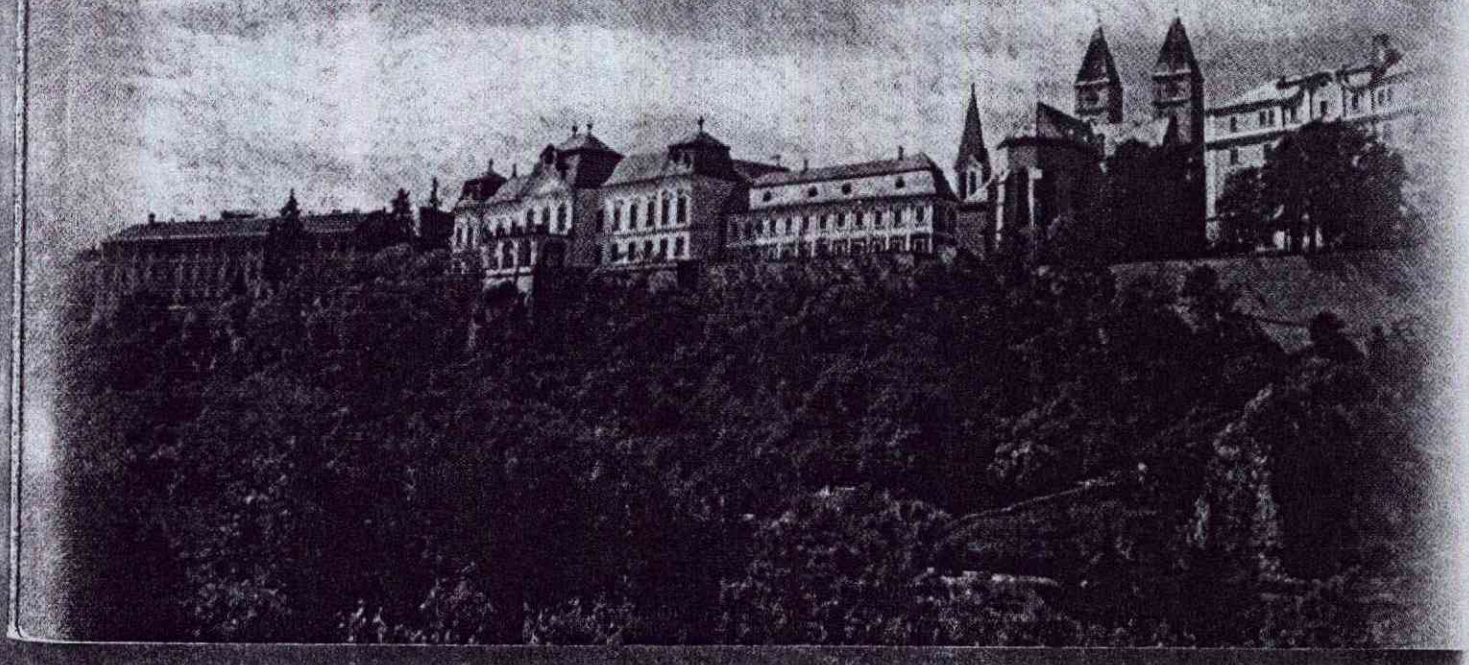
Second International Conference on ENVIRONMENTAL ENGINEERING

University of Veszprém, Hungary



Pair-Correlation Method with parametric and non-parametric test-statistics for variable selection. Description of computer program and application for environmental data case studies

K. Héberger[1] and R. Rajkó[2]

[1]Institute of Chemistry, Chemical Research Center,
Hungarian Academy of Sciences
H-1525 Budapest, POB 17, Hungary

[2]Department of Unit Operations and Environmental Engineering
Szeged University College of Food Industry
H-6701 Szeged, POB 433, Hungary

Abstract

Pair-Correlation Method (PCM) has been developed for choosing from among correlated descriptor variables provided the scatter is caused not only by random effects. This assumption is almost always valid for QSAR applications.
After initial heuristic use of this method [1-3], we have developed several test statistics. The following test statistics have been investigated and compared to *Fisher's Conditional Exact Test, McNemar's test, Chi-square test* and *Williams' t-test* using the analogy of the PCM table to a 2x2 contingency table.
We have constructed a macro based on the *MS Excel 8.0 Visual Basic for Applications (VBA)*, which yielded a user-friendly and easy-to-use program because of the spreadsheet properties.
In this paper we show the use of PCM in detail, as well as some case studies on the selection among several topological indexes for description of cAMP phospodiesterase inhibition by flavons, chlorobenzenes toxicity and mutagene characters.

1 Description of Pair Correlation Method

The PCM is a non-parametric method which can distinguish between statistically equivalent or seemingly equivalent variables. If the classic parametric methods are not able to find any difference then the non-parametric methods are expected to be able to make discrimination between variables.

