppe von Kältemittel-Deskriptoren angewendet. Zur Bewertung der Stabilität dieser Klassifikationen wurde der Cluster Index, eine Methode zur Quantifizierung der Ähnlichkeit von Klassifikationspaaren, entwickelt und auf die Klassifikation von Kältemitteln angewendet. Die Ergebnisse zeigen, dass lediglich bei der Verwendung der Deskriptoren der Umwelteigenschaften die Kältemittel Klassen bilden, die stabil gegenüber der Variation der Klassifikationsmethode sind.

Das chemotopologische Verfahren, eine Methode zur Untersuchung von Ähnlichkeitsbeziehungen, wurde auf die Klassifizierung der Umwelteigenschaften der Kältemittel angewendet. Die Ergebnisse

zeigen, dass die Moleküle der CFC-Klasse zu sich selber und auch zu 1,1,1,3,3,3-Hexafluorpropan, einem HFC, ähnlich sind. Die Substanzen Trichlorfluormethan und 1,1,2-Trichlor-1,2,2-trifluorethan weisen die größte Ähnlichkeit zu allen in der Studie betrachteten CFC auf. Die anderen Kältemittelfamilien zeigten eine Ähnlichkeit zu vielen anderen Substanzen. Demzufolge ist eine eindeutige Zuordnung der Kältemittel einer Familie zu einer bestimmten Klasse nicht möglich.

Es ergab sich ein Widerspruch zwischen der überwachten Klassifikation, die zu Kältemittelfamilien führt, und den drei nicht überwachten Klassifikationen (Umwelteigenschaften, thermodynamische Eigenschaften und Molekularstrukturen). Demzufolge, impliziert eine Klassifikation in Familien nicht automatisch die gleiche Klassifikation, wenn diese auf Umwelteigenschaften, thermodynamischen Eigenschaften und molekularen Deskriptoren der untersuchten Kältemittel beruht.

Eine weitere Kältemittelklassifikation wurde durchgeführt, die auf Ordnungsbeziehungen der Umwelteigenschaften der Kältemittel beruht. Hierbei wurde die Hasse Diagramm Technik, eine Methode die auf der Theorie partiell geordneter Mengen beruht, auf die 40 Kältemittel angewendet, die durch ihre Umwelteigenschaften charakterisiert wurden. Ein parameterfreies Verfahren zur Ordnung der Klassen basierend auf den Ordnungsbeziehungen ihrer Elemente wurde entwickelt: Hierzu wurden Dominanz- und Trennbarkeitsgrade eingeführt. Der Dominanzgrad quantifiziert den Umfang, in dem die Elemente der einen Klasse diejenigen der anderen dominieren. Der Trennungsgrad hingegen quantifiziert den Mangel an Ordnungsbeziehungen zwischen zwei Klassen. Dominanz- und Trennungsgrade wurden anhand eines Theorems in Beziehung zueinander gesetzt. Bei der Anwendung des Dominanz- und Trennungsgrades auf die Kältemittelfamilien konnten drei Hauptklassen ausfindig gemacht werden: problematische Stoffe, die CFC, Oktafluorcyclobutan, und Bromchlordifluormethan einschließen, wenig problematische Stoffe, wie HC, CM, Kohlenstoffdioxid, Trifluorjodmethan, Dimethylether und Ammoniak, und mäßig problematische Kältemittel, wie HCFC, HFC und HFE. Es zeigte sich dass einige HFE nicht durch CFC dominiert werden, was die Frage hinsichtlich ihrer Akzeptanz als umweltverträgliche Ersatzstoffe aufwirft.

METEOR, ein Verfahren für die Priorisierung von Deskriptoren und die Untersuchung ihrer Auswirkung auf die Ordnungsbeziehung der untersuchten Objekte wurde diskutiert. Auf die Kältemittel angewendet, wurde die Auswirkung der Priorisierung von ODP, GWP und ALT auf die Ordnungsbeziehung dieser Stoffe untersucht. Es zeigte sich, dass Pentafluordimethylether, ein HFE, innerhalb einer breiten Prioritätsspanne der betrachteten Umwelteigenschaften eines der problematischsten Kältemittel ist.

Aufgrund der mathematischen Allgemeingültigkeit der hier eingeführten Methoden sind diese nicht auf die Bewertung von Kältemittel beschränkt, sondern können zur Untersuchung verschiedener multivariat charakterisierter Objekte eingesetzt werden.

*) In der deutschen Zusammenfassung werden die englischen Abkürzungen verwendet.

List of manuscripts and author's contribution

This dissertation is presented in cumulative form. It comprises eight individual manuscripts, from which seven are published and one is in press. Author's contribution to each manuscript is given below.

Published

Restrepo, G.; Mesa, H.; Llanos, E. J. Three dissimilarity measures to contrast dendrograms. *J. Chem. Inf. Model.* **2007,** *47,* 761-770. (**Appendix A**)
Own contribution. Idea (50 %), calculations (40 %), writing (90 %).

Restrepo, G.; Brüggemann, R. Modelling the fate of alkanes in rivers. In *Recent progress in computational sciences and engineering*; Simos, T.; Maroulis, G., Eds.; VSP: Leiden, Netherlands, 2006; pp 1386-1389. (**Appendix B**)
Own contribution. Idea (30 %), calculations (100 %), writing (90 %).

Restrepo, G.; Brüggemann, R. Partially ordered sets in the analysis of alkanes fate in rivers. *Croat. Chem. Acta* **2007**, *80*, 261-270. (**Appendix C**)
Own contribution. Idea (30 %), calculations (100 %), writing (70 %).

Restrepo, G.; Weckert, M.; Brüggemann, R.; Gerstmann, S.; Frank, H. Refrigerants ranked by partial order theory. In *EnviroInfo 2007, 21st international conference on informatics for environmental protection*; Hryniewicz, O.; Studziński, J.; Szediw, A., Eds.; Shaker: Aachen, Germany, 2007; pp 209-217. (**Appendix D**)
Own contribution. Idea (70 %), calculations (100 %), writing (70 %).

Restrepo, G.; Brüggemann, R. Dominance and separability in posets, their application to isoelectronic species with equal total nuclear charge. *J. Math. Chem.* doi: 10.1007/s10910-007-9331-x (**Appendix E**)
Own contribution. Idea (60 %), calculations (100 %), writing (70 %).

Brüggemann, R.; Voigt, K.; Restrepo, G.; Simon, U. Concept of stability fields and hot spots in ranking of environmental chemicals. *Environ. Modell. Softw*. doi: 10.1016/j.envsoft.2007.11.001 (**Appendix F**)

Own contribution. Idea (50 %), calculations (60 %), writing (50 %).

Restrepo, G.; Brüggemann, R.; Weckert, M.; Gerstmann, S.; Frank, H. Ranking patterns, an application to refrigerants. *MATCH Commun. Math. Comput. Chem.* **2008**, *59*, 555-584. (**Appendix G**)
Own contribution. Idea (80 %), calculations (100 %), writing (70 %).

In press

Restrepo, G.; Weckert, M.; Brüggemann, R.; Gerstmann, S.; Frank, H. Ranking of refrigerants. *Environ. Sci. Technol.* (**Appendix H**)
Own contribution. Idea (70 %), calculations (100 %), writing (70 %).

Abbreviations

AFAE	Alkylfluoroalkylether
ALT	Atmospheric lifetime
BCF	Bromochlorofluorocarbon (Bromochlorodifluoromethane)
CFC	Chlorofluorocarbon
CM	Chloromethane
DFAE	Di(fluoroalkyl)ethers
DME	Dimethyl ether
FIM	Fluoroiodomethane (Trifluoroiodomethane)
GWP	Global warming potential
HC	Hydrocarbon
HCA	Hierarchical cluster analysis
HCFC	Hydrochlorofluorocarbon
HD	Hasse diagram
HDT	Hasse diagram technique
HFC	Hydrofluorocarbon
HFE	Hydrofluoroether
KIF	*K* inflation factor
METEOR	Method of evaluation by order theory
MW	Molecular weight
ODP	Ozone depletion potential
PFC	Perfluorocarbon (Octafluorocyclobutane)
QSAR	Quantitative structure-activity relationships
QSPR	Quantitative structure-property relationships
TH	Time horizon

Chapter 1: Refrigeration and refrigerants

1.1 Refrigeration

Refrigeration technology has changed with time; from early icehouses [1] to modern mechanical refrigerators [2]. Nowadays, the most widespread refrigeration method used in dwellings and automobiles is based upon the vapour-compression procedure, conceived by Cullen in 1748 [3] and further improved along the history. A brief description of the processes involved is given in the following.

1.1.1 *Vapour-compression system*

Four fundamental processes are included in this procedure, namely vaporisation, compression, condensation and expansion [2], in which a working fluid, called refrigerant, alternatively absorbs and releases energy experiencing changes in its pressure, temperature and/or phase. The sequence of these processes is depicted in Figure 1.1.

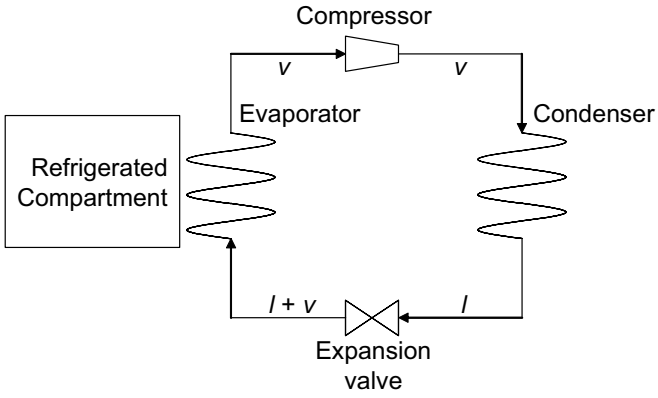


Figure 1.1. Scheme of a vapour-compression system; *v* and *l* stand for vapour and liquid, respectively.

Evaporator. It is a heat transfer coil that allows heat to be conducted from the compartment to be cooled to the refrigerant, which initially enters the evaporator as a liquid/vapour mixture with relatively low pressure and low temperature. During its transit through the evaporator, the absorption of latent heat from the thermal energy of the air in the refrigerated space turns the liquid refrigerant into vapour and causes a decrease of air temperature in the refrigerated compartment, that is the desired refrigeration effect.

Compressor. It is an electromechanical device used to develop and maintain the flow of the refrigerant vapour through the refrigeration system. In this compartment refrigerant's pressure is raised through a kinetic energy transfer that also raises vapour temperature to a level that further permits condensation at normal ambient temperatures.

Condenser. It is another heat transfer coil used to conduct heat from the hot refrigerant vapours leaving the compressor to the ambient surrounding the coil. This process allows the vapours to condense into their liquid phase delivering the latent heat of vaporisation absorbed in the evaporator and also the additional energy taken in the compressor.

Expansion valve. This device reduces the pressure of the liquid refrigerant causing the adiabatic evaporation of part of the liquid refrigerant, which drops the refrigerant temperature making it colder than the space to be refrigerated during its transit in the evaporator.

1.1.2 *Thermodynamic refrigerant requirements*

A substance must meet the following thermodynamic features in order to be considered as a refrigerant fluid in a vapour-compression system:

- High latent heat of vaporisation, which means that the substance must be able to absorb high amounts of energy during its change of phase from liquid to vapour.

- Low vapour specific volume, meaning that determined mass of refrigerant must occupy a reduced space.

- Low refrigerant specific heat in its liquid phase, whereas high values in its vapour phase.

First two conditions permit that the substance absorbs high amounts of energy without expanding to a big extent, therefore the energy used in the compression is low making the process energy efficient as well as reducing the size of the compressor.

First requirement of third condition guarantees that the refrigerant needs low amounts of energy to increase its temperature, which favours its vaporisation. Second requirement implies that the vapour needs large amounts of energy to rise its temperature, which makes the vapour condensation before approaching the compressor and the condenser more unlikely. These constraints on refrigerant specific heat favour the refrigeration effect and the efficiency of the process.

The form in which molecules arrange is a determining factor in the volume reached by a given amount of substance. In this situation the molecular shape is a feature determining this molecular packing. The heat of vaporisation and the specific heat are related to the energy involved to separate molecules, which is determined by the electrostatic forces between them. This situation is rather determined by the electronic density distribution on the molecules.

1.1.3 *Technical, safe and environmental refrigerant requirements*

Besides the physical properties mentioned before a refrigerant must fulfil the following conditions related to its technical use, environmental effects and end user security [2].

- Non-toxicity neither in its pure state nor when mixed with air. It must not contaminate products stored in the refrigerated compartment.

- Non-flammability neither in its pure state nor when mixed with air. It must also be non-explosive.

- Non-reactivity neither with the lubricating oil in the system nor with any material used in the equipment.

- Non-reactivity with the moisture present to some degree in all refrigerant systems.

- Economically suitable for its large scale production and environmentally safe, neither depleting stratospheric ozone layer nor increasing earth's temperature. Furthermore, its atmospheric reaction products must meet same environmental and toxicological requirements.

1.2 Refrigerants

1.2.1 *Chlorofluorocarbons, sunrise and sunset*

Prior to the 1930s, refrigerants commonly used were ammonia, chloromethane, carbon tetrachloride, isobutane and propane [4, 5]. Because of concerns about their toxicity and flammability, particularly in the home environment, Midgley and co-workers at General Motors started investigations to identify suitable replacement materials [6-8]. The requirements for volatility, low toxicity, stability and non-flammability led the research to concentrate on new substances based on some elements of groups 14-17 of the periodic table, namely C, N, O, F, S, Cl, Br and also H [6]. Further considerations on

flammability and toxicity permitted to anticipate the importance of fluorine for the desired substance. Hence, in 1930 Midgley's team came up with dichlorodifluoromethane [9, 10], the first of a series of chlorofluorocarbons (CFCs) which found, between 1930s and 1990s, not only applications as refrigerants but also as blowing agents for making foam, as cleaning fluids and as propellants [7, 8]. Their production and releases remained comparatively low until the 1950s, then they increased rapidly with refrigeration spreading in the developed world and as a consequence of their new applications.

There are several reasons for CFC applicability; some of them [7] present high vapour specific heats, high latent heats of vaporisation and low liquid specific heats which make them quite appropriate for refrigeration in thermodynamic terms. Some others hold low thermal conductivities and low permeation rates [11] making them suitable for insulating foam. Other CFCs [7] have low surface tension and low viscosity, ideal properties for cleaning agents because they can wet even tiny spaces easily; their high vapour densities guarantee no significant losses of the cleaning agent through evaporation. Because CFCs are non-toxic and non-flammable, they are safe to use in consumer applications. Additionally, CFCs can be easily produced on a large scale [12] in high purity. Unfortunately, one of their advantages, i.e. CFCs are extremely stable, has disastrous atmospheric consequences [13], which are treated in the ensuing discussion.

In the beginning of the 1970s, Lovelock and co-workers demonstrated that CFCs were trace constituents in the atmosphere [14-16]. By 1972, Dupont initiated a series of meetings with CFC manufacturers to discuss the environmental fate of these substances. McCarthy summarised the conclusions of that meeting in this way: "Fluorocarbons are intentionally or accidentally vented to the atmosphere world-wide at a rate approaching one billion pounds per year. These compounds may be either accumulating in the atmosphere or returning to the surface, land or sea, in pure form or as decomposition products. Under any of these alternatives it is prudent that we investigate any effects which the compounds may produce on plants or animals now or in the future" [8].

Lovelock's observations brought Molina and Rowland to determine the ultimate atmospheric fate of CFCs and in 1974 they argued that these substances could destroy stratospheric ozone [17, 18]. Their arguments were based on the inexistence of tropospheric sinks, their poor dissolution and oxidation in raindrops whereby they concluded that the only significant sink was solar ultraviolet photolysis in the stratosphere, producing chlorine atoms as one of the reaction products. They also explored the fate of these chlorine atoms and concluded that they react with ozone yielding oxygen and more chlorinated atoms as final products of a series of reactions through chlorine oxides. Hence, Molina and Rowland pointed out the threat to the ozone layer caused for these widespread substances.

After confirming these results, each CFC manufacturer initiated its own research programme to look for CFC replacements which keep the advantages of CFCs and could be used in the current equipment. Replacements were sought having properties close to CFC ones, which included non-flammability, non-toxicity, miscibility with acceptable lubricants, thermodynamic properties as close to CFC original refrigerants and environmentally acceptable properties [8].

Simultaneously, several countries unilaterally banned the use of CFCs in most aerosols but they were still used in applications such as cooling systems. In 1984 Farman and co-workers discovered a remarkable and totally unusual phenomenon, the so-called "ozone hole" [19] making CFCs the prime suspect. After several meetings and discussions, 24 countries negotiated the Montreal Protocol on Substances that Deplete the Ozone Layer in September 1987 [20], which originally mandated a 50% reduction in CFC production and consumption by 1 July 1999. Subsequently, it has been modified as the result of additional scientific investigations and nowadays is ratified by 165 countries [21].

Around 1990, global warming resulting from the release of anthropogenic gases became a major environmental concern. Although one of the largest contributors was and still is carbon dioxide from the burning of fossil fuels, it was estimated that CFCs accounted for 15% of global warming in the early 1980s [22, 23].

Considering all these aspects, Midgley's pool of elements was reduced [6]. Bromine is excluded on environmental grounds because of the high potential to deplete the ozone layer associated to its compounds; chlorine, although less problematic in this respect, can be problematic if its compounds remain long time in the atmosphere; sulphur substances are likely to hold high toxicities, so the set shrinks to H, C, N, O and F. Remarkably, these elements allow ammonia and hydrocarbons (HCs) as possible CFCs replacements, which were the toxic and flammable substances Midgley wanted to replace initially. During the 1980s, industry proposed as potential CFC alternatives a group of compounds from C, H, F and Cl, which reach a compromise among all replacement requirements; thus, hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs) showed up.

1.2.2 *Developing atmospheric environmental indicators*

Concerns about the potential of anthropogenic chemicals to alter the earth's global atmospheric environment led to the development of measures for comparing and quantifying the lifetimes of various compounds in the atmosphere as well as their effects on the stratospheric ozone layer and on the radiative balance of the atmosphere. In the ensuing discussion a brief description is given of these measures, which are amply described in references 24 and 25.

Atmospheric lifetime (ALT). The global atmospheric lifetime (τ_{RH}^{global}) of a gas RH characterises the time required to turn over the global atmospheric burden [26]. The lifetime depends on the chemistry and dynamics of the atmosphere, therefore it may depend on the location of the sources [25, 27]. For gases in steady state, τ_{RH}^{global} is calculated as follows

$$\tau_{RH}^{global} = \frac{C_{RH}^{global}}{L_{RH}^{global}} = \frac{C_{RH}^{global}}{P_{RH}^{global}} \qquad 1.1$$

where C_{RH}^{global} is the RH global atmospheric burden and L_{RH}^{global} and P_{RH}^{global} are the burden loss and production terms, respectively. Normally, production rates do not depend on the RH concentration in the atmosphere. In contrast, loss rates do depend on it and this relationship can even be proportional to the *n*-th power of the concentration, with $n \neq 1$ [25]. However, general lifetime calculations do not consider those cases and the RH losses associated to different atmospheric *j* removal processes of an atmospheric *i* region are regarded as first-order removal processes [24, 25]. Hence, the RH losses of a region *i*, L_{RH}^{i} , are described by

$$L_{RH}^{i} = \sum_{j} k_{j}^{i} C_{RH}^{i} \qquad 1.2$$

where k_{j}^{i} represents the removal process *j* within the region *i* and C_{RH}^{i} the RH burden in *i*. Therefore, the RH lifetime in the global atmosphere can be calculated as

$$\left(\tau_{RH}^{global}\right)^{-1} = \frac{\iiint \sum_{j} k_{j}(x,y,z) C_{RH}(x,y,z)\,dx\,dy\,dz}{C_{RH}^{global}} := \sum_{j} \left(\tau_{RH}^{j}\right)^{-1} \qquad 1.3$$

where τ_{RH}^{j} is the RH lifetime caused by its removal through the *j* process calculated for the global atmosphere.

In general, for well-mixed RH gases, $(\tau_{RH}^{global})^{-1}$ can be calculated through

$$(\tau_{RH}^{global})^{-1} = (\tau_{RH}^{trop})^{-1} + (\tau_{RH}^{strat})^{-1} \qquad 1.4$$

This equation is an example of the resistance approach, in which a total kinetic process is considered as an electrical network in which each one of the single subprocesses has associated a resistance [27].

Ozone depletion potential (ODP). The ozone depletion potential (ODP_{RH}) of a gas RH is the relative amount of degradation to the ozone layer RH can cause, with trichlorofluoromethane (R11) used as reference whose ODP is set to 1.0. Hence, ODP_{RH} is calculated as follows [29-31].

$$ODP_{RH} = \frac{\Delta O_3 \text{ for emission of a unit mass of RH}}{\Delta O_3 \text{ for emission of a unit mass of R11}} \qquad 1.5$$

This kind of calculation assumes that RH reacts in the stratosphere, as does trichlorofluoromethane (R11) and the CFCs, therefore it is suitable for ODP calculations of CFCs or stratospheric reactive gases. However, it has been applied to ODP calculations of gases that also react through the atmosphere and not uniquely in the stratosphere.

Another problem with the (steady-state) ODP definition in Eq. 1.5 stems from the fact that the relative effect of a gas emission on stratospheric ozone changes with time, which occurs because chemicals with different lifetimes accumulate at different rates in the atmosphere. This has required the definition of specific time horizons for model calculations of ODPs [30, 32]. Time dependent ODPs can be used to provide an indication of the effect on the ozone layer of a mix of compounds with different lifetimes; they are calculated using

$$ODP_{RH} = \left(\frac{F_{RH}}{F_{R11}}\right)\left(\frac{M_{R11}}{M_{RH}}\right)\left(\frac{n_{RH}}{3}\right)\alpha\left(\frac{\tau_{RH}^{global}}{\tau_{R11}^{global}}\right)\left(\frac{1-\exp\left(-[t-t_s]/\tau_{RH}^{global}\right)}{1-\exp\left(-[t-t_s]/\tau_{R11}^{global}\right)}\right) \qquad 1.6$$

where F_{RH} / F_{R11} is the measured fraction of RH injected into the stratosphere that has been dissociated relative to that of R11 [33]; τ_{RH}^{global}, τ_{CFC-11}^{global}, M_{RH} and M_{CFC-11} are the atmospheric lifetimes and molecular weights of RH and R11, respectively; n_{RH} is the number of chlorine or bromine atoms in a RH molecule; t_s is the time required to transport a RH molecule from the troposphere to the stratosphere region under consideration; and α is a factor required for RH containing bromine in order to account for the higher efficiency of this atom to catalyze ozone loss compared to chlorine.

Global warming potential (GWP). The global warming potential (GWP_{RH}) of a gas RH is the ratio of the time-integrated radiative forcing from the instantaneous release of 1 kg of RH relative to that of CO_2 [34, 35], and it is given by

$$GWP_{RH}(TH) = \frac{\int_0^{TH} a_{RH}[RH](t)dt}{\int_0^{TH} a_{CO_2}[CO_2](t)dt} = \frac{\int_0^{TH} a_{RH} M_{CO_2} \exp\left(-[t/\tau_{RH}^{global}\right)dt}{\int_0^{TH} a_{CO_2} M_{RH}[CO_2](t)dt} \qquad 1.7$$

where TH is the time horizon over which the calculation is performed, a_{RH} and a_{CO_2} are the radiative forcings per unit mass of RH and CO_2, respectively; their units are (W m^{-2} kg^{-1}). M_{RH}, M_{CO_2}, $[RH](t)$ and $[CO_2](t)$ are the molecular weights and abundance time-dependent decays after pulsed emissions of RH and CO_2, respectively.

Importance and relationships of ALT, ODP and GWP. Two additional comments are important regarding ALT, ODP and GWP. First one is related to their relevance in designing new refrigerants. An environmentally acceptable refrigerant must fulfil the thermodynamic and economic constrains required for its use but to avoid CFC problems; it is mandatory it holds low ALT, ODP and GWP values. Otherwise, its use could bring even worse results that those brought by CFC use.

Second comment concerns the relationship between ALT and both ODP and GWP. This can be seen in Eqs. 1.6 and 1.7 which predict that an ozone depleting substance with high ALT will persist in the atmosphere increasing its probability of degradate the ozone layer. Similarly, a chemical with high radiative forcing and high ALT will remain long time in the atmosphere increasing its potential to absorb infrared radiation and therefore significantly contribute to the global warming.

In the next sections some possible CFC replacements are discussed based upon these environmental indicators as well as some other technical and environmental aspects.

1.2.3 *CFC replacements*

Hydrofluorocarbons (HCFCs). One of the justifications for using HCFCs is that, in spite of containing chlorine, the presence of hydrogen makes them suitable for hydroxyl reactions in the troposphere [23]. Therefore, their potential to react with stratospheric ozone is reduced in comparison with that of CFCs. Some of the HCFCs reaction products are hydrogen chloride and hydrogen fluoride but according to estimations on HCFCs production the environmental impact of these acids is not likely to have any significance [34]. In contrast, HCFCs reaction product trifluoroacetic acid constitutes a major environmental problem. This substance, a strong carboxylic acid, can irritate tissue and skin and its concentration, ranging from 30 to 40,000 ng L^{-1} in rivers and lakes from around the world [37-41], already exceeds those estimated for 2010 having taken into account all anthropogenic

sources [42]. This suggests that a ‘‘natural’’ source of trifluoroacetate exists [43]. Harnisch and co-workers [44] have speculated that it could have a geological origin but clear evidence is still missing.

Although HCFCs are degraded in the troposphere, a small fraction of them can reach the stratosphere whereby they are still a problem taking into account the cascade of ozone reactions a single chlorine atom can produce. For this reason HCFCs were included in the Montreal Protocol and their phase-out is scheduled by 2020 [45]. Additionally, it has been shown that HCFCs contribute to the global warming [36].

Hydrofluorocarbons (HFCs). Since the early 1990s, HFCs have been used as acceptable alternatives to CFCs and also to HCFCs because HFCs hold several favourable characteristics [46] including near-zero ODPs [47], similar physical properties as CFCs and HCFCs. Beyond this, HFCs have short atmospheric lifetimes, are less- or non-flammable and their industrial production is not expensive [48]. Because of this, problematic HCFCs will be replaced by HFC-blends in refrigeration equipment before 2010 [49]. However, HFCs pose some environmental hazards and health risks. Their high volatility and very low solubility in water [46] make them mainly reside in the atmosphere where they photolytically react forming hydrogen fluoride and trifluoroacetic acid among other substances. Because HFCs have relatively high GWPs [50, 51] they were included in the Kyoto Treaty (nowadays Kyoto Protocol) together with carbon dioxide, methane, perfluorocarbons (PFCs), sulphur hexafluoride and nitrous oxide [52]. Countries signing this treaty committed to achieve a specific level of total global warming gas emissions in 2010 based on their emissions in 1992. These environmental problems made that new alternatives were mooted in the late 1980s and early 1990s [8]; some of the proposed substances were fluorinated ethers [53-56].

Hydrofluoroethers (HFEs). By the late 1980s, the US Environmental Protection Agency promoted investigations to synthesize fluorinated dimethyl ether derivatives and assess their properties [8]. Because of the attractive properties of several HFEs, particularly high volatility and hydrophobicity compared to similar chemicals such as saturated fluorocarbons and ethers, they began to be successfully developed in the mid of 1990s [56]. According to these properties the atmosphere is the most likely place for their emissions and one of their main degradation products is carbonyl fluoride (COF_2), an irritating gas which is easily hydrolysed to - for humans only moderately toxic – hydrogen fluoride [57]. The main concern regarding HFEs is related to the high GWP of some of them currently used as CFC replacements. Further research is conducted on these substances and it is expected that new HFEs may replace the problematic HFEs presently used.

Why not “natural” refrigerants? It is thought that the so-called ‘‘natural’’ refrigerants such as hydrocarbons (HCs), carbon dioxide and ammonia might be appropriate refrigerants for current

necessities [8]. According to Powell, a ‘‘natural’’ substance is one whose presence in the environment is the result of biological or geological processes; however, their commercial use for refrigeration implies their extraction from non-renewable sources: HCs from oil cracking; ammonia and carbon dioxide from natural gas. Furthermore, peak oil [58, 59] and the nearness of a peak in natural gas production [59] make this option intractable. It is remarkable that all these refrigerants were extensively used until the rapid growth of CFCs and HCFCs after 1945 and some of them are still used [8], for example ammonia has remained as the preferred refrigerant in large-scale food freezing plants, as well as some hydrocarbons.

If natural refrigerants are still in use and without considerable problems, then it is still possible to keep asking for the lack of their widespread application after the recognition of CFCs environmental problems. Powell [8] answers this question pointing out that the use of these “natural” refrigerants concerns mainly refrigerator/freezer systems, which surprisingly only account for 4% of the total use of refrigerants. The largest refrigerant application is automobile air-conditioning and the introduction of natural refrigerants in those systems is under discussion [8]. Carbon dioxide is being considered by some auto-manufacturers in Germany, although the very high pressures associated to its use require a radical engineering redesign [8]. Various hydrocarbon compositions have been offered both as retrofit replacements for CFCs used in vehicles, but some tests have shown that if these fluids escaped into the passenger compartment during an accident and ignited, the resulting explosion would cause serious injury [8]. This is the reason why in USA and in some Australian states HCs are banned from mobile air-conditioning units.

Carbon dioxide could be used in vehicles but some tests suggest that it is not energy efficient as some HFCs at high ambient temperatures wherein the air conditioning system is mainly needed; additionally, if the gas reached the cabin it would cause physiological effects that could be worse than those of most used HFC, e.g. CF_3CH_2F (R134a). On the other hand, regulations normally require that refrigeration systems, in direct contact with the general public, must not contain hazardous refrigerants [8].

A possible solution to this shortcoming can be the installation of secondary circuits containing glycol or calcium chloride brine to transport the ‘‘coolth’’ from the refrigeration plant to the building or display unit. At the end, this generates more CO_2 with serious global warming consequences [60]. In short, the simple replacement of current refrigerants by natural ones would bring little impact or in some cases could increase global warming [8].

1.3 Research purpose

As shown in the previous sections, the selection of appropriate replacements is not a simple issue since various factors need to be considered to make a decision. In the present dissertation these factors are analysed using elements of partial order theory, whose application to chemistry and environmental sciences is grounded on the comparison of the attributes characterising the objects to study, in the current case, refrigerant features. As Brüggemann has stated, partial order theory in its application aspects is the science of comparisons. This dissertation deals with comparison of refrigerant features as a mathematical tool supporting the environmental assessment of refrigerants. In this respect the current dissertation works on possible solutions to the question: which refrigerant is better, or which one is worse than the others?

Since chemical knowledge, as well as chemical substances, can be classified according to several criteria, the former question can be extended to: which class of refrigerants is better, or which one is worse than the others? To answer this question, different refrigerant classifications are introduced and studied with partial order theory.

Finally, it is shown a methodology to include priorities of the different features characterising the refrigerants studied. The method determines the needed priorities to ensure that a refrigerant is better than another one and it also yields the probability of that event.

In the following chapters each one of these procedures is introduced and further explained. However, a deep discussion is given in the manuscripts attached to this dissertation, which have been the result of this research.

Chapter 2: Unsupervised and supervised refrigerant classifications

Any classification equips a set with classes, which can be formed according to features of the elements in the set or created and imposed by the researcher; the former classifications are called unsupervised and the latter supervised ones [61]. In the particular case of refrigerants, one may classify them according to the similarity of their properties or the classes may be created based upon previous knowledge, for example the common classification according to the molecular structure into CFCs, HCFCs, HFCs. In this chapter unsupervised classifications are performed and their matching with supervised refrigerant classifications is studied.

When the environmental problems of CFC were recognised, first possible replacements were HCFCs and HFCs [8]. The chemical idea behind this solution was the searching for substances similar to CFCs, in fact the researchers proposing HCFCs and HFCs dealt with the issue of chemical similarity, which was rather the same done by Midgley when looking for ammonia and hydrocarbon substitutes in the 1930s [6].

Chemical similarity searching is a well established subject in chemical information studies [62]; it makes use of mathematical tools to look for classes of similar chemical objects, e.g. compounds, molecular fragments, etc. In the refrigerants' case these similarities were understood as close properties among CFCs and the possible replacements, for example non-flammability, non-toxicity, miscibility with lubricants, and thermodynamic features.

In general terms, any classification divides the set into different subsets, which may be disjoint or overlapped depending on the methodology used to find classes; for instance fuzzy cluster analysis [63] may yield overlapped classes whereas hierarchical cluster analysis disjoint ones [64]. Classification in chemistry is of special importance because it helps to save resources when the amount of data is too large, as often occurs in the current chemical investigations. For example, through classification it is possible to select a representative substance of each class for further study instead of analysing all substances within the class, whereby analysis and time spent on it are reduced.

2.1 Hierarchical cluster analysis (HCA)

In this dissertation hierarchical cluster analysis (HCA) was used as unsupervised classification technique; its first step is the characterisation of the objects to study by selecting various of their

attributes, e.g. thermodynamic properties or descriptors derived from molecular representations when the objects are substances. Afterwards, a similarity function is applied to calculate the nearness of objects' properties and finally the classes are formed by applying a grouping methodology [62]. Normally, HCA results are depicted in a tree called dendrogram whose branches represent clusters of similar objects. An exemplary dendrogram is depicted in Figure 2.1, where the most similar object to *a* is *b*; *c* is similar to *a* and *b*; additionally *d* is similar to *e*.

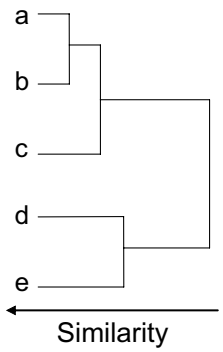


Figure 2.1. A dendrogram depicting similarity relationships among the objects *a*, *b*, *c*, *d* and *e*.

Cluster analysis permits to look for similarities to such an extent that they can be found even if the compounds actually are not similar. To solve this "similarity over-estimation" it is suggested to apply different clustering algorithms to assess whether the classes fluctuate or are stable under algorithm changes [65]. Therefore, if different classes result then no real similarities among chemicals hold; otherwise the classes actually exist and are not algorithmic dependent. Based on this idea, the cluster index was developed (Appendix A), a method to assess the similarity between classifications by contrasting their clusters.

2.2 Characterising substances

Currently, besides substances' experimental properties, more than 2000 features can be derived from their molecular representations [66]. These features are called molecular descriptors and started to be developed since the late 1940s when elements of discrete mathematics applied to chemistry began to be further studied [67-69]. Molecular descriptors can be classified into arithmetic, geometrical and topological ones. First of them count the presence of a particular feature within a molecule, e.g. chlorine atoms and bonds, and also calculate some values based upon those features, e.g. molecular weight. Geometrical descriptors represent information concerning three-dimensional features of molecules and are calculated from a molecular conformation of low energy [70], e.g. momenta of inertia, molecular volume and surface area. Topological descriptors are used to characterise the

constitution and configuration of a molecule by a single number [71]. To calculate them, molecules are regarded as graphs which can be analytically represented by matrices from which topological descriptors may be derived [72]. Examples of these descriptors include indices encoding size, shape, and branching of a molecule [73].

Molecular descriptors are used in Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) studies, in which a target property is related to different molecular descriptors in such a way that unknown target property values can be predicted from those descriptors. Some other applications appear in chemical retrieval information, where substances are classified according to their degree of descriptors' similarities [74-75]. Molecular descriptors can be efficiently calculated with various computer programmes; in the current dissertation MOLGEN-QSPR [76-79], software that combines structure generation with calculation of molecular descriptors and statistical treatment was used to characterise several refrigerants. Before describing some representative molecular descriptors for refrigerants, some fundamental terms are introduced.

In a molecular graph atoms are represented by points (vertices) and bonds by segments (edges) between vertices. This graph depicts the connectivity of atoms in a molecule irrespective of parameters representing the molecular geometry, e.g. interatomic distances, bond and torsion angles. In a H-suppressed molecular graph all hydrogen atoms are excluded. In the following some descriptors are described in more detail, which turned out to be most relevant due to the KIF-procedure, which is explained in section 2.5.2.

2.2.1 *Arithmetic descriptors*

Average atomic weight ($\overline{AW}$). It is given by

$$\overline{AW} = \frac{MW}{A} \qquad 2.1$$

here MW is the molecular weight of the studied substance and A is the number of atoms excluding hydrogens [66].

Relative number of X atoms (*Rel.* N_X). It is calculated according to

$$Rel.N_X = \frac{N_X}{A(incl.H)} \qquad 2.2$$

where N_X and $A(incl.\ H)$ is the number of X atoms, e.g. C, H, O, F, Cl, etc., and the total number of atoms in the molecule [66].

Number of methyl groups (T_m). It represents the number of methyl groups in the non H-suppressed molecular graph [66].

2.2.2 *Geometrical descriptors*

Steric energy. It corresponds to a stable spatial distribution of the atoms in a molecule. It is calculated using molecular mechanics, which considers a molecule as an ensemble of spheres (atoms) connected by springs (bonds). This calculation takes into account the ability of bonds to stretch, bend, and twist. It also accounts for interaction of non-bonded atoms through calculation of electrostatic forces [66].

Shadow indices. Before introducing these indices, principal moments of inertia are defined. The moment of inertia of a molecule is given by

$$I = \sum_{i=1}^{A(incl.H)} w_i \cdot r_i^2 \qquad 2.3$$

where r_i is the perpendicular distance of atom i with atomic weight w_i from a given axis. A molecule has three moments of inertia corresponding to its three axes in the three-dimensional space. This coordinate system can be transformed into another one based on three principal moments of inertia I_A, I_B and I_C, such that their origin is located at the molecular mass-centre. I_A is defined as the smallest moment, I_B as the intermediate one and I_C as the greatest moment; they define the principal inertia axes of the molecule, whose axes are aligned along their three principal inertia coordinates.

Shadow indices consider a molecule in a principal inertia system of coordinates; atoms are regarded as spheres of van der Waals radii, i.e. the atomic radii is calculated based upon the distance at which the attractive and repulsive forces between two non-bonded atoms are balanced [66]. Afterwards, the molecular surface is projected onto three mutually perpendicular planes *XY*, *XZ* and *YZ*, from which shadow descriptors are derived.

SHDW1: Area of molecular shadow in the *XY* plane.
SHDW2: Area of molecular shadow in the *XZ* plane.
SHDW3: Area of molecular shadow in the *YZ* plane.

$$SHDW4 = \frac{SHDW1}{L_X \cdot L_Y} \qquad SHDW5 = \frac{SHDW2}{L_X \cdot L_Z} \qquad SHDW6 = \frac{SHDW3}{L_Y \cdot L_Z} \qquad 2.4\text{-}2.6$$

where $L_X \cdot L_Y$ represent the area of the rectangle embedding the molecular *XY*-shadow. For {*SHDW1*, *SHDW2*, *SHDW3*}, *SHDWI* is the largest value, *SHDWII* the second largest one and *SHDWIII* the smallest value.

$$ssSHDW1 = SHDWI \qquad ssSHDW2 = SHDWII \qquad ssSHDW3 = SHDWIII \qquad 2.7\text{-}2.9$$

$$ssSHDW4 = \frac{SHDWI}{L_X \cdot L_Y} \qquad ssSHDW5 = \frac{SHDWII}{L_X \cdot L_Z} \qquad ssSHDW6 = \frac{SHDWIII}{L_Y \cdot L_Z} \qquad 2.10\text{-}2.12$$

In these six last descriptors, *ss* stands for size-sorted.

2.2.3 *Topological descriptors*

Hosoya index. It is given by

$$Z = \sum_{k=0}^{\lfloor A/2 \rfloor} a_k \text{ , with } \lfloor A/2 \rfloor = \begin{cases} A/2 \text{ for even } A \\ (A-1)/2 \text{ for odd } A \end{cases} \qquad 2.13$$

where a_k indicates the number of ways *k* edges may be selected from all *B* edges of the H-suppressed graph such that no two of them are adjacent. For any graph $a_0 = 1$ and $a_1 = B$ (number of bonds in the H-suppressed graph) [66, 80].

Connectivity indices. These descriptors are based upon *m*-th order subgraphs of a H-suppressed molecular graph and vertex degrees. A molecular subgraph is a subset of atoms and related bonds, it usually represents a molecular fragment; its order (*m*) is given by the number of edges within it. There are four types of molecular subgraphs: chain or ring (*ch*), cluster (*c*), path (*p*) and path-cluster (*pc*) (Figure 2.2). The type of molecular subgraph is determined as follows:

1. If the subgraph contains a cycle it is *ch*, for $m \geq 3$; otherwise
2. if every vertex degree is equal to one or greater than two, the subgraph is *c*, for $m \geq 3$; otherwise
3. if every vertex degree is equal to one or two, the subgraph is *p*, for $m \geq 2$; otherwise
4. the subgraph is *pc*, for $m \geq 4$.

The vertex degree of any vertex in a molecular subgraph is the number of neighbours the vertex has (Figure 2.2).

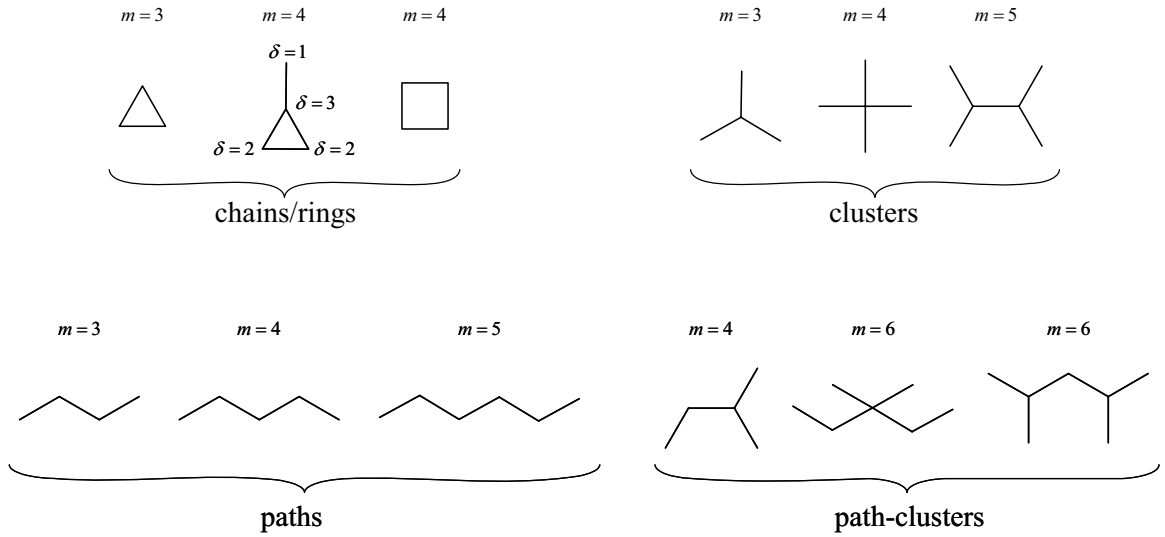


Figure 2.2. Examples of types of *m*-molecular subgraphs. Vertex degrees (δ) are shown for a 4-chain.

The connectivity indices of order *m* for graphs whose subgraphs are of type *q*, i.e. *ch*, *c*, *p* and *pc*, are calculated by

$$ {}^{m}\chi_q = \sum_{k=1}^{K(m,q)} \left(\prod_{i=1}^{n} \delta_i \right)_k^{-1/2} \qquad 2.14 $$

where *k* runs over all the *m*-th order subgraphs constituted by *n* atoms; *K* is the total number of *m*-th order subgraphs in the molecular graph. The product is performed over the simple vertex degrees δ_i of all vertices involved in each subgraph [66].

Perturbation connectivity index. These descriptors are calculated according to

$$ {}^{m}\chi_q^{v} = \sum_{k=1}^{K(m,q)} \left(\prod_{i=1}^{n} \delta_i^{v} \right)_k^{-1/2}, \quad \delta_i^{v} = \frac{Z_i^{v} - h_i}{Z_i - Z_i^{v} - 1} \qquad 2.15\text{-}2.16 $$

where δ_i^{v} is the valence vertex degree of atom *i* in a H-suppressed molecular graph, Z_i is the atomic number of atom *i*, Z_i^{v} is the number of valence electrons of atom *i* and h_i is the number of H atoms attached to atom *i* [66].

During 1993 and 1996 Randić, taking into account the large number of molecular descriptors developed and their mutual relationships, proposed the following requirements for claiming the existence of a new descriptor [81-83]:

- The descriptor must have a direct structural interpretation,
- must involve structural features that existing descriptors do not cover, and
- must have a high correlation with a substance or molecular property.

The elucidation of the molecular features related to the molecular descriptors is of utmost importance; however it is not always a reachable target because of the high diversity of features and properties the 32 799 436 known substances hold [84]. In order to assist and standardise the search for chemical meaning of molecular descriptors, the International Academy of Mathematical Chemistry recently suggested the use of benchmark data sets [85] for calculating descriptor values and relating them with specific molecular features of the molecules gathered in each data set. To cope with the comparison of descriptors, the same institution calculated the correlations of 735 molecular descriptors derived from 221 860 molecules from the National Cancer Institute dataset.

In general, it has been found that topological descriptors are related to some physico-chemical properties, e.g. melting point, boiling point, refractive index, molar volume and density. If the aim is the estimation of biological activities based upon molecular descriptors, topological and geometrical ones have resulted to be related to these properties [70].

2.3 Contrasting classifications: Cluster index

This methodology permits to measure the resemblance between classifications (dendrograms); a brief description of it is given in the following. Given two dendrograms D_i and D_j defined on a set P of n objects, their clusters are collected in CD_i and CD_j, respectively. The number of different clusters between D_i and D_j is calculated by the cardinality of the symmetric difference of CD_i and CD_j, $|C(D_i, D_j)|$, which is given by

$$\begin{aligned} |C(D_i, D_j)| &= |CD_i \cup CD_j| - |CD_i \cap CD_j| \\ &= |CD_i| + |CD_j| - 2|CD_i \cap CD_j| \end{aligned} \qquad 2.17$$

The number of clusters of a dendrogram is $2n - 1$ (Appendix A), therefore $|CD_i| = |CD_j| = 2n - 1$. Any two dendrograms have always all their n single clusters in common and also the cluster gathering the n

objects; these $n + 1$ clusters are called trivial clusters. Hence, if the trivial clusters are removed from the clusters of each dendrogram, then $|CD_i| = |CD_j| = 2n - 1 - (n + 1) = n - 2$, which yields

$$| C(D_i, D_j) | = 2(n-2) - 2 | CD_i \cap CD_j | \qquad 2.18$$

If c represents the number of common clusters between both dendrograms, then $|C(D_i, D_j)| = 2(n - 2 - c)$. As c takes values in the following interval $0 \leq c \leq n - 2$, then $|C(D_i, D_j)|$ can be normalised yielding $CI(D_i, D_j)$, the cluster index.

$$CI(D_i, D_j) = 1 - \frac{c}{n-2} \qquad 2.19$$

When $CI(D_i, D_j) = 0$, the contrasted dendrograms have all their clusters in common. If $CI(D_i, D_j) = 1$, all clusters are different. Details on cluster index, its mathematical properties and its comparison with some other methods to contrast dendrograms are found in Appendix A. As an example of application, cluster index for the dendrograms depicted in Figure 2.1 (D_1) and 2.3 (D_2) is calculated.

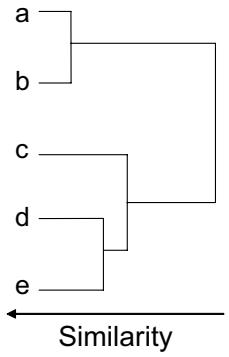


Figure 2.3. A dendrogram of five objects.

In this case, $CD_1 = \{\{a, b\}, \{d, e\}, \{a, b, c\}\}$ and $CD_2 = \{\{a, b\}, \{d, e\}, \{c, d, e\}\}$. By contrasting CD_1 and CD_2 it is concluded that $c = 2$, that is, there are two common clusters to D_1 and D_2, namely $\{a, b\}$ and $\{d, e\}$. Hence, $CI(D_1, D_2) = 1/3$, which means that D_1 and D_2 are 33 % dissimilar (66 % similar).

2.4 Chemotopology

The investigation of objects within a class is one of the most important aims of cluster analysis, many text books conclude its application by interpreting clusters [64]. Nevertheless, the collection of classes as a whole contains important information on the objects studied and their relationships. These

relationships are the starting point of chemotopology [86, 87], a mathematical procedure in which the collection of classes found by a classification process, e.g. HCA, becomes a topological basis. This mathematical view of the collection of clusters enables topological studies of different subsets belonging to the set under study [88]. Its application to a set *P* endowed with a dendrogram can be summarised as follows:

- Selection of clusters from the dendrogram, which are gathered in a topological basis (**B**).
- Selection of a subset *A* of *P* for studying its topological properties; some of them are closure ($\overline{A}$), boundary [b(*A*)] and interior [int(*A*)].

The closure of *A* contains the elements of *P* which are similar to *A*; the elements of *P* which are similar to *A* and simultaneously to elements not included in *A* constitute the boundary of *A*. The interior of *A* contains the elements of *P* which are completely similar to *A* and constitute the core of *A*. A further mathematical discussion on the topological properties and on their chemical meaning is given in references [88, 89]. An important aspect of the chemotopological approach is the generalised concept of similarity that can be derived by its application. Chemotopology permits to reach a deep understanding of the similarity relationships among members of a set, it permits to find elements of a class which are strongly related to the main features of the class, i.e. class representatives; additionally it is possible to find elements which share features of different classes and therefore are transition elements between different classes. Chemotopology has been applied to different chemical sets [86-91] and particularly to chemical elements [86, 89, 90] where it has permitted to make a step forward on the mathematisation of well-known relationships among these substances such as the concept of metallicity, non-metallicity and semi-metallicity.

2.5 HCA of refrigerants

In order to assess the similarity relationships among different refrigerants three classifications were performed, two based on experimental properties and another one on molecular descriptors. The 40 chemicals studied* are shown in Table 2.1, they constitute a diverse group of refrigerants used in the past, used presently, and some proposed substitutes (Appendix H). Because of the different scales of the features selected to characterise substances, each feature is normalised as follows [86]:

$$q_i(x) = \frac{q_i'(x) - \min q_i'}{\max q_i' - \min q_i'} \qquad 2.20$$

* Although some substances may have different isomers, it is known the specific isomer for which the properties shown in Table 2.1 are determined. A list of the molecular structures for the 40 refrigerants appears in Figure S.1 of appendix H.

where $q_i'(x)$ is the value of feature *i* for chemical *x*, and $\min q_i'$ and $\max q_i'$ are the minimum and maximum values of feature *i*, respectively.

Table 2.1. Refrigerants included in this study, their labels, chemical families, molecular formulae, chemical and non-proprietory names.

Label	Subset	Molecular formula	Chemical name	Non-proprietory name
1	CFC	CCl_3F	Trichlorofluoromethane	R11
2	CFC	CCl_2F_2	Dichlorodifluoromethane	R12
3	HCFC	$CHClF_2$	Chlorodifluoromethane	R22
4	HCFC	$C_2HCl_2F_3$	2,2-Dichloro-1,1,1-trifluoroethane	R123
5	HCFC	C_2HClF_4	2-Chloro-1,1,1,2-tetrafluoroethane	R124
6	HCFC	$C_2H_3Cl_2F$	1,1-Dichloro-1-fluoroethane	R141b
7	HCFC	$C_2H_3ClF_2$	1-Chloro-1,1-difluoroethane	R142b
8	HFC	CHF_3	Trifluoromethane	R23
9	HFC	CH_2F_2	Difluoromethane	R32
10	HFC	C_2HF_5	Pentafluoroethane	R125
11	HFC	$C_2H_2F_4$	1,1,1,2-Tetrafluoroethane	R134a
12	HFC	$C_2H_3F_3$	1,1,1-Trifluoroethane	R143a
13	HFC	$C_2H_4F_2$	1,1-Difluoroethane	R152a
14*	HFC	$C_3H_3F_5$	1,1,1,3,3-Pentafluoropropane	R245fa
15*	HFC	$C_3H_2F_6$	1,1,1,3,3,3-Hexafluoropropane	R236fa
16	HC	C_3H_8	*n*-Propane	R290
17	HC	C_4H_{10}	*n*-Butane	R600
18	HC	C_4H_{10}	Isobutane	R600a
19	HC	C_5H_{12}	*n*-Pentane	R601
20	HC	C_3H_6	Propene	R1270
21*	CO_2	CO_2	Carbon dioxide	R744
22	BCF	$CBrClF_2$	Bromochlorodifluoromethane	R12B1
23	PFC	C_4F_8	Octafluorocyclobutane	RC318
24*	HFC	C_3HF_7	1,1,1,2,3,3,3-Heptafluoropropane	R227ea
25	AFAE	$C_4H_3F_7O$	Heptafluoropropyl methyl ether	HFE-7000
26	AFAE	$C_5H_3F_9O$	Methyl nonafluorobutyl ether	HFE-7100
27	AFAE	$C_6H_5F_9O$	Ethyl nonafluorobutyl ether	HFE-7200

				HFE-569mccc
28	AFAE	$C_9H_5F_{15}O$	Ethyl pentadecafluoroheptyl ether	HFE-7500
29*	DFAE	C_2HF_5O	Pentafluorodimethyl ether	HFE-125
30*	DFAE	$C_2H_2F_4O$	1,1,1',1'-Tetrafluorodimethyl ether	HFE-134
31	CM	CH_2Cl_2	Methylene chloride	R30
32	CM	CH_3Cl	Methyl chloride	R40
33	CFC	$C_2Cl_3F_3$	1,1,2-Trichloro-1,2,2-trifluoro-ethane	R113
34	HCFC	$CHCl_2F$	Dichlorofluoromethane	R21
35	CFC	$C_2Cl_2F_4$	1,2-Dichloro-1,1,2,2-tetrafluoro-ethane	R114
36*	FIM	CF_3I	Trifluoroiodomethane	R13I1
37	DME	C_2H_6O	Dimethyl ether	-
38	NH_3	NH_3	Ammonia	R717
39*	AFAE	$C_2H_3F_3O$	Methyl trifluoromethyl ether	HFE-143
40*	AFAE	$C_3H_3F_5O$	Methyl pentafluoroethyl ether	HFE-245

* Not considered refrigerants in the thermodynamic classification.

2.5.1 *Classification based upon experimental properties*

Each refrigerant was characterised by its ozone depletion potential (ODP), global warming potential (GWP) and atmospheric lifetime (ALT); these values appear in Table 1 of Appendix H. Two similarity functions were applied, Hamming (H) and Euclidean (E) distance, both particular cases of Minkowski's metric families [86, 92].

$$d_p(x,y) = \left[\sum_{i=1}^{n} \left| q_i(x) - q_i(y) \right|^p \right]^{1/p} \quad p = 1 \text{ Hamming, } p = 2 \text{ Euclidean distances} \quad 2.21$$

Four grouping methodologies were used (Table 2.2), which are special cases of the Lance and Williams' formula [93].

$$f(L,i) = \alpha_A f(A,i) + \alpha_B f(B,i) + \beta f(A,B) + \gamma \left| f(A,i) - f(B,i) \right| \quad 2.22$$

where L is formed by merging clusters A and B and $f(L, i)$, $f(A, i)$, $f(B, i)$ and $f(A, B)$ are the nearness between clusters L and i, A and i, B and i and A and B, respectively.

Table 2.2. Grouping methodologies.

Methodology	α_A	α_B	β	γ
Single linkage (sing)	0.5	0.5	0	-0.5
Complete linkage (comp)	0.5	0.5	0	0.5
Unweighted average linkage (unav)	$\frac{n_A^*}{(n_A + n_B)}$	$\frac{n_B}{(n_A + n_B)}$	0	0
Ward's method (Ward)	$\frac{n_A + n_i}{n_A + n_B + n_i}$	$\frac{(n_B + n_i)}{(n_A + n_B + n_i)}$	$\frac{-n_i}{(n_A + n_B + n_i)}$	0

*n_j is the number of elements in the j group

Because each combination of similarity function and grouping methodology lead to a dendrogram, eight of them were obtained: E-sing, E-comp, E-unav, E-Ward; H-sing, H-comp, H-unav, and H-Ward. In order to study the resemblance of these classifications the cluster index was applied to each pair of dendrograms (Table 2.3). For each dendrogram D_i, an average cluster index ($\overline{CI}_i$) was calculated according to:

$$\overline{CI}_i = \frac{\sum_{j=1}^{t} CI(D_i, D_j)}{t} \qquad 2.23$$

where $CI(D_i, D_j)$ is the cluster index between dendrograms D_i and D_j and t is the number of dendrograms contrasted with D_i. These values show that the classifications are not so different which implies that the clusters found using environmental properties are stable under algorithmic variations. In average, the most similar dendrogram to the other ones is E-unav (Figure 2.4).

Table 2.3. Cluster index and average cluster index values for the contrast of dendrograms obtained using environmental properties.

	E-sing	E-comp	E-unav	E-Ward	H-sing	H-comp	H-unav	H-Ward
E-sing	0	0.342	0.289	0.447	0.105	0.342	0.289	0.447
E-comp	0.342	0	0.184	0.211	0.316	0.211	0.211	0.289
E-unav	0.289	0.184	0	0.316	0.342	0.289	0.184	0.263
E-Ward	0.447	0.211	0.316	0	0.447	0.342	0.263	0.158
H-sing	0.105	0.316	0.342	0.447	0	0.289	0.316	0.447
H-comp	0.342	0.211	0.289	0.342	0.289	0	0.289	0.342
H-unav	0.289	0.211	0.184	0.263	0.316	0.289	0	0.289
H-Ward	0.447	0.289	0.263	0.158	0.447	0.342	0.289	0
$\overline{CI}_i$	0.417	0.293	0.267	0.380	0.703	0.549	0.523	0.620

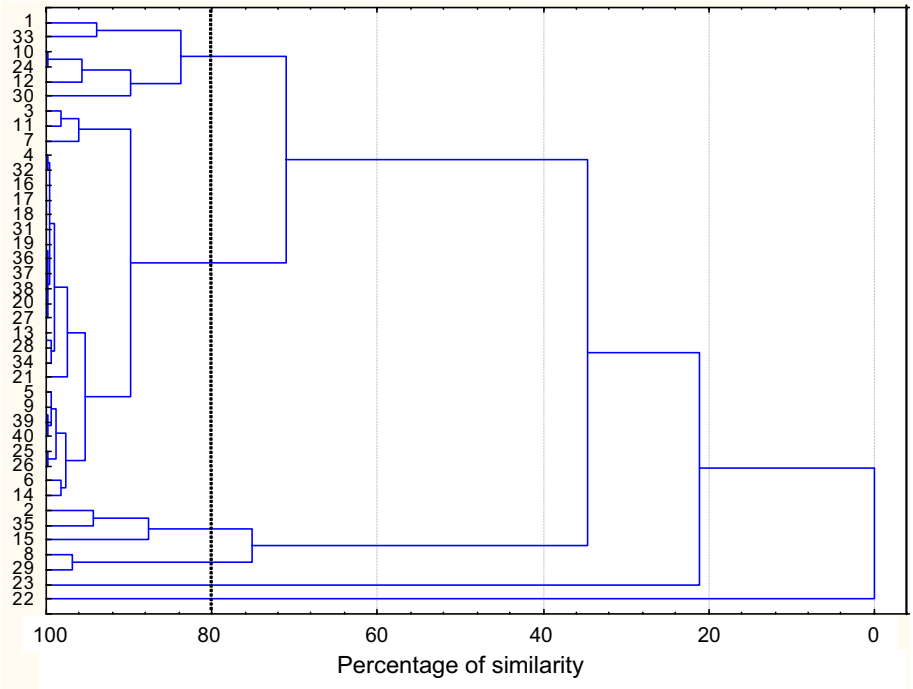


Figure 2.4. Dendrogram of 40 refrigerants based upon environmental properties and calculated with the Euclidean distance and the unweighted average linkage. A cluster selection level of 80 % of similarity is remarked.

Refrigerants are commonly classified into different families, namely CFCs, HCFCs and HFCs to name but a few [8]. This supervised classification is mainly based on the chemical composition and functional groups present in the molecules. By imposing this classification to the 40 refrigerants shown in Table 2.1, they are split into 13 families, labelled: CFC, HFC, HCFC, hydrocarbons (HC),

di(fluoroalkyl)ethers (DFAE), alkylfluoroalkylethers (AFAE), chloromethanes (CM), and the single-compound families trifluoroiodomethane (FIM), octafluorocyclobutane (PFC), carbon dioxide (CO_2), bromochlorodifluorobutane (BCF), dimethyl ether (DME) and ammonia (NH_3) (Table 2.1). None of these families appear as a complete cluster obtained by HCA of environmental properties. However, some of them belong to some clusters. In table 2.4 the frequency of appearance of each refrigerant family into the 141 different clusters obtained with the eight HCA methods is shown. Note that refrigerant families with only one chemical are not considered because they constitute trivial clusters, i.e. they appear in each dendrogram as single elements before any aggregation with another refrigerant. It is also shown (Table 2.4) the percentage of clusters in which a family appears.

Table 2.4. Frequency of appearance of refrigerant families into clusters obtained from environmental properties and from thermodynamic ones. Percentages of clusters containing a given family of refrigerants.

	Environmental properties		Thermodynamic properties	
Family	Frequency	Percentage (%)	Frequency	Percentage (%)
CFC	0/141	0	0/51	0
HCFC	0/141	0	0/51	0
HFC	0/141	0	4/51	8
HC	27/141	19	5/51	10
AFAE	0/141	0	2/51	4
DFAE	4/141	3	-	-
CM	24/141	17	4/51	8

These results show that although refrigerants form stable clusters when grouped using environmental properties, these clusters do not match with the common classification of refrigerants into different families, e.g. CFCs, HCFCs, etc. This implies that the environmental classification does not follow a one-to-one relation with the common partition of refrigerants into families.

In order to explore the relationship between refrigerants' families and their thermodynamic classification, heat of vaporisation, boiling point and heat capacity were considered as new refrigerant descriptors. Heat of vaporisation was selected because of their importance for the refrigeration process as mentioned in section 1.1.2. Boiling points and heat capacities were also selected for their relationship with heat of vaporisation and specific heat [94]. Specific volumes and specific heats, important properties for the refrigeration performance as mentioned in section 1.1.2, were not included due to the lack of information for the refrigerants studied. In fact, data for the three thermodynamic properties included were found only for 31 of the original 40 refrigerants. The labels of refrigerants

which were not studied in this thermodynamic classification are marked with an asterisk (*) in Table 2.1.

After performing the classifications with these new properties, the families of refrigerants were looked for into the clusters obtained (Table 2.4). It can be seen that the percentage of appearance of HFC and AFAE families increases from 0 % to 8 %, and from 0 % to 4 %, respectively, when thermodynamic properties are considered. However, this percentage decreases for HC and CM, from 19 % to 10 % and from 17 % to 8 %, respectively. Thereby, the matching of refrigerant families and clusters found by HCA is not improved when the classification is only performed considering thermodynamic properties. This result shows the lack of a one-to-one relation between the thermodynamic classification and the common refrigerant classification into different families. Hence, for instance, it is not possible to expect that all HCFCs hold similar thermodynamic properties just because they belong to HCFC family.

2.5.2 *Classification based upon molecular descriptors*

The pool of 708 molecular descriptors included in the software package MOLGEN-QSPR [76-79] was applied to the 40 refrigerants shown in Table 2.1. Several descriptors require a minimum molecular size to be calculated. Because of the small size of the molecules representing some refrigerants, 388 molecular descriptors could effectively be computed. There are several methods for selecting the most representative and independent descriptors [77, 95-98]. In the current dissertation, the information content of each descriptor was calculated [96-98] in order to select the most informative ones. Afterwards, the *K* inflation factor technique (*KIF*) was used to find the most representative descriptors [98]. This procedure finds and eliminates those descriptors with the highest multivariate correlation. *KIF* is based upon the *K* multivariate correlation index [97,98], given by

$$K = \frac{\sum_{m=1}^{p} \left| \frac{\lambda_m}{\sum_{m=1}^{p} \lambda_m} - \frac{1}{p} \right|}{\frac{2(p-1)}{p}} \qquad 2.24$$

where p is the number of descriptors and λ_m the eigenvalues obtained by the diagonalisation of the descriptors' correlation matrix. In the *KIF* method, the multivariate correlation index $K_{p/j}$ is calculated by removing the j-th descriptor from the original p ones. Therefore, $K_{p/j}$ is calculated on a correlation matrix obtained with n chemicals and $p - 1$ descriptors. The key of the algorithm is to look for and

eliminate the q descriptor with the highest multivariate correlation with those remaining. Hence, when q is excluded the remaining multivariate correlation derived from the remaining $p - 1$ descriptors is maximally decreased. Upon q is eliminated, the whole procedure is recursively repeated on the remaining $p - 1$ descriptors. The KIF_j value of the j-th descriptor is an inflation factor obtained by considering the total correlation K_p and the correlation $K_{p/j}$, scaled according to the different number of descriptors p and $p - 1$, respectively. KIF_j values range from 0 to 1, $KIF_j = 0$ when the eliminated j descriptor is uncorrelated with the remaining ones and its is 1 when the j-th descriptor is correlated with the remaining descriptors. For this reason, those descriptors with $KIF_j < 0.5$ or < 0.6 are normally kept, while the others are removed [98].

After applying the *KIF* methodology to the 388 molecular descriptors, they were reduced to 15 informative and uncorrelated ones whose *KIF* and mean correlation values are shown in Table 2.5. This result is understandable since most of the descriptors were highly correlated because of the small size of the majority of molecules considered.

Table 2.5. Relevant informative molecular descriptors for the 40 refrigerants shown in Table 2.1.

Descriptor	*KIF* value	Mean correlation value
Steric energy	0	0.126
Rel. NCl	0	0.214
ssSHDW5	0.026	0.157
ssSHDW6	0.063	0.259
Z	0.082	0.658
Rel. NO	0.157	0.051
Rel. NC	0.204	0.345
SHDW4	0.279	0.453
${}^{m}\chi_c$	0.268	0.616
ssSHDW1/SHDW2	0.308	0.314
Tm	0.299	0.103
${}^{m}\chi_c^{v}$	0.343	0.492
ssSHDW1/SHDW3	0.396	0.246
$\overline{AW}$	0.461	0.162
Rel. NF	0.506	0.383

Consequently, the 40 refrigerants are characterised by six geometrical descriptors, six arithmetic and three topological ones. The eight HCA methodologies mentioned in section 2.5.1 were applied to these descriptors; cluster index results are shown in Table 2.6.

Table 2.6. Cluster index and average cluster index values for the contrast of dendrograms obtained using molecular descriptors.

	E-sing	E-comp	E-unav	E-Ward	H-sing	H-comp	H-unav	H-Ward
E-sing	0	0.658	0.526	0.711	0.526	0.684	0.605	0.711
E-comp	0.658	0	0.474	0.316	0.737	0.658	0.605	0.684
E-unav	0.526	0.474	0	0.579	0.632	0.658	0.5	0.684
E-Ward	0.711	0.316	0.579	0	0.763	0.658	0.711	0.684
H-sing	0.526	0.737	0.632	0.763	0	0.658	0.579	0.658
H-comp	0.684	0.658	0.658	0.658	0.658	0	0.553	0.526
H-unav	0.605	0.605	0.5	0.711	0.579	0.553	0	0.553
H-Ward	0.711	0.684	0.684	0.684	0.658	0.526	0.553	0
$\overline{CI}_i$	0.632	0.590	0.579	0.632	0.650	0.628	0.586	0.643

These values show that these classifications are rather different, holding average cluster index greater than 0.5, which contrasts with the results obtained for classifications based on environmental properties. In the current case, clusters found depend on the clustering algorithm applied, therefore these clusters cannot be considered as neighbourhoods of compounds sharing actual similarities. In general, the similarities among refrigerants are more stable with respect to their environmental properties than with respect to their molecular structures. In the present classification based upon molecular descriptors, the dendrogram obtained with the Euclidean similarity function and the unweighted average linkage grouping methodology is the most similar to the other dendrograms (Figure 2.5).

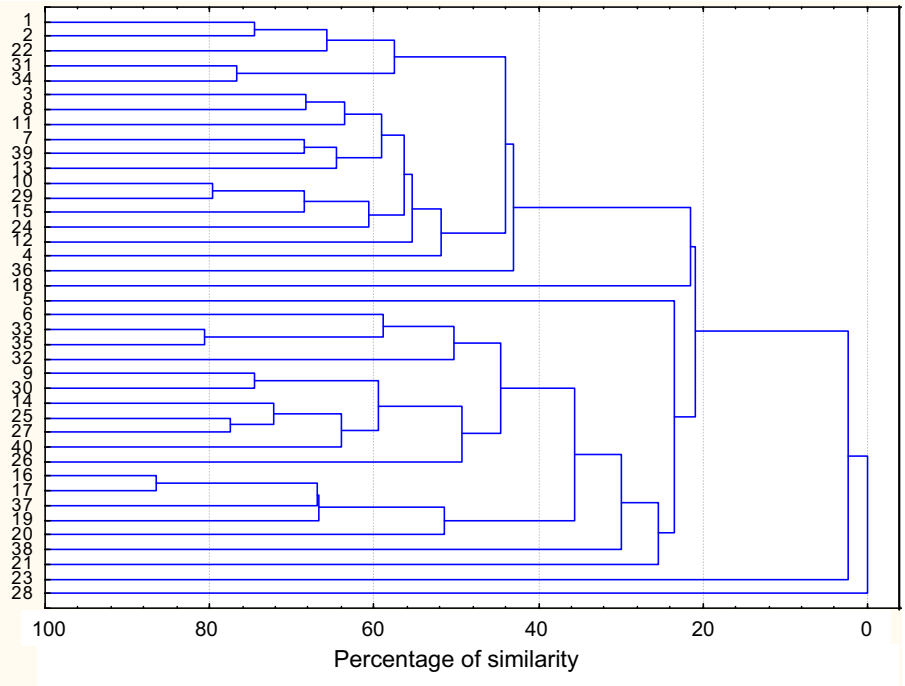


Figure 2.5. Dendrogram of 40 refrigerants based upon molecular descriptors and calculated with the Euclidean distance and the unweighted average linkage.

