MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **54** (2005) 643-670

ISSN 0340 - 6253

Application of multidimensional rank-correlation

Peter B. Sørensen[1*], Rainer Brüggemann[2], Marianne Thomsen[1], Dorte B. Lerche[1]

[1]: Department of Policy Analysis, The National Environmental Research Institute (NERI)

Frederiksborgvej 399, PostBox 358, DK-4000 Roskilde, Denmark

[2]: Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB)

Müggelseedamm 310, D - 12587 Berlin, Germany

[*]: Corresponding author (email: pbs@dmu.dk)

Abstract

A novel application of partial order theory for object rank correlation analysis of multiple variables is presented by the new software package *PO Correlation*. The design is made transparent for rank correlation analysis by a detailed mapping of the rank relations between all objects. In contrast to conventional rank correlation methods, it is possible to identify specific relations among the single variables using *PO Correlation*. The principle and application is described using a specific data set based on environmental monitoring of pesticide findings in small streams in Denmark, however, the *PO Correlation* is usable for many different correlation problems. It is shown how *PO Correlation* is effective for data interpretation by identifying a series of useful conclusions in relation to the mechanism of pesticide exposure under realistic conditions, which hardly can be found using conventional procedures of data analysis. The difference in usage pattern for different pesticides seems to be the main governing factor for pesticide exposure in small streams. This is not trivial, and challenges the conventional understanding of pesticide exposure, which claims that differences in physico-chemical properties including degradation and adsorption are important factors governing differences in pesticide exposure in surface waters.

1 Introduction

A novel application of partial orders for rank correlation analysis is presented by the new software *PO Correlation*. The design supports an easy, robust and transparent analysis of correlation details. Non-commercial use of the software for research is free if reference is given to this paper and a version can be made available by contacting the first author (email: pbs@dmu.dk). Two other freely available software products exist for application of Partial Order Theory in decision support: (1) the software WHASSE , Brüggemann et al. (1999); (2) The software ProRank (Vers. 1.0), Pudenz, (2004). The software *PO Correlation* differs from the other programs by focusing on the correlation analysis between partially ordered sets.

The method is an extended and improved version of the methodology for assessing ranking similarity as presented by Sørensen et al., (2003). This paper will not make a comprehensive review of ranking correlation methods, however, and for a more general discussion of ranking correlation see *e.g.* Brüggemann et al., (2001), Brüggemann et al., (1995), Bath et al., (1993), Moock et al., (1998), Conover, (1999), Gibbons, (1993), Pavan, (2003) and Pudenz, (1998).

2 Data background

The data set, selected for illustration, is taken from the Danish Monitoring Program (NOVA 2003) and includes pesticides finding during the year 2000 in small streams, see Table 1. The data set is based on 23 sampling stations, each covering a separated catchment area. At each station, 6 water samples were analysed for a series of pesticide active ingredients, in the following denoted pesticides. The detection frequency (*DetFreq*) is defined as the frequency for a pesticide to be detected above detection limit in the joint set of measurement from the 23 sample stations. If the set of stations is assumed representative for Danish conditions then the *DetFreq* is a measure for the propagation of a given pesticide in the stream water environment in Denmark. For each station and for each pesticide, the highest measured concentration level among the 6 single samples is identified. This yields 23 maximal concentration values and the median (*MedMax*) is subsequently calculated characterising the level of contamination.

Table 1. The data set used in the correlation analysis as input for *PO Correlation***.**

		Predicted variables *Set 1*		Predicting variables *Set 2*	
Id	Substances	*DetFreq* (%)	*MedMax* (ng/l)	Dose (g/ha)	SpArea (1000 ha)
1	2,4_D	2	40	0	0
2	Atrazine	9	30	0	0
3	Bentazone	36	20	523	91
4	Bromoxynil	6	80	383	110
5	Carbofuran	0	0	659	1
6	Chloridazon	1	380	0	0
7	Chlorsulforon	2	30	0	0
8	Cyanazin	2	200	0	0
9	Diclorprop	7	70	847	2
10	Dimethoat	2	40	304	81
11	Ethofumesat	5	90	491	31
12	Fenpropimorph	2	70	477	249
13	Glyphosat	76	220	1172	573
14	Ioxynil	6	30	349	113
15	Isoproturon	40	130	2750	4
16	Maleinhydrazid	1	10	1790	0.3
17	MCPA	20	140	1410	101
18	Mecoprop	17	30	900	13
19	Metamitron	8	90	2098	48
20	Metribuzine	1	50	250	27
21	Metsulfuron methyl	1	10	5	151
22	Pendimethalin	12	40	1368	178
23	Pirimicarb	4	30	135	7
24	Propiconazole	6	20	6837	3
25	Terbuthylazine	33	100	1500	22

So, the two numbers *DetFreq* and *MedMax* together form an eco-toxicological meaningful way of characterising occurrence by taking into account both propagation (*DetFreq*) and level (*MedMax*) as discussed by Sørensen et al., (2003).

The ranking of pesticides using data of *DetFreq* and *MedMax* together will be compared with two variables for the pesticide field application in the form of recommended dosage level (*Dose* in g/ha) and total sprayed area in Denmark (*SpArea* in 1000 ha). The application data are taken from the Danish sales statistics and the reported recommended dosage. The data set is shown in Table 1. Some of the pesticides in the monitoring program have been banned since year 1995 and are thus not used in the year 2000. They are identified in Table 1 as: *Dose*=0 and *SpArea*=0.
The primary topic in this correlation analysis is to investigate the coincidence between the ranking of pesticides based on the measured variable set *DetFreq* and *MedMax* on one side *(Set 1)* and the field application in terms of the variable set *Dose* and *SpArea* on the other side *(Set 2)*. The variables *DetFreq* and *MedMax* are denoted the **predicted** variables while *Dose* and *SpArea* are denoted the **predicting** variables.

3 Correlation analysis

A simple rank correlation measure is *Kendalls Tau* (Kendall, 1938). The principle in *Kendalls Tau* is partly linked to Partial Order Theory as explained in the following. For a set of two variables as *e.g.* the variables *DetFreq* and *MedMax* in Table 1, the ranking of two objects (two pesticides in Table 1) can be done using either the first or the second variable. If the ranking using the first variable is equivalent with the ranking using the second variable then the ranking is claimed to be concordant. A pair of objects is discarded if at least one of the variables is equal or equivalent. In Table 1, a concordant ranking is seen for the ranking of Id. 13 above Id. 8, while Id. 13 > Id. 8 for both the variables *DetFreq* (76>2 in Table 1) and *MedMax* (220>200 in Table 1). A discordant ranking appears when there is discordance between the single variables in the order. The variable pair formed by the Ids. 6 and 13 is an example of a discordant ranking, where *DetFreq* (1<73 in Table 1) and the *MedMax* (380>220 in Table 1) yields a different ranking of the two objects. The number of concordant rankings is denoted by

C and the number of discordant rankings is denoted by *D*. A modified *Kendalls Tau* is used in *PO Correlation* as suggested by Goodman and Kruskal, (1963):

$$\tau = \frac{C - D}{C + D}$$

The τ value is 1 for complete ranking agreement and –1 if we have complete disagreement between the two variables.

