Ecotoxicology and Environmental Safety 105 (2014) 29–35

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

CombiSimilarity, an innovative method to compare environmental and health data sets with different attribute sizes example: Eighteen Organochlorine Pesticides in soil and human breast milk samples

Rainer Bruggemann [a,*], Hagen Scherb [b], Karl-Werner Schramm [c,d], Ismet Cok [e], Kristina Voigt [b]

[a] Leibniz-Institute of Fresh Water Ecology and Inland Fisheries, Berlin, Germany
[b] Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Institute of Computational Biology, Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany
[c] Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Molecular EXposomics (MEX), Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany
[d] TUM, Wissenschaftszentrum Weihenstephan fuer Ernaehrung und Landnutzung, Department fuer Biowissenschaften, Weihenstephaner Steig 23, 85350 Freising, Germany
[e] Department of Toxicology, Faculty of Pharmacy, Gazi University, 06330 Ankara, Turkey

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form
25 March 2014
Accepted 27 March 2014

Keywords:
Partial order
Ranking
PyHasse software
Environmental health
Organochlorine Pesticides (OCPs)
Turkey

ABSTRACT

Human health and the health of the environment have entwined. In this paper we underpin this position by presenting a modeling approach named CombiSimilarity, which has been developed by the first author in the software tool PyHasse comprising a wide variety of partial ordering tools. A case study of 18 Organochlorine Pesticides (OCPs) detected in soil as well as in human breast milk samples in the Taurus Mountains in Turkey is carried out. Seven soil samples and 44 breast milk samples were measured. We seek to answer the question whether the contamination pattern in breast milk is associated with the contamination pattern in soil by studying the mutual quantitative relationships of the chemicals involved. We could demonstrate that there is a similarity with respect to the concentration profiles between the soil and breast milk pollution. Therefore the hypothesis may be formulated that the concentrations of chemicals in the milk samples are strongly related to the soil contamination. This supports the concept that soil could be a surrogate for human exposure at background locations.

1. Introduction

We find ourselves in an uneasy position due to the impact of the great industrial and technologic expansions. The mushrooming chemical industry, the wide application of nuclear energy, the revolutionary changes in food processing, and many other technological developments are affecting the environment and consequently the health of man, in multitudinous ways. The numerous benefits associated with these developments have been detracted from the recognition that the health factor may well be the limiting factor in the continuing development. It is amazing that these statements have already be recognized and published around 60 years ago (Ludwig, 1955). An approach to the identification of organic compounds hazardous to the environment and human health has already been initiated by the US National Science Foundation in the seventies (Stephenson, 1977). However, in the following decades environmental topics and health topics have not been regarded sufficiently in parallel. Only recently, these two interacting areas are acknowledged to be considered in closer conjunction. This implies that monitoring data are available not only on the environmental contamination side but also on the human tissue side.

The modeling aspects have been acknowledged by several approaches, e.g. an environmental health impact assessment approach (Knol et al., 2010), a hazard ranking model for environmental and human health hazard classifications for 55 plastic polymers (Lithner et al., 2011), human health risk assessment of chemicals at electronic waste sites in China (Chan and Wong, 2013).

The global pesticide use comes at the cost of its widespread occurrence in the environment and eventually in many cases in the human body. A special emphasis should be laid on the pesticides with endocrine effects. Endocrine disruptors are exogenous agents that interfere with the production, release, transport, metabolism, binding, action or elimination of the natural hormones from the body and are responsible for the maintenance of homeostasis and the regulation of developmental processes. Hence they can act like uncontrolled medicine (Birnbaum, 2013). In a recently performed

* Corresponding author.
E-mail address: brg_home@web.de (R. Bruggemann).

http://dx.doi.org/10.1016/j.ecoenv.2014.03.031
0147-6513/

Table 1
Eighteen OCPs detected in human and environmental samples in the Taurus Mountains, Turkey.

Nr.	Acronym	Standard abbreviation	Name	CAS-number
01	**AHCH**	alpha-HCH	alpha-Hexachlorcyclohexane	319-84-6
02	**BHCH**	beta-HCH	beta-Hexachlorcyclohexane	319-85-7
03	**GHCH**	gamma-HCH	gamma-Hexachlorcyclohexane	58-89-9
04	**PECB**	PCB	Pentachlorobenzene	608-93-5
05	**HCBE**	HCB	Hexachlorobenzene	118-74-1
06	**PPDT**	p, p′-DDT	p, p′-Dichlordiphenyltrichlorethane	50-29-3
07	**OPDT**	o, p′-DDT	o, p′-Dichlordiphenyltrichlorethane	789-02-6
08	**PPDD**	p, p′-DDD	p, p′-Dichlordiphenyldichlorethane	72-54-8
09	**OPDD**	o, p′-DDE	o, p′-Dichlordiphenyldichlorethane	53-19-0
10	**PPDE**	p, p′-DDD	p, p′-Dichlordiphenyldichlorethene	72-55-9
11	**OPDE**	o, p′-DDD	o, p′-Dichlordiphenyldichlorethene	3424-82-6
12	**OXYC**	Oxychlordane	Oxychlordane	27304-13-8
13	**CHCE**	cis-Heptachloroepoxide	cis-Heptachloroepoxide	1024-57-3
14	**DIEL**	Dieldrin	Dieldrin	60-57-1
15	**END1**	Endosulfan	Endosulfan-1	959-98-8
16	**END2**	Endosulfan	Endosulfan-2	33213-65-9
17	**MECH**	Methoxychlor	Methoxychlor	72-43-5
18	**MIRE**	Mirex	Mirex	2385-85-5

study the current knowledge of the potential endocrine impacts of 105 pesticides on human health is given (Mnif et al., 2011). Out of our test set of eighteen pesticides (see Table 1) seventeen have endocrine effects according to Mnif's study.