Refrigerant families discussed in section 2.5.1 were looked for into the clusters found using molecular descriptors. None of these families appears as a complete cluster. Afterwards, the presence of each family into the 155 different clusters obtained with the eight HCA was analysed. Table 2.7 condenses these results and it also shows the percentage of clusters in which a family appears are shown.

Table 2.7. Frequency of appearance of refrigerant families into clusters obtained from molecular descriptors. Percentages of clusters containing a given family of refrigerants.

	Molecular descriptors	
Family	Frequency	Percentage (%)
CFC	0/155	0
HCFC	0/155	0
HFC	3/155	2
HC	11/155	7
AFAE	0/155	0
DFAE	57/155	37
CM	37/155	24

These results suggest that the classification based upon molecular descriptors, as well as the one performed with environmental properties, does not match with the common classification of

refrigerants into different families; implying that the classification based upon refrigerant molecular structures does not follow a one-to-one relation with the partition of refrigerants into families. From the refrigerant families considered, namely CFC, HCFC, HFC, HC, AFAE, DFAE and CM, it can be said that HC, DFAE and CM are the only cases where, in the three classifications performed, they appear as part of the obtained clusters. However, these families do not appear as isolated clusters but being part of large clusters and combined with other refrigerants from other families.

2.6 Chemotopology of refrigerants

In this section the chemotopological procedure is applied to a representative dendrogram of those obtained with environmental properties since these clusters are the most stable under clustering algorithmic variations. By selecting a similarity level of 80 %, the clusters found from the dendrogram depicted in Figure 2.4 were collected in the topological basis $\mathbf{B}_{80}$.

$$\mathbf{B}_{80} = \left\{ \begin{array}{c} \{1,\ 33\},\ \{2,\ 15,\ 35\},\ \{8,\ 29\},\ \{10,\ 12,\ 24,\ 30\}, \\ \{22\},\ \{23\},\ \{3,\ 4,\ 5,\ 6,\ 7,\ 9,\ 11,\ 13,\ 14,\ 16,\ 17,\ 18,\ 19, \\ 20,\ 21,\ 25,\ 26,\ 27,\ 28,\ 31,\ 32,\ 34,\ 36,\ 37,\ 38,\ 39,\ 40\} \end{array} \right\}$$

For the sake of simplicity, we call

$$G = \left\{ \begin{array}{c} 3, 4, 5, 6, 7, 9, 11, 13, 14, 16, 17, 18, 19, 20, 21, \\ 25, 26, 27, 28, 31, 32, 34, 36, 37, 38, 39, 40 \end{array} \right\}$$

Then $\mathbf{B}_{80}$ = {{1, 33}, {2, 15, 35}, {8, 29}, {10, 12, 24, 30}, {22}, {23}, *G*}.

The topological results for the subsets *CFC*, *HCFC*, *HFC*, *HC*, CO_2, *BCF*, *PFC*, *AFAE*, *DFAE*, *CM*, *FIM*, *DME* and NH_3 are shown in the following. Note that these subsets are written in italics to remark their mathematical character.

CFC = {1, 2, 33, 35}	*HCFC* = {3, 4, 5, 6, 7, 34}
$\overline{CFC}$ = {1, 2, 15, 33, 35}	$\overline{HCFC}$ = *G*
b(*CFC*) = {2, 35, 15}	b(*HCFC*) = *G*
int(*CFC*) = {1, 33}	int(*HCFC*) = ∅

HFC = {8, 9, 10, 11, 12, 13, 14, 15, 24}

$\overline{HFC}$ = {2, 8, 10, 12, 15, 24, 29, 30, 35, *G*}

b(*HFC*) = {2, 8, 10, 12, 15, 24, 29, 30, 35, *G*}
int(*HFC*) = ∅

Same *HCFC* topological properties were found for *HC*, CO_2, *AFAE*, *CM*, *FIM*, *DME* and NH_3.

Regarding CFC it can be stated that these substances are similar to themselves and also to 15, a HFC (Table 2.1). CFC's boundary and interior show that refrigerants 1 and 33 constitute the core of CFCs, while 2, 35 and 15 are chemicals with similarities to the other CFCs and also to some other refrigerants different to CFCs.

The topological properties of *HCFC*, *HC*, CO_2, *AFAE*, *CM*, *FIM*, *DME* and NH_3 show that each family of refrigerants is similar to many other substances, in fact to those chemicals gathered in *G*, which contains HCFCs, AFAEs, HCs, FIM, DME, NH_3, CMs and CO_2. Thereby, there is no clear distinction of families for these classes of refrigerants. Similar conclusions can be drawn for *HFC*, whose chemicals are similar to *G* and to other substances like 2, 29, 30 and 35. The fact of finding an empty boundary for these subsets stresses the lack of identity of these refrigerant families, implying the impossibility of selecting a representative of these families. These results stress the lack of matching between the environmental refrigerant's classification and the common classification into refrigerant families. If there were an agreement between families and refrigerant clusters obtained from environmental properties, then the topological properties of these families would show empty boundaries and cores make from all chemicals in each particular family.

Chapter 3: Refrigerant classifications based upon order

Chemicals can hold different relationships, in chapter 2 it was shown that substances can be related by their similarities, which in mathematical terms is understood as the presence of a tolerance relation between chemicals of a set [99]. Once chemicals are endowed with a relation they can be further classified based upon such relation. It was shown in chapter 2 how a similarity relation can be used to this end. However, chemicals not only share similarities, they can also be endowed with some other relations. In this chapter the order relationships in refrigerants are explored and these substances are classified according to this mathematical relation.

3.1 Order relationships in chemistry

3.1.1 *Order relation*

Two elements x, y, characterised by their properties $q_1(x)$, $q_2(x)$,..., $q_n(x)$ and $q_1(y)$, $q_2(y)$,..., $q_n(y)$, respectively, are said to be comparable, if $q_i(x) \leq q_i(y)$ or $q_i(y) \leq q_i(x)$, for all $i = 1, 2,.., n$; in this case it is written $x \leq y$ or $y \leq x$, respectively. If $q_i(x) \leq q_i(y)$ not for all i, which implies the existence of at least one property j with $q_j(x) > q_j(y)$ and one property k with $q_k(x) < q_k(y)$, then x and y are said to be incomparable and it is written $x \parallel y$. Sets endowed with an order relation are called partially ordered sets (posets), a particular poset is a totally ordered set, whose elements are comparable to each other [100].

3.1.2 *Applications in chemistry*

The order relation is rather usual in chemistry [101], for example in the ordering of substances according to their boiling points, or aromaticities; or in the different order relations one finds in the periodic table [102]. An order relation underlies several environmental policies, for example the total order behind the Hazard Ranking System developed by the US Environmental Protection Agency to prioritise hazardous waste sites in the USA [103].

Particularly in environmental sciences and policy making, a big effort is done to obtain total orders which are introduced through different ranking procedures such as the Utility Function [104], PROMETHEE [105], Concordance Analysis [106], and many others described in reference 107. In general, these ranking methodologies perform mathematical operations on the chemical descriptors in

such a way that their different kinds of information are finally condensed into a number or ranking score. These mathematical changes hide the meaning and contribution of each descriptor to the final result. In addition, the mathematical functions employed to combine descriptors, also called aggregation functions, are often adjusted by the researcher implying a bias in the process.

3.2 Hasse Diagram Technique (HDT)

A different approach to endow a set with an order relation but without hiding descriptor effects is the Hasse diagram technique (HDT) [108, 109] which assigns to any pair of elements in the set either the relation $\leq$ or ||, introduced in section 3.1.1. Since this methodology is extensively applied in the present research, its application is illustrated using an example. Let us consider $P = \{a, b, c, d, e, f\}$ where each element is characterised by the descriptors q_1, q_2 and q_3 (Table 3.1); as a methodological precondition, all descriptors need to be consistently orientated in such a way that low descriptor values indicate, for example, low ranking and high values the contrary [103]; that is the convention adopted in the present dissertation. Order relations among elements in P may be represented by the graph $G(P, E)$ (Figure 3.1A), where each element in P is represented by a vertex in the graph and E is the set of edges (arrows) between any two $x, y \in P$ such that $x \leq y$; by convention, the arrow points to y. These arrows can be replaced by lines if for each pair $x \leq y$ (comparable elements), y is located higher than x in the drawing plane (Figure 3.1B). Hence, the graph depicted in Figure 3.1B is a graph of comparable pairs and is called a comparability graph. Normally, $G(P, E)$ contains unnecessarily many edges from which several ones can be omitted because the order relations they show are implied by some other edges; for example, the edge corresponding to $a \leq e$ can be dropped because this relation is already shown by the pairs $a \leq c$ and $c \leq e$. This edge reduction is called a "transitive reduction" and the resulting graph is called a Hasse diagram (HD) (Figure 3.1C) if it fulfils the following requirements [110]:

1) Each element $x \in P$ is represented by a labelled circle.

2) For all $x < y$, y is located at the top and x at the bottom of the drawing plane and they are connected by a line, taking into regard the transitive reduction.

3) Given $x, y \in P$, if $q_i(x) = q_i(y)$ for all i, then x and y are equivalent ($x \sim y$) and are represented by a double circle labelled either for x or for y.

4) For all $x \parallel y$ their circles are either no connected or connected through a sequence of lines in the upward-downward direction or downward-upward direction.

5) Circles are arranged into levels (numbers 1 to 4 in Figure 3.1C).

6) When possible, a circle is located at the highest position of the drawing plane as a convention.

In the HD depicted in Figure 3.1C *e* and *f* are ranked higher because their three descriptors are higher than those of the other chemicals (Table 3.1). An example of incomparable elements is the case of *a* and *d*, where $q_1(a) < q_1(d)$, $q_2(d) < q_2(a)$ and $q_3(a) = q_3(d)$ (Table 3.1, Figure 3.1C). In this case *a* is connected with *d* through a sequence of lines of the sort upward-downward direction.

Table 3.1. Properties q_1, q_2 and q_3 of the elements *a*, *b*, *c*, *d*, *e* and *f*.

	q_1	q_2	q_3
a	0	5	1
b	1	1	1
c	3	6	4
d	2	4	1
e	4	9	7
f	5	7	7

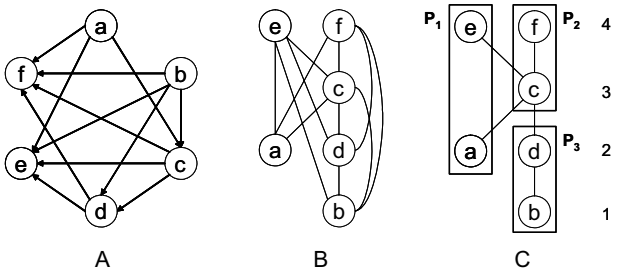


Figure 3.1. A) Graph $G(P, E)$, B) comparability graph and C) Hasse diagram of the elements in Table 3.1; the numbers in C represent the levels of the Hasse diagram; P_1, P_2 and P_3 are subsets (see text).

As can be seen in this example, HDT makes use of all descriptors for endowing *P* with an order relation but it does not consider any mathematical function combining them, thereby they contribute simultaneously to the ranking process without a bias.

The order relations depicted in a HD can be used to perform classifications based upon order in the set considered. For example the chemicals in Figure 3.1C can be classified into the following three subsets: $D = \{e, f\}$, $F = \{a, b\}$ and $I = \{c, d\}$, where *D*, *F* and *I* contain the most problematic, least problematic and intermediate elements, respectively.

In the following the results of the application of the HDT to the set of 40 refrigerants introduced in chapter 2 are summarised.

3.3 HDT applied to refrigerants

The HD of the 40 refrigerant under consideration (Table 2.1) is depicted in Figure 3.2, from which the following conclusions can be drawn:

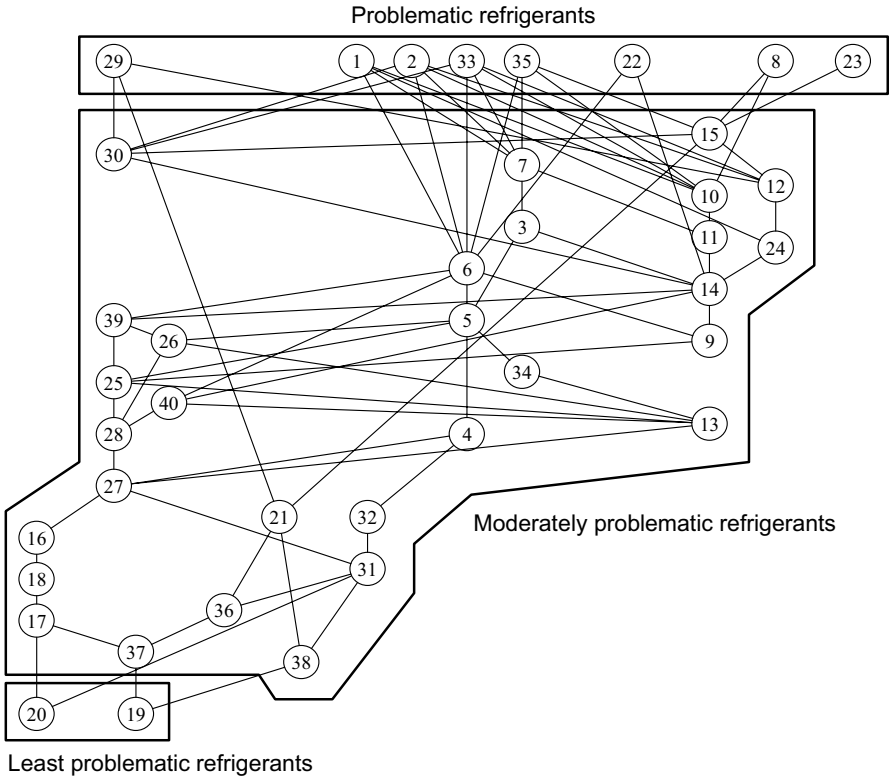


Figure 3.2. Hasse diagram of refrigerants shown in Table 2.1 and calculated from their ODPs, GWPs and ALTs. The three subsets correspond to classes obtained from refrigerant environmental properties.

Eight problematic refrigerants with high impact with respect to ODP, GWP and ALT, and two least problematic substances were found. The former are CCl_3F (1), CCl_2F_2 (2), CHF_3 (8), $CBrClF_2$ (22), C_4F_8 (23), C_2HF_5O (29), $CHClF_2$ (33), and $C_2Cl_2F_4$ (35); the least problematic ones are C_5H_{12} (19) and C_3H_6 (20); the numbers in parenthesis indicate the label of each substance (Table 2.1). The most problematic chemicals are well-known examples of ozone depleting substances such as CCl_3F (1), CCl_2F_2 (2), $CBrClF_2$ (22), $CHClF_2$ (33) and $C_2Cl_2F_4$ (35), which include CFCs, HCFCs and the bromochlorofluorinated substance $CBrClF_2$; these compounds are also characterised by high ALTs and relatively high GWPs. The other two environmentally problematic refrigerants are CHF_3 (8) and

C_4F_8 (23), a HFC and a perfluorocarbon (PFC), respectively, which are located on the top of the HD because of their high GWPs and in the case of C_4F_8 (23) also high ALT. The fact of having C_2HF_5O (29) as one of the most problematic substances is remarkable since this is a hydrofluoroether produced to replace the problematic CFCs, HCFCs and HFCs [57].

Regarding the least problematic substances, they are two HCs, namely *n*-pentane (19) and propene (20), whose ODPs, GWPs and ALTs are considerable lower than the values of the other refrigerants analysed in this study (Table 2.1). Further discussion on these results is found in Appendix H.

The moderately problematic refrigerants appear located in the central part of the HD (Figure 3.2) and they are characterised by their connections to problematic and least problematic refrigerants.

3.4 Order relations among refrigerant classes

Normally, a HD is interpreted by assessing the order relationships of an object in respect to others, for example looking for problematic or non problematic refrigerants in Figure 3.2. However, if the set under study is classified, an additional structure is given to the objects depicted in the HD, that is, the structure given by the classes. A question arising from these classifications concerns the possibility of studying the order relationships among classes based upon the order relations of the objects in each class.

In this section the supervised classification of the 40 refrigerants into 13 families (chapter 2) is analysed taking advantage of the order relations depicted in Figure 3.2. The refrigerant families are the following, whose respective class-members appear in Table 2.1.

CFC: chlorofluorocarbons.
HFC: hydrofluorocarbons.
HCFC: hydrocholorofluorocarbons.
HC: hydrocarbons.
DFAE: di(fluoroalkyl) ethers.
AFAE: alkylfluoroalkyl ethers.
CM: chloromethanes.
FIM: trifluoroiodomethane.
PFC: octafluorocyclobutane (a perfluorocarbon).
CO_2: carbon dioxide.
BCF: bromochlorodifluorobutane.
DME: dimethyl ether.

NH_3: ammonia.

Although the study of partially ordered sets, their features and properties are an important and active research field of combinatorics, to our knowledge there is little information on the study of order relations among supervised classes whose elements are partially ordered. Therefore, a mathematical procedure was developed to study these relations (Appendix E).

3.4.1 *Order relations among subsets of a poset*

The procedure is based on the definition of three measurements for each pair of subsets; two of them are called dominance degrees and the third one separability degree. Dominance degree indicates the extent to which members of one subset hold higher descriptor values than the members of the other subset; the separability degree quantifies the number of incomparabilities among the members of two subsets. Their mathematical description is the following.

Given a HD of a set P and two disjoint subsets $P_1, P_2 \subset P$, the dominance degree between P_1 and P_2 is given by $\mathrm{Dom}(P_1, P_2) = N_R / N_T$, where $N_R = |\{(x, y), x \in P_1, y \in P_2 \text{ and } y \leq x\}|$ and $N_T = |P_1| \cdot |P_2|$; $|X|$ means the cardinality or number of elements in a finite set X. The separability degree between P_1 and P_2 is given by $\mathrm{Sep}(P_1, P_2) = N_I / N_T$, where $N_I = |\{(x, y), x \in P_1, y \in P_2 \text{ and } y \| x\}|$. Dominance and separability degrees yield real values ranging from 0 to 1; $\mathrm{Dom}(P_1, P_2) = 1$ means that all elements in P_1 have descriptor values higher than the ones of the elements of P_2; in this case it is said that P_1 completely dominates P_2. $\mathrm{Dom}(P_1, P_2) = 0$ means that for no element x of P_1 and y of P_2 the relation $y \leq x$ holds; in this case P_1 does not dominate P_2. Furthermore, $\mathrm{Sep}(P_2, P_1) = 1$ means that all possible relations between P_1 and P_2 are incomparabilities; $\mathrm{Sep}(P_1, P_2) = 0$ means that there are no incomparabilities between P_1 and P_2, and therefore all their relations are ruled by $\leq$.

A theorem relating dominance and separability degrees in the following manner $\mathrm{Dom}(P_1, P_2) + \mathrm{Dom}(P_2, P_1) + \mathrm{Sep}(P_1, P_2) = 1$ was introduced (Appendix E). The results of dominance and separability degrees can be represented in a graph that may or not be represented as a HD depending on the fulfilling of the transitivity axiom, which implies that if $x \leq y$ and $y \leq z$ then $x \leq z$ (Appendix E). The dominance degree was initially introduced in the papers collected in Appendices B and C, where it was applied to the ranking of 35 alkanes according to their fate descriptors in two river scenarios, namely hilly and lowland rivers.

The conditions for fulfilling the transitivity axiom as well as the mathematical details of the dominance and separability degrees, their properties and their implication on the collection of subsets were collected in the paper shown in Appendix E.

As an example of application of dominance and separability degrees, their values for the three subsets shown in Figure 3.1 are the following:

$$\mathrm{Dom}(P_1, P_2) = \frac{|\{(e,c)\}|}{|\{(a,c),(a,f),(e,c),(e,f)\}|} = \frac{1}{4} = 0.25$$

$$\mathrm{Dom}(P_1, P_3) = \frac{|\{(e,d),(e,b)\}|}{|\{(a,d),(a,b),(e,d),(e,b)\}|} = \frac{2}{4} = 0.5$$

$$\mathrm{Dom}(P_2, P_1) = \frac{|\{(f,a),(c,a)\}|}{|\{(f,e),(f,a),(c,e),(c,a)\}|} = \frac{2}{4} = 0.5$$

$$\mathrm{Dom}(P_2, P_3) = \frac{|\{(f,d),(f,b),(c,d),(c,b)\}|}{|\{(f,d),(f,b),(c,d),(c,b)\}|} = \frac{4}{4} = 1$$

$$\mathrm{Dom}(P_3, P_1) = \frac{|\varnothing|}{|\{(d,e),(d,a),(b,e),(b,a)\}|} = \frac{0}{4} = 0$$

$$\mathrm{Dom}(P_3, P_2) = \frac{|\varnothing|}{|\{(d,f),(d,c),(b,f),(b,c)\}|} = \frac{0}{4} = 0$$

$$\mathrm{Sep}(P_1, P_2) = \frac{|\{(e,f)\}|}{|\{(a,c),(a,f),(e,c),(e,f)\}|} = \frac{1}{4} = 0.25$$

$$\mathrm{Sep}(P_1, P_3) = \frac{|\{(a,d),(a,b)\}|}{|\{(a,d),(a,b),(e,d),(e,b)\}|} = \frac{2}{4} = 0.5$$

$$\mathrm{Sep}(P_2, P_3) = \frac{|\varnothing|}{|\{(f,d),(f,b),(c,d),(c,b)\}|} = \frac{0}{4} = 0$$

The percentage of elements in a subset P_n dominating other elements (PD_n) can be calculated by adding the number of chemicals in all subsets P_i dominated by P_n and dividing the result by the number of elements that might be dominated, i.e. $|P| - |P_n|$:

$$PD_n = \frac{\sum |P_i|}{|P| - |P_n|} \cdot 100 \text{, for all } P_i \text{ holding Dom}(P_n, P_i) \geq 0.5 \qquad 3.1$$

The subsets analysed and their dominance relationships, Dom(P_n, P_m), can be depicted in the "dominance diagram" according to PD_n in such a way that a subset P_n is located higher than a subset P_m if $PD_n > PD_m$. The corresponding diagram for the subsets shown in Figure 3.1 is shown in Figure 3.3.

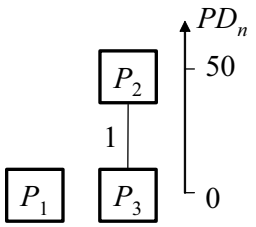


Figure 3.3. Dominance diagram of the classes shown in Figure 3.1. PD_n indicates the percentage of elements in a class dominating elements of other classes.

3.4.2 *Ordering refrigerant classes*

The HD of 40 refrigerants endowed with 13 chemical families is depicted in Figure 3.4. The dominance and separability degree results (Appendices D and H) for these families show that the transitivity axiom is fulfilled therefore its respective dominance diagram can be considered as a Hasse diagram (Figure 3.5). The four subsets DFAE, BCF, CFC and PFC appear as problematic subsets; whereas DME and NH_3 appear as the least environmentally problematic refrigerant classes.

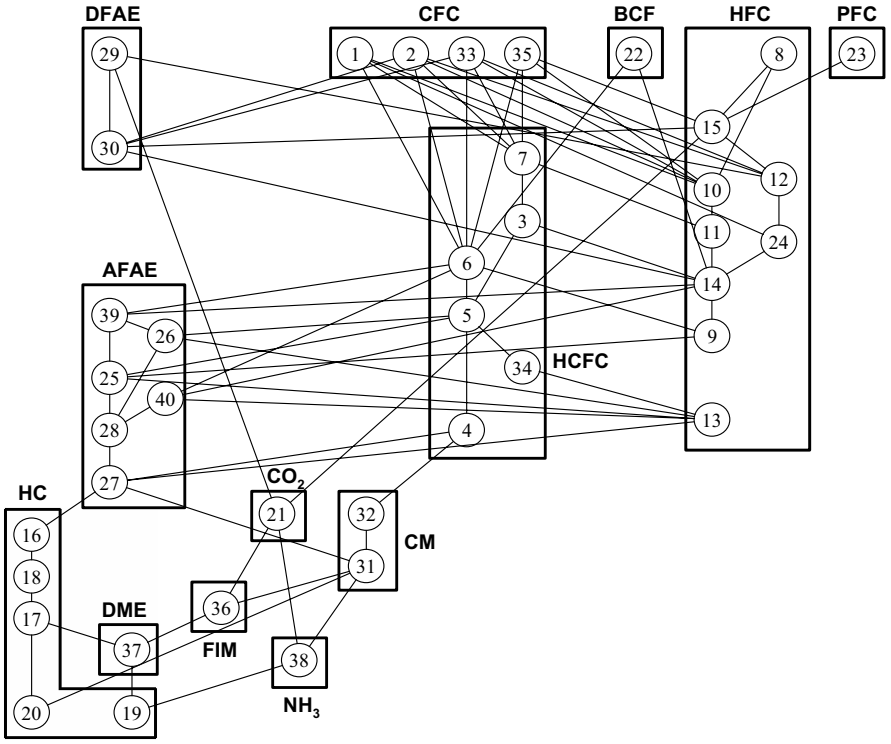


Figure 3.4. Hasse diagram of 40 refrigerants endowed with 13 chemical families.

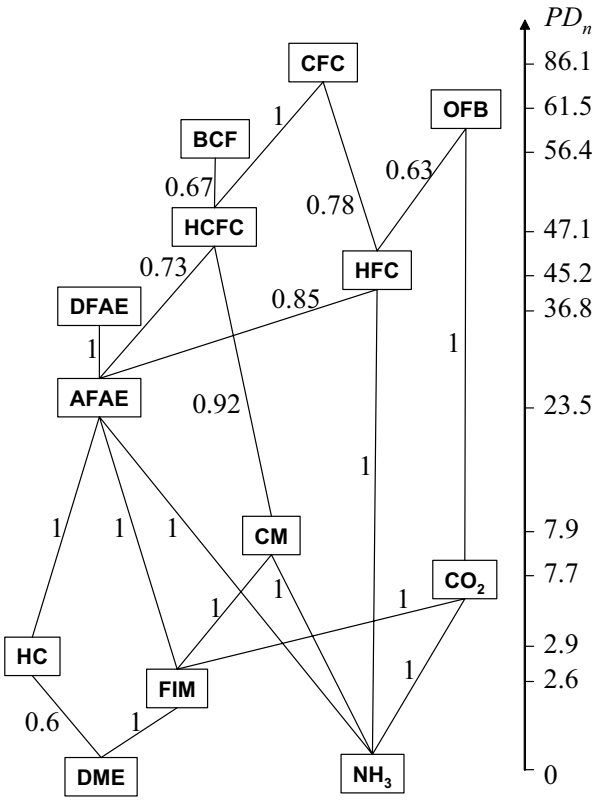


Figure 3.5. Dominance diagram of 13 refrigerant classes. PD_n indicate the percentage of elements dominated by refrigerants in a given class. The value close to each line correspond to Dom(A, B) where A is always a subset located higher than the subset B on the drawing plane.

According to Figure 3.5, CFC is the class that dominates most of the other substances. CFC, PFC and BCF dominate each more than half of the refrigerants considered. The second generation alternatives, HCFC and HFC, dominate less than half of the other substances, which means that they are environmentally better than CFC, PFC and BCF, substances that replace HCFC and HFC. Although problematic HCFCs will be replaced by HFC-blends in refrigeration equipment before 2010 [111], it is worthy to note that HCFC does not dominate HFC; therefore the environmental suitability of these replacements is questionable.

At lower PD_n values (Figure 3.5) appear DFAE and AFAE, both HFEs; DFAE dominates 37 % of the other chemicals, which is a value close to that one found for HFC (45 %). Additionally, none of the classes of chemicals studied dominates DFAE, not even CFC that accounts for the largest percentage of dominated chemicals. In contrast, AFAE, the other group of HFEs, is dominated by problematic refrigerants and also by DFAE. The fact that DFAE dominates AFAE indicates a possible relationship between the distribution of fluorine atoms in the molecular structure of these substances and their environmental properties. DFAE molecules hold fluorine substituents in both alkyl branches attached to the oxygen, whereas AFAE molecules have fluorine substituents in just one alkyl branch.

There are six classes of refrigerants with PD_n values (Figure 3.5) lower than 8 %, they are CM, CO_2, HC, FIM, DME and NH_3. They constitute the most environmentally acceptable refrigerants from the 13 classes studied.

A further discussion on these results was published in papers gathered in Appendices D and H.

Chapter 4: Classification, order and supervised structure: descriptor preferences

4.1 Looking for total orders

Normally, when comparing chemicals, a total order among them is desired because this facilitates the interpretation of the result. Hence, it is possible to find only one highest substance and only one lowest one in respect to the descriptors considered. By applying HDT incomparable objects may appear causing a partial order instead of a total order; this situation is caused by the presence of incomparable objects that arise from the presence of conflictive values among their descriptors. These conflictive values appear when some descriptors of an object reach high values while some others, for the same object, low values. Then, incomparabilities in a HD are the result of conflictive values among descriptors; to avoid these conflictive situations, different ranking procedures [107] opt to aggregate all descriptors at once, which yield the desired total order. These aggregations are often weighted-combinations of descriptors in such a way that the researcher participation is also included through the weights in the aggregation. However, the weighted aggregation implies descriptor compensation, for example a low value in a descriptor offsets large values in other ones. In addition, the ranking interpretation becomes troublesome since the weighted descriptors make the result almost non-transparent because the compensation takes place over all descriptors simultaneously. This problem is avoided by using METEOR (Method of evaluation by order theory) [112-114], an approach based upon HDT, which permits to solve the dilemma among obtaining a total order keeping the HDT transparency but allowing researcher participation. Contrary to some other ranking methods where the descriptor weighting-aggregation is carried out in one step, METEOR aggregates them in a step-by-step procedure permitting to analyse the effects of individual descriptor weights and their compensations. Therefore, the effect of all possible weights on the ranking can be systematically studied.

4.2 METEOR

This methodology is exemplified by its application to the data set shown in Table 3.1, whose values must be normalised to avoid dimensional conflicts when combining descriptors and to give all weights the same interval (0, 1).

Any incomparability $x \parallel y$ indicates a conflict among at least two descriptors for x and y; METEOR aggregates them into a new combined descriptor that is a linear aggregation function. One of the aggregation possibilities METEOR considers is to group similar descriptors into an aggregated one or to aggregate descriptors with a high degree of conflictive potential, which normally are anticorrelated ones. In this example conflictive descriptors are aggregated, namely q_1 and q_2, the least correlated ones according to the Spearman's rank correlation [115] $\rho = 0.8$. When looking at the matrix (Table 3.1) it can be seen that the three incomparabilities $a \parallel d$, $e \parallel f$ and $a \parallel b$ are due to conflictive values between q_1 and q_2 for each element. In general, if $x \parallel y$, because of conflictive values on q_1 and q_2, then there is an aggregated property $\varphi(x)$ for x and another $\varphi(y)$ for y as follows:

$$\varphi(x) = g \cdot q_1(x) + (1-g) \cdot q_2(x) \qquad 4.1$$

$$\varphi(y) = g \cdot q_1(y) + (1-g) \cdot q_2(y) \qquad 4.2$$

where g and $(1 - g)$ are the selected weights (preferences) for q_1 and q_2, respectively; note that the sum of the weights must be equal to 1. If $\varphi(x) = \varphi(y)$ then there must exist a particular g value which is represented by g_c and is called the "crucial value" for the pair x and y under the aggregation φ; this value is the weight for which the order relation between x and y changes and it is given by:

$$g_c = \frac{1}{1 - \dfrac{q_1(x) - q_1(y)}{q_2(x) - q_2(y)}} \qquad 4.3$$

with $q_2(x) - q_2(y) \neq 0$. If the incomparability $e \parallel f$ is considered then the aggregation of q_1 and q_2 yields the following combined properties and a value of $g_c = 0.56$; for $a \parallel d$ $g_c = 0.24$ and 0.71 for $a \parallel b$. The changes in the order relations between each one of the three incomparable pairs can be seen in Figure 5.1, where the diagrams are drawn using HDT, that is considering φ and q_3 without any aggregation among them. The diagrams depicted in Figure 4.1 are all of them linear orders because the aggregation of q_1 and q_2 breaks all the incomparabilities in the original HD (Figure 3.1C). Hence, the order of the elements regarding φ is the same as the order held by q_3.

A weight $g < 0.24$ always yields an order as depicted in Figure 4.1A; if the weight is shifted to $g > 0.24$ then the diagram shown in Figure 4.1B is obtained, where the change in the order relation between the pair a, d can be seen after passing the crucial value $g_c = 0.24$. In the same way, a value of $g > 0.56$ produces the diagram depicted in Figure 4.1C, where the change occurs over the pair e, f when surpassing $g = 0.56$, which is the crucial value for $e \parallel f$. Finally, the selection of $g > 0.71$ (Figure 4.1D) induces that the relation $a > b$ changes to $b < a$ because it has exceeded the crucial value for $b \parallel$

a. Note that each crucial value only affects the incomparable pair that generates it and the other pairs keep their mutual order relations. For example, for all *g* values greater than 0.24 $d > a$ is held even if $g > 0.56$ or 0.71 (Figure 4.1).

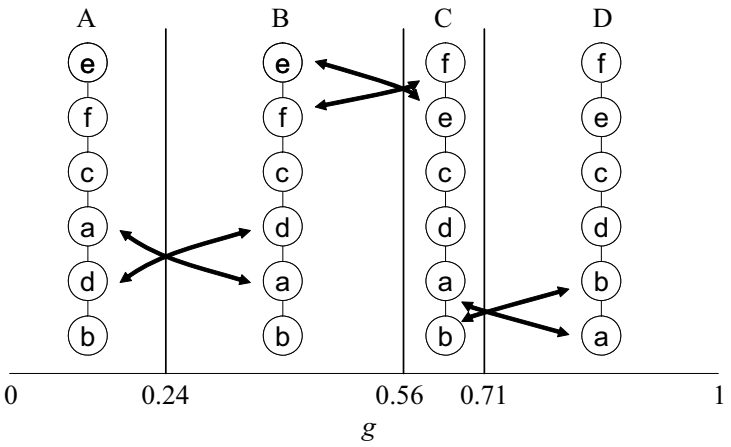


Figure 4.1. *g*-spectrum for the aggregation of q_1 and q_2 in the data matrix shown in Table 3.1; each HD represents the total order of its corresponding stability field. The arrows indicate the changes on the order relations.

The distribution of g_c values along $0 < g < 1$ is called the "*g*-spectrum" and each one of the regions where *g* can take different values without changing the order relations among the elements is called a "stability field" (Appendix F). Hence, in Figure 4.1 there are four stability fields and each one is characterised by one HD. In consequence, an important aspect of this method is that it permits to have an upper estimate of the number of changes in the original ranking when particular priorities of the properties are selected.

Often, the linear order is not reached after the first aggregation of conflictive descriptors and then a further aggregation can be performed and a series of stability fields arising from the previous ones result. This kind of situation is considered in the example shown in reference (Appendix G). Another aspect to take into consideration when aggregating descriptors is the fact that many g_c values may result because of many incomparabilities among elements in the set, therefore it is not informative to plot the entire diagrams for each stability field. In those cases clustering the g_c values is recommended in order to plot the diagrams between clusters of g_c values. Each one of these clusters is called a "hot spot" (Appendix F). Then, a hot spot is a region of the *g*-space where many order relations change; these hot spots may be regarded as zones of high order-instability, while a stability field represents a region where any of the weights covered by it does not affect the order.

An important aspect of this aggregation procedure is that by aggregation of all descriptors, always a linear order is found, which in fact is a linear extension of the original HD (without descriptor priorities). Hence, the procedure ensures that all order relations present in the original HD are preserved in the final linear extension, therefore changes produced in the HD are only those related to incomparable elements.

A mathematical discussion on METEOR and its properties appears in Appendix F.

4.3 Looking for totally ordered refrigerants

METEOR was applied to the set P = {1, 2, 6, 7, 8, 16, 21, 22, 23, 29, 32, 33, 35, 36, 37, 38, 39, 40} of maximal chemicals of each refrigerants subset (Table 2.1, Figure 3.2) in order to explore the order relations among these problematic substances. The most anticorrelated properties were ALT and ODP with a Spearman's rank correlation of – 0.1. Hence, the first aggregation for a refrigerant x was

$$\varphi(x) = g \cdot ALT(x) + (1-g) \cdot ODP(x) \qquad 4.4$$

The g_c-values found were clustered using HCA and 12 stability fields resulted along the weights g. The HD of each one of these stability fields was drawn based upon φ and GWP but none of them corresponded to a linear order, therefore another aggregation was performed, namely the one of φ with GWP, through the following function

$$\psi(x) = h \cdot \varphi(x) + (1-h) \cdot GWP(x) \qquad 4.5$$

The h-values associated to each stability field in the first aggregation were clustered using HCA and finally 109 stability fields were obtained as the result of the two aggregations. Their distribution along the values g and h can take is depicted in Figure 4.2, where black regions represent hot spots and the coloured ones stability fields. HCA was performed on these stability fields in order to look for their similarities, the results are also included in Figure 4.2 where stability fields equally coloured are similar ones. In the present work the similarity between any two stability fields is understood as the resemblance between their linear rankings, which can be calculated with the W-index [100, 108], a mathematical function quantifying the number of different order relations among the elements of two linear orders.

In total the 109 stability fields could be condensed into 27 clusters of similar stability fields. The average ranking for each cluster was calculated and the two average rankings associated to the two clusters gathering most of the stability fields are depicted (Figure 4.2).

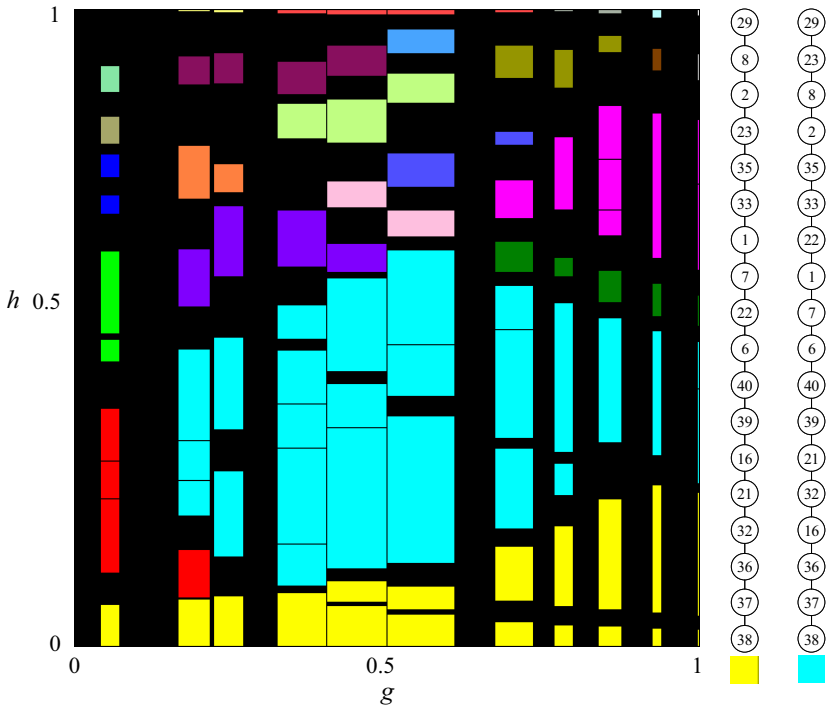


Figure 4.2. Stability fields found by the step-by-step aggregation of ALT, ODP and GWP. The two total orders depicted correspond to the two clusters of stability fields holding the major number of values that weights *g* and *h* can take.

For these two clusters of stability fields the most problematic substance is C_2HF_5O (29), which is a hydrofluoroether and the least problematic substance is ammonia (38). Further discussion and additional results on the application of METEOR to the refrigerants are found in Appendix G.

Chapter 5: Extended summary

5.1 On the developed methods

5.1.1 *Cluster index*

In a similarity study, clusters gather information regarding the resemblance of the elements studied, therefore a method contrasting similarity classifications must be grounded on the concept of cluster. Because of the lack of such procedures, the cluster index was developed (Appendix A). It permits:
(1) to assess the effect of element representations, that is, real number descriptors, such as topological indices and physicochemical properties; or fingerprints [62] representing the presence or absence of different attributes in the elements to classify.
(2) to study the influence of the similarity functions on the classification results, that is, the effect of different such as Euclidean and Hamming distances.
(3) to evaluate the effect of applying different grouping methodologies on the classification results, such as single and complete linkage.

In order to study the statistical distribution of cluster index values, it is important to analyse its values for random trees defined over different sets of various cardinalities.

5.1.2 *On dominance and separability degrees*

In the paper shown in Appendix E implications are discussed that particular values of dominance and separability degrees have over the collection of subposets of a HD. Hence, two binary relations based upon dominance and separability were introduced and some of their properties were studied. Special attention was dedicated to the lack of transitivity of the relation arising from the dominance degree and it was proved that, when the partial order is equipped with a collection of paths including comparable elements where at least one element of the sequence of subposets compared is considered, the relation becomes a transitive one on the collection of subsets where it is applied. This kind of transitivity dependent on the paths of comparable elements between the compared subposets keeps a certain resemblance to investigations on fuzzy transitive relations developed by De Baets [116, 117]. Hence, the study of the transitive relation as a fuzzy transitive one and its implications must be explored in

forthcoming investigations.

Similar researches for comparing subsets of a given set P have been conducted in observational studies where the relations between two subsets of P are measured by their coherence [118, 119]. In the present work was found that the functions used for measuring coherence are in fact functions of dominance and separability degrees. It is considered that application of dominance and separability degrees to observational studies would permit to give a detailed description of the cause-effect pattern these studies look for.

5.1.3 *On METEOR*

The aggregation procedure described therein includes nested aggregations, for example {{ALT, ODP}, GWP} in the refrigerant's case. In this situation, the g-spectrum is related to the one of the second aggregation. Hence, g determines the values h can take and if a further aggregation were necessary involving a weight k, k would be related to g and h. A different approach where the descriptors are not nested-aggregated can be found in Appendix F where four descriptors q_1, q_2, q_3 and q_4 were pair-aggregated, namely $\{q_1, q_2\}$ and $\{q_3, q_4\}$. Although a linear aggregation was performed in the current manuscript, METEOR is not restricted to this kind of combinations; in fact other aggregation functions must be explored and the assessment of their results on the ranking must be equally carried out.

In paper shown in Appendix G the use of prioritisation trees was introduced as a graphical tool to explore relationships among total orders after the performance of more than three aggregations. This representation was mooted because a graphical representation as the one shown in Figure 4.2 only accounts up to three aggregations, which is the case of a three-dimensional space of aggregation-weights. The use of prioritisation trees or graphical representations as that depicted in Figure 4.2 allows to analyse which kinds of priorities on different aggregations permit to find common rankings.

For METEOR applications involving equal or less than three aggregations, the graphical representations of the type shown in Figure 4.2 can be used to explore the similarities among stability fields. In this respect, the selection of different similarity levels among the stability fields allows to see how the similarities evolve with respect to descriptor priorities in such a way that for each similarity level a collection of stability field neighbourhoods can be constructed, whereby each one of these neighbourhoods systems becomes a topological basis permitting to study topological relationships among different subsets of stability fields through application of the chemotopological method.

It would be interesting to estimate refrigerant environmental properties by using QSPR methodologies. In fact, in this dissertation the molecular descriptors derived from MOLGEN-QSPR were used to predict ODP, GWP and ALT values but the small number of refrigerants for which these experimental values are available, less than 40, makes it difficult to have a statistical relevant molecular sample. Additionally, the high diversity of refrigerant molecular structures considered split the set of molecules into smaller subsets which makes it more difficult the application of QSPR methods. In order to perform a QSPR study it would be appropriate to consider homologous chemical series, e.g. floroalkanes and chloroflouoroalkanes to name but a few. However, for obtaining successful models, it is necessary to have accurate experimental property values for these substances, which are still lacking in the scientific literature.

Another mathematical tool that can be used either to derive conclusions or to raise hypothesis about refrigerants and their properties is the Formal Concept Analysis [120, 121], which makes use of the presence/absence of different properties of the objects to study and derives a lattice of concepts, namely a partially ordered set. A concept is a pair (B, C) where B is a subset of objects, e.g. refrigerants, and C a subset of properties. By the application of this technique it is also possible to obtain implications relating objects and properties. A simple implication that could be derived from the case of refrigerants is that the presence of chlorine atoms in a refrigerant with no hydrogen atoms is related to a high ozone depletion potential.

5.2 On refrigerants

5.2.1 *Classification*

Environmental properties. The results of contrasting eight classification methodologies of refrigerants characterised according to their environmental properties show that there are rather stable clusters, implying stable similarity relationships among environmental properties of the studied substances. The most similar classification was the one obtained using the Euclidean distance as similarity function and the unweighted average linkage.

None of the environmental clusters of refrigerants resulted to be a refrigerant family, i.e. CFC, HFC, HCFC, HC, DFAE, AFAE, CM, FIM, PFC, CO_2, BCF, DME and NH_3.This implies that from the partition of refrigerants into families is not possible to draw conclusions regarding refrigerant environmental properties.

Thermodynamic properties. Refrigerants classified based upon thermodynamic properties do not constitute stable clusters, therefore actual similarity relationships on their thermodynamic properties do not hold. None of the clusters derived from thermodynamic properties resulted to be a family of refrigerants; consequently, from these families cannot be derived conclusions on refrigerants' thermodynamic properties.

Molecular descriptors. Because of the high correlation among molecular descriptors caused by the small size of several refrigerants, the number of descriptors could be reduced from 388 to 15 informative and uncorrelated ones. Classifications derived from these descriptors yield unstable clusters; none of them matches with any refrigerant family, therefore classifications from molecular descriptors do not lead to the refrigerant classification into chemicals families.

Classifications derived from molecular descriptors and those yielding families of refrigerants share a common feature, both result from the analysis of the molecular structure of the refrigerants studied. Therefore, the agreement of these classifications is expected. However, results here shown disagree with this hypothesis. The reason lies on the different kind of information contained in the molecular descriptors and in the classification arguments considered to build refrigerant families. Molecular descriptors account for chemical composition, connectivity of atoms involved in the molecule and for the molecular geometry. In contrast, refrigerant families are built from the specific presence or absence of particular molecular structural features, namely functional groups. Hence, for example, if a molecule contains two alkyl or halo-alkyl groups bonded to one oxygen with hybridisation sp^3, then the molecule is considered as an ether without regard of its molecular shape or complete atomic connectivity. In order to find an agreement between both classifications, different descriptors are needed which account for the presence/absence of particular functional groups.

The most similar classifications among the eight considered based upon molecular descriptors was the one obtained with the Euclidean similarity function and the unweighted average linkage grouping methodology. This clustering methodology yielded also the most similar classification for environmental properties, which raises the question on its centrality in the mathematical space of classifications obtained from the set of refrigerants analysed. This study must be addressed in forthcoming investigations.

5.2.2 *Ordering*

When applying HDT to the set of 40 refrigerants the eight most problematic substances were halogenated compounds belonging to problematic families such as CFCs, PFC, HCFCs and HFCs. The appearance of C_2HF_5O, an HFE, in this group of substances is a matter of concern since these

substances were introduced to overcome the problems presented by CFCs and some of their first replacements, namely HCFCs and HFCs. The main reason of the high position of C_2HF_5O in the total order is its high GWP (14,800 units), which is associated to a high ALT (165 years). Tsai [57] has pointed out the GWP drawbacks of some HFEs, including C_2HF_5O; however, not all HFEs hold as high ALT and GWP as C_2HF_5O. In fact, some other HFEs are still under study as possible CFCs, HCFCs and HFCs replacements because of their favourable environmental and appropriate thermodynamic properties [122]. In the HD depicted in Figure 3.4 can be seen that there is only one most problematic HFE in the DFAE subset, namely C_2HF_5O and that the other HFEs, included in DFAE and AFAE subsets, are located lower than C_2HF_5O on the drawing plane, which stresses the conclusion on their not so problematic environmental properties. In fact, the calculations of the dominance degree (chapter 3) show that some HFEs, particularly AFAEs are dominated by or located lower than CFCs, HCFCs, HFCs and PFC.

In the study considering properties prioritisation (chapter 4), C_2HF_5O is the top of the linear orders for cases corresponding to high GWP priorities associated to any priority of ALT and ODP (Figure 4.2). This result remarks the effect of the problematic GWP of this HFE. It is worthy to note that the total orders shown in Figure 4.2 hold C_2HF_5O on the top of the ranking and this total orders correspond to the regions with highest probability of occurrence when randomly selecting *g* and *h* weights for the aggregations, which means that many priority combinations yield C_2HF_5O as a problematic refrigerant.

It is also important to remark the fact that DFAEs are more problematic than AFAEs, which suggests a relationship between the environmental properties and the distribution of fluorine atoms in the alkyl branches of these ethers. DFAEs possess fluorine atoms in both alkyl chains attached to the ethylic oxygen whereas AFAEs hold fluorine only in one of the chains. It would be important to perform an ordering study with more HFEs belonging to both subsets, DFAE and AFAE, in order to test this hypothesis.

In general, the total orders found from prioritisation among different properties are quite similar for high GWP priorities, which correspond to lowest h-values in the second aggregation (chapter 4). Dimethyl ether (37) is the least problematic substance for high GWP priorities, associated to high ALT and low ODP ones.

Ammonia (38) reaches the lowest place in the total orders for the majority of stability fields clusters where an intermediate GWP priority is hold which in turn is associated to any priority of ODP and ALT.

Refrigerant 22 (bromochlorodifluoromethane) is maximal for low GWP priorities, high ones of ODP and low ones of ALT; 23 (octafluorocyclobutane) is maximal for low GWP priorities associated to relatively high ALT and low ODP ones. Note that bromochlorodifluoromethane and octafluorocyclobutane are non-hydrogenated substances with a high degree of halogenation.

Figure 4.2 is a versatile tool to predict the effect of prioritising ALT, ODP and GWP, for example, if one is interested in a total order with 10 % priority for ALT, 40 % for ODP and 50 % for GWP, after simple algebraic manipulations of the weighting factors in Eq. 4.5, it results that $g = 0.2$ and $h = 0.5$, therefore it corresponds to the cluster of stability fields in Figure 4.2 where octafluorocyclobutane is the most problematic substance and ammonia the least one. If for example, a total order performed prioritising ODP is carried out, then the following priorities could be set up: ALT 5%, ODP 75% and GWP 20%, in such a case $g = 0.0625$ and $h = 0.8$, which corresponds to the cluster of stability fields in Figure 4.2 where bromochlorodifluoromethane is the most problematic refrigerant and ammonia the least one.

Classifications derived from molecular descriptors and classifications yielding families of refrigerants share a common feature, both result from the analysis of the molecular structure of the refrigerants studied. In such a case, it is expected an agreement in both classifications. However, the results here shown disagree with this hypothesis. The reason lies on the different kind of information contained in the molecular descriptors and in the classification arguments considered to build refrigerant families. Molecular descriptors account for chemical composition, connectivity of atoms involved in the molecule and for the molecular geometry. In contrast, refrigerant families are built from the specific presence or absence of particular molecular structural features, namely functional groups. Hence, for example, if a molecule contains two alkyl or halo-alkyl groups bonded to one oxygen with hybridisation sp^3, then the molecule is considered as an ether without regard of its molecular shape or complete atomic connectivity. In order to find an agreement between both classifications, different descriptors are needed which account for the presence/absence of particular functional groups.

Acknowledgements

The idea of applying the Hasse diagram technique to refrigerants' assessment is associated with the project "Comparative life cycle assessment of certain CFC-replacements in different technical applications" (Project number 81-00213381) of the Bavarian State Ministry of the Environment, Public Health and Consumer Protection (Germany).

This research would not have been possible without the support of Nubia, my girlfriend. She has given me not only affection but also scientific support in a field that is not directly related to her activities; that is one of the advantages of being so intelligent.

I owe a debt of gratitude too large to measure to Rainer Brüggemann for showing me the marvellous world of the partially ordered sets and many other blossoms of discrete mathematics. I am thankful for his support, for encourage me to develop my own ideas.

Special thanks are given to Hartmut Frank for providing the possibility to write this thesis at the Department of Environmental Chemistry and Ecotoxicology of the University of Bayreuth. I am thankful for permitting me to work on the refrigerant project together with Monika Weckert, a exceptional person, who helped me in the most simple and difficult things of my residence in Bayreuth. I hope, one day, to be able to be as kind with her as she was with me.

I really appreciate the excellent lessons of English grammar and writing style Silke Gerstmann gave me during the writing of several manuscripts.

I am also in debt with Adalbert Kerber at the Department of Mathematics of the University of Bayreuth. He gave me a wonderful office in his department and also access to the software MOLGEN-QSPR and to invaluable literature. But my gratitude mainly concerns with his vivid interest on previous researches I have developed with my team in Colombia, which I have applied in the current dissertation. I am glad to know we have common ideas and I appreciate so much the time he has spent in edifying mathematical discussions with me.

I would like to thank Ralph Gugisch for his valuable comments, time and coffee spent during interesting mathematical discussions about the possibilities our paper on ranking patterns opens for new research in the field of ranking.

I thank my colleagues Héber Mesa at the Departamento de Matemáticas of the Universidad del Valle (Colombia) and Eugenio J. Llanos at the Scio Corporación (Colombia) for the great scientific discussions we had during the development of the cluster index. I also express my gratitude to Peter Willett and Yogendra Patel at the Department of Information Studies of the University of Sheffield (UK), and Barnard Chemical Information Limited (nowadays, Digital Chemistry (UK)) for the access they permitted to the Tanimoto data matrix used in the paper where the cluster index was introduced. I did not produce any paper with José L. Villaveces during my time in Bayreuth, but I am quite sure his ideas are always behind all my writings. I thank so much that wonderful person for having sown his scientific ideas in me years ago and for keeping watering them.

I appreciate the scientific support given by Douglas J. Klein at the Department of Marine Sciences of the Texas A&M University at Galveston (USA) and Andrés Bernal at the Grupo de Química Teórica of the Universidad Nacional de Colombia at Bogotá (Colombia) for their valuable comments and suggestions about our paper on the mathematical properties of the dominance and separability degrees.

I cannot finish these acknowledgements without expressing my gratitude to quite important people of the Department of Environmental Chemistry and Ecotoxicology at the University of Bayreuth. I thank Agnes Bednorz for her support and familiar smile, Benjamin Schmidt for providing us technical assistance, Irmgard Lauterbach for her help in the academic procedures needed to start and end this study, Huong Ngo for her amiability and support and Abed A. Qader for his orientation and friendship upon arriving in Bayreuth.

Finally, I express my gratitude to the Universidad de Pamplona and COLCIENCIAS - Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología “Francisco José de Caldas” for the financial support received during the time spent in this research. I specially thank Álvaro Gonzalez Joves, rector of the Universidad de Pamplona, for his strongly decision on supporting mathematical chemistry activities.

References

1. James, P.; Thorpe, N. *Ancient inventions*; Ballantine Books: New York, USA, 1995. Chapter 7, pp. 320-323.

2. Dossat, R. J.; Horan, T. J. *Principles of refrigeration*; Prentice Hall: Upper Saddle River, USA, 2002.

3. Cullen, W. Of the cold produced by evaporating fluids and of some other means of producing cold. *Essays and Observations Physical and Literary Read Before a Society in Edinburgh and Published by Them II*, **1756**.

4. Thompson, R. J. Freon, a refrigerant. *Ind. Eng. Chem.* **1932**, *24*, 620-623.

5. Calm, J. M.; Didion, D. A. Proceedings of the ASHRAE/NIST refrigerants conference; Gaithersburg, USA, 6-7 October 1997.

6. Midgley, T. Jr. From the periodic table to production. *Ind. Eng. Chem.* **1937**, *29*, 241-244.

7. McCulloch, A. CFC and Halon replacements in the environment. *J. Fluorine Chem*. **1999**, *100*, 163-173.

8. Powell, R. L. CFC phase-out: Have we met the challenge? *J. Fluorine Chem*. **2002**, *114*, 237-250.

9. Midgley, T. Jr.; Henne, A. L.; McNary, R. R. (Frigidaire Corporation, USA). Heat transfer. US Patent 1,833,847; filed Feb. 8, 1930; patented Nov. 24, 1931.

10. Midgley, T. Jr.; Henne, A. L. Organic fluorides as refrigerants. *Ind. Eng. Chem.* **1930**, *22*, 542-545.

11. Perkins, R.; Cusco, L.; Howley, J.; Laesecke, A.; Matthes, S.; Ramires, M. L. V. Thermal conductivities of alternatives to CFC-11 for foam insulation. *J. Chem. Eng. Data* **2001**, *46*, 428-432.

12. Daudt, H. W.; Youker, M. A. (Kinetic Chemicals Inc., USA). Production of fluorine-containing carbon compounds. US Patent 2,062,743; filed Jun. 12, 1935; patented Dec. 1, 1936; also US Patents 2,005,710; 2,005,705; 2,005,706; 2,005,707; 2,005,708; 2,005,709.

13. Pool, R. The elusive replacements for CFCs. *Science* **1988**, *242*, 666-668.

14 Lovelock, J. E.; Maggs, R. J.; Adlard, E. R. Gas-phase coulometry by thermal electron attachment. *Anal. Chem.* **1971**, *43*, 1962-1965.

15. Lovelock, J. E.; Maggs, R. J.; Wade, R. J. Halogenated hydrocarbons in and over the Atlantic. *Nature* **1973**, *241*, 194-196.

16. Lovelock, J. E. Atmospheric halocarbons and stratospheric ozone. *Nature* **1974**, *252*, 292-294.

17. Molina, M. J.; Rowland, F. S. Stratospheric sink for chlorofluoromethanes: Chlorine atom-catalysed destruction of ozone. *Nature* **1974**, *249*, 810-812.

18. Rowland, F. S.; Molina, M. J. Ozone depletion - 20 years after the alarm. *Chem. Eng. News* **1994**, *72*, 8-13.

19. Farman, J. C.; Gardiner, B. G.; Shanklin, J. D. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature* **1985**, *315*, 207-210.

20 UNEP. *Montreal protocol on substances that deplete the ozone layer*; United Nations Environment Programme: Nairobi, Kenya, 1987.

21. UNEP. *Montreal protocol on substance that deplete ozone layer*; United Nations Environment Programme: Montreal, Canada, 1998.

22. Dekant, W. Toxicology of chlorofluorocarbon replacements. *Environ. Health Persp*. **1996**, *104*, 75-83.

23. Hayman, G.; Derwent, R. D. Atmospheric chemical reactivity and ozone-forming potentials of potential CFC replacements. *Environ. Sci. Technol*. **1997**, *31*, 327-336.

24. Kurylo, M. J.; Orkin, V. L. Determination of atmospheric lifetimes via the measurement of OH radical kinetics. *Chem. Rev.* **2003**, *103*, 5049-5076.

25. Ravishankara, A. R.; Lovejoy, E. R. Atmospheric lifetime, its application and its determination: CFC-substitutes as a case study. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 2159-2169.

26. Bolin, B.; Rodhe, H. Note on concepts of age distribution and transit-time in natural reservoirs. *Tellus* **1973**, *25*, 58-62.

27. Trapp, S.; Matthies M. Generic one-compartment model for uptake of organic chemicals by foliar vegetation. *Environ. Sci. Technol*. **1995**, *29*, 2333-2338.

28. Prather, M. J. Time scales in atmospheric chemistry: Theory, GWPs for CH_4 and CO, and runaway growth. *Geophys. Res. Lett.* **1996**, *23*, 2597-2600.

29. Wuebbles, D. J. Chlorocarbon emission scenarios - potential impact on stratospheric ozone. *J. Geophys. Res.-Oc. Atm.* **1983**, *88*, 1433-1443.

30. Solomon, S.; Wuebbles, D. Ozone depletion potentials, global warming potentials, and future chlorine/bromine loading; Scientific assessment of ozone depletion: 1994; Report No. 37; Global ozone research and monitoring project; World Meteorological Organization: Geneva, 1995.

31. Madronich, S.; Velders, G. J. M. Halocarbon scenarios for the future ozone layer and related consequences; Scientific assessment of ozone depletion: 1998; Report No. 44; Global ozone research and monitoring project; World Meteorological Organization: Geneva, 1999.

32. Solomon, S.; Albritton, D. L. Time-dependent ozone depletion potentials for short-term and long-term forecasts. *Nature* 1992, 357, 33-37.

33. Daniel, J. S.; Solomon, S.; Albritton, D. L. On the evaluation of halocarbon radiative forcing and global warming potentials. *J. Geophys. Res.-Atmos.* **1995**, *100*, 1271-1285.

34. *Climate Change 1990: The intergovernmental panel on climate change scientific assessment*; Cambridge University Press: New York, USA, 1990.

35. *Climate Change 2001: The scientific basis; contribution of working group I to the third assessment report of the intergovernmental panel on climate change*; Cambridge University Press: New York, USA, 2001.

36. Tsai, W-T. A Review of environmental hazards and adsorption recovery of cleaning solvent hydrochlorofluorocarbons (HCFCs). *J. Loss Prevent. Proc.* **2002**, *15*, 147-157.

37. Frank, H.; Renschen, D.; Klein, A.; Scholl, H. Trace analysis of airborne haloacetates. *HRC-J. High Res. Chrom.* **1995**, *18*, 83-88.

38. Frank, H.; Klein, A.; Renschen, D. Environmental trifluoroacetate. *Nature* **1996**, *382*, 34.

39. Zehavi, D.; Seiber, J. N.; Wujcik, C. Development and application of a method for analyzing trifluoroacetic-acid in environmental-samples. *Am. Chem. Soc. Environ. Chem. Bull.* **1995**, *209*, 22.

40. Zehavi, D.; Seiber, J. N. An analytical method for trifluoroacetic acid in water and air samples using headspace gas chromatographic determination of the methyl ester. *Anal. Chem.* **1996**, *68*, 3450-3459.

41. Wujcik, C. E.; Zehavi, D.; Seiber, J. N. Trifluoroacetic acid levels in 1994-1996 fog, rain, snow and surface waters from California and Nevada. *Chemosphere* **1998**, *36*, 1233-1245.

42. Boutonnet, J. C.; Bringham, P.; Calamari, D.; de Rooij, C.; Franklin, J.; Kawano, T.; Libre, J. M.; McCulloch, A.; Malinverno, G.; Ocom, J. M.; Rusch, G. M.; Smythe, K.; Sobolev, I.; Thompson, R.; Tiedije, J. M. Environmental risk assessment of trifluoroacetic acid. *Hum. Ecol. Risk Assess.* **1999**, *5*, 59-124.

43. Jordan, A.; Frank, H. Trifluoroacetate in the environment: Evidence for sources other than HFC/HCFCs. *Environ. Sci. Technol*. **1999**, *33*, 522-527.

44. Harnisch, J.; Frische, M.; Borchers, R.; Eisenhauer, A.; Jordan, A. Natural fluorinated organics in fluorite and rocks. *Geophys. Res. Lett.* **2000**, *27*, 1887-1890.

45. Berends, A. G., de Rooij, C. G.; Shin-Ya, S.; Thompson, R. S. Biodegradation and ecotoxicity of HFCs and HCFCs. *Arch. Environ. Contam. Toxicol*. **1999**, *36*, 146-151.

46. Tsai, W-T. An overview of environmental hazards and exposure risk of hydrofluorocarbons (HFCs). *Chemosphere* **2005**, *61*, 1539-1547.

47. Ravishankara, A. R.; Turnipseed, A. A.; Jensen, N. R.; Barone, S.; Mills, M.; Howard, C. J.; Solomon, S. Do hydrofluorocarbons destroy stratospheric ozone? *Science*, **1994**, *263*, 71-75.

48. Good, D.A.; Francisco, J. S. Atmospheric Chemistry of alternative fuels and alternative chlorofluorocarbons. *Chem. Rev.* **2003**, *103*, 4999-5023.

49. Tullo, A. H. The switch is on for refrigerants. *Chem. Eng. News* **2006**, *84*, 24-25.

50. Wallington, T. J.; Nielsen, O. J. Atmospheric chemistry and environmental impact of hydrofluorocarbons (HFCs) and hydrofluoroethers (HFEs). In *The handbook of environmental chemistry*; Neilson, A. H., Ed.; Springer: Berlin, Germany, 2002; vol. 3, Part N-Organofluorines.

51. WMO. *Scientific assessment of ozone depletion: 2002*; World Meteorological Organization: Geneva, Switzerland, 2002.

52. UNFCCC. *Kyoto protocol to the United Nations framework convention on climate change*; United Nations Framework Convention on Climate Change: Kyoto, Japan, 1997.

53. Devotta, S.; Gopichand, S.; Pendyala, V. R. Comparative assessment of some HCFCs, HFCs and HFEs as alternatives to CFC11. *Int. J. Refrig.* **1994**, *17*, 32-39.

54. Sekiya, A.; Misaki, S. A continuing search for new refrigerants. *Chem. Tech*. **1996**, *26*, 44-48.

55. Bivens, D. B.; Minor, B. H. Fluoroethers and other next generation fluids. *Int. J. Refrig.* **1998**, *21*, 567-576.

56. Sekiya, S.; Misaki, S. The potential of hydrofluoroethers to replace CFCs, HCFCs and PFCs. *J. Fluorine Chem*. **2000**, *101*, 215-221.

57. Tsai, W-T. Environmental risk assessment of hydrofluoroethers (HFEs). *J. Hazard. Mater.* **2005**, *A119*, 69-78.

58. Lardelli, M. Scientists need to confront economists about peak oil. *Nature* **2007**, *446*, 257.

59. Schindler, J.; Zittel, W. Alternative world energy outlook 2006: A possible path towards a sustainable future. *Adv. Solar En*. **2007**, *17*, 1-44.

60. Shine, K. P.; Sturges, W. T. CO_2 is not the only gas. *Science* **2007**, *315*, 1804-1805.

61. Wang, X. Z.; McGreavy, C. Automatic classification for mining process operational data. *Ind. Eng. Chem. Res*. **1998**, *37*, 2215-2222.

62. Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. In *Reviews in computational chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; Wiley-VCH: Weinheim, Germany, 2002; Vol. 18, pp 1-40.

63. Höppner, F.; Klawonn, F.; Kruse, R.; Runkler, T. *Fuzzy cluster analysis*; Wiley: Chichester, UK, 1999. Chapters 1 and 2, pp 1-60.

64. Everitt, B. S. *Cluster analysis*; Edward Arnold: Bristol, UK, 1993; Chapter 1, pp 1-10.

65. Handl, J.; Knowles, J.; Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201-3212.

66. Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

67. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.

68. Trinajstic, N.; Gutman, I. Mathematical chemistry. *Croat. Chem. Acta* **2002**, *75*, 329-356.

69. Kerber, A.; Laue, R.; Wieland, T. Discrete mathematics for combinatorial chemistry. Workshop on discrete mathematical chemistry, Dimacs Center, Piscataway, USA, 1998.

70. Katritzky, A. R.; Gordeeva, E. V. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835-857.

71. Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520-525.

72. Lahuerta, L.; Fuster, Y.; Duart, M. J.; Antón, G. M.; García, R.; Gálvez, J.; Martínez, J. Prediction of the chemiluminescent behavior of pharmaceuticals and pesticides. *Anal. Chem.* **2001**, *73*, 4301-4306.

73. Mosier, P. D.; Counterman, A. E.; Jurs, P. C.; Clemmer, D. E. Prediction of peptide ion collision cross sections from topological molecular structure and amino acid parameters. *Anal. Chem*. **2002**, *74*, 1360-1370.

74. Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503-511.

75. Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: A simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686-698.

76. Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN-QSPR, a software package for the study of quantitative structure property relationships. *MATCH Commun. Math. Comput. Chem.* **2004**, *51*, 187-204.

77. Rücker, C.; Meringer, M.; Kerber, A. QSPR using MOLGEN-QSPR: The example of haloalkane boiling points. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2070-2076.

78. Rücker, C.; Meringer, M.; Kerber, A. QSPR using MOLGEN-QSPR: The challenge of fluoroalkane boiling points. *J. Chem. Inf. Model.* **2005**, *45*, 74-80.

79. MOLGEN QSPR. http://www.mathe2.uni-bayreuth.de/molgenqspr (accessed September 5, 2007).

80. Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332-2339.

81. Randić, M. Novel molecular descriptor for structure-property studies. *Chem. Phys. Lett.* **1993**, *211*, 478-483.

82. Randić, M. Molecular bonding profiles. *J. Math. Chem*. **1996**, *19*, 375-392.

83. Molecular descriptors, the free online resource. http://www.moleculardescriptors.eu/tutorials/T3_moleculardescriptors_requirements.pdf (accessed September 9, 2007).

84. CAS. http://www.cas.org/cgi-bin/cas/regreport.pl (accessed September 8, 2007).

85. Molecular descriptors, the free online resource. http://www.moleculardescriptors.eu/dataset/dataset.htm (accessed September 9, 2007).

86. Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 68-75.

87. Restrepo, G.; Villaveces, J. L. From trees (dendrograms and consensus trees) to topology. *Croat. Chem. Acta* **2005**, *78*, 275-281.

88. Restrepo, G.; Mesa, H.; Villaveces, J. L. On the topological sense of chemical sets. *J. Math. Chem*. **2006**, *39*, 363-376.

89. Restrepo, G.; Llanos, E. J.; Mesa, H. Topological space of the chemical elements and its properties. *J. Math. Chem*. **2006**, *39*, 401-416.

90. Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. In *The mathematics of the periodic table*; King, R. B.; Rouvray, D. H., Eds.; Nova: New York, USA, 2006; Chapter 5, pp 75-100.

91. Daza, M. C.; Restrepo, G.; Uribe, E. A.; Villaveces, J. L. Quantum chemical and chemotopological study of fourth row monohydrides. *Chem. Phys. Lett*. **2006**, *428*, 55-61.

92. Mendelson, B. *Introduction to topology,* 3rd Ed.; Dover: New York, USA, 1990; Chapter 2, pp 29-69.

93. Lance, G. N.; Williams, W. T. A general theory of classificatory sorting strategies, 1. hierarchical systems. *Comput. J.* **1967**, *9*, 373-380.

94. Gopinathan, N.; Saraf, D. N. Predict heat of vaporization of crudes and pure components revised II. *Fluid Phase Equilibr.* **2001**, *179*, 277-284.

95. Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311-320.