The correlation between two variables can also be graphically displayed as a partial order using a Hasse Diagram. This is illustrated in the following for the variables *DetFreq* and *MedMax* in Figure 1. This diagram ranks two pesticides if they are neither equal nor discordantly ranked. They are equal if they have the same values for both *DetFreq* and *MedMax*. A discordant ranked pair of objects in the Hasse diagram does not have downward connecting lines between the two objects as *e.g.* seen for the object pair Id. 6 and Id. 8 in Figure 1. A more detailed discussion about these relationships can be seen in Brüggemann and Bartel, (1999). The value for *C* and *D* in case of the ranking using *DetFreq* and *MedMax* is respectively: *C*= 169 and *D*= 97 and τ=0.24, indicates a positive but not strong correlation.

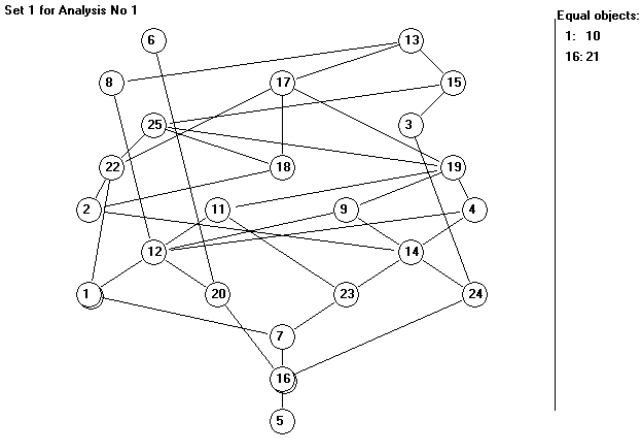


Figure 1. Hasse Diagram using *DetFreq* and *MedMax* as parameters. The numbering refers to the Id's in Table 1. The equal objects, between which both the variables *DetFreq* and *MedMax* are equal, are listed in the right column**.**

Table 2. τ values for all parameter combinations using the complete data set in Table 1.

		Predicted variables		Predicting variables	
		DetFreq	*MedMax*	*Dose*	*SpArea*
Predicted variables	*DetFreq*	1,00	0,27	0,42	0,24
	MedMax	0.27	1,00	0,07	0,07
Predicting variables	*Dose*	0,42	0,07	1,00	0,20
	SpArea	0,24	0,07	0,20	1,00

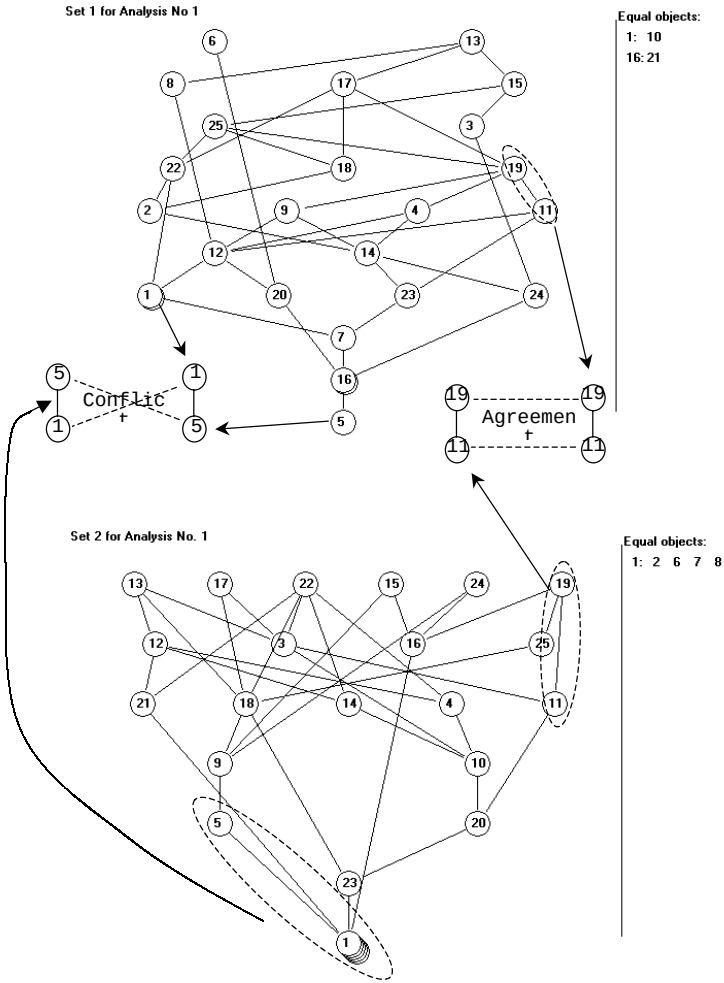


Figure 2. Comparison of two partially ordered sets (Hasse Diagrams). *Set* 1 is the partially ordered set for respectively *DetFreq* and *MedMax* and *Set* 2 is the partially ordered set for respectively the *Dose* and the *SpArea*. All pesticides in Table 1 are included.

The value of τ is calculated for all combinations of variable pairs in *PO Correlation* as shown in Table 2. A general positive correlation is seen in Table 2 for *Dose* and *DetFreq*. However a very weak correlation is seen between *MedMax* and respectively *Dose* and *SpArea*.

The τ -values in Table 2 show only the correlation between pairs of a predicting and a predicted variable, while more complex correlation showing combined ranking of several variables will be investigated in the following. Two partially ordered sets are defined: (1) *Set* 1, composed by *DetFreq* and *MedMax* as also shown by the Hasse Diagram in Figure 1.; (2) *Set* 2 composed by *Dose* and *SpArea*. Both concordant and discordant rankings are compared between the two sets for all pair of objects. This procedure is illustrated for two pairs of objects in Figure 2.

The agreement in rankings between the sets can be graphically shown in an *Agreement Diagram* as a Hasse diagram where all variables in the two sets are applied for ranking in one diagram as defined by Sørensen et al., (2003). This is shown in Figure 3, where all the variables *DetFreq*, *MedMax, Dose* and *SpArea* are used simultaneously.