The occurrence of environmental chemicals in the Taurus Mountains in Turkey motivated an international study on POPs (Persistent Organic Pollutants) in environmental and human media. Eighteen OCPs (Organochlorine Pesticides) in samples of soils as well as in human breast milk were analyzed in different regions in the Taurus Mountains in Turkey (Turgut et al., 2012). The soil samples were taken in seven different geographical heights and composited according to the methods of the German Environment Specimen Bank (http://www.umweltprobenbank.de/en/documents/publications/15883). At each height only one soil sample was retained. Concerning the breast milk samples, women at five different heights were considered. At each height a different number (from three to fourteen) of human breast milk samples was analyzed.

The contamination of soils has already been evaluated and published in this journal by Turgut et al. (2012). The determination of the occurrence of OCPs in breast milk samples was the aim of a recently published study (Voigt et al., 2013a). Voigt et al. (2013b) demonstrated that in Finland, Demark and Turkey the concentration profiles (values of the concentrations in an ordered tuple of samples) of OCPs in breast milk samples are similar to that of the soil samples.

In our current study we evaluate the same number of chemicals, namely eighteen pesticides, in breast milk samples as well as in soil samples and aiming to find out as to how far, concentration profiles between the environmental soil samples (seven samples) can be considered as similar (see Section 2) to those of the human breast milk samples (44 samples). An appropriate data analysis method to answer such is the discrete mathematical method called the Hasse diagram technique (HDT) (Bruggemann et al., 2001). The software package used is the PyHasse software (Bruggemann et al., 2014). This software is written in Python by the first author and it is under constant development. It comprises more than 100 modules which are of great support especially in the data evaluation of environmental health data.

2. Material and methods

2.1. Data matrix

In this approach we want to examine the occurrence of eighteen OCPs presented in Table 1 in a human medium, namely breast milk, as well as in an environmental medium, namely mountain soil. In Table 1 we can see the eighteen chemicals with their used acronym, standard abbreviation, name, and CAS-number. The list comprises persistent organic pollutants and their degradation products. Most of these chemicals have already been banned worldwide in the Stockholm Convention (United Nations, 2013). This convention is a global treaty to protect human health and the environment from persistent organic pollutants (POPs). The Stockholm Convention focuses on eliminating or reducing releases of twelve POPs, the so-called Dirty Dozen. The twelve key POPs that are targeted by the Convention include Aldrin, Chlordane, DDT, Dieldrin, Dioxins, Endrin, Furans, Hexachlorobenzene, Heptachlor, Mirex, PCBs and Toxaphene. Alfa-hexachlorocyclohexane, beta-hexachlorocyclohexane, gamma-hexachlorocyclohexane (lindane) and pentachlorocyclohexane are now included as POPs in Annexes A and C of the Stockholm Convention (2009). Mirex is a good example for POPs traveling long distances in the air and be deposited in areas far from where they were released because Mirex has never been produced and used in Turkey.

These listed organochlorine chemicals are known to pose a serious threat to the environment and consequently to human health. Especially the endocrine disruption potential of some of these chemicals should initiate action worldwide. In a review concerning the history of the discovery of the widespread toxicity of chlorinated hydrocarbons by Rosner and Markowitz (2013), the authors conclude the enormous lag between identification of danger and ultimate regulation of these products which is still a major public challenge.

2.2. Ranking

2.2.1. Partial order relation

Often a ranking aim is not directly measurable. As a proxy for the ranking aim suitable indicators are introduced, which can be measured or calculated by mathematical models. Taken many objects to be ranked and a set of indicators a data matrix results. The data matrix can be considered as a multi-indicator system (abbr.: mis) (Bruggemann and Patil, 2011). The columns of this data matrix represent indicator values expressing a non-measurable ranking aim. That means, the indicators are oriented in a manner that increases in values express increase with respect to the ranking aim.

To rank objects, whose ranking aim is expressed by several indicators as proxies is by far not trivial. The inherent difficulties will be evident when the manifold of different ranking aggregation methods is inspected. Munda and Nardo (2008) write that there is indeed an inherent ambiguity, which can be traced back to the works of Borda and Condorcet in the eighteenth century (Borda, 1784; Condorcet, 1785). Here we apply simple elements of partial order theory, which do not aggregate the indicators and which are therefore not affected by the above-mentioned built-in ambiguity.

Partial orders in multi-indicator systems can be introduced in many different ways. Most obvious is to set:

Let X be an object set and $x, y \in X$, and let $q_i\,(i=1,...,m)$ be the m indicators which we conveniently consider as elements of a set too, namely of the "information base", IB and $m=|IB|$ (Bruggemann et al., 1995). Note we use objects and elements of a set interchangeably. Elements of a set will not always be objects; hence both notions are needed. We define then:

$$x \le y : \Leftrightarrow q_i(x) \le q_i(y) \text{ for all } q_i \in IB. \quad (1)$$

In application of Eq. (1) we assume that objects equivalent to each other with respect to their profiles $(q_1(),q_2(),...,q_m())$ are identified and only representatives of the equivalence classes are retained. If needed, the equivalent elements are taken into consideration appropriately.

By Eq. (1) a partial order is established among the elements of X, i.e. X is a partially ordered set (poset) by means of Eq. (1). The partial order relation is

- reflexive (i.e. $x \in X$ can be compared with itself),
- antisymmetric (i.e if $x \le y$ and $y \le x$, then $x=y$),
- transitive (i.e. if $x \le y$ and $y \le z$ then $x \le z$).

The resulting partially ordered set is denoted as $P=(X, IB)$, in order to refer to the actually applied mis.

The application of Eq. (1) can lead to the following situation, which may bother decision makers or stakeholders:

Object x may have the profile: (1,1,1,...,1,2) and object y the profile (2,2,2,.....,2,1). It seems to be obvious that should be $x \le y$; however the last indicator $q_{IB|}(x)$ and $q_{IB|}(y)$ does not allow stating an order relation between x and y. If namely the indicator $q_{IB|}$ is considered as element of IB, i.e. considered as a proxy for the ranking aim, and if there is no causal argument to aggregate $q_{IB|}$ with other indicators in a convincing manner, then indeed there is a conflict between x and y and partial order theory alerts us about this fact.