96. Char-Tung, R. Application of information theory to select relevant variables. *Math. Biosci.* **1971**, *11*, 153-161.

97. Todeschini, R.; Cazar, R.; Collina, E. The chemical meaning of topological indices. *Chemometr. Intell. Lab.* **1992**, *15*, 51-59.

98. Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: theory development and its application in chemometrics. *Chemometr. Intell. Lab.* **1999**, *46*, 13-29.

99. Basak, S. C.; Grunwald, G. D. Tolerance space and molecular similarity. *SAR QSAR Environ. Res.* **1995**, *3*, 265-271.

100. Brüggemann, R.; Bartel, H. G. A theoretical concept to rank environmentally significant chemicals. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 211-217.

101. Brüggemann, R.; Carlsen, L. *Partial order in environmental sciences and chemistry*; Springer: Berlin, Germany, 2006.

102. Klein, D. J.; Babić, D. Partial orderings in chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 656-671.

103. Doty, C. B.; Travis, C. C. Is EPA's national priorities list correct? *Environ. Sci. Technol.* **1990**, *24*, 1778-1780.

104. Schneeweiß, C. *Planung 1 - Systemanalytische und entscheidungstheoretische Grundlagen*; Spinger: Berlin, Germany, 1991.

105. Brans, J. P.; Vincke, P. H.; Mareschal, B. How to select and how to rank projects: the PROMETHEE method. *Eur. J. Oper. Res.* **1986**, *24*, 228-238.

106. Opperhuizen, A.; Hutzinger, O.; Multi-criteria analysis and risk assessment. *Chemosphere* **1982**, *11*, 675-678.

107. Davis, G. A.; Swanson, M.; Jones, S. Comparative evaluation of chemical ranking and scoring methodologies; EPA order No. 3N-3545-NAEX, 1994.

108. Brüggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C. E. W. Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 918-925.

109. Lerche, D.; Sørensen, P. B.; Bruggemann, R. 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.* **2003**, *43*, 1471-1480.

110. Brüggemann, R.; Halfon, E. Comparative analysis of nearshore contaminated sites in Lake Ontario: Ranking for environmental hazard. *J. Environ. Sci. Health.* **1997**, *A32*, 277-292.

111. Reisch, M. S. Hot times ahead for refrigerants. *Chem. Eng. News* **2005**, *83*, 23-24.

112. Simon, U.; Brüggemann, R.; Mey, S.; Pudenz, S. METEOR - application of a decision support tool based on discrete mathematics. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 623-642.

113. Voigt, K.; Brüggemann, R. Water contamination with pharmaceuticals: data availability and evaluation approach with Hasse diagram technique and METEOR. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 671-689.

114. Simon, U.; Brüggemann, R.; Behrendt, H.; Shulenberger, E.; Pudenz, S. 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.* **2006**, *34*, 126-136.

115. Krzanowski, W. J. *Principles of multivariate analysis: A user's perspective*; Oxford university press: Oxford, UK, 2003, p 407.

116. De Meyer, H.; Naessens, H.; De Baets, B. Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. *Eur. J. Oper. Res.* **2004**, *155*, 226-238.

117. De Baets, B.; De Meyer, H.; De Schuymer, B.; Jenei, S. Cyclic Evaluation of transitivity of reciprocal relations. *Soc. Choice Welfare* **2006**, *26*, 217-238.

118. Rosenbaum, P. R. *Observational studies*; Springer: New York, USA, 1995.

119. Gefeller, O.; Pralle, L. A nonparametric test for evaluating coherent alternatives in nonrandomised studies. In *Nonrandomized comparative clinical studies. Proceedings of the international conference on nonrandomized comparative clinical studies, 10–11 April 1997, Heidelberg, Germany*; Abel, U.; Koch, A., Eds.; Symposion Publishing: Düsseldorf, Germany, 1997.

120. Kerber, A. Contexts, concepts, implications and hypotheses. In *Partial order in environmental sciences and chemistry*; Brüggemann, R.; Carlsen, L., Eds.; Springer: Berlin, Germany, 2006; Chapter 6, pp 355-365.

121. Ganter, B.; Wille, R. *Formal concept analysis – Mathematical foundations*; Springer: Berlin, Germany, 1999.

122. Eltringham, W.; Catchpole, O. J. Relative permittivity measurements of trifluoromethyl methyl ether and pentafluoroethyl methyl ether. *J. Chem. Eng. Data* **2007,** *52,* 1095-1099.

Appendices

Appendix A

Three Dissimilarity Measures to Contrast Dendrograms

Guillermo Restrepo

Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia, and Environmental Chemistry and Ecotoxicology, University of Bayreuth, Bayreuth, Germany

Héber Mesa

Departamento de Matemáticas, Universidad del Valle, Cali, Colombia

Eugenio J. Llanos

Corporación SCIO, Carrera 13 N*o*. 13 - 24 oficina 721, Bogotá, Colombia

Received November 19, 2006

We discussed three dissimilarity measures between dendrograms defined over the same set, they are triples, partition, and cluster indices. All of them decompose the dendrograms into subsets. In the case of triples and partition indices, these subsets correspond to binary partitions containing some clusters, while in the cluster index, a novel dissimilarity method introduced in this paper, the subsets are exclusively clusters. In chemical applications, the dendrograms gather clusters that contain similarity information of the data set under study. Thereby, the cluster index is the most suitable dissimilarity measure between dendrograms resulting from chemical investigation. An application example of the three measures is shown to remark upon the advantages of the cluster index over the other two methods in similarity studies. Finally, the cluster index is used to measure the differences between five dendrograms obtained when applying five common hierarchical clustering algorithms on a database of 1000 molecules.

INTRODUCTION

Hierarchical cluster analysis,[1] HCA, has become a standard method in searching for similarities among data sets;[2,3] its applications are related to the partitioning of a set into similarity classes[4] that are represented as clusters. HCA constitutes a method for classifying the original set with which it is possible to study the behavior of a member of determined class and finally generalize such knowledge to the other members of the class. This procedure endows the set under study with a mathematical structure,[5] namely, a topology.[6−12] In general, a HCA study begins defining the set Q of work by means of the features of its elements and then looking for the (dis)similarities among the elements using a (dis)similarity function, DF. Afterward, when a grouping methodology, GM, is used, similar elements are clustered and represented graphically in a dendrogram (rooted acyclic-connected binary graph; Figure 1), where clusters appear as branches of the dendrogram.

The total number of dendrograms,[13] $|F|$, which can be defined over Q, whose cardinality $|Q|$ is n, grows with n according to

$$|F| = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

When applying different DFs and GMs (different HCA algorithms) over Q, several and different results might come up as a consequence of some bias of clustering algorithms toward particular cluster properties[2,3] or as the effect of the lack of "natural clusters"[3] in Q. Then, a question arising from this discussion is, how can we measure the dissimilarity between these results (it is between the corresponding dendrograms)? A contrary situation may occur if the similarity relationships among the elements are strong enough and almost invariant to different HCA algorithms; in this case, even for a set of large cardinality with a large number $|F|$ of possible dendrograms, it is likely to find similar clusters (natural ones) in their resultant trees. Hence, a possible answer to the question on the contrast of HCA results can be addressed to the contrast of their respective dendrograms by the comparison of their clusters.

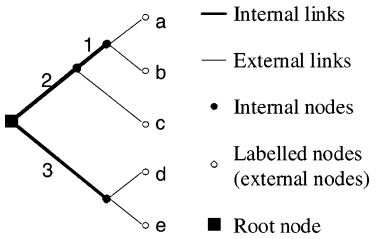


Figure 1. A dendrogram and types of nodes and links characterizing it (numbers label internal links).

Two main mathematical methods have been proposed since the 1960s[14] for contrasting dendrograms. One of them defines a new tree representing a consensus or area of agreement,[14] and the other method defines for any pair of dendrograms a (dis)similarity measure indicating the extent of (dis)agreement.[14] Since this paper deals with dissimilarity measures, we describe them in detail. A detailed discussion on consensus techniques appears in refs 15 and 16. The majority of methods designed to measure dissimilarity between dendrograms have been developed for applications to biological trees (evolutionary trees in the majority of

10.1021/ci6005189 CCC: $37.00
Published on Web 04/28/2007

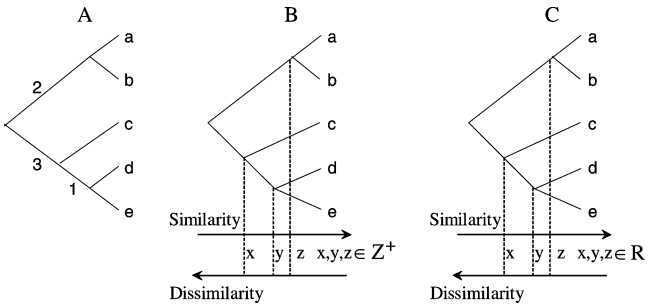


Figure 2. Types of structural dendrograms: (A) bare, (B) ranked, and (C) valued dendrograms.

cases), and several of these techniques are related to each other[17] because of the resemblances among structural units of the dendrograms they assess. According to Steel and Penny,[18] there are different structural features of a tree because: "[...] there is no "obvious" or "natural" way to measure the distance between two trees, unlike the comparison of two numbers (where one subtracts the smaller number from the larger)". Some examples of dissimilarity measures are partition metric[19] or symmetric difference,[15] quartets distance,[14] triples distance,[20] nearest neighborhood interchange metric,[21,22] and some others based on differences in the lengths of the paths between pairs of elements in Q.[15,23] Although none of these methods actually consider clusters as units to contrast dendrograms, it is crucial, particularly in chemistry, to assess the resemblance between dendrograms using their clusters because they are the pieces of the dendrograms containing the similarity information of Q. When one applies HCA to a data set, the interpretation of dendrograms is carried out on their clusters; therefore, it would be important to contrast dendrograms on the basis of their clusters. In this paper, we developed a new dissimilarity method dealing exclusively with clusters. This method and the other two discussed in this paper consider a dendrogram as a structure (a graph) able to be decomposed into substructures (subgraphs). A structural classification of dendrograms[23] is given in the following.

Bare Dendrograms: They only show the similarity relationships among the elements of Q without a scale of (dis)similarity (Figure 2A).

Ranked Dendrograms: They have internal nodes ranked on an ordinal scale of (dis)similarity (Figure 2B).

Valued Dendrograms: They possess internal nodes which have been assigned to a continuous (dis)similarity scale in the real numbers with at least an interval-scale interpretation (Figure 2C).

Bare dendrograms are topological in nature because the relationships (or links) among the elements they cluster are the only features of interest in such structures. Hence, it is irrelevant if the links joining two nodes are longer or shorter. In these kinds of dendrograms, it can only be stated either that "a is related to b", a and b elements being in the same cluster, or that "a is not related to b", otherwise. Consequently, the notion of an equivalence relation can be attached to the cluster definition, and then a cluster becomes an equivalence class where the mathematical relation is a similarity relationship.[24,25] On the other hand, ranked and valued dendrograms are regarded as geometrical objects where the membership of an element to a cluster is ruled by the presence or absence of links and by the (dis)similarity

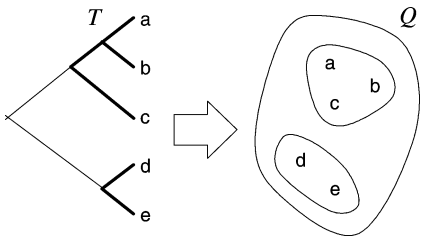


Figure 3. Three subtrees (T and the two bold graphs) and their associated subsets (Q and $\{a, b, c\}$ and $\{d, e\}$).

scale. Although they have their differences, these three kinds of dendrograms hold, besides their common graph-classification as complete secondary trees,[26] a metric relationship; all of them are ultrametric trees because any pair (r,s) of elements in Q is related by $d(r,s)$, d being a function fulfilling the metric properties[27] and also the ultrametric property:[23] $d(r,t) \leq \max\{d(r,s),d(s,t)\}$ with $r, s, t \in Q$. For example, for valued dendrograms, $d(r,s)$ may be defined as the (dis)similarity value of the closest internal node connecting r and s; for ranked trees, it may correspond to the order of the internal nodes and for bare dendrograms to the number of external nodes connected to the internal ones.[23]

Dissimilarity measures between dendrograms, in general, can be divided into two classes:[23] those transforming one dendrogram into another one and those considering a dendrogram as a simpler structure (i.e., sets, partitions, or incidence matrices). The disadvantage of the first ones is that they are often very hard to compute in contrast to the second ones, which are generally quite tractable.[16,23] In the following, we discuss three dissimilarity measures dealing specifically with bare and ranked dendrograms. A discussion on valued trees appears in refs 15 and 23. Two of the analyzed methods in this paper are reported in the literature, and the third one is a novel dissimilarity measure based on the contrast of clusters, which makes it highly appropriate for chemical similarity applications.

MEASURING DISSIMILARITY AMONG SETS OF DENDROGRAMS

For the sake of clarity, we define some relevant terms; Q is the set of objects to classify using HCA; $|Q|$, the cardinality of Q, is represented by n; T is a dendrogram (tree) on Q; C is a cluster of T; and $a|bc|...$ is the representation of any partition $\{\{a\},\{b,c\},...\}$. When considering T as a graph (tree), then a cluster may be regarded as a subgraph or subtree of T. Normally, the extraction of a cluster from T has two viewpoints: one is to consider the dendrogram as a graph; in that case, the dendrogram becomes a rooted acyclic-connected binary graph (tree), and its clusters may be regarded as its subtrees (A1); the other viewpoint comes from the set theory, and T is considered as a subset of Q. Note that a cluster, besides being a proper set of $Q(C \subset Q)$, can also be the same set $Q(C = Q)$. For that reason, in general, we write $C \subseteq Q$. Then, having a subtree from T, a subset in Q can be associated to it. In this paper, when we refer to a cluster, it is because it has a subtree in T and also an associated subset in Q. We show in Figure 3 three subtrees of T and their corresponding associated subsets. Note that T is a subtree (A1), and its associated subset is Q.

Triples Index. This method[20] was designed to deal with rooted binary trees (dendrograms), in contrast to the similar

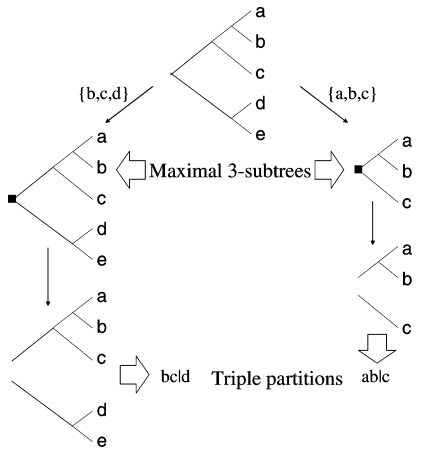


Figure 4. Partitioning the triples $\{a,b,c\}$ and $\{b,c,d\}$ according to the triples index method (in the example of application of the three dissimilarity indices appears the complete list of triple partitions).

methodology, quartets distance,[14] developed to treat unrooted trees. In triples index,[20] $h = \{i, j, k\}$ is defined as a **triple**, where $i, j, k \in Q$. The total number of triples in Q is $t = \binom{n}{3}$. Given a triple h, the next step of the procedure is to look for the maximal 3-subtree (A2) containing the elements in h. Afterward, the root node of the maximal 3-subtree is deleted, inducing a binary partition $ij|k$ on the triple h, which is called triple partition TP (Figure 4). Because this dissimilarity measure contrasts pairs of dendrograms, the procedure described previously is carried out over the dendrograms of interest T_1 and T_2. In order to contrast the triple partition of the triple h in $T_1[\mathrm{TP}(T_1)]$ and $T_2[\mathrm{TP}(T_2)]$, the symmetric difference (A3) given by $\mathrm{SD} = \mathrm{TP}(T_1)\ \Delta\mathrm{TP}(T_2)$ is calculated. SD shows the different subsets between T_1 and T_2 regarding the triple h whose number is given by $|\mathrm{SD}| = |\mathrm{TP}(T_1)| + |\mathrm{TP}(T_2)| - 2|\mathrm{TP}(T_1) \cap \mathrm{TP}(T_2)|$ (A3). In general, if two binary partitions TP_1 and TP_2 are built on a set of three elements, there exist only two possible values |SD| can take, namely, 0 or 4; 0 corresponds to equal partitions ($\mathrm{TP}_1 = \mathrm{TP}_2$) and 4 to different partitions ($\mathrm{TP}_1 \cap \mathrm{TP}_2 = \varnothing$). Now, the indicator function I_h is defined as

$$I_h = \begin{cases} 1 \text{ if } |\mathrm{SD}| = 4 \\ 0 \text{ if } |\mathrm{SD}| = 0 \end{cases}$$

I_h yields a value of 1 if the partition of the considered triple h in T_1 is different from the partition of h in T_2. If those partitions are the same, then $I_h = 0$. The triples distance between dendrograms T_1 and T_2 is defined as

$$S(T_1,T_2) = \sum_{h=1}^{t} I_h$$

Hence, $S(T_1,T_2)$ counts how many triples are different out of the total t. In order to restrict the values $S(T_1,T_2)$ can take to the interval $[0,1] \in R$, we normalized $S(T_1,T_2)$ looking for the maximum and minimum values it can take. Hence, $\max[S(T_1,T_2)] = t$, which means that all t triple partitions in T_1 and T_2 are different. On the other hand, $\min[S(T_1, T_2)] = 0$, meaning that all the t triple partitions in T_1 and T_2 are the same. Having these maximum and minimum values for $S(T_1,T_2)$, we define the triples index as

$$\bar{S}(T_1,T_2) = \frac{S(T_1,T_2)}{t}$$

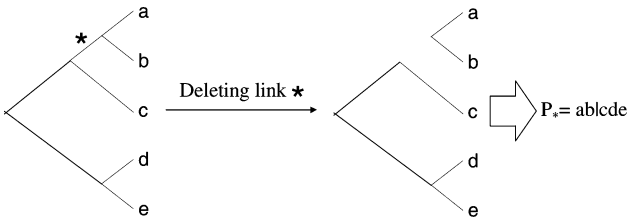


Figure 5. Deletion of a link in the partition index method.

Table 1. Partitions for T_1 and T_2 (Figures 1 and 2, Respectively) by Deleting Their Internal Links

rth deleted link	$P_r(T_1)$	$P_r(T_2)$
1	*ab|cde*	*abc|de*
2	*abc|de*	*ab|cde*
3		

Thus, $\bar{S}(T_1,T_2)$ is a dissimilarity measure between T_1 and T_2. In this method, it is considered that a dendrogram is completely determined by the way in which its triples are partitioned in Q. However, it might be possible to consider other kinds of subtrees, namely, quartets, quintets, and in general s-tets. Note that $s = 2$ is not informative, since we look for partitions of the s-tets to do contrasts; in this case, a duo always yields the same partitions, and then they cannot be used to look for differences between trees. A generalization of this procedure is to look for all the $\binom{n}{3}$ s-tets in Q and to study how they are partitioned in the two trees. We are preparing another paper regarding this family of s indices and their features. We show an example of the application of this method in the next section.

Partition Index. This method[19] was conceived to contrast unrooted as well as rooted binary trees (dendrograms). Nowadays, it has been called partition metric and is one of the most known measures of dissimilarity between trees. In fact, it is included in several software packages, for instance, COMPONENT,[28,29] PHYLIP,[30] and PAUP.[31] Its procedure is based on the contrast of partitions generated by removing internal links (Figure 1) of a dendrogram. It does not remove external links (Figure 1) because the generated partitions do not contribute to differentiate the contrasted dendrograms (A4). The first step in this method is the deletion of an internal link r from T, $n - 2$ being the total number of internal links in a dendrogram.[19] The deletion of this link produces two subgraphs of T whose associated subsets become a binary partition P_r of Q. Consider, for instance, the dendrogram shown in Figure 5; if the link marked * is removed, then two disjoint subsets are produced: $\{a,b\}$ and $\{c,d,e\}$, which are gathered in P^* (Figure 5). Note that the deletion of a link always produces a partition $A|A^C$, where either A or A^C is a cluster (subtree) (A1). The key of this method is the contrast of these binary partitions P_r's, and for that reason, it is important to know their number. Although $n - 2$ internal links yield $n - 2$ P_r's, this is not the number of partitions to contrast because there is always a redundant partition. We can explain this analyzing the P_r's of the dendrogram in Figure 1, which appear in Table 1 (second column), where $P_2(T_1) = P_3(T_1)$. In general, there are always two equal P_r's produced by deleting the connected links to the root node (Figure 1). Then, the total number of partitions to contrast is $n - 2 - 1 = n - 3$. Now, PT is defined as the collection of partitions to contrast and $P(T_1\ T_2) = \mathrm{PT}_1\ \Delta\mathrm{PT}_2$ as their symmetric difference (A3), where

$P(T_1,T_2)$ yields the different partitions between T_1 and T_2. The partition metric is defined as the cardinality of $P(T_1,T_2)$,[19] which is given by $|P(T_1,T_2)| = 2(n-3) - 2|\mathrm{PT}_1 \cap \mathrm{PT}_2|$ because of A3 and $|\mathrm{PT}_1| = |\mathrm{PT}_2| = n-3$. If m is assumed as the number of common partitions to T_1 and T_2, then $|P(T_1,T_2)| = 2(n-3-m)$. In order to obtain a dissimilarity index for two dendrograms ranging between 0 and 1, we normalized $|P(T_1,T_2)|$, which depends on m because n is a fixed value for T_1 and T_2. Hence, the normalization factor of $|P(T_1,T_2)|$ depends on the minimum and maximum values m can take. max(m) is reached when all partitions are equal for both trees, then max(m) $= n-3$, and min(m) occurs when T_1 and T_2 have the minimum number of common partitions between them; it is min(m) $= 0$. Now, min(m) and max(m) determine the maximum and minimum values of $|P(T_1,T_2)|$, respectively. Thus, $0 \le |P(T_1,T_2)| \le 2(n-3)$ is obtained, from which $|P(T_1,T_2)|$ is normalized to the partition index $\mathrm{PI}(T_1,T_2)$ given by

$$\mathrm{PI}(T_1,T_2) = 1 - \frac{m}{n-3}$$

When the partition index takes a value of zero, the contrasted dendrograms have all their partitions in common; if $\mathrm{PI}(T_1,T_2) = 1$, it is because there are no common partitions to T_1 and T_2. In appendix A5, we prove that partition index results are the same if the link deletion includes either all the links in T or only the internal ones. We show an example of the application of this method in the next section.

Remarks on the Concept of Cluster. Penny et al.[19] justify the use of partitions as objects to contrast in the partition metric following the results of Waterman and Smith,[22] to whom a tree is represented by the binary partitions produced in the partition metric method. In general, any method extracting structural units from a dendrogram can be considered as a relation R between T and the power set of Q. However, there is not a unique R; moreover, it can be possible to have several relations between T and different aggregations of the $2^n - 1$ nonempty subsets in Q. Then, in principle, it can be possible to find several subsets or structural units representing T. Nevertheless, if the interest in the definition of T is not concerned exclusively with its representation but also with the similarities shown by T, then the most appropriate units for reaching this goal are the clusters of T. In the following, we show how the concept of a cluster can be regarded, first, as a structural unit describing and reconstructing T and, second, as an equivalence class containing similarity information.

Gusfield[32] has noted that a dendrogram is represented by its clusters, a statement that can be formalized through the concept of intersection graphs[33] as follows: let J be a collection of sets; the intersection graph of J is the graph obtained by assigning to each set in J a distinct vertex. A line is drawn between two vertices if the intersection between the two sets associated with each vertex is nonempty. Hence, the dendrogram shown in Figure 1 can be considered as the intersection graph (Figure 6) of the subsets A to I shown in Figure 6. Additionally, according to hypergraph theory,[34] the clusters can be regarded as the vertices of the hypergraph (dendrogram) and the hyperedges as the intersection between any two different clusters.

Although the reconstruction of a graph from the collection of its one-vertex-deleted subgraphs is still an open question

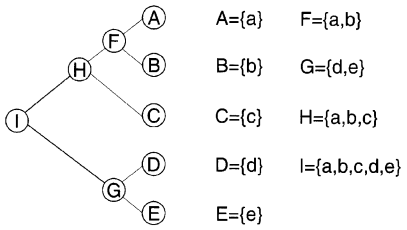


Figure 6. A dendrogram as an intersection graph of the subsets A to I and as a hypergraph of the subsets A to I and their intersections.

in graph theory (reconstruction[35,36] or Ulam's conjecture),[37] its particularization to the case of trees[38] and some other special graphs has been proved.[37] Since a dendrogram is a tree (complete secondary tree)[26] and it can be defined as an intersection graph or as a hypergraph of its clusters, it can therefore be reconstructed from its clusters, which are also obtained from the one-vertex-deleted subgraphs.

The second, and most relevant, viewpoint of a cluster is its ability to show similarities between the elements in Q, which is the main reason why cluster analysis is broadly used in drug discovery processes and molecular diversity studies.[2] Restrepo and Brüggemann[39] recently showed that, if the similarity between the elements in Q is regarded as an equivalence relation R, then the elements in a cluster constitute an equivalence class, R being a similarity relation. In that case, the set of all the clusters in Q becomes the quotient of Q by R (Q/R). In summary, clusters determine similarity classes or similarity neighborhoods in Q,[6,10] and also they contain structural information of the dendrogram; for these reasons, they can be used to characterize a dendrogram and also to characterize the similarity relationships in the set Q, which is one of the targets of HCA in chemistry. In the following, we propose a new method for measuring the dissimilarity between two dendrograms on the basis of the contrast of their clusters.

Cluster Index. We understand the question about the dissimilarity between two dendrograms as "how dissimilar are their clusters", because, as we have remarked, the applications of HCA in chemistry are related to the notion of similarity, expressed in the clusters. The cluster index, here described, follows the counting ideas shown in the previous methods using symmetric difference, but it considers as raw material the clusters of the contrasted dendrograms. The total number of clusters in a dendrogram is $2n-1$, as we show in the following proposition.

Proposition. Let Q be a finite set, T a dendrogram over Q, and RT the collection of clusters of T; then, the number of clusters in T is given by $|\mathrm{RT}| = 2n - 1$.

Proof. Let us use induction over $|Q|$. If $Q = \{x_1\}$, then there exists only one dendrogram with a unique cluster $\{x_1\}$. In this case, $|\mathrm{RT}| = 1 = 2(1) - 1 = 2|Q| - 1$. Suppose $|Q| = k$ with $k \le n$, then for any dendrogram T, the number of its clusters is given by $|\mathrm{RT}| = 2k - 1 = 2|Q| - 1$. Let Q be a set such that $|Q| = n + 1$ and T be a dendrogram. Every T has a root node (Figure 1) splitting T into subtrees (clusters) T_1 and T_2, where T_1 is a dendrogram over a set $A \subset Q$ and T_2 a dendrogram over A^C. Since $|\mathrm{RT}|$ is the number of clusters in T, then $|\mathrm{RT}| = |\mathrm{RT}_1| + |\mathrm{RT}_2| + 1$. $|\mathrm{RT}|$ counts the clusters in T_1 and T_2 given by $|\mathrm{RT}_1|$ and $|\mathrm{RT}_2|$ respectively and also the largest cluster corresponding to the complete dendrogram T that is counted by the addition of 1 in $|\mathrm{RT}|$.

On the other hand, $|A| < |Q| = n + 1$ and $|A^C| < |Q| = n + 1$; then, $|A| \leq n$ and $|A^C| \leq n$. Using the hypothesis of induction, we have that $|\mathrm{RT}_1| = 2|A| - 1$ and $|\mathrm{RT}_2| = 2|A^C| - 1$. Note that $|A| + |A^C| = |Q|$ because A and A^C are disjoint; then, $|\mathrm{RT}| = 2|A| - 1 + 2|A^C| - 1 + 1 = 2|Q| - 1$. Now, since Q is made from $n + 1$ elements, then $|Q| = n + 1$ and $|\mathrm{RT}| = 2(n + 1) - 1$.

When contrasting the set of clusters of T_1 with the ones of T_2, there are always $n + 1$ common trivial clusters, which are the n single clusters $\{x\}$, $x \in Q$, and the complete set Q. For this reason, if the goal is to measure the difference between two dendrograms, then these trivial clusters must not be considered, and the total number of clusters to contrast becomes $n - 2$. We call CT_1 and CT_2 the clusters to contrast for T_1 and T_2, respectively, and their contrast is carried out by the symmetric difference $C(T_1,T_2) = \mathrm{CT}_1 \, \Delta \mathrm{CT}_2$, which yields the different clusters between T_1 and T_2. Now, we call the cluster metric the cardinality of $C(T_1,T_2)$ (A3), given by $|C(T_1,T_2)| = 2(n - 2) - 2|\mathrm{CT}_1 \cap \mathrm{CT}_2|$. If we assume c as the number of clusters common to CT_1 and CT_2, then $|C(T_1,T_2)| = 2(n - 2 - c)$, which depends on c because n is a fixed value for T_1 and T_2. For that reason, we studied the maximum and minimum values c can reach. $\max(c) = n - 2$, meaning that all the possible clusters in T_1 are present in T_2 and $\min(c) = 0$. These maximum and minimum values of c determine the minimum and maximum values of $|C(T_1,T_2)|$, respectively, yielding $0 \leq |C(T_1,T_2)| \leq 2(n - 2)$. Now, we call the cluster index, $\mathrm{CI}(T_1,T_2)$, the rank normalization of $|C(T_1,T_2)|$, which is given by

$$\mathrm{CI}(T_1,T_2) = 1 - \frac{c}{n-2}$$

When the cluster index reaches a value of zero, it is because the contrasted dendrograms have all their clusters in common; if that value is 1, all of their contrasted clusters are different. One question arising from the $\mathrm{CI}(T_1,T_2)$ expression is whether or not it changes when considering one, two, or in general k trivial clusters. We show in A6 that $\mathrm{CI}(T_1,T_2)$ yields the same result, when adding k trivial clusters to the clusters to contrast.

EXAMPLE OF APPLICATION OF THE THREE DISSIMILARITY INDICES

In order to show the way in which the three dissimilarity measures here discussed work, we propose the following example. Suppose the dendrograms shown in Figures 1 and 2 defined over $Q = \{a,b,c,d,e\}$; since methods here-described treat bared and ranked trees, then whatever dendrogram, either Figure 2A or B, can be considered in this example.

Triples Index. In this case $t = 10$; then, Q has 10 possible triples, which are listed in Table 2. Partitions TP_h for T_1 and T_2 are shown in the third and fourth columns in Table 2, respectively. In addition, $\mathrm{TP}(T_1)$ and $\mathrm{TP}(T_2)$ are the sets gathering the cells in the third and fourth columns of Table 2, respectively.

The column labeled I_h shows the indicator function values for the symmetric difference between each hth couple of partitions. Thus, it is clear that there are four triples out of 10 for which the partitions are different. Hence, $\bar{S}(T_1,T_2) = 4/10 = 0.4$, which means that the dendrograms in Figures 1 and 2 are 40% dissimilar.

Table 2. Triples, Their Partitions and Indicator Function Values for Dendrograms of Figures 1 and 2

h	triples	$\mathrm{TP}_h(T_1)$	$\mathrm{TP}_h(T_2)$	I_h
1	*abc*	*ab*|*c*	*ab*|*c*	0
2	*abd*	*ab*|*d*	*ab*|*d*	0
3	*abe*	*ab*|*e*	*ab*|*e*	0
4	*bcd*	*bc*|*d*	*cd*|*b*	1
5	*bce*	*bc*|*e*	*ce*|*b*	1
6	*cde*	*de*|*c*	*de*|*c*	0
7	*cda*	*ac*|*d*	*cd*|*a*	1
8	*edb*	*ed*|*b*	*ed*|*b*	0
9	*ace*	*ac*|*e*	*ce*|*a*	1
10	*ade*	*de*|*a*	*de*|*a*	0

Partition Index. In this case, $n - 3 = 2$; then, each tree T_1 and T_2 has two possible nonredundant partitions of Q by removing their internal links. These partitions P_r are shown in Table 1.

Thus, $\mathrm{PT}_1 = \{ab|cde, abc|de\}$ and $\mathrm{PT}_2 = \{abc|de, ab|cde\}$, and according to Table 1, we conclude that $m = 2$, and then $\mathrm{PI}(T_1,T_2) = 0$. It means that the two partitions are common to the dendrogram in Figures 1 and 2. In other words, they are 0% dissimilar (100% similar). We give further details about the interpretation of this result in the next section.

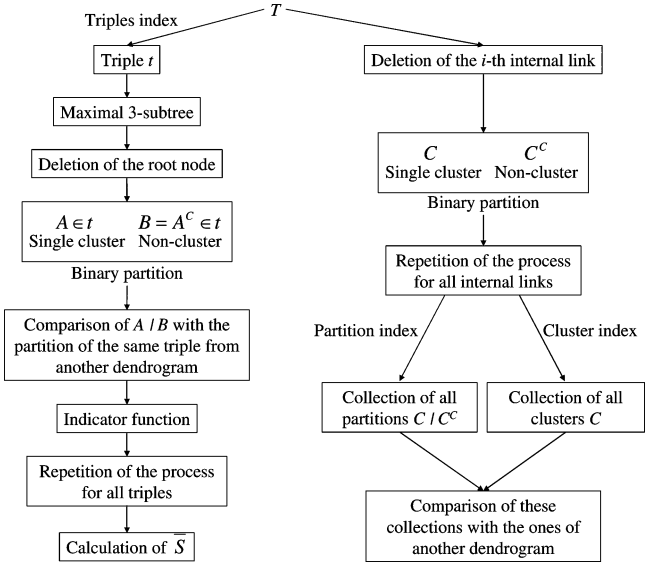
Cluster Index. The collections of nontrivial clusters for T_1 and T_2 are $\mathrm{CT}_1 = \{\{ab\},\{de\},\{abc\}\}$ and $\mathrm{CT}_2 = \{\{ab\},\{de\},\{cde\}\}$. From CT_1 and CT_2, we conclude that $c = 2$ since $\{a,b\}$ and $\{d,e\}$ are common clusters to CT_1 and CT_2. Finally, $\mathrm{CI}(T_1,T_2) = 1/3 = 0.\overline{3}$. In other words, dendrograms in Figures 1 and 2 are about 33% dissimilar.

In summary, the results of the application of the three methods here described are $\bar{S}(T_1,T_2) = 0.4$, $\mathrm{PI}(T_1,T_2) = 0$, and $CI(T_1,T_2) = 0.\overline{3}$, which show that these methods assess different aspects of the structure of a dendrogram, as we have remarked above.

It is surprising to have a value of $\mathrm{PI}(T_1,T_2) = 0$ for the above example because T_1 and T_2 are topologically different and one expects $\mathrm{PI}(T_1,T_2) > 0$. That is, nodes in T_1 and T_2 have different connectivities. Nevertheless, it is important to state that all the methods here-described assess the dissimilarity between dendrograms, decomposing each tree into subsets and then contrasting those subsets. For that reason $\mathrm{PI}(T_1,T_2) = 0$ does not necessarily mean $T_1 = T_2$; what it really means is $\mathrm{PT}_1 = \mathrm{PT}_2$. That is, the binary partitions of T_1 are the same as the ones of T_2. In general, each dissimilarity measure here-discussed can be regarded as a function $d(T_1,T_2)$, assessing a particular structural aspect of T_1 and T_2; then, $d(T_1,T_2) = 0$ is reached if the "contrasted units" (not the dendrograms as they are) are exactly the same for T_1 and T_2. Dendrograms are mathematically represented by subsets, and a complete match of these subsets ought to be interpreted only in terms of the subsets. It is not correct to state that two dendrograms are the "same" because one particular dissimilarity measure yields $d(T_1,T_2) = 0$. Perhaps, if the idea of a holistic dissimilarity measure is searched, that is, if we want to consider several structural aspects of the dendrogram, then a first attempt for having such a measure is the combination of f specific dissimilarity measures in this way:

$$D(T_1,T_2) = \frac{\sum_{i=1}^{f} w_i d_i(T_1,T_2)}{\sum_{i=1}^{f} w_i}$$

Scheme 1. Relationships between Triples, Partition, and Cluster Indices



where $d_i(T_1,T_2)$ is a dissimilarity measure and w_i is a weight factor sizing the priority or the importance of the *i*th dissimilarity measure (note that $D(T_1,T_2)$ is based on the general similarity coefficient suggested by Gower).[40]

SIMILARITIES AND DIFFERENCES AMONG THE THREE METHODS

It is possible to study the relationships between the three dissimilarity methods, analyzing their procedures. We summarize in Scheme 1 the three methodologies, showing how each one works on a dendrogram *T*. The triples index contrasts dendrograms, studying their triples binary partitions $A|B$, where *A* is always a single cluster (Scheme 1) and the other part, *B*, is not a cluster. The partition index also contrasts binary partitions $C|C^C$ of *Q*, and *C* is always a cluster. Thus, both, partition and triples indices, always contrast partitions including at least one cluster. Nevertheless, the whole contrast is not based exclusively on clusters but on a mixture of 50% clusters and 50% nonclusters. In other words, 50% of the raw material of triples and partition indices contains similarity information, and the other 50% does not. For this reason, their results cannot be interpreted in terms of similarities between the elements in *Q*. On the other hand, according to Scheme 1, partition and cluster indices can be regarded as similar in the initial step. If we consider the process of looking for clusters as a link deletion, then both methods remove links on the dendrogram and produce partitions $C|C^C$ where *C* is a cluster and C^C is not. The difference between these two methods begins when the partition index considers, as units to contrast two dendrograms, the complete partition $C|C^C$; thereby, it is attached to the similarity information, expressed as *C*, nonsimilarity information, given by C^C. On the contrary, the cluster index only takes clusters *C* as units to contrast, discarding C^C, that is, discarding the nonsimilarity information of *Q*. We mentioned above some remarks on the concept of cluster and its importance in expressing neighborhood and similarity relationships among its elements. Accordingly, if we want to build a method looking for dissimilarity between dendrograms on the basis of their clusters, then the most appropriate of the three methods here discussed is the cluster index because it is the only one completely based on the contrast of clusters, without adding other kinds of mathematical objects which "contaminate" the similarity expressed in the clusters. In few words, the cluster index is created to interpret a dendrogram as a collection of clusters, and it always contrasts exclusively clusters.

By the analysis of dendrograms T_1 and T_2 shown in Figures 1 and 2, respectively, we can see that $CT_1 = \{ab,de,abc\}$ and $CT_2 = \{ab,de,cde\}$. Hence, these dendrograms are differentiated by the clusters *abc* and *cde*. However, if we attach to the clusters in CT_1 and CT_2 their complements, then we obtain PT_1 and PT_2 in the partition metric, which are the sets including the binary partitions of the first and second columns of Table 1, respectively. Then, the discriminatory power of *abc* and *cde* is lost since *abc* (in the cluster metric) becomes $abc|de$ (in the partition metric) and *cde* becomes $cde|ab$. The ability to differentiate disappears because $abc|de = P_2(T_1)$ (Table 1) is also present in $T_2[P_1(T_2)]$ (Table 1). Similarly, $P_1(T_1) = ab|cde$ (Table 1) becomes $P_2(T_2)$ in T_2 (Table 1). Thus, the discriminatory power embedded in the clusters is lost when attaching their complements to them. In contrast to the partition index, the cluster index keeps the discriminatory power of the clusters.

In summary, all three methods work with clusters, the triples index with single clusters and the partition and cluster indices with the same clusters; however, triples and partition indices join to their clusters some other subsets not corresponding to clusters. On the other hand, although cluster and partition indices use the same clusters, the attachment of nonclusters to the clusters, in the case of the partition index, makes the results and their meaning change. Thus, the advantage of contrasting trees using the *similarity sense of clusters is lost when applying triples and partition indices*, and *it is kept in cluster index.*

CHEMICAL APPLICATION OF THE CLUSTER INDEX

In the previous discussion the argument arose that, in HCA chemical applications, any contrast of dendrograms must deal with the contrast of their clusters because the clusters are the entities gathering the similarity chemical information. In this section, we calculate the dissimilarity of different HCA algorithms over a chemical database; since triples and partition indices do not operate entirely on the clusters of the HCA results, then these methods are not considered in this example.

A set of 1000 molecules was randomly selected from the National Cancer Institute, NCI, database,[41] and they were represented by 1024-bit Barnard Chemical Information, BCI, fingerprints;[42−44] their similarities were calculated using the Tanimoto coefficient.[4] Five different GMs were applied to the Tanimoto similarity matrix yielding five dendrograms, which, because of their large size, cannot be displayed in this manuscript; however, their electronic ASCII files can be requested from G. Restrepo. The GMs employed were[9] single (sing), complete (comp), centroid (cent), and unweighted average (unav) linkages and Ward's method, all of them members of the sequential agglomerative hierarchical nonoverlapping methods.[45] The cluster index values for the contrast of the 10 pairs of dendrograms appear in Table 3.

There are 1999 clusters in each dendrogram, 1001 of which are trivial ones, and 998 are considered in the calculations

Table 3. Cluster Index Results for the Contrast of Five Dendrograms Obtained from the Combination of the Tanimoto Coefficient and Five Grouping Methodologies

	centroid linkage	unweighted average linkage	complete linkage	single linkage	Ward's method
centroid linkage	0				
unweighted avg linkage	0.679	0			
complete linkage	0.618	0.545	0		
single linkage	0.769	0.791	0.765	0	
Ward's method	0.669	0.511	0.537	0.787	0

of the cluster dissimilarity index. From Table 3, it can be seen that the five HCA results yield rather different outcomes since their dissimilarities are grater than 0.5, which means that more than 50% of the clusters in each dendrogram are different from those in the other four trees. In other words, more than 499 clusters are different in any contrast of dendrograms. Keeping in mind these differences, the lowest difference occurs for the couple unav−Ward (51% dissimilarity), with 488 common clusters, while the largest difference results for the couple unav−sing (79% dissimilarity), with 209 common clusters. According to Table 3, the ranges of dissimilarities are Ward, [0.511, 0.787]; sing, [0.765, 0.791]; comp, [0.537, 0.765]; unav, [0.511, 0.791]; and cent, [0.618, 0.769]. Their standard deviations are sing, 0.011; cent, 0.054; comp, 0.091; Ward, 0.110; and unav, 0.111. Hence, the most spread dissimilarities are those of unweighted average linkage, and the least ones are those corresponding to a single linkage. The following order of the dissimilarity spread can be set up: sing < cent < comp < Ward < unav. Although sing is the GM with the least spread dissimilarities, it is, in general, the most dissimilar GM when combined with the Tanimoto coefficient because of its high cluster index dissimilarity values (Table 3).

SUMMARY, CONCLUSIONS, AND OUTLOOK

We described three different methods (triples, partition, and cluster indices) for calculating dissimilarities between trees; each one assesses the dissimilarity between two dendrograms by the analysis of different structural aspects of a tree. In general, the triples index looks for all the possible sets of three members (triples) contained in Q, and it contrasts the structural connections of them (maximal 3-subtrees) through the analysis of their binary partitions. This contrast is carried out counting the different partitions between the dendrograms considered. The mathematical background of this method opens the possibility of considering other kinds of h-tets (quartets, quintets, and so on) to scan the structural dissimilarity between two trees. Nevertheless, these dissimilarities cannot be fully interpreted in terms of the resemblances among the elements in Q because they are not completely based on the contrast of clusters because some other subsets, which are not clusters, are considered.

The second dissimilarity method was the partition index, which builds all the possible binary partitions of a dendrogram by deleting its internal links, and then it contrasts those partitions counting the different numbers of them. Hence, its result is a measure of how many partitions are different between the two given dendrograms.

The cluster index, a novel dissimilarity measure introduced in this paper, was developed taking into consideration the lack of methods based on the contrast of the clusters present in two dendrograms. We considered it important to have such kinds of dissimilarity measures because of the two important aspects of the concept of the cluster, namely, the possibility of reconstructing a dendrogram from its clusters and the similarity information gathered in each cluster regarding the elements contained in it. In fact, a dendrogram is mainly used in chemistry for searching similarity classes (clusters) in a given set Q. Hence, the cluster index is based on the consideration that a dissimilarity measure between dendrograms must be addressed to the assessment of the dissimilarity between their clusters.

On the other hand, through the application of a method contrasting clusters, it is possible to evaluate the following: (1) the effect of the chemical representations, that is, assessing to what extent the HCA results change if the molecules are described by dataprints,[2] that is, real number descriptors, such as topological indices and physicochemical properties, or if they are described by fingerprints,[2] that is, binary strings representing the presence or absence of 2D structural fragments or 3D pharmacophores; (2) the effect of clustering a data set Q using different dissimilarity functions and similarity coefficients,[4,46] such as Hamming and Soergel distances and Tanimoto and Dice coefficients; (3) the effect of applying different grouping methodologies, such as those mentioned in this paper.

Furthermore, it is possible to assess the combined effect of points 1−3; these contrasts can be done directly measuring the behavior of the similarity neighborhoods in a dendrogram, that is, in their clusters.

We introduced the cluster index as a dissimilarity measure evaluating the behavior of the clusters in two dendrograms. In its mathematical development, it was proved that the number of clusters to contrast between two dendrograms is always $n - 2$ because the total number of clusters in a dendrogram is $2n - 1$, where $n + 1$ out of $2n - 1$ are always trivial clusters present in all couples of dendrograms.

By the comparison of the three methods described in this paper, we found that, although the three methods initially consider clusters as units to contrast dendrograms, triples and partition indices mix them with some other mathematical objects different than clusters, which causes them to not be recommended for chemical applications where the final aim is the searching of similarities in a data set, similarities that are contained in the clusters. It was shown that the only method dealing exclusively with clusters is the cluster index; thereby, it is the recommended dissimilarity method to be applied in chemical studies.

A common characteristic of the procedure followed by the three dissimilarity measures here discussed is the use of the operation of symmetric difference. This resemblance underlies the fact of describing each dendrogram as a collection of subsets, which is a constant feature of the methods discussed here. In that case, a mathematical tool for looking for particular differences in sets is the symmetric difference. In fact, this operation has been used in several structural dissimilarity measures, and it is not only restricted to the case of dendrograms; for example, Brüggemann and co-workers[47] have defined a dissimilarity measure between posets (partially ordered sets), describing a poset as a collection of subsets and using the symmetric difference as the tool for looking for differences.

We mentioned that each dissimilarity measure analyzes different structural aspects of a dendrogram and also discussed the possibility of having a general dissimilarity index through the weighted linear combination of different dissimilarity measures.

A chemical application of the cluster index was carried out when analyzing five HCA algorithms which combine the Tanimoto coefficient and five common grouping methodologies; the database selected was a collection of 1000 molecules from the National Cancer Institute. The dissimilarity values showed a high sensitivity against the change of grouping methodology (more than 50% of the clusters in the contrasted dendrograms were different); similar results were obtained by Adamson and Badwen when analyzing a data set of 36 chemicals.[48,49] The high variability found for the Tanimoto−unweighted average linkage algorithm indicates that, in the space of 10 dendrograms here considered, this dendrogram has an intermediate dissimilarity in respect to the other nine dendrograms. The most dissimilar HCA algorithm was the Tanimoto−single linkage, which means it was the dendrogram with more changes in the similarity relationships shown by its clusters.

The potential application of the cluster index was mentioned to assess the effect of varying the parameters defining a HCA study, such as the chemical representation, the dissimilarity measure, and the grouping methodology. Some other applications of this dissimilarity index are, for instance, the quantification of the effect of the ties in proximity in a given chemical data set, where the high probability of having ambiguities in the HCA results is well-known when the size of the data set increases and the (dis)similarity function employed has statistical bias toward particular proximity values.[50] In this particular case, the effect of a decision to overcome a tie can be assessed directly on the effect on the clusters, that is, how many clusters are affected for a particular tie.

An aspect to explore, after defining the cluster index, is the study of the distribution of its values for random trees defined over a particular set Q. Additionally, it must be proved that $d(sT_i,sT_j) = 0$ if $sT_i = sT_j$, where sT_i and sT_j are structural units of T_i and T_j. This particular proof must be developed for all the methods here discussed and also for the suggested h-tets indices.

Finally, triples, partition, and cluster indices deal with bare or ranked trees, that is, with dendrograms where the branch lengths are not considered. In this case, all three methods work on the topology of the trees to be contrasted. However, several applications of HCA and the determination of the number and sort of clusters extracted from a dendrogram are based on the geometrical viewpoint of the dendrogram (valued dendrograms). In those cases, the branch lengths in the dendrogram are important, and it is interesting to develop a dissimilarity measure including these structural aspects of the contrasted dendrograms. There are only two methods[15] measuring dissimilarity between trees that consider branch lengths, but none of them deal exclusively with clusters. A first attempt for reaching this goal is to attach to each cluster in a dendrogram the corresponding length of its associated subtree (the ultrametric height of the cluster). In such a case, the contrast between two dendrograms keeps being based on the clusters and its similarity information but now also includes the geometrical structure of each cluster.

It was mentioned that the representation of a dendrogram as another mathematical object, different from a graph, is computationally more tractable. In this paper, the three dissimilarity measures consider a dendrogram as a collection of subsets, but it is also possible to describe it using adjacency matrices.[23] These kinds of graph representations are well-known in mathematical chemistry and have been widely used in molecular dissimilarity calculations.[51−53] Hence, the description of a dendrogram as an adjacency matrix or as a collection of topological invariants is an alternative possibility for calculating dendrogram dissimilarities, and it would be important to contrast its advantages and disadvantages when compared to the methods here described.

ACKNOWLEDGMENT

G.R. specially thanks P. Willett and Y. Patel at the Department of Information Studies, University of Sheffield (U. K.), and Barnard Chemical Information Limited (nowadays, Digital Chemistry) for the access they permit to the Tanimoto data matrix used in the chemical example of application of the cluster index. The authors thank the valuable comments of the reviewers of this paper. G.R. thanks COLCIENCIAS and the Universidad de Pamplona in Colombia for the grant offered during the development of this research, and H.M. thanks the Universidad del Valle for the financial support.

APPENDIX

A1. Let C be a subgraph of the dendrogram T. It is said that C is a subtree iff either $C = T$ or (1) C does not contain the root node and (2) there is a node p in T whose degree is different than 1 such that C corresponds to one of the connected subgraphs obtained by subtracting p from T.

The idea of defining T as a subtree can be better understood if we consider a tree defined over a subset of Q; then, when that subset is Q, the subtree becomes T. T can be considered as a subtree also under the following argument. If a subtree (cluster) is the structural unit grouping elements of Q, then the subtree grouping all the elements in Q is T, which is a subtree.

A2. In order to define maximal 3-subtree, we first define 3-subtree. A 3-subtree is a subtree (A1) whose associated set has a cardinality less than or equal to 3. A maximal 3-subtree is any 3-subtree such that it is not possible to find another 3-subtree containing it. Two examples of maximal 3-subtrees for the dendrogram in Figure 1 appear in Figure 3 (bold graphs in the corresponding dendrogram).

A3. We define the operation of symmetric difference between two nonempty sets A and B by the identity $A \Delta B = (A \cup B) - (A \cap B)$. The cardinality of the symmetric difference is given by $|A \Delta B| = |A \cup B| - |A \cap B| = |A| + |B| - 2|A \cap B|$.

A4. The deletion of the external link joining the element x [external node (Figure 1)] to the dendrogram produces a partition $x|(Q - x)$. Because there are n elements in a dendrogram, then there are n partitions of the form $x|(Q - x)$. These partitions are always common to all dendrograms defined over Q and do not contribute to differentiate them.

A5. *Proposition.* Partition index $PI(T_1,T_2)$ yields the same result if either all the links in T or only the internal ones are deleted.

Proof. The number of external and internal links in T is n and $n - 2$, respectively. Then, the total number of links in T is $2n - 2$. These links yield $2n - 2$ binary partitions. However, if we avoid the repeated binary partition produced by deleting the internal link connected to the root node, then the number of nonredundant binary partitions obtained by deleting links is $2n - 3$. Hence, $|P(T_1,T_2)| = 2(2n - 3 - m')$, m' being the number of common partitions between T_1 and T_2. Now, $\max(m') = 2n - 3$ and $\min(m') = n$, which means that $\max[|P(T_1,T_2)|] = 2(n - 3)$ and $\min[|P(T_1,T_2)| = 0$. If we call $PI'(T_1,T_2)$ the value of the partition index when deleting all the links, then $PI'(T_1,T_2) = (2n - 3 - m')/(n - 3)$. But $m' = m + n$, where m represents the number of common binary partitions by deleting internal links from T. Then, we found that $PI'(T_1,T_2) = 1 - [m/(n - 3)]$, which is the same result found when deleting only internal links in T $[PI(T_1,T_2)]$.

A6. T has $2n - 1$ total clusters (including the $n + 1$ trivial ones), but if the trivial clusters are not considered as objects to contrast, then $CT_i = n - 2$ for a given T_i. We called c the number of clusters common to T_1 and T_2, and we found $0 \leq c \leq n - 2$, which yields $0 \leq |C(T_1,T_2)| \leq 2(n - 2)$. Now, if k trivial clusters are added to the clusters of T_i to contrast, then $CT_i' = n - 2 + k$, CT_i' being the set of clusters of T_i to contrast after adding k trivial clusters. Now, if we call c' the number of common clusters between T_1 and T_2, then $|C'(T_1,T_2)| = 2(n - 2 + k - c')$, where $|C'(T_1,T_2)|$ is the cluster metric for this case. In this situation, we have $k \leq c' \leq n - 2 + k$, and for that reason, $0 \leq |C'(T_1,T_2)| \leq 2(n - 2)$. Now, if we call $CI'(T_1,T_2)$ the cluster index when adding k trivial clusters to the ones to contrast, then $CI'(T_1,T_2) = (n - 2 + k - c')/(n - 2)$. But $c' = c + k$; then, $CI'(T_1,T_2) = 1 - [c/(n - 2)] = CI(T_1,T_2)$.

REFERENCES AND NOTES

(1) Everitt, B. S. *Cluster Analysis*; Edward Arnold: Bristol, U. K., 1993; Chapter 1, pp 1−10.

(2) Downs, G. M.; Barnard, J. M. *Clustering Methods and Their Uses in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Weinheim, Germany, 2002; Vol. 18, pp 1−40.

(3) Handl, J.; Knowles, J.; Kell, D. B. Computational Cluster Validation in Post-Genomic Data Analysis. *Bioinformatics* **2005**, *21*, 3201−3212.

(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(5) Potter, M. *Set Theory and Its Philosophy*; Oxford University press: Oxford, U. K., 2004; p 72.

(6) Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological Study of the Periodic System. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 68−75.

(7) Restrepo, G.; Villaveces, J. L. From Trees (Dendrograms and Consensus Trees) to Topology. *Croat. Chem. Acta* **2005**, *78*, 275−281.

(8) Restrepo, G.; Mesa, H.; Villaveces, J. L. On the Topological Sense of Chemical Sets. *J. Math. Chem*. **2006**, *39*, 363−376.

(9) Restrepo, G.; Llanos, E. J.; Mesa, H. Topological Space of the Chemical Elements and Its Properties. *J. Math. Chem*. **2006**, *39*, 401−416.

(10) Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological Study of the Periodic System. In *The Mathematics of the Periodic Table*; King, R. B., Rouvray, D. H., Eds.; Nova: New York, 2006; pp 75−100.

(11) Daza, M. C.; Restrepo, G.; Uribe, E. A.; Villaveces, J. L. Quantum Chemical and Chemotopological Study of Fourth Row Monohydrides. *Chem. Phys. Lett*. **2006**, *428*, 55−61.

(12) Mesa, H.; Restrepo, G. On Dendrograms and Topologies. *J. Math. Chem.* **2007**, Submitted.

(13) Felsenstein, J. The Number of Evolutionary Trees. *Syst. Zool*. **1978**, *27*, 27−33.

(14) Estabrook, G. F.; McMorris, F. R.; Meacham, C. A. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Syst. Zool*. **1985**, *34*, 193−200.

(15) Felsenstein, J. *Inferring Phylogenies*; Sinauer Associates: Sunderland, MA, 2004; Chapter 30, pp 528−535.

(16) Day, W. H. E. Optimal Algorithms for Comparing Trees with Labeled Leaves. *J. Class.* **1985**, *2*, 7−28.

(17) Penny, D.; Hendy, M. D. The Use of Tree Comparison Metrics. *Syst. Zool.* **1985**, *34*, 75−82.

(18) Steel, M. A.; Penny, D. Distributions of Tree Comparison Metrics—Some New Results. *Syst. Biol*. **1993**, *42*, 126−141.

(19) Penny, D.; Foulds, L. R.; Hendy, M. D. Testing the Theory of Evolution by Comparing Phylogenetic Trees Constructed from Five Different Protein Sequences. *Nature* **1982**, *297*, 197−200.

(20) Critchlow, D. E.; Pearl, D. K.; Qian, C. The Triples Distance for Rooted Bifurcating Phylogenetic Trees. *Syst. Biol.* **1996**, *45*, 323−334.

(21) Robinson, D. F. Comparison of Labeled Trees with Valency Three. *J. Comb. Theory* **1971**, *11*, 105−119.

(22) Waterman, M. S.; Smith, T. F. On the Similarity of Dendrograms. *J. Theor. Biol.* **1978**, *73*, 789−800.

(23) Boorman, S. A.; Olivier, D. C. Metrics on Spaces of Finite Trees. *J. Math. Psychol.* **1973**, *10*, 26−59.

(24) Rouvray, D. H. Definition and Role of Similarity Concepts in the Chemical and Physical Sciences. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 580−586.

(25) Restrepo, G.; Brüggemann, R. Ranking Regions through Cluster Analysis and Posets. *WSEAS Trans. Inf. Sci. Appl.* **2005**, *2*, 976−981.

(26) Chartrand, G.; Lesniak, L. *Graphs & Digraphs*.; Wadsworth & Brooks/Cole: Monterey, CA, 1986; pp 77−83.

(27) Deza, M. M.; Deza, E. *Dictionary of Distances*; Elsevier: Amsterdam, 2006; p 6.

(28) Page, R. D. M. *COMPONENT User's Manual (Release 1.5)*; University of Auckland: Auckland, New Zealand, 1989; Chapter 4, pp 4.1−4.7. URL: http://taxonomy.zoology.gla.ac.uk/rod/cpw.html (accessed Apr 2007).

(29) Slowinski, J. B. Review. *Cladistics* **1993**, *9*, 351−353.

(30) Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **1989**, *5*, 164−166.

(31) Swofford, D. L. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1*; Illinois Natural History Survey: Champaign, IL, 1993; p 53.

(32) Gusfield, D. Efficient Algorithms for Inferring Evolutionary Trees. *Networks* **1991**, *21*, 19−28.

(33) Golumbic, M. C.; Trenk, A. N. *Tolerance Graphs*; Cambridge University Press: Cambridge, U. K., 2004; Chapter 1, p 4.

(34) Bollobás, B. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability*; Cambridge University Press: Cambridge, U. K., 1988; p 1.

(35) Gross, J.; Yellen, J. *Graph Theory and Its Applications*; CRC Press: Boca Raton, FL, 1999; Chapter 2, pp 64−65.

(36) Chartrand, G.; Lesniak, L. *Graphs & Digraphs*; Chapman & Hall: London, 1996; pp 50−51.

(37) Kratsch, D.; Hemaspaandra, L. A. On the Complexity of Graph Reconstruction. *Math. Syst. Theory* **1994**, *27*, 257−273.

(38) Kelly, P. J. A Congruence Theorem for Trees. *Pac. J. Math*. **1957**, *7*, 961−968.

(39) Restrepo, G.; Brüggemann, R. Ranking Regions Through Cluster Analysis and Posets. *WSEAS Trans. Inf. Sci. Appl.* **2005**, *2*, 976−981.

(40) Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857−874.

(41) The NCI database is available at URL http://dtp.nci.nih.gov/ (accessed Apr 2007).

(42) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 644−649.

(43) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.

(44) The BCI software is available from Digital Chemistry Ltd. at URL http://www.digitalchemistry.co.uk (accessed Apr 2007).

(45) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; WH Freeman: San Francisco, CA, 1973.

(46) Gower, J. C. Measures of Similarity, Dissimilarity and Distance. In *Encyclopaedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Read, C. B., Eds.; Wiley: Chichester, U. K., 1982; pp 397−405.

(47) Brüggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C. E. W. Applying the Concept of Partially Ordered Sets on the Ranking

of Near-Shore Sediments by a Battery of Tests. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 918−925.
(48) Adamson, G. W.; Bawden, D. Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 204−209.
(49) Downs, G. M.; Willett, P. *Similarity Searching in Databases of Chemical Structures*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, U.S.A., 1996; Vol. 7, pp 1−66.
(50) MacCuish, J.; Nicolaou, C.; MacCuish, N. E. Ties in Proximity and Clustering Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 134−146.
(51) Kvasnicka, V.; Pospichal, J. Fast Evaluation of Chemical Distance by Tabu Search Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1109−1112.
(52) Diudea, M. V. Molecular Topology. 16. Layer Matrices in Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1064−1071.
(53) Rücker, C.; Rücker, G.; Meringer, M. Exploring the Limits of Graph Invariant- and Spectrum-Based Discrimination of (Sub)structures. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 640−650.

CI6005189

Appendix B

Brill Academic Publishers
P.O. Box 9000, 2300 PA Leiden
The Netherlands

Lecture Series on Computer and Computational Sciences
Volume 6, 2006, pp. 1-3

Measuring dissimilarity between dendrograms*

G. Restrepo[a,b,1], H. Mesa[c] and E. J. Llanos[d]

[a]Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia
[b]Lehrstuhl für Umweltchemie & Ökotoxikologie, Universität Bayreuth, 95440 Bayreuth, Deutschland
[c]Departamento de Matemáticas, Universidad del Valle, Cali, Colombia
[d]Observatorio Colombiano de Ciencia y Tecnología, Bogotá, Colombia

Received 24 July, 2006; accepted in revised form 25 July, 2006

Abstract: We discuss two dissimilarity measures for quantifying the dissimilarity between dendrograms defined over the same set. They are: partition metric and cluster metric. The first one is a standard method comparing dendrograms through their partitions obtained by deleting links of the dendrogram. The cluster metric is a novel method counting the number of common clusters between two dendrograms. We describe the two methods in a similar mathematical background. Finally, we remark the fact that cluster metric is a measure based on the concept of cluster that is the structural unit of a dendrogram and which contain all the structural information of the dendrogram.

Keywords: Hierarchical cluster analysis, resemblance between trees, partition, cluster.

Mathematics SubjectClassification: 05C75, 05C05

1. Introduction

Hierarchical Cluster Analysis (HCA) has become an important method of searching for similarities. It is used in chemistry for the understanding and classification of the chemical information (chemical database, substances). Normally, HCA studies begin defining the set Q of work by means of the features of its elements and then looking for the (dis)similarities among them using a (dis)similarity function, DF. Afterwards, the sets of similar elements are clustered using a grouping methodology, GM. A graphical representation of these clusters is a dendrogram (rooted acyclic-connected binary graph) (Figure 1), where the clusters of similar elements are represented as branches in the dendrogram.

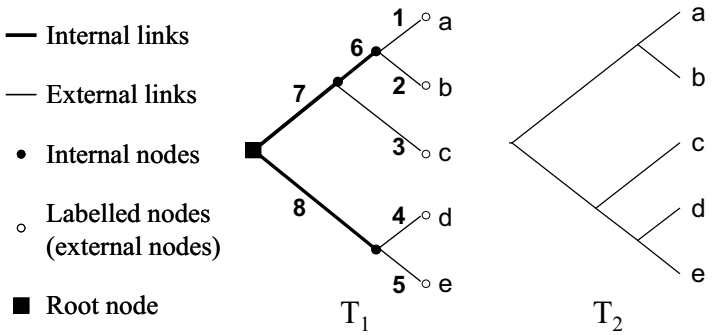


Figure 1: Two dendrograms. Types of nodes and links characterising them.

The number of dendrograms that can be defined over a set Q of cardinality $|Q|=n$ grows with n according to $|F|=(2n-3)!/[2n-2(n-2)!]$ [1]. Then, a natural question is, how can we measure the

*Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

[1] Corresponding author. E-mail: grestrepo@unipamplona.edu.co, guillermorestrepo@gmail.com

resemblance between dendrograms defined over a set Q? Two main mathematical methods have been proposed to answer this question. One is to define for any pair of dendrograms a measure of (dis)similarity indicating the extent of agreement [2]. The other is to define, as a function of the number of compared dendrograms, a new tree representing their consensus or area of agreement [2]. In this paper we discuss the first method, particularly two of these methods: partition metric and cluster metric indices, where the last one is a novel methodology. For the sake of clarity, a dendrogram in HCA is considered as a rooted acyclic-connected binary graph with labelled external nodes (Figure 1). It is important to state that the methods here discussed deal with bare and ranked dendrograms [3] that do not consider a numerical degree of similarity among the elements of Q but binary belonging relationships of the elements to branches (clusters) of the dendrogram.

2. Measuring the resemblance between dendrograms

We call Q the set of objects to be classified using HCA, the cardinality of the set Q is represented by $|Q|=n$ and T is a dendrogram (tree) on Q.

Partition metric index

Originally [4] this procedure was based on the comparison of partitions generated by deleting internal links (Figure 1) of a dendrogram. However, that procedure can be generalised to the comparison of partitions produced by the deletion of all the links [5] (internal and external ones) (Figure 1). The mathematical description of the method is as follows: A dendrogram T_i has $2n$-2 links (Figure 1). Let P_r be the partition (A1) of Q obtained by deleting the r-th link from T_i. Thus, $2n$-3 is the number of different partitions of Q, given T_i (A2). Let $CPT_i=\{P_r\}$ the collection of different partitions of T_i. The partition-symmetric difference, $P(T_1,T_2)$, is defined as $P(T_1,T_2)=CPT_1 \Delta CPT_2$ (A3). Thus, $P(T_1,T_2)$ yields the different partitions between T_1 and T_2 by deleting $2n$-2 links. Now, the partition metric is defined as the cardinality of $P(T_1,T_2)$. Thus, $|P(T_1,T_2)|=2(2n\text{-}3)\text{-}2|CPT_1 \cap CPT_2|$ (A4) and if we assume m as the number of common link partitions to T_1 and T_2, then $|P(T_1,T_2)|=2(2n\text{-}3\text{-}m)$. $\max|P(T_1,T_2)|=4n\text{-}6$ and $\min|P(T_1,T_2)|=0$ are the maximum and minimum values of $|P(T_1,T_2)|$, respectively. In this way the normalisation of $|P(T_1,T_2)|$, called partition metric index, $PM(T_1,T_2)$, is $PM(T_1,T_2)=1\text{-}[m/(2n\text{-}3)]$.

Cluster metric index

Let R be a cluster [6] (A5) of Ti in such a way that $R \subseteq Q$. Let CRT_i the collection of clusters of T_i (A6). Let $C(T_1,T_2)= CRT_1 \Delta CRT_2$ the cluster-symmetric difference, then $C(T_1,T_2)$ yields the different clusters between T_1 and T_2. Let cluster metric the cardinality of $C(T_1,T_2)$, $|C(T_1,T_2)|$, as follows (A7): $|C(T_1,T_2)|=2(n\text{-}2)\text{-}2|CRT_1 \cap CRT_2|$ and if we assume c as the number of common clusters to CRT_1 and CRT_2, then $|C(T_1,T_2)|=2(2n\text{-}1\text{-}c)$, where $\max|C(T_1,T_2)|=4n\text{-}2$ and $\min|C(T_1,T_2)|=0$ are the maximum and minimum values of $|C(T_1,T_2)|$, respectively. Now we call cluster metric index, $CM(T_1,T_2)$, the normalisation of $|C(T_1,T_2)|$. Thus, $CM(T_1,T_2)=1\text{-}[c/(2n\text{-}1)]$. Then, when the cluster metric reaches a value of zero it is because the compared dendrograms have all their clusters in common. In contrast, if the value of the cluster metric is one then the dendrograms are completely different, which occurs when they do not have common clusters.

Normally HCA is used to show clusters of similar elements. In general, clusters are branches of the dendrogram and also can be the dendrogram itself. But the importance of the clusters is not only related to the similarity concept. As Gusfield [7] has noted, a dendrogram is uniquely defined by its clusters. It means that the clusters not only determine the similarity neighbourhoods of the elements belonging to them [8-10] but they are in fact the fundamental pieces containing the structural information of the dendrogram. For this reason they can be used to characterise a dendrogram and also to measure the resemblance between couples of them. Knowing the clusters of a dendrogram it is possible to compare the common or different clusters between two dendrograms over the same set. A similar procedure is followed by the consensus methods [11-12] which look for common clusters among trees for producing a consensus tree.

As an example of application of the methods here discussed we calculated the dissimilarity between the two dendrograms of Figure 1. We obtained these results: $PM(T_1,T_2)=0$ and $CM(T_1,T_2)=0.1$. From these results it is possible to conclude that partition metric does not differentiate between isomorphic dendrograms. In contrast, the target is reached by the cluster metric.

Acknowledgments

The authors wish to thank COLCIENCIAS for the PhD grant given to one of the authors. Special thanks are offered to the Universidad de Pamplona in Colombia and Dr. A. González, rector of that University, for their financial support during this research.

Appendix

A1. Let Q be a non-empty set and P a collection of subsets of Q. P is called a partition of Q iff:

1. $Q = \cup_{p \in P} p$
2. If p_1 and $p_2 \in P$, then $p_1 \cap p_2 = \varnothing$

A2. Since $2n$-2 is the number of total links from a dendrogram, then the total number of P_r is also $2n$-2. However, the total number of "different" P_r's is not $2n$-2. We can explain this statement analysing the P_r's of the dendrogram T_1 in Figure 1, which are: $P_1=\{a|bcde\}$, $P_2=\{b|acde\}$, $P_3=\{c|abde\}$, $P_4=\{d|abce\}$, $P_5=\{e|abcd\}$, $P_6=\{ab|cde\}$, $P_7=\{abc|de\}$ and $P_8=\{abc|de\}$. We can see that there are two equal partitions P_r ($P_7 = P_8$). In general, there will always be two equal P_r's and each one of them arises by deleting each one of the connected links to the root node (Figure 1). Then, in order to actually have "different" P_r's we must select just one of these two P_r's related to the root node. Hence, the total number of different partitions is $2n$-2-1=$2n$-3.

A3. We define the operation of symmetric difference between two non-empty sets A and B by the identity: $A\Delta B=(A \cup B)-(A \cap B)$.