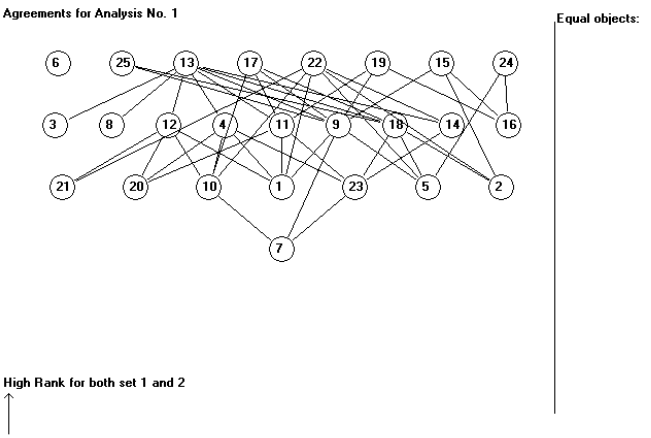


Figure 3. Graphical displays of the agreements (*Agreement Diagram*) between *Set* 1 and *Set* 2 (the sets shown in Figure 2).

The *Agreement Diagram* has a complementary diagram denoted the *Conflict Diagram* (Sørensen et al., 2003) in where the *Set* 1 parameters are ranked upward while the *Set* 2 parameters are ranked downward (inverse rank). Such a diagram is shown in Figure 4, where the variables: *DetFreq*, *MedMax,* negative *Dose* and negative *SpArea* are used. No ranking can exist simultaneously in both the *Agreement Diagram* and the *Conflict diagram*. These two diagrams show important elements of the correlation for each single object in relation to the other objects. The *Conflict Diagram* maps the conflicting ranking for each object. In this way Id. 8 is seen to be ranked above several objects in the *Conflict Diagram* telling that *Set* 1 will like to rank Id. 8 upward while the variables of *Set* 2 rank Id. 8 downward. The top objects in the *Conflict Diagram* having multiple comparisons to other objects are dominated by pesticides banned in 1995 and thus not used in 2000 (The Ids.: 1, 2, 6, 7, 8). Hence the *Conflict Diagram* indicates that the correlation between occurrence and field application is damaged by the fact that some of the pesticides have not been used since 1995. It also tells that there are only a few conflicts between the pesticides, which are still in use.

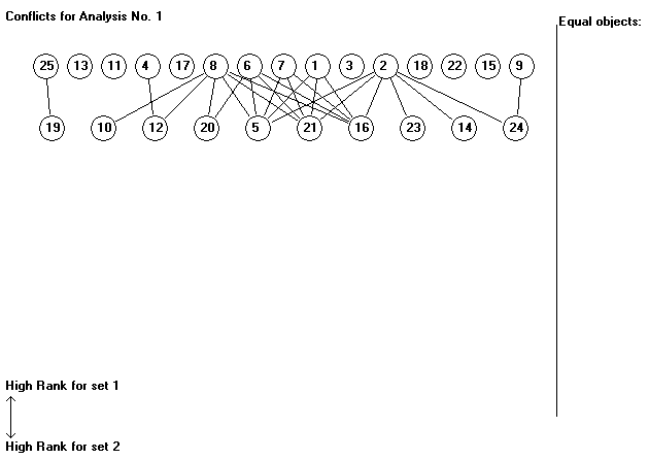


Figure 4. Graphical displays of the conflicts (*Conflict diagram)* between *Set* 1 and *Set* 2 (The sets shown in Figure 2).

Many different indices can be used for similarity between two partial ordered sets. However, it is important to make clear that a single number for similarity is unable to capture all information. The similarity between partial orders is a multi-dimensional problem and any one-dimensional representation will discard information. The principle of the modified Tanimoto-index as a similarity index, *T(...,...)* and the linkage to other concepts are shown in Sørensen et al., (2003).

The quantification of similarity in *PO Correlation* is based on the following counting of object pairs which are:

a: Concordant ranked in both sets having the same ranking (concordant ranked) between the two sets (see Figure 2: agreement).

b: Concordant ranked in both sets but discordant ranked between the two sets (see Figure 2: conflict).

c: Concordant ranked in *Set 1* and discordant ranked in *Set 2*.

d: Concordant ranked in *Set 2* and discordant ranked in *Set 1*.

e: Concordant ranked in *Set 1* and concordant ranked in *Set 2* simultaneously (*e*=*a*+*b*).

f: Discordant ranked in both sets simultaneously.

g: Equivalent in *Set 1* and not equivalent in *Set 2*.

h: Equivalent in *Set 2* and not equivalent in *Set 1*.

i: Equivalent in both *Set 2* and *Set 1* simultaneously.

These definitions will be used in the following tables showing the correlation between the two sets. The similarity of *Set* 1 and *Set* 2 is analysed as shown in Table 3.

The number of agreements between *Set* 1 and *Set* 2 is relatively high (*a*=82) compared with the number of disagreements (*b*=25). Obviously some positive correlation seems to exist between the predicting and predicted variables as also indicated in Table 2. All the pairs of objects that contribute to the value *a* are ranked in the *Agreement Diagram*, Figure 3, while all the pairs contributing to *b* are ranked in the *Conflict Diagram*, Figure 4. The significance of the numbers is tested by an estimate of the probability for a randomly formed value to be equal or larger than the actual value. This procedure will be explained below. If the probability estimate is close to zero the actual value is "relatively large" and probability estimates close to one shows that the actual value is "relatively low". The *T*(0,0) value of 0.77 is seen to be relatively large by having a probability of 0.014 for a random estimate to be larger than or equal to the actual value.

There is a tendency for that a concordant ranking in one set is discordant in the other set as seen in Table 3, where the *c* and *d* values are significantly high (related probability estimate of 0.030). This is supported by relatively low values for *e* and *f*.

Table 3. The correlation results between *Set* 1 and *Set* 2.

Comparison n=25	Counting	Probability for larger or equal value
a	82	0.067
b	25	0.993
c	91	0.030
d	72	0.030
e	107	0.980
f	25	0.980
g	2	0.932
h	10	0.932
i	0	1.000
$T(0,0) = a/(a+b)$	0.77	0.014
$T(1,0) = a/(a+b+c)$	0.41	0.078
$T(0,1) = a/(a+b+d)$	0.46	0.078
$T(1,1) = a/(a+b+c+d)$	0.30	0.124

This shows some degree of correlation within respectively the pairs (*DetFreq*, *MedMax*) and (*SpArea*, *MedMax*), which is not reflected in a corresponding correlation between the pairs.

The similarity between *Set 1* and *Set 2* is governed by the following three factors:

1. The value setting of the descriptors, within the sets. This is illustrated in Table 3 where the value setting of both *Dose* and *SpArea* shows several zero values in pairs and thus many equal objects in *Set 2*. This tends to reduce the number of concordant rankings and thus the potential number of rankings, which can be compared with rankings done in *Set 1*.

2. The inter-correlation (both negative and positive) between the descriptors within the sets. This is important for the number of concordant rankings within each set and thus also for the potential similarity of the two sets.
3. Correlation between the ranking of the predicted descriptors in *Set 1* compared to the ranking of the predicting descriptors in *Set 2*.

