ENVIRONMETRICS
Environmetrics 2004; **15**: 577–596 (DOI: 10.1002/env.653)

Data analysis of environmental air pollutant monitoring systems in Europe

Kristina Voigt[1*,†], Gerhard Welzl[2] and Rainer Brüggemann[3]

[1]*GSF—National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Department of Biostatistics, 85764 Neuherberg, Germany*
[2]*GSF—National Research Center for Environment and Health, Institute of Development Genetics, 85764 Neuherberg, Germany*
[3]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Department of Ecohydrology, 12587 Berlin, Germany*

SUMMARY

Public access to information about the environment is being strengthened across Europe. The concept of public's right to information gives the basis for the access to environmental information. In this paper the quality of air pollutant monitoring systems in the 15 European member states is analyzed. For pragmatic reasons only the capitals are looked upon. Comprehensive data on environmental monitoring programs concerning air pollutants like ozone (O_3), nitrogen dioxide (NO_2), nitrogen oxide (NO), carbon dioxide (CO_2) and carbon monoxide (CO), and sometimes suspended dust, benzene and other environmental chemicals are available on the free Internet. As different monitoring information systems exist in the European member states a comparison of these systems with their pros and cons is of great interest to the public. Environmental air pollutant monitoring systems in the capitals of the 15 EEC member countries (objects) are evaluated by applying 5 evaluation criteria for the differentiation of these systems. The scores run from 0 = insufficient, 1 = medium, to 2 = excellent.

Different data-analysis methods will be applied. As order theory is still not sufficiently presented in the scientific literature, a short overview about the so-called Hasse diagram technique and POSAC method is outlined. In several steps the data-matrix is analyzed coming to the conclusion that all methods (additionally PCA is used) identify one criterion as specifically important. Not unexpected, each method has its own advantage. The aim of this data-analysis is the evaluation of the publicly available air quality monitoring systems in Europe with their pros and cons. This might help the interested public to find and understand the information given on the Internet. Furthermore our evaluation approach might give some recommendations for an improvement of the air quality monitoring systems.

KEY WORDS: air pollution monitoring systems; multi-variate statistics; discrete mathematics; Hasse-diagram technique; POSAC; PCA

1. INTRODUCTION: INFORMATION OVERFLOW AND FREEDOM OF ENVIRONMENTAL INFORMATION

We are living in a world of information proliferation. A very common saying in this respect is the following: We are drowning in information but starving for knowledge (Roger, 2003). Another true saying is:

*Correspondence to: K. Voigt, GSF—National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Department of Biostatistics, 85764 Neuherberg, Germany.
†E-mail: kvoigt@gsf.de

Published online 21 June 2004
Received 17 March 2003
Accepted 18 October 2003

Where is the knowledge? Lost in the information.
Where is the information? Lost in the data.
Where is the data? Lost in the database.
Where is the database? Lost in the Internet (Streuff, 2000).

Public access to information about the environment is being strengthened across Europe. The concept of the public's right to information replaces the freedom of information, a principle enshrined in the European Union's 1990 directive (Directive 90/313/EEC, of 7 June 1990), on access to environmental information. This concept provides the basis for the free access to all environmental information measured by public authorities throughout the EEC.

Concerning environmental information, a proposal for a Directive of the European Parliament and of the Council on public access to environmental information exists (Eu-Lex, 2002). Environmental objectives to be achieved by this directive are:

preserving, protecting and improving the quality of the environment;
protecting human health;
prudent and rational utilization of natural resources;
promoting measures at the international level to deal with regional or world-wide environmental problems.

Giving public access to environmental information is essential to achieve these aims. The second, but not less important, step is the analysis of these environmental data. Mathematical and statistical groups should actively influence the interpretation of environmental and chemical data given in and extracted from the Internet in order to receive effective and correct results and support decisions derived from these data.

After a brief discussion of air pollution monitoring systems in Europe the data matrix is introduced by which the evaluation of the monitoring systems will be performed. As order theory is still not sufficiently presented in the scientific literature, a short overview about the so-called Hasse diagram technique and POSAC method is outlined. In several steps the data matrix is analyzed, producing the conclusion that all methods (additionally PCA is used) identify one criterion as specifically important. Not unexpectedly, each method has its own advantage.

2. AIR POLLUTION MONITORING SYSTEMS IN EUROPE

Providing public access to environmental air monitoring systems and their results is seen as a major topic in fulfilling the above-mentioned directive. However, it is very time-consuming to find these air quality monitoring systems in Europe and to effectively use the information found in these systems. A helpful URL is the following: http://www.stadtklima.de/stuttgart/s-luft/links.htm. This site provides links to all environmental air monitoring systems in the 16 states of Germany as well as links to a few European countries. The evaluation of the contents and quality of the German air quality monitoring systems is given by Voigt *et al.* (2002).

In this article the quality of air pollutant monitoring systems in the 15 European member states is analyzed. Comprehensive data on environmental monitoring programs concerning air pollutants like ozone (O_3), nitrogen dioxide (NO_2), nitrogen oxide (NO), carbon dioxide (CO_2) and carbon monoxide (CO), and sometimes suspended dust, benzene and other environmental chemicals, are available on the free Internet. As different monitoring information systems exist in the European member states a comparison of these systems with their advantages and disadvantages is of great interest to the public. This analysis of the 15 systems with different methods hopefully leads to an improvement of the

Table 1. Air-monitoring systems in European capitals

Country	Abb.	Name of capital	URL
Austria	AUS	Vienna	http://www.ubavie.gv.at/index.htm
Belgium	BEL	Brussels	http://www.irceline.be/ ~ celinair/english/homeen_nojava.html
Denmark	DEN	Copenhagen	http://www.dmu.dk/AtmosphericEnvironment/aq_aar/aovers.htm
Finland	FIN	Helsinki	http://www.ytv.fi/english/air/index.html
France	FRA	Paris	http://www.airparif.asso.fr/
Germany	GER	Berlin	http://www.met.fu-berlin.de/senum/
Greece	GRE	Athens	http://www.minenv.gr/welcome_en.html
Ireland	IRE	Dublin	http://www.environ.ie/environ/envindex.html
Italy	ITA	Rome	http://192.167.230.2:80/disia/Eng/DisiaEng.htm
Luxembourg	LUX	Luxembourg	http://www.restena.lu/meteo_lcd/
Netherlands	NET	The Hague	http://www.lml.rivm.nl/
Portugal	POR	Lisbon	http://www.drarn-lvt.pt/cga/l/qualidade/index.html
Spain	SPA	Madrid	http://vaxc.middlesex.ac.uk/emma/project/air_maps/24hr_for/m24_no2.htm
Sweden	SWE	Stockholm	http://www.ivl.se/en/miljo/projects/urban/
United Kingdom	UNK	London	http://www.erg.kcl.ac.uk/london/asp/home.asp

systems. Environmental air pollutant monitoring systems in the capitals of the 15 EEC member countries (objects) were evaluated by applying 5 evaluation criteria for the differentiation of these systems (see Table 1).

3. DATA-MATRIX: AIR MONITORING INFORMATION SYSTEMS IN 15 EUROPEAN COUNTRIES, 5 EVALUATION CRITERIA

The quality of air pollutant monitoring systems in 15 European countries (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden and the United Kingdom) is analyzed. For pragmatic reasons only the capitals are considered. The analysis and the searches on the Internet were performed in the months of March, April 2002. The authors are aware of the fact that the Internet offers are constantly changing and improving and that it is quite possible that an air monitoring information system was not found or unavailable during the time of the search or that existing systems were improved in the meantime. We would like to encourage all those experts who have further knowledge on the systems to inform us.

In order to make distinctions among the air monitoring systems we need evaluation criteria. Not only chemical parameters but also parameters concerning the presentation and explanation of the results are chosen as criteria.