A4. By A3 $|P(T_1,T_2)|=|CPT_1|+|CPT_2|-2|CPT_1 \cap CPT_2|$ and having (A2) $|CPT_i|=2n-3$, then: $|P(T_1,T_2)|=2(2n-3)-2|CPT_1 \cap CPT_2|$.

A5. Given a dendrogram T, an external node x (Figure 1) is called a descendant of a node y if the path from x to the root node (Figure 1) passes through y [6].
We call a cluster the set of external nodes (Figure 1) that are descendant of a node v in T.

A6. If we consider the dendrogram T_1 shown in Figure 1, we have $CRT_1=\{\{a\},\{b\},\{c\},\{d\},\{e\},\{a,b\},\{d,e\},\{a,b,c\},\{a,b,c,d,e\}\}$.

A7. According to A3, $|C(T_1,T_2)|=|CRT_1|+|CRT_2|-2|CRT_1 \cap CRT_2|$. For calculating $|C(T_1,T_2)|$ we must know $|CRT_i|$.
Proposition.
Let Q be a finite set and T a dendrogram over Q, then $|CRT_i|=2n-1$.
Proof.
If $Q=\{a_1\}$, then there exists just one dendrogram with a unique cluster $\{a_1\}$. In this case $|CRT|=1=2(1)-1=2-1$. Let us suppose a $k \leq n$ and $|Q|=k$, then for every dendrogram T we have $|CRT_i|=2|Q|-1$. Let Q be a set such that $|Q|=n+1$ and T be a dendrogram. Every T has a root node (Figure 1) splitting T into subtrees (clusters) T_x and T_y, where T_x is a dendrogram over a set $A \subset Q$ and T_y a dendrogram over $A^C=Q-A$. Thus, $|CRT|=|CRT_x|+|CRT_y|+1$, $|A|<|Q|=n+1$ and $|Q-A|<|Q|=n+1$. Then $|A| \leq n$ and $|Q-A| \leq n$. Using the hypothesis of induction we have that $|CRT_x|=2|A|-1$ and $|CRT_y|=2|Q-A|-1$. Note that $|A|+|Q-A|=|Q|$ since A and $Q-A$ are disjoint, then $|CRT|=2|A|-1+2|Q-A|-1+1=2|Q|-1$. Now, since Q is made from n elements, then $|Q|=n$ and $|CRT|=2n-1$. ■
Using the above result we obtain $|C(T_1,T_2)|=2(n-2)-2|CRT_1 \cap CRT_2|$.

References

[1] J. Felsenstein, The number of evolutionary trees, *Syst. Zool*. **27**, 27-33(1978).

[2] G.F.Estabrook, F.R. McMorris and C.A. Meacham, Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units, *Syst. Zool*. **34**, 193-200(1985).

[3] S.A. Boorman and D.C. Olivier, Metrics on spaces of finite trees, *J. Math. Psych*. **10**, 26-59(1973).

[4] D. Penny, L.R. Foulds and M.D. Hendy, Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*. **297**, 197-200(1982).

[5] G. Restrepo, H. Mesa and E.J. Llanos, On the resemblance between dendrograms. Submitted to *J. Chem. Inf. Model.*

[6] D. Bryant, Building trees, hunting for trees, and comparing trees. Theory and methods in phylogenetic analyisis, PhD thesis, The University of Canterbury, (1997).

[7] D. Gustfield, Efficient algorithms for inferring evolutionary trees. *Networks*, **21**, 19-28(1991).

[8] G. Restrepo, H. Mesa, E.J. Llanos and J.L. Villaveces, Topological Study of the Periodic System. *J. Chem. Inf. Comput. Sci*. **44**, 68-75(2004).

[9] G. Restrepo and J.L. Villaveces, From Trees (Dendrograms and Consensus Trees) to Topology. *Croat. Chem. Acta* **78**, 275–281(2005).

[10] G. Restrepo, H. Mesa, E.J. Llanos and J.L. Villaveces. Topological Study of the Periodic System. In *The Mathematics of the Periodic Table*; R.B. King and D.H. Rouvray Eds.; Nova: New York, 2006.

[11] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, 2004.

[12] W.H.E. Day, Optimal algorithms for comparing trees with labeled leaves. *J. Class*. **2**, 7-28(1985).

Appendix C

CROATICA CHEMICA ACTA
CCACAA **80** (2) 261–270 (2007)

ISSN-0011-1643
CCA-3168
Original Scientific Paper

Partially Ordered Sets in the Analysis of Alkanes Fate in Rivers*

Guillermo Restrepo,[a,b,] Rainer Brüggemann,[c] and Kristina Voigt[d]**

[a]*Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia*

[b]*University of Bayreuth, Environmental Chemistry and Ecotoxicology, Bayreuth, Germany*

[c]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany*

[d]*GSF-National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Neuherberg, Germany*

RECEIVED NOVEMBER 22, 2006; REVISED APRIL 15, 2007; ACCEPTED APRIL 25, 2007

Dominance degree is introduced as a mathematical procedure to quantify the order relations between a pair of subsets contained in a partially ordered set obtained from the features of its elements. Dominance degree summarizes the partial order relations of the members of two subsets. If a member of one subset follows an order relation to a member of another subset, then the dominance degree informs how far this relation can be transferred to all elements of the two subsets. Dominance degree was applied to the study of 35 acyclic alkanes (from C_5H_{12} to C_8H_{18}) in two river-scenarios: hilly regions and lowland rivers. Each chemical was defined by three fate descriptors estimated by applying the module EXWAT from the E4CHEM package. It was found that C_nH_{2n+2} dominates C_mH_{2m+2} if $n > m$, which means that when considering the fate descriptors simultaneously, those of C_nH_{2n+2} are higher than those of C_mH_{2m+2}. Finally, some particular results were found for the linear isomer of each subset.

Keywords
partially ordered sets
alkanes
environmental scenarios
Hasse diagrams
WHASSE software
E4CHEM software

INTRODUCTION

Alkanes have been detected in several rivers around the world[1,2] and their presence is derived from natural biogenic, geologic and industrial sources.[3–6] In fact, it was estimated in 1991 that approximately 750,000 tons of hydrocarbons are annually transported by rivers to the Mediterranean Sea[7] and a large proportion of them are alkanes. Furthermore, in natural aquatic systems, for instance rivers, the freely dissolved fractions of hydrophobic organic contaminants, like alkanes, generally have the greatest impact on aquatic organisms representing the most ecotoxicologically relevant environmental residues.[8] Hence, studies of the distribution and fate of these chemicals in rivers are of the utmost environmental importance.

In this work, we use the module EXWAT from the software package E4CHEM in order to assess the risk of 35 acyclic alkanes in rivers. E4CHEM (available from the second author) consists of a system of modules describing the behaviour of chemicals in different environmental targets and depending on different stages of data availability. E4CHEM makes it possible to study the fate of chemicals in different targets (troposphere, stratosphere, plants, soil and rivers)[9] by the application of single

* Dedicated to Professor Haruo Hosoya in happy celebration of his 70th birthday.

** Author to whom correspondence should be addressed. (E-mail: grestrepo@unipamplona.edu.co)

simulation models for each target. Especially for rivers, E4CHEM includes the model EXWAT, which in an appropriate way combines environmental parameters of the river where the chemical is present with the substance properties. It is important to note that the use of EXWAT is supported by the agreement obtained between EXWAT predictions and experimental results for some other cases of chemicals in rivers.[10]

We consider two different river scenarios, each defined by its special features: a river in a hilly region and a lowland river. In this way, we can obtain descriptors for the fate of chemicals in each scenario that allow a comparison of the behaviour of the substances involved. This procedure may be considered as a ranking process of the chemicals and it can be studied by applying the concept of partially ordered sets (posets), as Brüggemann has shown in several studies.[11] The use of partial orders as a data exploring concept is called the Hasse diagram technique (abbreviated HDT) and here it is applied to an environmental chemistry case. By the application of the HDT, a Hasse diagram of a set under study is found. In chemical applications this type of diagram show which chemical/s is/are the most pollutant or the environment friendliest substance/s as well as which chemicals are in--between these substances. We show in this paper how some subsets of the chemicals under study can be analyzed by characterization of their order relationships, which are represented in the structure of the Hasse diagram.

Exposure Model EXWAT

A study of the fate of a substance in an environmental target cannot be based only on substance data but must also include environmental parameters of the media where the chemical is present. Thus the substance properties and environmental parameters are coupled by a deterministic mathematical exposure model (stationary). Such a model must be based on the differential mass balance,

$$dc / dt = \text{Input}\ (p, q) - \text{Output}(p, q). \qquad (1)$$

where p is the tuple of environmental parameters and q is the tuple of chemical properties. The Input(p, q) term includes the input due to the upstream concentration as well as the input by human activity into the first box modelling the river stretch.

However, real cases, such as a river for a particular case of EXWAT, have different targets. For instance, a river has two targets, sediment and water body of surface water. In these cases, a differential equation is needed for each target, which indeed is considered by EXWAT (mathematical details on the particular mass balance equations for these two targets are given in reference 10). Once the stationary concentration in the outflow of one river segment is determined, the inputs of the downstream section can be calculated. As we are interested in studying the fate of alkanes, we modelled each river scenario just by one segment, consisting of a water (W) and sediment (S) body (Figure 1) where all relevant processes are adequately described. Representation of a river segment, according to EXWAT, is depicted in Figure 1B.

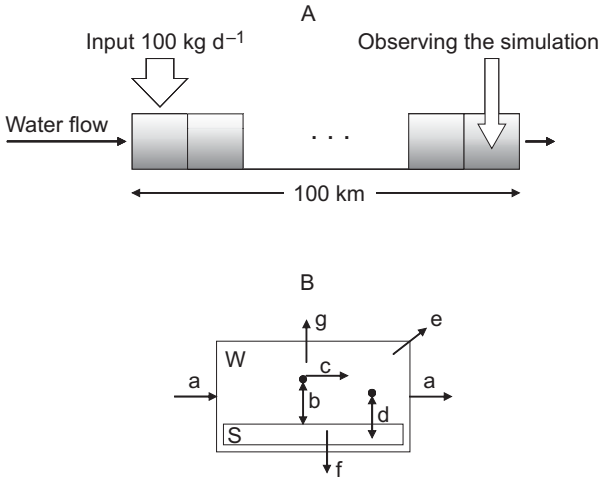


Figure 1. A) Partition of a river into several segments (boxes) in EXWAT, and B) the processes considered within each box (see text).

There is water inflow (a) with an upstream concentration of the substance and water outflow (a) with the resulting concentration due to different processes within the compartment. In W, suspended material that can be deposited or resuspended (b) is transported (c) (black circles). It is assumed that the dissolved substance is in equilibrium with its sorbed form on the suspended material. By dispersive forces, the dissolved chemical enters the interstitial water (d), which is assumed to be approximately in the order of amount of the molecular diffusion coefficient. Processes of degradation (e) can be included in the model; however, we considered the chemicals as conservative, *i.e.,* without degradation. Sediment burial (f) and volatilization (g) are considered as sinks; metabolites are not considered.

Once the missing physicochemical properties of the substance have been estimated by DTEST[12] (an E4CHEM module giving a high degree of automatic estimation of required chemical properties), the model EXWAT couples them with environmental data and physical parameters of the river. Some of the physicochemical properties estimated by DTEST are water solubility, vaporization entropy, vapour pressure and the partitioning coefficients K_{OW}, K_{OC}, and K_{AW}. Some of the river parameters of EXWAT are the ones listed in Table I and the concentration of suspended solids, temperature, pH, porosities, water discharge, and some others. Having this information and ignoring the temporal behaviour of the environmental system, EXWAT yields chemical concentrations in: the fluid phase (water and suspended matter), sediment, water (not including suspended matter), sediment matrix, pore water, suspended sediments and biomass.

TABLE I. Parameters input to EXWAT describing two scenarios of a river. H and L stand for the river in a hilly region and the lowland river, respectively.

River parameters	H	L
River length / km	100	100
Box length[a] / km	2	2
Volume flow / $m^3\ s^{-1}$	500	1000
Water body depth / m	2.5	3.5
Sediment depth / m	0.05	0.05
Width / m	150	300
Wind / $m\ s^{-1}$	5	5
Suspended matter content / $g\ m^{-3}$	100	100
w(Organic carbon in suspended matter)[b]	0.02	0.04
w(Organic carbon in sediment matrix)[b]	0.02	0.04
Sinking velocity of suspended matter / $m\ d^{-1}$	10	15

[a] See Figure 1A.
[b] w = mass fraction.

These concentrations can be regarded as fate descriptors or can be combined with flux parameters in order to yield additional descriptors.[13] Further, EXWAT, as a simple stationary model, provides a set of linear equations in its state variables; these equations may be mathematically related to each other and allows additionally the derivation of descriptor-descriptor relations. However, our interest in this paper is not to go into the details of those relationships but to show how the chemical fate is related to a posetic structure.

METHODS

A Chemical in Two River Scenarios

The river we studied was divided into two different scenarios: 1) river in a hilly region (H) and 2) lowland river (L). Parameters defining each scenario are given in Table I.

We selected three fate descriptors from the EXWAT results:

D_1: Total concentration of chemicals in the fluid phase, γ_w / $\mu g\ L^{-1}$;

D_2: Total concentration of chemicals in sediment, γ_s / $\mu g\ L^{-1}$;

D_3: Deposition flux: Concentration of sorbed chemicals on suspended sediment, γ_{ws}, times deposition velocity, Depos. $D_3 = (\gamma_{ws} * \text{Depos.}) / (\mu g \cdot m)\ (L \cdot d)^{-1}$.

Note that the values of γ_w and γ_s refer to different compartments; for example, in the hilly scenario γ_w refers to the water body with a volume of $7.5 \times 10^5\ m^3$ whereas γ_s refers to the sediment compartment with a volume of $1.5 \times 10^4\ m^3$.

Each descriptor was calculated by considering as the input rate of alkanes into the river a constant value of 100 $kg\ d^{-1}$ in order to differentiate the descriptor values of alkanes in the river (Figure 1A). The three concentrations estimated by EXWAT were performed in the box shown in Figure 1A. Note that our interest concerns the fate of chemicals and its methodological evaluation rather than the modelling of real amounts of alkanes in rivers. Our modelled river must be considered as a fictitious system.

General Remarks on the Hasse Diagram Technique

We introduce some definitions in order to illustrate some basic functionalities of the Hasse diagram technique,[11,14] implemented in the WHASSE software, available from the second author. WHASSE makes it possible to draw Hasse diagrams and to explore the influence of different parameters on them.

Definition 1. – We call x a chemical and G the ground set that is the set of chemicals.

Definition 2. – $D_i(x)$ is the numerical value of the i-th fate descriptor of the chemical x.

According to EXWAT, we have $D_i(x) = f[p, q(x)]$, where p is a tuple of environmental and physical parameters of the river and $q(x)$ is a tuple of properties of the chemical x. Then, $D_i(x)$ values characterize the fate of the chemical x in the river considered. In order to rank the chemicals according to their $D_i(x)$, the procedure followed by the Hasse diagram technique is to compare the fate descriptors of all chemicals.

Definition 3. – Let $x, y \in G$, then $x \le y$ if $D_i(x) \le D_i(y)$ for all i. This specific order relation is called a product- (or component-wise-) order and obeys the following axioms of order:

i) reflexivity: $\forall\ x \in G,\ x \le x$ (a chemical can be compared with itself);

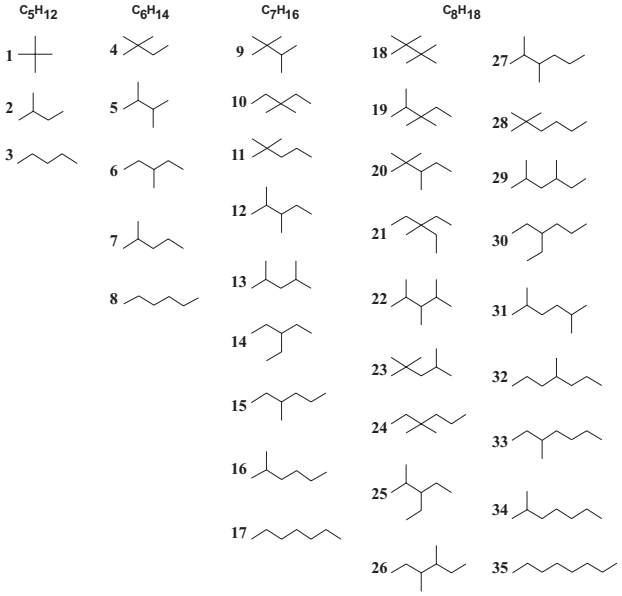
ii) antisymmetry: $\forall\ x, y \in G,\ x \le y$ and $y \le x \Rightarrow x = y$ (if x is better than y, then y is worse than x);

iii) transitivity: $\forall\ x, y, z \in G,\ x \le y$ and $y \le z \Rightarrow x \le z$ (if x is better than y and y is better than z, then x is better than z).

Note that in most mathematical textbooks the symbol $(G, \le)$ is used for a partially ordered set.[15] However, Brüggemann and co-workers have introduced the notation (G, D), where D is called the "information base" and is the set of D_i descriptors.[16] The reason for writing D instead of $\le$ is to emphasize that the order relation between the chemicals depends on the descriptors selected. Thus, the fact of having certain order relations between the chemicals in one scenario does not imply that those chemicals will have the same order relations in another. The cause of this behaviour is that $D_i(x)$ depends, besides chemical properties, on the river parameters, as mentioned above.

If $D_i(x) \le D_i(y)$ for some indices i and $D_i(y) \le D_i(x)$ for one or some other indices, then x and y are "incomparable", denoted as $x \parallel y$. A graph P representing the order relations found in G can be drawn,[17] where the order relation $\le$ is represented by an arrow going, for instance, from the better chemical to the worse. But P contains unnecessarily many edges, which can be avoided by a transitive reduction[18] eliminating all edges that arise solely from the transitivity axiom. After such "transitivity reduction", a more parsimonious graph H, called the Hasse diagram, can be drawn.

TABLE II. Molecular graphs and labels for the 35 alkanes considered in the fate analysis.



RESULTS

The set G of chemicals in this study is made from the complete set of 35 acyclic alkanes ranging from C_5H_{12} to C_8H_{18}: three C_5H_{12}, five C_6H_{14}, nine C_7H_{16} and eighteen C_8H_{18} isomers (Table II). The physicochemical properties of each alkane (water solubility, vapour pressure, melting point, boiling point and octanol/water partition coefficient) were taken from the Chemical Properties Handbook[19] and the Handbook of Physical Properties of Organic Chemicals;[20] the missing values were estimated using the module DTEST of E4CHEM. Having the complete pool of physicochemical properties coming from the literature and from estimations by DTEST, we use the EXWAT model of E4CHEM in order to generate the three fate descriptors D_1, D_2 and D_3 for each scenario.

General Dependences of the Descriptors

The alkane labels were assigned following the increasing values of the Wiener index[21] (as a measure of branching index) of each molecule (Table II). The values of the three fate descriptors for each alkane appear in Table III. Note that concentrations D_1(H) and D_1(L) relate to the volume of the water body while those of D_2(H) and D_2(L) relate to the volume of the sediment. A simple equilibrium calculation shows that, due to the small volume of the sediment, the variation of D_1 can be quite low whereas that of D_2 can be rather high.

Before discussing the results obtained using the Hasse diagram technique, we analyze separately the behaviour of each fate descriptor for the 35 alkanes in both, H and L, scenarios.

We found that the trends present in H are also present in L (Figure 2). We observed that D_1 (chemical concentration in the fluid phase) is mainly determined by the molecular weight of the molecules. Thus, we classified D_1 into four subsets of values corresponding to C_5H_{12}, C_6H_{14}, C_7H_{16} and C_8H_{18} isomers, respectively. We found that D_1 values for both scenarios fulfil this order relationship: $C_8H_{18} > C_7H_{16} > C_6H_{14} > C_5H_{12}$. In all the cases, the D_1 values of isomers are nearly the same; however, with the increase of the molecular weight, the linear isomer of each subset increases its D_1 value (Table III, Figure 2). We observed, for the case of alkanes in L, that the linear isomer of C_7H_{16} reached the value of D_1 corresponding to the C_8H_{18} isomers (Figure 2). This result

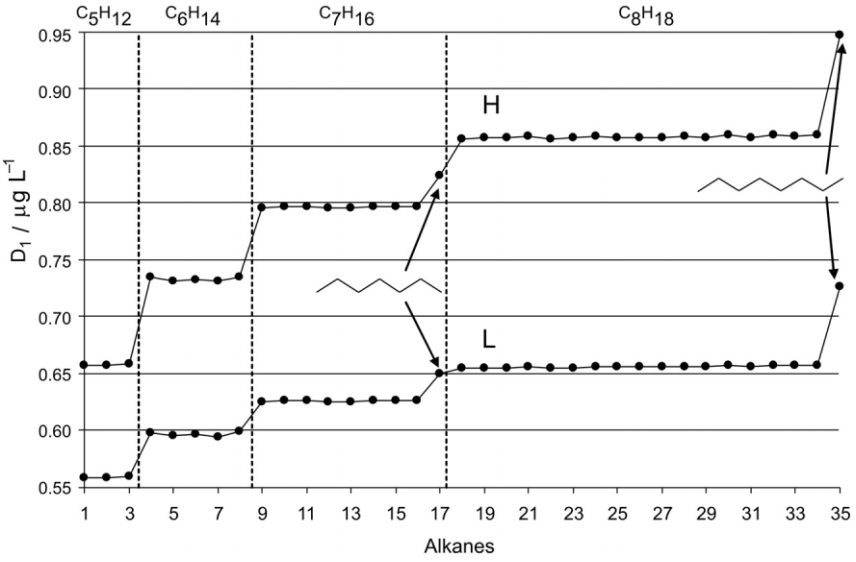


Figure 2. Total chemical concentration in the fluid phase (D_1) of 35 alkanes in hilly region (H) and in lowland (L) rivers. Some structures are drawn (see text).

TABLE III. Values of the fate descriptors for the alkanes studied. D_i(H) and D_i(L) stand for the D_i values in the hilly region (H) and lowland river (L) scenarios, respectively

Alkanes	D_1(H)	D_2(H)	D_3(H)	D_1(L)	D_2(L)	D_3(L)
	$\mu g\ L^{-1}$	$\mu g\ L^{-1}$	$(\mu g \cdot m)\ (L \cdot d)^{-1}$	$\mu g\ L^{-1}$	$\mu g\ L^{-1}$	$(\mu g \cdot m)\ (L \cdot d)^{-1}$
1	0.658	0.078	0.010	0.559	0.065	0.018
2	0.657	0.088	0.002	0.559	0.045	0.003
3	0.658	0.092	0.020	0.560	0.093	0.034
4	0.735	0.169	0.060	0.598	0.209	0.096
5	0.732	0.105	0.024	0.596	0.104	0.038
6	0.733	0.127	0.036	0.596	0.140	0.058
7	0.731	0.091	0.015	0.595	0.077	0.024
8	0.736	0.191	0.072	0.599	0.245	0.116
9	0.796	0.136	0.038	0.626	0.144	0.060
10	0.796	0.150	0.046	0.626	0.166	0.072
11	0.796	0.150	0.046	0.626	0.166	0.072
12	0.796	0.143	0.042	0.626	0.154	0.065
13	0.796	0.143	0.042	0.626	0.154	0.065
14	0.797	0.158	0.050	0.626	0.178	0.079
15	0.797	0.158	0.050	0.626	0.178	0.079
16	0.797	0.158	0.050	0.626	0.178	0.079
17	0.824	0.882	0.448	0.651	1.279	0.682
18	0.856	0.284	0.117	0.655	0.361	0.178
19	0.857	0.305	0.129	0.656	0.392	0.195
20	0.857	0.305	0.129	0.655	0.392	0.195
21	0.859	0.346	0.151	0.657	0.453	0.229
22	0.856	0.284	0.117	0.655	0.361	0.178
23	0.857	0.305	0.129	0.655	0.392	0.195
24	0.859	0.346	0.151	0.657	0.453	0.229
25	0.858	0.322	0.138	0.656	0.417	0.209
26	0.858	0.322	0.138	0.656	0.417	0.209
27	0.858	0.322	0.138	0.656	0.417	0.209
28	0.858	0.346	0.151	0.657	0.453	0.229
29	0.858	0.322	0.138	0.656	0.417	0.209
30	0.860	0.372	0.166	0.658	0.493	0.251
31	0.858	0.322	0.138	0.656	0.417	0.209
32	0.860	0.372	0.166	0.658	0.493	0.251
33	0.859	0.372	0.166	0.658	0.493	0.251
34	0.860	0.372	0.166	0.658	0.493	0.251
35	0.947	2.938	1.572	0.727	4.001	2.172

may suggest that when considering isomers of the set C_9H_{20} (not studied here), perhaps the linear isomer of C_8H_{18} could reach the values of D_1 for C_9H_{20} isomers, which is supported by the high D_1 value of the linear C_8H_{18} isomer.

In order to relate D_1 with some molecular structural parameter, we calculated the complete pool of 708 molecular descriptors available in the MOLGEN-QSPR software (arithmetical, topological, electrotopological, and geometrical descriptors).[22] After these calculations, we found a high Pearson correlation ($R > 0.9$) between D_1 and several molecular branching indices (W, $^1\chi$, MTI, and MTI'), which are in turn highly correlated to molecular weight. Thus, the fate of alkanes in the fluid phase is determined mainly by the molecular weight of the substances.

Regarding D_2 and D_3, we found a high correlation between these fate descriptors ($R > 0.9$). However, in

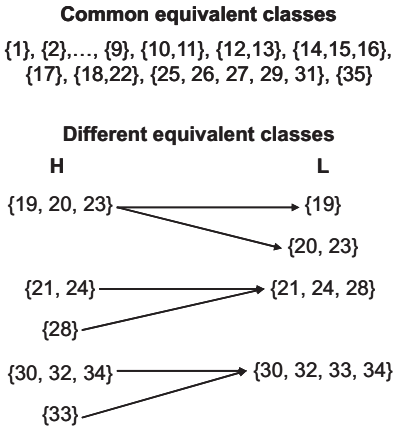
contrast to D_1, we found no clear distinction between groups according to molecular weight (Figure 3). When looking for correlations between D_2 and D_3 through our pool of molecular descriptors, we did not find any relevant ($R > 0.8$) relationship. This result suggests that D_2 and D_3, contrary to D_1, are not related to the molecular parameters of alkanes. The high correlation between D_2 and D_3 suggests a similar trend in the alkane concentrations in sediments and also in suspended sediments. Note that D_3 contains the term γ_{ws}, the concentration of chemicals on suspended sediment. Despite the lack of correlation between the degree of branching and D_2 and D_3, it is important to note the high D_2 and D_3 values of alkanes 17 and 35, which correspond to the linear structures of C_7H_{16} and C_8H_{18}, respectively (Figure 3). This trend is not observed for the linear structures of the light alkanes C_5H_{12} and C_6H_{14}. A similar behaviour was observed for the same linear alkanes when considering D_1.

Having described each fate descriptor separately, we can discuss the effect on each descriptor of changing the river parameters from H to L. We observe that D_1 decreases when we change from H to L (Figure 2). This means that the concentration of alkanes in fluid phase is lower in lowland rivers than in rivers in hilly regions. The reason is the high dilution due to higher discharge in the lowland river. Now, considering D_2, we observe a small increment in L compared to H. On the other hand, D_3 increases in L compared to H, because the deposition of alkanes on suspended sediments is faster in L than in H. In general, the change of scenario, from H to L, makes D_1 decrease in contrast to increasing D_2 and D_3. All in all, even if we consider structurally simple alkanes, it is difficult to oversee their fate in different environmental scenarios. Here, the concept of partially ordered sets is helpful and is applied in the next section.

Hasse Diagram of Alkanes in Hilly Region and Lowland Rivers

It was mentioned in the above section that some alkanes share some fate descriptor values; it means that two alkanes x, $y \in G$ may have $D_i(x) = D_i(y)$ for $i = 1, 2, 3$. We say that then x and y belong to an equivalence class K, and we select one representative of such a class. These selected chemicals together with the chemicals for which $D_i(x) \neq D_i(y)$ for i's, are gathered in the set T of representatives. Thus, we draw the Hasse diagram over the set T of representatives. The equivalence classes for both scenarios are shown in Scheme 1.

The set of representatives for scenario H is { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 17, 18, 19, 21, 25, 28, 30, 33, 35} and the one for L is { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 17, 18, 19, 20, 21, 25, 30, 35}. We note four subsets of isomers (C_5H_{12}, C_6H_{14}, C_7H_{16} and C_8H_{18}) in each Hasse diagram (Figure 4) and we will discuss some of their features in the following text.



Scheme 1. Equivalence classes in H and L and their relationships.

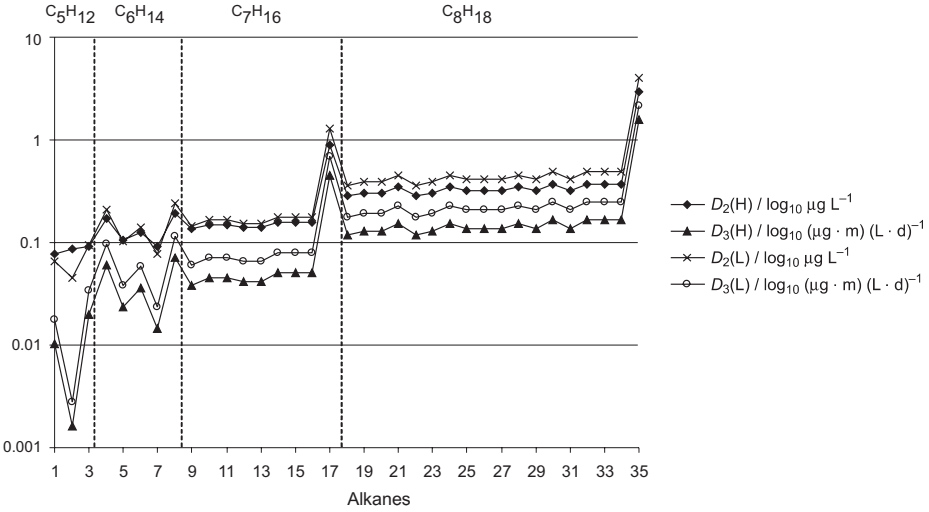


Figure 3. Total chemical concentration in sediment (D_2) and deposition flux (D_3) of 35 alkanes in hilly region (H) and in lowland rivers (L). D_i(H) and D_i(L) stand for the values of the $\log_{10}$ of D_i in the H and L scenarios, respectively.

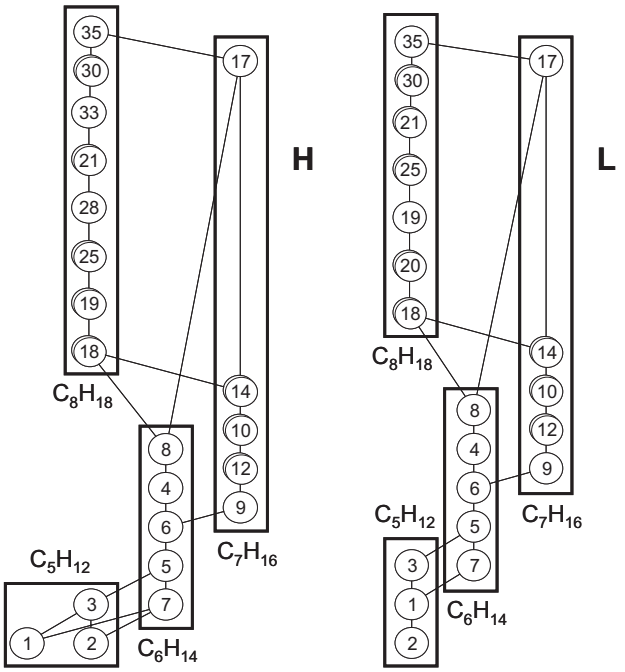


Figure 4. Hasse diagrams of 35 alkanes in the river scenarios H (hilly regions) and L (lowland). Double circles indicate equivalence classes with more than one alkane.

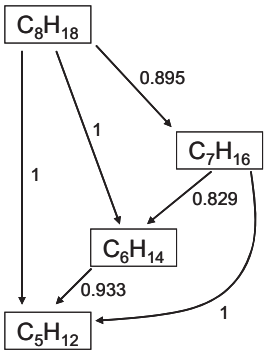


Figure 5. Dominance diagram: The scheme corresponds to the location of subsets in Figure 4 (see text).

General Observations

Brüggemann and co-workers have demonstrated the versatility of using Hasse diagrams in ranking[17,23] the chemicals in a given environmental space defined by descriptors.

In our particular case, the rank is built from fate descriptors (D_1, D_2 and D_3) and in all the cases their high values (upper part of the diagram) may imply a hazard: either by being transported downstream with adverse effects on aquatic organisms or because of accumulation in sediments (chemical time bomb effect). In contrast, if D_1, D_2 and D_3 have low values for an alkane, then this substance is "better" or "less unfriendly" regarding the environment and is located in the lower part of the diagram. Hence, the diagrams shown in Figure 4 can be interpreted as a rank of alkanes in the given scenario. If we consider the H diagram, we can see that the most pollutant alkanes are the C_8H_{18} isomers. Then, going down in the diagram, we found C_7H_{16} isomers, then C_6H_{14}, and finally C_5H_{12}. In summary, having classified the alkanes into four isomer subsets, it seems possible to establish one ranking according to their fate descriptors. In order to do that, we introduce the concept of dominance degree (Dom). Let us assume that $G' \subset G$ and $G'' \subset G$ with $G' \cap G'' = \varnothing$; if $\forall x \in G'$, $\forall y \in G''$, $x > y$ then G' dominates G'' and we write $G' \blacktriangleright G''$. In the practice of empirical posets the condition "for all" is too hard. Therefore we are introducing the dominance degree Dom (G', G'') = N_R / N_T, where N_R (N realized) = $|\{(x, y), x \in G', y \in G''$ and $x > y\}|$ and $N_T = |G'| \cdot |G''|$. Note that the counting is based on the complete object set (35 objects) rather than that on T, because different equivalence classes appear in H and L. If Dom (G', G'') > 0.5, then we write $G' \blacktriangleright G''$. We show schematically in Figure 5 the Dom (G', G'') results for each pair of isomer groups. For example, the calculation of Dom (C_6H_{14}, C_5H_{12}), in both scenarios, is performed by determining $N_R = |\{(4, 1), (4, 2), (4, 3), (5, 1), (5, 2), (5, 3), (6, 1), (6, 2), (6, 3), (7, 1), (7, 2), (8, 1), (8, 2), (8, 3)\}| = 14$ and $N_T = |C_6H_{14}| \cdot |C_5H_{12}| = 15$. Then, Dom ($C_6H_{14}$, C_5H_{12}) = 14 / 15 = 0.933.

An arrow $\blacktriangleright$ is drawn for each dominance relation; each of these relations is characterized by its dominance degree and, in this case:

$$\text{Dom } (C_nH_{2n+2}, C_mH_{2m+2}) \text{ is } \begin{cases} > 0.5 & \text{for all } n > m \\ = 0 & \text{for all } n < m \end{cases}.$$

Note that the case $n = m$ is not considered because the subsets compared ought to be disjoint (by definition). It is important to note that the same dominance diagram holds for both scenarios. In summary, we find $C_8H_{18} \blacktriangleright C_7H_{16} \blacktriangleright C_6H_{14} \blacktriangleright C_5H_{12}$, which generalizes our finding with only one descriptor (D_1), as discussed above. Further discussion on the mathematical properties of the dominance degree is given in reference 24; another application of this concept to environmental studies can be found in reference 25.

Particular Object Related Observations

For each isomer subset, the linear alkane is the chemical presenting simultaneously high values of its fate descriptors. They are 35 for C_8H_{18}, 17 for C_7H_{16}, 8 for C_6H_{14} and 3 for C_5H_{12} (compare Table II). Now, from a general analysis of the Hasse diagrams we can say that the maximal[15] element is 35, which is also the greatest[15] element. There are two minimal[15] elements, 1 and 2, for the H diagram and only one, 2, for the L diagram, which becomes the smallest element of this diagram. This means that the linear C_8H_{18} alkane is the substance from the complete set of 35 alkanes whose fate descriptors make it the most potentially problematic compound in environmental terms.

Similarly, the fact of having two minimal elements in H means that there is no alkane with simultaneous fate descriptors lower than 1 and 2. When considering the L diagram, 2 becomes the least alkane.

Comparing H and L Hasse Diagrams

In the H diagram, each subset of isomers (C_5H_{12}, C_6H_{14}, C_7H_{16} and C_8H_{18}) appears as a chain or as belonging to a chain,[15] except for the C_5H_{12} subset where there is no linear order[15] between its members. When we analyze the effect of changing the river parameters from hilly regions to a lowland river, we found two general changes:

i) C_5H_{12} subset becomes a linear order, a chain.

ii) There are some internal rearrangements within C_7H_{16} and C_8H_{18} subsets.

The reason for i) can be explained first by mentioning the reason why 1 and 2 in H are incomparable and then why it changes in L. Chemical 1 is incomparable with chemical 2 ($1 \parallel 2$) in H because $D_1(1) > D_1(2)$, $D_2(1) < D_2(2)$ and $D_3(1) > D_3(2)$ (Table III); hence D_2 is the cause of incomparability. The reasons are difficult to explain because there are many competitive processes, which, on the one side, depend on the chemical properties and, on the other side, on the environmental ones. For example, high accumulation in the sediment need not necessarily be implied by a high deposition velocity of suspended matter. When we analyze the order relations for these two alkanes in the L diagram, we find that $D_1(1) > D_1(2)$, $D_2(1) > D_2(2)$ and $D_3(1) > D_3(2)$, hence $1 > 2$. In summary, the linear order of the C_5H_{12} subset in the L diagram is due to the change in the D_2 order relation for 1 and 2. In other words, the change in the concentration of alkanes 1 and 2 in sediments is the cause of the linear order in the C_5H_{12} subset.

The second change in Hasse diagrams, when comparing H and L, is caused by the redistribution of some equivalence classes (Scheme 1), which do not alter the order relations among the chemicals. This is due to small numerical variations of the descriptors defining each chemical in each scenario. These variations normally occur just in one descriptor while the remaining two keep their order relations. Moreover, these variations are within the limits of discriminatory power of the descriptors since they occur in the last decimal position. For instance, the relation $\{28\} < \{21, 24\}$ is found in H and these chemicals are rearranged to $\{21, 24, 28\}$ in L. The cause of $\{28\} < \{21, 24\}$ in H is that $D_1(28) = 0.858$ is lower than $D_1(21) = D_1(24) = 0.859$ since $D_2(28) = D_2(21) = D_2(24)$ and $D_3(28) = D_3(21) = D_3(24)$. When varying the river conditions from H to L, then the small numerical difference between $D_1(28)$ and $D_1(21) = D_1(24)$ becomes an equality and 28 joins 21 and 24 in an equivalence class. In general, all the rearrangements within isomer subsets obey these kinds of small numerical differences.

CONCLUSIONS

The combination of basic chemical fate properties with partial ordering concepts is an interesting tool for drawing general conclusions on simultaneous analysis of fate descriptors. The dominance degree was introduced in this paper as a mathematical tool able to quantify the simultaneous effect of different descriptors on the general ranking of subsets of chemicals. The dominance degree is a measure of the number of real comparabilities between the members of two different subsets and the theoretical number of comparabilities holding if all the members in one subset are "greater" or "lower" than the members of the other subset. By applying this measure to the hilly and lowland Hasse diagrams of alkanes we found that the isomers with highest molecular weight dominate, or are more problematic than the rest of the isomer subsets following this relationship: $C_8H_{18} \blacktriangleright C_7H_{16} \blacktriangleright C_6H_{14} \blacktriangleright C_5H_{12}$, where $G' \blacktriangleright G''$ means that the subset G' dominates the subset G''. According to our definition of Dom $(G',G'') > 0.5$, the above result means that more than 50 % of the C_8H_{18} isomers dominate the C_7H_{16} ones, more than 50 % of the C_7H_{16} isomers dominate the C_6H_{14} ones, etc.

The order relationships found in the dominance of heavy alkanes over the light ones suggest the possibility of interdependence between the dominance degree and the molecular weight of the alkanes. To test this hypothesis, it would be interesting to consider more acyclic alkanes as objects of study.

It was found that, within each isomer subset, the linear alkane is the most environmentally problematic substance because of its relatively high concentrations in the water and the sediment bodies of the river scenarios considered.

Analysis of fate descriptors allows the conclusions that 1) the concentration of alkanes in the fluid phase of each scenario was determined mainly by the molecular weight, 2) the chemical concentration in sediments and the deposition flux were not related to the molecular weight, nor to any molecular parameter of the alkanes, and 3) the change of the river parameters from a river in hilly regions to a lowland scenario caused the chemical concentration in the fluid phase to decrease while the concentration in sediments and the deposition flux increased.

A general feature of the Hasse diagram technique is that it is based on the qualitative comparison of the descriptors characterizing the objects. Hence, the fact of having two chemicals x and y with $x < y$ does not necessarily exclude that their actual concentrations might be so close that an experimental determination might yield identical values for both x and y. Then, the practical importance of the posetic results such as the ones shown in this manuscript are the relations in the graph, rather than the geometrical ones. This enables, when assessing che-

micals, to determine pollutant substances, or potentially problematic ones. However, these results must not be interpreted from a geometrical point of view where, for instance, $x < y < z$ means 1 ppb < 2 ppb < 3 ppb. In fact, $x < y < z$ might mean 1.001 ppb < 1.002 ppb < 1.003 ppb, and if the uncertainty of the measure is ± 0.002, then x, y and z become an equivalence class. A similar case as the one described here are the concentrations shown in Table III, where the aqueous concentrations are close to each other for the majority of the alkanes belonging to a particular isomer subset. This situation causes minor variations within the subsets for the other two descriptors to be responsible of the comparabilities found between isomers but it does not mean that the "higher" chemical represents a markedly different chemical concentration when compared to a "lower" chemical in the ranking.

In this research, we considered just two river scenarios with the aim of checking how the order relations between chemicals change from scenario to scenario. However, this methodology can be applied to new river scenarios, perhaps defined by local parameters pertaining to particular rivers, particular sets of pollutants and particular input patterns. It may also be applied to chemicals characterized by some other risk-relevant factors such as toxicities or some other combinations of chemical attributes.

Several authors[14,26] have pointed out that posetic structures are present in different fields of chemistry and particularly Brüggemann[14] has shown the advantages of their study in environmental chemistry. The procedure developed here to deal with the order relations between subsets of objects may be applied to any poset and it is interesting to go into more details of its application when considering chemical posets like those developed by Randić[27] and Daza and Bernal,[28] among others.

Acknowledgements. – The authors thank A. Kerber of the Department of Mathematics, Universität Bayreuth (Germany), for permitting access to the MOLGEN-QSPR software; they are also grateful for the valuable comments of the reviewers of this paper. G. Restrepo thanks COLCIENCIAS and the Universidad de Pamplona in Colombia for the grant offered during the development of this research.

REFERENCES

1. B. O. Ekpo, O. E. Oyo-Ita, and H. Wehner, *Naturwissenschaften* **92** (2005) 341–346.
2. O. P. Heemken, B. Stachel, N. Theobald, and B. W. Wenclawiak, *Arch. Environ. Contam. Toxicol.* **38** (2000) 11–31.
3. B. R. T. Simoneit, in: J. P. Riley and R. Chester (Eds.), *Chemical Oceanography*, Vol. 7, 2nd ed., Academic Press, New York, 1978, pp. 233–311.
4. M. A. Mazurek and B. R. T. Simoneit, in: L. H. Keith (Ed.) *Identification and Analysis of Organic Pollution in Air*, ACS Symposium, Ann Arbor Science Publishers/Butterworth, Woburn, 1983, p. 353.
5. T. A. T. Aboul-Kassim and B. R. T. Simoneit, *Environ. Sci. Technol*. **29** (1995) 2473–2483.
6. T. A. T. Aboul-Kassim and B. R. T. Simoneit, *Marine Chem*. **54** (1996) 135–158.
7. I. Bouloubassi and A. Saliot, *Mar. Pollut. Bull*. **22** (1991) 588–594.
8. Y. Wang, Y. Huang, J. N. Huckins, and J. D. Petty, *Sci. Technol*. **38** (2004) 3689–3697.
9. R. Brüggemann and U. Drescher-Kaden, *Einführung in die modellgestützte Bewertung von Umweltchemikalien – Datenabschätzung, Ausbreitung, Verhalten, Wirkung und Bewertung*, Springer-Verlag, Berlin, 2003.
10. R. Brüggemann and S. Trapp, *Chemosphere* **17** (1988) 2029–2041.
11. R. Brüggemann, E. Halfon, G. Welzl, K. Voigt, and C. Steinberg, *J. Chem. Inf. Comp. Sci*. **41** (2001) 918–925.
12. R. Brüggemann, G. Restrepo, and K. Voigt, *WSEAS Trans. Inf. Sci. Appl*. **2** (2005) 1023–1033.
13. R. Brüggemann, G. Restrepo, and K. Voigt, *J. Chem. Inf. and Model*. **46** (2006) 894–902.
14. R. Brüggemann and L. Carlsen, *Partial Order in Environmental Sciences and Chemistry*, Springer Verlag, Berlin, 2006.
15. W. T. Trotter, *Combinatorics and Partially Ordered Sets Dimension Theory,* John Hopkins Series in the Mathematical Science, The J. Hopkins University Press, Baltimore, 1991.
16. R. Brüggemann, J. Schwaiger, and R. D. Negele, *Chemosphere* **30** (1995) 1767–1780.
17. R. Brüggemann and H.-G. Bartel*, J. Chem. Inf. Comput. Sci*. **39** (1999) 211–217.
18. K. Simon, *Efficient Algorithms for Perfect Graphs*, B. G. Teubner, Stuttgart, 1992.
19. C. L. Yaws, *Chemical Properties Handbook*, McGraw-Hill, New York, 1999.
20. P. H. Howard and W. M. Meylan, *Handbook of Physical Properties of Organic Chemicals*, CRC, Boca Raton, 1996.
21. H. Wiener, *J. Am. Chem. Soc*. **69** (1947) 17–20.
22. C. Rücker, M. Meringer, and A. Kerber, *J. Chem. Inf. Comput. Sci.* **44** (2004) 2070–2076. http://www.mathe2.uni-bayreuth.de/molgenqspr/start.html
23. G. Restrepo and R. Brüggemann, *WSEAS Trans. Inf. Sci. Appl*. **7** (2005) 976–981.
24. G. Restrepo and R. Brüggemann, Submitted to *J. Math. Chem.*
25. G. Restrepo, M. Weckert, R. Brüggemann, S. Gerstmann, and H. Frank, Submitted to *Environ. Sci. Technol.*
26. D. J. Klein, *J. Math. Chem*. **18** (1995) 321–348.
27. M. Randić, *Chem. Phys. Lett.* **55** (1978) 547–551.
28. E. E. Daza and A. Bernal, *J. Math. Chem.* **38** (2005) 247–263.

SAŽETAK

Parcijalno uređeni skupovi u analizi sudbine alkana u rijekama

Guillermo Restrepo, Rainer Brüggemann i Kristina Voigt

Kao matematičku mjeru uređaja za podskupove parcijalno uređenog skupa uveli smo stupanj dominacije koji se izvodi iz svojstava njihovih elemenata. U stupnju dominacije sažima se parcijalni uređaj parova elemenata iz dvaju podskupova. Stupanj dominacije pokazuje koliko je uređaj između neka dva elementa iz različitih podskupova, svojstven svim parovima njihovih elemenata. Stupanj dominacije primijenjen je u komparativnoj analizi sudbine 35 acikličkih alkana (od C_5H_{12} do C_8H_{18}) prema riječnim scenarijima za brdska i nizinska područja. Svakom kemijskom spoju pridružena su tri deskriptora sudbine, određena pomoću modula EXWAT iz programskog paketa E4CHEM. Utvrđeno je da C_nH_{2n+2} dominira nad C_mH_{2m+2} kad je $n > m$, što znači da deskriptori za C_nH_{2n+2} imaju uglavnom veće vrijednosti od onih za C_mH_{2m+2}. Određeni rezultati dobiveni su za linearne izomere iz svakog podskupa.

Appendix D

Refrigerants ranked by Partial Order Theory

Guillermo Restrepo[1,2], Monika Weckert[1], Rainer Brüggemann[3],
Silke Gerstmann[1], Hartmut Frank[1]

Abstract

Forty refrigerants used in the past, used presently, and some proposed substitutes, were studied in respect to their ozone depletion potential, global warming potential, and atmospheric life times. They were ranked using the Hasse diagram technique, a mathematical method which permits to draw diagrams representing order relations among chemicals. The refrigerants were divided into 13 chemical classes (subsets) of which the most prominent ones are the chlorofluorocarbons (CFC), hydrofluorocarbons, hydrochlorofluorocarbons and hydrofluoroethers. Order relations among these subsets were calculated applying dominance and separability degrees. The dominance degree is a measure indicating the extent to which descriptors of members of one subset are higher than those of members of other subsets; the separability degree is a measure for the incomparability or lack of order relations between elements of two subsets. By application of these measures to the 13 chemical subsets it was found that more than half of the order relations among them are complete dominances; this means a high degree of comparability among subsets permitting to find the ones most problematic in environmental terms. This is the case for the CFC and for some of the hydrofluoroethers.

1. Introduction

Chlorofluorocarbons (CFC) have been used as refrigerants (Stemmler et al., 2004) but due to environmental problems, i.e. their high ozone depletion potential (ODP), their global warming potential (GWP) and their long atmospheric life times (ALT) (Molina and Rowland 1974, Rowland 1994, UNEP 1987, 1998, UNFCCC 1997) industry looked for substitutes; hydrochlorofluorocarbons (HCFC) and hydrofluorocarbons (HFC) became the first-generation alternatives (Haymann and Derwent, 1997). However, the latter still are not fully satisfactory; therefore, the search continues (Stemmler et al., 2004; UNFCCC 1997). Some of the newly proposed substances are chlorine-free fluorinated ethers, hydrocarbons (HC), alcohols, amines, and mixtures thereof (Sekiyaa and Misaki, 2000; Swaminathan, 2005a, b; Bivens, 1998). Prior to commencement of large-scale production, industry and regulatory agencies assess potential substitutes in respect ODP, GWP, ALT, toxicity, insulating ability, flammability, physical and chemical stability, solubility, cost, and other aspects of technical applicability (Sumantran et al., 1999; Swaminathan and Visco, 2005a, b;WMO 1991, 1992, 1995).

Normally an assessment implies that a decision is made based upon ranking of the substances under consideration (Lerche et al., 2002; Brüggemann, 1999); for this purpose, the Hasse diagram technique (HDT) (Brüggemann et al., 1993, 1994, 2001) is one of the most general and least subjective procedures (Lerche et al., 2002). In the present work, the HDT is applied to a set of 40 refrigerants taking ODP, GWP and ALT into consideration. After ranking, the refrigerants are divided into 13 subsets, and their order relations are studied by calculation of the two measures of comparability and incomparability, the dominance and the separability degrees.

[1] Environmental Chemistry and Ecotoxicology, University Bayreuth, Germany
email: guillermo.restrepo@uni-bayreuth.de, Internet: www.uni-bayreuth.de/departments/umweltchemie/
[2] Laboratorio de Química Teórica, Universidad de Pamplona, Colombia
[3] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

2. Hasse diagram technique, dominance and separability degrees

2.1 Hasse Diagram Technique (HDT)

Important concepts of the HDT are illustrated in the following example; a detailed description of the technique is found in Brüggemann et al. (1994, 2001). Let us assume set P contains 6 substances: $P = \{a, b, c, d, e, f\}$ of which each is described by three properties: q_1, q_2 and q_3 (Table 1). The HDT permits to compare any two chemicals considering simultaneously all their properties. A substance x is ranked higher than another y ($y \leq x$) when all its properties are higher than those of y, as is the case for $c \leq e$ (Table 1, Figure 1); in this case it is said that the two substances are "comparable". If x, y and z are substances and if $x \leq y$ and $y \leq z$, then $x \leq z$; for example $c \leq e$, $d \leq c$, then $d \leq e$ (Figure 1). If the property q_i of x is higher than q_j for y and the value of the property q_j for x is lower than q_j for y, then x and y are said to be "incomparable" ($x \parallel y$), for instance $e \parallel f$ (Table 1, Figure 1). Two substances are in a "cover-relation" if they are comparable and if there is no third one in between; all the pairs of substances with cover-relations are graphically represented in a Hasse diagram (Figure 1) drawn and analysed with the software WHASSE (Brüggemann et al., 1995). In general, the HDT permits to study order relations among the elements of a set by analysing the partially ordered set given by the couple (P, $\leq$), where $\leq$ is the order relation discussed above.

Table 1:
Properties q_1, q_2 and q_3 of the chemicals a, b, c, d, e and f.

	q_1	q_2	q_3
a	0	5	1
b	1	1	1
c	3	6	4
d	2	4	1
e	4	9	7
f	5	7	7

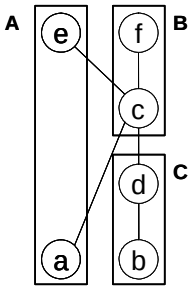


Figure 1: Hasse diagram of the substances in Table 1; the three boxes represent three subsets of chemicals (see text).

If the chemicals are described by properties whose values increase with the extent of their adverse impact, as is the case of ODP, GWP and ALT, then these substances located at the top of the diagram will be the "most hazardous" ones; in Figure 1 these substances correspond to *e* and *f*; in contrast, the chemicals found at the bottom (*a* and *b*) will represent the "least hazardous" ones.

Normally, a Hasse diagram is interpreted by analysing the order relations of single elements of *P*, but in some cases the order relations among different subsets is also of interest, for example among classes of similar chemicals. For doing this, two measures have been developed (Restrepo et al., 2007a, b, c), one indicating the extent to which members of one subset hold higher values in their descriptors than the members of other subsets, called the dominance degree (Restrepo et al. 2006a, 2007a, b, c); the second measure quantifies the number of incomparabilities among the members of two compound subsets and is called the separability degree (Restrepo et al., 2007c).

2.2 Dominance and separability degrees

Given a Hasse diagram of $(P, \leq)$ and two disjoint subsets $P_1, P_2 \subset P$, the dominance degree between P_1 and P_2 is given by Dom $(P_1, P_2) = N_R / N_T$, where $N_R = |\{(x, y), x \in P_1, y \in P_2$ and $y \leq x\}|$ and $N_T = |P_1| \cdot |P_2|$; $|X|$ means the cardinality or number of elements in a set *X*. The separability degree between P_1 and P_2 is given by Sep $(P_1, P_2) = N_I / N_T$, where $N_I = |\{(x, y), x \in P_1, y \in P_2$ and $y \| x\}|$.

Dom (P_1, P_2) and Sep (P_1, P_2) range from 0 to 1; Dom $(P_1, P_2) = 1$ means that all elements in P_1 have descriptor values higher than the ones of the elements of P_2; in this case it is said that P_1 completely dominates P_2 (Restrepo et al. 2007c). Dom $(P_1, P_2) = 0$ means that for no element *x* of P_1 and *y* of P_2 the relation $y \leq x$ holds; in this case P_1 does not dominate P_2. Because of the antisymmetry of $\leq$ (Trotter 1992), Dom (P_1, P_2) is not necessarily equal to Dom (P_2, P_1) (Restrepo et al. 2007b, c). Furthermore, Sep $(P_2, P_1) = 1$ means that all possible relations between P_1 and P_2 are incomparabilities; Sep $(P_1, P_2) = 0$ means that there are no incomparabilities between P_1 and P_2, and therefore all their relations are ruled by $\leq$. Additionally, it was proved that Dom (P_1, P_2) + Dom (P_2, P_1) + Sep $(P_1, P_2) = 1$ (Restrepo et al. 2007c). The values of dominance and separability degrees for the three subsets shown in Figure 1 are the following: Dom $(A, B) = 1 / 4 = 0.25$, Dom $(B, A) = 2 / 4 = 0.5$, Sep $(A, B) = 1 / 4 = 0.25$; Dom $(B, C) = 4 / 4 = 1$, Dom $(C, B) = 0 / 4 = 0$, Sep $(B, C) = 0 / 4 = 0$; and Dom $(A, C) = 2 / 4 = 0.5$, Dom $(C, A) = 0 / 4 = 0$, Sep $(A, C) = 2 / 4 = 0.5$.

The set *P* of 40 refrigerants (Table 2) was divided into the following 13 subsets: CFC, HFC, HCFC, HC, di(fluoroalkyl)ethers (DFAE), alkylfluoroalkylethers (AFAE), chloromethanes (CM), and the single-compound subsets trifluoroiodomethane (FIM), octafluorocyclobutane (PFC), carbon dioxide (CO_2), bromochlorodifluorobutane (BCF), dimethyl ether (DME) and ammonia (NH_3). These subsets were formed taking into consideration the common classification of the refrigerants into different chemical families. Each refrigerant is represented by its ALT, ODP and GWP (Restrepo et al. 2007b).

Table 2:
Refrigerants included in this study, their labels, chemical subsets, molecular formulae and non-proprietory names

Label	Subset	Molecular formula	Non-proprietory name	Label	Subset	Molecular formula	Non-proprietory name
1	CFC	CCl_3F	R11	**21**	CO_2	CO_2	R744
2	CFC	CCl_2F_2	R12	**22**	BCF	$CBrClF_2$	R12B1
3	HCFC	$CHClF_2$	R22	**23**	PFC	C_4F_8	RC318
4	HCFC	$C_2HCl_2F_3$	R123	**24**	HFC	C_3HF_7	R227ea
5	HCFC	C_2HClF_4	R124	**25**	AFAE	$C_4H_3F_7O$	HFE-7000
6	HCFC	$C_2H_3Cl_2F$	R141b	**26**	AFAE	$C_5H_3F_9O$	HFE-7100
7	HCFC	$C_2H_3ClF_2$	R142b	**27**	AFAE	$C_6H_5F_9O$	HFE-7200/ HFE-569mccc
8	HFC	CHF_3	R23	**28**	AFAE	$C_9H_5F_{15}O$	HFE-7500
9	HFC	CH_2F_2	R32	**29**	DFAE	C_2HF_5O	HFE-125
10	HFC	C_2HF_5	R125	**30**	DFAE	$C_2H_2F_4O$	HFE-134
11	HFC	$C_2H_2F_4$	R134a	**31**	CM	CH_2Cl_2	R30
12	HFC	$C_2H_3F_3$	R143a	**32**	CM	CH_3Cl	R40
13	HFC	$C_2H_4F_2$	R152a	**33**	CFC	$C_2Cl_3F_3$	R113
14	HFC	$C_3H_3F_5$	R245fa	**34**	HCFC	$CHCl_2F$	R21
15	HFC	$C_3H_2F_6$	R236fa	**35**	CFC	$C_2Cl_2F_4$	R114
16	HC	C_3H_8	R290	**36**	FIM	CF_3I	R13I1
17	HC	C_4H_{10}	R600	**37**	DME	C_2H_6O	
18	HC	C_4H_{10}	R600a	**38**	NH_3	NH_3	R717
19	HC	C_5H_{12}	R601	**39**	AFAE	$C_2H_3F_3O$	HFE-143
20	HC	C_3H_6	R1270	**40**	AFAE	$C_3H_3F_5O$	HFE-245

3. Results and discussion

The Hasse diagram of the considered refrigerants is depicted in Figure 2 with the marked boxes representing the 13 subsets mentioned in Table 2.

Since high values of ALT, ODP and GWP indicate a refrigerant with adverse effects on the environment, the most problematic substances are 1, 2, 8, 22, 23, 29, 33 and 35, while 19 and 20 are considered as benign ones. Problematic substances belong to DFAE, CFC, BCF, HFC and PFC subsets, and the benign ones to the HC subset. It is important to note the identification of 29, a difluoroalkyl ether, as hazardous refrigerant although this chemical has been introduced as potential replacement of CFC, HFC and HCFC (Tai, 2005), while 31, although a chlorocarbon, is relatively benign.

From the total of $13 \times 13 = 169$ ordered pairs (P_i, P_j), dominance and separability degrees are defined for 156 (169 minus 13 self comparisons $P_i = P_j$). Hence, for each of the 78 non-ordered pairs $\{P_i, P_j\}$, Dom (P_i, P_j), Dom (P_j, P_i) and Sep (P_i, P_j) were calculated (Table 3).

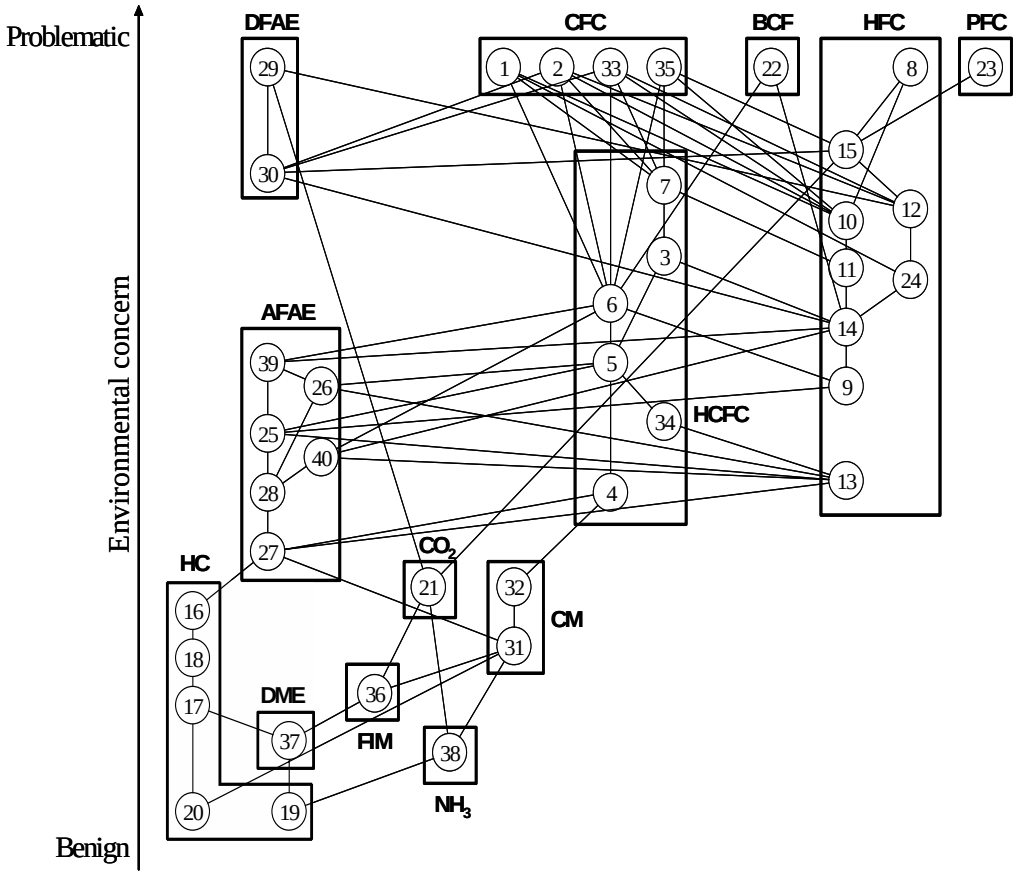


Figure 2: Hasse Diagram of 40 refrigerants and its subsets shown as boxes. CFC: chlorofluorocarbons, HFC: hydrofluorocarbons, HCFC: hydrochlorofluorocarbons, HC: hydrocarbons, DFAE: di(fluoroalkyl)ethers, AFAE: alkylfluoroalkylethers, CM: chloromethanes, FIM: trifluoroiodomethane, PFC: octafluorocyclobutane, BCF: bromochlorodifluorobutane, DME: dimethyl ether, CO_2 and NH_3.

Table 3:
Dominance and separability degrees for the 13 subsets shown in Figure 2

$\{P_i, P_j\}$	Dom (P_i, P_j)	Dom (P_j, P_i)	Sep (P_i, P_j)
{DFAE, AFAE}, {DFAE, HC}, {DFAE, FIM}, {DFAE, DME}, {DFAE, NH_3}, {BCF, AFAE}, {BCF, CM}, {BCF, HC}, {BCF, FIM}, {BCF, DME}, {BCF, NH_3}, {CFC, HCFC}, {CFC, AFAE}, {CFC, CM}, {CFC, HC}, {CFC, FIM}, {CFC, DME}, {CFC, NH_3}, {PFC, AFAE}, {PFC, CO_2}, {PFC, HC}, {PFC, FIM}, {PFC, DME}, {PFC, NH_3}, {HCFC, HC}, {HCFC, FIM}, {HCFC, DME}, {HCFC, NH_3} {HFC, HC}, {HFC, FIM}, {HFC, DME}, {HFC, NH_3}, {AFAE, HC}, {AFAE, FIM}, {AFAE, DME}, {AFAE, NH_3}, {CM, FIM}, {CM, DME}, {CM, NH_3}, {CO_2, FIM}, {CO_2, DME}, {CO_2, NH_3}, {FIM,DME}	1	0	0
{HCFC, CM}	0.92	0	0.08
{HFC, AFAE}	0.85	0.07	0.08
{CFC, HFC}	0.78	0	0.22
{HCFC, AFAE}	0.73	0	0.27
{BCF, HCFC}	0.67	0	0.33
{PFC, HFC}	0.63	0	0.37
{HC, DME}	0.6	0.2	0.2
{DFAE, CM}, {DFAE, CO_2}, {PFC, CM}, {HFC, CM}, {AFAE, CM}, {PFC, DFAE}	0.5	0	0.5
{DFAE, HFC}	0.44	0.11	0.45
{CM, HC}	0.4	0	0.6
{CFC, DFAE}	0.38	0	0.62
{BCF, HFC}	0.33	0	0.67
{CFC, CO_2}	0.25	0	0.75
{HFC, CO_2}	0.22	0	0.78
{HCFC, HFC}, {CO_2, HC}, {FIM, HC}, {NH_3, HC}	0.2	0	0.8
{DFAE, BCF}, {DFAE, HCFC}, {BCF, CFC}, {BCF, PFC}, {BCF, CO_2}, {CFC, PFC}, {PFC, HCFC}, {HCFC, CO_2}, {AFAE, CO_2}, {CM, CO_2}, {FIM, NH_3}, {DME, NH_3}	0	0	1

According to Table 3, for 55% of all pairs $\{P_i, P_j\}$ there is complete dominance of one subset over another, they are not separable (Sep (P_i, P_j) = 0). This indicates a high degree of comparability among the studied subsets, implying the possibility of finding order relations among the majority of subsets. Accor-

ding to a previous investigation (Restrepo et al., 2007b), one of the subsets dominating the majority of the subsets is CFC (Table 3) with high values of dominance over the other subsets.

When none of the compared subsets dominates any others, these are completely separated and therefore all their elements are incomparable (Restrepo et al., 2007c); in this case the dominance degree drops to 0 and the separability degree grows to 1. One example is the pair {CFC, PFC} (Figure 2). In this study, 15% of the pairs have this distribution for the three calculated parameters (Table 3); this means that less than a sixth of the relations among subsets do not follow an order relation, thereby it is not possible to find a most hazardous susbset for these cases because of their mutual incomparability. A strategy for looking for comparabilities in these cases is to prioritise the properties of the chemicals in order to aggregate them into a new superdescriptor which yields a linear order of the original Hasse diagram. The application of this strategy to the refrigerants will be published in a forthcoming paper.

The remaining pairs of Table 3, 30% of the total, are representing intermediate situations where one subset partially dominates the other one, both being partially separable. One example is { DFAE, HFC }, where 15% of their possible relations are of the kind $y \leq x$, with $x \in$ HFC and $y \in$ DFAE; 40% hold $x \leq y$, and 45% are incomparabilities between x and y.

The fact that more than half of the refrigerant subsets dominate others can be regarded as evidence of the relationship between the ranking of single compounds and the ranking of their families. This also suggests a relationship between the criteria for selecting members of each class and their ranking. Since the criteria of forming the subsets is the common chemical classification, i.e. chlorofluorocarbons, hydrofluorocarbons, and so on, it is interesting to form the classes by unsupervised classification, such as hierarchical cluster analysis using molecular descriptors for describing the refrigerants. It will be interesting to compare the resulting ranking of subsets with the ones found in the current work.

Acknowledgements

This study was financially supported by the Bavarian Environmental Agency. G. Restrepo thanks the Universidad de Pamplona and COLCIENCIAS in Colombia for a grant received during this research.

Bibliography

Bivens, D.B., Minor, B. (1998): Fluoroethers and other next generation fluids. Int. J. Refrig. 21: 567 - 576.

Brüggemann, R., Münzer, B. (1993): A graph-theoretical tool for priority setting of chemicals. Chemosphere, 27: 1729-1736.

Brüggemann, R., Münzer, B., Halfon, E. (1994): An algebraic/graphical tool to compare ecosystems with respect to their pollution - the German river “Elbe” as an example - I: Hasse-diagrams. Chemosphere, 28: 863-872.

Brüggemann, R., Halfon, E., Bücherl, C. (1995): Theoretical base of the program “Hasse”, GSF–Bericht 20/95, Neuherberg. (WHASSE available from R. B.).

Brüggemann, R., Bücherl, C., Pudenz, S., Steinberg, C.E.W. (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.E.W. (2001): Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests. J. Chem. Inf. Comput. Sci., 41: 918-925.

Haymann, G.D., Derwent, R.G. (1997): Atmospheric Chemical Reactivity and Ozone-Forming Potentials of Potential CFC Replacements. Environ. Sci. Technol., 31: 327-336.

Lerche, D., Brüggemann, R., Sørensen, P., Carlsen, L., Nielsen, O. J. (2002): A Comparison of Partial Order Technique with Three Methods of Multi-Criteria Analysis for Ranking of Chemical Substances. J. Chem. Inf. Comput. Sci., 42: 1086-1098.

Molina, M.J., Rowland, F.S. (1974): Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone. Nature, 249: 810-812.

Restrepo, G., Brüggemann, R. (2006): Modelling the Fate of Alkanes in Rivers. In Recent Progress in Computational Sciences and Engineering; Simos, T.; Maroulis, G., Eds.; VSP: Leiden, 1386-1389.