Only the third factor is important when the confidence of correlation between *Set 1* and *Set 2* are going to be assessed. So, the challenge is to design a statistical test that can keep the two first factors constant (take them as conditions) and only test the correlation between the two set of descriptors. Keeping the structure of the Hasse Diagram of Set 2 constant as a condition for the test is one way to solve this problem. Such a procedure is shown in the following by using a simple example.

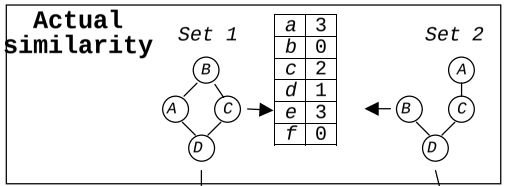
The following simple example will use two, small and arbitrary chosen, partially ordered sets for illustration of the probability estimates, see Figure 5. Consider two partially ordered sets: *Set 1* and *Set 2*. The ranking of the objects named A, B, C and D is done in *Set 1* using predicted variables (like *e.g DetFreq* and *MedMax*) and the same objects are ranked in *Set 2* using predicting variables (like *e.g. SpArea*, *MedMax*). The box marked in top of Figure 5 shows the two Hasse Diagrams of respectively *Set 1* and *Set2* and the small table between them shows the actual values for the parameters defined in Table 3. The comparison of the two sets can be illustrated in matrix form as follows:

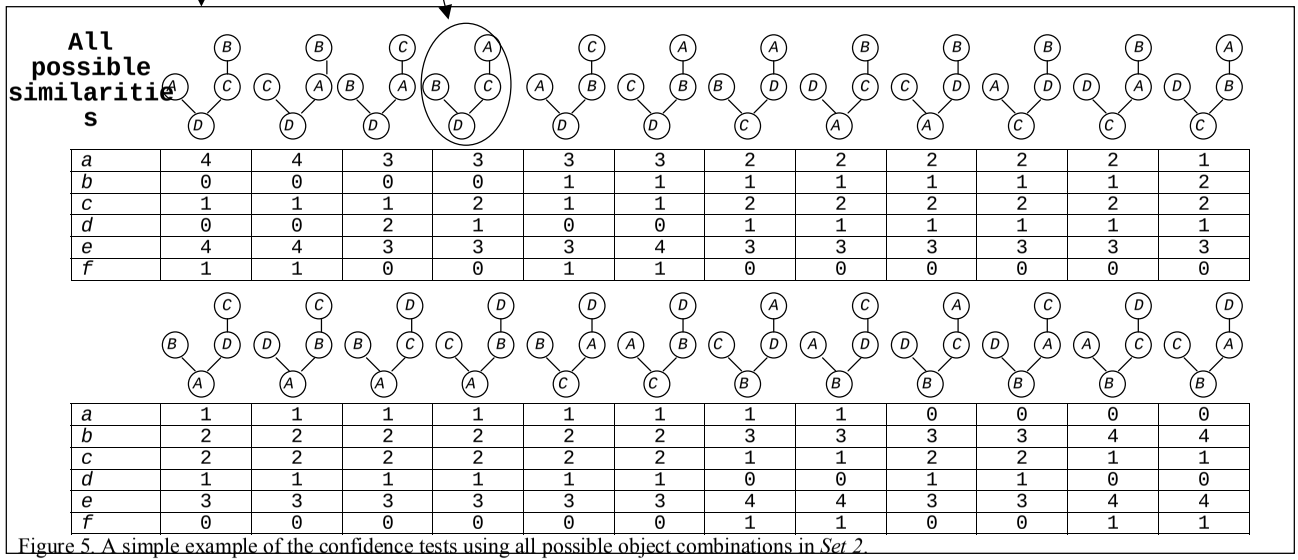
$$\begin{bmatrix} A & B & C & D \end{bmatrix}$$

$$C = \begin{pmatrix} - & <,\| & \|,> & >,> \\ & - & >,\| & >,> \\ & & - & >,> \\ & & & - \end{pmatrix} \qquad \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}$$

The entries of the matrix are to be read as follows: For example first row , second column: A < B in *Set 1* and A || B in *Set 2*. Three agreements (quantity *a* =3), no conflicts, i.e. no entry like >,<, or <,>; therefore *b* = 0; concordant rankings only in *Set 1* (entries $C_{1,2}$ and $C_{2,3}$), thus *c*=2, one concordant ranking only in *Set 2*: entry $C_{1,3}$. therefore *d*=1. Furthermore it is found from the Comparison matrix that *e* = 3 , *f* = 0. Equivalent objects are not included in this simple example. They will not serve any purpose for illustration and this excludes *g*, *h* and *i*.

The concept of testing is based on a fixed Hasse Diagram structure for *Set 2*, where positioning of the objects in this Hasse Diagram is considered as not given. This mean that all possible namings of objects are allowed in the Hasse Diagram and the procedure is to test all these combinations for similarity to *Set 1*. This yield 24 possible Hasse Diagrams in Figure 5, listed from high correlation towards lower correlation measured positive for a high value for *a* and a low value for *b*. In this way it is seen that *Set 2* belongs to the Hasse Diagram, which is among the four best similarity diagrams. The probability for a *a* value to be equal of larger than the actual *a* value of 3 is seen to be 6/24 and the probability for a *b* value to be equal or larger than the actual value of 0 is 20/24.





a	4	4	3	3	3	3	2	2	2	2	2	1
b	0	0	0	0	1	1	1	1	1	1	1	2
c	1	1	1	2	1	1	2	2	2	2	2	2
d	0	0	2	1	0	0	1	1	1	1	1	1
e	4	4	3	3	3	4	3	3	3	3	3	3
f	1	1	0	0	1	1	0	0	0	0	0	0

a	1	1	1	1	1	1	1	1	0	0	0	0
b	2	2	2	2	2	2	3	3	3	3	4	4
c	2	2	2	2	2	2	1	1	2	2	1	1
d	1	1	1	1	1	1	0	0	1	1	0	0
e	3	3	3	3	3	3	4	4	3	3	4	4
f	0	0	0	0	0	0	1	1	0	0	1	1

Figure 5. A simple example of the confidence tests using all possible object combinations in *Set 2*.