- **Number of chemicals monitored in the system (NU)**. Many air-pollutant monitoring systems only measure the common air pollutants like ozone, carbon dioxide (CO_2), carbon monoxide (CO), nitrogen dioxide (NO_2) and nitrogen oxide (NO). Furthermore, other monitoring stations provide measurements for organic chemical substances like benzene, toluene, xylene, polyaromatic hydrocarbons, inorganic substances like cadmium, lead, mercury, as well as suspended dust. This is a *quantitative* criterion which describes the number of pollutants. The type of pollutant, however, would need a *qualitative* criterion. Thus NU is on one side a quantitative measure; however, by the selection of the kind of chemicals some qualitative arguments are introduced.
- **Type and duration of measurements (ME)**. The actual measurements are given most of the time. Additionally, the developments and trends of the air pollution are of great concern to the interested

public. This means that data of recent measurements and measurements of former months and years are provided. The user can distinguish trends and can see whether the pollution worsens or improves. This criterion describes the *temporal* aspect of the monitoring situation.

- **Measurement stations in capitals (ST)**. In many capitals several monitoring stations exist. For pragmatic reasons no correlation between the size of the capital and the number of monitoring stations is made. For the user, however, this evaluation criterion is rather important, as he/she wants to know the state of the air pollution in his/her vicinity. The more stations are available the more informative the whole system in the capital is. This criterion describes the *spatial* aspect of the monitoring systems.
- **Method of data presentation on the Internet (PR)**. The presentation of the monitoring results varies a lot from information system to information system. Often only tables are given. These are, of course, less informative than those systems which give graphical illustrations. The best systems are those which offer graphics as well as tables. This is a *descriptive* criterion.
- **Background material (BM)**. The user of these information systems needs to understand the data and information given. In general, users are no experts in monitoring systems and environmental pollution. They are not familiar with the limit values and the background laws and regulations. That is the reason why the explanation and background material has to be offered online. This has to be done in a comprehensive manner. In this respect it is very important to know what measures have to be taken, if the limit values of air pollutants are exceeded, etc. This criterion is also a *descriptive* one.

The given evaluation criteria encompass the three dimensions air quality indices show, namely time, space and type(s) of pollutant(s). A unified strategy for building up air quality indices is given by Bruno and Cocchi, 2002.

These evaluation criteria, together with the objects, form the background of an evaluation system. This is a very simple first evaluation approach. We use the scores good = 2, medium = 1 and bad = 0. If, for example, many chemicals are monitored, a higher score is given than for a monitoring system which only encompasses very few standard chemicals. The evaluation criteria and their corresponding scores are given in Table 2.

In Table 3 the scores are allocated to the monitoring systems in the 15 capitals. For the explanations of the abbreviations (Abb.) see Tables 1 and 2.

4. APPLIED DATA ANALYSIS METHODS

Different data analysis methods will be applied: the Hasse diagram technique, a method derived from discrete mathematics and the partially ordered scalogram analysis with coordinates (POSAC) method, a multivariate statistics approach. These methods are compared and advantages and disadvantages will be elaborated.

The aim of this data analysis is the evaluation of the publicly available air quality monitoring systems in Europe with their pros and cons. This helps the interested public to find and understand the information given on the Internet. Furthermore, our evaluation approach might provide some recommendations for an improvement of the air quality monitoring systems.

4.1. Hasse diagram technique (HDT)

4.1.1. Introduction. The Hasse diagram technique is well explained in different journals. A recent description can be found in Brüggemann *et al.*, (2001). Therefore some aspects are picked out, which

Table 2. Evaluation criteria and their scores

Number of chemicals monitored/NU	Score
≤ 4	0
5–7	1
>7	2
Types and duration of measurements/ME	
Only daily values, yesterday's values	0
daily + past week, past 2 weeks	1
daily + past weeks + past 6 months, past year(s)	2
Measurement stations in capital/ST	
1–2	0
3–5	1
>6	2
Method of presentation on the Internet/PR	
Table	0
Graphic	1
Table + graphic	2
Background material/BM	
Explanations, reports	0
+limit values, laws and regulations	1
direct communication, measures	2

Table 3. Data matrix 15 European monitoring stations evaluated by 5 criteria

Abb.	NU	ME	ST	PR	BM
AUS	1	2	1	2	1
BEL	1	2	1	2	2
DEN	2	2	0	2	2
FIN	1	1	2	1	1
FRA	1	2	2	2	2
GER	2	2	2	2	1
GRE	1	0	2	0	0
IRE	0	0	0	1	1
ITA	2	0	2	1	1
LUX	0	0	0	1	0
NET	1	1	0	2	1
POR	1	0	2	0	1
SPA	0	0	1	1	0
SWE	1	2	0	1	0
UNK	1	0	2	1	2

will be useful in the subsequent application. Hasse diagrams visualize the order relations within objects: two objects (also called elements, if the aspect of belonging to sets is important), x, y of an object set are considered as being ordered if all scores of x are less than or equal to those of y. Hasse diagrams are acyclic digraphs and objects are drawn as small circles together with an appropriate identifier. In our applications the circles near the top of the page (of the Hasse diagram) indicate objects that are the 'better' objects according to the criteria used to rank them: the objects not 'covered' by other objects are called *maximal objects*. Objects which do not cover other objects are called *minimal objects*. In some diagrams there also exist *isolated objects* which can be considered as maximal and minimal objects at the same time.

The WHasse program is developed, improved and updated by Rainer Brüggemann and is available for non-commercial use from the third author. For commercial applications it is recommended to contact the company Criterion—Evaluation and Information Management (Criterion, 2003).

4.1.2. Hasse diagrams as mathematical objects. The basis of the Hasse diagram technique (named HDT for short) is the assumption that a ranking can be performed while avoiding the use of an ordering index (Halfon and Reggiani, 1986). For an evaluation of the objects they must be compared. The comparison is done by examining characteristic properties (attributes, descriptors) of these objects. If the evaluation is aimed to assess criteria, then the attributes or (synonyms: descriptors) are thought of as measures of how well a criterion is fulfilled. Note that the concepts 'criterion' and an assigned measure 'attribute' should kept well separated. Attributes are, in the case of the object 'x', denoted as $q(1, x), q(2, x), \dots, q(m, x)$ and often written as a tuple $\mathbf{q}(x)$. We avoid the term vector, because the properties of a linear space are not needed in the HDT. Often the properties are gathered to a set without reference to actual values realized by the objects. This set of properties is called an *information base*. Consider now two objects x and y; then we say $y \geq x$ (with respect to the m properties of interest) if $q(i, x) \leq q(i, y)$ *for all* $i = 1, 2, \dots, m$ and there is at least one i^* for which $q(i^*, x) < q(i^*, y)$ (because of the demand 'for all', this definition is denoted as a 'generality principle'). If $q(i, x) \leq q(i, y)$ for all $i = 1, \dots, m$ then the objects x and y are comparable. The mere fact that x is comparable with y is often denoted as $x \perp y$.

Often, however, one finds:

$q(i, x) < q(i, y)$ for one index set I' and
$q(i, x) > q(i, y)$ for another index set I'' with $\mathrm{I}' \cap \mathrm{I}'' = \emptyset$.

In that case, the objects x and y are incomparable and one writes: $x \parallel y$. There are many other ways to define order relations. The order relation defined here is known as product order.
The main frame of HDT is therefore (the four-point-program (fpp)):

1. Selecting a set of elements of interest which are to be compared, E, called the ground set.
2. Selecting a set of properties, by which the comparison is performed, called the information base IB.
3. Finding a common orientation for all properties, according to the criteria they are assigned.
4. Analyzing $x, y \in E$ when one of the following relations is valid:

   - $x \sim y$ (equivalence; we call the corresponding equivalence relation R the equality of two tuples $\mathbf{q}(x)$, $\mathbf{q}(y)$)
   - $x \leq y$ or $x \geq y$ (comparability)
   - $x \parallel y$ (incomparability; there is a 'contradiction in the data of x and y').

The relation defined above among all objects is indeed an order relation, because it fulfils the axioms of order, namely:

- reflexivity (one can compare each object with itself)
- antisymmetry (if x is preferred to y then the reverse is only true if the two objects are equal (or (quasi order) equivalent)
- transitivity (if x is better than y, and y is better than z, then x is better than z).

A set E equipped with an order relation $\leq$ is said to be an ordered set (or partially ordered set) or briefly 'poset' and is denoted as $(E, \leq)$.