Restrepo, G., Brüggemann, R. (2007a): Partially ordered sets in the analysis of alkanes fate in rivers. Croat. Chem. Acta. In press.

Restrepo, G., Weckert, M., Brüggemann, R., Gerstmann, S., Frank, H. (2007b): Order relationships between sets of refrigerants. Submitted to Environ. Sci. Technol.

Restrepo, G., Brüggemann, R. (2007c): Dominance and separability in posets, their application to isoprotonic-isoelectronic species. Submitted to J. Math. Chem.

Rowland, F.S., Molina, M.J. (1994): Ozone depletion: 20 years after the alarm. Chem. Eng. News, 8-13.

Sekiyaa, A., Misaki, S. (2000): The potential of hydrofluoroethers to replace CFCs, HCFCs and PFCs. J. Fluorine Chem., 101: 215-221.

Stemmler, K., O'Doherty, S., Buchmann, B., Reimann, S. (2004): Emissions of the Refrigerants HFC-134a, HCFC-22, and CFC-12 from Road Traffic: Results from a Tunnel Study (Gubrist Tunnel, Switzerland). Environ. Sci. Technol., 38: 1998-2004.

Sumantran, V., Khalighi, B., Saka, K., Fischer, S. (1999): An Assessment of Alternative Refrigerants for Automotive Applications based on Environmental Impact. SAE International Automotive Alternate Refrigerant Systems Symposium. Scottsoale, http://www.sae.org/altrefrigerant/presentations/present 99.htm

Swaminathan, S., Visco, D.P. Jr. (2005a): Thermodynamic Modeling of Refrigerants Using the Statistical Associating Fluid Theory with Variable Range. 1. Pure Components. Ind. Eng. Chem. Res., 44: 4798-4805.

Swaminathan, S.; Visco, D.P. Jr. (2005b): Thermodynamic Modeling of Refrigerants Using the Statistical Associating Fluid Theory with Variable Range. 2. Applications to Binary Mixtures. Ind. Eng. Chem. Res., 44: 4806-4814.

Tsai, W.T. (2005): Environmental risk assessment of hydrofluoroethers (HFEs). J. Hazard. Material., 119: 69-78.

Trotter, W.T. (1992): Combinatorics and Partially Ordered Sets Dimension Theory; The Johns Hopkins University Press: Baltimore.

UNEP. (1987): Montreal Protocol on Substances that Deplete the Ozone Layer; United Nations Environment Programme: Nairobi, Kenya.

UNEP. (1998): Montreal Protocol on Substance that Deplete Ozone Layer; United Nations Environment Programme: Montreal, Canada.

UNFCCC. (1997): Kyoto Protocol to the United Nations Framework Convention on Climate Change, United Nations Framework Convention on Climate Change.

WMO. (World Meteorological Organization) (1991): Scientific Assessment of Stratospheric Ozone: 1989; Volume II, Appendix AFEAS Report; Global Ozone Research and Monitoring Project Report 20, Geneva, Switzerland.

WMO. (World Meteorological Organization) (1992): Scientific Assessment of Stratospheric Ozone: 1991; Global Ozone Research and Monitoring Project Report 25, Geneva, Switzerland.

WMO. (World Meteorological Organization) (1995): Scientific Assessment of Stratospheric Ozone: 1994; Global Ozone Research and Monitoring Project Report 37, Geneva, Switzerland.

Appendix E

Dominance and separability in posets, their application to isoelectronic species with equal total nuclear charge

Guillermo Restrepo[a,c], Rainer Brüggemann[b]

[a] Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia
[b] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[c] Environmental Chemistry and Ecotoxicology, University of Bayreuth, Bayreuth, Germany

Abstract

We developed the dominance and separability degrees as two new mathematical tools measuring the amount of comparabilities and incomparabilities among pairs of disjoint subposets in a parent poset and we have related them through a theorem. Their mathematical properties when these measures are constrained to be higher than 0.5 have been studied. We have shown that variations of dominance and separability degrees from values in the real interval (0.5, 1] permit to "tune" the level of detail on the comparabilites and incomparabilities among the subsets studied. The lack of transitivity of dominance and separability degrees is established, along with the special requirement, needed on the poset, to have a transitivity of these measures.

As a chemical application, the Hasse diagram of Born-Oppenheimer molecular total energies of the complete set of isoelectronic species with total nuclear charge 10 in their minimum energy configurations has been studied. We partition this set into ten subsets, each one containing all the species with the same number of nuclei. By the calculation of the dominance and separability degrees a relation between the number of atoms in any ensemble and the Born-Oppenheimer energies is established.

Key Words: partially ordered sets, posets, ranking, Born-Oppenheimer energies
MSC: 06A06, 62F07, 92E99

1. Introduction

Partially ordered sets (posets) are mathematical structures based on a comparison among the elements of a set [1]. If these elements are defined by their properties, then a poset is the result of the simultaneous comparison of the elements through their properties. Comparisons are usual in human activities, examples of their presence are these kinds of questions: which is the best kind of economic system?, which is the best university?, which is the best quantum-chemical level of calculation?, which is the most hazardous substance? In order to reply to these questions it is always necessary to make comparisons, and a poset is the mathematical structure behind them.

There are several posetic studies in different fields of knowledge [1,2]; in economy, for instance, Annoni recently ranked and classified a set of European countries according to their public level of satisfaction regarding different public services [3]; in ecology, Solomon has pointed out the posetic character of the abundance vectors used to define a community in diversity studies [4]; in thermodynamics and quantum mechanics it has been shown the relation of the Young diagrams lattice (a poset) with the mixing character [5-7] and the generation of wave functions satisfying the Pauli exclusion principle [8]. Particularly in chemistry, instances of posetic studies started at the end of 1960s with Ruch and his investigations into the algebraic description of chirality [9,10]; afterwards Randić and some others made important researches on chemical structure and its posetic description [11-15]. Another relevant and nowadays quite explored aspect of posets in chemistry was initiated by Halfon and Reggiani in 1986 when they ranked substances in environmental hazard studies [16]; this line of research has been deeply studied by Brüggemann and coworkers [1,17-29], who regularly organise workshops about posets in chemistry and environmental sciences. Some other instances of these mathematical structures in chemistry are found in references [1,28-36]. A more general statement on posets has been set up by Klein and Babić [35,36], who have pointed out that posets may be deeply related to experimental sciences through the measuring process. According to these authors, ambiguities resulting from measurements might be explained as the result of measuring elements, which in reality must be considered as incomparable. Hence, the measuring method may force the

incomparabilities to be comparable and, because of the different possibilities to do this [37,38], the outcomes may be different, therefore "ambiguous", in a systematically controlled way.

Once a poset is detected or constructed on a given set, its analysis permits to draw conclusions on the order relations among the elements considered, for instance maximal and minimal elements or order intervals, ideals or filters [39]. The study of these posetic features and their properties are an important and active research field of mathematics, mainly carried out in combinatorics. However, to our knowledge, there is little information on the study of posetic properties of subsets of elements belonging to a poset. That is to ask, if the original set is partitioned into different subsets, what is the behaviour of the order relations among these subsets? In a chemical framework this question can be exemplified as: given a poset of organic molecules, which are the order relations between alkanes-alkenes, alkanes-alkynes, alkynes-amines and so on? In this paper we deal with that question and we develop two measures, one related to the comparability and another one with the incomparability between a pair of subsets; the first measure is called the dominance degree and the second one the separability degree. We also describe their mathematical properties and their relation through a novel theorem. Finally, these measurements are applied to different subposets of a poset of isoelectronic chemical species with equal total nuclear charge.

2. Methodology

For the sake of clarity we introduce some terms useful for the understanding of the paper:

Definition 1. An ordered pair (P, R) is called a *structure* if R is a relation on the non-empty set P which is called a ground set and is here considered finite.

Definition 2. A binary relation $\leq$ on P is called a *partial order* if:
1. $x \in P \Rightarrow x \leq x$,
2. $x, y \in P, x \leq y$ and $y \leq x \Rightarrow x = y$,
3. $x, y, z \in P, x \leq y$ and $y \leq z \Rightarrow x \leq z$.
Then $\leq$ is respectively reflexive, antisymmetric and transitive on P. A ground set equipped with a partial order is called a *poset* (partially ordered set) and it is denoted as $(P, \leq)$.

Definition 3. Let P' be a subset of P, with the inherited order relation $\leq$, then $(P', \leq)$ is a subposet of $(P, \leq)$.

In some cases the fact of having $x \leq y$ and $y \leq x$ does not necessarily imply $x = y$. In those situations it is said that x and y are related by an equivalence relation different than equality, for instance a similarity relation [40] in which case $\leq$ is called a quasi order [19]. This may occur when the elements of P are described by means of their features. In the case with a quasi-order one may define an equivalence relation $\approx$ such that the equivalence class of $x \in P$ is $\{y : x \leq y \text{ and } y \leq x\}$. Hence, if one wants to order the elements of P according to $\leq$, it is possible to select a representative element of each equivalence class to perform the ordering, instead of considering all the elements in P, including equivalent ones. In that case the relation $\leq$ is not applied to the complete set P but to a set P'' of representative elements of each equivalence class. In order to avoid cumbersome notation, we keep calling P the reduced set P'' of representatives.

Definition 4. Two elements $x, y \in P$ are said to be *comparable* if either $x \leq y$ or $y \leq x$. We say that x *surpasses* y if $y \leq x$.

Definition 5. Given two elements $x, y \in P$, we say that y is *covered* by x, denoted $y \leq: x$, if $y \leq x$ and there is no $z \in P$ for which $y < z$ and $z < x$. If $y \leq: x$, it is said that x *covers* y.

The existence of $(P, \leq)$ does not guarantee the comparability between every pair $x, y \in P$. For those "incomparable" elements a new relation is defined.

Definition 6. For all $x, y \in P$, x and y are *incomparable* ($x \parallel y$) iff not $x \leq y$ and not $y \leq x$. The *incomparability* relation $\parallel$ is a binary relation on P fulfilling these properties:

1. $x \in P \Rightarrow$ not $x \parallel x$,
2. $x, y \in P, x \parallel y \Rightarrow y \parallel x$.
Then $\parallel$ is an irreflexive and symmetric relation on P.

Definition 7. Let $G_{\leq} = (P, E_{\leq})$ the *comparability graph* of $(P, \leq)$, where $E_{\leq}$ is the set of edges containing the comparable pairs in P.

Definition 8. Let $G_{\leq:} = (P, E_{\leq:})$ the *cover graph* of $(P, \leq)$, where $E_{\leq:}$ is the set of edges containing the cover pairs in P.

$G_{\leq:}$, as well as $G_{\leq}$, is an undirected graph which offers more information about comparabilities and incomparabilities if it is oriented taking advantage of the antisymmetry of $\leq$ [2].

Definition 9. Let $H = (P, d(E_{\leq:}))$ a directed graph of $(P, \leq)$ where $d(E_{\leq:})$ is the set of directed edges containing the cover pairs in P. H is called the *Hasse diagram* of $(P, \leq)$ if it is drawn in the Euclidean plane whose horizontal/vertical coordinate system requires that the vertical coordinate of $x \in P$ be larger than the one of $y \in P$ if $y \leq: x$.

Definition 10. Let $G_{\parallel} = (P, E_{\parallel})$ the *incomparability graph* of $(P, \leq)$, where $E_{\leq:}$ is the set of edges containing the incomparable pairs in P.

2.1. Order relations among subsets of a poset

There are two ways for studying the order relations among subsets of a poset $(P, \leq)$. The first one clusters the elements of P and defines pseudo-objects as centres of the clusters and finally analyses the resulting partial order on the set of pseudo-objects [20]. The second possibility considers all the order relations between members of different subsets of P, which can arise from external knowledge. For example, chemicals may be ordered due to a set of properties. There may still be information, which is not used for ordering the chemicals but which can be used to define subsets within the partially ordered set. This methodology and its properties are studied in this paper by defining two new structural parameters of $(P, \leq)$; one dealing with the $\leq$-relation between subsets, called dominance[1] degree, and another one studying the $\parallel$-relation, called separability degree. Since these two measures depend on the number of comparabilities and incomparabilities among the elements of any two subsets of a Hasse diagram, we introduce an indicator function useful for counting them.

Definition 11. Let $(P, \leq)$ be a poset with P_i and $P_j \subset P$ such that $P_i \cap P_j = \varnothing$. Then for all $x \in P_i, y \in P_j$ it is defined the *indicator function* $L_{xy}^{(i,j)}$ as follows:

$$L_{xy}^{(i,j)} = \begin{cases} 1 \text{ if } y \leq x \\ -1 \text{ if } x \leq y \\ 0 \text{ if } x \parallel y \end{cases} \qquad (1)$$

Whenever it holds a comparability between x and y, $L_{xy}^{(i,j)}$ assigns a value of 1 or –1, being 1 when x surpasses y and –1 when y surpasses x; $L_{xy}^{(i,j)}$ yields a value of zero when the pair is incomparable (recall that equivalences are excluded). This kind of indicator function is used in observational studies [42] and it is further described by Rosenbaum [42,43].

In order to have an account of the number of comparabilities ($y \leq x$ and $x \leq y$) and incomparabilities ($x \parallel y$) in P, the statistics $T_{j \leq i}$, $T_{i \leq j}$ and $T_{i \parallel j}$ are created.

[1] The concept of dominance developed in this paper is not directly related to the one of dominating set, which is as follows [41]: A *dominating set* is a set of vertices $D \subseteq V$ in a graph $G = (V, E)$ having the property that every vertex $v \in V - D$ is adjacent to at least one vertex in D.

Definition 12. Let P_i and P_j be two disjoint subsets with $x \in P_i$, $y \in P_j$ and respective cardinalities n_i and n_j. The statistics $T_{j \le i}$, $T_{i \le j}$ and $T_{i \| j}$ among all possible $n_i \cdot n_j$ relations are defined as:

$$\begin{aligned} &T_{j \le i} \text{ is a count of all } L_{xy}^{(i,j)} = 1, \\ &T_{i \le j} \text{ is a count of all } L_{xy}^{(i,j)} = -1, \\ &T_{i \| j} \text{ is a count of all } L_{xy}^{(i,j)} = 0. \end{aligned} \qquad (2)$$

In the following we introduce the dominance and separability degrees.

Definition 13. Let $(P, \le)$ be a poset with $P_i, P_j \subset P$ such that $P_i \cap P_j = \varnothing$ and $n_i = |P_i|$, $n_j = |P_j|$. Then for all $x \in P_i$ and $y \in P_j$, the *dominance degree* of P_i over P_j is given by

$$Dom(P_i, P_j) = \frac{T_{j \le i}}{n_i \cdot n_j} \qquad (3)$$

Hence, $Dom(P_i, P_j)$ counts the number of ordered pairs where an element of P_i surpasses an element of P_j and divides it by all the possible relations between P_i and P_j. Therefore, *Dom* yields a real value ranging from 0 to 1; $Dom(P_i, P_j) = 1$ means that all the elements in P_i surpass all those in P_j. In contrast, if $Dom(P_i, P_j) = 0$, it means that no element of P_i surpasses an element of P_j. Note that, because of the antisymmetry of $\le$ (Definition 2), $Dom(P_i, P_j)$ is not necessarily equal to $Dom(P_j, P_i)$ and the equality only occurs when the number of pairs $x \le y$ is equal to the number of pairs $y \le x$ (see Corollary 1).

The relations defined on P are of two types: comparabilities ($\le$) and incomparabilities ($\|$). Since $Dom(P_i, P_j)$ represents the fraction of relations of P_i over P_j such that $y \le x$ with $x \in P_i$ and $y \in P_j$, it is possible that the rest of the relations correspond to either comparabilities where the elements of P_j surpass those of P_i, or incomparabilities among them. Hence, given a value of $Dom(P_i, P_j)$ it is natural to ask for $Dom(P_j, P_i)$ and also for the proportion of incomparabilities. These incomparabilities may be gathered in a mathematical expression similar to and, as we show later in Theorem 1, related to dominance degree.

Definition 14. Given a poset $(P, \le)$ with $P_i, P_j \subset P$ such that $P_i \cap P_j = \varnothing$ and $n_i = |P_i|$, $n_j = |P_j|$, then for all $x \in P_i$, $y \in P_j$, the *separability degree* between P_i and P_j is given by

$$Sep(P_i, P_j) = \frac{T_{i \| j}}{n_i \cdot n_j} \qquad (4)$$

$Sep(P_i, P_j)$ is the result of the division of the number of incomparabilities between the elements of P_i and P_j and the number of order relations between P_i and P_j. Note that $Sep(P_i, P_j) = Sep(P_j, P_i)$ because of the symmetry of $\|$ (Definition 6). Separability degree takes values in the real interval [0, 1]; $Sep(P_i, P_j) = 1$ means that all the possible relations between P_i and P_j are incomparabilities; in contrast, a value of $Sep(P_i, P_j) = 0$ means that there are no incomparabilities between P_i and P_j, thereby all their relations are comparabilities and they are counted in $Dom(P_i, P_j)$ and $Dom(P_j, P_i)$. Hence, there is a mathematical relation between $Dom(P_i, P_j)$, $Dom(P_j, P_i)$ and $Sep(P_i, P_j)$, which is set up in Theorem 1. Before introducing this theorem and its consequences, we show an example of calculation of dominance and separability degrees.

Example 1. Let $P = \{a, b, c, d, e, f, g, h\}$, $P_1 = \{a, b, c, d\}$, $P_2 = \{e, f, g, h\}$ and the Hasse diagram depicted in Figure 1. In this case $Dom(P_1, P_2) = 8/16 = 0.5$, $Dom(P_2, P_1) = 4/16 = 0.25$ and $Sep(P_1, P_2) = 4/16 = 0.25$.

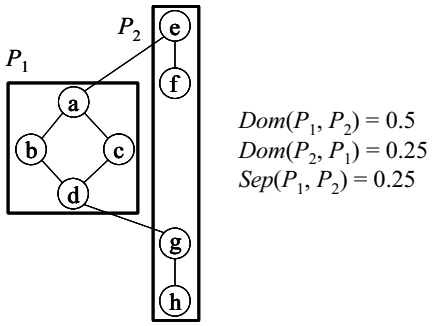


Figure 1. A Hasse diagram on the set $P = \{a, b, c, d, e, f, g, h\}$; two "boxed" subsets $P_1 = \{a, b, c, d\}$ and $P_2 = \{e, f, g, h\}$ with $n_1 = n_2 = 4$; and their respective dominance and separability degrees.

Note that the only requirement for calculating dominance and separability degrees is that P_i and P_j be disjoint subsets. This implies that their internal relations, $\leq$ and $\|$, are not necessary for building up P_i or P_j. In fact, they might be antichains, chains or mixtures of them and this fact does not address the membership to P_i or P_j.

Theorem 1. Let $(P, \leq)$ be a poset with $P_i, P_j \subset P$ such that $P_i \cap P_j = \varnothing$. The dominance (*Dom*) (Definition 13) and separability (*Sep*) (Definition 14) degrees for P_i and P_j satisfy $Dom(P_i, P_j) + Dom(P_j, P_i) + Sep(P_i, P_j) = 1$.

Proof. $Dom(P_i, P_j)$, $Dom(P_j, P_i)$ and $Sep(P_i, P_j)$ are defined on $P_i, P_j \subset P$, where $n_i = |P_i|$ and $n_j = |P_j|$. There are two relations defined on P, namely $\leq$ (Definition 2) and $\|$ (Definition 6), which are in turn defined on the subposets $(P_i, \leq)$ and $(P_j, \leq)$. The relation $\leq$ can be split into the relations $\leq_{ji}$ and $\leq_{ij}$, where $\leq_{ji} = \{(x, y) : x \in P_i, y \in P_j \text{ and } y \leq x\}$ and $\leq_{ij} = \{(x, y) : x \in P_i, y \in P_j \text{ and } x \leq y\}$. Then, the set $\{\leq_{ji}, \leq_{ij}\}$ is a partition of $\leq$ because $\leq = \leq_{ji} \cup \leq_{ij}$ and $\leq_{ji} \cap \leq_{ij} = \varnothing$. From this, and from Definition 6, follows that $\leq \cap \| = \varnothing$ and $\mathscr{R} = \leq \cup \|$, where $\mathscr{R} = \{(x, y) : x \in P_i, y \in P_j$ and either $y \leq x$ or $x \leq y$ or $x \| y\}$, which is the set of ordered pairs $(x, y) \in P_i \times P_j$ fulfilling the relations $\leq$ and $\|$. Then, $\{\leq_{ji}, \leq_{ij}, \|\}$ is a partition of $\mathscr{R}$ because $\leq_{ji} \cap \leq_{ij} \cap \| = \varnothing$ and $\mathscr{R} = \leq_{ji} \cup \leq_{ij} \cup \|$. Since $|\mathscr{R}| = n_i \cdot n_j$, then $|\leq_{ji}| + |\leq_{ij}| + |\|| = n_i \cdot n_j$, and, according to Definition 11, this is equivalent to $T_{j \leq i} + T_{i \leq j} + T_{i \| j} = n_i \cdot n_j$. Thus, from Definitions 13 and 14 it follows that $Dom(P_i, P_j) + Dom(P_j, P_i) + Sep(P_i, P_j) = 1$ □

The dominance degree, $Dom(P_i, P_j)$, is a measurement of the extent of comparability between any two disjoint subposets of P. In observational studies [42-44], whose goal is to measure the effect of a cause, for instance the effect of a medical treatment on patients, a set P of observations (patients) is divided into two subsets P_i (control) and P_j (treatment). When the observations are described by more than one outcome, then P may become a poset and the coherence of the cause-effect hypothesis is measured by the degree of dominance of one of the two considered subsets in the poset. This measurement is carried out by a statistic operating on the set of relations between the two compared subsets and it considers simultaneously comparabilities and incomparabilities. The statistic used in these studies [44] is:

$$\hat{\zeta}_C = \frac{\sum_{l=1}^{n_1}\sum_{m=1}^{n_2} L_{lm}^{(1,2)}}{n_1 \cdot n_2}, \text{ with } L_{lm}^{(1,2)} = \left\{ \begin{array}{l} 1 \text{ if } y_m \leq x_l \\ -1 \text{ if } x_l \leq y_m \\ 0 \text{ if } x_l \| y_m \end{array} \right\} \left\{ \begin{array}{c} l \in I_1, m \in I_2 \\ I_j \equiv \text{index set for } P_j \end{array} \right. \qquad (5)$$

Since $\hat{\zeta}_C$ operates over all possible values of $L_{lm}^{(1,2)}$, it does not distinguish between $L_{lm}^{(1,2)} = 1, -1$ or 0, thereby it does not differentiate between $Dom(P_i, P_j) = Dom(P_j, P_i)$, $Sep(P_i, P_j) = 1$; and $Dom(P_i, P_j) = Dom(P_j, P_i)$, $Sep(P_i, P_j) = 0$; yielding a value of zero for both cases. It can be also noted that $\hat{\zeta}_C =$

$Dom(P_i, P_j) - Dom(P_j, P_i)$. Then, the advantage of studying individually $Dom(P_i, P_j)$, $Dom(P_j, P_i)$ and $Sep(P_i, P_j)$ makes it possible to go into the details of the comparability and incomparability relations between P_i and P_j. Note that $\hat{\zeta}_C$ is equivalent to the statistic suggested by Rosenbaum [42,43].

In a work on observational studies developed by Gefeller and Pralle [44], it is stated that, what here it is defined as, $Dom(P_i, P_j)$ and $Dom(P_j, P_i)$ must fulfil $Dom(P_i, P_j) + Dom(P_j, P_i) \leq 1$. This inequality becomes an equality by adding the term of separability between P_i and P_j, as in the statement of Theorem 1.

Corollary 1. $Dom(P_i, P_j) = Dom(P_j, P_i)$ iff $T_{j\leq i} = T_{i\leq j}$.

This corollary states that P_i dominates P_j and P_j dominates P_i to the same extent only if the number of pairs where $y \leq x$ is equal to the number of pairs where $x \leq y$, having $x \in P_i$ and $y \in P_j$.

Since $Dom(P_i, P_j)$ depends on the number of comparabilities between P_i and P_j, where the elements of P_i surpass the ones of P_j, then $Dom(P_i, P_j)$ can be related to a matrix of comparabilities of this kind.

Definition 15. Given a Hasse diagram of $(P, \leq)$, P_i and $P_j \subset P$ holding $P_i \cap P_j = \varnothing$; then for all $x \in P_i, y \in P_j$ the *indicator function* $M_{xy}^{(i,j)}$ is defined as follows:

$$M_{xy}^{(i,j)} = \begin{cases} 1 \text{ if } y \leq x \\ 0 \text{ otherwise} \end{cases} \qquad (6)$$

Definition 16. Let $\mathbf{M}^{(i,j)} = \left[M_{xy}^{(i,j)}\right]_{n_i \times n_j}$ the matrix representing the $M_{xy}^{(i,j)}$-values of $(x, y) \in P_i \times P_j$, where $|P_i| = n_i$ and $|P_j| = n_j$.

This matrix can be regarded as an adjacency matrix describing the $\leq$-relations among the elements in P_i and those in P_j. Note that $\mathbf{M}^{(i,j)}$ is not a symmetric matrix because of the antisymmetry of $\leq$ (Definition 2). Since the statistic $T_{j\leq i}$ (Definition 12) can be derived from this matrix, then $Dom(P_i, P_j)$ (Definition 13) can also be related to $\mathbf{M}^{(i,j)}$.

Each couple of disjoint subsets in P can be described by an $\mathbf{M}$ matrix and it is possible to study the relation between subsets by the multiplication of these matrices.

Before describing the meaning of the standard matrix product of $\mathbf{M}$ matrices, we define the collection of subsets of P and the $\leq$-paths.

Definition 17. Let P be a non-empty set and $\mathscr{P}$ a collection of subsets of P. $\mathscr{P}$ partitions P iff:

1. $P = \cup_{P_i \in \mathscr{P}} P_i$,
2. If P_1 and $P_2 \in \mathscr{P}$, then $P_1 \cap P_2 = \varnothing$.

Definition 18. Let $P_i, P_k, \ldots, P_l, P_j \in \mathscr{P}$ and $r \in P_i$, $s \in P_k, \ldots, t \in P_l$, $u \in P_j$. Any sequence of comparabilities $r \leq s \leq \cdots \leq t \leq u$ is called a $\leq$-path.

Proposition 1. Let $(P, \leq)$ be a poset; $S \subseteq \mathscr{P}$; $P_i \in S$, $|P_i| = n_i$; and $G_{\leq}(S)$ the cover graph of $(S, \leq)$. Let $\mathbf{M}^{(i,j)}$ be the associated matrix to any pair $P_i, P_j \in S$. Let be the standard matrix product of an arbitrary number of $\mathbf{M}$ matrices yielding a matrix $\mathbf{C}$, defined as follows: $\mathbf{C}^{(i,j)} = \mathbf{M}^{(i,k)} \cdot \ldots \cdot \mathbf{M}^{(l,j)} = \left[C_{ru}^{(i,j)}\right]_{n_i \times n_j}$, with i, j indicating $P_i, P_j \in S$ and $r \in P_i, u \in P_j$.

If these conditions hold, then each $C_{ru}^{(i,j)}$ represents the number of $\leq$-paths $r \leq s \leq \cdots \leq t \leq u$ between $r \in P_i$ and $u \in P_j$ in $G_{\leq}(S)$ such that these $\leq$-paths pass through at least one element of each subset $P_i, P_k, \ldots, P_l, P_j \in S$ considered in the matrix product.

Proof. We shall prove that given a standard matrix product of **M** matrices, the elements of the final matrix **C** indicate the number of $\leq$-paths passing through at least one element of each subset in the matrix product.

Let us start assuming, without loss of generality, $P = \{a, b, c, d, e, f, g, h\}$ and $\mathscr{P} = \{P_1, P_2, P_3, P_4\}$ with $P_1 = \{a, b\}$, $P_2 = \{c, d\}$, $P_3 = \{e, f\}$ and $P_4 = \{g, h\}$. Let us consider $S = \mathscr{P}$ and the following arbitrary **M** matrices.

$$\mathbf{M}^{(1,2)} = \begin{matrix} \\ a \\ b \end{matrix}\!\!\begin{matrix} \begin{matrix} c & \quad d \end{matrix} \\ \begin{bmatrix} M_{ac}^{(1,2)} & M_{ad}^{(1,2)} \\ M_{bc}^{(1,2)} & M_{bd}^{(1,2)} \end{bmatrix} \end{matrix} \quad \mathbf{M}^{(2,3)} = \begin{matrix} \\ c \\ d \end{matrix}\!\!\begin{matrix} \begin{matrix} e & \quad f \end{matrix} \\ \begin{bmatrix} M_{ce}^{(2,3)} & M_{cf}^{(2,3)} \\ M_{de}^{(2,3)} & M_{df}^{(2,3)} \end{bmatrix} \end{matrix} \quad \mathbf{M}^{(3,4)} = \begin{matrix} \\ e \\ f \end{matrix}\!\!\begin{matrix} \begin{matrix} g & \quad h \end{matrix} \\ \begin{bmatrix} M_{eg}^{(3,4)} & M_{eh}^{(3,4)} \\ M_{fg}^{(3,4)} & M_{fh}^{(3,4)} \end{bmatrix} \end{matrix}$$

and their product $\mathbf{C}^{(1,4)} = \mathbf{M}^{(1,2)} \cdot \mathbf{M}^{(2,3)} \cdot \mathbf{M}^{(3,4)}$

$$\mathbf{C}^{(1,4)} = \begin{matrix} \\ a \\ b \end{matrix}\!\!\begin{matrix} \begin{matrix} g & \quad h \end{matrix} \\ \begin{bmatrix} C_{ag}^{(1,4)} & C_{ah}^{(1,4)} \\ C_{bg}^{(1,4)} & C_{bh}^{(1,4)} \end{bmatrix} \end{matrix} \text{ with}$$

$$C_{ag}^{(1,4)} = M_{ac}^{(1,2)}M_{ce}^{(2,3)}M_{eg}^{(3,4)} + M_{ad}^{(1,2)}M_{de}^{(2,3)}M_{eg}^{(3,4)} + M_{ac}^{(1,2)}M_{cf}^{(2,3)}M_{fg}^{(3,4)} + M_{ad}^{(1,2)}M_{df}^{(2,3)}M_{fg}^{(3,4)}$$
$$C_{ah}^{(1,4)} = M_{ac}^{(1,2)}M_{ce}^{(2,3)}M_{eh}^{(3,4)} + M_{ad}^{(1,2)}M_{de}^{(2,3)}M_{eh}^{(3,4)} + M_{ac}^{(1,2)}M_{cf}^{(2,3)}M_{fh}^{(3,4)} + M_{ad}^{(1,2)}M_{df}^{(2,3)}M_{fh}^{(3,4)}$$
$$C_{bg}^{(1,4)} = M_{bc}^{(1,2)}M_{ce}^{(2,3)}M_{eg}^{(3,4)} + M_{bd}^{(1,2)}M_{de}^{(2,3)}M_{eg}^{(3,4)} + M_{bc}^{(1,2)}M_{cf}^{(2,3)}M_{fg}^{(3,4)} + M_{bd}^{(1,2)}M_{df}^{(2,3)}M_{fg}^{(3,4)}$$
$$C_{bh}^{(1,4)} = M_{bc}^{(1,2)}M_{ce}^{(2,3)}M_{eh}^{(3,4)} + M_{bd}^{(1,2)}M_{de}^{(2,3)}M_{eh}^{(3,4)} + M_{bc}^{(1,2)}M_{cf}^{(2,3)}M_{fh}^{(3,4)} + M_{bd}^{(1,2)}M_{df}^{(2,3)}M_{fh}^{(3,4)}$$

Now, each element $C_{ag}^{(1,4)}$, $C_{ah}^{(1,4)}$, $C_{bg}^{(1,4)}$, $C_{bh}^{(1,4)}$ of the matrix $\mathbf{C}^{(1,4)}$ is of the form $\sum M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$, with $r \in P_1$, $s \in P_2$, $t \in P_3$ and $u \in P_4$.

According to Definition 15, for any $x \in P_i$ and $y \in P_j$ with $P_i \cap P_j = \varnothing$, $M_{xy}^{(i,j)}$ is equal to 1 if $y \leq x$ and 0 otherwise, therefore each term $M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$ in $\sum M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$ is either equal to 1 or to 0. It is 1 if all M_{xy} have the value 1; that is, if $r \geq s$, $s \geq t$ and $t \geq u$. It is 0 if at least one $M_{xy} = 0$; that is, if at least one of these incomparabilities $r \parallel s$, $s \parallel t$ or $t \parallel u$ holds. Hence, each term $M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$ in $\sum M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$ indicates if it is possible to find a $\leq$-path of the form $r \geq s \geq t \geq u$ through the particular elements *r*, *s*, *t* and *u*. Consequently, $\sum M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$ indicates the number of such paths between $r \in P_1$ and $u \in P_4$.

Now, we have to prove that these paths pass through at least one element of the subsets P_i considered in the matrix product.

Because each $M_{xy}^{(i,j)}$ always considers only the element *x* of P_i and only the element *y* of P_j, the finding of $M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)} = 1$ guarantees the existence of a path $r \geq s \geq t \geq u$ passing exclusively through the elements *r*, *s*, *t* and *u* in the order given by the product. Hence, each element of the matrix **C**, given by $C_{ru}^{(1,4)} = \sum M_{rs}^{(1,2)}M_{st}^{(2,3)}M_{tu}^{(3,4)}$, accounts for all the theoretical paths between *r* and *u* passing through each single element of P_s and P_t. Thereby, if $C_{ru}^{(1,4)} = 0$, then none of the theoretical paths is realised. In contrast, if $C_{ru}^{(1,4)} \geq 1$, the inequality is met if more than one of the theoretical paths between *r* and *u* passing through single elements of P_s and P_t is realised. $C_{ru}^{(1,4)} = 1$ if at least one of the theoretical paths including single elements of P_s and P_t exists.

In conclusion, the $\leq$-paths pass through at least one element of the subsets P_s and P_t, which are considered in the matrix product.

Because of the properties of the standard matrix product regarding the generality of the elements of a **C** matrix obtained by the finite product of arbitrary **M** matrices, it is possible to extend this result to any finite set *P* partitioned into different disjoint subsets gathered in $\mathscr{P}$ with a subset $S \subseteq$

$\mathcal{P}$, $S = \{P_1, P_2, P_3, \ldots, P_{n-1}, P_n\}$ for which arbitrary matrices are defined in such a way that $\mathbf{C}^{(i,j)} = \mathbf{M}^{(i,k)} \cdot \ldots \cdot \mathbf{M}^{(l,j)}$, for any $P_i, P_k, \ldots, P_l, P_j \in S$. The elements of $\mathbf{C}^{(i,j)}$ are of the form $\sum M_{rs} \cdots M_{tu}$ with $r \in P_i$, $s \in P_k$, $t \in P_l$ and $u \in P_j$. Therefore, each element of $\mathbf{C}^{(i,j)}$ indicates the number of $\leq$-paths passing through at least one element of each subset in the matrix product□

Example 2. Let be the Hasse diagram depicted in Figure 1 and the new subsets $P_1 = \{a, e\}$, $P_2 = \{c, d\}$ and $P_3 = \{g\}$. In this case there are $3(3-1) = 6$ possible matrices **M**:

$$\mathbf{M}^{(1,2)} = \begin{array}{c} \\ a \\ e \end{array}\!\!\begin{array}{c} \begin{array}{cc} c & d \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array} \qquad \mathbf{M}^{(1,3)} = \begin{array}{c} \\ a \\ e \end{array}\!\!\begin{array}{c} \begin{array}{c} g \end{array} \\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{array} \quad \mathbf{M}^{(2,1)} = \begin{array}{c} \\ c \\ d \end{array}\!\!\begin{array}{c} \begin{array}{cc} a & e \end{array} \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{M}^{(2,3)} = \begin{array}{c} \\ c \\ d \end{array}\!\!\begin{array}{c} \begin{array}{c} g \end{array} \\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{array} \quad \mathbf{M}^{(3,1)} = \begin{array}{c} \\ g \end{array}\!\!\begin{array}{c} \begin{array}{cc} a & e \end{array} \\ \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{M}^{(3,2)} = \begin{array}{c} \\ g \end{array}\!\!\begin{array}{c} \begin{array}{cc} c & d \end{array} \\ \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array}$$

If we calculate $\mathbf{M}^{(1,2)} \cdot \mathbf{M}^{(2,3)}$, for instance, the matrix $\mathbf{C}^{(1,3)}$ obtained is:

$$\mathbf{C}^{(1,3)} = \begin{array}{c} \\ a \\ e \end{array}\!\!\begin{array}{c} \begin{array}{c} g \end{array} \\ \begin{bmatrix} 2 \\ 2 \end{bmatrix} \end{array}.$$

The entries of $\mathbf{C}^{(1,3)}$ indicate that there are two $\leq$-paths ($C_{ag}^{(1,3)} = 2$) of the form $g \leq x \leq a$ with $a \in P_1$, $x \in P_2$ and $g \in P_3$; these two $\leq$-paths are $g \leq c \leq a$ and $g \leq d \leq a$. For $C_{eg}^{(1,3)} = 2$ the corresponding $\leq$-paths are $g \leq c \leq e$ and $g \leq d \leq e$. The paths can be easily visualised if the comparability graph of the poset is drawn.

If we consider another product, for example $\mathbf{M}^{(1,3)} \cdot \mathbf{M}^{(3,2)}$, then $\mathbf{C}^{(1,2)} = [0]_{2\times 2}$ is obtained, which means that there are no possible $\leq$-paths of the form $x \leq g \leq a$, with $x \in P_2$, $g \in P_3$ and $a \in P_1$, between an element of P_1 and an element of P_2 passing through an element of P_3.

Example 3. Let $P_1 = \{a, b\}$, $P_2 = \{c, d\}$, $P_3 = \{e\}$, $P_4 = \{f, g, h\}$, $P_5 = \{i, j\}$ and the Hasse diagram shown in Figure 2. In this case there are 20 **M** matrices; we show 4 of them and their associated $\mathbf{C}^{(1,5)}$ matrix.

$$\mathbf{M}^{(1,2)} = \begin{array}{c} \\ a \\ b \end{array}\!\!\begin{array}{c} \begin{array}{cc} c & d \end{array} \\ \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{M}^{(2,3)} = \begin{array}{c} \\ c \\ d \end{array}\!\!\begin{array}{c} \begin{array}{c} e \end{array} \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{array} \quad \mathbf{M}^{(3,4)} = \begin{array}{c} \\ e \end{array}\!\!\begin{array}{c} \begin{array}{ccc} f & g & h \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \end{array} \qquad \mathbf{M}^{(4,5)} = \begin{array}{c} \\ f \\ g \\ h \end{array}\!\!\begin{array}{c} \begin{array}{cc} i & j \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{C}^{(1,5)} = \mathbf{M}^{(1,2)} \cdot \mathbf{M}^{(2,3)} \cdot \mathbf{M}^{(3,4)} \cdot \mathbf{M}^{(4,5)} = \begin{array}{c} \\ a \\ b \end{array}\!\!\begin{array}{c} \begin{array}{cc} i & j \end{array} \\ \begin{bmatrix} 3 & 2 \\ 0 & 0 \end{bmatrix} \end{array}$$

Here, $C_{ai}^{(1,5)} = 3$ which means that there are three $\leq$-paths of the form $i \leq z \leq y \leq x \leq a$ between a and i passing through at least one element x of P_2, one y of P_3 and one z of P_4; they are $i \leq f \leq e \leq d \leq a$, $i \leq g \leq e \leq d \leq a$ and $i \leq h \leq e \leq d \leq a$. There are, $|P_2| \cdot |P_3| \cdot |P_4| = 6$ theoretical paths between a and i passing through at least one element x of P_2, one y of P_3 and one z of P_4, the remaining three paths are $i \leq f \leq e \parallel c \leq a$, $i \leq g \leq e \parallel c \leq a$ and $i \leq h \leq e \parallel c \leq a$, which are not possible because of the incomparability $c \parallel e$.

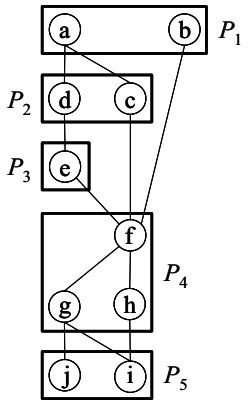


Figure 2. A Hasse diagram on the set $P = \{a, b, c, d, e, f, g, h, i, j\}$; four "boxed" subsets $P_1 = \{a, b\}$, $P_2 = \{c, d\}$, $P_3 = \{e\}$, $P_4 = \{f, g, h\}$ and $P_5 = \{i, j\}$.

2.2. Properties of dominance and separability degrees

Once the dominance and separability degrees are calculated, a critical value of dominance or separability may be selected for stating that one subset dominates or is separable from another one. Since dominance degree comes from the total number of possible comparabilities among the elements of the compared subsets, then it is said that the subset P_i dominates P_j, when more than half of the possible relations among the elements of P_i and P_j are comparabilities $y \leq x$ with $x \in P_i$ and $y \in P_j$. Thus, we are interested in values of $Dom(P_i, P_j) > 0.5$, which, according to Theorem 1, guarantees that $Dom(P_j, P_i) + Sep(P_i, P_j) < 0.5$; for that reason if $Dom(P_i, P_j) > 0.5$ then, P_j cannot dominate P_i. The limiting value for $Dom(P_i, P_j)$ could also be shifted to high scores, for example 0.9, in which case we look for subsets P_i and P_j for which 90% of the possible relations among elements of P_i and P_j are comparabilities where $y \leq x$ with $x \in P_i$ and $y \in P_j$.

Definition 19. We say P_i *ε-dominates* P_j iff $Dom(P_i, P_j) > \varepsilon$ with $\varepsilon \in [0.5, 1)$. In that case it is written $P_j \prec_\varepsilon P_i$.

It is important to note the meaning of $Dom(P_i, P_j) > \varepsilon$ with $\varepsilon \in [0.5, 1)$. It implies to have dominance degree values greater than an ε in the interval [0.5, 1), which means to have dominance degree values in the interval (0.5, 1].

In order to explore the properties of $\prec_\varepsilon$, we display six general properties of binary relations.

Definition 20. Let $X \neq \varnothing$ and R a binary relation on X. Then six possible properties of R are:

1. $x \in X \Rightarrow x\,R\,x$ (reflexive),
2. $x \in X \Rightarrow$ not $x\,R\,x$ (irreflexive),
3. $x, y \in X, x\,R\,y \Rightarrow y\,R\,x$ (symmetric),
4. $x, y \in X, x\,R\,y \Rightarrow$ not $y\,R\,x$ (asymmetric),
5. $x, y \in X, x\,R\,y$ and $y\,R\,x \Rightarrow x = y$ (antisymmetric),
6. $x, y, z \in X, x\,R\,y$ and $y\,R\,z \Rightarrow x\,R\,z$ (transitive).

Proposition 2. From the properties shown in Definition 20, $\prec_\varepsilon$ is only irreflexive and asymmetric on $\mathscr{P}$.

Proof. 1,2. $\prec_\varepsilon$ is irreflexive and not reflexive because it is a binary relation defined on $\mathscr{P}$, whose elements are disjoint subsets (Definition 17)□

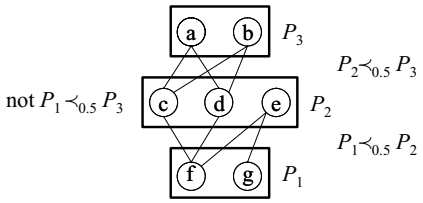
3. If $\prec_\varepsilon$ is to be a symmetric relation, then $P_j \prec_\varepsilon P_i \Rightarrow P_i \prec_\varepsilon P_j$, with $P_i, P_j \in \mathscr{P}$. If $P_j \prec_\varepsilon P_i$, then, from Definition 13, $T_{j \leq i} > \varepsilon(n_i \cdot n_j)$. According to Theorem 1, the maximum number of $x \leq y$ relations, for

all $x \in P_i$ and $y \in P_j$, is given by $T_{j\leq i} + T_{i\leq j} = n_i \cdot n_j$ ($T_{i\|j} = 0$). Knowing that $T_{j\leq i} > \varepsilon(n_i \cdot n_j)$ then $T_{i\leq j} < n_i \cdot n_j(1-\varepsilon)$. Hence, $T_{i\leq j}$ not $> \varepsilon(n_i \cdot n_j)$, therefore P_i not $\prec_\varepsilon P_j$□

4. If $\prec_\varepsilon$ is asymmetric then $P_j \prec_\varepsilon P_i \Rightarrow$ not $P_i \prec_\varepsilon P_j$ as was shown before (3)□

5. In order to prove that $\prec_\varepsilon$ is not antisymmetric, let $P_i, P_j \in \mathscr{P}$ and $n_i = |P_i|$, $n_j = |P_j|$. The two initial conditions of the antisymmetry are $P_j \prec_\varepsilon P_i$ and $P_i \prec_\varepsilon P_j$. It is enough to prove that they cannot be given simultaneously in $\mathscr{P}$. If $P_j \prec_\varepsilon P_i$ then, from Theorem 1, $T_{j\leq i} > (T_{i\leq j} + T_{i\|j})$; if $P_i \prec_\varepsilon P_j$ then $T_{i\leq j} > (T_{j\leq i} + T_{i\|j})$; thereby $T_{i\|j} < T_{j\leq i} - T_{i\leq j} < -T_{i\|j}$, which is a contradiction and $P_j \prec_\varepsilon P_i$, $P_i \prec_\varepsilon P_j$ cannot hold simultaneously□

6. An example showing the lack of transitivity of $\prec_\varepsilon$ is illustrated in Figure 3.



$Dom(P_3, P_2) = 0.67$
$Dom(P_2, P_1) = 0.67$
$Dom(P_3, P_1) = 0.5$

Figure 3. A Hasse diagram showing the lack of transitivity of $\prec_\varepsilon$.

$P_j \prec_\varepsilon P_i$ means $Dom(P_i, P_j) > \varepsilon$ with $\varepsilon \in [0.5, 1)$. In this example (Figure 3), $Dom(P_3, P_2)$, $Dom(P_2, P_1) > 0.5$, then $P_2 \prec_{0.5} P_3$ and $P_1 \prec_{0.5} P_2$ but $Dom(P_3, P_1) = 0.5$, therefore not $P_1 \prec_{0.5} P_3$.

If we consider again the Hasse diagram of Figure 3 and we add to it the comparability $g \leq d$, it keeps holding $P_2 \prec_{0.5} P_3$ and $P_1 \prec_{0.5} P_2$ but now $P_1 \prec_{0.5} P_3$, in fact $Dom(P_3, P_1) = 1$. Why does it occur? Because $g \leq d$ permits the additional comparabilities $g \leq a$ and $g \leq b$. On the other hand $P_2 \prec_{0.5} P_3$ and $P_1 \prec_{0.5} P_2$ are maintained together with $P_1 \prec_{0.5} P_3$ because more than half of the relations among elements of P_3 and P_1 correspond to $x \leq y \leq z$, where $x \in P_1$, $y \in P_2$ and $z \in P_3$. It is, more than half of the pairs x, z are related by a $\leq$-path $x \leq y \leq z$ passing through some element of P_2. Then, the existence of these paths is determinant for the $\leq$-relation of two subsets having a third one in between.

The above statement makes one think that $\prec_\varepsilon$ may become a transitive relation if it is endowed with $\leq$-paths. That is correct but then the transitivity is not a property of $\prec_\varepsilon$, as we show in Proposition 2, but of the structure ($\prec_\varepsilon$, $\leq$-paths).

In Figure 4 we show five Hasse diagrams, three subposets and their dominance relations for $\varepsilon = 0.5$; additionally we show the presence or absence of $\leq$-paths between those subposets. We write no $\leq$-paths (Figure 4) if the number of $\leq$-paths $x \leq y \leq z$ between P_i and P_k is less than or equal than half of $n_i \cdot n_k$.

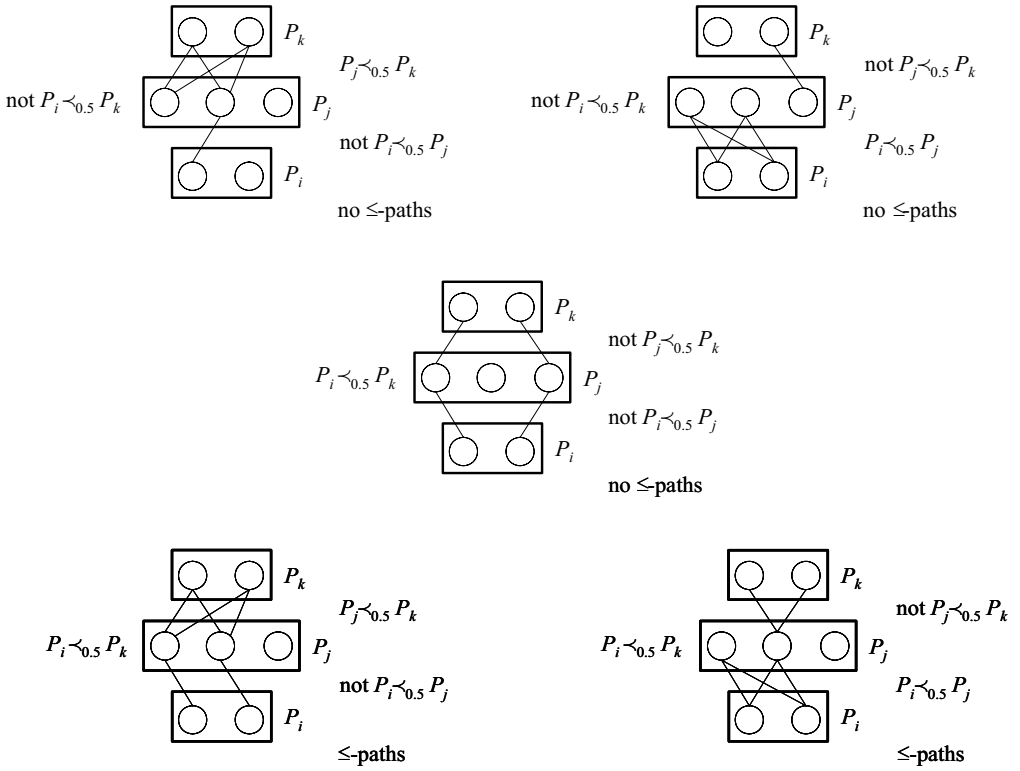


Figure 4. Four Hasse diagrams, $\prec_{0.5}$ relations for its subposets and the presence/absence of ≤-paths among them.

From the Hasse diagrams shown in Figure 4 and from the discussion about Figure 3, it can be concluded that always the presence of more than $(n_i \cdot n_k)/2$ ≤-paths $x \le y \le z$ guarantees that $P_i \prec_{0.5} P_k$. However, those ≤-paths alone are not a guarantee of $P_i \prec_{0.5} P_j$ and $P_j \prec_{0.5} P_k$. Then, in order to have an implication similar to $P_i \prec_{0.5} P_j$, $P_j \prec_{0.5} P_k \Rightarrow P_i \prec_{0.5} P_k$, it is necessary to meet $P_i \prec_{0.5} P_j$, $P_j \prec_{0.5} P_k$ and to guarantee the existence of more than $(n_i \cdot n_k)/2$ ≤-paths of the form $x \le y \le z$ with $x \in P_i$, $y \in P_j$ and $z \in P_k$.

Theorem 2. Let P_i, P_j, $P_k \in \mathscr{P}$ such that there are more than $\varepsilon(n_i \cdot n_k)$ ≤-paths of the sort $x \le y \le z$, with $x \in P_i$, $y \in P_j$ and $z \in P_k$, and $P_i \prec_\varepsilon P_j$, $P_j \prec_\varepsilon P_k$. If this is satisfied, then $P_i \prec_\varepsilon P_k$.

Proof. $P_i \prec_\varepsilon P_j$ ensures that there are more than $\varepsilon(n_i \cdot n_j)$ relations $x \le y$ with $x \in P_i$, $y \in P_j$. In the same way $P_j \prec_\varepsilon P_k$ guarantees the existence of more than $\varepsilon(n_j \cdot n_k)$ relations $y \le z$ with $z \in P_k$. Since $\varepsilon \in [0.5, 1)$, then both $P_i \prec_\varepsilon P_j$ and $P_j \prec_\varepsilon P_k$ ensure the existence of at least one ≤-path $x \le y \le z$. If there are more than $\varepsilon(n_i \cdot n_k)$ ≤-paths $x \le y \le z$, then $T_{i \le k} > \varepsilon(n_i \cdot n_k)$, thereby $Dom(P_k, P_i) > \varepsilon$ and $P_i \prec_\varepsilon P_k$. It is, the number of ≤-paths required to state $P_i \prec_\varepsilon P_k$ depend on n_i and n_k but not on n_j. Because of this, from the number of these ≤-paths cannot be inferred if $T_{j \le k} > \varepsilon(n_j \cdot n_k)$ neither if $T_{i \le j} > \varepsilon(n_i \cdot n_j)$; or in other words whether $P_j \prec_\varepsilon P_k$ or not. The inclusion of $P_j \prec_\varepsilon P_k$ and $P_i \prec_\varepsilon P_j$ as additional conditions of this theorem then are a guarantee that $P_i \prec_\varepsilon P_k$□

Definition 21. Two sets P_i, $P_j \in \mathscr{P}$ are said to be *ε-dominable* iff either $P_i \prec_\varepsilon P_j$ or $P_j \prec_\varepsilon P_i$. We say that P_i *ε-dominates* P_j if $P_j \prec_\varepsilon P_i$.

Definition 22. Given two sets $P_i, P_j \in \mathscr{P}$, we say that P_i is *covered by ε-dominance* by P_j, denoted $P_i \prec_\varepsilon: P_j$, if $P_i \prec_\varepsilon P_j$ and there is no $P_k \in \mathscr{P}$ for which $P_i \prec_\varepsilon P_k$ and $P_k \prec_\varepsilon P_j$. If $P_i \prec_\varepsilon: P_j$, it is said that P_j *covers by ε-dominance* P_i.

Definition 23. Given $P_i, P_j \in \mathscr{P}$, it is said that P_i is *δ-separated* from P_j or P_j is *δ-separated* from P_i iff $Sep(P_i, P_j) > \delta$ with $\delta \in [0.5, 1)$. In that case it is written $P_i \|_\delta P_j$ and we say that P_i and P_j are *δ-separable*.

Proposition 3. From the properties shown in Definition 20, $\|_\delta$ is irreflexive and symmetric on $\mathscr{P}$.
Proof. 1,2. The relation $\|_\delta$ is irreflexive because the elements of $\mathscr{P}$ are mutually disjoint subsets; for the same reason it is not reflexive□
3,4. $\|_\delta$ is symmetric by definition since $Sep(P_i, P_j) = Sep(P_j, P_i)$ (Definition 13); hence it is not asymmetric□
5. Although $P_i \|_\delta P_j$ and $P_j \|_\delta P_i$ can always coexist, it does not imply that $P_i = P_j$ because $P_i \cap P_j = \varnothing$, then $\|_\delta$ is not antisymmetric□
6. In order to show the no transitivity of $\|_\delta$, let us assume that the poset in Figure 3 is an incomparability graph (Definition 10), then each link in it is a $\|_\delta$ relation where it holds $P_1 \|_\delta P_2$, $P_2 \|_\delta P_3$ but not $P_1 \|_\delta P_3$. Therefore $\|_\delta$ is not a transitive relation on $\mathscr{P}$.

Then $\|_\delta$ is defined as follows:

Definition 24. The *δ-separability* relation $\|_\delta$ is a binary relation on $\mathscr{P}$ fulfilling these properties:
1. $P_i \in \mathscr{P} \Rightarrow$ not $P_i \|_\delta P_i$,
2. $P_i, P_j \in \mathscr{P}, P_i \|_\delta P_j \Rightarrow P_j \|_\delta P_i$.

Definition 25. Let $G_{\prec_\varepsilon} = (\mathscr{P}, E_{\prec_\varepsilon})$ the *ε-dominance graph*, where $E_{\prec_\varepsilon}$ is the set of edges containing all the ε-dominable pairs in $\mathscr{P}$.

Definition 26. Let $G_{\|_\delta} = (\mathscr{P}, E_{\|_\delta})$ the *δ-separability graph*, where $E_{\|_\delta}$ is the set of edges containing all the *δ*-separable pairs in $\mathscr{P}$.

Definition 27. Let $G_{\prec_\varepsilon:} = (\mathscr{P}, E_{\prec_\varepsilon:})$ the *cover by ε-dominance graph*, where $E_{\prec_\varepsilon:}$ is the set of edges containing all the cover by ε-dominance pairs in $\mathscr{P}$.

In general, because of the lack of transitivity of $\prec_\varepsilon$ it is not possible to associate a Hasse diagram to $\mathscr{P}$, but if the comparabilities and incomparabilities among the elements of $\mathscr{P}$ permit the existence of ≤-paths, as described in Theorem 2, then a Hasse diagram on the elements of $\mathscr{P}$ can be drawn.

Definition 28. Let $\mathscr{H} = (\mathscr{P}, d(E_{\prec_\varepsilon:}))$ a directed graph of $(\mathscr{P}, \prec_\varepsilon, C)$ where $d(E_{\prec_\varepsilon:})$ is the set of directed edges containing the cover by ε-dominance pairs in $\mathscr{P}$. $\mathscr{H}$ is called the *ε-dominance Hasse diagram* of the structure $(\mathscr{P}, \prec_\varepsilon, C)$, where C is the collection of ≤-paths described in Theorem 2, if $\mathscr{H}$ is drawn in the Euclidean plane whose horizontal/vertical coordinate system requires that the vertical coordinate of $P_i \in \mathscr{P}$ be larger than the one of $P_j \in \mathscr{P}$ if $P_j \prec_\varepsilon: P_i$.

3. Application to chemical posets

3.1. Ordering molecular total energies of isoelectronic species with equal total nuclear charge

The molecular total energy $E(Z, R)$ can be considered as a function of the nuclear geometry R and the nuclear charges Z in such a way that energy relations for different molecular species can be reached by variations of R and Z. By energy relations Mezey [45], Villaveces, Daza and Bernal [32,33,46] refer to order relationships between the total energy of molecular species. However, $E(Z, R)$ is mathematically complicated and it is usual to restrict the study of such relationships to particular cases of R and Z, e.g. Z fixed while R changes and R fixed while Z changes, both cases considering isoelectronic species [45]. The constraint of R fixed and Z variable, together with the Born-Oppenheimer approximation has led to obtain general expressions showing order relationships between total energies of isoelectronic molecular species [32]. This kind of studies were initiated by Thirring, Narnhofer, Lieb and Simon [47-49] in the 1970s and further extended by Mezey in the 1980s [45,50-53]. Villaveces, Daza and Bernal [32,33,46] have generalised these ideas and have developed elegant theorems to order isoelectronic molecular species in their minimum energy configurations. A brief description of these results is given as follows.

Two isoelectronic species $Z^{(A)}$ and $Z^{(B)}$ with equal total nuclear charge N are called *isoelectronic-isoprotonic species* and are represented by nuclear charge vectors $Z^{(A)} = (Z_1^{(A)}, Z_2^{(A)}, \ldots, Z_N^{(A)})$ and $Z^{(B)} = (Z_1^{(B)}, Z_2^{(B)}, \ldots, Z_N^{(B)})$, respectively, where $Z_i^{(k)}$ is the i-th component of the vector k, which corresponds to the i-th nucleus in the species k.

A set S of vectors in "general position" [46] constitutes the vertices of a polyhedron $P(S)$ in the space of isoelectronic-isoprotonic species. An example of vectors in general position is constituted by the atomic vectors (N, 0, 0,…, 0), (0, N, 0,…, 0),…, (0, 0, 0,…, N), which are atomic vectors of nuclear charge N. In general, $S = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}\}$ and the polyhedron is defined as

$$P(S) = \left\{ Z : Z = \sum_{k=1}^{p} \alpha_k Z^{(k)},\ Z^{(k)} \in S,\ \sum_{k=1}^{p} \alpha_k = 1,\ \alpha_k \geq 0,\ k = 1,2,\ldots,p \right\} \quad (7)$$

$P(S)$ contains all the isoelectronic-isoprotonic species that can be generated by linear combinations of isoelectronic-isoprotonic atomic vectors, if these vectors are selected as the generating vertices of the polyhedron.

The Born-Oppenheimer, BO, operator that generates the BO molecular total energy of any isoelectronic-isoprotonic species in the polyhedron can be expressed in terms of polyhedron vertices when all molecules in such a polyhedron hold the same nuclear configuration R. This energy is bounded by:

$$E_R(Z) \geq \sum_k^p \alpha_k E_R(Z^{(k)}) + Q \quad (8)$$

where Q depends on the vertices generating the polyhedron. If $Q \geq 0$, it can be removed from the inequality without altering it. It has been shown [32] that a set of $Z^{(k)}$'s yielding $Q \geq 0$ is a subset S_a of vertices, in which any two of them can be obtained by permutations of its components. Additionally, these permutations must be equal to a product of disjoint transpositions (details are given in reference [32,33]).

In general, if the minimum energy configurations of two isoelectronic-isoprotonic species Z and $Z^{(i)}$ are selected, and if Z can be obtained by permuting components of $Z^{(i)}$, then the following inequality holds [32,33]:

$$\min_R E(Z) \geq \min_R E(Z^{(i)}) \quad (9)$$

Hence, the BO molecular total energy of any two isoelectronic-isoprotonic species $Z^{(A)}$ and $Z^{(B)}$ in their minimum energy configurations can be ordered and one of these two order relations may result:

$$\min_R E(Z^{(A)}) \geq \min_R E(Z^{(B)}) \quad (10)$$

$$\min_R E(Z^{(B)}) \geq \min_R E(Z^{(A)}) \quad (11)$$

To check if Eq. 10 holds, the set of permutations of $Z^{(B)}$ is obtained, that is $S_B = \{Z^{(B)1}, Z^{(B)2},\ldots, Z^{(B)p}\}$. If $Z^{(A)}$ belongs to the polyhedron generated by S_B, then Eq. 10 is satisfied. In particular, this implies to solve the following set of linear equations:

$$\begin{aligned} Z_1^{(A)} &= \sum_i^p \alpha_i Z_1^{(B)i} \\ Z_2^{(A)} &= \sum_i^p \alpha_i Z_2^{(B)i} \\ &\vdots \\ Z_N^{(A)} &= \sum_i^p \alpha_i Z_N^{(B)i} \end{aligned} \qquad (12)$$

Thus, if each component of $Z^{(A)}$ is actually generated by linear combinations of vectors obtained by permutations of components of $Z^{(B)}$, then the BO molecular total energy of $Z^{(A)}$ is higher than the one of $Z^{(B)}$ in their minimum energy configurations. If it is not possible to solve the linear equations, then the other possibility (Eq. 11) must be tested. If none of these sets of linear equations can be solved, then it is said that the BO molecular total energies of $Z^{(A)}$ and $Z^{(B)}$ in their minimum energy configurations are incomparable. In that case it may be written

$$\min_R E(Z^{(A)}) \,||\, \min_R E(Z^{(B)}) \qquad (13)$$

In the following we apply the dominance and separability degrees to the BO molecular total energies of the complete set of isoelectronic-isoprotonic species with total nuclear charge 10.

3.2. Molecular total energies of isoelectronic-isoprotonic species with total nuclear charge 10

The Hasse diagram of the set P of BO molecular total energies of 42 isoelectronic-isoprotonic species with charge 10 was recently published by Daza and Bernal [32,33] (Figure 5). In this diagram, objects holding high and low energies are respectively located at the top and bottom of the diagram. We partition P into 10 subsets containing, each one, all the objects with same number of nuclei; these subsets are: P_1 = {Ne}, P_2 = {HF, HeO, NLi, CBe, B_2}, P_3 = {H_2O, NHeH, CLiH, CHe_2, BBeH, BLiHe, Be_2He, $BeLi_2$}, P_4 = {NH_3, $CHeH_2$, $BLiH_2$, BHe_2H, Be_2H_2, BeLiHeH, $BeHe_3$, Li_3H, Li_2He_2}, P_5 = {CH_4, $BHeH_3$, $BeLiH_3$, $BeHe_2H_2$, Li_2HeH_2, $LiHe_3H$, He_5}, P_6 = {BH_5, $BeHeH_4$, Li_2H_4, $LiHe_2H_3$, He_4H_2}, P_7 = {BeH_6, $LiHeH_5$, He_3H_4}, P_8 = {LiH_7, He_2H_6}, P_9 = {HeH_8} and P_{10} = {H_{10}}. Hence, $\mathscr{P}$ = $\{P_i : i = 1, 2,\ldots, 10\}$.

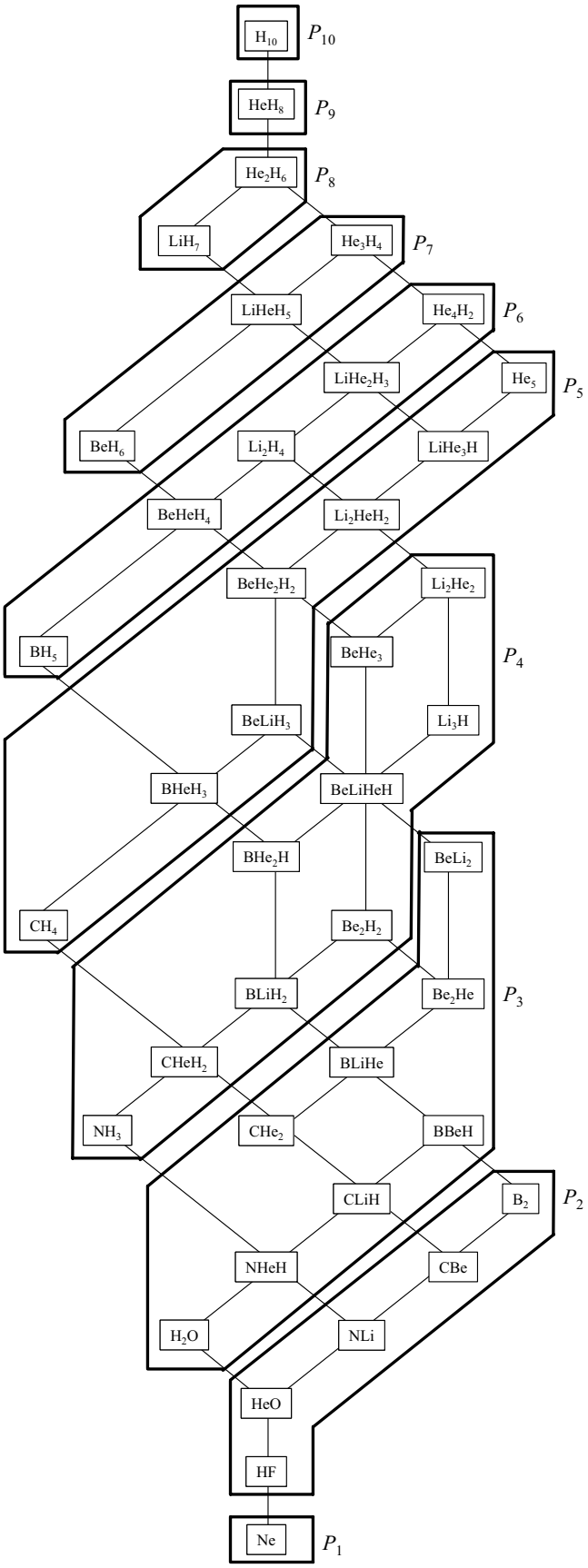


Figure 5. Hasse diagram of the 42 isoelectronic species with total nuclear charge 10. The boxes represent species with same number of nuclei.

This Hasse diagram, in fact a lattice, shows $P_{10} = \{H_{10}\}$ as the maximal subset, it is that the hydrogen cluster has the maximum BO energy of all the isoelectronic-isoprotonic species with total nuclear charge 10. In contrast, the minimal subset is $P_1 = \{Ne\}$, it is the Neon atom. Hence, the BO energies of all the isoelectronic-isoprotonic species in their minimum energy configurations with total nuclear charge 10 are in-between the energy of Ne and H_{10}. This result has been mathematically formalised and generalised by Daza and Bernal [32,33] to any set of isoelectronic-isoprotonic species in their minimum energy configurations.

From the total of 100 ordered pairs $(P_i, P_j) \in \mathscr{P} \times \mathscr{P}$, the dominance and separability degrees are defined for 90 of them because of the condition of having disjoint subsets (Definitions 14 and 15). Thus, for each one of the 45 sets $\{P_i, P_j\}$ (non-ordered pairs) the parameters $Dom(P_i, P_j)$, $Dom(P_j, P_i)$ and $Sep(P_i, P_j)$ were calculated (Table 1).

Table 1. Dominance and separability degrees for 10 subsets of the Hasse diagram depicted in Figure 5.

$\{P_i, P_j\}$[a]	$Dom(P_i, P_j)$	$Dom(P_j, P_i)$	$Sep(P_i, P_j)$
{1,2},{1,3},{1,4},{1,5},{1,6},{1,7}, {1,8},{1,9},{1,10},{2,6},{2,7},{2,8}, {2,9},{2,10},{3,7},{3,8},{3,9},{3,10}, {4,8},{4,9},{4,10},{5,9},{5,10},{6,9}, {6,10},{7,9},{7,10},{8,9},{8,10},{9,10}	0	1	0
{2,5}	0	0.97	0.03
{3,6}	0	0.95	0.05
{2,4},{4,7},{5,8}	0	0.93	0.07
{6,8}	0	0.9	0.1
{3,5}	0	0.89	0.11
{4,6}	0	0.84	0.16
{2,3},{7,8}	0	0.83	0.17
{5,7}	0	0.81	0.19
{3,4}	0	0.79	0.21
{4,5},{6,7}	0	0.73	0.27
{5,6}	0	0.69	0.31

[a] We renamed the set P_i and P_j as i and j, respectively.

It is particularly interesting to note that $Dom(P_i, P_j) = 0$ and $Dom(P_j, P_i) > Sep(P_i,P_j)$ in all the cases, then any comparison of two subsets P_i and P_j, where P_i contains objects with fewer nuclei than P_j, shows that $P_i \prec_\varepsilon P_j$, and because $Dom(P_i, P_j) = 0$ then there are no cases where a species x having fewer nuclei than another y has more energy than y as has been proved by Daza and Bernal [32,33].