This simple example shows that there exists a variety of interrelations between the single comparison parameters. The four objects have six possible ranking relations between each other (generally for n as the number of objects: $n \cdot (n\text{-}1)/2$). Every relation can either be a concordant or a discordant ranking (Note: we have neglected equivalence). Thus the vertical sum of the comparison parameters (neglecting *e*, which just is the sum $a+b$) will always be 6. The total number of discordant rankings is 3 (one in *Set 1* and two in *Set 2*), so the relation $c+d+f=3$ needs to be valid due to the constancy of the Hasse Diagram structure. Similarly, the number of concordant rankings in *Set 2* is 4, so $a+b+d= 4$ and for *Set 1* there are 5 concordant rankings yielding $a+b+c= 5$. Another example of interaction is seen for *d* and *e*, where $d= 1$ for $e=4$ and $d=2$ for $e=3$. These simple interrelations become more complex when some objects are equivalent in *Set 1* and/or *Set 2*. However, there will still be a close interrelation between the single comparison parameters as seen in Table 3, where the probability estimates for respectively (*c*,*d*), (*e*,*f*) and (*g*,*h*) in pairs are equivalent.

The number of possible Hasse Diagram versions for testing of *Set 2* is $n!$ (4!=24 diagrams in Figure 5). So, it will not be advisable to try to test more than about 10 objects using the outlined method directly and an approximate method is to be applied in order to solve this problem. In this procedure, the full number of possible *Set 2* versions is estimated using a random sampled subset, see Figure 6. Every random sample is found by mixing up the object Ids in *Set 2*. More precisely first Id. 1 is selected and subsequently by random choice a selection is made of *e.g.* Id. 5 and exchange is made between respectively all the descriptors for Id. 1 and Id. 5. This procedure is repeated for all the other objects ending up in a Hasse Diagram like *Set 2*, but where the object naming is randomly distributed. The correlation between *Set* 1 and *Set 2* is first calculated yielding the "actual" correlation result (*AC*). Then subsequently a random Hasse Diagram is generated as explained above and a comparison with Set 1 generates a correlation estimate (RC). A sum (*sum* in Figure 6) sums up how many times $RC>AC$ is true out of totally *I* randomly formed Hasse Diagrams. The ratio *I*/*sum* is an estimate for the probability for a randomly formed correlation to exceed the actual correlation.

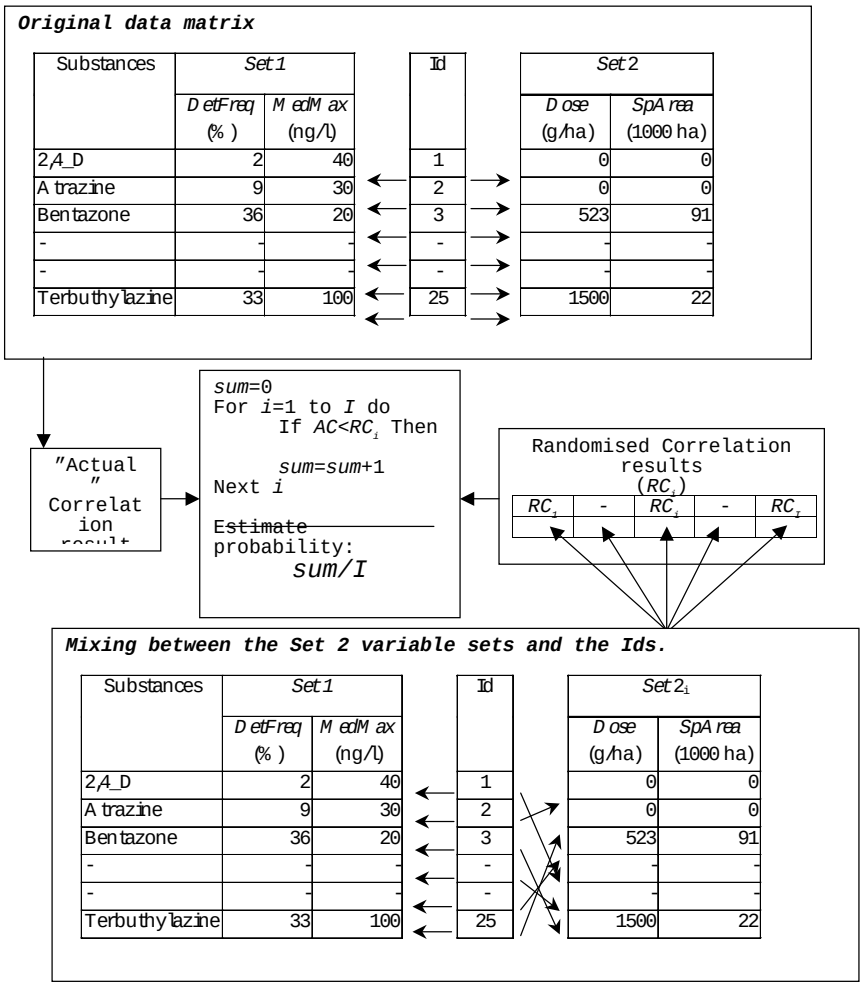


Figure 6. The principle of the probability estimate for a randomly formed comparison between the two sets to be of higher or equal value compared to the true comparison. A.randomised version of *Set 2* denoted *Set* 2_i is formed by mixing the Ids (names) for the rows in the data set of *Set 2*.

The value of *I* needs to be high enough to assure a robust probability estimate, however, the analysis has an upper limited for a meaningful increment of the *I* value around the factorial value for the number of objects (*n*!).For higher values of *I* no further information will be gained by further increasing the value of *I*. However, in case of 25 objects as in Table 3, the factorial value is about $1.6 \cdot 10^{25}$ and thus far above any realistic value for *I*. Different values of *I* are tested for the data set in Table 1 and the results is shown in Figure 7

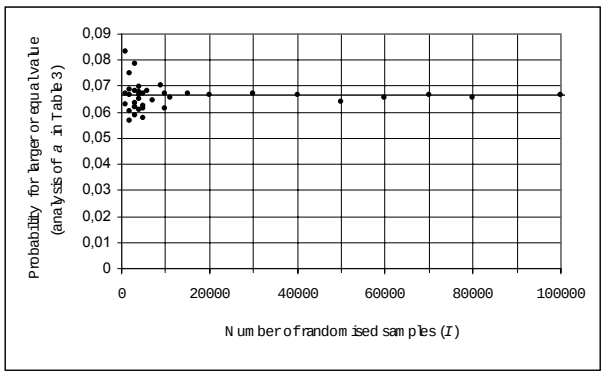


Figure 7. The principle shown in Figure 6 applied for different values of *I* and for testing *a*=82 from Table 3. The line shows an estimate of the "true" probability using $I=10^6$.

The probability estimate is graphically shown in *PO Correlation*, in form of a *Significance Plot*, see Figure 8. This figure shows two numbers for the correlation quality and every point is formed based on *Set* 2_i for i=1,2..,*I*. The number on the x-axis is *T*(0,0) as a measure for the quality of correlation. The quality number on the y-axis describes the total completeness of tested correlation. This quality number is calculated as the ratio between the number of rankings, which can be included as either an agreement or a conflict (*a* + *b* in Table 3), and the total number of possible ranking relations in the data set ($n \cdot (n\text{-}1)$)/2. The marked circle is the actual correlation estimate. In Figure 8 the actual estimate is located in the high end of the point cluster in the direction of *T*(0,0) (to the right from the middle), which indicates some degree of positive correlation. On the y-axis the actual estimate is located close to the lower edge

of the point cluster showing some misfit between the comparability in *Set* 1 and in *Set* 2. This was also seen in Table 3 as discussed above, where the random probability (equal or larger value) for the *e* value of 117 is quite high (0.980).