Note that Eq. (1) has a counterpart in fuzzy systems, where the user has a control as to how far numerical differences in indicator values should be considered better as equivalence or not (Bruggemann et al., 2011). Furthermore, P can be considered as a set of pairs (x, y), $x, y \in X$, with $x \le y$.

2.2.2. Some notational remarks

a) Objects, for which $x \le y$, or $x \ge y$ (according to Eq. (1)) are called comparable.
b) Objects, for which Eq. (1) does not hold, are called incomparable. Object x, being incomparable with y, is denoted as $x \parallel y$. In analogy to ameba diagrams we call the fact that data profiles lead to incomparability a "crisscrossing", because in ameba diagrams for instance the lines belonging to two objects are crossing each other.
c) Chains:

Let $C \subseteq X$, if all $x, y \in C$ obey (1) then C is called a chain. (2)

d) Antichain:
Let $AC \subseteq X$, if for all x, $y \in AC$ is valid $x \parallel y$, then AC is called an antichain.
e) Cover relations:
Let x, y, $z \in X$, when $x < y < z$, then y "is between" x and z. If $x < z$ without any element $\in X$, which is between x and z then z "covers" x or x is covered by z. A cover relation is denoted by $x <: z$.
f) The cover relations are the basis to draw Hasse diagrams; see for instance Davey and Priestley (1990) and Halfon and Reggiani (1986). From the use of Hasse diagrams to identify chains and antichains, the often used name Hasse diagram technique (HDT) is derived.
g) Weak order: a sequence of objects, where ties are allowed. For example

$$a < b < c \cong d < e \quad (3)$$

h) Average height $hav(x)$: a characteristic quantity derived from a poset according to Winkler (1982). The average height can be used to derive a weak order of objects. Therefore this quantity often is called "average rank" in applied studies.

2.2.3. Application of partial orders

The easiest way to apply partial orders is to draw a Hasse diagram, which is based on the cover relations and which needs some additional heuristics in some cases, which however do not affect the poset P.

Then we can start

a) to analyze why an object x is located in the Hasse diagram at a certain position and analyze this position in terms of the profile of x. This close interaction between the location of an object within an algebraic graph (in fact a Hasse diagram is a directed, acyclic, transitively reduced graph) and its profile can be formalized; see Wolski (2004). However, here this formalization is not in the focus of our paper. In this context the identification of an object x being a
   - maximal element (no other elements are covering x)
   - minimal element (no other elements are covered by x)
   - isolated element (x is at the same time a minimal and a maximal element)

   is of primary interest.
b) The next step is to consider x as an element of a chain. A suitable long chain (i.e. having many elements of X) is of interest, because a chain indicates that indicators are simultaneously not decreasing, when e.g. started at the bottom of the chain and proceed upwards ("vertical analysis").
c) Another step is to see x as an element of an antichain ("horizontal analysis"). Why is x not comparable with say y? Which indicators cause this incomparability? Is the incomparability caused by only one pair of indicators or by several pairs and if yes, what is common for these indicators?
d) Objects may be considered to be elements of subsets of X which are not related by each other by order relations or only by few ones. We speak of separated subsets and an analysis may be performed to identify (for example by the use of tripartite graphs (Bruggemann and Voigt, 2012).

Often Hasse diagrams are not clear enough then the results (a)–(d) can be obtained by calculating the corresponding information in form of tables, bar diagrams etc., or by using posetic coordinates (see Myers and Patil, 2014). The worst situation however appears when X is an antichain. In this case one should reconcile the multi-indicator system.

2.2.4. Modeling

Let $x \le y$ with respect to a multi-indicator system (mis_1) P_1 (we write $x \le_1 y$), and consider x, y with respect to another multi-indicator system mis_2 and its corresponding partial order P_2. Then it is of interest as to how far $x \le_1 y$ is realized by $x \le_2 y$. We call this a "modeling" because a similarity between two mis allows hypothesizing appropriate mechanism. The corresponding data matrices are called dm_1 and dm_2. It is convenient to introduce three concepts:

The number of common $\le -$relations of P_1 and P_2 is called the number of isotone relations. The fact that for $x,y \in X$ $x \le_1 y$ and $x \le_2 y$ is called an isotone relation.

The fact that for $x, y \in X$ $x \le_1 y$ and $y \le_2 x$ is called an antitone relation and the number of antitone relations is an important quantity too.

Finally: the fact that for $x,y \in X$ $x \parallel_i y$ and $x \le_j y$ $(i, j=1, 2)$ or $x \parallel_1 y$ and $x \parallel_2 y$ is called an indifference relation.

Technically the modeling is performed by the similarity (or proximity) analysis of two posets. In Section 2.3 we first consider the similarity analysis and in Section 2.3.3 we consider the special case that the number of indicators in the two different mis is very different.

2.3. Similarity analysis

2.3.1. Distance based measures

First of all it is to be clarified that here the focus is on similarity between different posets, based on different mis. The similarity of objects within one given poset is not of interest (see Klein, 1995). The study of similarity among a set of posets is of considerable interest, not only in the mathematical literature but in increasing amount also in the applied one. See for example Fattore and Grassi (2014). Whereas Fattore and Grassi apply a lattice theoretical approach, we understand that in our study the statistical aspects are prevailing. Similarity among posets which aim to quantify the role of each indicator of one and only one mis is well established (see Bruggemann and Patil, 2011). Here the similarity between two posets is of interest which are based on two different mis.

Basically there are two sets to be compared, P_1 and P_2 and therefore measures of set-similarities, such as Tanimoto-Index could be useful (see for instance Bock, 1974). Indeed this kind of measure was applied in a pesticide monitoring study (Sørensen et al., 2003). The problem is that the Tanimoto Index and other scalars seem to be too highly aggregated to be informative for a comparison of two posets, because not only isotone-, or antitone-relations are to be checked, but also the indifference relations.