From Table 1 it can be seen that 66.7% of the pairs $\{P_i, P_j\}$ correspond to the complete dominance of P_j over P_i, it is $Dom(P_j, P_i) = 1$ and $Sep(P_i, P_j) = 0$. These dominance and separability values are related by $P_i \prec_\varepsilon: P_j$, with ε taking all the possible values in the real interval (0.5, 1]. Additionally, these subsets do not present incomparabilities between their members, meaning that all the species in P_j have higher BO energies than all the species in P_i. This situation occurs for pairs of subsets located up and down in the Hasse diagram, for instance P_8 and P_2. The maximum values of separability degrees occur for adjacent pairs of subsets holding the highest number of incomparable BO energies between their objects, the pair of subsets with maximum separability is P_5, P_6 ($Sep(P_5, P_6) = 0.31$). Although this is the maximum separability degree value it cannot be stated that $P_5 \|_\delta P_6$ because the condition $Sep(P_5, P_6) > 0.5$ does not hold.

Since the structure of the Hasse diagram guarantees the existence of $\leq$-paths of the sort discussed in Theorem 2, it is possible to draw ε-dominance Hasse diagrams for $(\mathscr{P}, \prec_\varepsilon, C)$. We show in Figure 6 four ε-dominance Hasse diagrams.

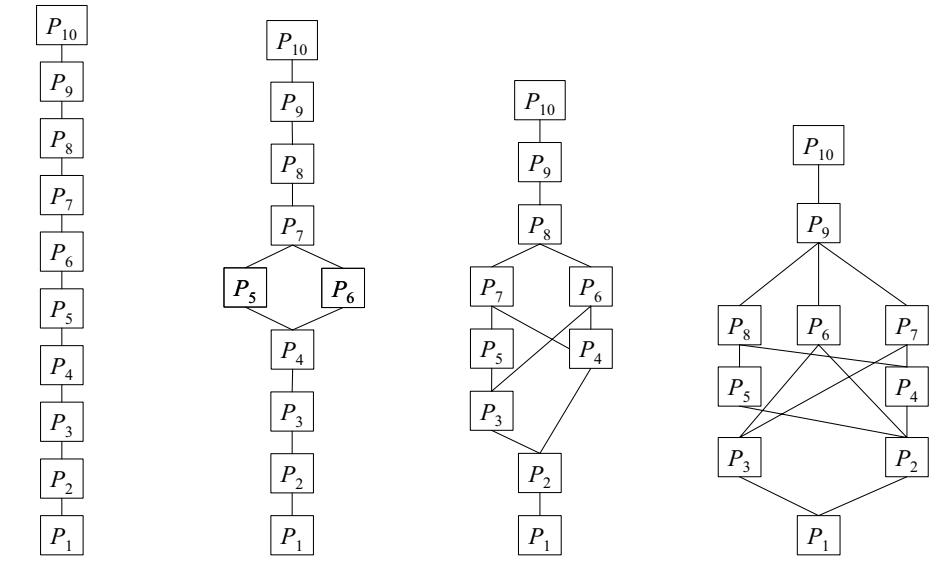


Figure 6. A) 0.5-, 0.6-; B) 0.7-; C) 0.8- D) 0.9-dominance Hasse diagrams of the subposets remarked in Figure 5.

Different ε-values yield different kinds of information regarding the comparabilities among the considered subposets. In general low ε-values give a broad landscape of the order relations and high ε-values permit going into the details of the relations. For example, the diagram A (Figure 6) shows dominance degree values greater than 0.5 and also greater than 0.6; from its linear order it can be concluded that a direct relationship between the BO energies and the number of nuclei in the species holds for these levels of dominance degrees. Using the dominance degree values it can be found the maximum value of ε for which the linear order of Figure 6A holds. In this case the linear order is kept only up to values of dominance degree equal or less than 0.7 (Figure 6B); this result can be interpreted as: given two subsets P_i and P_j containing, respectively, species with i and j nuclei ($i > j$), then at least 70% of the pairs $(x, y) \in P_i \times P_j$ holds that the energy of x is higher than the energy of y.

Additionally, the possibility of varying ε allows adjusting the level of detail we want to explore concerning the order relations; for instance, if $\varepsilon = 0.7$, then P_6 and P_5 become incomparable. That is, when we look for subsets where more than 70% of the pairs $(x, y) \in P_i \times P_j$ holds that the energy of x is higher than the energy of y, then P_6 and P_5 are not regarded because of their incomparability (not $P_5 \prec_{0.7} P_6$), it means that no more than 70% of the pairs $(x, y) \in P_6 \times P_5$ holds that the energy of x is higher than the one of y.

If ε is shifted to $\varepsilon = 0.8$ then more subsets become incomparable (Figure 6C) and it can be seen that they start to appear around P_5 and P_6 (the subsets having more incomparabilities per total relations). At this level of dominance degree the following relations can be seen: $P_3 \prec_{0.8} P_5 \prec_{0.8} P_7$, $P_2 \prec_{0.8} P_4 \prec_{0.8} P_6 \prec_{0.8} P_8$. In general a subset P_i in $\{P_2, P_3, \ldots, P_8\}$ dominates P_{i-2} in $\{P_2, P_3, \ldots, P_8\}$. This mainly occurs because the least energetic species in P_i is able to dominate more than 80% of the species in P_{i-2} but not more than 80% of the objects in P_{i-1}. This is the case for BeH_6, BH_5, CH_4 and NH_3.

Figure 6D shows the 0.9-dominance Hasse diagram, and it presents those subsets for which more than 90% of their species present more BO energies than others in different subsets. In general the amount of incomparabilities increase showing that the linear order depicted in Figure 6A is in fact caused by the low level of dominance considered in that case ($\varepsilon = 0.5, 0.6$).

Now if we consider high ε-values and compare the corresponding ε-diagrams with the ones of low ε-values, it can be seen that the dominances present in the diagram with high ε-values are kept in

the diagrams of low ε-values, for example $P_3 \prec_\varepsilon P_7$ in all the diagrams shown in Figure 6. This relation between Hasse diagrams is known as order preserving, in this case dominances present in high ε-dominance Hasse diagrams are preserved in the diagrams of low ε-values. Formally, given $\varepsilon > \varepsilon'$, the mapping $(\mathscr{P}, \prec_\varepsilon, C) \rightarrow (\mathscr{P}, \prec_{\varepsilon'}, C)$, $\prec_{\varepsilon'} \subset \prec_\varepsilon$ is order preserving if any $P_i \prec_\varepsilon P_j \in (\mathscr{P}, \prec_\varepsilon, C)$ implies $P_i \prec_{\varepsilon'} P_j \in (\mathscr{P}, \prec_{\varepsilon'}, C)$.

The incomparabilities between subsets that begin to be prominent for ε grater than 0.7, occur because of the distribution of the comparabilities among the objects in the subposet, as expected. All the relations among subsets (Figure 5), except those between P_{10}, P_9, P_2 and P_1 are characterised 1) for the surpassing of a high energetic species of P_j over all the species of P_i, where P_j contains objects with more nuclei than those in P_i; and 2) for the surpassing of the lowest energetic species of P_j over the 2 lowest energetic species of P_i. An example of 1) is the surpassing of $He_3H_4 \in P_7$ over all the objects in P_6 and an example of 2) is the surpassing of $CH_4 \in P_5$ only over $CHeH_2$ and NH_3, which belong to P_4. A remark extracted from 2) is that the deletion of the least energetic object of the subsets P_8 to P_3 would strengthen the dominance relations between the subsets and this effect would be especially notorious for high BO energy subsets. For example, the effect of removing LiH_7 in P_8 (a high BO energy subset) would cause that $Dom(P_8, P_7)$ change from 0.83 to 1 and the effect of deleting H_2O in P_3 (a low BO energy subset) would shifted $Dom(P_3, P_2)$ just from 0.83 to 0.89. This study opens the possibility of determining the most influent species for the dominance of their respective subsets when compared with others.

Regarding the separability degree results (Table 1), they are not a determining structural factor of the diagrams depicted in Figure 6 because of their low values. According to Theorem 1, these separability degree values are compensated for the high results of dominance degree which in turn yield many ε-dominances in the ε-dominance Hasse diagrams.

4. Conclusion and outlook

In this paper we present the mathematical background of a methodology that permits to draw conclusions on the relations of comparabilities and incomparabilities between pairs of disjoint subposets in $(P, \leq)$. Particularly, the dominance degree measures the extent of comparabilities in the considered subsets while the separability degree considers the corresponding incomparabilities. An advantage of this method over the analysis of the poset of class-representatives after clustering the elements in P is that the method here presented does not reduce the cardinality of the ground set P of the original poset, therefore all the relations between all the pairs of the two compared subsets are considered. It is, the dominance and separability of two subposets is assessed by evaluating all the elements of both subposets and their possible relations.

Similar researches have been developed in observational studies where the relations between two subsets of P are measured by their coherence. We have found that the functions used for measuring coherence are in fact functions of the dominance and separability degrees introduced in this paper. We consider that the application of dominance and separability degrees to observational studies would permit to give a detailed description of the cause-effect pattern these studies look for.

One of the uses of posets in chemistry is the ranking and prioritisation of individual objects (chemicals, regions and databases, for instance) when they have been defined by more than one of their properties. Dominance and separability degrees allow prioritising subsets of P based on the relations found in the given Hasse diagram of P. In this case we may rank complete subsets of similar chemicals, for instance, and explore their behaviour considering their order relations. Another chemical application of the concept of dominance degree was recently reported in the environmental ranking of families of refrigerants [54].

After defining and studying the mathematical properties of the dominance and separability degrees we discussed the implications that particular values $\varepsilon = [0.5, 1)$ of dominance and $\delta = [0.5, 1)$ of separability have over the collection of subposets of a Hasse diagram. Hence, $\prec_\varepsilon$ and $\|_\delta$ were introduced as binary relations and some of their properties were studied. Special attention was dedicated to the lack of transitivity of the dominance relation $\prec_\varepsilon$ and it was proved that when $\prec_\varepsilon$ is equipped with a collection of $\leq$-paths which consider at least one element of the sequence of subposets

compared, then the new relation ($\prec_\varepsilon$, C) becomes a transitive one on the collection of subsets where it is applied. This kind of transitivity dependent on the $\leq$-paths between the compared subposets keeps certain resemblance with the investigations on fuzzy transitive relations developed by De Baets [55-58]. Hence the study of ($\prec_\varepsilon$, C) as a fuzzy transitive relation and its implications must be explored in forthcoming investigations.

From the Hasse diagram of the Born-Oppenheimer, BO, molecular total energies of the complete set of isoelectronic-isoprotonic species with total nuclear charge 10 and from its partitioning into ten subsets containing, each one, all the objects with same number of nuclei, the following conclusions can be drawn:

1. More than half of the subsets dominate completely the others and these dominances correspond to subsets of species with more nuclei over subsets with species having fewer nuclei. This occurs because the energy of all the objects with more nuclei is greater than the energy of objects with few nuclei in more than half of the comparisons between subsets.
2. When looking for the maximum ε-value of dominance degree necessary to have a linear order showing the direct relationship between number of nuclei and BO energies it was found that it corresponds to $\varepsilon = 0.7$. This means that when considering all the 10 subposets, at least 70% of the pairs $(x, y) \in P_i \times P_j$ are cases where the energy of x is higher than the one of y, with P_i gathering species with more nuclei than those collected in P_j. This result sharps the general conclusion drawn by Daza and Bernal [32,33] on the direct relationship between the number of nuclei and the BO energies.
3. We found that, in the majority of cases, the removing of the least energetic species of a subset P_i increases the dominance degree of P_i over other subsets. This finding suggests a systematic study of the effect of removing species on the dominance degree values. Thus, it might be analysed which objects affect in a big extent the dominance relations among groups. Studies of this sort can be regarded as the searching for "hubs" in the poset ($\mathscr{P}$, $\prec_\varepsilon$, C) and it would be interesting to find a connection between these dominance posetic hubs and the hubs studied in network theory.

When the conditions are given for having a ε-dominance Hasse diagram (Theorem 2), some of these diagrams correspond to lattices, for example those with $\varepsilon = 0.5$, 0.6 and 0.7 in Figure 6, while some other ε-values yield no lattices. It is interesting to explore the relationship between ε and the lattice character of the ε-dominance Hasse diagram.

Once calculated the dominance degrees for the subposets of a Hasse diagram, it is possible to study particular ε-values as we did in this paper when selecting $\varepsilon = 0.5$, 0.6, 0.7, 0.8 and 0.9 (Figure 6). However, is also possible, and rather interesting, to plot each ε-Hasse diagram for each dominance degree value, for example, according to Table 1 it would be worthy to select $\varepsilon = Dom(P_j, P_i)$, it is 0.69, 0.73, 0.79,… to 0.97. Thus, it is possible to check which pairs of subsets become incomparable when increasing ε. Although this procedure is interesting, it may be intractable because the dominance degree values might be disperse on the real interval (0.5,1]. In such a case it is recommended clustering the dominance degree values in order to group near values in different regions of the (0.5,1] interval. Then, the analysis of the step-by-step changes in the diagrams can be replaced by the analysis of the changes when selecting an ε-value from each cluster. For example, if a given clustering process groups the values of $Dom(P_j, P_i)$ (Table 1) into these four clusters: [0.9,1], [0.81,0.89], [0.73,0.79], [0.69]; then it might be interesting to select a representative ε-value from each cluster and to draw the corresponding ε-dominance Hasse diagrams in order to compare them.

In spite of having found $Dom(P_i, P_j) = 0$ and $Dom(P_j, P_i) > Sep(P_i, P_j)$ for all the subposets considered in Figure 5, it does not mean that this is a general result attached to the dominance and separability degrees. Those values are strictly depended on the $\leq$-relations among the elements in the Hasse diagram and it is usual to find values where $Dom(P_i, P_j) > 0$ for some subposets and $Dom(P_j, P_i) > 0$ for some others, as well as $Sep(P_i, P_j) > 0$ for others. This diversity of dominances and separabilites makes possible to represent each ordered pair (P_i, P_j) as a point $(Dom(P_i, P_j), Dom(P_j, P_i), Sep(P_i, P_j))$ in a Cartesian space. Following the same idea drawn before, there may be found different clusters of similar dominated and separated subposets for which is interesting to study the ε-dominance Hasse diagrams among them. An example of a Hasse diagram over which is possible to apply this procedure is the one shown in reference [54].

A chemical application of the measurements here developed is to the chemical elements. Klein has suggested [35] that they may be regarded as a poset and several results by Restrepo and coworkers [59-62] have shown that the groups of chemical elements correspond to similarity classes. It is worthy to calculate the dominance and separability degrees among these chemical groups in order to check the ε-dominances and δ-separabilities among them and their possible relationship with their chemical behaviour.

In general, the dominance and separability degrees are useful mathematical tools for exploring the landscape of comparabilities and incomparabilities among subsets. Making use of them it is possible to "tune" the level of detail we want to achieve in our investigations on the order relations among subposets and this is achieved just by varying the ε- and δ-values.

Bernal [63] has pointed out the resemblance of the dominance and separability formalism with that one of blockmodel used in social network analysis [64] where a graph (a poset in the current case) is given, a partition is defined and relations between elements of the partition arising from relations between elements in the parts of the partition are analysed. That set of relations between parts of the partition defines in turn a graph (a diagram in the present work) which is further analysed in order to simplify the initial graph. In short, with this procedure "one can classify the objects of the graph and, even more; one can explore the relations between classes" [63].

Acknowledgments

G. Restrepo thanks COLCIENCIAS and the Universidad de Pamplona for the grant offered during this research. G. Restrepo thanks D. J. Klein from the Texas A&M University at Galveston (U.S.A.) and A. Bernal from the Universidad Nacional de Colombia at Bogotá (Colombia) for their valuable comments and suggestions. A. Bernal is especially thanked for the detailed explanation of his work.

References

[1] R. Brüggemann and L. Carlsen, *Partial Order in Environmental Sciences and Chemistry* (Springer, Berlin, 2006).

[2] I. Rival, *Algorithms and Order* (Kluwer, Dordrecht, 1989).

[3] P. Annoni, submitted to Environ. Ecol. Statist. Preprint available at: http://services.bepress.com/unimi/statistics/art15/

[4] D. L. Solomon, in: *Ecological diversity in theory and practice*, eds. J. F. Grassle, G. P. Patil, W. Smith and C. Taillie (International Co-operative Publishing House, Fairland, 1979), pp. 29-35.

[5] E. Ruch, Theoret. Chim. Acta (Berl.) 38 (1975) 167.

[6] E. Ruch and A. Mead, Theoret. Chim. Acta (Berl.) 41 (1976) 95.

[7] E. Ruch and B. Lesche, J. Chem. Phys. 69 (1978) 393.

[8] F. A. Matsen and D. J. Klein, J. Phys. Chem. 75 (1971) 1860.

[9] E. Ruch and A. Schönhofer, Theoret. Chim. Acta (Berl.) 19 (1970) 225.

[10] E. Ruch, Accounts Chem. Res. 5 (1972) 49.

[11] I. Gutman and M. Randić, Chem. Phys. Lett. 47 (1977) 15.

[12] M. Randić, Chem. Phys. Lett. 55 (1978) 547.

[13] M. Randić, J. Math. Chem. 4 (1990) 157.

[14] M. Randić, J. Chem. Educ. 69 (1992) 713.

[15] S. El-Basil and M. Randić, Adv. Quantum Chem. 24 (1992) 239.

[16] E. Halfon and M. G. Reggiani, Environ. Sci. Technol. 20 (1986) 1173.

[17] R. Brüggemann and B. Münzer, Chemosphere 27 (1993) 1729.

[18] R. Brüggemann, B. Münzer and E. Halfon, Chemosphere 28 (1994) 863.

[19] R. Brüggemann and H. G. Bartel, J. Chem. Inf. Comput. Sci. 39 (1999) 211.

[20] S. Pudenz, R. Brüggemann, B. Luther, A. Kaune and K. Kreimes, Chemosphere 40 (2000) 1373.

[21] K. Voigt, J. Gasteiger and R. Brüggemann, J. Chem. Inf. Comput. Sci. 40 (2000) 44.

[22] R. Brüggemann, E. Halfon, G. Welzl, K. Voigt and C. E. W. Steinberg, J. Chem. Inf. Comput. Sci. 41 (2001) 918.

[23] D. Lerche, R. Brüggemann, P. Sørensen, L. Carlsen and O. J. Nielsen, J. Chem. Inf. Comput. Sci. 42 (2002) 1086.

[24] D. Lerche, P. Sørensen and R. Brüggemann, J. Chem. Inf. Comput. Sci. 43 (2003) 1471.

[25] R. Brüggemann, G. Welzl and K. Voigt, J. Chem. Inf. Comput. Sci. 43 (2003) 1771.

[26] R. Brüggemann, P. B. Sørensen, D. Lerche and L. Carlsen, J. Chem. Inf. Comput. Sci. 44 (2004) 618.

[27] R. Brüggemann, G. Restrepo and K. Voigt, J. Chem. Inf. Comput. Sci. 46 (2006) 894.

[28] G. Restrepo and R. Brüggemann, in: *Recent Progress in Computational Sciences and Engineering*, eds. T. Simos and G. Maroulis (VSP, Leiden, 2006), pp. 1386-1389.

[29] G. Restrepo and R. Brüggemann, submitted to Croat. Chem. Acta.

[30] J. Gabarro-Arpa, J. Math. Chem. First online (August 29, 2007).

[31] J. R. Dias, J. Math. Chem. 4 (1990) 17.

[32] E. E. Daza and A. Bernal, J. Math. Chem. 38 (2005) 247.

[33] A. Bernal, *Ordenamientos moleculares basados en la energía* (BSc Thesis, Universidad Nacional de Colombia, Bogotá, 2004).

[34] Papers in MATCH Commun. Math. Comput. Chem. 42 (2000) pp. 7-290 and 54 (2005) pp. 489-689.

[35] D. J. Klein, J. Math. Chem. 18 (1995) 321.

[36] D. J. Klein and D. Babić, J. Chem. Inf. Comput. Sci. 37 (1997) 656.

[37] I. Rival and N. Zaguia, Congressus numerantium 55 (1986) 199.

[38] G. Brightwell and P. Winkler, Order 8 (1991) 225.

[39] W. T. Trotter, *Combinatorics and Partially Ordered Sets, Dimension Theory* (The Johns Hopkins University Press, Baltimore, 1992).

[40] G. Restrepo and R. Brüggemann, WSEAS Trans. Inf. Sci. Appl. 2 (2005) 976.

[41] S. T. Hedetniemi and R. C. Laskar, *Topics on Domination* (North-Holland, Amsterdam, 1991).

[42] P. R. Rosenbaum, *Observational studies* (Springer, New York, 1995).

[43] P. R. Rosenbaum, Ann. Stat. 19 (1991) 1091.

[44] O. Gefeller and L. Pralle, in: *Nonrandomized comparative clinical studies. Proceedings of the International Conference on Nonrandomized Comparative Clinical Studies, 10–11 April 1997, Heidelberg, Germany*, eds. U. Abel and A. Koch (Symposion Publishing, Düsseldorf, 1997).

[45] P. G.Mezey, J. Am. Chem. Soc. 107 (1985) 3100.

[46] E. E. Daza and J. L. Villaveces, J. Chem. Inf. Comput. Sci. 34 (1994) 309.

[47] W. Thirring, Acta Phys. Aust. Suppl. 14 (1975) 631.

[48] H. Narnhofer and W. Thirring, Acta Phys. Aust. 41 (1975) 281.

[49] E. H. Lieb and B. Simon, J. Phys. B. 11 (1978) 1537.

[50] P. G.Mezey, Theor. Chim. Acta. 59 (1981) 321.

[51] P. G.Mezey, Int. J. Quant. Chem. 22 (1982) 101.

[52] P. G.Mezey, Mol. Phys. 47 (1982) 121.

[53] P. G.Mezey, J. Chem. Phys. 80 (1984) 5055.

[54] G. Restrepo, W. Weckert, R. Brüggemann, S. Gerstmann and Hartmut Frank, submitted to Environ. Sci. Technol.

[55] B. De Baets and H. De Meyer, Soft Comput. 7 (2003) 210.

[56] B. De Baets and H. De Meyer, Inform. Sciences 152 (2003) 167.

[57] H. De Meyer, H. Naessens and B. De Baets, Eur. J. Oper. Res. 155 (2004) 226.

[58] B. De Baets, H. De Meyer, B. De Schuymer and S. Jenei, Soc. Choice Welfare 26 (2006) 217.

[59] G. Restrepo, H. Mesa, E. Llanos and J. L. Villaveces, J. Chem. Inf. Comput. Sci. 44 (2004) 68.

[60] G. Restrepo and J. L. Villaveces, Croat. Chem. Acta 78 (2005) 275.

[61] G. Restrepo, E. J. Llanos and H. Mesa. J. Math. Chem. 39 (2006) 401.

[62] G. Restrepo, H. Mesa, E. Llanos and J. L. Villaveces, in: The Mathematics of the Periodic Table, eds. R. B. King and D. H. Rouvray (Nova, New York, 2006), pp. 75-100.

[63] A. Bernal. Private communication.

[64] W. de Nooy, A. Mrvar and V. Batagelj, *Exploratory Social Network Analysis with Pajek* (Cambridge, Cambridge, 2006).

Appendix F

Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software xx (2007) 1–13
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Concept of stability fields and hot spots in ranking of environmental chemicals

Rainer Brüggemann [a], Kristina Voigt [b,*], Guillermo Restrepo [c,d], Ute Simon [e]

[a] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, 12587 Berlin, Germany
[b] GSF-National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
[c] Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia
[d] University of Bayreuth, Chair Environmental Chemistry and Ecotoxicology, Bayreuth, Germany
[e] Leibniz-University Hanover, Institute of Meteorology and Climatology, Hanover, Germany

Received 12 February 2007; received in revised form 13 November 2007; accepted 18 November 2007

Q1 Abstract

In contrast to conventional multicriteria decision aids, such as the well known PROMETHEE approach, AHP or the different versions of ELECTRE, we support the basic assumption of environmetrics: let first the data speak, and then let us include subjective preferences in order to get a unique decision. In the present paper we introduce and discuss the decision support system METEOR (Method of Evaluation by Order Theory). The basis of the method is a data matrix. The rows are defined by the objects which are to be evaluated; the columns are defined by the attributes, which characterize the objects with respect to the evaluation problem. By means of the attributes a partial order is derived. In subsequent steps attributes are aggregated by a weighting procedure, allowing a high degree of participation of stakeholders and other participants of the planning process. The aim is to enrich the partial order stepwise, until the objects of interest can be compared. The software WHASSE written in Delphi is available for scientific purposes from the first author.

As an example we evaluate 12 high production volume chemicals (HPVC) which have been detected in the environment by four attributes and discuss the enriched partial order after introducing some weights. It turns out that in some cases the weights have almost no influence concerning the evaluation result, whereas in some other cases slight variations of weights drastically change the evaluation result. Therefore, the metric space spanned by weights can be partitioned in so-called ''stability fields'' where the evaluation result is invariant and in so-called ''hot spots'', where the evaluation is strongly changing. This characterisation of the space of weights is helpful for stakeholders to express their preferences.

Keywords: HPV chemicals; Environmental chemicals; Ranking; Posets; Hasse diagram technique (HDT); Decision support systems; Cluster analysis; Method of Evaluation by Order Theory (METEOR); Stability fields; Environmetrics; Environmental software

1. Introduction

Multi-criteria decision making becomes more and more important in environmental sciences and hence quite a few research projects focus on this topic. For example the MULINO Decision Support System (mDSS) has been developed for implementing the European Water Framework Directive, namely integrating environmental, social and economic concerns (Giupponi, 2007). Another example concerning the Integrated Water Resources Management (IWRM), is a multi Objective Decision Support System (MODSS) which has been developed and applied to the planning of the Lake Maggiore (Castelletti and Soncini-Sessa, 2006). The Elbe-Decision Support System is a computer based system for integrated river basin management of the German river Elbe basin and is therefore another example for an environmental decision

* Corresponding author. Tel.: +49 89 3187 4029; fax: +49 89 3187 3029.
E-mail addresses: brg_home@web.de (R. Brüggemann), kvoigt@gsf.de (K. Voigt).

1364-8152/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.envsoft.2007.11.001

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

support system (Berlekamp et al., 2007). A methodology based on a hybrid approach that combines fuzzy inference systems and artificial neural networks has been used to classify the ecological status in surface waters. This methodology is applied to sampling sites in the Ebro river basin and can support decision makers in evaluation and classification of ecological status, as required by the EU Water Framework Directive (Ocampo-Duque et al., 2007). The chemical speciation model BIOCHEM comprises ecotoxicological transfer functions for uptake of metals (As, Cd, Cu, Ni, Pb, and Zn) by plants and soil invertebrates and is another example for a flexible and dynamic decision support system (DSS) to analyse natural or anthropogenic changes that occur in river systems (Vink and Meeussen, 2007). A further interesting work including the spatial factor is a multi-criteria decision making approach applied to urban water management (Makropoulos and Butler, 2006). Concepts for the use of techniques of decision analysis to structure scientist and stakeholder involvement in river rehabilitation decisions are published by Reichert et al. (2007). The software, named proDEX is also applied as a multi-criteria decision support model in environmental sciences (Znidarsc et al., 2006).

Decisions concerning risk assessment of chemicals are to be supported by information about exposure and effects of chemicals. Both, exposure and effects are used as attributes/indicators to evaluate the chemicals under investigation. For the subsequent evaluation of chemicals, many methodological approaches are available, requiring in principle the same working steps, which are discussed in more detail in Simon et al. (2005) and Klauer et al. (2001). One step, namely the evaluation algorithm is often almost disregarded in real evaluations. The chosen evaluation approach however influences the evaluation result and the participation of stakeholders. The efficiency of participation of stakeholders and the acceptance of the decision result in turn depends on the transparency of the evaluation procedure. For example: decisions about complex problems such as chemical risk assessment will include conflicting attributes. To solve such conflicts, the most commonly used approaches include the methodological step of attributes' aggregation. The benefit of the aggregation step is that finally a linear ranking of the objects (here: chemicals) can be obtained, identifying *one* best solution, e.g. the chemical with the lowest risk. Aggregation often implies a trade off among attributes: a bad evaluation in one or more attribute(s) can be compensated for by a good evaluation in other attributes. However, attributes can represent fundamentally different aspects such as accumulation, mobility and toxicity. Therefore the methodological idea followed in this paper is to take first a purely statistical explorative point of view (i.e. "let first the data speak") and to include additional knowledge, e.g. the preferences of the stakeholder, in separate steps in order to keep a maximal control over the effect of including additional knowledge.

The paper is organized as follows: in Section 2 the example of 12 high-production volume chemicals (HPVC) is introduced, the methods Hasse diagram technique (HDT) and Method of Evaluation by Order Theory (METEOR) and the concept of crucial weights together with their analysis toward the introduction of "*g*-spectra, stability fields and hot spots" are explained. Whereas for the sake of demonstration a simpler example is used, Section 3 shows the application of METEOR on the HPVC-data matrix. A detailed discussion about possible extensions of the method concludes the paper. Additionally, there are appendices 1–4, where abbreviations, symbols and concepts are listed (Appendix 1) and where some counting formulas are explained (Appendices 2–4).

2. Materials and methods

2.1. Data preprocessing

With publication of the White Book of the EU (EEC, 2001) and of the REACH-procedure (European Commission, 2006) the interest in ranking of chemicals as a preparatory step is renewed: here the data matrix (12 high production volume chemicals) define the rows, and 4 attributes define the columns), first published by Lerche (2002a) is taken as a ranking example and is more extensively described in the Section 3. We are calling the set of objects (i.e. of chemicals) C.

"Results". Note, that we refer to 'objects' instead of chemicals as long we are not discussing the real life example.

Often it is necessary to transform a data matrix into the appropriate form i.e. into the closed interval [0,1] for an evaluation:

(i) a normalization by

$$\overline{q}_i(j) := \frac{q_i(j) - q_i(\min)}{q_i(\max) - q_i(\min)}, \quad i = 1, \ldots, 4, \quad j \in C$$

(ii) check for a common orientation (high numerical value indicates a high risk) by multiplying attributes − if necessary − with −1 or another appropriate transformation

(iii) shifting negative values to positive entries by adding a positive number to the attribute values.

The subjective preferences of stakeholders are expressed by weights, which are taken from the closed interval [0,1]. We consider the weights as 'external knowledge', whereas the data matrix expresses the basic information taken from measurements or modelling.

2.2. Hasse diagram technique

Several well-known evaluation algorithms are available such as PROMETHEE (Brans and Vincke, 1985), AHP (Saaty, 1994), MAUT (Schneeweiss, 1991), ELECTRE (Roy, 1990) or NAIADE (Matarazzo and Munda, 2001). All these methods include an aggregation of attributes by including subjective preferences and therefore cannot be considered as purely data explorative methods. Beyond this it is difficult to trace back how the evaluation result was influenced by parameters to run those algorithms. Hence we consider these high sophisticated methods on the one side as efficient, as they deliver a unique decision, but on the other side as not transparent and difficult to handle as all preferences must be at hand simultaneously.

An alternative approach is provided by simple elements of partial order theory, such as Hasse diagram technique (HDT) (Brüggemann and Voigt, 1995; Brüggemann and Welzl, 2002; Brüggemann and Carlsen, 2006; Brüggemann et al., 1994, 2001, 2006a; Voigt et al., 2004a,b, 2006). For the sake of clarity we define some important concepts used in this paper.

Definition 1. We call x an object and C the ground set that is the set of objects.

Definition 2. $q_i(x)$ is the ith attribute of the object x and $IB = \{q_i | i = 1, 2, \ldots, m\}$ the set of m attributes (information base).

Definition 3. Let $x, y \in C$ and $q_i \in IB$, then $x \leq y$ if $q_i(x) \leq q_i(y)$ for all $i = 1, 2, \ldots, m$. We say that x and y are comparable. If the orientation does not play a role, we write $x \perp y$ to express that x and y are comparable.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

By definition 3 a product- (or component-wise-) order is given and it obeys the following axioms of order:

i) reflexivity (an object can be compared with itself)
ii) antisymmetry (if an object x is better that y, then y is worse than x)
iii) transitivity (if $x, y, z \in C$ and $x \leq y$ and $y \leq z$ then $x \leq z$).

Definition 4. Two objects $x, y \in C$ are called incomparable ($x||y$) if there is at least one q_i with $q_i(x) > q_i(y)$ and one q_j with $q_j(x) < q_j(y)$.

Sometimes we add further information to the order relation. For example $b||_{q1,q2}d$ expressing that b is incomparable with d with respect to the attribute values of q_1 and q_2. The evaluation result, a partially ordered set, is visualized in a Hasse diagram (HD) (see e.g. Fig. 1). We show in Table 1 an example of a data matrix used to evaluate a set of 5 objects $C = \{a, b, c, d, e\}$ characterized by two attributes (q_1 and q_2).

Hasse diagrams are digraphs, which have no cycles (because of the order-axiom of antisymmetry) and — as ordinary graphs — have no triangles (because of the order-axiom of transitivity). The software realization of HDT, "WHASSE" (Brüggemann et al., 1999) provides several tools for convenient and detailed data analysis such as the sensitivity of the structure of the digraph with respect to different attributes (Brüggemann et al., 2001; Brüggemann and Welzl, 2002). The software is available for scientific purposes from the first author. WHASSE is written in Delphi, equipped with a comfortable GUI running under the operation system Windows NT and XP. As no compensation among attributes is carried out at all, conflicting evaluations of attributes cannot be methodologically removed. Consequently multiple favourable objects can be identified as incomparable winners. In our example (Table 1, Fig. 1) assuming that low values are favourable there are two incomparable objects, namely a and c. Altogether we find three incomparabilities in the Hasse diagram, symbolically written as: $b||d$, $b||c$ and $a||c$ and five cover-relations denoted by the symbol "$\cdot>$" and lines in Fig. 1, (details, see Brüggemann et al., 1994) $e\cdot > b$, $e\cdot > d$, $b\cdot > a$, $d\cdot > a$ and $d\cdot > c$.

2.3. A new concept to solve the problem of incomparable objects: METEOR

2.3.1. Overview

Partial order theory provides many concepts to derive linear orders without any additional introduction of (stakeholder's) preferences (Lerche et al., 2002b, 2003; De Loof et al., 2006). As no subjective weighting is involved the linear order obtained from a partial order is called a "canonical (linear) order" (Brüggemann et al., 2004, 2005). In contrast to derive canonical linear orders, METEOR (Method of Evaluation by Order Theory) attempts to resolve the incomparabilities among objects by inclusion of external knowledge. METEOR intends to obtain a clear decision (one best object), maintaining transparency and allowing participation [see for details Simon et al. (2005) and Voigt and Brüggemann (2005)]. It is conceptually based on the well known and often used concept of a hierarchy of criteria in multi-criteria decision aids (as e.g. in the AHP method (Saaty, 1994)). Basically METEOR allows a step-by-step aggregation of attributes by forming e.g. weighted sums about subsets of attributes. Principally non-linear aggregation (non-linear with respect to attributes) is also possible but still has not worked out because of its inherent complexity.

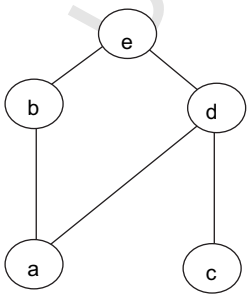


Fig. 1. Hasse diagram of the data matrix of Table 1.

Table 1
Data matrix as an example to construct a Hasse diagram (Fig. 1)

Objects	q_1	q_2
A	15	5
b	25	35
c	10	60
d	20	70
e	60	80

The possibility of a step-by-step aggregation of attributes provides the freedom to thoroughly analyse the effects of attribute weights and compensation. Furthermore, preferences (attribute weights) which are most sensitive to the evaluation result can easily be identified.

One may consider the data matrix characterizing the objects by attributes as primary knowledge, which is based on measurements, mathematical models, causal relations. Inclusion of knowledge beyond the data matrix means that relations are supposed among the attributes and implies external information: for example, one may consider one attribute to be more important than another one (Brüggemann et al., 2006b). Here we introduce the notion of "importance" of attributes from a technical, pragmatic point of view: importance is expressed by weights. In Fig. 2 HDT, METEOR and conventional algorithms like PROMETHEE, ELECTRE, etc. are compared: it is schematically shown how the inclusion of weights (external knowledge) reduces the transparency and the objectivity of the decision process (dashed line), whereas the efficiency (i.e. the ability to identify uniquely a best (or worst) object) of the decision process (continuous line) is enhanced. Methods, such as HDT may be located at the high transparency and low efficiency side, and methods like PROMETHEE at the high efficiency and low transparency side. METEOR may be located in between these two extreme cases.

2.3.2. METEOR as iterative application of HDT

The kernel of METEOR is the Hasse diagram technique (HDT). METEOR is discussed in detail by Simon et al. (2005) and was developed to solve conflicts among objects stepwise. Taking a look at two objects x, y characterized by m attributes it often occurs that one of them is evaluated better in one attribute and worse in the other one and the other way round. The incomparabilities $x||y$ between any two objects indicate the conflict among at least two of their attributes and one may decide that compensation is useful. Then a new "aggregated attribute", for example by a weighted sum of two original attributes may be constructed. The new information base IB' consists now instead of originally m attributes of only m-1 and the number of comparabilities increases. If the aggregation is done in such a way that a weak positive monotonous function f of attributes is found, then this aggregation is equivalent with

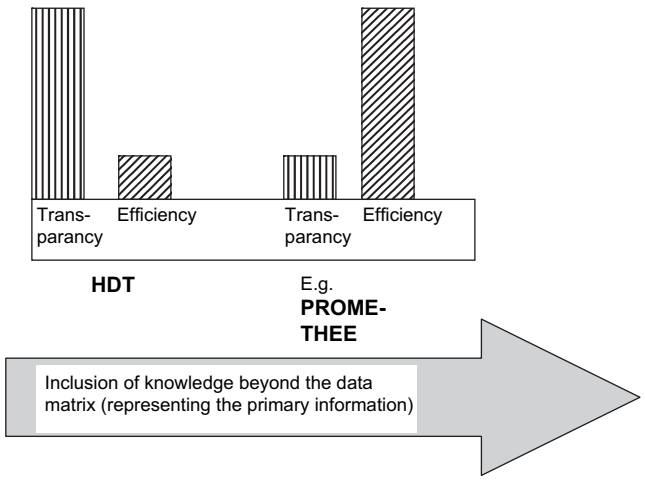


Fig. 2. Behaviour of transparency, objectivity and efficiency of a decision process (see text for further explanation).

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

an order preserving map. Presently in METEOR specifically a linear function is selected, however, non-linear aggregations are not excluded but more difficult to evaluate by explicit analytical expressions (see later). The step-by-step aggregation can be carried out until one single "aggregated hyper-attribute" is obtained, which is considered as a weighted sum of all original m attributes, and which will lead to a linear ranking of all objects (in technical terms: a weak linear order (because equivalent objects may appear)).

2.3.3. Aggregation strategies

If m attributes are considered then basically 2^m-m-2 attribute subsets can be potential candidates for forming new aggregated attributes (see Appendix 2). It is hypothesized that it is not meaningful to include subsets which are not disjoint, hence a more suitable base to discuss the stepwise aggregation is to analyze the partitions of IB (Appendix 3). It is well known that the number of partitions one can obtain, can be calculated after the Stirling numbers of the second kind if the number of classes is known (Appendix 4) (see for example Brüggemann and Drescher-Kaden, 2003, p. 95 f) or – if the number of classes is left open – by the Bell formula (Appendix 5) (Bock, 1974). Taking for example four attributes and assuming two classes for the partitioning, we get 7 partitions (one of these, for example, is $\{(q_1, q_2), (q_3, q_4)\}$ another one $\{(q_1, q_2, q_3), (q_4)\}$).

If, for example, 20 attributes are considered, then by applying the Bell formula 10^{13} partitions are possible, i.e. 10^{13} disjoint subsets of attributes can be formed in order to aggregate the attributes. Therefore some kind of heuristics is needed to find a way through the jungle of possible aggregations.

Clearly from a logical point of view one should start with attributes belonging to one sub-criterion. If for example chemicals are to be evaluated, one may consider exposure attributes on one side as candidates for an aggregation and effect attributes as candidates for another aggregation, obtaining two super-attributes "Exposure" and "Effects". This point of view is comfortable for stakeholders as it allows them first to consider general aspects and then – perhaps – to go into details. We had in mind this procedure, when we first established METEOR. From the point of evaluation we might call this procedure a *bottom-up procedure* (from the basis of detailed information, to more generalized concepts via sub-criteria). However similar attributes are often well correlated (indeed one may even define similarity by the correlation behaviour) and their aggregation has little effect on the poset and is consequently of little use for decision making. More efficient is to aggregate those attributes which have a high degree of conflicting potential. Those attributes are often anti-correlated. Hence their aggregation will rather efficiently reduce the incomparabilities. This kind of procedure one may call a *top-down procedure*: first reduce the most conflicting attribute subsets and then analyze the results by applying partial order. Even if we have decided which procedure we will follow it is not clear how the aggregation functions should look like (linear or non-linear). Before we proceed, some more definitions and notations are needed:

Definition 5. We call $S(k) = \{q_i | i < m\}$ the set of aggregated attributes and $n(k)$ its cardinality.

The corresponding super-attribute, φ_k based on $S(k)$, is calculated as

$$\varphi_k = \sum_{i=1}^{n(k)} g_i q_i$$

together with the normalization:

$$\sum_{i=1}^{n(k)} g_i = 1 \quad \text{and} \quad q_i \in S(k).$$

If we call $n(k) < m$ the number of attributes actually aggregated, then any super-attribute has (because of the weights' normalisation) $n(k) - 1$ "freedom" of freely varying the scalars $g_i \in [0,1]$. We call $[0,1]^{n(k)-1}$ the g-space of the kth super-attribute. Therefore we associate to any super-attribute a metric space of weights with the dimension $n(k) - 1$ and any aggregation step in METEOR is accompanied by the product of all g-spaces, which we call the G-space. In general $n(k)$ may vary and may depend on the intuition of the researcher, applying METEOR. Here, however, we restrict ourselves on aggregation schemes with freedom 1, i.e. we analyze in the subsequent parts of the paper for any super-attribute a g-space of dimension 1. If we combine for example

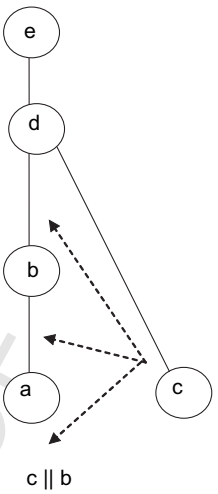


Fig. 3. Scheme explaining the "local" condition derived from $c||b$. For object c above b there is only one accessible position; for c below b there are two positions available.

(as we will do and describe later) four attributes pairwise to two super-attributes, the two linear g-spaces are combined, forming a two-dimensional space $[0,1] \times [0,1]$. As we also will see later in the text, the restriction to freedom = 1 simplifies considerably the procedure and we call a procedure, based on a purely pairwise combination of attributes the "orthogonal-METEOR" (abbr.: o-METEOR). In another paper Restrepo et al. describe a non-orthogonal METEOR procedure, with refrigerants as an example (Restrepo et al., 2007a).

Finally – depending on the task of the decision procedure – a good idea is not to perform aggregations until a linear order is generated, but to stop the aggregation if at least a greatest or a least element is found or if any two objects of specific interest can be compared. We call this strategy the "*extremal case – procedure*".

Careful analysis is needed if attributes which will be combined by a weighted sum are not on the same scaling level, if for example continuous variables are used in the evaluation process together with linguistic ones, even if one gives them an ordinal or metric interpretation. The best strategy in such cases is, just to stop the aggregation process before attributes of different scaling levels are numerically combined. We see the possibility of taking care of different scaling levels as a main advantage of the step-wise procedure in METEOR. Here, however for the sake of demonstration the role and structure of g-spaces we consider all attributes as metric quantities.

2.4. The concept of crucial weights

Imagine that four attributes, i.e. IB = $\{q_1, q_2, q_3, q_4\}$ are pairwise aggregated as follows:

$$\varphi_1 = g_1 * q_1 + (1 - g_1) * q_2 \tag{1a}$$

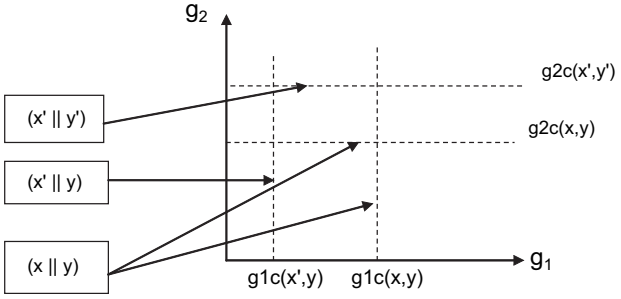


Fig. 4. Exemplifying conclusion 7 in the case of two dimensions (i.e. aggregating attributes pairwise to two super-attributes).

Table 2
Towards H_k: if there are two attributes gathered in k then for each pair of objects one has to calculate Δ_i, Q and gkc

k index	Pair	Δ_1	Δ_2	Q	gc
1	ac	$15-10=5$	$5-60=-55$	-0.09	0.917
2	bc	$25-10=15$	$35-60=-25$	-0.6	0.625
3	bd	$25-20=5$	$35-70=-35$	-0.143	0.875

$H_1(0.917)=1$, $H_2(0.625)=1$, $H_3(0.875)=1$.

and

$$\varphi_2 = g_2 * q_3 + (1-g_2) * q_4 \tag{1b}$$

Now, assume that the object $x \in C$ is incomparable with object y due to:

$$q_1(x) > q_1(y) \quad \text{and} \quad q_2(x) < q_2(y).$$

For this case we write: $x||_{q1,q2}y$.

If $x||_{q1,q2}y$ then the result of aggregation (1a) for those particular objects x and y depends on the weight g_1. Obviously the equation

$$\varphi_1(x) = \varphi_1(y) \tag{2}$$

determines the g_1 value where the character of order relation between x and y changes. Note firstly that Eq. (2) is the reason that non-linear aggregation functions will be more difficult to evaluate: instead of an analytical expression derived from Eq. (2) a numerical procedure maybe needed. Secondly, note that Eq. (2) is a "local" condition regarding x and y, as it only determines $x > y$ or $x < y$ but not necessarily the actual order relationships of all objects from C. A scheme (Fig. 3) may be useful for a better understanding.

Eq. (2) determines the transition from $x < y$ to $x > y$ but not the final position or the final resulting configuration. The number of all configurations is less than 2^U, with U the number of incomparable pairs. A correct application of Eq. (2) has to regard all incomparable pairs and the final configuration must be constructed from all possible outcomes under the constraints of transitivity (see below).

After introducing

$$\Delta_i^{x,y} := q_i(x) - q_i(y) \tag{3a}$$

and

$$Q_{ij}^{x,y} := \frac{\Delta_i^{x,y}}{\Delta_j^{x,y}}. \tag{3b}$$

we find because of the supposed linearity of the aggregation function from Eq. (2):

$$g_1^c(x,y) = \frac{\Delta_2^{x,y}}{\Delta_2^{x,y} - \Delta_1^{x,y}} \tag{4a}$$

or

$$g_1^c(x,y) = \frac{1}{1 - Q_{1,2}^{x,y}}. \tag{4b}$$

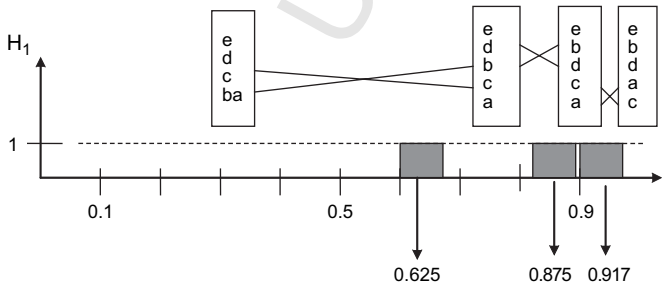


Fig. 5. g-spectrum derived from the data matrix of Table 1.

Table 3
Signs of Δ_i and their possible role if a boundary of a stability field is passed

Case	Δ_1	Δ_2	Δ_3	Δ_4	"Reaction" g_1 small[1)] → large[1)]
1	+	−	+	+	$x\|\|_{\varphi1,\varphi2}y \rightarrow x > {}_{\varphi1,\varphi2}y$
2	+	−	−	−	$x < {}_{\varphi1,\varphi2}y \rightarrow x\|\|_{\varphi1,\varphi2}y$
3	+	−	0	+	$x\|\|_{\varphi1,\varphi2}y \rightarrow x > {}_{\varphi1,\varphi2}y$
4	+	−	0	−	$x < {}_{\varphi1,\varphi2}y \rightarrow x\|\|_{\varphi1,\varphi2}y$
5	+	−	0	0	$x < {}_{\varphi1,\varphi2}y \rightarrow x > {}_{\varphi1,\varphi2}y$

One pair (x, y) is assumed and $\Delta_i = q_i(x) - q_i(y)$.

To simplify the notation we also write gkc if just "crucial weights" are mentioned and $g(k)$ (x, y) (omitting the index "c" for "crucial") if we relate to the subset S_k and the objects x and y.

There are some observations, namely that

1. Crucial values for the weights depend on the pair of objects, whose order relation is to be examined.
2. Crucial weights have only values within the closed interval [0,1] if Q is ≤ 0. If x, y are comparable, Q becomes a positive number and the crucial weight would get values larger than 1. Therefore
3. Eqs. (4a) or (4b) is only meaningful if we discuss incomparable objects.
4. As all incomparable pairs are to be taken into account the set of all crucial weights is important if o-METEOR is to be applied.
5. In o-METEOR the crucial weight of two objects does not depend on the values of other weights. Hence in the product-space of weights, the G-space, each condition of type (4) defines parallel or orthogonal (hyper-) planes. (If only two super-attributes are formed we obtain parallel or orthogonal lines in the g_1, g_2-positive orthant).
6. It is possible that several pairs (x, y), (x', y'), $\ldots \in C \times C$ have the same gkc-values. This is especially the case if the data matrix consists of integers. Hence it is of interest to count the pairs belonging to one numerical value of gkc. The count is summarized by the $H_k(g(k)^c$ function (see the next section, Eq. (7)).
7. One pair (x, y) can only have exactly one gkc-value for a fixed aggregation of the selected attributes in the set S_k. If different attributes are aggregated and the same pair (x, y) is considered, then it can be build a set of {g1c, g2c, …} in the G-space gathering particular gkc for those particular aggregations. In that latter case one pair must have an intersection of several (hyper-) planes. A scheme (Fig. 4) may be helpful to explain this.

In Fig. 4 it is assumed that the incomparabilities of (x', y') and (x, y) resp. are associated with crucial weights for φ_2 (Eq. (1b)). In contrast, the incomparability of (x', y) is related to φ_1 (Eq. (1a)). Note that in Fig. 4 the pair (x, y) has two crucial weights, which necessarily must be assigned to different g-spaces, i.e. to φ_1 and to φ_2.

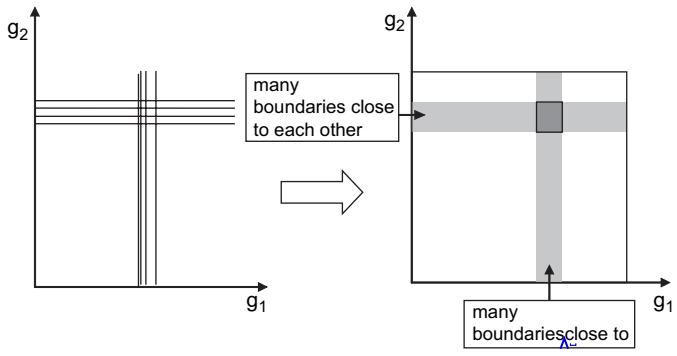


Fig. 6. Instead of a single line several single lines may appear which are close to each other. A hot spot in the G-space at the intersection points (dark rectangle).

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

Table 4
Primary information

Name and abbr.	PV: production volume as score	LC50: Acute fish toxicity [mg/L]	Log Kow	BD: [% degradation/day]
1-chloro-nitrobenzene (CNB)	4	1.5	2.6	0.2
4-nitroaniline (4NA)	2	35	1.4	0
4-nitrophenol (4NP)	1	7	1.9	0.1
Atrazin (ATR)	2	4.3	2.5	0.5
Chlormequat chloride (CHL)	2	80	−2.2	1
Diazinon (DIA)	1	2.6	3.3	0
Dimethoate (DIM)	2	7.5	0.7	0
Ethofumesate (LIN or ETH)	1	11	2.7	0.4
Glyphosphate (GLY)	2	52	0.002	0.3
Isoproturon (ISO)	2	3	2.5	30
Malathion (MAL)	3	0.04	2.7	100
Thiram (THI)	2	0.3	1.7	0

Here it may be a good place to demonstrate the role of Eqs. (2) and (3) by revisiting Fig. 3. In Fig. 3 there are two incomparabilities ($c \,||\, a$) and ($c \,||\, b$). Hence in general we will obtain two crucial weights ${g_1}^c$ (b, c) and ${g_1}^c$ (a, c) (if the simplest case of two attributes is supposed). Starting with the equation for ($c \,||\, b$) it depends on the selected value for the weight whether $c > b$ and hence $c > a$, or $c < b$ is obtained. If $c < b$ is found, we need another information, namely resulting from the Eq. (2) for ($a \,||\, c$). The outcome is once again twofold: $c > a$ or $c < a$, therefore two inequalities imply two equations of type (2), four possible orders for (c, b) and (c, a) resp., but only three final configurations.

2.5. Stability fields and hot spots

It is oversimplified to consider that any pairwise aggregation leads to a well separated set of g_k^c-values. An example of well separated parallel, orthogonal gkc-values can be found in Brüggemann et al. (2006b) (here also an example of non-orthogonal METEOR was given). In order to pave the way of handling the case of many gkc-values, we introduce some further concepts:

If m attributes are pairwise aggregated and those aggregations are disjoint then the G-space has the dimension p.

$$p := \begin{cases} m/2 & \text{if} \quad m = 2*n, \quad n = 1, 2, 3, \ldots \\ (m-1)/2 & \text{if} \quad m = 2*n+1, \quad n = 1, 2, 3, \ldots \end{cases} \tag{5}$$

The calculation of gkc values refers to pairs of objects x, y which are incomparable (see Eq. (2) in Section 2.4). Hence sets of pairs $x||_{qi,qj}y$ play a basic role. Then we call

Table 5
After normalization, orientation and shifting of the data

Chem	PV	LC	Log Kow	BD
CNB	1	0.98	0.87	1
4NA	0.33	0.56	0.65	1
4NP	0	0.91	0.74	1
ATR	0.33	0.95	0.85	1
CHL	0.33	0	0	0.99
DIA	0	0.97	1	1
DIM	0.33	0.91	0.53	1
LIN	0	0.86	0.89	1
GLY	0.33	0.35	0.40	1
ISO	0.33	0.96	0.85	0.7
MAL	0.67	1	0.89	0
THI	0.33	1	0.71	1

Table 6
Pearson correlation matrix

	PV	LC	Log Kow
LC	0.074		
Log Kow	0.000	0.896	
BD	−0.364	−0.251	−0.258

$$\mathrm{IC}_k := \{(x,y) | x, y \in \mathrm{C}, x||_{q_i,q_j} y, q_i, q_j \quad \text{belong to set } k\} \tag{6}$$

the set of incomparable objects given the pairwise aggregation k. As we mentioned above, we indexed k by 1, 2, …, p.

Any pair $(x, y) \in \mathrm{IC}_k$ has exactly one gkc if S_k is held fixed (linearity of the aggregation and conclusion 7). However one value of a crucial weight may represent several pairs like (x, y), (x', y') (Fig. 4). The set of gkc-values in any single g-space can be ordered and we call

$$H_k(g(k)) := |\{(x,y) \in \mathrm{IC}_k \text{ having the same gkc} - \text{value}\}| \tag{7}$$

the "$g(k)$-spectrum". H_k is a function operating on the gkc-values of the kth g-space.

In more technical terms: for the pairs $(x||y) \in \mathrm{C} \times \mathrm{C}$ an equivalence relation $\mathrm{R}(k)$ is introduced as follows:

$$(x||y)\mathrm{R}(k)(x'||y') : \Leftrightarrow g(k)(x,y) = g(k)(x',y'), \quad k - \text{fixed} \in \{1,2,\ldots,p\} \tag{8}$$

Hence given the set $\{(x||y)\} \subset \mathrm{C} \times \mathrm{C}$, then this set is partitioned into k-equivalence classes. Each equivalence class is characterized by one and only one gkc-value. Correspondingly $H_\mathrm{k}(g(k))$ counts the elements of any of the k-equivalence classes and orders them for increasing values of gkc along the g-axis.

Example: we take the data of Table 1, then the dimension of the g-space $= p = 1$. $\mathrm{IC}_1 = \{(b, d), (b, c), (a, c)\}$. In Table 2 the calculation is performed.

Hence formally we can draw a spectrum, as shown in Fig. 5.

Performing the aggregation means that we discuss the order relation as a function of g_1. Clearly $g_1 = 0$ is a projection onto q_2, hence a linear order results: $a < b < c < d < e$. As long as g_1 is less than 0.625 there will be no change. Passing this value a change occurs, which refers to the pair (b, c). The next change in the order relation refers to 0.875, which is assigned to the pair (b, d), finally a change happens when g_1 passes the value 0.917, which is assigned to the pair (a, c). The resulting four linear orders are found in Fig. 5 too. Theoretically five linear orders are possible, however by checking Eq. (2) only four linear orders are obtained. The configuration $e > d > b > a > c$ which is compatible with the partial order shown in Fig. 2, is not obtained (for more details, see Section 4). In order to introduce the concept of stability fields we define.

Definition 6. The G-space is generated by the space of all the weights coming from different aggregations.

Definition 7: Let C be a non-empty object set and IB a set of attributes. The weighted pairwise aggregation of attributes in IB implies that the order relationships of the objects in C change or not, depending on the weights g selected. We call a "stability field" those regions in the G-space where the

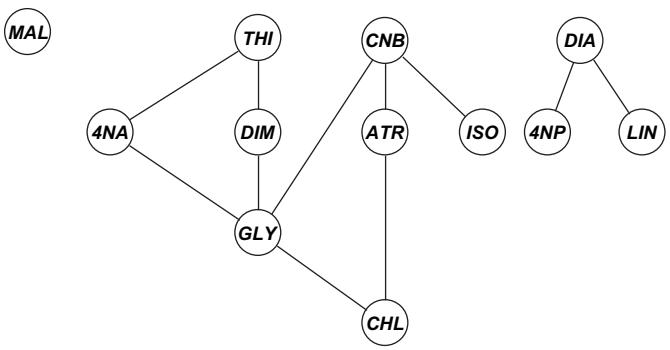


Fig. 7. Hasse diagram (C, IB), $C = \{\mathrm{ATR}, \ldots\}$, IB = {PV, LC50, log Kow, BD}.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

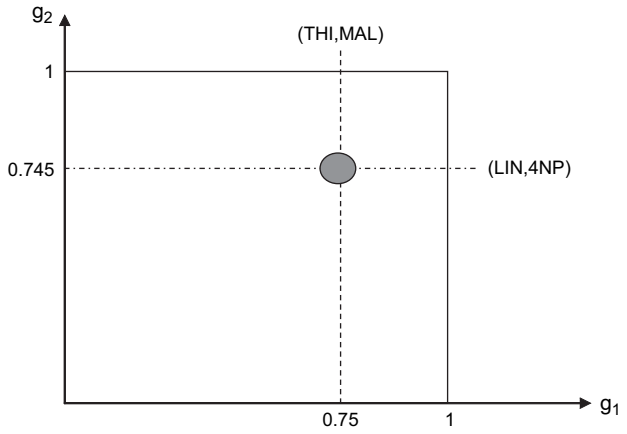


Fig. 8. Two-dimensional *G*-space, and stability fields and hot spots (grey circle).

changes of the weights of aggregation do not change the order relationships of the crucial object-pairs in IC_1, IC_2, …, IC_p.

Definition 8: Stability fields are separated by linear spaces of lower dimensions. In a two-dimensional *G*-space the separating spaces are just lines. The separating linear spaces and their intersections are called "hot spots".

In simple words: hot spots are the regions in the *G*-space where transition from one configuration into another appear; stability fields are those regions in the *G*-space where the configuration of the poset is invariant. In the next sections this is discussed in more detail.

2.6. Change of order relations at crucial g-subspaces

As it was shown in (4) the crucial weights depends on the Δ-values found for all pairs of objects which are incomparable if the original IB is applied.

From the example of Table 3 we deduce that each boundary is to be discussed with respect to

- the pairs of objects belonging to this boundary;
- the reactions (in terms of Table 3) related to each of the (x, y)-pairs.

2.7. Stripes at hot spots as small regions in the G-space

Up to now we discussed some few and well separated gkc-values. If there are N objects then the upper bound for $|IC_k|$ is $N^*(N-1)/2$. Even if $p = 2$ we have to discuss many gkc-values in g_1- and in g_2-direction, corresponding to IC_k. Hence instead of having discrete lines in the *G*-space it may be more convenient to group the lines to stripes as it is shown in Fig. 6.

Therefore it is of primary importance to generalize the concept of the *g*-spectrum: instead of a discrete distribution as formalized by $H_k(g(k))$ (Eq. (7)) one may discuss a quasi-continuous one. A cluster analysis applied to all the gkc-values seems to be the appropriate statistical tool: let us imagine that we consider a clustering for each *g*-space in the o-METEOR on a high similarity level, then a cluster may contain some pairs $\in IC_k$ and each of the pairs is characterized by its gkc-value. Selecting a cluster an interval of gkc-values can be found. Each interval defines a *stripe*. Instead of discrete lines representing discrete gkc-values *transition zones* (consisting of a series of lines close to each other) appear where a small variation of a weight will change the order relations of many (x, y)-pairs. Between the stripes there may be rectangular areas in the *G*-space where the position of originally incomparable objects does not change if weights are varied. Therefore these areas (or hypercubes if $p > 2$) are called as before *stability fields*. Furthermore the bundle of hot spots is no more a bundle of linear spaces of a low dimension but may get as a whole a measure $\neq 0$ in the *G*-space. Nevertheless we call the area defined by a bundle of gkc-values as a whole a "hot spot".

3. Results

3.1. The chemicals

In Table 4 the twelve chemicals are listed up, the attributes are briefly described and the entries are shown. More information about these chemicals, background information to the selection criteria can be found in Lerche et al. (2002b).

3.2. Preprocessing of the data and aggregation

As discussed in Section 2 the data matrix (Table 4) needs several preprocessing steps. The final resulting matrix is shown in Table 5.

The Pearson correlation matrix is shown in Table 6; and we start with those two attributes, which have the highest degree of anti-correlation, that are BD and PV and combine the remaining other two attributes, namely LC50 and log Kow. As the leading principle is to find out the highest degree of anti-correlation which dictates the kind of aggregation, we consider this as a top-down-procedure.

$$\varphi_1 = g_1{}^*\mathrm{PV} + (1 - g_1)^*\mathrm{BD}$$

$$\varphi_2 = g_2{}^*\mathrm{LC50} + (1 - g_2)^*\log \mathrm{Kow}$$

In the first example we discuss two pairs of chemicals out of Table 6.

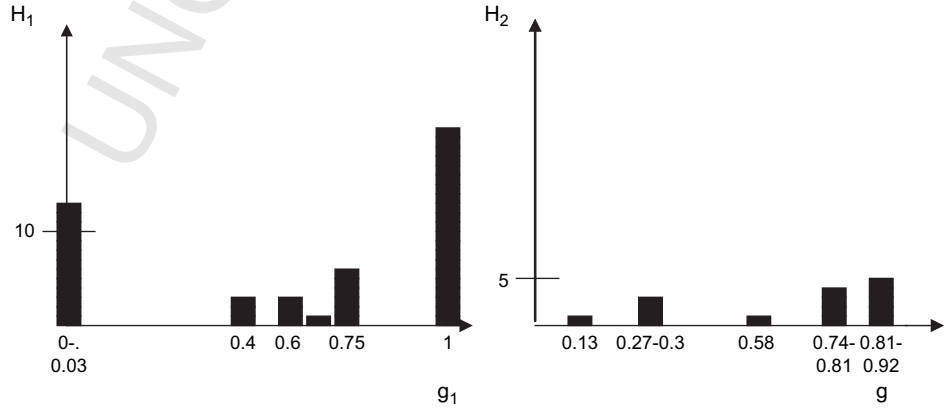


Fig. 9. g_1-spectrum [PV, BD] and g_2-spectrum [LC50, log Kow]. Ordinate axes are H_1 (left side) and H_2 (right side). Abscissa are g_1 and g_2.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

Table 7
g1c and g2c – cluster

Pairs due to g1c-values	Intervals g1c	Pairs due to g2c-values	Intervals g2c
7	0.68…0.75	4	0.74…0.81
6	0.47…0.60	4	0.89…0.92
22	1	1	0.58
13	0	4	0.27…0.30
		1	0.13

If $p = 2$ or greater, then one has to discuss p different weights in p 1-dimensional g-spaces. For example if $p = 2$ a graphic as shown in Fig. 8 may result. For example, exemplifying the procedure we start with two incomparabilities, namely 4NP||LIN and THI||MAL in the original Hasse diagram (Fig. 7).

Regarding 4NP||LIN we may calculate at least one crucial weight, according to (3). If we aggregate as follows: PV and BD on the one side (φ_1) and LC50 and log Kow on the other side (φ_2) then we see that with respect to PV and BD there is no incomparability: LIN $<_{PV,BD}$ 4NP. Then, the incomparability is due to antagonistic attributes (Simon et al., 2004) LC50 and log Kow. For that reason we calculate the crucial weight of LC50 and log Kow for LIN and 4NP. We find g2c = 0.745. For THI and MAL we find: THI$||_{PV,BD}$MAL but THI $<_{LC50,log\ Kow}$MAL. Therefore we associate with the pair (THI, MAL) the crucial weight g1c (separating the g_1-space) and its value turns out to be g1c = 0.75. A graphical representation for the pairs (LIN, 4NP) and (THI, MAL) is shown in Fig. 8.

Within the G-space [0,1] × [0,1] there are four fields, which arise from the separating lines due to g1c = const and g2c = const'. Within these four stability fields (see Section 4 below), the variation of the weights will not affect the order relations between THI-MAL on the one side and LIN-4NP on the other side. Hence we are speaking of a "structure" in the G-space. If $p > 2$ then a similar consideration leads to the generalization of stability (hyper-) cubes. Crossing a boundary (i.e. a line in $p = 2$, or a hyper-plane in $p > 2$) changes the relation for those pairs which belong to the corresponding crucial weight (see (6) in Section 2.4). If the variation of the weights crosses more than one crucial (hyper)plane of gkc-values then correspondingly many pairs of chemicals are affected in their order relations. Therefore crossing of (hyper-) planes are of special interest and are called "hot spots" in the "g-space" as explained in Section 2.5.

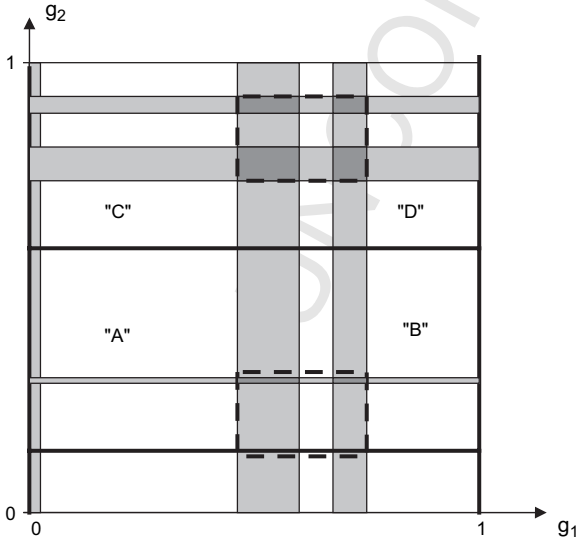


Fig. 10. Transition zones and stability fields for twelve chemicals. A, B, C, D are identifiers of stability fields (see Section 3.3).

3.3. g-spectra, stability fields and hot spots for twelve chemicals

We calculate the two g-spectra and represent them as histograms (see Fig. 9).

By a cluster analysis (complete linkage, squared Euclidian distance) a series of partitionings can be obtained. The cut level is selected so that an aggregation of non-trivial clusters is avoided. The partitioning in case of g1c contained 7, 6, 22, 13 object pairs, in case of g2c 4, 4, 1, 4, 1 object pairs. From any cluster the interval of its crucial g-values is determined. Hence the stability- and transition fields are found as follows (Table 7).

The diagrammatic representation of the results of the cluster analysis (Table 9 and Fig. 9) is shown in Fig. 10: the fields between the stripes are the stability fields. Q2

Fig. 10 shows us that there are 18 stability fields (blank rectangular areas in Fig. 10) which can be characterized by just one Hasse diagram and there are 4 hot spots (dark rectangles in Fig. 10) which may also be merged to bigger hot spots (dashed lines). It should be clear that each transition zone contains a series of small stability fields, which are neglected in the course of the generalization. Variation of weights (by keeping the scheme of pairwise aggregation) will not change the relative positions of incomparable chemicals within a stability field (if the g-spectra allow defining such fields). If however the variation of weights crosses transition zones (the series of crucial gkc-values, grouped to stripes) the order relation of many pairs changes. A perhaps useful picture is that of a phase transition: varying the weights within a stability field the configuration will be invariant, crossing hot spots will change the configuration.

3.4. Typical Hasse diagrams

o-METEOR tends to reduce the problematic work of finding all weights simultaneously by a step-by-step procedure. Taking the data from the publication of Lerche et al. (2002a) the Hasse diagram shown in Fig. 7 and all subsequent Hasse diagrams are obtained. Here we show how after the introduction of the first two weights a set of possible posets will be obtained, so that only 18 typical Hasse diagrams are to be considered.