The banned pesticides (Ids. 1,2, 6, 7, 8) are now excluded from the data set and the correlation analysis is repeated. The results for the τ correlation are shown in Table 4. The product of *Dose* and *SpArea* could be an effective variable, having unit of used amount per year (kg/year). Sørensen et al. (2003) show, for the Swedish data, that this product is far from being complete, however, this statement will be investigated using the Danish data set also. The correlation in Table 4 has changed substantially compared with Table 2. The correlation between the predicted and predictive variables has improved. The product *Dose·SpArea* has the best correlation to both *DetFreq* and *MedMax*.

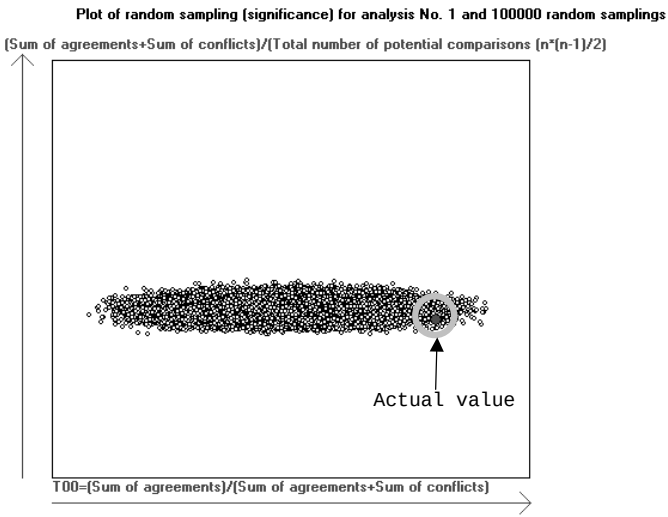


Figure 8. *Significance Plot* based on *Set* 2_i for i=1,2,..,I (I=100000) using the data set in Table 1. The red circle is equal to the actual value from Table 3.

Table 4. τ values for all parameter combinations using the data set in Table 1, where the banned pesticides, Ids. 1,2, 6, 7 and 8, are excluded.

		Predicted variables		Predicting variables		
		DetFreq	*MedMax*	Dose	SpArea	*Dose+SpArea*
Predicted variables	*DetFreq*	1.00	0.49	0.41	0.14	0.46
	MedMax	0.49	1.00	0.22	0.21	0.43
Predicting variables	*Dose*	0.41	0.22	1.00	-0.22	0.17
	SpArea	0.15	0.22	-0.22	1.00	0.61
	Dose+SpArea	0.46	0.43	0.17	0.61	1.00

The correlation between *Set 1* and *Set 2* is recalculated for the data set without the banned pesticides (Ids. 1, 2, 6, 7 and 8) and the results are displayed in Table 5.

The numbers in Table 5 are smaller compared to Table 3 due to a smaller number of pesticides. However, a much more confident positive correlation is seen having only 3 conflicting rankings and thus a $T(0,0)$ value of 0.94. The *Significance Plot* also shows an improved confidence compared to Figure 8, as displayed in Figure 9. A better separation between the actual correlation (marked circle) and the cluster of points is seen. It is also seen in Figure 9 that the actual correlation is placed in the centre of the cluster in the direction on the y-axis. This indicates absence of information within the sets, which is not reflected in the correlation between the two sets. The latter is also seen in Table 4, where the probability level for a higher e value has moved away from unity (0.980 in Table 3 down to 0.815 in Table 5).

Table 5. Correlation results between *Set* 1 and *Set* 2, where the banned pesticides having the Ids. 1,2, 6, 7 and 8 are withdrawn from the analysis leaving 19 pesticides in the correlation analysis.

Comparison n=19	Counting	Probability for larger or equal value
a	51	0.003
b	3	1.000
c	91	0.273
d	20	0.273
e	54	0.815
f	25	0.786
g	1	1.000
h	0	1.000
i	0	1.000
$T(0,0) = a/(a+b)$	0.94	0.000
$T(1,0) = a/(a+b+c)$	0.35	0.003
$T(0,1) = a/(a+b+d)$	0.69	0.003
$T(1,1) = a/(a+b+c+d)$	0.31	0.006

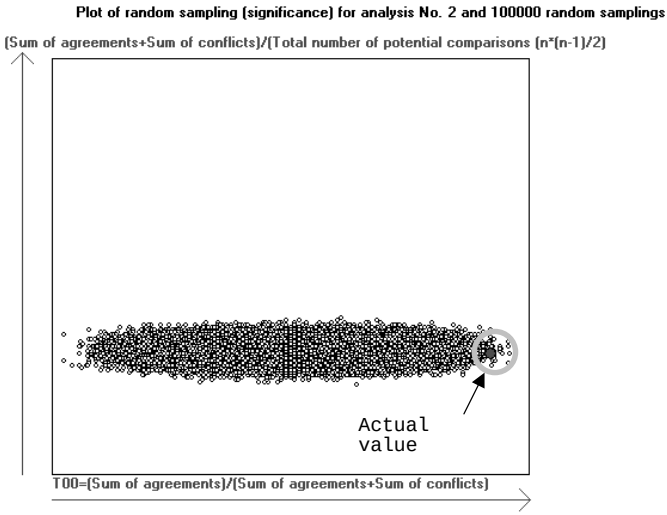


Figure 9. *Significance Plot* formed by random testing for the data set (I=100000), where the banned pesticides are neglected (Ids. 1, 2, 6, 7, 8).

The 3 conflicts (b = 3) in Table 5 need to be analysed using the *Conflict Diagram* before a final conclusion is possible, see Figure 10. The *Conflict Diagram* in Figure 10, which arises from that of Figure 4 by subtracting the banned pesticides from the set of chemicals, shows separated pairs of rankings, where no pesticide is connected to more than one single other pesticide. This indicates that there is no single responsible pesticide for all conflicts and thus no strong discrepancy for specific pesticides. The inclusion of other variables like physico-chemical parameters and the analytical detection limit is discussed by Sørensen et al., (2003) in relation to South Sweden monitoring data. However, any further addition of ranking variables will increase the number of discordant rankings in *Set* 2 and thus tends to damage the completeness of the correlation analysis. In this way there is a trade off between the number of variables to be included and the completeness of the correlation analysis. So, it seems irrelevant to consider any additional variables and the simple information about field application seems to be rather powerful for describing the occurrence of current used pesticides in streams. This is not a trivial conclusion because the main paradigm for the mechanism

of pesticides exposure is based on the hypothesis that basic physico-chemical properties including degradation and adsorption are governing factors for the quantification of differences in exposure between different pesticides.