Note: A set E equipped with a partial order is often written as $(E, \leq)$. Because the $\leq$-comparison depends on the selection of the information base (and of the data representation (classified or not, rounded, etc.)) we also write (*E*, *IB*) to denote this important influence of the *IB* for any rankings (Brüggemann and Welzl, 2002).

Sometimes it is useful to refer to the quotient set, which is induced by the equivalence relation of equality, *R* (see Brüggemann and Bartel, 1999). Sometimes there is no path between parts of the non-directed graph; these parts, supplied with the inherited order relation, are called (isolated) hierarchies or—if only one object (one equivalence class) constitutes a part of the Hasse diagram—an *isolated element*. Different hierarchies indicate quite specific data structures.

If empirical posets are to be examined it is important to establish orientation rules, i.e. which value of attributes is considered to contribute to badness and which values to goodness. Concerning the evaluation of the ecotoxicity of environmental chemicals by lethal concentrations, i.e. LC_{50} values, for example, the orientation is the other way round. Here the following situation arises:

small values: 'good', relatively unhazardous
large values: 'bad', relatively hazardous.

4.1.3. Characterizing numbers of the Hasse diagram. In order to interpret a Hasse diagram some further terms have to be introduced:

A *chain* is a set of mutually comparable objects. An *antichain* is a set of mutually incomparable objects. An *articulation point* is a vertex of a graph whose elimination would increase the number of hierarchies.

Levels: A first screening and a partitioning of set *E* according to increasing values of the attributes. They are defined by the longest chain within the Hasse diagram (see below). They are not unique from the point of view of order theory, but uniquely defined, if additional rules are introduced (for example: conservativity, i.e. in HDT objects are assigned to the highest possible level). The set of levels together with the $\leq$ relation forms a new poset $(L, \leq)$. Both the empirical poset $(E, \leq)$ and $(L, \leq)$ are related by an order preserving map.

4.1.4. Comparability and incomparability. N: the elements of the set *E* (*E* is assumed to be a finite set); *NA*: the number of elements of the set *E* which are contained in non-trivial equivalence sets; *NECA*: the number of equivalence classes with more than one object, also called the number of non-trivial equivalence classes; *NL*: the number of levels; *NMAX*: the number of maximal elements (called number of maximal equivalent classes because this information is related to *E/R*); *NMIN*: the number of minimal elements (notation as for the maximals); *Z*: the number of all equivalence classes, including singletons (that are sets with only one element). Note that the following simple equation relates *Z*, *NECA* and *NA*:

$$N = NA + Z - NECA \tag{1}$$

DIM(E, IB): The dimension of the poset (*E, IB*). The concept of dimension of partial orders (see for details, Trotter, 1992) is very useful in combination with applications of POSAC, because one can decide how many latent attributes will suffice to get an isomorphic Hasse diagram. In general, it is difficult to calculate the dimension. If, however, the Hasse diagram can be considered as a planar lattice, then the dimension of the poset is 2.

The total number of *comparabilities V* and *incomparabilities U* and their local analogues (i.e. the number of comparabilities *V*(*x*) and incomparabilities *U*(*x*) of a certain element *x* are useful tools for the documentation of the Hasse diagram. (Brüggemann and Welzl, 2002).

4.1.5. W-matrix: dissimilarity-matrix. The W-matrix describes the influence of the attributes on the Hasse diagram. The entries of the W-matrix are a measure for the metric distance among posets, based on the same ground set of objects, but induced by different subsets of IB of $m-1$ attributes, i.e. subset generated by $IB-\{q_i\}$, $i=1,\ldots,m$. More details can be found in Brüggemann *et al.* (2001). For further reading we refer to background publications by Brüggemann and Halfon (2000) and Brüggemann and Welzl (2002).

4.2. Partially ordered scalogram analysis with coordinates (POSAC)

The importance of multivariate statistics in chemometrics and environmetrics is emphasized by just the existence of this journal! Many textbooks covering this field of multivariate statistics are available; see, for example, Stoyan *et al.* (1997), Einax *et al.* (1997) and Millard and Neerchal (2001). An overview article on the chemometrical and environmetrical research of the research group Biostatistics at the GSF—National Research Center for Environment and Health, Institute of Biomathematics and Biometry, is given by Welzl *et al.* (2003).

There are many statistical approaches of condensing a data matrix by the creation of new variables. This process, called ordination, is often used to visualize relationships in two dimensions. These new variables, which are derived from the original variables, are constructed to optimise some specific criteria. For example, principal component analysis (PCA) creates new axes to explain as much as possible of the variance of the data matrix. This idea can be applied when order relations (comparability as well as incomparability) are considered as the essential aspect of the data to be preserved in the analysis. This method—construction of new axes which presents correctly as many as possible of the order relations—is called partially ordered scalogram analysis with coordinates (POSAC). POSAC is integrated in the program package SYSTAT 10 (SPSS Science, 2001) under the feature of statistics, data reduction. In POSAC, order relations (comparability as well as incomparability) are considered as the essential empirical-substantive aspect of the data to be preserved in the data analysis (Borg and Shye, 1995). For a better interpretation of the new axes, correlations between old and new variables can be calculated (Borg and Shye, 1995).

The background of the POSAC method as well as the mathematics in it is described in Shye (1985). Beyond this, the starting point is the same as in the HDT; therefore we give only a very brief introduction. There is a set of N objects. Once again, each object x is characterized by an m-tuple of attribute values, $\mathbf{q}(x)$. The tuple $\mathbf{q}(x)$ is also called a profile. Under these conditions, the set A' of all observable profiles forms a partially ordered set which is the same as in HDT. That is, some pairs of profiles may be ordered comparable, while some pairs of profiles may be unordered, or incomparable.

Now an order preserving mapping can be constructed which preserves all order relations of the empirical poset and which contains a lower dimensional tuple of arbitrary scores. As the empirical poset contains incomparabilities, at least 2-tuples (2-score profiles) of scores must be found. By a dimension analysis of the empirical poset it is possible to find out the minimum number of needed scores.

The set of all 2-score profiles may be thought of as a two-dimensional Cartesian coordinate space with the one coordinate, X, indicating the first score and the other, Y, indicating the second score. Conversely, the two coordinates of each point in the XY space can be regarded as a two-score profile, so that all points in the plane form a partially ordered set. The essential thing to notice is that for every given point in the coordinate space, three different regions in the space are determined:

(I) a region of points that are less than the given point
(II) a region of points that are greater than the given point
(III) a region (made up of two disconnected parts) of points that are incomparable to the given point.

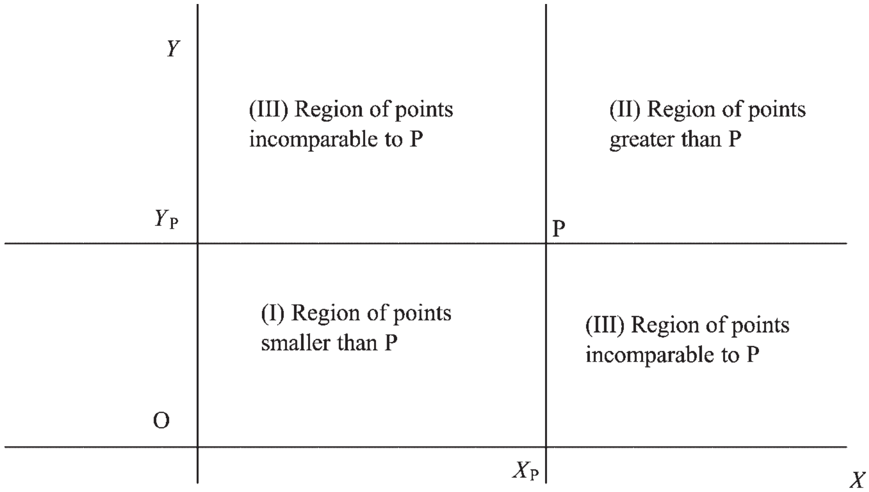


Figure 1. Regions in the coordinate space of POSAC

This is illustrated in Figure 1.