2.3.2. Distance between two weak orders

In many publications the construction of weak orders is described (see for instance Bruggemann et al. (2004) and Bruggemann and Carlsen (2011)). Hence, we may assume that two weak orders derived from two mis are available. Then, clearly, different association measures are available. Once again the results do not regard the manifold of posetic structures. Therefore the idea is to describe the proximity of two posets by a tuple of numbers (tuple as a generalization of pair, triple, quadruple), as can be seen in the next section. In several publications Voigt et al. (2013a) and (2013b) demonstrated how to compare different partial orders, without the need of an aggregation to two weak orders. The basic theoretical idea is published, and detailed information can be found in Bruggemann and Patil (2011).

2.3.3. Similarity of mis with different number of indicators

2.3.3.1. Statistical thesis. The following concepts are technically realized in the module CombiSimilarity: consider a poset with m indicators. Then the set P may have $n_m=f(m)$ pairs (x,y), describing a relation $x \le y$. When a new indicator is added then $n_{m+1} \le n_m$ and the effect of a new inserted indicator is twofold: additional information contextually (otherwise it should not be considered as an element of IB) but also in general reducing the number of comparabilities according to the levels of correlation with the indicators already included in the mis. As can be supposed, the general impact of the new indicator *alone* does not contradict the ranking aim; however within the context of all the other indicators some order relations may be broken, so that finally the contribution of the $\parallel$-pairs is dominating. Hence, when additional indicators are inserted into the mis they imply at the same time noise. Noise in turn will hamper a comparison of two mis. The more the two mis differ in their number of indicators, with $m_1 \ll m_2$, the more the matrix dm_2 will reveal

incomparability relations. Therefore, the idea is to introduce a statistical approach to the proximity analysis: the null-hypothesis is a discrete random variable D, appropriately approximated by a continuous random variable (see for instance Lehmann (1986)) describing the two mis, supports that the two sets of order relations are belonging to the same (but unknown) set. Following this idea there are two steps needed:

1) How can we get D and
2) with which theoretical distribution is D to be compared.

2.3.3.2. Methods to obtain the random variable D. This section refers to step 1) described above. Let λ be the data matrix with m_1 indicators and dm_2 that with m_2 indicators and let have both data matrices n rows, corresponding to n objects. We assume without violating the generality that $m_1 < m_2$. Then there are two possibilities, which seem to be computationally tractable:

1) Take a column out of dm_1 and one column out of dm_2 and construct with these two columns a new data matrix, which we call dm_{12}. Then we can get $m_1 * m_2$ different dm_{12}—n by 2—matrices. For each single dm_{12} the number of order relations and of $x \| y$ relations as well as of $x = y$ − relations can be counted. Let us select the number of order relations C_{12}. Then $m_1 * m_2$ C_{12}-values are obtained. The distribution of C_{12} is considered as a realization of the random variable D.
2) Another variant, which however up to now is less intensively studied is to take $m_{21} = (m_2/m_1)$ submatrices from dm_2. From each of these n by m_1 submatrices, *dmsub*, the partial order, based on (1) can be derived and the number of comparabilities can be checked. The number of m_{21} comparisons C'_{12} can be once again considered as a distribution and hence as a realization of D.
The random variable D can therefore basically realized as Fig. 1 shows.

Both variants have the disadvantage, not to consider both mis with all their indicators simultaneously, as it is the case if the partial order is constructed. Therefore regarding the restriction on less indicators may lead to an optimistic view on the similarity. Other variants, such as analyzing all possible submatrices are tasks for the future; however the computational realization will be difficult.

2.3.3.3. Simulated (theoretical) distribution. The random variable D (realized by C_{12}) is to be compared with a simulated one. In the case of the variant of $m_1 * m_2$ n by 2-matrices the simulated (theoretical) distribution can be easily derived: the simulated distribution is independent of the specific values chosen. It is sufficient to take for instance the numbers $(1,2,\ldots,n)$. Let s be a vector of length of the numbers $1,2,\ldots,n$ and $\pi(s)$ a permutation of s. The series of 2 by n-matrices formed by the columns s and $\pi(s)$ is obtained by a Monte-Carlo-like procedure. From each of these matrices the number C is determined and leads to a simulated distribution $C_{12}^{(t)}$.

Null-hypothesis: the two samples, described by C_{12} and $C_{12}^{(t)}$ have the same number of isotone relations in the average (i.e. $\leq_1$ and at the same time $\leq_2$-relations between $x,y \in X$).

If the null-hypothesis cannot be rejected (t-test) then the distribution of C_{12} is considered as random distribution and there is no statement about proximity possible.

If the null-hypothesis is rejected (t-test) then we want to know: is the expectation value of $C_{12} > C_{12}^{(t)}$ or not. Finally we want to measure with which probability, p the null-hypothesis is erroneously rejected. Fig. 2 shows the statistical decision situation.

2.3.3.4. Technical realization. Performing the t-test needs on the one side an estimation of the degree of freedoms fg, based on means and of standard deviations once of the empirical and once of the theoretical distribution of C_{12} and $C_{12}^{(t)}$, respectively. In order to let the formulas not too troublesome, we introduce:

m_1 for mean of C_{12}, sd_1 for standard deviation of C_{12}
m_2 for mean of $C_{12}^{(t)}$, sd_2 for standard deviation of $C_{12}^{(t)}$.