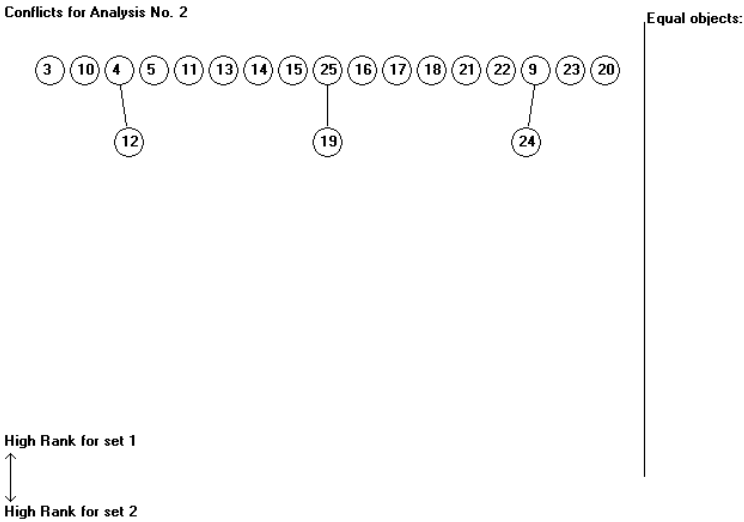


Figure 10. *Conflict Diagram* for the data set where the banned pesticides are neglected (Ids. 1,2, 6, 7, 8).

The correlation of the single predictive variables including their interaction will be investigated in the following. This is easy to do using *PO Correlation* by repeating the correlation analysis only for respectively *Dose* and *SpArea* as single variable in *Set* 2. Inspection of the *Conflict Diagram*s shows how the two variables are acting alone and in relation to each other. In Figure 11 the *Conflict Diagram* is displayed, where only the *Dose* is included as single variable in *Set* 2. The three pesticides (Ids. 5, 16 and 24) are ranked strongly downward having many concordant rankings. This tells us that the *Dose* variable tends to rank these pesticides upward while downward rankings are more likely to happen for the variables *DetFreq* and *MedMax*. The values in Table 1 also show that these pesticides are characterised by having a relatively high *Dose* value and a low *SpArea* value. Hence the occurrence seems limited (low value for *DetFreq* and

MedMax) due to rare application even though the dose level is high for the few events where application takes place in the field. The Id. 13 is concordant ranked above 6 other pesticides and also associated with a very high *SpArea* value, which is not reflected in a high *Dose* value.

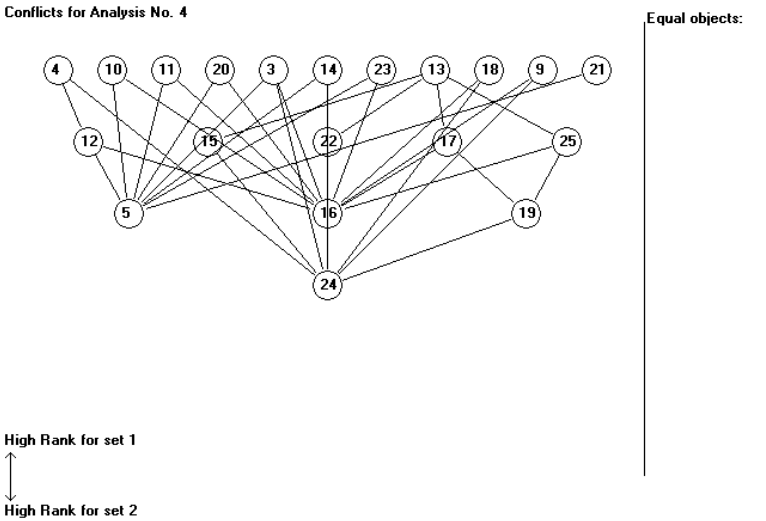


Figure 11. *Conflict Diagram* for the data set where the banned pesticides are neglected (1, 2, 6, 7, 8) and where the only predicting variable used in *Set 2* is *Dose*.

The *Conflict Diagram* using only the *SpArea* variable for *Set* 2 is shown in Figure 12. Two strongly top ranked pesticides are identified (Ids. 15 and 9) having many concordant rankings downward. They both have a high *Dose* value and small *SpArea* value in Table 1 and thus pesticides where a low rank due to *SpArea* is in conflict with the occurrence because of a high dose level. The complementary situation is also seen in Figure 12 for Id. 21. This pesticide is ranked strongly downward in Figure 12 and also a pesticide, which has low dose level and large sprayed area in according to Table 1. The

Id. 21 is a very low dosage pesticide and this low dose level seems to prevent the pesticide to occur in the stream water even though the sprayed area is rather large.

The common set of rankings for Figures 11 and 12 is displayed in Figure 10, so only three rankings are in common between the Figures 11 and 12. This shows that the two variables *Dose* and *SpArea* are working well together by describing different parts of the information captured by *DetFreq* and *MedMax*. This is supported by the negative correlation between *SpArea* and *DetFreq* in Table 4, which indicates that the two variables are not reproducing each other.

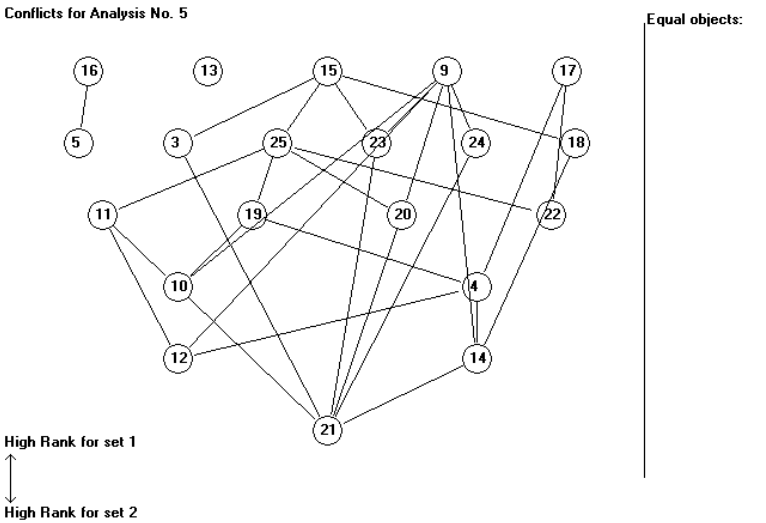


Figure 12. *Conflict Diagram* for the data set where the banned pesticides are excluded (Ids. 1, 2, 6, 7, 8) and where the only predicting variable used in *Set 2* is *SpArea***.**