Basically, POSAC is concerned with the following question:

Given a set A' of observed profiles in N tests (equal to Z if no equivalences are found (equation 1)), we can assign two scores (that is, a point in the coordinate plane) to each profile in A', so that for any two observed profiles their observed relation would be represented correctly by their corresponding two-coordinate profile?

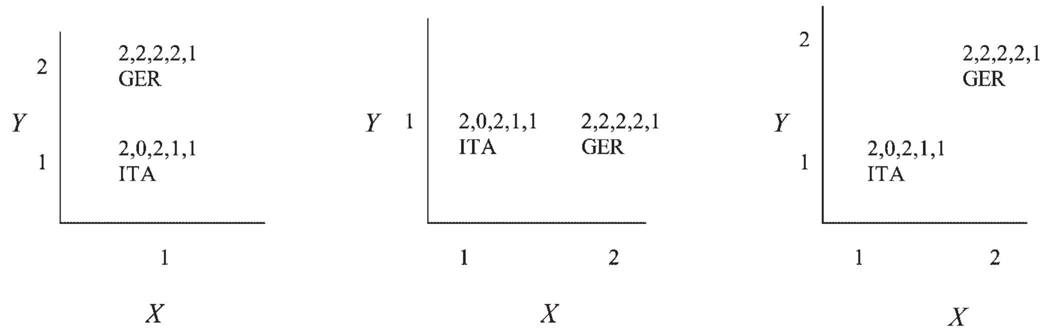
Let us see what would the required representation look like in the case of the three profiles just considered: GER {2,2,2,2,1}, ITA{2,0,2,1,1} and DEN {2,2,0,2,2}.

We draw a Cartesian coordinate axis X_1 and X_2 and place one profile, say ITA {2,0,2,1,1}, anywhere, say at the point (1,1).

To represent that GER {2,2,2,2,1} is greater than ITA {2,0,2,1,1}, we assign to GER {2,2,2,2,1} a point anywhere in the square whose left lower corner is (1,1). There are many possibilities. Three are shown in Figure 2.

Now, let us add profile DEN {2,2,0,2,2} to the picture. It is incomparable to both GER {2,2,2,2,1} and ITA {2,0,2,1,1}. So it must be represented within the region of points that are incomparable to both 11 and 22 (that is, within the intersection of regions of points incomparable to 11 and of points incomparable to 22). Taking, for instance, the third depiction above, any point in the shaded area in Figure 3 would do.

For example, one could place DEN {2,2,0,2,2} at (0,3). Additional five-score profiles can be plotted as points in the two-dimensional coordinate space, so that their order relations with profiles already plotted would be represented by their coordinates. But can any new profile be plotted in this described way in the plane? In general, the answer, is, of course, no. After plotting profiles from some list of observed profiles, it may well happen that a profile is encountered which cannot be located anywhere in the two-dimensional coordinate space without misrepresenting some of its relations with the profiles already plotted. Also, in general, there may not be a way of perfectly representing in the plane all order relations among existing profiles of a given set.



For these examples possible score assignments are:

2,0,2,1,1 ⇒ 11	2,0,2,1,1 ⇒ 11	2,0,2,1,1 ⇒ 11
2,2,2,2,1 ⇒ 12	2,2,2,2,1 ⇒ 21	2,2,2,2,1 ⇒ 22

Figure 2. Two profiles in a Cartesian coordinate space

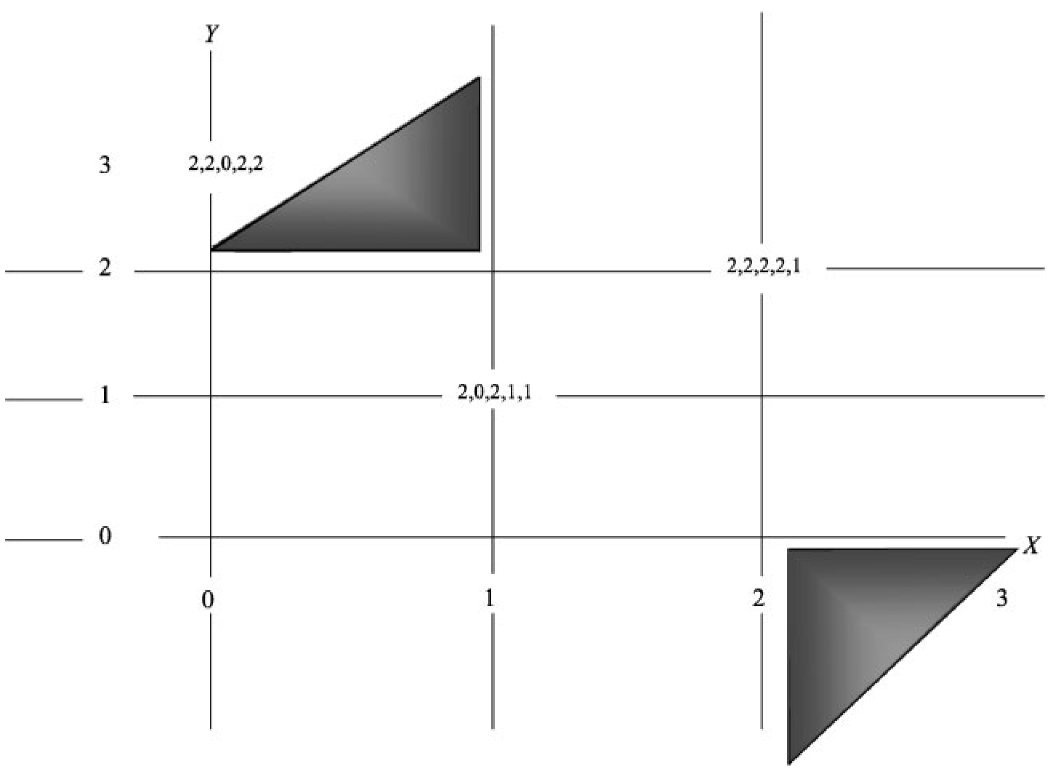


Figure 3. Three profiles in a two-dimensional coordinate space

Here the connection to HDT is obvious. In HDT the dimension of the empirical poset is of importance. If this dimension is >2, then any reduction to a graphical display in a plane will only be an approximation.

The practical POSAC problem is this: Given a set A' of observed profiles, what mapping of these profiles into the two-dimensional coordinate spaces would best preserve all their relations, 'greater than' ('less than') and 'incomparable to'? In order to deal with this question, a criterion must be specified by which one could determine whether a proposed mapping is better than another. A natural criterion is simply the proportion of profile pairs, out of all profile pairs, whose empirically observed order relation is correctly represented by their two twofold coordinates. The POSAC program provides, as its main output, a solution to the order-preserving mapping problem. POSAC provides also a coefficient based on the proportion of profiles correctly represented. The possible multiplicity of subjects with a given profile is taken, of course, into account. However, the program algorithm itself is based on the minimization of a different, computationally more manageable function proposed by Louis Guttman. This function as well as the entire POSAC/LSA algorithm is presented in detail by Shye (1980) and Shye and Amar (1985). When the output profile set is indeed two-dimensional or close to it, the POSAC solution is an approximation to the space of reduced dimension (Shye, 1985).

The POSAC method has already been applied on data matrices in environmental sciences and chemistry. Welzl examined regions polluted with metals (Welzl *et al.*, 1998). Pesticide Internet resources were analyzed with chemical and environmental evaluation criteria by Voigt *et al.* (2000), environmental and chemical search engines were ranked by Glander-Höbel *et al.* (2001) and Voigt and Welzl (2001), and drinking water analysis systems were evaluated by Voigt and Welzl (2002).

5. DATA ANALYSIS OF 15 × 5 MATRIX

The data analysis is done in two steps. First the Hasse diagram technique is applied. In a further step the attribute reduction is performed by applying the POSAC method, and the Hasse diagram based on the reduced attribute set is calculated. The procedure of the data analysis is outlined in Figure 4.

5.1. Hasse diagram technique applied on 15 × 5 data matrix

5.1.1. Interpretation of Hasse diagram. The Hasse diagram technique described in Section 4.1 is now applied on the given 15 × 5 data matrix.