$$H_i = sd_i^{\,2}/n_i \tag{4}$$

$$N := H_1 + H_2 \tag{5}$$

$$t = \left| \frac{m_1 - m_2}{\sqrt{N}} \right| \tag{6}$$

$$fg = \frac{N^2}{(H_1^2/(n_1-1)) + (H_2^2/(n_2-1))} \tag{7}$$

The null-hypothesis will be tested by t, whereas to calculate the p-value needs both t and the greatest least integer of fg. The probability p is the probability for the type I error. Although the calculation details can be found in Abramowitz and Stegun (1968), it may be convenient to see the details here, especially as these formulas are applied in the software PyHasse:

$$\Theta := \arctan\left(\frac{t}{fg}\right)$$

a. $fg = 1$

$$A(t, fg) := \frac{2}{\pi} \cdot \Theta$$

b) $fg > 1$ and odd:

$$\begin{aligned} \Xi &:= \cos\Theta + \left(\tfrac{2}{3}\right) \cdot \cos^3\Theta + \left(\tfrac{2\cdot 4}{1\cdot 3}\right) \cdot \cos^5\Theta + \ldots + \tfrac{2\cdot 4\cdot \ldots (fg-3)}{1\cdot 3\cdot \ldots (fg-2)} \cdot \cos^{fg-2}\Theta \\ A(t, fg) &= \tfrac{2}{\pi} \cdot (\Theta + \sin\Theta \cdot \Xi) \end{aligned} \tag{8}$$

c) $fg > 1$ and even:

$$\begin{aligned} \Psi &= 1 + \tfrac{1}{2} \cdot \cos^2\Theta + \tfrac{1\cdot 3}{2\cdot 4}\cos^4\Theta + \ldots + \tfrac{1\cdot 3\cdot 5\cdot \ldots (fg-3)}{2\cdot 4\cdot \ldots (fg-2)} \cdot \cos^{fg-2}\Theta \\ p &= A(t, fg) = \sin\Theta \cdot \Psi \end{aligned} \tag{9}$$

$A(t,fg)$ is the probability that the two mean-values are different and is called p.

2.3.3.5. Software. In PyHasse the calculation of the statistical test quantities is performed. For this purpose the module CombiSimilarity7_3.py was developed. The values of $C_{12}^{(t)}$ obtained after 1000 runs for each single n-value are stored as a matrix for a series of different numbers of $n = |X|$, i.e. $n = 1, 2, \ldots, 10, 20, 30, \ldots, 100, 200, 300, \ldots, 500$. It is convenient to write $stdev(n)$ and $mean(n)$ of $C_{12}^{(t)}$ as a function of n, instead of storing internally the data $mean(n)$ and $stdev(n)$, implying the need for interpolation:

$$\begin{aligned} std(n) &= P^{\sqrt{2}} \\ P &= a_0 \cdot n^3 + a_1 \cdot n^2 + a_2 \cdot n + a_3 \end{aligned} \tag{10}$$

$$mean(n) = b_0 \cdot n^2 + b_1 \cdot n + b_2 \tag{11}$$

The values of the coefficients a_i and b_i (in both regressions: $R^2 = 1$) are $a_0 = -0.2046 * 10^{-6}$, $a_1 = 2.0833 * 10^{-4}$, $a_2 = 0.3605$, $a_3 = -0.2728$, $b_0 = 0.2502$, $b_1 = -0.2889$, and $b_2 = 1.1319$. The distribution C_{12} and its statistical characteristics such as *stdev* and *mean* are calculated and called "empirical". The values of t, fg, and $p = A(t,fg)$ are calculated as described in the previous section.

3. Application example: eighteen chemicals in soil and breast milk samples

Eighteen OCPs (Organochlorine Pesticides) in samples of soils as well as in human breast milk were analyzed in different regions of the Taurus Mountains in Turkey. Taurus Mountain soils were suggested for this study because of their potential to act as sink for organic pollutants, and it is expected that the atmospheric pollution is present in Turkey as well as in neighboring countries, e.g. Arabia, Africa and Russia. Furthermore this remote region has a very low population density and the women are mainly fixed to their living area nearby the soil sampling locations. The geographical distribution of the samples represents a stretch from the Mersin Sea level to the Northern top mountains. The sampling sites were designed as an altitude profile at 121, 408, 981, 1225, 1373, 1639 and 1881 m, respectively (Turgut et al., 2012). Hence, in case of the soil samples seven samples were taken and analyzed, whereas 44 breast milk samples (from women, living in five different heights) have to be looked upon. Concerning the breast milk samples, only healthy women who had no occupational history to OCP's and PCBs were selected for this study (see the references in the study of Cok et al. (2012)). The study protocol was reviewed and approved by the Ethical Committee of Mersin University for Human Studies of the School of Medicine (ethical committee number: 15.05.2009-5/110). The applied methodologies for sampling and analysis are compliant with standardized procedures and generated in accredited laboratory environment at MEX (Molecular Exposomics Laboratory). Per plot at least five

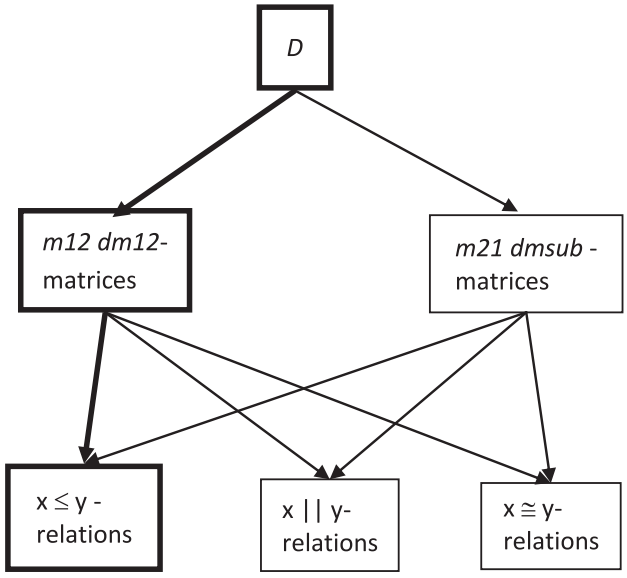

Fig. 1. Two variants of the random variable *D*. The text is focused on the bold drawn parts of the scheme.