There are many incomparabilities which hamper a unique decision, albeit one may begin with the maximal elements {MAL, THI, CNB, DIA}. It is not meaningful to show every

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

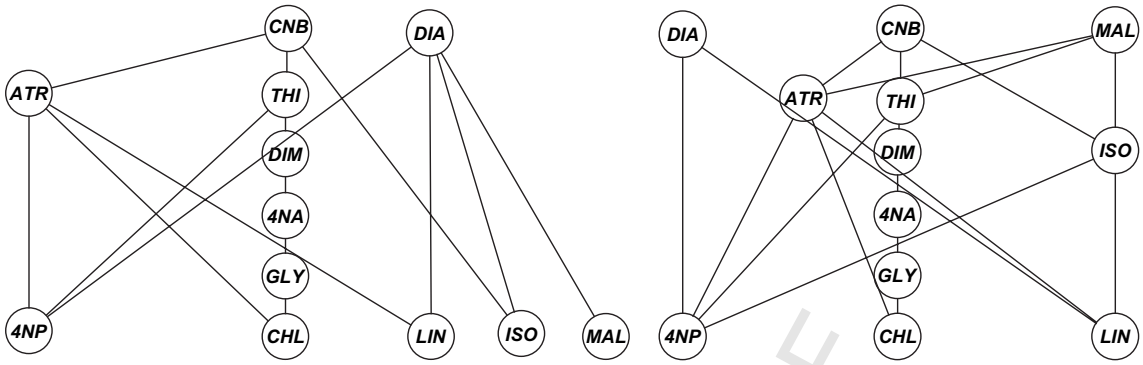


Fig. 11. Hasse diagram in stability field A (left), stability field B (right).

Hasse diagram. We only show the four most important ones; the importance we derive – as shown above – from the volumes of the (hyper-) cubes.

In stability field A ($0.03 < g1c < 0.47$) ($0.30 < g2c < 0.58$) the Hasse diagram is shown in Fig. 11 (left side), in stability field B in Fig. 11 (right side).

It is interesting to note that all three former hierarchies are now related to each other and in case of stability field A: that Malathion (MAL), which is isolated in (C, IB) is now comparable with Diazinon (DIA). This Hasse diagram results with a low weight for PV and a medium weight for LC50.

In the case of stability field B, where PV is considered as very important and BD as less important, Malathion (MAL) becomes a maximal element in the poset and is worse than many other chemicals. Comparing the Hasse diagram of field A with that of field B one observes many changes. This is consistent with the fact that between both stability fields a rather big transition zone is located, which implies that there are many pairs of incomparable elements, changing their order relations within the aggregation. Furthermore it should be noted that there are order preserving maps from the poset, shown in Fig. 7 to the posets shown in Fig. 11, but no order preserving map between the two Hasse diagrams of the stability fields A and B. This is consistent with the finding, summarized in Table 9 and schematically drawn in Fig. 12. In Fig. 13 the Hasse diagrams corresponding to the fields C and D (Fig. 10) are shown.

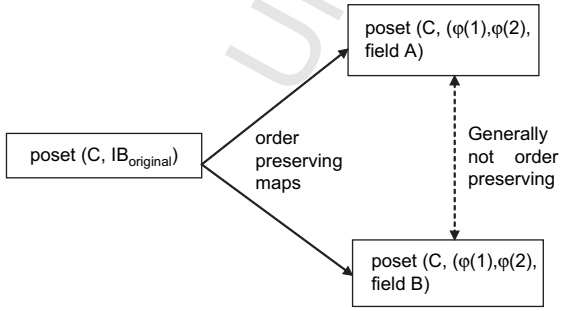


Fig. 12. Relation between original poset and posets after aggregation using different weights.

Once again, passing the large transition zone (from A/C to B/D) implies many changes, however the boundary which separates A from C and B from D is just one chemical pair. Hence it is clear that the vertical transition will only exhibit one change. Namely LIN||THI is converted to THI > LIN and vice versa. It is interesting to note how an aggregation affects the position of the chemical MAL. In the original Hasse diagram (Fig. 7) it is an isolated element, hence any aggregation may change the position of MAL drastically (compare Brüggemann et al., 2001). In stability field C the chemical MAL is a minimal element, whereas in stability field D it is a maximal one. This finding also shows, how crucial a weighting can be and how important it is, to analyze object sets by partial order set theory!

4. Discussion and conclusion

In contrast to its kernel, the HDT, METEOR allows participation of stakeholders and provides the stepwise introduction of weights. The expectation is that often just some few steps will be helpful for the decision (here: which chemical is hazardous to the environment). For example incomparable chemicals as shown in Fig. 7 are now related to each other in a systematic, i.e. order preserving way. The example shows an intermediate state of o-METEOR, namely after introduction of only two weights, whereas for a linear ranking three weights would be needed (weights are normalized so that their sum equals 1). The advantages associated with discrete approaches such as the HDT, which provide high transparency throughout the whole evaluation process is combined with the flexible use of weights, which model the subjective preferences.

From the more mathematical/statistical and software technical point of view there are many questions open, which are to be solved in the future:

1) What is the most efficient strategy for aggregation? Will the bottom-up strategy always be rather inefficient, whereas the top-down strategy resolves in its first steps the most important conflicts? Here it may be useful to discuss the aggregation procedure in terms of conflict diagrams as introduced by Sørensen et al. (2005).

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

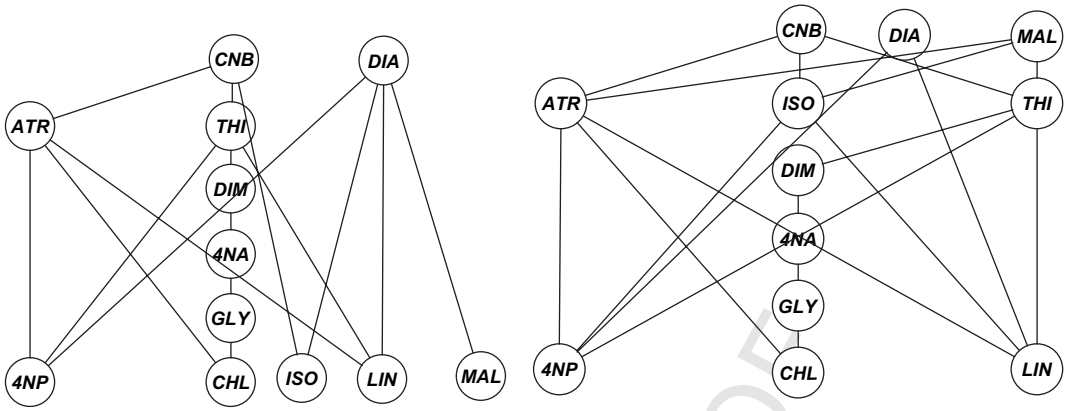


Fig. 13. Left side: Hasse diagrams of stability field C; Right side: stability field D.

2) As discussed in Section 2.4 the Eq. (2) alone leads only to a lower bound of possible order theoretical extensions. In order to obtain all extensions (not necessarily only the linear ones) each configuration must be expressed by a bundle of inequalities describing the order relations. This however, on the one side in large posets is a computational problem and on the other side not all linear extensions will be obtained by just a linear combination of the attributes. Therefore we think and suggest that equations of type (2) (Section 2.4) are a good compromise.
3) Which relations can be found among the posets obtained in intermediate steps of aggregation? Certainly there must be a set of order preserving maps between the original poset and the derived posets due to different aggregations. However the degree of enrichment of comparabilities will be different, as could be seen in Fig. 7 in comparison to the posets shown in Figs. 11 and 13. Hence it is of interest to characterize aggregation schemes at least in terms of similarity within the set of crucial weights.
4) How far the concept of stability fields (or "phases") and phase transitions can be further applied? We have seen that the broadness of transition zones corresponds to the number of changes by which the poset will be affected. Hence the stripes and their geometrical configurations are of main interest.
5) If we do not know any weight. Which stability field should be examined first? The volume (or in technical terms: measure) of the stability hypercubes (here planes) may give a useful advice. Here the largest stability field is field A: $(0.47 - 0.03) * (0.58 - 0.30 = 0.132)$ followed by field B: $(1 - 0.75) * (0.58 - 0.30) = 0.07$. Therefore we started in Section 3.3 with the largest stability field "A" by determining the Hasse diagrams and continue with three stability fields of lower measures.
6) What's about the generalization to $p > 2$. In this paper we exemplified the ideas by $p = 2$, many graphical schemes are based on a two-dimensional representation. In real life p will be by far greater than 2. If the o-METEOR approach is followed then p one-dimensional g-spaces are to be calculated and characterized by their gkc-values. If more general aggregations (including two- or higher dimensional g-spaces) are to be used, then the transition zones have to be calculated and it is more difficult to present them graphically.
7) Can we always expect stability fields? Yes and no! Clearly we get with a finite set of objects (i.e. of chemicals) only discrete sets of crucial weights. Hence one may find within two adjacent crucial weights always a more or less large field of invariant $<$ or $>$-relations. However, if the crucial weights in all dimensions are approximately homogenously filling out the interval [0,1] in any g-space then any small change of weights everywhere in the G-space will lead to a phase transition. In that case one may perform a classification of the original attributes into scores or define an aggregation which does not pairwise combine the original attributes. This however, leads to a theoretically more complex system, which is still open for further research.
8) METEOR may be seen as one example of the g-posets. We speak of g-posets, if the attributes by which a componentwise order is defined, are dependent on a set of continuous varying parameters. If for example a poset and its visualization by Hasse diagrams is used to exhibit structure-fate relations of chemicals in the environment (Brüggemann et al., 2006a) then a natural question arises how the poset depends on environmental parameters, if the characterizing attributes are fate descriptors. In METEOR and especially in o-METEOR the relations can be considered as relatively simple because of the linearity of the φ-functions with respect to the weights. In the case of structure–fate relationships the descriptors in general will non-linearly depend on parameters like water discharge, organic carbon content etc. First attempts are under development (Restrepo et al., 2007b). However the field of g-posets needs the joint work of many scientists in the future. We hope that in the series of workshops about partially ordered sets in chemistry and environmental sciences (initialized by the first author) this joint cooperation can be enhanced.
9) o-METEOR is intended to serve as a decision support tool for environmental sciences too. However, before

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

we recommend its application as a decision support tool we have to compare it with several other well-known procedures, e.g. PROMETHEE. That means we have to continue the work begun by Lerche et al. (2002a) where HDT was compared with four other decision support tools.

Appendices

Appendix 1

a. Abbreviations (alphabetically sorted)

Abbreviation	Explanation
AHP	Analytic hierarchy process
Bottom-up-procedure	Starting from very detailed information and aggregate those attributes, which belong to the same subcriterion
ELECTRE	(French): Elimination et Choix Traduisant la Realité
Extremal case procedure	The aggregation procedure in METEOR will be stopped if a) a greatest or least element is found or b) if two elements of interest can be compared.
HD	Hasse diagram
HDT	Hasse diagram technique
HPVC	High production volume chemicals
MAUT	Multiattribute utility function theory
MCDA	Multi-criteria decision aids
METEOR	Method of evaluation by order theory
NAIADE	Novel approach to imprecise assessment and decision environment
o-METEOR	Like METEOR, however a specific aggregation scheme
poset(s)	Partially ordered set(s)
PROMETHEE	Preference ranking organisation method for enrichment evaluation
REACH	Registration, evaluation, authorisation of chemicals
Top-down-procedure	Aggregate that pair of attributes which is most anticorrelated, then the next pair of attributes, etc.
WHASSE	Hasse for Windows

b. Symbols and concepts

Symbol	Explanation	Remarks
g_k^c	Crucial weight	Eq. (4) (often simply written gkc)
$\Delta_i^{x,y}$	$=q_i(x) - q_i(y)$	
$Q_{ij}^{x,y}$	$= \Delta_i^{x,y} / \Delta_j^{x,y}$	
φ_k	Weighted sum of $q_i \in k$	
$\|\|$	Sign to denote incomparability	
BD	Biodegradation	
C	Set of objects	In Section 3: set of Chemicals
g_i	The weights	They are representing in this study the external knowledge
g-space	$g \in [0,1]^p \subset IR^p$	IR^p set of p-tuples of real numbers. In o-METEOR the g-spaces are one-dimensional
G-space	Product of all g-spaces	$\subset IR^m$
$H_k(g(k))$	Number of pairs of ICk, having the same gkc-value	g-spectrum: H_k versus $g_i \in [0,1]$
Hot spots	Subspaces of the G-space where a change of weights changes the relative positions of two incomparable objects	
IB	Set of attributes, characterizing objects in order to perform an evaluation	
IC_k	Set of pairs of incomparable objects, due to set k	Equation 6
LC50	The dose of a substance which is fatal to 50% of the test animals	
log Kow	log of n-octanol/water partitioning coefficient	
m	Number of attributes	$m =$ card IB
N	Number of objects	$N =$ card C
p	Dimension of the G-space	
PV	production volume	
q_i	ith attribute	The attributes q_i represent in our study the primary knowledge
$q_i(x)$	The value of the ith attribute for object x	
q_i(max)	The maximum value of q_i within a set of objects	
q_i(min)	The minimum value of q_i within a set of objects	

(continued on next page)

Acknowledgments

G. Restrepo thanks COLCIENCIAS for the PhD grant given during this research. Special thanks are offered to the Universidad de Pamplona in Colombia and Dr. A. González, rector of that University, for their financial support.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

Appendix 1 (*continued*)

Symbol	Explanation	Remarks
S_k	A set of those attributes which are to be combined by a weighted sum	
Stability fields	Subspaces of the *G*-space where a change of weights does not change the relative positions of two incomparable objects	
Stripes	If gkc-values are close to each other one may define an interval and represent all these values by just one area in the *G*-space, where-crossing this area by variation of weights-many changes in the order relations appear	Also called a transition zone

Appendix 2

Given the set IB of m attributes the total number of possible subsets of it is given by the cardinality of the power set of IB (P(IB)), which corresponds to 2^m. We write 2^m-m-2 since we do not consider the original attributes in IB either the empty set either the subset containing all the attributes as simultaneously aggregated. If we suppose IB $= \{q_1, q_2, q_3\}$ then the possible number of subsets of those three attributes is $2^3 = 8$, which are P(IB) $= \{\{q_1\}, \{q_2\}, \{q_3\}, \{q_1, q_2\}, \{q_1, q_3\}, \{q_2, q_3\}, \{q_1, q_2, q_3\}, \varnothing\}$. From P(IB) just $\{q_1, q_2\}$, $\{q_1, q_3\}$, and $\{q_2, q_3\}$ can be considered as aggregation of the attributes q_1, q_2 and q_3. Hence, their number is 2^3-3-2 $= 3$. Note, however that these three possible aggregations are not disjoint.

Appendix 3

Given the set IB of m attributes, a partition D of IB is a collection of subsets of IB such that:

i) If $S_1, \ldots, S_m \subset$ IB, then $\bigcap_{k=1}^{m} S_k = \varnothing$

ii) If $S_1, \ldots, S_m \subset$ IB, then $\bigcup_{k=1}^{m} S_k =$ IB

Thus, if we IB $= \{q_1, q_2, q_3\}$ we have the following partitions: D1 $= \{q_1, q_2, q_3\}$, D2 $= \{\{q_1\}, \{q_2, q_3\}\}$, D3 $= \{\{q_2\}, \{q_1, q_3\}\}$, D4 $= \{\{q_3\}, \{q_1, q_2\}\}$ and D5 $= \{\{q_1\}, \{q_2\}, \{q_3\}\}$.

Appendix 4

The number of ways a set IB of m attributes can be partitioned into k non-empty sets is $S_2(m, k)$, which is called a Stirling number of the second kind.

$$S_2(m,k) = (1/k!)\sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^m.$$

If IB $= \{q_1, q_2, q_3\}$ and we decide to aggregate its elements in two classes, then $S(3,2) = 3$ and the partitions are D2, D3 and D4 form A2.

Appendix 5

The number of ways a set IB of m attributes can be partitioned into non-empty subsets is called a Bell's number ($B(m)$). $B(m) = \sum_{k=1}^{m} S_2(m,k)$, being $S_2(m, k)$ defined in A4. If IB $= \{q_1, q_2, q_3\}$, then $B(3) = 5$ and the partitions appear in A2.

References

Berlekamp, J., Lautenbach, S., Graf, N., Reimer, S., Matthies, M., 2007. Integration of MONERIS and GREAT-ER in the decision support. Environ. Model. Software 22, 239–247.

Bock, H.H., 1974. Automatische Klassifikation. Vandenhoeck&Ruprecht, Göttingen. 6–480.

Brans, J.P., Vincke, P.H., 1985. A preference ranking organisation method (The PROMETHEE method for multiple criteria decision-making). Manag. Sci. 31, 647–656.

Brüggemann, R., Carlsen, L., 2006. Partial Order in Environmental Sciences and Chemistry. Springer-Verlag, Berlin. 1-406.

Brüggemann, R., Drescher-Kaden, U., 2003. Einführung in die modellgestützte Bewertung von Umweltchemikalien – Datenabschätzung, Ausbreitung, Verhalten, Wirkung und Bewertung. Springer-Verlag, Berlin. 1-519.

Brüggemann, R., Voigt, K., 1995. An evaluation of online databases by methods of lattice theory. Chemosphere 31 (7), 3585–3594.

Brüggemann, R., Welzl, G., 2002. Order theory meets statistics-Hasse diagram 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., Münzer, B., Halfon, E., 1994. An algebraic/graphical tool to compare ecosystems with respect to their pollution – the German River 'Elbe' as an example – I: Hasse-Diagrams. Chemosphere 28, 863–872.

Brüggemann, R., 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., Sørensen, P.B., Lerche, D., Carlsen, L., 2004. Estimation of averaged ranks by a local partial order model. J. Chem. Inf. Comp. Sci. 44, 618–625.

Brüggemann, R., Simon, U., Mey, S., 2005. Estimation of averaged ranks by extended local partial order models. Match – Comm. Math. Co. 54, 489–518.

Brüggemann, R., Restrepo, G., Voigt, K., 2006a. Structure-fate relationships of organic chemicals derived from the software package E4CHEM and WHASSE. J. Chem. Inf. Model 46 (2), 894–902.

Brüggemann, R., Simon, U., Nützmann, G., 2006b. Analyzing water management strategies in urban regions by directed graphs. In: Studzinski, J., Hryniewicz, O. (Eds.), Modelling Concepts and Decision Support in Environmental Systems, Vol. 45. Polish Academy of Science, Warsaw, pp. 111–124.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

Castelletti, A., Soncini-Sessa, R., 2006. A procedural approach to strengthening integration and participation in water resource planning. Environ. Model. Software 21, 1455–1470.

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

EEC, 2001. WHITE PAPER, Strategy for a Future Chemicals Policy, Brussels, 27.2.2001, COM(2001) 88 final. http://www.reachcentrum.eu/media/whitepaper.pdf.

European Commission, 2006. REACH in brief, September 2006, http://ecb.jrc.it/DOCUMENTS/REACH/REACH_in_brief_council_comm_pos_060905.pdf.

Giupponi, C., 2007. Decision Support Systems for implementing the European Water Framework Directive: The MULINO approach. Environ. Model. Software 22, 248–258.

Klauer, B., Messner, F., Drechsler, M., Horsch, H., 2001. Das Konzept des integrierten Bewertungsverfahrens. In: Horsch, H., Herzog, F. (Eds.), Nachhaltige Wasserbewirtschaftung und Landnutzung. Methoden und Instrumente der Entscheidungsfindung und Umsetzung. Metropolis, Marburg, pp. 75–99.

Lerche, D., Brüggemann, R., Sørensen, P.B., Carlsen, L., Nielsen, O.J., 2002a. A comparison of partial order technique with three methods of multicriteria analysis for ranking of chemical substances. J. Chem. Inf. Comp. Sci. 42, 1086–1098.

Lerche, D., Sørensen, P.B., Larsen, H.L., Carlsen, L., Nielsen, O.J., 2002b. Comparison of the combined monitoring – based and modelling – based priority setting scheme with partial order theory and random linear extensions for ranking of chemical substances. Chemosphere 49, 637–649.

Lerche, D., Sørensen, P.B., Brüggemann, R., 2003. Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual probability. J. Chem. Inf. Comp. Sci. 53, 1471–1480.

Makropoulos, C.K., Butler, D., 2006. Spatial ordered weighted averaging: incorporating spatially variable attitude towards risk in spatial multi-criteria decision-making. Environ. Model. Software 21, 69–84.

Matarazzo, B., Munda, G., 2001. New approaches for the comparison of L-R fuzzy numbers: a theoretical and operational analysis. Fuzzy Sets Syst. 118, 407–418.

Ocampo-Duque, W., Schuhmacher, M., Domingo, J.L., 2007. A neural-fuzzy approach to classify the ecological status in surface waters. Environ. Pollut. 148, 634–641.

Reichert, P., Borsuk, M., Hostmann, M., Schweizer, S., Spoerri, C., Tockner, K., Truffer, B., 2007. Concepts of decision support for river rehabilitation. Environ. Model. Software 22, 188–201.

Restrepo, G., Brüggemann, R., Weckert, M., Gerstmann, S., Frank, H., 2007a. Paper under preparation.

Restrepo, G., Brüggemann, R., Voigt, K., 2007b. Partially ordered sets in the analysis of alkanes' fate in rivers. Croatia Chem. Acta 80 (2), 261–270.

Roy, B., 1990. The outranking approach and the foundations of the ELECTRE methods. In: Bana e Costa, C.A. (Ed.), Readings in Multiple Criteria Decision Aid. Springer, Berlin, pp. 155–183.

Saaty, T.L., 1994. How to make a decision: the analytical hierarchy process. Interfaces 24, 19–43.

Schneeweiss, C., 1991. Planung 1 – Systemanalytische und entscheidungstheoretische Grundlagen. Springer, Berlin.

Simon, U., Brüggemann, R., Pudenz, S., 2004. Aspects of decision support in water management – example Berlin and Potsdam (Germany) I – spatially differentiated evaluation. Water Res. 38, 1809–1816.

Simon, U., Brüggemann, R., Mey, S., Pudenz, S., 2005. METEOR – application of a decision support tool based on discrete mathematics. Match – Comm. Math. Co. 54, 623–642.

Sørensen, P.B., Brüggemann, R., Thomsen, M., Lerche, D., 2005. Applications of multidimensional rank-correlation. Match – Comm. Math. Co. 54, 643–670.

Vink, J.P.M., Meeussen, J.C.L., 2007. BIOCHEM-ORCHESTRA: a tool for evaluating chemical speciation and ecotoxicological impacts of heavy metals on river flood plain systems. Environ. Pollut. 148, 833–841.

Voigt, K., Brüggemann, R., 2005. Water contamination with pharmaceuticals: data availability and evaluation approach with Hasse diagram technique and METEOR. Match – Comm. Math. Co. 54, 671–689.

Voigt, K., Brüggemann, R., Pudenz, S., 2004a. Chemical databases evaluated by order theoretical tools. Anal. Bioanal. Chem. 380, 467–474.

Voigt, K., Welzl, G., Brüggemann, R., 2004b. Data analysis of environmental air pollutant monitoring systems in Europe. Environmetrics 15, 577–596.

Voigt, K., Brüggemann, R., Pudenz, S., 2006. A multi-criteria evaluation of environmental databases using the Hasse diagram technique (ProRank) software. Environ. Model. Software 21, 1587–1597.

Znidarsc, M., Bohanec, V., Zupan, B., 2006. ProDEX – A DSS tool for environmental decision-making. Environ. Model. Software 21, 1514–1516.

Please cite this article in press as: Brüggemann, R. et al., Concept of stability fields and hot spots in ranking of environmental chemicals, Environ. Model. Softw. (2007), doi:10.1016/j.envsoft.2007.11.001

Appendix G

Ranking patterns, an application to refrigerants

Guillermo Restrepo[a,b], Rainer Brüggemann[c], Monika Weckert[a], Silke Gerstmann[a] and Hartmut Frank[a]

[a] Environmental Chemistry and Ecotoxicology, University of Bayreuth, Bayreuth, Germany
E-mail: guillermo.restrepo@uni-bayreuth.de, monika.weckert@uni-bayreuth.de, silke.gerstmann@uni-bayreuth.de, hartmut.frank@uni-bayreuth.de

[b] Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia
E-mail: grestrepo@unipamplona.edu.co

[c] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
E-mail: brg_home@web.de

(Received October 31, 2007)

Abstract

Ranking methods are of great importance for assessing the relative importance or potential impact of different objects, e.g. chemicals, screening methods, etc. When a ranking procedure is applied it is desired to obtain a total order of the ranked objects. With most methods, this is achieved through the mathematical combination of descriptors characterising the objects, often under the inclusion of descriptor priorities. However, such priority settings affect the final ranking and the contribution of each descriptor to the final result becomes obscure. In this paper, METEOR (*Met*hod of *e*valuation by *or*der theory) is described, a ranking procedure which allows to explore the complete space of possible descriptor priorities in such a way that total orders are obtained. METEOR permits 1) to study the total order resulting from any descriptors' prioritisation, 2) to determine the priorities necessary to obtain a particular total order, 3) to calculate the probability of having a particular total order, and 4) to calculate the similarity between different total orders.

METEOR is applied to 18 refrigerants used in the past, presently used, and some proposed substitutes, characterised by their ozone depletion potentials, global warming potentials and atmospheric lifetimes. The results show that pentafluorodimethyl ether, a proposed replacement for the problematic fully halogenated refrigerants, has a probability of 68 % of being an environmentally problematic substance of the selection of refrigerants considered in this paper.

Introduction

Ranking is the process of positioning elements of a set on an ordinal scale in relation to each other. There are different ranking methods[1], developed for priority setting in decision making processes. Their applications cover areas such as the assessment of the performance of health systems[2], the selection of biodiverse ecosystems for governmental protection[3], and even the assessment of research institutions, scientists[4,5], and publications[6]; the results can have direct impacts on governmental research budgets[7]. Ranking processes are important in document retrieval[8,9] and are behind the search engines employed to gather relevant information from the World Wide Web[8,10] such as the PageRank[11] of Google[12].

In chemistry, ranking is used for chemical information retrieval[9,13]; different algorithms and procedures have been developed to manage large chemical data sets. Another application is the interpretation of spectra[14]; thereby, for a given set of spectra the possible substances are ranked according to their degree of spectra fitting. Ranking is also used in lead-discovery programs in virtual screening procedures[15,16], in order to rank molecules for running biological assays[15]; others are employed in lead-discovery processes for the selection of molecular descriptors for potential drugs[16]. In the latter cases, rankings are performed on pools of descriptors characterising the substances to select those which are ranked highest. Ranking procedures are also important for risk assessment studies in environmental sciences[17-19].

In general, a ranking procedure can be summarised as shown in Figure 1; the first step consists in collecting the elements and forming the dataset; the second step is the selection of descriptors characterising the objects; in the third step the ranking performance is considered, which can be done in principle by two ways: Including additional information (priorities) as shown in Figure 1 or trying to deduce a linear order from the properties of the partially

ordered set alone. The second possibility is not followed in this paper but it is described in reference 20. As final step the ranking is interpreted and applied. Each of these steps includes different procedures and decisions, for example the importance of diverse sets of chemicals for lead-discovery procedures, the kind of descriptors characterising the molecules in respect to the aim of the ranking, the advantages and disadvantages of different algorithms, and the design of rules for interpretation of the results for application.

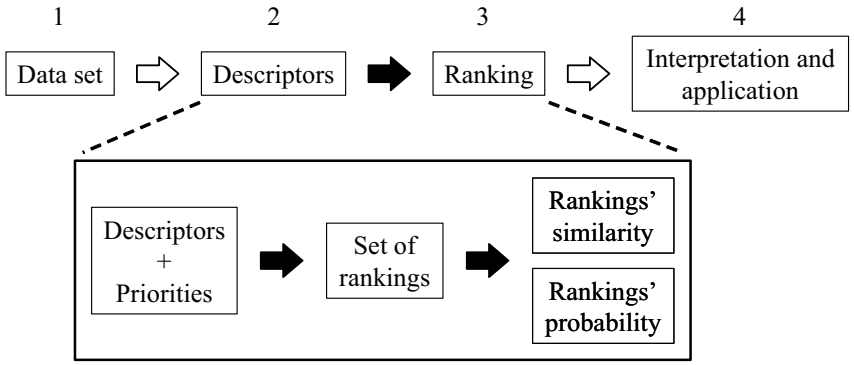


Figure 1. Ranking procedure. The bold box represents the scope of the present research.

Although there are several studies[1,21] on the individual steps of a ranking procedure (Figure 1) and their influence on the final ranking, the relationship between steps 2 and 3 has not been studied in respect to descriptor prioritisation. This is crucial because normally it is necessary to assess the importance of each descriptor for the final ranking, for example assessing the weight of exposure relative to effect descriptors in an environmental study. In this paper, METEOR[22-24], a mathematical method for assessing descriptor prioritisation and its effect on the ranking, is described. A methodology for calculating similarity among rankings obtained by different descriptor prioritisation is introduced. This allows to answer the following questions: How is the ranking affected when descriptor priorities change systematically? What is the probability of arriving at a certain ranking through descriptor prioritisation?

The current paper is organised as follows: After ranking objects using partial order theory, the concept of step-by-step aggregation is introduced. Hence, order preserving maps among partial orders are obtained. Further application of the step-by-step aggregation leads to linear orders and a similarity analysis of them is performed.

Ranking objects

A ranking method including the steps shown in Figure 1 should be flexible to permit "researcher participation"[25]. Also, one must cope with the fact that conflictive values may occur among the selected descriptors. This latter situation arises when some descriptors of an object have high values while others are low. "Researcher participation" means that subjective descriptor preferences are involved in the ranking. To avoid conflictive values, in some ranking procedures the aggregation of all descriptors is done at once[26-28] to yield a total order[29] (linear order) of the objects and to allow the identification of a single high-ranking object. Such aggregations may include descriptor priorities weights, representing the "researcher participation" in the process. Aggregation, however, entails descriptor compensation since a low value in one particular descriptor offsets large values in others. As each descriptor represents a particular aspect of the objects, compensation is regarded as comparison of "chalk and cheese"[24]. In such a situation, ranking interpretation becomes questionable as the descriptors' influences are hidden because compensation takes place over all descriptors simultaneously, often referred to as "weighting camouflage" in the ranking process[30]. To evade this problem, the Hasse Diagram Technique (HDT)[18,19,31-33] can be used to explore different aspects of the ranked set, such as ranking stability under addition and deletion of descriptors[34] and their influence on the ranking[17,31]. In the following, a brief description is given.

Hasse Diagram Technique (HDT)

In the HDT, different descriptors q_1, q_2, ..., q_i are simultaneously used to rank the objects *a*, *b*,… of a set *P*. As a methodological condition, all descriptors need to be oriented[17] in such a way that low descriptor values indicate low ranking and high values indicate high ranking. If an object $x \in P$ is characterised by the descriptors $q_1(x)$, $q_2(x)$, ..., $q_i(x)$ and another object $y \in P$ by $q_1(y)$, $q_2(y)$, ..., $q_i(y)$, *x* and *y* are compared by contrasting their individual descriptors. If all descriptors of *x* are higher or equal to those of *y* ($q_i(x) > q_i(y)$ for all *i*) or at least one descriptor is higher for *x* while all the others are equal ($q_j(x) > q_j(y)$ for some *j* and $q_i(x) = q_i(y)$ for the rest of descriptors), *x* is ranked higher than *y* ($x \geq y$). In this case, *x* and *y* are said to be comparable. If $q_i(y) = q_i(x)$ for all *i*, then *x* and *y* have identical rank and become equivalent objects ($x \sim y$). When at least one property fulfils $q_j(x) < q_j(y)$ while the others follow the relation $q_i(x) \geq q_i(y)$, *x* and *y* are called incomparable ($x \parallel y$); thus, they are not ordered with

respect to each other. Normally, several objects are mutually incomparable, and P is not totally but partially ordered[29] and is called a *p*artially *o*rdered *set* (poset)[29]. This can be represented as a directed acyclic graph whose vertices are the objects in P, and each edge represents the comparability among the linked objects; higher-ranked objects are given a higher vertical position[29]. Because the graph often contains edges for each pair of objects, it may contain trivial relations (i.e. if $x \geq y$ and $y \geq z$ then $x \geq y \geq z$) which can be simplified by drawing next neighbour edges only; such parsimonious graph is known as Hasse diagram (HD)[33]. Figure 2 depicts the HD of the set $P = \{a, b, c, d, e\}$ resulting from descriptors gathered in the corresponding data matrix.

HDT makes the ranking process transparent, besides other advantages[17-19,31-33]. On the other hand, the lack of descriptor weights is seen as an absence of researcher participation on the process and is usually regarded as disadvantage[24], stressed by the fact that several high-ranked objects may coexist (e.g. *b* and *d* in Figure 2). This is caused by the absence of aggregating functions, with weights as parameters, which would allow to remove conflictive descriptor values. Because of this disadvantage, Brüggemann and coworkers[22-24] developed METEOR (*Met*hod of *e*valuation by *or*der theory) as an extended procedure which permits to solve the dilemma by obtaining a single high-ranked object, keeping the HDT transparency and allowing "researcher participation". Contrary to other ranking methods such as PROMETHEE[27], NAIADE[35] or AHP[36] where descriptor weighting-aggregation is carried out in one step, with METEOR it is performed in a step-by-step procedure permitting to discern the effects of descriptor weights and their compensations. Compensation is restricted to the descriptors which are actually aggregated in a subset, and the effect of all possible weights on the ranking can be systematically studied. METEOR permits to find total orders which arise from the original HDT and which are called linear extensions[18,29]. In the following, these total orders are briefly discussed.

Linear extensions

A partial order corresponds to a HD with incomparable objects; since the majority of ranking methods is directed towards total orders, it is important to relate the concept of linear extension to that of partial order.

A linear extension is a projection of a partial order into a total order, keeping the order relations of the partial order. In mathematical terms it is an order-preserving mapping of a partial order[29]. The incomparabilities of a HD must be changed to comparabilities to obtain a linear extension (Figure 2, upper right). The total set of linear extensions can be found by a combinatorial procedure where the incomparable objects are systematically given an order with respect to each other[37].

For a set of linear extensions, the ranking frequency[25] r_{mn} can be calculated as the occurrence of object n at the rank m; additionally, r_{mn} divided by the number of linear extensions yields the ranking probability[25] p_{mn} of having n at the rank m. Finally, $\overline{r}_n$, the average rank[25] of n, can be calculated as

$$\overline{r}_n = \sum_m m \cdot p_{mn} \qquad (1)$$

The parameters r_{mn}, p_{mn} and $\overline{r}_n$ are also shown in Figure 2 (bottom half)[38]. The $\overline{r}_n$ values can be used to draw a consensus ranking, the most probable ranking for the set of linear extensions considered (Figure 2, bottom right). Note that a linear consensus ranking is only obtained when only one $\overline{r}_n$ corresponds to each $x \in P$; for example b and d, related by an automorphism[29], have the same value of $\overline{r}_n$ (4.5) and are therefore equivalent objects in the consensus ranking. Such equivalent rankings can be avoided by descriptor weighting and a linear order may be obtained, as is shown in the following.

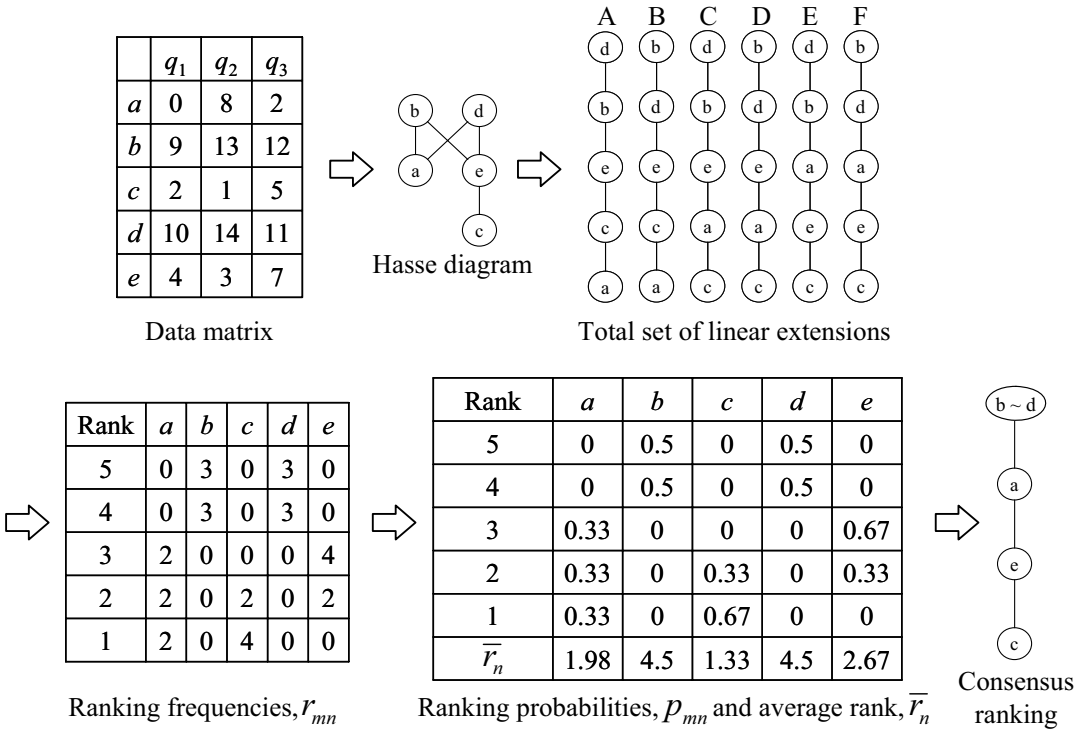


Figure 2. A Hasse diagram of five elements based on three descriptors (Data matrix), their linear extensions, ranking probabilities of each element, and the consensus ranking.

METEOR

Descriptors must be normalised to avoid dimensional conflicts when they are aggregated by using METEOR[22-24]. In the present work, the normalised *i* descriptor value of *x*, i.e. $q_i(x)$, can be calculated as follows[39]

$$q_i(x) = \frac{q_i'(x) - \min q_i'}{\max q_i' - \min q_i'} \qquad (2)$$

where $q_i'(x)$ is the value of descriptor *i* for *x*, and $\min q_i'$ and $\max q_i'$ are the minimum and maximum values.

The step-by-step aggregation using positive monotonous functions (like the linear ones with positive weights) performed by METEOR may be carried out until all descriptors are aggregated in a weighted sum. The final ranking is always a linear extension of the original

HD, i.e. each aggregation keeps the comparabilities of the previous ranking. Hence, the effect of aggregation is to add new comparabilities to those already existing.

This is exemplified by application to the data set shown in Figure 2. In general, incomparabilities in the original HD will be changed into comparabilities by descriptor aggregations; the stepwise aggregation performed for a subset of descriptors enriches the partial order by new comparabilities. One possibility is to group similar descriptors into an aggregated one, another is to aggregate descriptors with a high degree of conflictive potential which are normally anticorrelated[40]. In the example, the conflictive descriptors q_2 and q_3 are aggregated which are least correlated according to the Spearman's rank correlation[41] $\rho = 0.6$. In the data matrix (Figure 2), the three incomparabilities $a \parallel c$, $b \parallel d$, and $a \parallel e$ are due to conflictive values between q_2 and q_3. If $x \parallel y$, there is an aggregated property $\varphi(x)$ for x and another $\varphi(y)$ for y which can be defined as follows:

$$\varphi(x) = g \cdot q_2(x) + (1-g) \cdot q_3(x) \qquad (3)$$

$$\varphi(y) = g \cdot q_2(y) + (1-g) \cdot q_3(y) \qquad (4)$$

where g and $(1 - g)$ are the selected weights (priorities) for q_2 and q_3, respectively; the sum of weights must be equal to 1. An important value of g is achieved when $\varphi(x) = \varphi(y)$ indicating that the incomparability between x and y is turned into an equivalence $x \sim y$. This particular g value is called "crucial weight"[40] for the pair $\{x, y\}$ and is represented by g_c as deduced from Eqs. 3 and 4:

$$g_c = \frac{1}{1 - \dfrac{q_2(x) - q_2(y)}{q_3(x) - q_3(y)}} \qquad (5)$$

with $q_3(x) - q_3(y) \neq 0$. The g_c values for the three incomparabilities $a \parallel c$, $b \parallel d$, and $a \parallel e$ accounted by conflictive values between q_2 and q_3 are 0.357, 0.556 and 0.562, respectively. The change in the order relations between each of the three incomparable pairs can be seen in Figure 3a, where four new HDs are shown based upon q_1 and φ as descriptors. A weight $0 < g < 0.357$ always yields the partial order S1, and if the weight is shifted to $0.357 < g < 0.556$, S2 is obtained. Note that $a < c$ in S1, which turns to be $a \parallel c$ when passing the crucial value g_c

= 0.357. In the same way, a value of $0.556 < g < 0.562$ produces S3, and a value $0.562 < g < 1$ yields S4. The only change between S2 and S3 concerns the pair $\{b, d\}$ while the transition from S3 to S4 changes the order relation of the pair $\{a, e\}$. It can be seen that the transformations S4 $\rightarrow$ S3 $\rightarrow$ S2 $\rightarrow$ S1 are always order-preserving. It is important to note that each passage of the crucial value affects only the incomparable pair generating it while the other pairs keep their mutual relations.

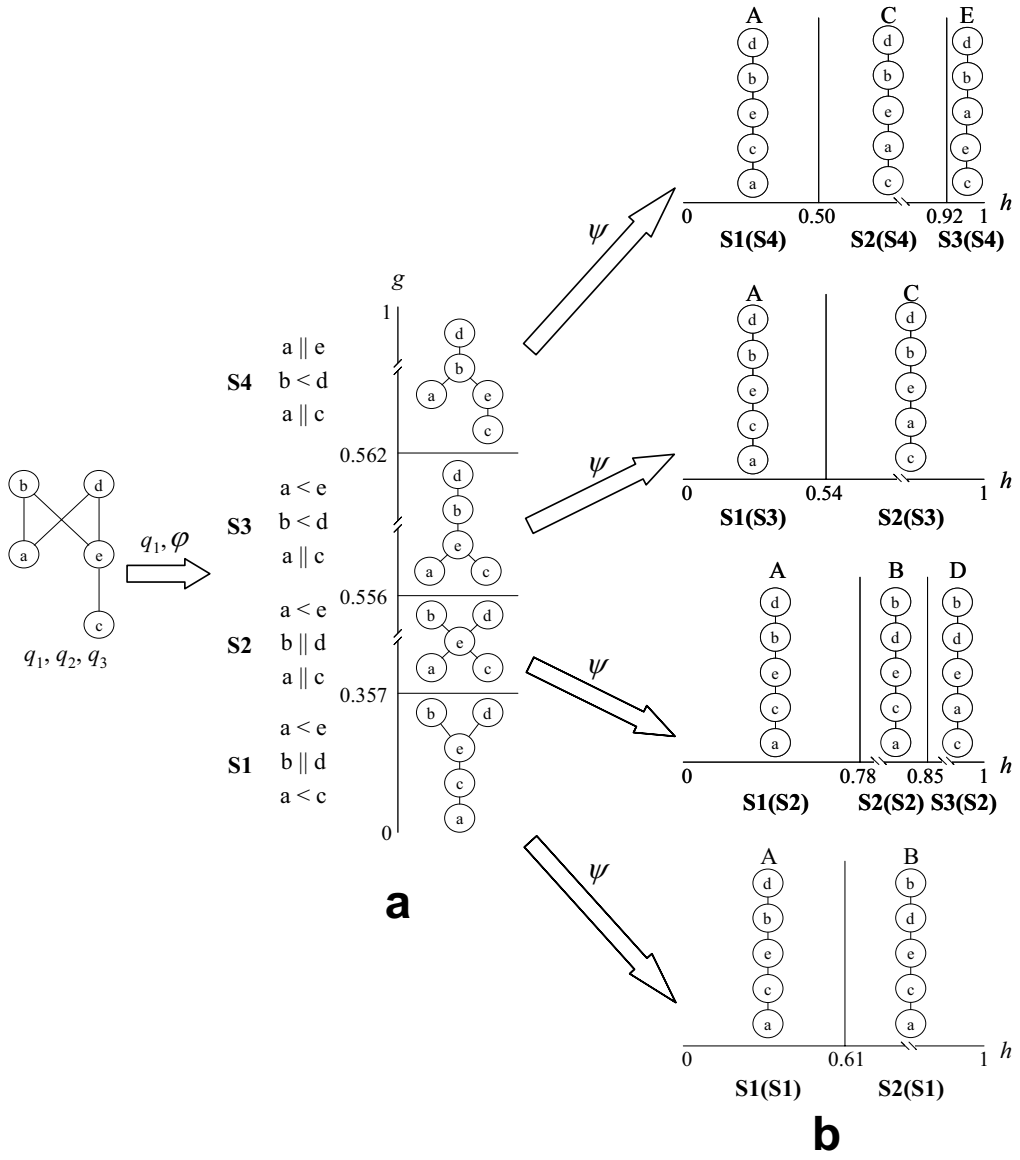


Figure 3. Hasse diagram, a) its four stability fields Si produced by aggregation of q_2 and q_3, and b) its ten stability fields Sj(Si) produced after a second (and in this example final) aggregation including q_1. A, B, C, D and E are the linear extensions labels shown in Figure 2.

The distribution of g_c values along the space g of weights is called the g-spectrum[40], and each region in which different g values yield the same poset is called a "stability field"[40]. Hence, in the case shown in Figure 3a there are four stability fields (S1 to S4) in the g-space, each one characterised by only one poset. In consequence, an important advantage of METEOR is that the changes in the ranking are shown for a complete range of descriptor weights.

Although it is interesting to plot the poset for each stability field, occasionally, when there are many incomparabilities, the various g_c values become close to each other making it difficult to plot them. In such case it is recommended to cluster g_c values and plot the posets between the clusters; such clusters are called "hot spots"[40].

When giving different priorities to q_2 and q_3, none of the HDs (Figure 3a) yields a linear ranking. To obtain a linear order and simultaneously evaluate the effect of q_1, φ must be further aggregated with q_1 yielding a new combined descriptor ψ :

$$\psi(x) = h \cdot \varphi(x) + (1-h) \cdot q_1(x) \qquad (6)$$

$$\psi(x) = (1-h) \cdot q_1(x) + hg \cdot q_2(x) + h(1-g) \cdot q_3(x) \qquad (7)$$

where h and (1 - h) are the weights for φ and q_1, respectively. As described for g-weights, there are also crucial values h_c. This new aggregation yields new stability fields Sj(Si) along the h axis for each one of the Si stability fields of the first aggregation (Figure 3b). Because in this example three descriptors are considered, the second aggregation including all descriptors yields linear extensions from the original HD. Some of the stability fields in a particular h-space hold the same linear extension of other stability fields from a different h-space; for example $a < c < e < b < d$ appears in all four h-spaces (S1(S1), S1(S2), S1(S3) and S1(S4)), resulting from low weights of q_1. The linear extensions B and C appear twice. In total, among the ten stability fields distributed along four h-spaces, there are five linear rankings from a total of six in Figure 2. The missing linear order F (Figure 2) cannot be obtained by any $\varphi \rightarrow \psi$ aggregation, which means that there are no linear weighted-aggregation resulting in a linear order. In general, the sequence of aggregations and their results may be regarded as a prioritisation scheme (Figure 4). Since in this example only two aggregations were required to obtain the linear extensions, a 2-dimensional representation can be drawn (Figure 5).

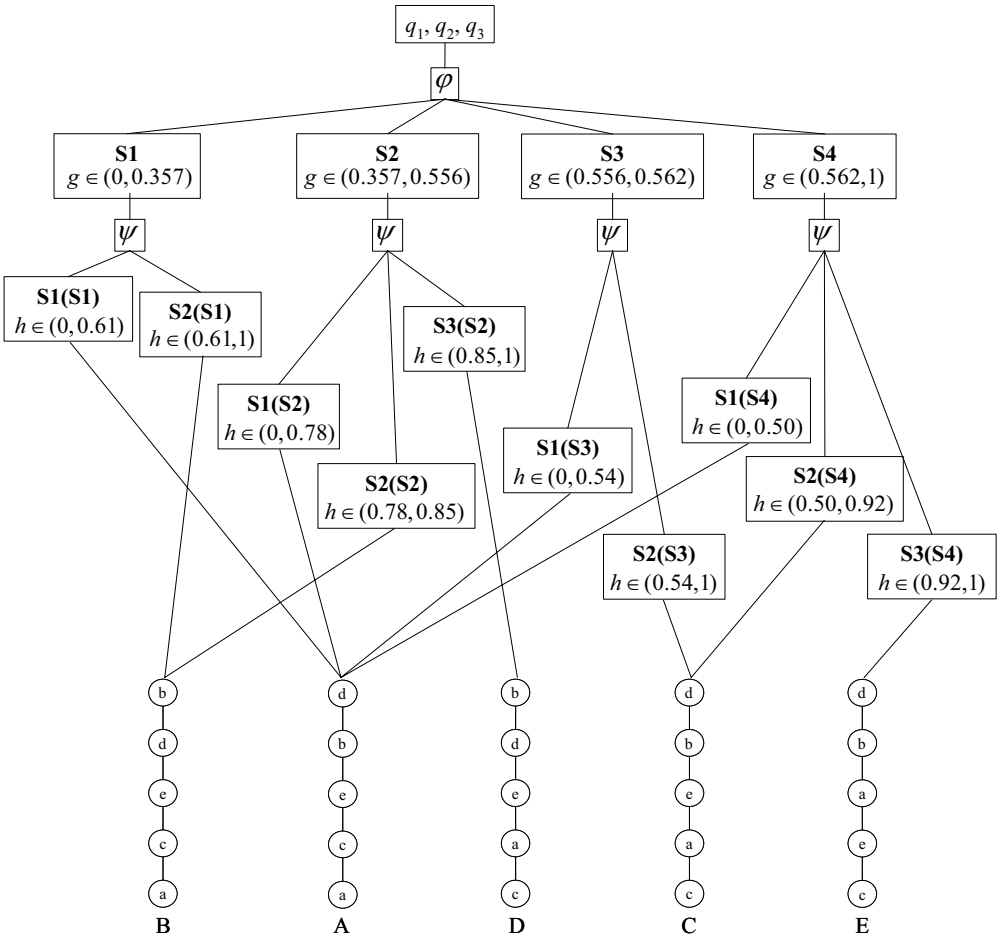


Figure 4. Prioritisation scheme corresponding to a first aggregation using φ (Eqs. 3 and 4) and to a second one by ψ (Eqs. 6 and 7). A, B, C, D and E represent the linear extensions depicted in Figure 2.

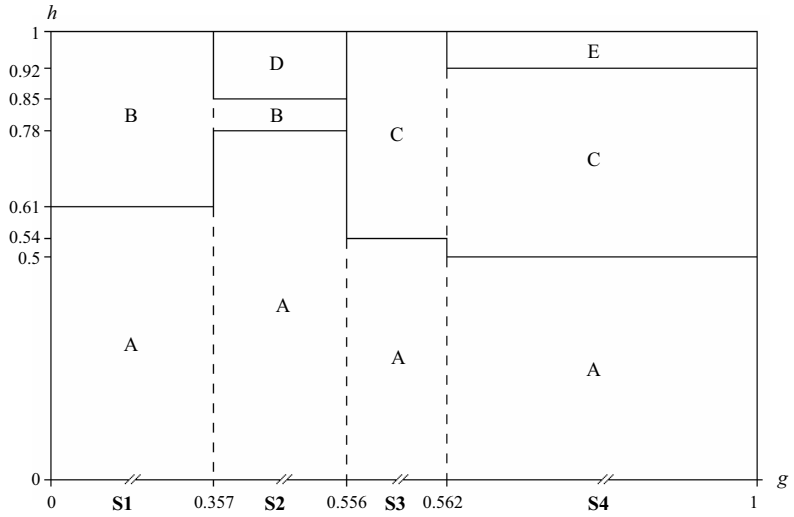


Figure 5. Two-dimensional representation of the effect of selecting particular *g* and *h*-weights on the stability fields. A, B, C, D and E represent the linear extensions depicted in Figure 2.

In the prioritisation scheme (Figure 4), each path from $\{q_1, q_2, q_3\}$ to a particular linear extension shows a specific set of priorities yielding the selected linear extension. It can be seen that different paths can lead to the same linear extension.

Weights representing property priorities can also be expressed in percentages. For example, the linear extension E is obtained according to Eq. 7 when priorities of q_1 are lower than 8 % [$(1 - h) < 0.08$], priorities of q_2 are greater than 52 % [$hg > 0.52$] and priorities of q_3 are lower than 44 % [$h\,(1 - g) < 0.44$] in such a way that the sum of the three selected priorities is 100 %, i.e. $(1 - h) + hg + h\,(1 - g) = 1$ (Eq. 7). The other stability fields are in the same way bounded by certain values of descriptor priorities. This result motivates to introduce a probability concept in the space of weights: Considering Figure 5 the linear extension A is the most probable of all five (Figure 5), its stability fields covering 60 % of all possible (*g*, *h*) combinations; the second most probable is C with 17 %. E and D are the most improbable ones.

Total orders in environmental ranking of refrigerants

In a recent work[42] we have ranked 40 refrigerants used in the past, used presently, and some proposed substitutes; they were described by their ozone depletion potentials (ODP), global warming potentials (GWP), and atmospheric life times (ALT). The interest in such a study was to analyse the order relations of 13 different subsets: chlorofluorocarbons (CFC),

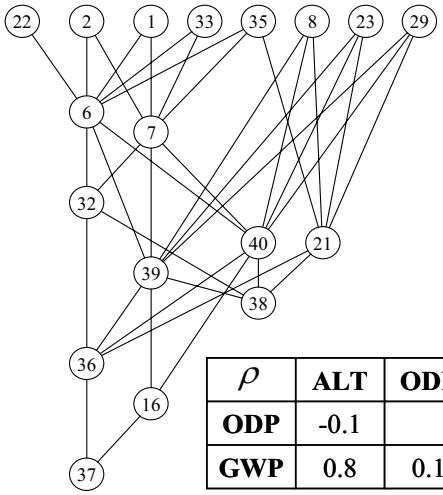
hydrofluorocarbons (HFC), hydrochlorofluorocarbons (HCFC), hydrocarbons (HC), di(fluoroalkyl)ethers (DFAE), alkyl-fluoroalkylethers (AFAE), chloromethanes (CM), and the single-compound subsets trifluoroiodomethane (FIM), octafluorocyclobutane (PFC), carbon dioxide (CO_2), bromochlorodifluorobutane (BCF), dimethyl ether (DME) and ammonia (NH_3). The HD showed many incomparabilities among refrigerants making it difficult to decide which refrigerant is the most problematic one; in fact, 8 maximal[29] chemicals resulted. A decision on the least problematic substance was also not possible because two minimal[29] refrigerants appeared.

When applying METEOR to obtain linear orders among refrigerants, a set *P* (Table 1) of maximal chemicals of each subset was selected. The HD and Spearman's rank correlation of these 18 refrigerants in *P* in respect to the three properties are depicted in Figure 6. *P* contains more than 13 substances because CFC, HCFC and AFAE have more than one maximal refrigerant (Table 1).

Table 1. Labels, chemical subsets, molecular formulae and non-propietory names of the refrigerants in *P*.

Label*	**Subset**	**Molecular formula**	**Non-proprie-tory name**
1	CFC	CCl_3F	R11
2	CFC	CCl_2F_2	R12
6	HCFC	$C_2H_3Cl_2F$	R141b
7	HCFC	$C_2H_3ClF_2$	R142b
8	HFC	CHF_3	R23
16	HC	C_3H_8	R290
21	CO_2	CO_2	R744
22	BCF	$CBrClF_2$	R12B1
23	PFC	C_4F_8	RC318
29	DFAE	C_2HF_5O	HFE-125
32	CM	CH_3Cl	R40
33	CFC	$C_2Cl_3F_3$	R113
35	CFC	$C_2Cl_2F_4$	R114
36	FIM	CF_3I	R13I1
37	DME	C_2H_6O	-
38	NH_3	NH_3	R717
39	AFAE	$C_2H_3F_3O$	HFE-143
40	AFAE	$C_3H_3F_5O$	HFE-245

*Labels correspond to those used in reference 42.



ρ	ALT	ODP
ODP	-0.1	
GWP	0.8	0.1

Figure 6. Hasse diagram of 18 maximal subset refrigerants and Spearman's rank correlation among their properties.

There are 65,362,464 linear extensions associated with this HD; in the following is shown how many and which are obtained under prioritisations of ALT, ODP and GWP. Initially, for a refrigerant *x*, ALT and ODP are aggregated in $\varphi(x)$ because they are weakly anticorrelated properties for the 18 refrigerants studied (Figure 6).

$$\varphi(x) = g \cdot ALT(x) + (1-g) \cdot ODP(x) \qquad (8)$$

There are 43 incomparable pairs due to conflictive values between ALT and ODP; their respective g_c values are shown in Table 2. Because several of these values are close to each other, they were clustered using hierarchical cluster analysis with Hamming distance as similarity function and unweighted average linkage as grouping method. A level of similarity of 90 % among g_c values was selected, and 11 clusters corresponding to hot spots (Hi) were detected resulting in 12 stability fields (Si) in this *g*-space (Table 2). A histogram showing the distribution of g_c values at the hot spots is shown in Figure 7. The differences among these stability fields can be calculated through the W-index[17,18], a dissimilarity function used to quantify the disagreement between two posets taking into account the order relationships among their objects. The W-values for all comparisons among the 12 stability fields are shown in Table 3, the *g*-spectrum and its corresponding stability fields in Figure 8.

Table 2. Pairs of refrigerants with conflictive values between ODP and ALT, and their corresponding clustering into hot spots Hi.

Hot spot	Pair	g_c	Hot spot	Pair	g_c
H1	**(8, 23)**	8.5651E-5	H8	**(1, 8)**	0.7360
	(23, 32)	0.0039		**(32, 39)**	0.7404
	(7, 23)	0.0127		**(2, 8)**	0.7516
	(6, 23)	0.0231		**(8, 33)**	0.7532
	(8, 32)	0.0438		**(2, 33)**	0.7699
H2	**(29, 32)**	0.0712	H9	**(6, 7)**	0.8005
	(21, 32)	0.0956		**(32, 40)**	0.8229
	(33, 35)	0.12733		**(1, 29)**	0.8395
	(7, 8)	0.1385	H10	**(29, 33)**	0.8759
	(2, 23)	0.1423		**(2, 29)**	0.8878
	(23, 33)	0.1535		**(1, 21)**	0.8932
	(23, 35)	0.1553		**(22, 35)**	0.9022
	(1, 23)	0.1659		**(8, 22)**	0.9251
H3	**(7, 29)**	0.2171	H11	**(21, 33)**	0.9416
	(6, 8)	0.2235		**(22, 29)**	0.9541
H4	**(1, 35)**	0.2696		**(2, 21)**	0.9626
	(7, 21)	0.2854		**(21, 22)**	0.9671
	(6, 29)	0.3260		**(2, 22)**	0.9679
H5	**(6, 21)**	0.4048		**(22, 33)**	0.9727
H6	**(22, 23)**	0.5009		**(1, 22)**	0.9870
H7	**(1, 33)**	0.6107		**(7, 22)**	0.9978
	(1, 2)	0.6725			

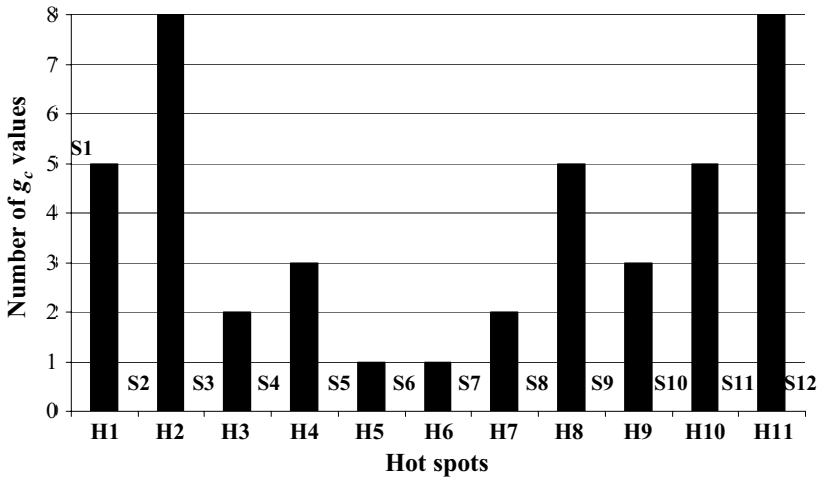


Figure 7. Histogram showing the g_c-population of each hot spot, with the stability fields Si between the hot spots.

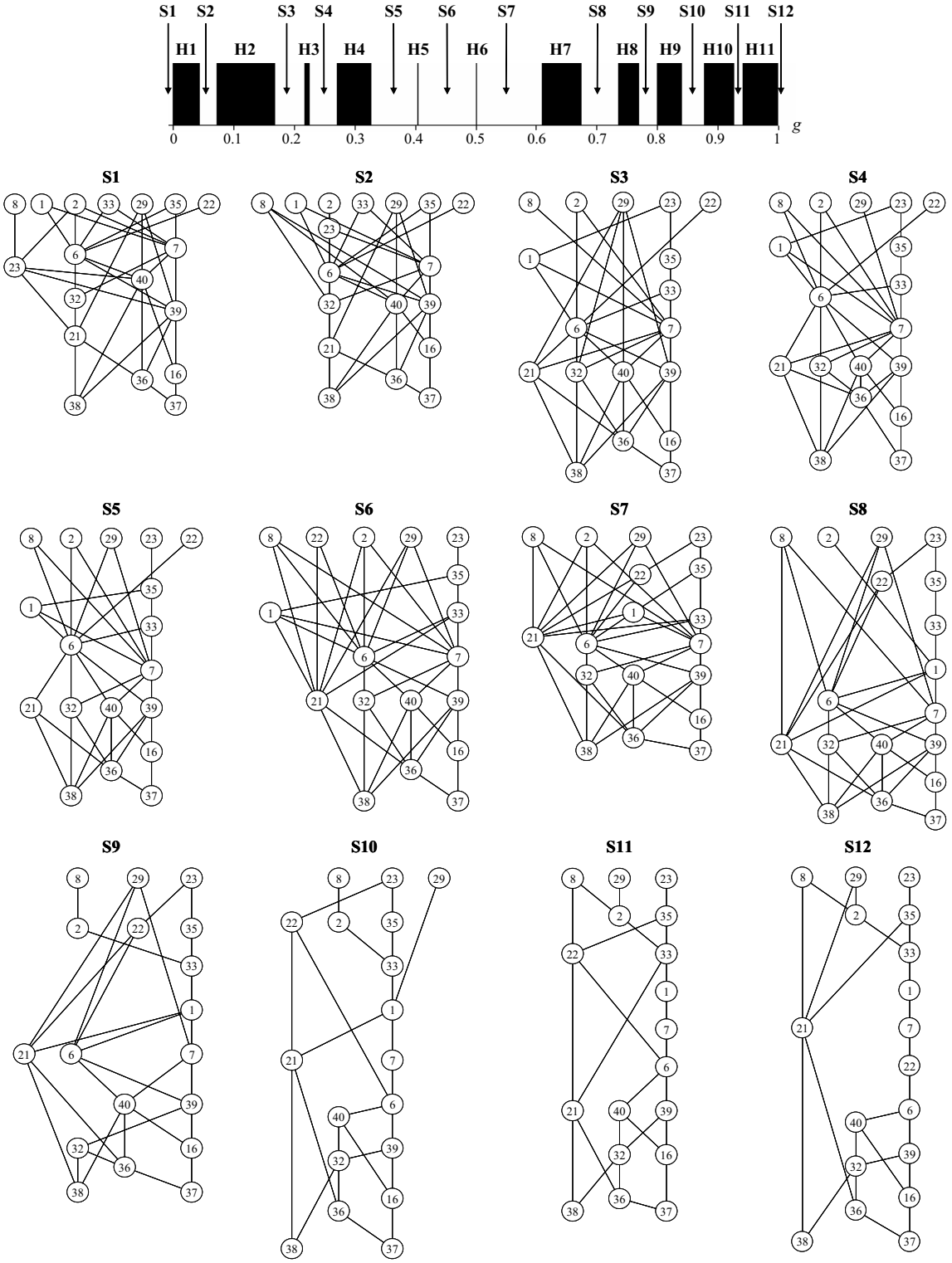


Figure 8. Stability fields (Si) and hot spots (Hi) on the *g*-spectrum produced by aggregation of ALT and ODP. Hasse diagrams of each Si calculated from GWP and φ.

Table 3. Dissimilarities (W-values) among the Hasse diagrams of the stability fields (Si) in Figure 8. W-index values of adjacent Si's are given in bold italics.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**
S1	0											
S2	***5***	0										
S3	13	***8***	0									
S4	15	10	***2***	0								
S5	18	13	5	***3***	0							
S6	19	14	6	4	***1***	0						
S7	20	15	7	5	2	***1***	0					
S8	22	17	9	7	4	3	***2***	0				
S9	27	22	14	12	9	8	7	***5***	0			
S10	30	25	17	15	12	11	10	8	***3***	0		
S11	35	30	22	20	17	16	15	13	8	***5***	0	
S12	43	38	30	28	25	24	23	21	16	13	***8***	0

As all 12 stability fields still hold incomparabilities due to conflictive values between φ and the non-aggregated property GWP, a second step is necessary to break the remaining incomparabilities and to assess the effect of prioritising GWP together with ODP and ALT. The second aggregation was carried out through the following combination:

$$\psi(x) = h \cdot \varphi(x) + (1-h) \cdot GWP(x) \quad (9)$$

$$\psi(x) = hg \cdot ALT(x) + h(1-g) \cdot ODP(x) + (1-h) \cdot GWP(x) \quad (10)$$

Hence, for each chemical x in the 12 stability fields in g, an aggregated function $\psi(x)$ was calculated. All incomparabilities yield corresponding h_c values which were clustered using Hamming distance and unweighted average linkage. Clusters were formed by h_c-values holding 90 % or more similarity, thus becoming hot spots of the h-space. To each Si in the first aggregation corresponds a collection of hot spots H(Si) in the second aggregation. The h-intervals of each H(Si) are shown in Table 4. A two-dimensional representation of the g and h-spaces is depicted in Figure 9a where 109 stability fields (white and coloured regions) are shown as well as the corresponding hot spots (black regions) separating them. Note that the stability fields arising from S1 are not visible in Figure 9 because the g-interval of S1 is too small in comparison to the other Si's (Table 2).

Table 4. 97 Hot spots on the *h*-space displayed for each stability field Si from the first aggregation.

H(S1)	**H(S4)**	**H(S7)**	**H(S10)**
(0.0443, 0.0750)	(0.0810, 0.1407)	(0.0510, 0.0580)	(0.0314, 0.0568)
(0.2212, 0.2783)	(0.2747, 0.3406)	(0.0943, 0.1303)	(0.2318, 0.3210)
(0.3702, 0.4313)	(0.4867, 0.5811)	(0.3640, 0.3940)	(0.5175, 0.5404)
(0.4678, 0.4770)	(0.6920, 0.7130)	0.4736	(0.5909, 0.6464)
(0.6015, 0.6696)	(0.7580, 0.8841)	(0.6234, 0.6444)	0.6844
(0.7609, 0.8198)	(0.9338, 0.9965)	(0.6860, 0.7217)	0.7653
(0.9070, 0.9216)		(0.7761, 0.8528)	(0.8494, 0.9332)
(0.9531, 1)		(0.9021, 0.9313)	(0.9602, 0.9939)
		(0.9696, 0.9921)	
H(S2)	**H(S5)**	**H(S8)**	**H(S11)**
(0.0658, 0.1140)	(0.0839, 0.0948)	(0.0390, 0.0712)	(0.0287, 0.0516)
0.2317	0.1618	(0.1583, 0.1841)	(0.2542, 0.3007)
0.2907	0.3109	(0.3111, 0.3269)	(0.4964, 0.5194)
(0.3733, 0.4464)	0.3818	0.4979	(0.5703, 0.6103)
(0.4838, 0.4926)	(0.4654, 0.4834)	(0.5664, 0.5887)	(0.8396, 0.9051)
(0.6204, 0.6793)	(0.5367, 0.5965)	(0.6376, 0.6718)	(0.9388, 0.9868)
(0.7100, 0.7372)	(0.6849, 0.7978)	(0.7343, 0.7864)	
(0.7751, 0.7892)	(0.8525, 0.8683)	(0.8105, 0.8924)	
(0.8339, 0.8719)	(0.9189, 0.9948)	(0.9450, 0.9958)	
(0.9131, 0.9992)			
H(S3)	**H(S6)**	**H(S9)**	**H(S12)**
(0.0759, 0.0769)	(0.0647, 0.0703)	(0.0346, 0.0628)	(0.0266, 0.0479)
(0.1517, 0.2049)	(0.1042, 0.1218)	(0.1901, 0.2376)	(0.2415, 0.2550)
0.2606	0.3442	(0.2883, 0.3034)	0.4051
0.3244	(0.4126, 0.4302)	(0.5395, 0.5816)	(0.4794, 0.5025)
(0.4666, 0.5330)	(0.5805, 0.5877)	(0.6121, 0.6851)	(0.5536, 0.5906)
(0.6243, 0.7028)	(0.6326, 0.6896)	(0.8007, 0.8760)	0.7266
(0.7880, 0.8827)	(0.7323, 0.7916)	(0.9368, 0.9980)	(0.8279, 0.8887)
(0.9277, 0.9982)	(0.8598, 0.8962)		(0.9316, 0.9917)
	(0.9452, 0.9935)		

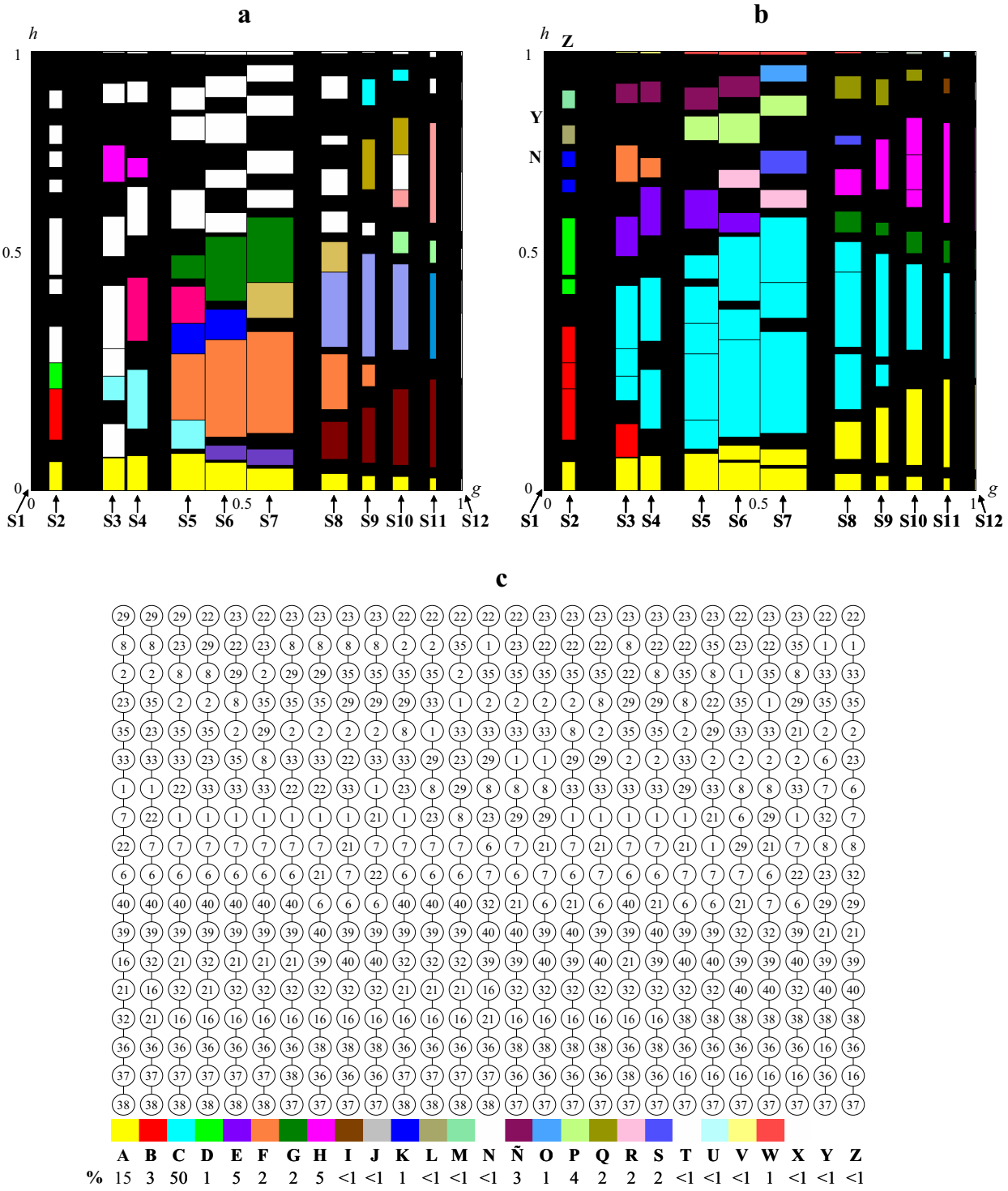


Figure 9. Two-dimensional representation of *g* and *h*-spaces. a) Stability fields equally coloured hold the same linear extension; each white region contains a different linear extension. b) Stability fields equally coloured contain linear extensions whose similarity is equal or greater than 90 %. Symbols N and Y refer to two stability fields obtained from S1 and Z to the stability field from S2 with highest *h*-values. c) Consensus ranking of each

cluster (coloured region in b); colours in c are those used in b. A to Z are the labels used to identify the 27 consensus rankings.