The general evaluation of this diagram is as follows:

The objects (air monitoring systems) DEN (Denmark), FRA (France) and GER (Germany) are maximal objects. They are 'better' than those objects with which they are connected in the downward direction.

The minimal objects in this evaluation approach are LUX (Luxembourg) and GRE (Greece). The differences of these two minimal objects can easily be seen in Table 3 and Figure 5: LUX has only the score 1 in attribute 4 (PR), which means the method of presentation on the Internet. GRE (Greece) has score 1 in attribute 1 (NU) and score 2 in attribute 3 (ST). Neither equivalent objects nor isolated objects are given in this diagram.

The number of comparabilities is $V = 57$ and the number of incomparabilities $U = 48$. Both numbers express that the criteria are neither mutually rank-correlated nor mutually rank-anticorrelated. The level of degeneracy $K = 0$ simply expresses that there are no equivalences.

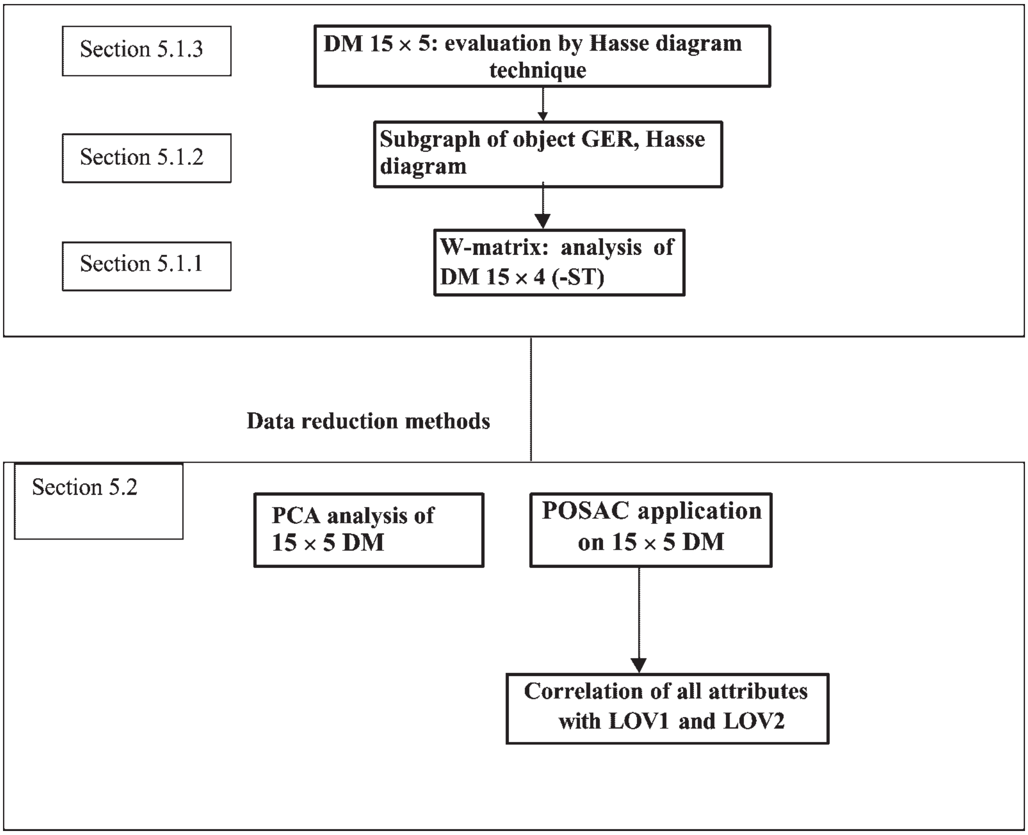


Figure 4. Steps of the data analysis of air pollution monitoring systems in Europe

The diagram is very structured: More comparabilities than incomparabilities are given. Furthermore we can see six levels illustrated in Figure 6, and the distribution of all objects over these levels can also be found in Figure 6.

This kind of representation of a partial ordered set is useful as the order of levels can be seen as the result of an order preserving map (Brüggemann and Welzl, 2002).

5.1.2. Subgraphs. Taking a closer look at the object GER (Germany), there are 10 objects y for which we can write $y \leq GER$, called successors. The corresponding subset inherits the order relation of the ground set and is called an order ideal generated by GER. As in general fewer objects are elements of the order ideal, we can change the kind of representation and show explicitly the profiles of the attributes. This form of presentation is called a super-graphic (Figure 7).

In Figure 7 it can be clearly demonstrated why the objects are positioned lower than GER (Germany). Germany has higher scores in attributes 1 and 3 than Austria (AUS) and the same scores in attributes 2 and 4 (see also Table 3). The object GER has scores at least as high as those of 10 other objects. These successors are: AUS, ITA, FIN, SWE, NET, SPA, POR, IRE, LUX, GRE. The incomparability of objects, e.g. Austria (AUS) and Italy (ITA) is obvious; Austria has a lower score for attributes 1 and 3 than Italy. On the other hand, Italy shows a lower score in attributes 2 and 4 than Austria.

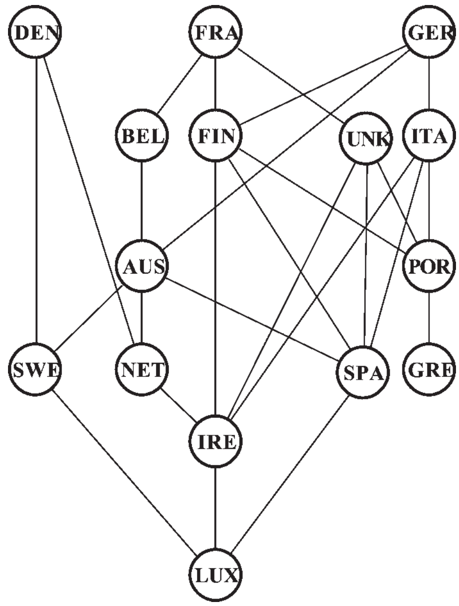


Figure 5. Hasse diagram for the 15 × 5 data matrix

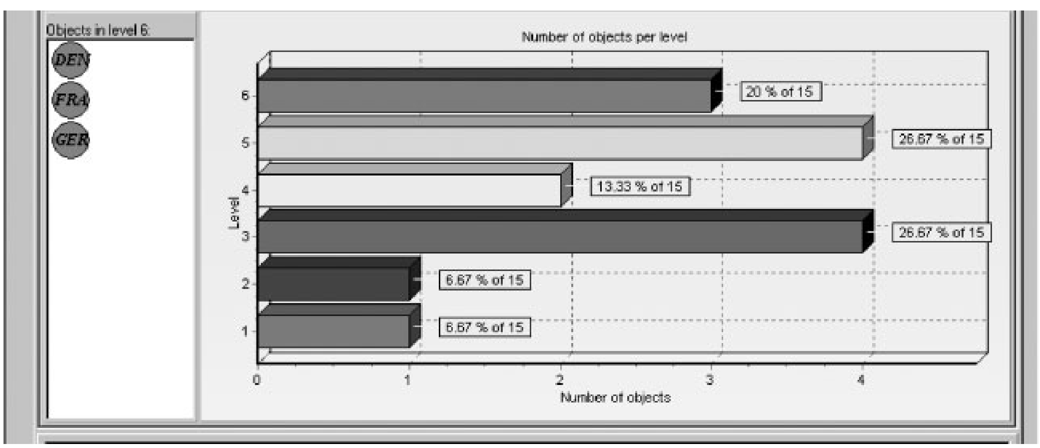


Figure 6. Level structure of the Hasse diagram in Figure 5

5.1.3. W-matrix: dissimilarity matrix. The W-matrix describes the influence of the attributes NU, ME, ST, PR and BM on the Hasse diagram. As we have five attributes, $m+1=6$ cases with either the full attribute set or attribute sets, where one attribute is left out, are examined and listed in Table 4.