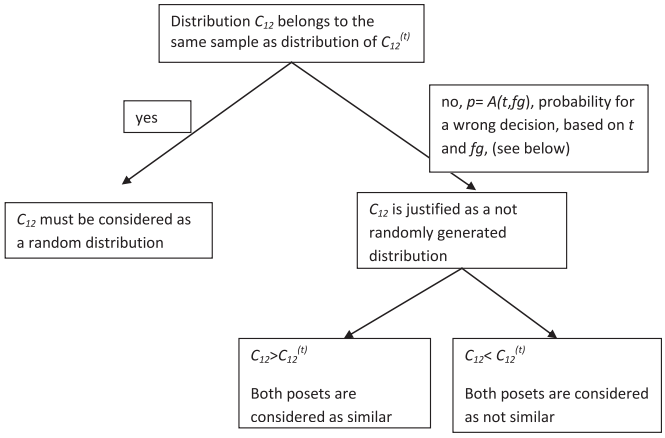

Fig. 2. Decision tree on the basis of C_{12} and $C_{12}^{(t)}$.

samples were collected and mixed to get an average sample. Thus sample errors are kept as small as possible.

The two data sets (18 chemicals × 7 soil samples versus 18 chemicals × 44 breast milk samples) offer an example for the application of the CombiSimilarity analysis described above. A brief description of this example has already been given in the Workshop Simulation in Environmental and Geo Sciences in Leipzig, Germany in April 2013 (Voigt et al., 2013c).

3.1. Main partial order model: Hasse diagrams for two data matrices

3.1.1. Remark

In the following analysis the data matrices are analyzed as follows:

objects: the chemicals,
indicators: the samples (soil: seven samples; milk: 44 samples),
concentration profile: ordered set of seven and 44 samples and the concentration values of the considered chemical.

3.1.2. The two partial orders

First we calculate the Hasse diagram for the 18 chemicals × 7 soil samples and 18 chemicals × 44 breast milk applying the main Hasse diagram technique module of the PyHasse software (mHDCl2_7); see Fig. 3.

One of the maximal objects in both diagrams is the chemical PPDE, the first degradation product of the pesticide DDT, and the chemical MECH is a minimal object in both diagrams. Also some analogies in the chains, e.g. PPDE > HCBE > PECE or PPDT > GHCH > MECH can be seen in both diagrams. However, no precise answer can be given, whether or not the pollution of soil is similar to the pollution pattern of breast milk with respect to the order relations among the concentration profiles. i.e. how far a relation chemical $x <$ chemical y in soil can imply the same relation with respect to breast milk samples.

3.2. CombiSimilarity analysis of soil samples versus breast milk samples

We calculate the similarity using the attribute–attribute ("att_att") analysis for the isotone relation which is the second variant offered by the CombiSimilarity module. This procedure leads to an empirical distribution of $m_1 * m_2$ quantities of isotone relations. The expected value of this distribution and the standard deviation are obtained, and called *mean*(empir) and *stdev*(empir). These parameters are further examined by appropriate statistical tests. As usual, the Null-hypothesis is that we have to test for "belonging to the same distribution".

Theoretically, a distribution can be calculated assuming that the values of each column of the n by 2 matrices are permutated. Theoretically the range of the degree of isotone varies from 0to 1. The data rendered in Table 2 are obtained by the methods described above. Table 2 is the basis for a decision whether or not two posets can be considered as similar with respect to their order relations among concentration profiles. Beyond this it is of interest to visualize the theoretical distribution together with the empirical one. Hereto a simulation with 100,000 runs was performed to obtain this theoretical distribution (Mathematica). Note, the aforementioned simulation with 1000 runs was used to derive the functions *stdev*(n) and *mean*(n).)

From the theoretical distribution, the expected value (*mean* (theor)) and the standard deviation (*stdev*(theor)) will be determined too. The two distributions are shown in Fig. 4.

The test of the Null-hypothesis is performed by the Students-*t*-test allowing for unequal variances (Sachs, 2002). As seen in Fig. 4, the distributions of isotone random and isotone soil/milk are different (*t*-test) with $p < 0.01$. The distributions of soil/milk show higher isotone relations. This means similarity. High isotone character (statistically significant) differs considerably from the theoretically expected distribution approach. These calculations demonstrate that soil/milk Hasse diagrams are similar.

4. Summary, discussion and conclusion

Many monitoring studies concerning the fate of organic pollutants are available. Of special interest is when monitoring results are concerned with the same set of chemicals but with different targets. In our study the contamination of eighteen Organochlorine Pesticides is investigated and the targets are "soil" and "breast milk". The question arises: Is there a causal argument that the contamination of soil has an influence on the contamination of the human breast milk? The best way to get an answer would be a deterministic modeling approach, which calculates the transport paths from emitters to soil and emitters to humans on the one side, and the uptake of chemicals described by paths from soil to humans on the other side. Then a mass balance taking account the different paths should allow an answer. Appropriate mathematical models would be at hand; however, the background information

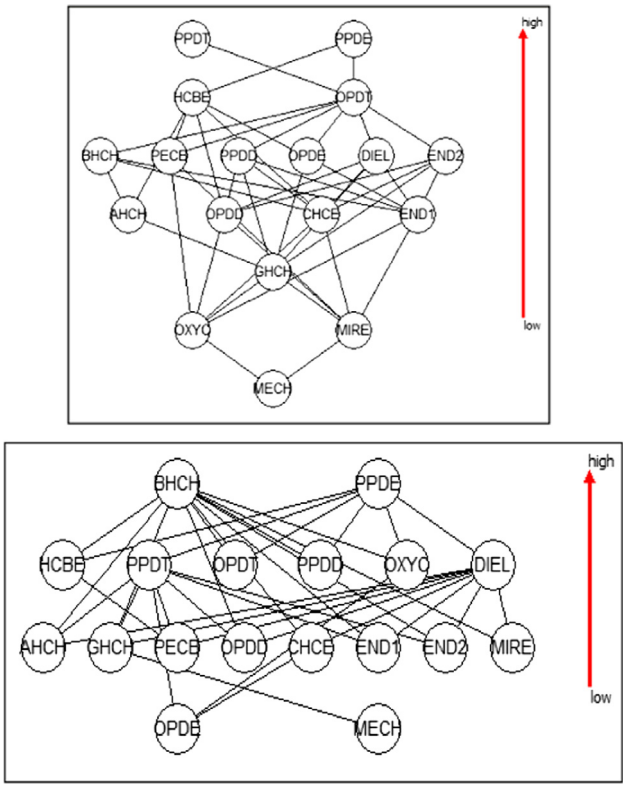

Fig. 3. Hasse diagrams for 18×7 soil (top) versus 18×44 breast milk samples (bottom).