In Figure 13 a graphical display is made for a series of correlation analyses using different combinations of variables. The x and y axis are similar as the axis in the *Significance Plot* (Figures 8 and 9). Six different correlation analyses are performed for

the pesticides, which have not been banned in 2000. The numbered circles show the correlation results for a series of different variable combinations used for *Set 1* and *Set 2* and the numbering refers to the columns in the Table to the right in Figure 13. The small circles in the Table to the right identify the included variables in every combination as numbered. So *e.g.* for analysis no. 1, *Set 1* consist of the two variables *DetFreq* and *MedMax* and *Set 2* consist of only *Dose*. The variable denoted PRODUCT is the product between *Dose* and *SpArea* as presented by Table 4. The analyses 1, 2 and 3 is all correlation analyses with *Set 1* (*DetFreq* and *MedMax*) using respectively *Dose* (analysis 1), *SpArea* (analysis 2) and PRODUCT (analysis 3) as single predictive variable in *Set 2*. The product is seen to be best as single predictive variable, because the circle numbered 3 has in comparison with numbers 1 and 2 the highest *T(0,0)* value. The next best correlation is seen for *Dose* (number 1 in the figure) and *SpArea* has the lowest correlation. It is possible to improve the correlation by using more than one predicting descriptor in *Set 2* as seen for number 4, where the descriptors *Dose* and *SpArea* are included. However, the completeness of the correlation decreases (downward direction in the figure) when two variables are included in *Set* 2 instead of only a single, because of the emergence of discordant rankings in *Set 2*. The use of respectively *DetFreq* and *MedMax* as single variable in *Set* 1 is also tested as respectively no. 5 and no. 6. Neither analysis no. 5 nor 6 can make the same good correlation measured by *T*(0,0) as analysis no. 4.

4 Conclusion

Application of partial ordering for rank correlation analysis including multiple variables has been shown effective by the software package *PO Correlation*. The design is shown transparent for correlation analysis, where the correlation of every object is characterised in detail in relation to all other objects. The principle and use of the software is shown and described using a specific data set based on environmental monitoring of pesticide findings in small streams in Denmark. It is shown how *PO Correlation* is valid for robust data interpretation.

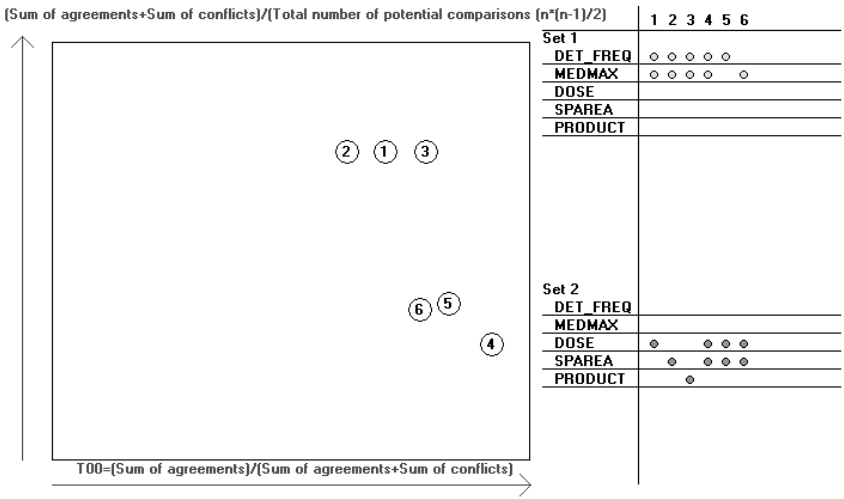


Figure 13. A graphical display of the correlation result for a series of different variable combinations. The numbered circles reefers to the numbering in the top row in the table to the right. Each number is a correlation analysis and the small circles in the tables indicate which variables that have been used in respectively *Set 1* and *Set 2*. The x and y axis is similar to the axis in the *Significance Plots* as explained for Figure 9.

A series of useful conclusions are shown to come out simply and clear. It is seen how the difference in the field application seems to be the main governing factor for pesticides exposure in stream water. The data analysis has shown how *PO Correlation* can make detailed data mining into single object relations using multiple variables. Using this novel methodology, it possible to identify specific interactions between the variables, which hardly can be reproduced by more conventional methods.

Reference

Bath PA, Morris CA, Willet P. (1993). Effects of Standardization on Fragment-Based Measures of Structural Similarity. J Chemom 7: 543-550.

Brüggemann R, Halfon E, Welzl G, Voigt K, Steinberg C (2001). Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests. J.Chem.Inf.Comp.Sci. 41:918-925

Brüggemann R, Bücherl C, Pudenz S, Steinberg CEW. (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. Acta Hydrochim Hydrobiol 23: 170-178.

Brüggemann R, Bartel HG, (1999) A theoretical concept to rank environmentally significant chemicals, J.Chem.Inf.Comp.Sci. 39 (2): 211-217

Brüggemann R, Zelles L, Bai QY, Hartmann A. (1995). A use of Hasse Diagram Technique for evaluation of phospholipid fatty acids distribution in selected soils. Chemosphere 30: 1209-1228.

Conover, W.J., (1999), Practical nonparametric statistics, John Wiley & Sons, ISBN 0-471-16068-7.

Gibbons, J.D., (1993), Nonparametric measures of association, SAGE University papers, series: quantitative applications in the social sciences, ISBN 0-8039-4664-3.

Goodmann, L. A. and Kruskal, W.H., (1963), Measures of association for cross-classifications, III: Approximate sample theory. Journal of American Statistical Association. 58: 310-364.

Kendall, M. G., (1938), A new measure of rank correlation, Biometrika. 30: 81-93.

Moock TE, Grier DL, Hounshell WD, Grethe G, Cronin K, Nourse JG, Theodosiou J. (1998). Similarity searching in the organic reaction domain. Tetrahedron Computer Methodology 1: 117-128.

Pavan, M., (2003), Total and partial ranking methods in chemical sciences, PhD. in Chemical Sciences, Cycle XVI, University of Milano – Bicocca.

Pudenz, (2004), Stefan Pudenz, Criterion - Evaluation and Information Management, Mariannenstr. 33, D-10999 Berlin, Germany (web: www.criteri-on.de).

Pudenz S, Brüggemann R, Komossa D, Kreimes K. (1998). An algebraic / graphical tool to compare ecosystems with respect to their pollution by Pb,Cd III: comparative regional analysis by applying a similarity index, Chemosphere. 36: 441-450.

Sørensen, P. B., Brüggemann R., Carlsen L., Mogensen, B. B., Kreuger J. and Pudenz, S., (2003), Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory - Data interpretation and model development, Environmental Toxicology and Chemistry, 22(3): 661-670.