This W-matrix analysis shows that, leaving out the attribute ST, most changes in the diagram take place. This means that the attribute ST is the most important one in this analysis. We now calculate the Hasse diagram for 15 objects with only four attributes (NU, ME, PR, BM), leaving out the attribute ST. This procedure is documented in Figure 8 and the results presented in Table 5, second and third columns.

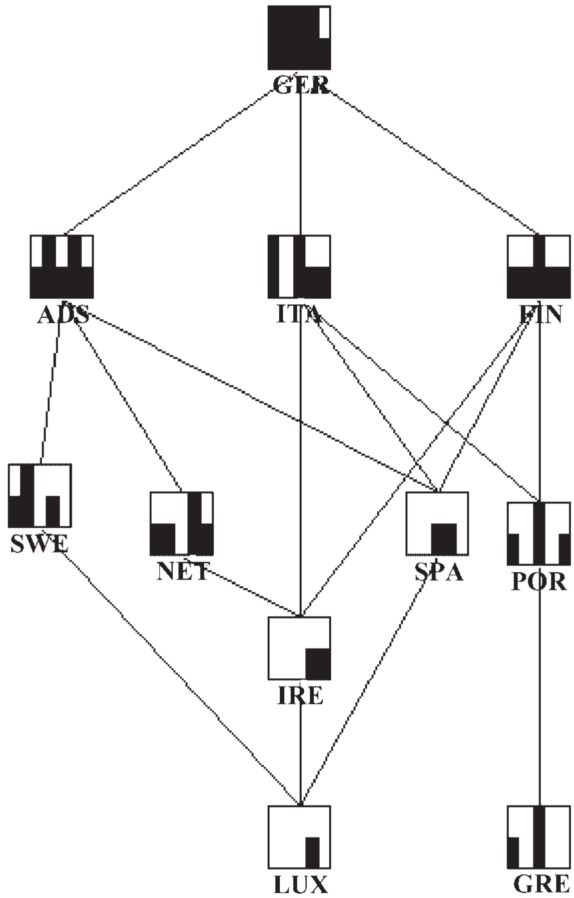
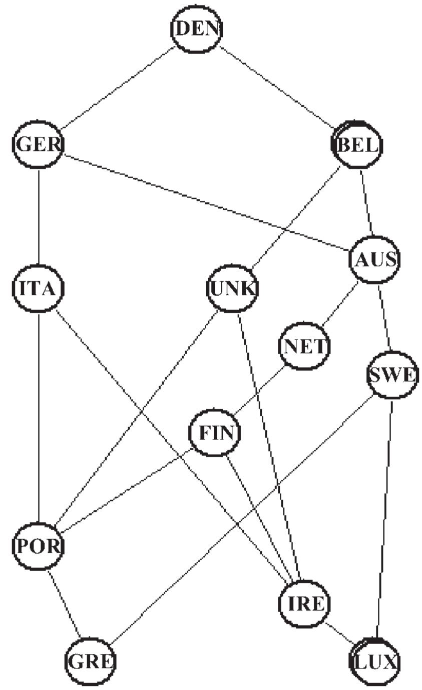


Figure 7. Subgraph of the object GER in the super-graphic

Comparing the Hasse diagram of the complete data set of Figure 5 with the reduced data matrix, 15×4 considerable differences can be detected. These are not only in the maximal and minimal objects but also with respect to the levels, comparabilities, incomparabilities, etc. In this diagram two equivalent objects, {BEL;FRA} {LUX;SPA}, are found. This means, for example, that the minimal object LUX is equivalent to the object SPA.

Table 4. Evaluation of the W-matrix for all elements

Case	Leaving out	Number of changes in HD
1	NU = number of chemicals monitored	6
2	ME = types and duration of measurements	6
3	ST = measurement stations in capital	26
4	PR = method of presentation on the Internet	6
5	BM = background material	11



Equivalent objects: {BEL;FRA} {LUX;SPA}

Figure 8. Hasse diagram of Case 3 (leaving out the attribute ME = measurement stations in capital)

5.2. *Data reduction methods applied on the 15 × 5 data matrix*

5.2.1. Principle component analysis (PCA). PCA is a procedure for analyzing multivariate data which transforms the original variables into new ones that are uncorrelated and account for decreasing proportions of the variance in the data. The aim of the method is to reduce the dimensionality of the data. The new variables, the principal components, are defined as linear functions of the original variables. If the first few principal components account for a large percentage of the variance of the observations (say >70%) they can be used both to simplify subsequent analyses and to display and summarize the data in a parsimonious manner (Everitt, 1998).

Table 5. Comparison of results in Hasse diagrams with different attribute sets

Hasse information/ data matrix	15 × 5	15 × 4 (Case 3)	15 × 2 (POSAC)
Number of levels	6	7	6
Elements in largest level	4	3	5
Comparabilities	57	83	60
Incomparabilities	96:2 = 48	48:2 = 24	90:2 = 45
Maximal objects	{DEN}, {FRA}, {GER}	{DEN}	{DEN}, {FRA}, {GER},
Minimal objects	{GRE}, {LUX}	{GRE}, {LUX}, {SPA}	{GRE}, {LUX}, {POR}
Equivalent objects	—	{BEL; FRA} {LUX; SPA}	—

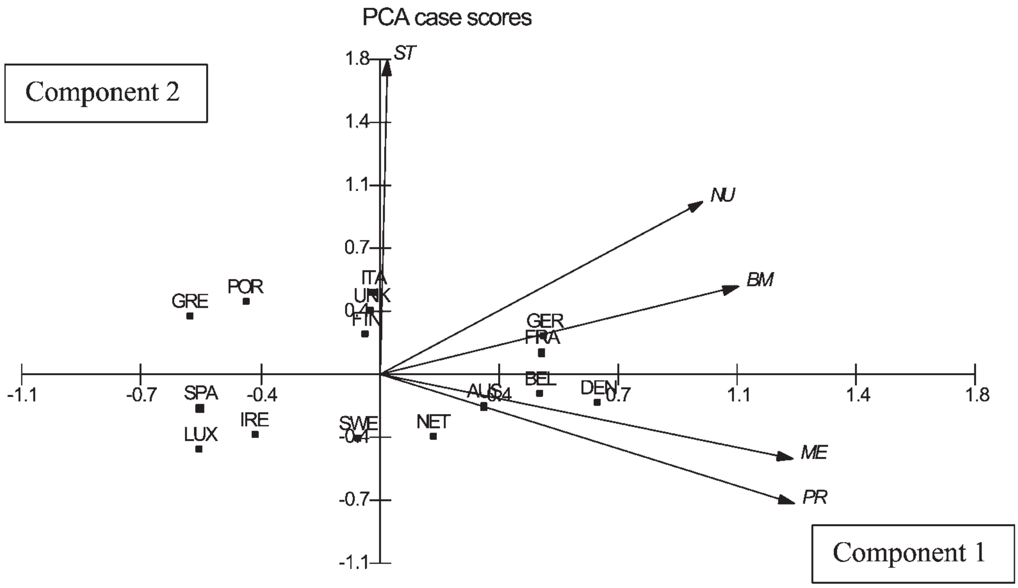


Figure 9. PCA biplot of 15×5 data matrix

PCA is applied on the 15×5 data matrix. The PCA Biplot is given in Figure 9. The software package used is MVSP Statistics Package (Kovach Computing Services, 2002).

The variables *PR*, *ME*, *BM*, *NU* have high loadings on component 1, whereas component 2 is mostly influenced by the variable *ST*. The declared variance with 2 components is 77.5%.

5.2.2. POSAC method applied on 15 × 5 data matrix. Now we perform the POSAC (partially ordered scalogram analysis with coordinates) analysis on the given data matrix. This method reduces the data matrix in finding a two-dimensional space preserving most of the original comparabilities. A given percentage of information is lost by this method.

The proportion of the order relations which are correctly represented is 93.4%. This means that the Hasse diagram can be approximated by using the two latent order variables (LOVs) instead of the five initial variables. The objects with the maximal tuple (22222) as well as with the minimal tuple (00000) is automatically added by the POSAC program (see Figure 10). The object FRA and GER can be found at the upper right side of the POSAC plot, which means that they have high values in the latent order variable 1 as well as in the latent order variable 2. The object DEN (also a maximal object in the original Hasse diagram of the data matrix 15×5 (see Figure 5) shows a high value in LOV2 but only a medium value in LOV1. The object GRE (minimal object in the Hasse approach) gives a high value in LOV1 and a low value in LOV2. Contrary to GRE the country LUX has a medium value in LOV2 and a small value in LOV1.