Table 2
Results of CombiSimilarity (PyHasse software).

mean(theor) $C_{12}^{(t)}$	76.99	
stdev(theor)	13.44	
mean(empirical) C_{12}	107.84	
stdev(empirical)	6.33	
fg	1100	Eq. (7)
t	55.34	Eq. (6)
$A(t, fg)$	0.00	Eq. (9)

on the emitters of chemicals and where and when the emission takes place, as well as many model parameters are missing. Hence statistical modeling is needed.

Accepting the usefulness of statistical modeling, one may start on the modest level, namely what can be deduced from the order relations of chemicals referring to the soil and those of the breast milk samples. This is the task where this paper is starting from. When the target "soil" is considered, the partially ordered set allows many insights (only few of them are mentioned here), whereas the target "breast milk" leads to a poset, where incomparabilities are dominating. Any human sample is slightly different from the other implying many crisscrossings of the concentration profiles although the study design targeted already humans which are living and feeding mostly from their habitat for a long time and had been less mobile than individuals in cities or other areas. Therefore the comparison of the two posets is hampered by the many $\leq$-relations of the soil—target whose counterparts in breast-milk samples are predominantly $\|$-relations. The fraction of isotone-relation is so low, and that of indifference relation so high that the posets should be considered as dissimilar. However, taken into account that many crisscrossings are only induced by slight deviations in the numerical concentration values and that there is a natural variability in the samples a more refined concept is needed. Here for the first time a statistical test-concept applied on partial orders is presented. The result is that the evidence is that the two posets should be considered as similar. Returning to the question posed, it can be hypothesized that either the transport path from soil to human is strong enough to cause this similarity or both targets are the sink of two paths from the emitters, which are strong enough to cause a related pattern in concentration values concerning the eighteen OCPs. The hypothesis is supported by the fact that the chemicals investigated in our study are indeed long-range transported. These chemicals transported into the Taurus Mountain region are not applied locally. They can be expected to represent persisting impact toward soil as well as humans as suitable recipients.

Unfortunately, this analysis only ends with hypotheses. Hence, one may ask what can be done to refine the finding of this study. First of all, other variants as shown in Fig. 1 should also be applied. It is clear that by taking one column of the one and one column from the other data matrix does not encompass the simultaneous effects of all indicators of both mis. By avoiding noise induced by all indicators, the result of analysis performed here may be too

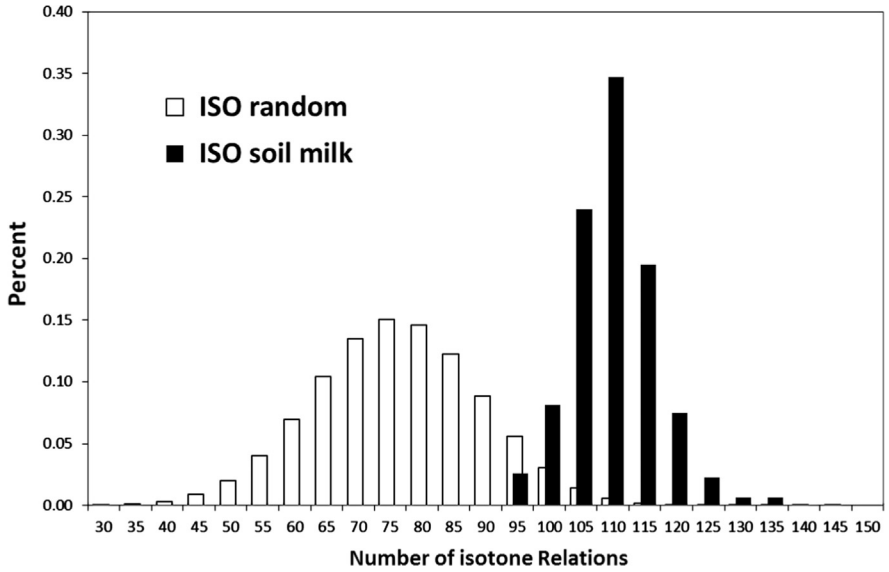

Fig. 4. Calculation of isotone relations random versus soil/milk data.

optimistic. Other variants to obtain the random variable D must be studied.

Furthermore, instead of checking the isotone-relations alone also the distribution of antitone - and indifference-relations should be studied. Although it is not expected that the results will contradict each other further insights confirming our hypotheses may be obtained.

When the numerical values of the concentrations lead to non-relevant incomparabilities, the application of fuzzy partial order is of interest. Whereas the fuzzy partial order is established in PyHasse (Bruggemann et al., 2011), the needed and suitable statistical test machinery is still not implemented. These mentioned and partially planned activities are purely related to the data situation as given at the moment. An interesting aspect would be if the data analysis could be repeated where the set of breast-milk samples is stratified according to age, health, and other traits in humans as well as that of the target soil (may be according to the geographical height). Corresponding investigations will be performed in the future.

References

Abramowitz, M., Stegun, I.A., 1968. Handbook of Mathematical Functions. Dover Publications, New York.

Birnbaum, L.S., 2013. When environmental chemicals act like uncontrolled medicine. Trends Endocrinol. Metab. 24, 321–323.

Bock, H.H., 1974. Automatische Klassifikation. Vandenhoeck and Ruprecht, Goettingen, Germany.

Borda, J.C., 1784. Mémoire sur les élections au scrutin. Histoire de l'Academie Royale des Sciences, Paris.

Bruggemann, R., et al., 1995. Use of Hasse diagram technique for evaluation of phospholipid fatty acids distribution as biomarkers in selected soils. Chemosphere 30, 1209–1228.

Bruggemann, R., et al., 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.