As was shown in the METEOR example, different regions of the two-dimensional representation may hold identical linear orders. To trace similarities among these 109 linear orders, the W-index for each pair was calculated, resulting in 39 identical linear orders; equally coloured regions in Figure 9a represent identical orders (white regions hold different orders). Hence, the 109 stability fields were reduced to 70 different ones implying different linear rankings when prioritising ALT, ODP and GWP. They were grouped using cluster analysis where the similarity function is the W-index transformed into a similarity value[43]; the grouping method was the unweighted average linkage. After a selection of those clusters whose members share 90 % or more similarity, 27 linear extensions are obtained (Figure 9b).

The average ranking $\overline{r_n}$ of each refrigerant n within each cluster of similar linear extensions was calculated using

$$\overline{r_n} = \sum_S r_S \cdot p_S \qquad (11)$$

where S represents a stability field within the studied cluster, r_S the ranking of n in S, and p_S the probability of having the linear extension associated to S when all the stability fields in the cluster are considered. Hence, $p_S = A_S / A_C$, A_S being the area of the stability field in the g-h plot and A_C the sum of areas of all stability fields in the studied cluster. The consensus rankings for the 27 stability field clusters were calculated using the $\overline{r_n}$ values (Figure 9c). The probabilities p_C of having a particular consensus-ranking were also calculated, considering the area of each cluster of stability fields A_C relative to the total area A_T of all clusters, according to $p_C = A_C / A_T$. These probabilities are given in percentages at the bottom of Figure 9c. A_T equals 39 %, meaning that 61 % of the g-h space is covered by hot spots.

Discussion

On METEOR

There are several ranking methods, for example Utility function[26], PROMETHEE[27] and Concordance analysis[28]; all of them include priorities in their procedures but none of them allows to study the whole set of priorities and their effects on the ranking. This is done by

METEOR, as was shown in the ranking of refrigerants; it allows to explore the space of priorities as a whole or to trace back the priorities needed to have a particular ranking. Another advantage of METEOR is the possibility of calculating the probability of having a particular linear order by descriptor prioritisation, which in the end saves time and resources since it avoids the "trial and error" selection of different descriptor priorities. METEOR yields concrete intervals of priorities for which the ranking is stable, a fact that through the usual ranking methods is only achievable after a large number of trials.

The aggregation procedure described in this paper includes nested aggregations, as $\{\{q_2, q_3\}, q_1\}$ in the first example, or {{ALT, ODP}, GWP} in the refrigerant case. In these examples the spectrum of the first aggregation weights is related to the spectrum of the second aggregation weights. Hence, g determines which values h can take; if a further aggregation were necessary involving a weight k, then k would be related to g and h. Instead of nest-aggregating four descriptors q_1, q_2, q_3 and q_4, it is also possible to pair-aggregate them, namely $\{q_1, q_2\}$ and $\{q_3, q_4\}$[40]. Although a linear aggregation was performed in the current work, METEOR is not restricted to this kind of combinations; in fact other aggregation functions may be explored and the assessment of their results on the ranking may be carried out.

Prioritisation schemes allow to analyse which kind of priorities on different aggregations permit to find common rankings. In a similar manner, a two-dimensional plot combining results of two subsequent aggregations can be used to explore the similarities among stability fields. In this respect, the selection of different similarity levels allows to see how similarities evolve in respect to descriptor priorities. Thereby, a collection of stability field neighbourhoods can be constructed for each similarity level. Each of these neighbourhood systems represents a basis permitting to study topological relationships among different subsets of stability fields. This kind of topological approach, based on the notion of similarity, is called chemotopology[39,44]; its application to stability field neighbourhoods will be published in a forthcoming paper.

Although the two-dimensional plot presenting the space of priorities is a versatile tool for analysing priority effects, it is restricted to the number of aggregations performed during the process. In fact, for a set P characterised by more than 4 descriptors and aggregated in the nested manner shown in this paper, a graphical representation of the priority space is not

possible. In such case the analysis must be carried out through prioritisation schemes, independent of the number of aggregations and always two-dimensional.

Ranking of refrigerants

Regarding environmental descriptors used in this research it should be noted that ODP and GWP are related to ALT[45]. It is also possible to perform regression analysis studies in order to obtain ODP = *f* (ALT), but this is out of scope of this manuscript. The current measure of ODP is based upon the comparison of the respective alternative refrigerant with trichlorofluoromethane which is decomposed in the stratosphere. Therefore, such ODP calculations are particularly appropriate for substances with similar reactivity, but they are also applied to substances which react in the lower atmosphere[45]. Similar considerations are valid for GWP calculations taking CO_2 as reference[45]. In this manuscript, because of the comparative aim behind any ranking methodology, ODP and GWP values are referred to trichlorofluoromethane and CO_2, respectively[42].

First aggregation: {*ALT*, *ODP*}.

According to the histogram shown in Figure 7 and Table 3, the majority of order relation changes occur at low and high *g*-values, whereas few changes are observed for intermediate *g*-values. The hot spots gathering most g_c-values are H2 and H11; obviously, the stability fields adjacent to each of them undergo many changes in order relations when compared. For example, S2 and S3, adjacent to H2, hold the maximum dissimilarity value if adjacent pairs of Si's are considered; the same situation occurs for S11 and S12. In contrast, few order changes occur for adjacent Si's separated by hot spots with only one g_c-value, as is the case for S5 - S6 and S6 - S7, separated by H5 and H6, respectively. These results show that low and high priorities of ALT or high and low priorities of ODP influence the ranking strongly. If the priorities of ALT and ODP are similar, relatively few changes entail. In general, aggregation of ALT and ODP prioritising ALT over ODP produces an increasing number of comparabilities among refrigerants (e.g. compare S1 with S12).

All Si's show that refrigerants 8 (trifluoromethane) and 29 (pentafluorodimethyl ether) are maximal substances. When considering the influence of GWP and all the weighted combinations of ALT and ODP, these substances are the most problematic ones. This is not surprising as HFCs are substances with relatively high values in these three properties[42]. The identification of 29, a di(fluoroalkyl) ether, as a maximal[29] substance in all Si's is remarkable

because these compounds have been proposed as replacements of CFCs, HCFCs and HFCs. In a recent publication[42] was shown that 29, even when ranked without prioritising ALT, ODP and GWP, is a problematic refrigerant. The ALT and ODP aggregation carried out here shows that 29 is invariant to prioritisation which is caused by its property values; its ALT (165 years) is one of the highest of the refrigerants studied[42], and additionally its high GWP value (14,800 relative to CO_2 with 100 years of time horizon) places 29 at the top of the ranking. At high priorities of ODP (low *g*-values) emphasizing its low ODP value (0 with respect to CCl_3F), it is still placed high in the ranking due to its high non-aggregated GWP. Thus, its GWP stresses the importance of ALT for high *g* values, the main reason for the high ranking at low *g*-values.

In all Si's, 37 (dimethyl ether) and 38 (ammonia) are the minimal[29] refrigerants because their ALTs, ODPs and GWPs values are low in respect to others (ALT(38) = 0.25, ALT(37) = 0.015 years; ODP(38) = ODP(37) = 0 relative to CCl_3F; GWP(38) = 0, GWP(37) = 1 relative to CO_2 with 100 years time horizon). Note that the incomparability between 37 and 38 cannot be broken by any weighted aggregation of ODP and ALT followed by a non-weighted ranking with GWP. In fact, the incomparability arises because ALT(38) > ALT(37) and GWP(38) < GWP(37), awaiting the second aggregation step.

Second aggregation: {{*ALT, ODP*}, *GWP*}.

In general, the stability fields Sj(Si) arising from each Si (first aggregation) are quite similar for low *h*-values, i.e. high GWP priorities. Hence, many of the incomparabilities in each Si are mapped into similar linear extensions for low *h*-values. In general, 37 is the least problematic substance for high GWP priorities, associated to high ALT and low ODP ones, corresponding to the consensus rankings G, H,… , O, P,… , Y, Z (Figure 9c), while 38 reaches the lowest ranking for the consensus A, B, C, D, E, F, K, L, M and N, all of them related to intermediate priorities of GWP and any priority of ODP and ALT. While 8 (trifluoromethane) and 29 (pentafluorodimethyl ether) are the most problematic substances in the first aggregation, after performing the second aggregation 22 (bromochlorodifluoromethane) and 23 (octafluorocyclobutane) turn out to be the most problematic ones. Refrigerant 22 is maximal for low GWP priorities, high ODP and low ALT priorities; 23 is maximal for low GWP priorities associated to high ALT and low ODP priorities. Note that 22 and 23 are substances without hydrogen atoms and a high degree of

halogenation; 22 contains halogens which promote its ODP, i.e. 1 bromine and 1 chlorine atoms, besides 2 fluorine atoms.

Figure 9c can be used to explore the ranking of particular substances; for example 29 (pentafluorodimethyl ether) is the maximal element for A, B and C, corresponding to high GWP priorities associated to any priority of ALT and ODP. This emphasizes the problematic nature of this hydrofluoroether as regulatory and environmental agencies are today focusing their attention on green house gases[46] after having stabilised the ozone depleting substances[47]. Another fact stressing the problematic position of 29 is that A and C are the two stability field clusters with the highest probability of occurrence (Figure 9b, cumulative probability of 68 %) which means that many priority combinations yield 29 as a problematic refrigerant.

A versatile tool to predict the effect of prioritising ALT, ODP and GWP is presented (Figure 9b). For example, if one is interested in a ranking with 10 % priority for ALT, 40 % for ODP and 50 % for GWP, simple algebraic manipulations of the weighting factors in Eq. 10 result in g = 0.2 and h = 0.5, corresponding to S3 in the first aggregation and to the consensus ranking E (Figure 9c), where 23 (octafluorocyclobutane) turns out to be the most problematic and ammonia the least problematic substances. If for example, a ranking is performed emphasizing ODP with the following priorities: ODP 75 %, GWP 20 % and ALT 5 %, then g = 0.0625 and h = 0.8 are obtained. This corresponds to S2 in the first aggregation and to the consensus ranking L in the second. In this case 22 (bromochlorodifluoromethane) turns to be the most problematic and ammonia the least problematic refrigerants.

Conclusions and outlook

In the ranking of objects often the following questions arise: How does the rank look like if one prioritises certain object descriptors? Can one trace back the priorities needed to have a particular ranking? What is the probability of obtaining a certain ranking by prioritisation? METEOR and the method shown in this work to draw and analyse ranking similarities give answers to these questions.

It was shown the aggregation of descriptors by linearly combining them (Eq. 3, 4, 6 and 7). In general, the aggregation function selected must provide a way to “dial” descriptor priorities and to assess their effect on the final ranking. This dialling can be done by using different

mathematical functions, not only linear functions; therefore it is important to explore their application. In the same manner, the method here described is not restricted to the use of hierarchical clustering, in fact any clustering method might be used.

Although some other ranking methods like Utility function[26], PROMETHEE[27] and Concordance analysis[28] do not permit to study the effect of a step-by-step descriptor aggregation, it is important to compare METEOR with these procedures in order to stress the advantages of METEOR. This study will be published in a forthcoming paper.

The application of METEOR to the environmental ranking of 18 representatives of refrigerants used in the past, presently used, and some proposed substitutes showed that pentafluorodimethyl ether (29), a potential HCFC and HFC replacement, is the most problematic refrigerant for 68 % of the total descriptor prioritisations studied. This result warns us about the adverse environmental impact before the large scale production of this substance is commenced.

Although environmental ranking is an important step towards the selection of acceptable refrigerants, an even more expansive ranking must be performed in order to include some other relevant aspects such as energy efficiency, toxicity, insulating ability, flammability, physical and chemical stability, solubility, costs, and other technical properties. In those cases, METEOR and the methods here developed can play an important role.

Acknowledgments

The authors thank the Bavarian Environmental Agency for supporting this study under the Research project 81-00213381. G. Restrepo specially thanks COLCIENCIAS and the Universidad de Pamplona for the grant offered during this research.

References and notes

1. Davis, G. A.; Swanson, M.; Jones, S. Comparative evaluation of chemical ranking and scoring methodologies; EPA order No. 3N-3545-NAEX, 1994.

2. Jamison, D. T.; Sandbu, M. E. WHO Ranking of health system performance. *Science* **2001**, *293*, 1595-1596.

3. Roberts, L. Ranking the rain forests. *Science* **1991**, *251*, 1559-1560.

4. Taubes, G. Measure for measure in science. *Science* **1993**, *260*, 884-886.

5. Ball, P. Index aims for fair ranking of scientists. *Nature* **2005**, *436*, 900-900.

6. Shewchuk, R. M.; O'Connor, S. J.; Williams, E. S.; Savage, G. T. Beyond rankings: using cognitive mapping to understand what health care journals represent. *Soc. Sci. Med.* **2006**, *62*, 1192-1204.

7. Finkel, E. Australia's proposed U.K.-style merit ranking stirs debate. *Science* **2006**, *312*, 176.

8. Broder, A. Z.; Lempel, R.; Maghoul, F.; Pedersen, J. Efficient PageRank approximation via graph aggregation. *Inf. Retrieval* **2006**, *9*, 123-138.

9. Willett, P. Textual and chemical information retrieval: different applications but similar algorithms. Inform. Res. 2000, 5(2), URL: http://InformationR.net/ir/5-2/infres52.html (accessed Jul. 2007).

10. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. Assoc. Comput. Mach.* **1999**, *46*, 604-632.

11. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank citation ranking: bringing order to the web. URL: http://dbpubs.stanford.edu/pub/1999-66 (accessed Jul. 2007).

12. Google, Inc., Mountain View, CA. http://www.google.com (accessed Jul. 2007).

13. Swamidass, S. J.; Baldi, P. Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952-964.

14. Fontana, P.; Pretsch, E. Automatic spectra interpretation, structure generation, and ranking. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 614-619.

15. Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469-474.

16. Lin, T-H.; Chiu, S-H.; Tsai, K-C. Supervised feature ranking using a genetic algorithm optimized artificial neural network. *J. Chem. Inf. Model*. **2006**, *46*, 1604-1614.

17. Brüggemann, R.; Bartel, H. G. A theoretical concept to rank environmentally significant chemicals. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 211-217.

18. Brüggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C. E. W. Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 918-925.

19. Lerche, D.; Sørensen, P. B.; Bruggemann, R. 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.* **2003**, *43*, 1471-1480.

20. Brüggemann, R.; Simon, U.; Mey, S. Estimation of averaged ranks by extended local partial order models. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 489-518.

21. Davis, G.; Fort, D.; Hansen, B.; Irwin, F.; Jones, B.; Jones, S.; Socha, A.; Wilson, R.; Haaf, B.; Gray, G.; Hoffman, B.; Swanson, M. B.; Socha, A. C. Framework for chemical ranking and scoring systems. In *Chemical ranking and scoring: guidelines for relative assessment of chemicals*; Swanson, M. B.; Socha, A. C., Eds.; SETAC Press: Pensacola, 1997; pp 1-30.

22. Simon, U.; Brüggemann, R.; Mey, S.; Pudenz, S. METEOR - application of a decision support tool based on discrete mathematics. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 623-642.

23. Voigt, K.; Brüggemann, R. Water contamination with pharmaceuticals: data availability and evaluation approach with Hasse diagram technique and METEOR. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 671-689.

24. Simon, U.; Brüggemann, R.; Behrendt, H.; Shulenberger, E.; Pudenz, S. 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.* **2006**, *34*, 126-136.

25. Lerche, D.; Brüggemann, R.; Sørensen, P.; Carlsen, L.; Nielsen, O. J. A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1086-1098.

26. Schneeweiß, C. *Planung 1 - Systemanalytische und entscheidungstheoretische Grundlagen*; Spinger-Verlag: Berlin, 1991.

27. Brans, J. P.; Vincke, P. H.; Mareschal, B. How to select and how to rank projects: the PROMETHEE method. *Eur. J. Oper. Res.* **1986**, *24*, 228-238.

28. Opperhuizen, A.; Hutzinger, O.; Multi-criteria analysis and risk assessment. *Chemosphere* **1982**, *11*, 675-678.

29. Trotter, W. J. *Combinatorics and partially ordered sets Dimension theory*; The Johns Hopkins university press: Baltimore, 1992.

30. Strassert, G. *Das Abwägungsproblem bei multikriteriellen Entscheidungen – Grundlagen und Lösungsansatz unter besonderer Berücksichtigung der Regionalplanung*; Peter Lang Frankfurt: Frankfurt am Main, 1995.

31. Brüggemann, R.; Münzer, B.; Halfon, E. An algebraic/graphical tool to compare ecosystems with respect to their pollution - the German river “Elbe” as an example - I: Hasse-diagrams. *Chemosphere* **1994**, *28*, 863-872.

32. Brüggemann, R.; Sørensen, P. B.; Lerche, D.; Carlsen, L. Estimation of averaged ranks by a local partial order model. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 618-625.

33. Brüggemann, R.; Restrepo, G.; Voigt, K. Structure-fate relationships of organic chemicals derived from the software packages E4CHEM and WHASSE. *J. Chem. Inf. Model.* **2006**, *46*, 894-902.

34. Brüggemann, R.; Voigt, K. Stability of comparative evaluation, -example: environmental databases. *Chemosphere* **1996**, *33*, 1997-2006.

35. Matarazzo, B.; Munda, G. New approaches for the comparison of L-R fuzzy numbers: a theoretical and operational analysis. *Fuzzy Set. Syst*. **2001**, *118*, 407-418.

36. Saaty, T. L. Decision-making with the AHP: why is the principal eigenvector necessary. *Eur. J. Oper. Res*. **2003**, *145*, 85-91.

37. Pruesse, G.; Ruskey, F. Generating linear extensions fast. *SIAM J. Comput.* **1994**, *23*, 373-386.

38. These features of the HDs and the corresponding drawings can be performed with WHASSE, a software programme freely available from R. B. and introduced in: Brüggemann, R.; Halfon, E.; Bücherl, C. 1995. Theoretical base of the program “Hasse”, GSF–Bericht 20/95, Neuherberg.

39. Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 68-75.

40. Brüggemann, R.; Voigt, K.; Restrepo, G.; Simon, U. Concept of stability fields and hot spots in ranking of environmental chemicals. *J. Environ. Softw*. (In press).

41. Krzanowski, W. J. *Principles of multivariate analysis: A user's perspective*; Oxford university press: Oxford, 2003, p 407.

42. Restrepo, G.; Weckert, W.; Brüggemann, R.; Gerstmann, S.; Frank, H. Ranking of refrigerants. Submitted to *Environ. Sci. Technol.*

43. Gordon, A. D. *Classification*; Chapman & Hall/CRC: Boca Raton, 1999, p 13.

44. Restrepo, G.; Mesa, H.; Villaveces, J. L. On the topological sense of chemical sets. *J. Math. Chem.* **2006**, *39*, 363-376.

45. Kurylo, M. J.; Orkin, V. L. Determination of atmospheric lifetimes via the measurement of OH radical kinetics. *Chem. Rev.* **2003**, *103*, 5049-5076.

46. IPCC. Climate Change 2007: Mitigation. Contribution of working group III to the fourth assessment report of the Intergovernmental Panel on Climate Change; Metz, B. Davidson, O. R. Bosch, P. R. Dave, R. Meyer, L. A., Eds.; Cambridge university press: Cambridge, New York, 2007.

47. Reisch, M. S. Hot times ahead for refrigerants. *Chem. Eng. News* **2005**, *83*, 23-24.

Appendix H

Ranking of refrigerants

Guillermo Restrepo[a,c], Monika Weckert[a], Rainer Brüggemann[b], Silke Gerstmann[a] and Hartmut Frank[a]

[a] Environmental Chemistry and Ecotoxicology, University of Bayreuth, Bayreuth, Germany
[b] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[c] Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

Abstract

Environmental ranking of refrigerants is of need in many instances, Here, the ranking is based upon ozone depletion potential (ODP), global warming potential (GWP), and atmospheric life time (ALT). The aim is to find the environmental hazard posed by 40 refrigerants including those used in the past, presently used, and some proposed substitutes. This is achieved by applying the Hasse Diagram Technique, a mathematical method which permits to assess order relationships of chemicals. The refrigerants are divided into 13 classes of which the most prominent ones are chlorofluorocarbons, hydrofluorocarbons, hydrochlorofluorocarbons, hydrofluoroethers, and hydrocarbons. The dominance degree, a method for measuring order relationships among classes, is discussed and its application to the 13 classes of refrigerants is performed. The results show that some hydrofluoroethers are as problematic as the hydrofluorocarbons. The hydrocarbons and ammonia are the least problematic refrigerants regarding the three selected properties.

Introduction

Over the years several chemicals have been used as refrigerants and the change to new substances has been addressed avoiding the disadvantages of the previous ones (*1*). Currently, the adverse environmental properties of chlorofluorocarbons (CFCs), hydrofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs) has lead to regulate their production and consumption (*2-4*) and further research is needed in order to find environmental acceptable refrigerants (*5*).

The main drawback of CFCs and HCFCs is their potential to deplete the atmospheric ozone layer (*6*); HFCs are of concern as they contribute to global warming (*4*). Recognition of these problems has led to develop indicators for quantifying and comparing them, namely ODP (*8*) (Ozone Depletion Potential) and GWP (*9*) (Global Warming Potential); these two indicators are closely related to the atmospheric lifetime (ALT) (*7*) of these substances.

From an environmental point of view, an optimal refrigerant must have low values of ODP, GWP and ALT. However, the selection of a suitable alternative is not an easy task because there is no chemical embracing the lowest indicators at the same time. Therefore, the most appropriate substances in respect to ODP, GWP and ALT must be selected through the application of a methodology which compares the impact of the environmental indicators simultaneously and independently. This is reached by the ranking procedure based upon partial order theory, as is shown in this paper.

Normally, refrigerants are classified into different families based on their molecular structure, for example CFCs, HCFCs, HFCs, hydrocarbons (HCs) and hydrofluoroethers (HFEs). Hence, for a given set of refrigerants it is possible to rank those different classes in such a

way that the classes are identified which are less problematic than others or which ones present overall more adverse environmental impacts than others.

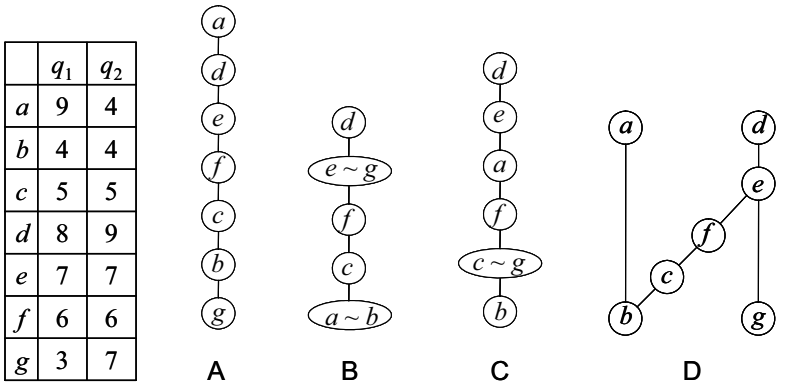
Materials and methods

Ranking

In a ranking procedure different descriptors q_1, q_2, ..., q_i are used to rank objects *a*, *b*,… that are gathered in a set *G*. For example, a set of chemicals $G = \{a, b, c, d, e, f, g\}$ may be described as appears in the data matrix shown in Figure 1. A linear ranking is easily obtained if only one property q_i is considered; for instance, the linear ranking A is obtained if q_1 is uniquely analysed, whereas B results if only q_2 is regarded (Figure 1). Because the descriptor q_2 of *a* is equal to that of *b* [$q_2(a) = q_2(b)$] and the one of *e* is equal to that of *g* [$q_2(e) = q_2(g)$] each one of these pairs is considered as equivalent in the ranking B, i.e. $a \sim b$ and $e \sim g$. If q_1 and q_2 are, for example, environmental properties whose values increase with the extent of adverse impact, ranking A shows that *a* is the "most hazardous" substance, whereas in the ranking B *d* is the most problematic. In real cases, the objects to rank are described by more than one descriptor, and all of them have to be considered simultaneously in the ranking process. Many ranking methods (*11*) perform a weighted combination of descriptors yielding a new superdescriptor; an example is the Utility Function (*14*) which calculates $\Gamma(x)$ for each object *x*, giving a weight g_i to each q_i descriptor; a simple version of $\Gamma(x)$ is the following:

$$\Gamma(x) = \sum g_i \cdot q_i(x) \qquad (1)$$

If equal priorities are assigned to q_1 and q_2, $\Gamma(x)$ values can be depicted in a linear order (Figure 1C) where objects with high $\Gamma(x)$ scores are located high in the ranking. Although the descriptors are simultaneously used, the mathematical form of $\Gamma(x)$ and its weights are still a source of subjectivities. A different ranking method avoiding these drawbacks is the Hasse diagram technique.



	q_1	q_2
a	9	4
b	4	4
c	5	5
d	8	9
e	7	7
f	6	6
g	3	7

Figure 1. Data matrix of 7 chemicals described by q_1 and q_2. Rankings according to A) q_1 and B) q_2. C) Ranking due to a weighted combination of q_1 and q_2 (aggregation). D) Hasse diagram.

Hasse Diagram Technique (HDT)

In HDT (*12*, *13*) two objects *x* and *y* characterised by descriptors $q_1(x)$, $q_2(x)$, ..., $q_i(x)$ and $q_1(y)$, $q_2(y)$, ..., $q_i(y)$, respectively, are compared by contrasting their individual descriptors in such a way that *x* is ranked higher than *y* ($x \geq y$) if all *x* descriptors are higher than those of *y* ($q_i(x) > q_i(y)$ for all *i*) or at least one descriptor is higher for *x* regarding *y* while all the others

are equal ($q_j(x) > q_j(y)$ for some j and $q_i(x) = q_i(y)$ for the rest of descriptors). In this case, x and y are said to be comparable. If all descriptors for x are equal to those of y ($q_i(x) = q_i(y)$ for all i), both substances are equivalent (*12*). It further follows that if $x \geq y$ and $y \geq z$ then $x \geq z$. In case at least one property q_j fulfils $q_j(x) < q_j(y)$ while the others meet the relation $q_i(x) \geq q_i(y)$, x and y are called incomparable and they are not ordered with respect to each other (*12*). Two objects are in a "cover-relation" if they are comparable and if there is no third one in between. The results of the order relationships between the objects of a set G are graphically presented in a Hasse diagram (HD) which can be drawn and analysed with the software WHASSE (*15*) (available from R. B.). Figure 1D depicts the HD for the matrix shown in Figure 1.

The richness of a HD lies in the lines connecting objects. An object with lines only in the downward direction indicates the highest rank or maximal object (*16*), for example *a* and *d* in Figure 1D. Objects with lines only in the upward direction are called minimal objects (*16*) and correspond to the lowest ranked objects, as *b* and *g* in Figure 1D. The absence of a line between two objects means that they are either incomparable or that there is a sequence of lines connecting them and keeping the same direction. For example, *f* and *b* are comparable ($f \geq b$), but no direct line is drawn between them because it is already contained in the path $f \geq c$, $c \geq b$ (*13*). According to the matrix shown in Figure 1, *f* and *b* are comparable because $q_1(b) < q_1(f)$ and $q_2(b) < q_2(f)$. The pair *a* and *c* is an example of incomparable chemicals because $q_1(a) > q_1(c)$, but $q_2(a) < q_2(c)$. In general, such pairs are recognised in a HD because there are no lines connecting them or they are connected by lines that do not follow the same direction, as is the case of *b* and *g* in Figure 1D.

According to the HD (Figure 1D), *a* is more problematic than *b*. Any comparison of *a* with another chemical requires additional knowledge about importance of the used descriptors. For example, the results obtained using eq 1 (Figure 1C) show the same relation between *a* and *b*, but also $a > f$, $a > c$ and $a > g$, which is caused by the weighted aggregation of these method. With the HD it is also possible to state that *d* is more problematic than all the other compounds, except *a*, whereas the eq 1 result states that *d* is more problematic than all substances, including *a*. The presence of two maximal objects in Figure 1D, *a* and *d*, is an indication of how risky it is to perform a weighted aggregation of descriptors because particular weights may lead to rankings with either *a* or *d* as the most problematic substances.

Avoidance of an aggregation function allows to prevent statistical overlap of descriptors, a fact often observed. Applying aggregation functions, descriptors must be statistically independent in order to avoid overdescription of certain features, in addition to the weights used in the aggregation.

The dominance degree

A set of substances G sometimes contains several "classes" of chemicals which can be found either by unsupervised classification methodologies such as cluster analysis or in a supervised manner. The questions arise whether it is possible to rank classes of compounds. This can be done with standard techniques of statistics (calculating medians, or means and perform a ranking based on them) but the order-theoretical approach of dominance degree *(17)* is preferable as it extends the parameter-free method of HDT to the ranking of classes. Two disjoint subsets G_n and G_m in G are formed of which G_n completely dominates G_m if for all x in G_n and for all y in G_m it holds that $y \leq x$. This condition "for all" implies that all objects in G_n are ranked higher than those in G_m. In practice, some objects of G_n may be incomparable with some of G_m, some objects in G_n may be ranked higher than some ones in G_m, while some others are ranked lower. Hence, it is necessary to quantify the number of objects in G_n ranked

higher in respect to those in G_m; this is called the dominance of G_n over G_m, measured as dominance degree.

The dominance degree is defined as $\mathrm{Dom}(G_n, G_m) = N_R / N_T$, where $N_R = |\{(x, y), x \in G_n, y \in G_m$, and $y \leq x\}|$ and $N_T = |G_n| * |G_m|$ ($|X|$: cardinality of the set X). Hence, $\mathrm{Dom}(G_n, G_m)$ is the fraction of total theoretical order relationships (N_T) for which the objects of G_n are ranked higher than those in G_m. $\mathrm{Dom}(G_n, G_m)$ ranges from 0 to 1; $\mathrm{Dom}(G_n, G_m) = 1$ means that all objects in G_n are ranked higher than those in G_m, i.e. subset G_n dominates subset G_m. When $\mathrm{Dom}(G_n, G_m) = 0$, no object in G_n is ranked higher than an object in G_m. In this work, values of $\mathrm{Dom}(G_n, G_m) > 0.5$ have been used for assessing the dominances between classes of refrigerants, meaning that more than half of the relations between G_n and G_m express a ranking where a compound in G_n is ranked higher than one in G_m.

To demonstrate the application of the dominance degree concept, the set G mentioned in Figure 1 is divided into three subsets, namely $G_1 = \{a, b, c\}$, $G_2 = \{d, e\}$ and $G_3 = \{f, g\}$ (Figure 2A). The resulting dominance degree values are: $\mathrm{Dom}(G_1, G_2) = 0 / (3 \times 2) = \mathrm{Dom}(G_1, G_3) = 0 / (3 \times 2) = \mathrm{Dom}(G_3, G_2) = 0 / (2 \times 2) = 0$, $\mathrm{Dom}(G_2, G_1) = 4 / (2 \times 3) = 0.67$, $\mathrm{Dom}(G_2, G_3) = 4 / (2 \times 2) = 1$, and $\mathrm{Dom}(G_3, G_1) = 2 / (2 \times 3) = 0.33$. In consequence, G_1 and G_3 do not dominate any subset because their values are lower than 0.5, while G_2 dominates G_1 and G_3. Dominance relationships are presented in the dominance diagram (Figure 2B) where a line is only drawn between subsets when $\mathrm{Dom}(G_n, G_m) > 0.5$ and G_n is located higher than G_m. The results indicate that 67 % of the chemicals in G_2 are more problematic than those in G_1, and that all elements in G_2 are more problematic than those in G_3. The percentage of chemicals dominated by G_n (PD_n) can be calculated by adding the number of chemicals in all subsets G_i dominated by G_n, dividing the result by the number of substances that might be dominated, i.e. $|G| - |G_n|$:

$$PD_n = \frac{\sum |G_i|}{|G| - |G_n|} \qquad (2)$$

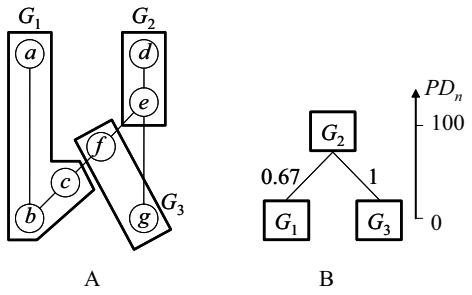


Figure 2. A) Hasse diagram endowed with three subsets. B) Dominance diagram, the numbers next to the lines are dominance degree values and the subsets are distributed according to their percentage of dominated substances (PD_n).

Classes of refrigerants and their properties

In this work, a set G comprising 40 refrigerants (Table 1) is divided into 13 subsets: CFC, HFC, HCFC, hydrocarbons (HC), di(fluoroalkyl)ethers (DFAE), alkylfluoroalkylethers (AFAE), chloromethanes (CM), and the single-compound subsets trifluoroiodomethane (FIM), octafluorocyclobutane (PFC), carbon dioxide (CO_2), bromochlorodifluorobutane (BCF), dimethyl ether (DME) and ammonia (NH_3).

ODP was originally defined (*8*) to represent the amount of ozone destroyed as a result of the emission of a gas. The numerical value is obtained by integrating over its entire atmospheric lifetime. Therefore, ODP is related to ALT (*7*); for example a chlorinated substance with a high ALT has a high probability of reacting with the stratospheric ozone. The compound R11 is used as reference for the calculation of ODP because it completely reacts in the stratosphere. Hence, ODP is only particularly appropriate for substances reacting in the same manner as R11, without tropospheric OH reaction. Additionally, the fact that the relative effect of a compound's emission on stratospheric ozone changes with time has lead to the definition of specific time horizons for ODP calculations (*18*).

GWP (*9*) is an index measuring the relative greenhouse efficiency of a gas in respect to carbon dioxide. GWP is also related to ALT in such a way that a chemical with high ALT and high infrared absorption holds high GWP. Considerations regarding the appropriateness of carbon dioxide as the reference substance have lead to propose new indices such as the Halocarbon Global Warming Potential (HGWP), considering R11 as reference substance (*19*).

Due to the comparative aim of the present paper, the refrigerants studied have to be described by indices, each one with a common reference substance. Therefore, chemicals studied are characterised by ODP relative to R11, GWP relative to CO_2, and ALT. All values were collected from the literature (Table 1). In the following the application of the HDT to these refrigerants is discussed as well as the dominance relationships among the 13 classes. As HDT does not use any aggregating function to combine descriptors but considers them simultaneous and independently, the fact that ODP and GWP are related to ALT does not result in an overestimation of the latter.

Table 1. Refrigerants included in this study, their labels, chemical subsets, molecular formulae, chemical and non-proprietory names, and their ODP, GWP, and ALT values.

Label	Subset	Molecular formula	Chemical name	Non-proprietory name	ODP [relative to R11]	GWP relative to CO_2 [100 yr time horizon]	ALT [yr]
1	CFC	CCl_3F	Trichloro-fluoromethane	R11	1[a]	4680[a]	45[a]
2	CFC	CCl_2F_2	Dichlorodi-fluoromethane	R12	0.82[a]	10720[a]	100[a]
3	HCFC	$CHClF_2$	Chlorodifluoro-methane	R22	0.05[a]	1780[a]	12[a]
4	HCFC	$C_2HCl_2F_3$	2,2-Dichloro-1,1,1-trifluoro-ethane	R123	0.022[a]	76[a]	1.3[a]
5	HCFC	C_2HClF_4	2-Chloro-1,1,1,2-tetrafluoro-ethane	R124	0.022[b]	599[a]	5.8[a]
6	HCFC	$C_2H_3Cl_2F$	1,1-Dichloro-1-fluoroethane	R141b	0.12[a]	713[a]	9.3[a]
7	HCFC	$C_2H_3ClF_2$	1-Chloro-1,1-difluoroethane	R142b	0.065[a]	2270[a]	17.9[a]
8	HFC	CHF_3	Trifluoro-methane	R23	0.0004[b]	14310[a]	270[a]
9	HFC	CH_2F_2	Difluoro-	R32	0[c]	670[a]	4.9[a]

			methane				
10	HFC	C_2HF_5	Pentafluoro-ethane	R125	0.00003[b]	3450[a]	29[a]
11	HFC	$C_2H_2F_4$	1,1,1,2-Tetrafluoro-ethane	R134a	0.000015[c]	1410[a]	14[a]
12	HFC	$C_2H_3F_3$	1,1,1-Trifluoroethane	R143a	0[c]	4400[a]	52[a]
13	HFC	$C_2H_4F_2$	1,1-Difluoroethane	R152a	0[d]	122[a]	1.4[a]
14	HFC	$C_3H_3F_5$	1,1,1,3,3-Pentafluoro-propane	R245fa	0[f]	950[e]	7.2[e]
15	HFC	$C_3H_2F_6$	1,1,1,3,3,3-Hexafluoro-propane	R236fa	0[f]	9400[e]	220[e]
16	HC	C_3H_8	*n*-Propane	R290	0[c]	20[c]	0.041[a]
17	HC	C_4H_{10}	*n*-Butane	R600	0[c]	20[c]	0.018[a]
18	HC	C_4H_{10}	Isobutane	R600a	0[d]	20[d]	0.019[a]
19	HC	C_5H_{12}	*n*-Pentane	R601	0[g]	0[h]	0.01[a]
20	HC	C_3H_6	Propene	R1270	0[c]	3[i]	0.001[a]
21	CO_2	CO_2	Carbon dioxide	R744	0[d]	1[b]	120[j]
22	BCF	$CBrClF_2$	Bromochloro-difluoro-methane	R12B1	5.1[a]	1300[e]	11[e]
23	PFC	C_4F_8	Octafluoro-cyclobutane	RC318	0[f]	10000[f]	3200[f]
24	HFC	C_3HF_7	1,1,1,2,3,3,3-Heptafluoro-propane	R227ea	0[f]	3500[e]	33[e]
25	AFAE	$C_4H_3F_7O$	Heptafluoro-propyl methyl-ether	HFE-7000	0[a]	450[k]	4.7[k]
26	AFAE	$C_5H_3F_9O$	Methyl-nonafluoro-butyl ether	HFE-7100	0[a]	410[k]	5[k]
27	AFAE	$C_6H_5F_9O$	Ethyl-nonafluoro-butyl ether	HFE-7200/ HFE-569mccc	0[a]	60[k]	0.77[k]
28	AFAE	$C_9H_5F_{15}O$	Ethyl-pentadeca-fluoro heptyl-ether	HFE-7500	0[a]	100[k]	2.2[k]
29	DFAE	C_2HF_5O	Pentafluoro-dimethyl ether	HFE-125	0[a]	14800[k]	165[k]
30	DFAE	$C_2H_2F_4O$	1,1,1',1'-Tetrafluoro-dimethyl ether	HFE-134	0[a]	5760[k]	27.25[k]
31	CM	CH_2Cl_2	Methylene-chloride	R30	0[f]	10[a]	0.38[a]
32	CM	CH_3Cl	Methyl-chloride	R40	0.02[a]	16[a]	1.3[a]
33	CFC	$C_2Cl_3F_3$	1,1,2-Trichloro-1,2,2-trifluoro-	R113	0.9[a]	6000[f]	85[a]

			ethane				
34	HCFC	$CHCl_2F$	Dichlorofluoro-methane	R21	0.01[f]	210[e]	2[e]
35	CFC	$C_2Cl_2F_4$	1,2-Dichloro-1,1,2,2-tetrafluoro-ethane	R114	0.85[f]	9800[e]	300[a]
36	FIM	CF_3I	Trifluoroiodo-methane	R13I1	0[f]	1[e]	0.1[f]
37	DME	C_2H_6O	Dimethyl ether		0[f]	1[a]	0.015[a]
38	NH_3	NH_3	Ammonia	R717	0[c]	0[i]	0.25[a]
39	AFAE	$C_2H_3F_3O$	Methyl-trifluoromethyl ether	HFE-143	0[a]	656[k]	5.7[k]
40	AFAE	$C_3H_3F_5O$	Methyl-pentafluoro-ethyl ether	HFE-245	0[a]	697[k]	4[k]

References: [a] *20*, [b] *21*, [c] *22*, [d] *23*, [e] *9*, [f] *24*, [g] *25*, [h] *26*, [i] *27*, [j] *28*, [k] *29*.

Results and discussion

In order to determine the diversity of the set divided into 13 classes, the Simpson diversity index *D* (*30*) has been calculated. The obtained value, $D = 0.99$, shows that the set is large-diverse, which ensures no overpopulation of one class in comparison with others. The HD of the 13 subsets is shown in Figure 3 where substances at the top of the diagram are the most problematic ones, those at the bottom are the least problematic ones.

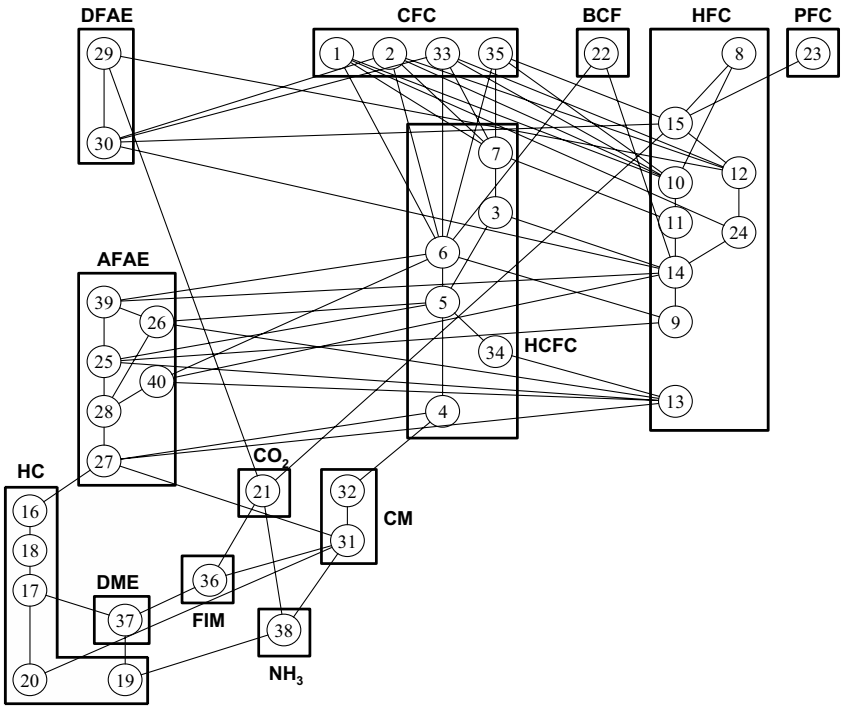


Figure 3. Hasse diagram of 40 refrigerants and its 13 classes shown as boxes.

Eight maximal refrigerants with high impact in respect to ALT, ODP and GWP, and two minimal substances are shown. The maximal ones belong to DFAE, CFC, BCF, HFC and

PFC, the minimal ones to HC. Not all the members of these subsets are maximal substances, for example 1,1,1',1'-tetrafluoro-dimethyl ether, **30**, is not a maximal one, although it is a DFAE, while **29** is a maximal substance. A similar situation is found for HFC, with the only maximal substance **8**. In contrast, all CFCs are maximal chemicals. In the same way, *n*-pentane, **19**, and propene, **20**, are minimal substances but this does not apply to all hydrocarbons.

Dominance degrees among the 13 classes are calculated. These values appear in Table 2 as a square matrix. The dominance values in Table 2 correspond to Dom(G_n, G_m) where G_n is always a class labeling a column and G_m a class labeling a row. The matrix is not symmetrical due to the order properties on which it is based; therefore Dom(G_n, G_m) can be different to Dom(G_m, G_n).

Table 2. Dominance matrix of the 13 classes of refrigerants.

	CFC	HFC	PFC	DFAE	AFAE	HCFC	CM	FIM	HC	CO_2	BCF	DME	NH_3
CFC	ND[a]	0	0	0	0	0	0	0	0	0	0	0	0
HFC	0.78	ND	0.63	0.44	0.07	0.2	0	0	0	0	0.33	0	0
PFC	0	0	ND	0	0	0	0	0	0	0	0	0	0
DFAE	0.38	0.11	0.5	ND	0	0	0	0	0	0	0	0	0
AFAE	1	0.85	1	1	ND	0.73	0	0	0	0	1	0	0
HCFC	1	0	0	0	0	ND	0	0	0	0	0.67	0	0
CM	1	0.5	0.5	0.5	0.5	0.92	ND	0	0	0	1	0	0
FIM	1	1	1	1	1	1	1	ND	0	1	1	0	0
HC	1	1	1	1	1	1	0.4	0.2	ND	0.2	1	0.2	0.2
CO_2	0.25	0.22	1	0.5	0	0	0	0	0	ND	0	0	0
BCF	0	0	0	0	0	0	0	0	0	0	ND	0	0
DME	1	1	1	1	1	1	1	1	0.6	1	1	ND	0
NH_3	1	1	1	1	1	1	1	0	0	1	1	0	ND

[a] The dominance degree for diagonal elements is not defined (ND).

Potentially, there are 156 dominance relationships among the 13 classes ($13 \times 13 = 169$ minus 13 diagonal elements, Table 2), one third of which corresponds to Dom(G_n, G_m) > 0.5, and two thirds to Dom(G_n, G_m) ≤ 0.5. There are 27.6% of total dominances (Dom(G_n, G_m) = 1) and 55.8% of no dominances (Dom(G_n, G_m) = 0). The corresponding dominance diagram is shown in Figure 4.

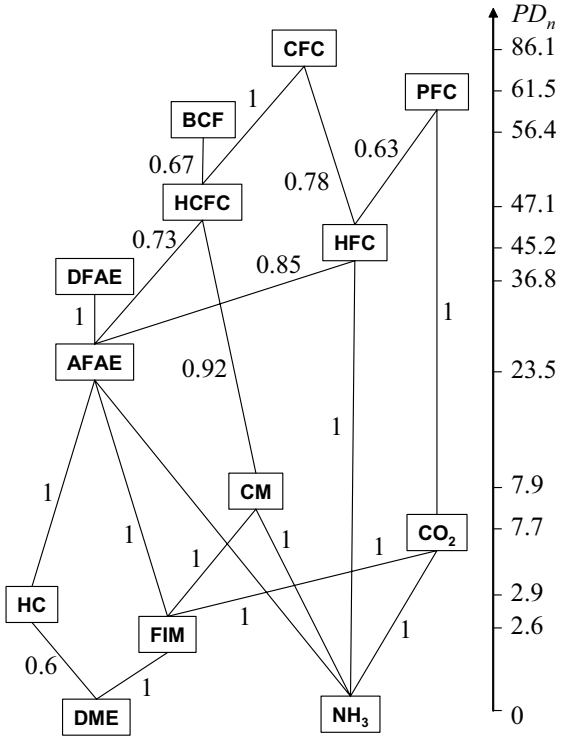


Figure 4. Dominance diagram for the 13 classes of refrigerants (Dom(G_n, G_m) > 0.5); the numbers next to the lines are the dominance degree values and the classes are distributed according to their PD_n (Eq. 2).

Depending on the particular order relationships among the considered classes, a dominance diagram may or may not fulfil the transitivity axiom, i.e. if class *A* dominates class *B* and *B* dominates class *C* then *A* dominates *C* (*31*). If the axiom is met, the dominance of *A* over *C* is graphically represented by the dominance of *A* over *B* and of *B* over *C*. In the present case of 13 refrigerant classes, the transitivity axiom is fulfilled.

According to Figure 4, CFC is the class that dominates most other substances. Each of the classes CFC, PFC and BCF dominate more than half of the refrigerants considered in respect to ODP, GWP and ALT. The second generation alternatives, HCFC and HFC, dominate less than half of the other substances, which means that they are environmentally better than CFC, PFC and BCF, substances that replace HCFC and HFC. Although problematic HCFCs will be replaced by HFC-blends in refrigeration equipment before 2010 (*32*), it is worthy to note that HCFC does not dominate HFC; therefore the environmental suitability of the latter as replacements is questionable. Particularly three blends, namely R410A, R407C and R404A, will replace chlorodifluoromethane (**3** in Table 1 and Figure 4). R410 is a blend of difluoromethane, **9**, and pentafluoroethane, **10**; R407C a blend of **9**, **10** and 1,1,1,2-tetrafluoroethane, **11**; and R404A a blend of **10**, **11** and 1,1,1-trifluoroethane, **12**. Only **9** is ranked lower than **3** whereas the other HFCs used in the blends are incomparable with **3** (Figure 3).

The classes DFAE and AFAE, both hydrofluoroethers, appear at lower PD_n values (Figure 4). DFAE dominates 37% of the other chemicals, which is a value neighboured to that one of HFC (45%) in the PD_n axis. None of the classes of chemicals studied dominates DFAE, not even CFC accounting for the largest percentage of domination. Therefore, it is not possible to state that chemicals belonging to DFAE are less problematic than CFC, PFC or BCF, although these hydrofluoroethers were introduced as CFC replacements. On the other hand,

AFAE, the other group of hydrofluoroethers, is dominated by the most problematic refrigerants including the DFAE. This is possibly caused by the particular distribution of fluorine atoms along the molecules; DFAE compounds hold fluorine substituents in both alkyl groups, whereas AFAE compounds have fluorine substituents on only one of the alkyl groups. This seems to be an important structural aspect related to the environmental properties of hydrofluoroethers. This finding indicates a need to carry out studies in this direction; some preliminary structure-property relationships investigations have been done on their tropospheric lifetimes (*33*, *34*).

There are six classes of refrigerants with PD_n values (Figure 4) lower than 8%, which are CM, CO_2, HC, FIM, DME and NH_3. They constitute the environmentally most acceptable refrigerants. It is particularly important to note that CFC, HCFC, HFC, DFAE and AFAE dominate HC and NH_3, two substances which earlier were considered as problematic and which motivated the development of CFC in the 1930s (*35*). Therefore, when comparing HC and NH_3 with their replacements in respect to the three descriptors ODP, GWP and ALT, the former are better. However, in order to develop a more general ranking, other aspects must be considered, such as energy efficiency, toxicity and flammability, properties which are also important for practical applications. Qualitatively it can be foreseen that in respect to flammability, DME and HC are the most problematic, that in respect to energy efficiency carbon dioxide and HC are the least recommendable, and that particular attention must be paid to toxicity in case of ammonia. Nevertheless, their simultaneous analysis by applying a mathematical technique such as HDT is an additional, non subjective, instrument for finding the least problematic compounds.

Information on the relative order among classes is based on the ranking of chemicals which in turn depends on the numerical values of the properties selected for their description. Small variations of these values may potentially affect the order relationships among them. In order to study this effect, each of the three environmental properties, continuous in concept, was classified and the effect on the dominance degree values was studied. The three properties were transformed into 37 scores by partitioning each property into 37 equidistant intervals. Differences between the original dominance degree values and those obtained after classification were calculated; the average variation of these differences was 0.11, indicating that the effect of classification on the dominance relationships is about 11 %, i.e. 89 % of the dominance relationships remained invariant towards property-classification. Thus, dominance relationships found in this research are robust (*12*) with respect to numerical noise.

The main aim of this manuscript was to explore the order relationships among classes of chemicals; there are some other studies (*12*) that can be done based on HD, such as (a) stability analysis (*36*) of the diagram under addition or deletion of properties, (b) study of the most influential properties on the structure of the diagram (sensitivity analysis (*12*, *37*)); (c) application of dimension analysis (*16*) to know if the same diagram can be obtained combining some non-redundant properties; and (d) step-by-step weighted aggregation of descriptors to obtain a linear ranking. Results on the application of these methodologies to the refrigerants will be published in forthcoming papers.

Although we consider in this research some representative refrigerants, the method described can be applied to any number of substances. In fact, a HD compares objects' descriptor values without regarding the number of objects. Therefore, such dominance degree calculations are not affected by the size of the classes.

Acknowledgements

The authors thank the Bavarian Environmental Agency for supporting this study under the Research Project 81-00213381. G. Restrepo specially thanks COLCIENCIAS and the Universidad de Pamplona for the grant offered during the development of this research.

References

(1) Calm, J. M.; Didion, D. A. Proceedings of the ASHRAE/NIST Refrigerants Conference, Gaithersburg, 6-7 October 1997.

(2) UNEP. *Montreal Protocol on Substances that Deplete the Ozone Layer*; United Nations Environment Programme: Nairobi, Kenya, 1987.

(3) UNEP. *Montreal Protocol on Substance that Deplete Ozone Layer*; United Nations Environment Programme: Montreal, Canada, 1998.

(4) Kyoto Protocol to the United Nations Framework Convention on Climate Change, United Nations Framework Convention on Climate Change, 1997.

(5) Bovea, M. D.; Cabello, R.; Querol, D. Comparative Life Cycle Assessment of Commonly Used Refrigerants in Commercial Refrigerants Systems. *Int. J. Life Cycle Ass.* **2007**, *12*, 299-307.

(6) Molina, M. J.; Rowland, F. S. Stratospheric Sink for Chlorofluoromethanes: Chlorine Atom-Catalysed Destruction of Ozone. *Nature* **1974**, *249*, 810-812.

(7) Kurylo, M. J.; Orkin, V. L. Determination of Atmospheric Lifetimes via the Measurement of OH Radical Kinetics. *Chem. Rev*. **2003**, *103*, 5049-5076.

(8) Wuebbles, D. J. Chlorocarbon Emission Scenarios - Potential Impact on Stratospheric Ozone. *J. Geophys. Res*. **1983**, *88*, 1433-1443.

(9) *Climate Change 2001: The Scientific Basis; Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: New York, 2001.

(10) Bolin, B.; Rodhe, H. A Note on the Concepts of Age Distribution and Transit Time in Natural Reservoirs. *Tellus* **1973**, *25*, 58-62.

(11) Davis, G. A.; Swanson, M.; Jones, S. *Comparative Evaluation of Chemical Ranking and Scoring Methodologies*; EPA Order No. 3N-3545-NAEX, 1994.

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

(13) Brüggemann, R.; Münzer, B. A Graph-Theoretical Tool for Priority Setting of Chemicals. *Chemosphere* **1993**, *27*, 1729-1736.

(14) Seip, K. L. Restoring Water Quality in the Metal Polluted Soerfjorden, Norway. *Ocean Coast. Manage.* **1994**, *22*, 19-43.

(15) Brüggemann, R.; Halfon, E.; Bücherl, C. 1995. Theoretical Base of the Program "Hasse", GSF–Bericht 20/95, Neuherberg.

(16) Trotter, W. T. *Combinatorics and Partially Ordered Sets Dimension Theory*; The Johns Hopkins University Press: Baltimore, 1992.

(17) Restrepo, G.; Brüggemann, R. Partially Ordered Sets in the Analysis of Alkanes Fate in Rivers. *Croat. Chem. Acta* **2007**, *80*, 261-270.

(18) Solomon, S.; Albritton, D. L. Time-Dependent Ozone Depletion Potentials for Short- and Long-Term Forecasts. *Nature* **1992**, *357*, 33-37.

(19) Fisher, D. A.; Hales, C. H.; Wang, W.-C.; Ko, M. K. W.; Sze, N. D. Model Calculations of the Relative Effects of CFCs and their Replacements on Global Warming. *Nature* **1990**, *344*, 513-516.

(20) IPCC/TEAP. (Intergovernmental Panel on Climate Change / Technology and Economic Assessment Panel). *Special Report on Safeguarding the Ozone Layer and the Global Climate System; Issues related to Hydrofluorocarbons and Perfluorocarbons;* Cambridge University Press: Cambridge, 2006.

(21) WMO. (World Meteorological Organization). *Scientific Assessment of Ozone Depletion: 2002*; Global Ozone Research and Monitoring Project - Report No. 47, Geneva, Switzerland , 2003.

(22) FKW. (Forschungszentrum für Kältetechnik und Wärmepumpen GmbH). *Ersatz des Kältemittels R22 in bestehenden Kälte- und Klimaanlagen* - Aktueller Stand - Studie im Auftrag des Umweltbundesamtes, Hannover, 2000.

(23) Devotta, S.; Padalkar, A. S.; Sane, N. K. Performance Assessment of HC-290 as a Drop-In Substitute to HCFC-22 in a Window Air Conditioner. *Int. J. Refrig.* **2005**, *28*, 594-604.

(24) Calm, J. M.; Hourahan, G. C. Refrigerant Data Summary. *Engineered Syst*. **2001**, *18*, 74-88.

(25) von Cube, H. L.; Steinle, F.; Lotz, H.; Kunis, J. *Lehrbuch der Kältetechnik*, Band 1, 4. Auflage; C. F. Müller Verlag: Heidelberg, 1997.

(26) Galvin, J. B.; Marashi, F. *n*-Pentane. *J. Toxicol. Env. Heal. A* **1999**, *58*, 35-56.

(27) Bitzer International. 2004. Kältemittelreport 13. Auflage A-500-13.

(28) Nieto de Castro, C. A.; Mardolcar, U. V.; Matos Lopes, M. L. Thermophysical Properties of Environmentally Acceptable Refrigerants. In *Stratospheric Ozone Depletion/ UV-B Radiation in the Biosphere*; Biggs, R. H.; Joyner, M. E. B., Eds.; Springer: Berlin, 1994; pp 27-34, Vol. 18.

(29) Tsai, W. T. Environmental Risk Assessment of Hydrofluoroethers (HFEs). *J. Hazard. Material.* **2005**, *119*, 69-78.

(30) Kotz, S., Johnson, N. L. *Encyclopedia of Statistical Sciences; Volume 2;* John Wiley & Sons: New York. 1982.

(31) Restrepo, G.; Brüggemann, R. Dominance and Separability in Posets, Their Application to Isoprotonic-Isoelectronic Species. Submitted to *J. Math. Chem.* **2007**.

(32) Tullo, A. H. The Switch Is On For Refrigerants. *Chem. Eng. News* **2006**, *84*, 24-25.

(33) Cooper, D. L.; Cunningham, T. P.; Allan, N. L.; McCulloch, A. Potential CFC Replacements: Tropospheric Lifetimes of C_3 Hydrofluorocarbons and Hydrofluoroethers. *Atmos. Environ. A-Gen*. **1993**, *27*, 117-119.

(34) Güsten, H.; Medven, Ž.; Sekušak, S.; Sabljić, A. Predicting Tropospheric Degradation of Chemicals: From Estimation to Computation. *SAR & QSAR Environ. Res*. **1995**, *4*, 197-209.

(35) Powell, R. L. CFC Phase-Out: Have We Met the Challenge? *J. Fluorine Chem.* **2002**, *114*, 237-250.

(36) Brüggemann, R.; Voigt, K. Stability of Comparative Evaluation, -Example: Environmental Databases. *Chemosphere* **1996**, *33*, 1997-2006.

(37) Brüggemann, R.; Münzer, B.; Halfon, E. An Algebraic/Graphical Tool to Compare Ecosystems with Respect to Their Pollution - the German River “Elbe” as an Example - I: Hasse-Diagrams. *Chemosphere* **1994**, *28*, 863-872.

Curriculum Vitae

Guillermo Restrepo

Birthday: 12th of August 1976
Place of birth: Bogotá, Colombia

Education:

Winter 2005-	Dr. rer. nat., University of Bayreuth, Germany (scheduled graduation)
2003	MSc Chemistry, Universidad Industrial de Santander, Colombia
1998	Chemistry, Universidad Industrial de Santander, Colombia

Work experience:

2004-	Chemistry professor, Departamento de Química, Universidad de Pamplona, Colombia
2002-2003	Chemistry professor, Facultad de Ingenierías, Universidad Pontificia Bolivariana, Colombia
1999-2004	Chemistry professor, Escuela de Química, Universidad Industrial de Santander, Colombia

Academic awards:

2005	PhD grant from the Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología "Francisco José de Caldas" (COLCIENCIAS) and the Universidad de Pamplona.
2005	Invited speaker, Physics Department, Southern Adventist University, USA

2005 Guest editor, WSEAS Transactions on Information Science and Applications, Greece

2004 Grant from the Chemical Structure Association Trust (UK) and the Molecular Graphics and Modelling Society (UK) for presenting research results at the "Third Joint Sheffield Conference on Chemoinformatics", The University of Sheffield, UK

2003 Laureated MSc thesis, Universidad Industrial de Santander, Colombia

1995-1997 Distinguished student, Universidad Industrial de Santander, Colombia

Participation in scientific events:

EnviroInfo 2007, 21st Environmental Informatics and Systems Research, Warsaw, Greece, September 12-14, 2007.

The Twenty-second International Course & Conference on the Interfaces among Mathematics, Chemistry & Computer Sciences, Dubrovnik, Croatia, June 11-16, 2007.

International Conference of Computational Methods in Sciences and Engineering 2006, Chania, Greece, October 30-November 1, 2006.

Workshop on ranking methods and multicriteria decision analysis in environmental sciences, Verbania, Italty, October 2-3, 2006.

International Congress of Mathematicians 2006, Madrid, Spain, August 22-30, 2006.

XXXI International Congress of Theoretical Chemists of Latin Expression, Margarita Island, Venezuela, October 2-6, 2005.

9th WSEAS International Conference on COMPUTERS, Atenas, Greece, July 14-16, 2005.

Fourth Indo-US Workshop on Mathematical Chemistry, Pune, India, January 8-12, 2005.

International Conference of Computational Methods in Sciences and Engineering 2004, Athens, Greece, November 19-23, 2004.

The Nineteenth International Course & Conference on the Interfaces among Mathematics, Chemistry & Computer Sciences, Dubrovnik, Croatia, June 21-26, 2004.

Third Joint Sheffield Conference on Chemoinformatics, Sheffield, UK, April 21-23, 2004.

V Seminars of Advances Studies on Molecuar Design and Bioinformatics, La Habana, Cuba, February 2-6, 2004.

4th Northeast Simposium of Mathematics, Bucaramanga, Colombia, December 9-12, 2003.

Regional Conference on Learning Assessment and University Teaching, Bucaramanga, Colombia, December 5-7, 2003.

1st National Chemistry Week, Bucaramanga, Colombia, October 27-31, 2003.

The Second Harry Wiener International Memorial Conference: The Periodic Table, into the 21st century, Banff, Canada, July 14-20, 2003.

Own scientific events organisation:

Special Session on Mathematical Chemistry, at the 9th WSEAS International Conference on COMPUTERS, Athens, Greece, July 14-16, 2005.

Course: Quantum Similarity, Bogotá, Colombia, taught by Professor Ramón Carbó-Dorca (Girona Universitat, Gent Universitaet), April 18-22, 2005.

1st National Workshop of Theoretical Chemists, Pamplona, Colombia, August 8-12, 2004.

Course: Mathematical Chemistry: Periodicity, Pamplona, Colombia, taught by Professor Ray Hefferlin (Southern Adventist University), July 12-16, 2004.

Course: Mathematical Chemistry, Bucaramanga, Colombia, taught by Professors José L. Villaveces and Edgar E. Daza (Universidad Nacional de Colombia), November 12-16, 2001.

Publications:

Restrepo, G.; Weckert, M.; Brüggemann, R.; Gerstmann, S.; Frank, H. Refrigerants ranked by partial order theory. In *EnviroInfo 2007, 21st international conference on informatics for environmental protection*; Hryniewicz, O.; Studziński, J.; Szediw, A., Eds.; Shaker: Aachen, Germany, 2007; pp 209-217.

Restrepo, G.; Brüggemann, R. Partially ordered sets in the analysis of alkanes fate in rivers. *Croat. Chem. Acta* **2007**, *80*, 261-270.

Restrepo, G.; Mesa, H.; Llanos, E. J. Three dissimilarity measures to contrast dendrograms. *J. Chem. Inf. Model.* **2007**, *47,* 761-770.

Restrepo, G., Pachón, L. Mathematical aspects of the periodic law. *Found. Chem.* **2007**, *9*, 189-214.

Restrepo, G.; Brüggemann, R. Modelling the fate of alkanes in rivers. In *Recent progress in computational sciences and engineering*; Simos, T.; Maroulis, G., Eds.; VSP: Leiden, Netherlands, 2006; pp 1386-1389.

Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. In *The mathematics of the periodic table*; King, B.; Rouvray, D., Eds.; Nova: New York, USA, 2006; Chapter 5, pp. 75-100.

Brüggemann, R.; Restrepo, G.; Voigt, K. Structure-fate relationships of organic chemicals derived from the software packages E4CHEM and WHASSE. *J. Chem. Inf. Model.* **2006**, *46*, 894-902.

Restrepo, G.; Llanos, E. J.; Mesa, H. Topological space of the chemical elements and its properties. *J. Math. Chem.* **2006**, *39*, 401-416.

Restrepo, G.; Llanos, E. J.; Mesa, H. On the topological sense of chemical sets. *J. Math. Chem.* **2006**, *39*, 363-376.

Daza, M. C.; Restrepo, G.; Uribe, E. A.; Villaveces, J. L. Quantum chemical and chemotopological study of fourth row monohydrides. *Chem. Phys. Lett.* **2006**, *428*, 55-61.

Restrepo, G.; Brüggemann, R. Ranking regions through cluster analysis and posets. *WSEAS Trans. Inf. Sci. Appl.* **2005**, *2*, 976-981.

Brüggemann, R.; Restrepo, G.; Voigt, K. Towards an evaluation of chemicals. *WSEAS Trans. Inf. Sci. Appl.* **2005**, *2*, 1023-1033.

Uribe, E. A.; Daza, M. C.; Restrepo, G. Chemotopological study of the fourth period mono-hydrides. *WSEAS Trans. Inf. Sci. Appl.* **2005**, *2*, 1085-1090.

Restrepo, G.; Villaveces, J. L. From trees (dendrograms and consensus trees) to topology. *Croat. Chem. Acta* **2005**, *78*, 275-281.

Restrepo, G. Química matemática y la Universidad de Pamplona. *Bistua* **2005**, *3*, 61-76.

Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 68-75.

Restrepo, G. Los elementos químicos, su matemática y relación con el sistema periódico. *Bistua* **2004**, *2*, 91-98.

Restrepo, G. Reseña: Para enseñar no basta con saber la asignatura. *Doc. Univ. UIS* **2003**, *4*, 157-159.

Restrepo, G. ¿Cambio físico o químico? La clasificación en el camino del aprendizaje. *Doc. Univ. UIS* **2003**, *3*, 57-64.

Kouznetsov, V.; Zubkov, F.; Palma, A.; Restrepo, G. A simple synthesis of spiro-C_6-annulated hydrocyclopenta[g]indole derivatives. *Tetrahedron Lett.* **2002**, *43*, 4707-4709.

Pregraduate thesis advised

Emilbus A. Uribe. Estudio quimiotopológico de los monohidruros del cuarto periódo. 2005. Universidad Industrial de Santander (Colombia).

Erklärung zur vorgelegten schriftlichen Leistung

Hiermit erkläre ich an Eides statt,

dass ich die vorliegende Dissertationsschrift selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Hiermit erkläre ich an Eides statt,

dass ich weder die vorliegende noch eine gleichartige Doktorprüfung an einer anderen Hochschule endgültig nicht bestanden habe.

(gez. Guillermo Restrepo)