5.2.3. Interpretation of the latent order variables (LOVs). In order to explore the influence of the attributes on the whole analysis, we perform a correlation analysis of the two latent order variables given in the POSAC plot. This is done by applying the analyses of variance (ANOVA) with LOV1 and LOV2 as dependent variables and the original five attributes as factors.

The following F-statistics are calculated:

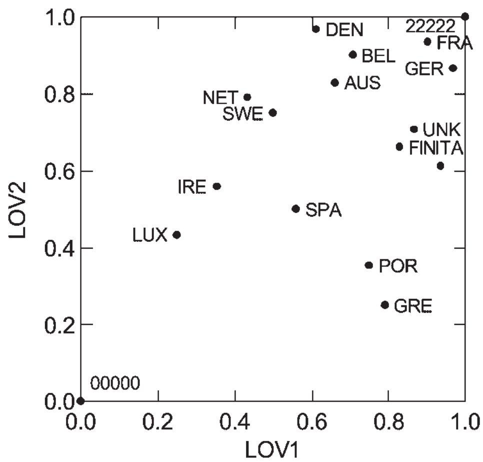
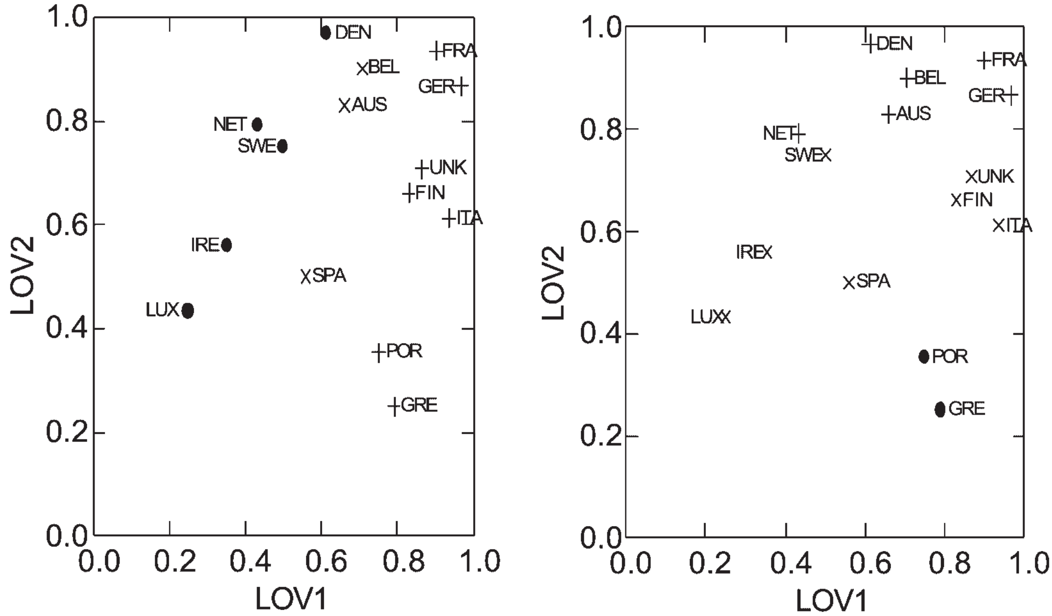


Figure 10. POSAC plot of data matrix 15×5



●. = 0, x=1, + =2

Figure 11. Scatter plots of the variables ST and PR

LOV1: ST: 26.556, NU: 6.169, BM: 1.475, PR: 0.520, ME: 0.237
LOV2: PR: 32.723, ME: 15.861, BM: 5.150, NU: 1.338, ST: 0.315

The latent order variable 1 is mainly described by ST, whereas the latent order variable 2 is mainly influenced by PR. The scatter plots of the variables ST and PR are given in Figure 11.

A comparison with the results of the PCA shows that the interpretation of Component 1 is similar to the interpretation of LOV2 in POSAC. This will say that PR, ME, BM and NU show a special

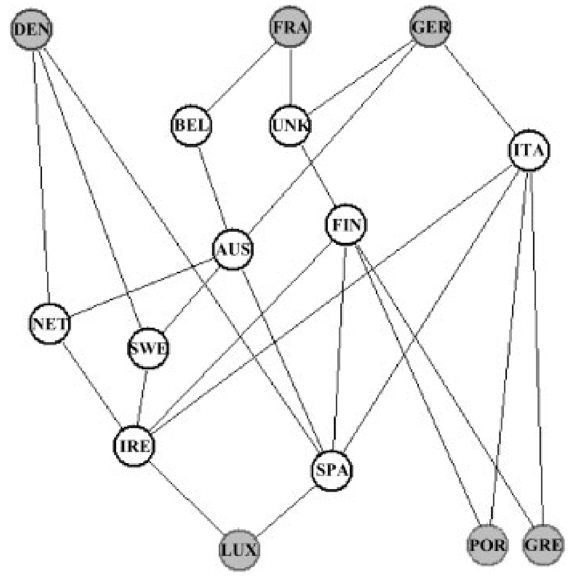


Figure 12. Hasse diagram of reduced data matrix 15×2 (LOVs from POSAC)

importance. On the other hand, Component 2 in PCA and LOV1 in POSAC demonstrate the relevance of the variable ST.

The high values (+2) are found on the upper right side of the scatter plot, that is to say high values for LOV1 and for LOV2.

In the next step of the evaluation procedure for this data matrix we calculate a Hasse diagram based on the two latent order variables found by the POSAC method. The result of this analysis is shown in Figure 12.

It is obvious that the Hasse diagram (Figure 12) is only a different way of presenting the results of the POSAC analysis given in Figure 10. It supports the ordering of the Hasse diagram. In this Hasse diagram, due to the rearrangements of the objects performed on the POSAC results, the following structure can be detected. The left side of the diagram in Figure 12 shows objects which have the score 0 or 1 for the criterion ST, whereas the objects situated at the right hand side of the diagram have the score 2 for ST.

In Table 5 the differences of the original Hasse diagram 15×5 and the 15×2 diagram (as well as the already discussed 15×4 data matrix) are listed. Comparabilities have slightly increased and incomparabilities have slightly decreased. The maximal objects are DEN, FRA, GER in the original data set as well as in the 15×2 data matrix, and the minimal objects are GRE and LUX in the 15×5 data-matrix and additionally POR in the 15×2 data matrix. This data reduction on the attribute side only shows minimal differences. The dimension reduction to two dimensions does not lead to a more comprehensive Hasse diagram than the initial Hasse diagram of the 15×5 data matrix given in Figure 5 but the rearranging of objects and the construction of LOVs enables an easier interpretation. Dimension reductions in leading to more than two LOVs should be envisaged. The POSAC program in SYSTAT does not provide the facility for processing data into higher dimensions. In the HUDAP package POSAC was originally also technically limited to two dimensions, until recently when a new algorithm allowed for processing POSAC in any dimensionality. The related computer module is called MPOSAC, namely multidimensional POSAC (Hebrew University of Jerusalem, 2002).

6. DISCUSSION AND OUTLOOK

The aim of this data analysis is the evaluation of the publicly available air quality monitoring systems in Europe with their pros and cons. This might help the interested public to find and understand the information given on the Internet. Furthermore, our evaluation approach might give some recommendations for an improvement of the air quality monitoring systems.

In our data analysis approach of the air monitoring information systems in Europe we found out in the initial analysis applying the Hasse Diagram Technique that the systems of Germany (GER), France (FRA) and Denmark (DEN) showed excellent results that means better results than many other information systems. In this initial approach all five chosen attributes (evaluation criteria) were considered. The air quality monitoring systems of Luxemburg (LUX) and Greece (GRE) gave rather weak results. In the latter case, only marginal information could be found on the Internet. Further analyses showed that some evaluation criteria had a high impact on the Hasse diagram. This is the attribute ST, the criterion which describes the spatial aspect of the analysis. The data analysis by the POSAC method reduces the data matrix into a two-dimensional plot. The new variables are called latent order variables. It can be demonstrated that the objects France (FRA), Denmark (DEN), Germany (GER) and Belgium (BEL) had high values in latent order variable 1 as well as in latent order variable 2.