Bruggemann, R., et al., 2004. Estimation of averaged ranks by a local partial order model. J. Chem. Inf. Comput. Sci. 44, 618–625.

Bruggemann, R., Carlsen, L., 2011. An improved estimation of averaged ranks of partial orders. MATCH: Commun. Math. Comput. Chem. 65, 383–414.

Bruggemann, R., et al., 2011. Ranking objects using fuzzy orders, with an application to refrigerants. MATCH: Commun. Math. Comput. Chem. 66, 581–603.

Bruggemann, R., Patil, G.P., 2011. Ranking and Prioritization for Multi-indicator Systems. Springer, Berlin.

Bruggemann, R., Voigt, K., 2012. Antichains in partial order: pollution in a German region by lead, cadmium, zinc and sulfur in the herb layer. MATCH: Commun. Math. Comput. Chem. 67, 731–744.

Bruggemann, R., Carlsen, L., Voigt, K., Wieland, R., 2014. PyHasse software for partial order analysis. In: Bruggemann, R., Carlsen, L., Wittmann, J. (Eds.), Multi-Indicator Systems and Modelling in Partial Order. Springer, New York, pp. 389–423.

Chan, J.K., Wong, M.H., 2013. A review of environmental fate, body burdens, and human health risk assessment of PCDD/Fs at two typical electronic waste recycling sites in China. Sci. Total Environ. 463–464, 1111–1123.

Cok, I., Mazmanci, B., Mazmanci, M.A., Turgut, C., Henkelmann, B., Schramm, K.-W., 2012. Analysis of human milk to assess exposure to PAHs, PCBs and organochlorine pesticides in the vicinity Mediterranean city Mersin Turkey. Environ. Int. 40, 63–69.

Condorcet, M.d., 1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la probabilité des voix. De l'Imprimerie Royale, Paris.

Davey, B.A., Priestley, H.A., 1990. Introduction to Lattices and Order. Cambridge University Press, Cambridge, UK.

Fattore,M., Grassi.R., 2014. Measuring dynamics and structural change of time-dependent socio-economic networks, 10.1007/s11135-013-9861-1, http://dx.doi.org/10.1007/s11135-013-9861-1.

Halfon, E., Reggiani, M.G., 1986. On ranking chemicals for environmental hazard. Environ. Sci. Technol. 20, 1173–1179.

Klein, D.J., 1995. Similarity and dissimilarity in posets. J. Math. Chem. 18, 321–348.

Knol, A.B., et al., 2010. Assessment of complex environmental health problems: framing the structures and structuring the frameworks. Sci. Total Environ. 408, 2785–2794.

Lehmann, E.L., 1986. Statistical Hypotheses, 2nd ed. Wiley, New York.

Lithner, D., et al., 2011. Environmental and health hazard ranking and assessment of plastic polymers based on chemical composition. Sci. Total Environ. 409, 3309–3324.

Ludwig, H.F., 1955. Chemicals and environmental health. Am. J. Public Health Nations Health 45, 874–879.

Mnif, W., et al., 2011. Effect of endocrine disruptor pesticides: a review. Int. J. Environ. Res. Public Health 8, 2265–2303.

Munda, G., Nardo, M. 2008. Noncompensatory/nonlinear composite indicators for ranking countries: a defensible setting. In: J. R. C., European Commission, Institute for the Protection and Security of the Citizen (IPSC), Econometrics and Statistical Support Antifraud (G-09), 21020 Ispra (VA), Italy, (Eds.), Applied Economics, 2008, pp. 1–11.

Myers, W.L., Patil, G.P., 2014. Coordination of contrariety and ambiguity in comparative compositional contexts: balance of normalized definitive status in multi-indicator systems. In: Bruggemann, R., Carlsen, L., Wittmann, J. (Eds.), Multi-Indicator Systems and Modelling in Partial Order, 2014 Springer, New York, pp. 167–196.

Rosner, D., Markowitz, G., 2013. Persistent pollutants: a brief history of the discovery of the widespread toxicity of chlorinated hydrocarbons. Environ. Res. 120, 126–133.

Sachs, L., 2002. Angewandte Statistik. Springer-Verlag, Berlin.

Sørensen, P.B., et al., 2003. Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory. Environ. Toxicol. Chem. 22, 661–670.

Stephenson, M.E., 1977. An approach to the identification of organic compounds hazardous to the environment and human health. Ecotoxicol. Environ. Saf. 1, 39–48.

Stockholm Convention, 2009 ⟨http://chm.pops.int/Convention/ThePOPs/ListingofPOPs/tabid/2509/Default.aspx⟩ (accessed: 22.10.2013).

Turgut, C., et al., 2012. The occurrence and environmental effect of persistent organic pollutants (POPs) in Taurus Mountains soils. Environ. Sci. Pollut. Res. Int. 19, 325–334.

United Nations,2013. The Stockholm Convention on persistent organic pollutants (POPs), United Nations Industrial Development Organization, Geneva, Switzerland, ⟨http://www.unido.org/index.php?id=5279⟩.

Voigt, K., et al., 2013a. Evaluation of organochlorine pesticides in breast milk samples in Turkey applying features of the partial order technique. Int. J. Environ. Health Res. 23, 226–246.

Voigt, K., et al., 2013b. Discrete mathematical data analysis approach: a valuable assessment method for sustainable chemistry. Sci. Total Environ. 454–455C, 149–153.

Voigt, K., Bruggemann, R., Scherb, H., Cok, I., Mazmanci, B., Mazmanci, M.A., Turgut, C., Schramm, K.-W., 2013c. Organochlorine pesticides in the environment and humans: Necessity for comparative data evaluation. Pages 9-22. In: Wittmann, J., Müller, M. (Eds.), Simulation in Umwelt- und Geowissenschaften; Workshop Leipzig 2013. Shaker-Verlag, Aachen, Germany.

Winkler, P.M., 1982. Average height in a partially ordered set. Discret. Math. 39, 337–341.

Wolski, M., 2004. Galois connections and data analysis. Fundam. Inform. 60, 401–415.