Concerning the data analysis aspects the following can be stated. The original Hasse diagram (see Figure 5) shows more comparabilities than incomparabilities. The considerable high value of comparabilities of the Hasse diagram technique might be the reason why the POSAC analysis is not giving considerably different results. POSAC substantially supports those results given by the Hasse diagram technique. By this POSAC method the initial 15×5 data matrix can be reduced to 15×2 (latent order variables). The Hasse diagram for this reduced data matrix shows only very few changes to the original Hasse diagram; however, the rearranging of objects and the construction of LOVs enables an easier interpretation. On the basis of the two latent order variables a correlation analysis is performed. The variables measurement stations in the capitals and the method of presentation on the Internet have an important meaning in the data analysis. The spatial criterion ST also attracted attention in the W-matrix analysis of the Hasse approach as well as in the PCA biplot. The PCA analysis and the POSAC method show similar results. PR, ME, BM and NU hold a special position in the interpretation of Component 1 in PCA as well as in the results of LOV2 in POSAC. On the other hand, Component 2 in PCA and LOV1 in POSAC demonstrate the relevance of the variable ST.

The analysis and the searches on the Internet were performed in the months of March and April 2002. The authors are aware of the fact that the Internet offers are constantly changing and improving and that it is quite possible that an air monitoring information system was not found or was unavailable during the time of the search. We would like to encourage all those experts who have further knowledge on the systems to inform us.

REFERENCES

Bollobas B. 1986. *Combinatorics*. Cambridge University Press: Cambridge.

Borg I, Shye S. 1995. *Facet Theory, Form and Content*. Sage Publications: Thousand Oaks; 111–112.

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

Brüggemann R, Halfon E. 2000. Introduction to the general principles of the partial order ranking theory. In *Order Theoretical Tools in Environmental Sciences, Proceedings of the Second Workshop October 21st, 1999 Held in Roskilde, Denmark,* *NERI—Technical Report No.* 318, Soerensen PB, Carlsen L, Mogensen BB, Brüggemann R, Luther B, Pudenz S, Simon U, Halfon E, Voigt K, Welzl G, Rediske G (eds). National Environmental Research Institute: Roskilde, Denmark; 7–44.

Brüggemann R, Welzl G. 2002. Order theory meets statistics—Hasse diagram technique. In *Order Theoretical Tools in Environmental Sciences, Order Theory (Hasse Diagram Technique) Meets Multivariate Statistics*, Voigt K, Welzl G (eds). Shaker-Verlag: Aachen; 9–40.

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

Bruno F, Cocchi D. 2002. A unified strategy for building up simple air quality indices. *Environmetrics* **13**: 243–261.

Criterion. 2003. WHasse Software, http://www.criteri-on.de/hdtws-site/index.html.

Einax JW, Zwanziger HW, Geiß S. 1997. *Chemometrics in Environmental Analysis*. VCH-Verlag: Weinheim; 1–16.

Eu-Lex. 2002. *Proposal for a Directive of the European Parliament and of the Council on public access to environmental information COM(2000)* 402. http://europa.eu.int/comm/environment/docum/00402_en.htm.

Everitt BS. 1998. *The Cambridge Dictionary of Statistics*. Cambridge University Press: Cambridge.

Glander-Höbel C, Voigt K, Welzl G. 2001. Comparing search engines with respect to environmental chemistry using statistical and mathematical evaluation methods. In *Online Information 2001, Proceedings, 4–6 December 2001 London*, Graham C (ed.). Learned Information Europe Ltd: Oxford; 187–195.

Halfon E, Reggiani MG. 1986. On ranking chemicals for environmental hazard. *Environ. Sci. Technol.* **20**(11): 1173–1179.

Hebrew University of Jerusalem, Computation Centre. 2002. *The HUDAP Package*. http://help.mscc.huji.ac.il/hudap.html.

Kovach Computing Services. 2002. *MVSP Statistical Package*, Anglesey, Wales. http://www.kovcomp.co.uk/mvsp/.

Millard SP, Neerchal NK. 2001. *Environmental Statistics with S-Plus*. CRC Press: Boca Raton.

Roger RD. 2003. *TPCN great quotations*. http://www.cyber-nation.com/victory/quotations/authors/quotes_roger_rutherfordd.html.

Shye S. 1980. *Partial Ordered Scalogram and Dimensional Analysis of Items by their Scalogram Generic Roles*. The Israel Institute of Applied Social Research: Jerusalem.

Shye S. 1985. *Multiple Scaling*. Elsevier Publishers: Amsterdam.

Shye S, Amar R. 1985. Partial order scalogram analysis by base coordinates and lattice mapping of the items by their scalogram roles. In *Facet Theory*, Canter D (ed.). Springer-Verlag: New York; 63–71.

SPSS Science 2001. *SYSTAT 10*, http://www.spssscience.com/SYSTAT/index.html.

Stoyan D, Stoyan H, Jansen U. 1997. *Umweltstatistik, Statistische Verarbeitung und Analyse von Umweltdaten*, Teubner BG (ed.). Verlagsgesellschaft: Stuttgart, Leipzig.

Streuff H. 2000. Umweltinformationen im Internet—Chance oder Chimäre—Können Hypermediatechniken einen Beitrag zur ökologischen Modernisierung leisten? In *Hypermedia im Umweltschutz, 3. Workshop, Ulm 2000*, Tochtermann K, Riekert W-F (eds). Metropolis-Verlag: Marburg; 15–22.

Trotter WT. 1992. *Combinatorics and Partially Ordered Sets Dimension Theory*. The Johns Hopkins University Press: Baltimore, Maryland.

Voigt K, Welzl G. 2001. Evaluation of search engines concerning environmental terms. In *Sustainability in the Information Society, 15th International Symposium Informatics for Environmental Protection, Zürich 2001*, Hilty LM, Gilgen PW (eds). Metropolis-Verlag: Marburg; 683–690.

Voigt K, Welzl G. 2002. Drinking water analysis systems in German cities: an evaluation approach combining Hasse diagram technique with multivariate statistical methods. In *Order Theoretical Tools in Environmental Sciences, Order Theory (Hasse Diagram Technique) Meets Multivariate Statistics*, Voigt K, Welzl G (eds). Shaker-Verlag: Aachen; 113–128.

Voigt K, Welzl G, Rediske G. 2000. Environmetrical approaches to evaluate Internet databases. In *Order Theoretical Tools in Environmental Sciences, Proceedings of the Second Workshop October 21st, 1999 held in Roskilde, Denmark*, NERI—Technical Report No. 318, Sorensen PB, Carlsen L, Mogensen BB, Brüggemann R, Luther B, Pudenz S, Simon U, Halfon E, Voigt K, Welzl G, Rediske G (eds). National Environmental Research Institute: Roskilde, Denmark; 135–144.

Voigt K, Welzl G, Brüggemann R, Pudenz S. 2002. Bewertungsansatz von Umweltmonitoring-Systemen in Ballungsräumen. *UWSF—Umweltwissenschaften und Schadstoff Forschung* **14**(1): 58–64.

Welzl G, Voigt K, Rediske G. 1998. Visualisation of environmental pollution—Hasse diagram technique and explorative statistical methods. In *Proceedings of the Workshop on Order Theoretical Tools in Environmental Sciences held on November 16th, 1998 in Berlin, Berichte des IGB 1998, Heft 6, Sonderheft I*, Institut für Gewässerökologie und Binnenfischerei (ed.). IGB: Berlin; 101–110.

Welzl G, Faus-Kessler T, Scherb H, Voigt K. 2003. In print. Biostatistics. In *EOLSS Encyclopedia of Life Support Systems*, Sydow A (ed.). EOLSS Publishers Co. Ltd.: Oxford.