1.1 Test statistics for application of PCM

Let us consider three vectors:

$$\mathbf{y}(i), \mathbf{X1}(i), \mathbf{X2}(i) \qquad i = 1, 2, \ldots m$$

where $\mathbf{y}(i)$ is the response (dependent variable) and $\mathbf{X1}(i)$, $\mathbf{X2}(i)$ are descriptors (factors, independent variables).
Further on, it is assumed that $\mathbf{y}(i)$, $\mathbf{X1}(i)$ as well as $\mathbf{y}(i)$, $\mathbf{X2}(i)$ are positively correlated. A negative correlation cannot cause serious limitation. Namely, if one of them (or both) is (are) negatively correlated, the converse of it (them) has to be used: i.e. a multiplication by (-1).
Our aim is to discriminate between **X1** and **X2** descriptors.
1) Let a pair of data points be selected and ordered:

$$\mathbf{y}(i) > \mathbf{y}(j) \text{ i.e. } \mathbf{y}(i) / \mathbf{y}(j) > 1.0 \qquad i,j = 1, 2, \ldots m$$

and $\mathbf{y}(i) \neq \mathbf{y}(j)$

2) Let us examine the differences: $\Delta X1 = X1(i) - X1(j)$ and $\Delta X2 = X2(i) - X2(j)$ belonging to y(i) and y(j). Assuming that every $\Delta X1$ and $\Delta X2 \neq 0$ (i.e., no repeated measurements), only the four possibilities (boxes) shown in Table 1 exist.
3) Let us consider all possible pair of data points {n = m*(m-1)/2} and count the cases A, B, C, D. The events and frequencies are summarized in Table 1

Table 1 Ordering the possibilities (events A, B, C and D) in a table.
Frequencies are: k_A, k_B, k_C, and k_D

	ΔX1>0	ΔX1<0
ΔX2>0	A : k_A	C : k_C
ΔX2<0	B : k_B	D : k_D

The analogy of the table above to a 2×2 contingency table is apparent.
If there is significant difference between descriptors **X1** and **X2** then the frequencies of B and C events should also be significantly different. The task is to prove the significance with well chosen test statistics.

A generalization for negative correlations, should include a rearrangement of boxes, as well. Ordering the frequencies in boxes cannot necessarily be made equivocally. Several (five) rearrangements of boxes can be carried out where the contingency tables will be equivalent [5].

Two principles can govern the rearrangement. The value of k_A should be the largest, i.e. the directions of **y** vs. **X1** and **y** vs. **X2** associations should be the same whenever possible. The values of k_D should be the smallest, i.e. worsening of the direction stochastically should be the smallest.

If k_B is significantly larger than k_C then **X1** is superior over **X2** and vice versa.
It is assumed for application of the test statistics that **X1** and **X2** do not differ significantly, i. e. our null-hypotheses is:

$$H_0: k_B = k_C$$

The alternative hypotheses is:

$$H_A: k_B \neq k_C$$

The contingency table (2 × 2) is introduced for application of test statistics [5]
The following test statistics have been investigated and compared McNemar's test, Chi-square test, Fisher's Conditional Exact Test, and Williams' t-test [4,5,6]. The first three tests are non-parametric, whereas the last one is a parametric test.
This study is devoted to present the computer program for easy-to use application of PCM. Later the generalization of the method will be shown for more than two descriptor variables. Environmental examples will suggest the large capacity and usefulness of the method.

1.2 Presentation of the computer program: PCM.xla

The algorithm of PCM was formulated in a macro language: Visual Basic for Applications (VBA) of MS Excel. Excel offers easy data handling, easy data transfer and compatibility possibilities.

The PCM computer program can be opened as a table saved earlier in the Excel program. The name of the file (PCM.xla) should be chosen in the menu "File, Open".
Clicking the OK button of the input window, the window disappears and a new menu point appears showing the PCM. There are two sub-menu points "Start" and "Quit".
The PCA does not need any pre-treatment of data. However, one row above the numbers should be empty or preferably it should contain the names of variables (e.g. Y, X1, X2, ..., etc.). PCM will generate labels for variables automatically in case of an empty row. PCM can be started after highlighting the data. It starts with a general setup window, shown below:

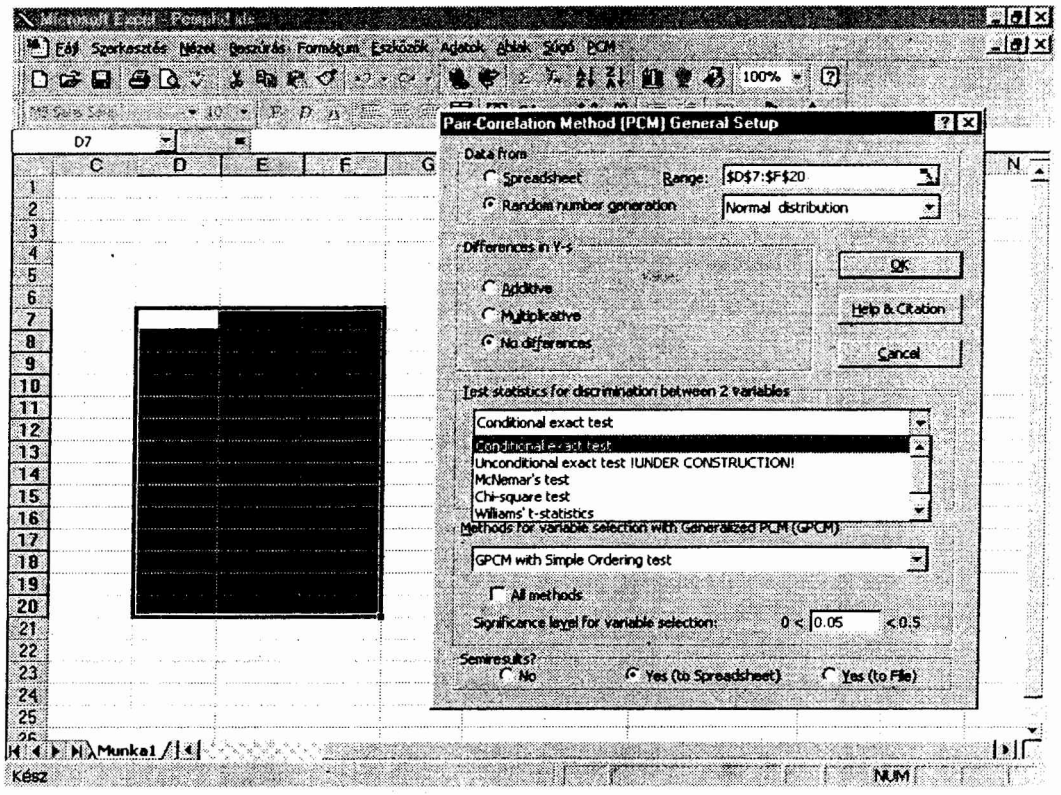


Figure 1 General setup window

Data can be used from spreadsheet or can be generated by a random number generator for advanced users. At present, vectors with multivariate normal distribution are generated but uniform and Cauchy distribution is also planned to be included.

The program offers the possibility of taking into account the experimental error. It is of crucial importance for the method that the values of dependent variable are different. If the values of dependent variable are equal within the experimental error the differentiation cannot be completed. Therefore, it is possible to give the error level. Two kind of error can be chosen additive and multiplicative (proportional). The error level (Differences in **y**-s) can be given numerically. Accepting the settings can be done by the OK button. The help file is not yet ready but the following information is available by clicking at the help button:

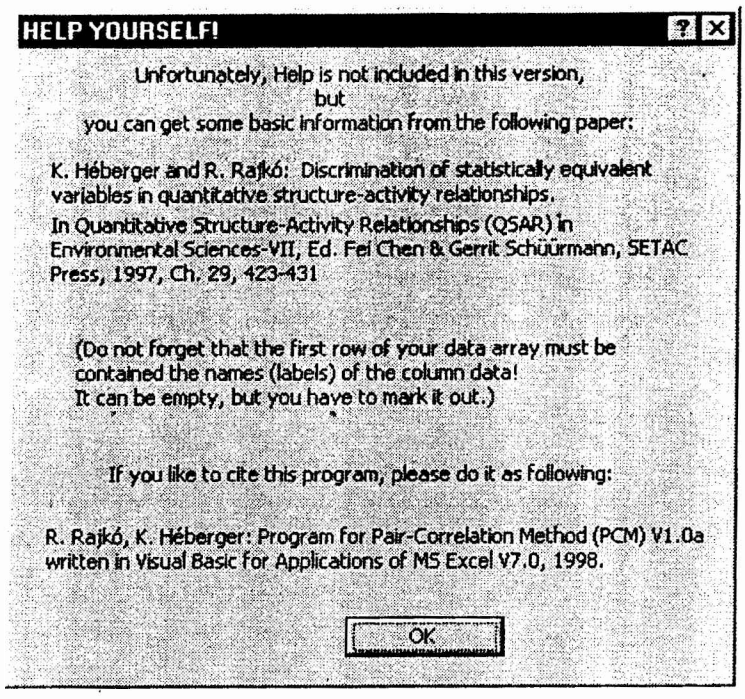


Figure 2 Information for citation of the method.

The selection criteria are enumerated in the scroll bar in the middle entitled "Test statistics for discrimination between 2 variables". The required significance level can be set similarly. There is a possibility to apply all the selection criteria at once by activating the "all test statistics" bar. In this case the program calculates the results using all the statistical tests.

The next scroll bar is devoted to the ordering of variables, another words to the generalization for more variables. Three ordering methods were elaborated: 1) simple ordering according to the numbers of "superiority" (wins); 2) ordering according to the difference between numbers of superiority and inferiority (wins minus losses); 3) as above (point 2) but weighted by probability.

The last question group entitled "Semi-results?" sets the way to giving intermediate results. There are three possibilities: "No" – the intermediate results will be lost, "Yes (to Spreadsheet)", "Yes (to File)" are self explanatory.

2 Case studies

2.1 Flavone derivatives

Flavonoids (compound with structure based on that of flavone (2-phenyl-chromone) are wide-spread compounds in the plant word. Many flavonoids exhibit pharmacological activity. They take part in co-pigmentation, in protection the plants against viral infection in inhibition or activation of different enzyme system.
Trinajstic et al. published a detailed QSAR study with generation of total 34 descriptors [7]. They selected nine favorable descriptors as follows [7]:

khi0: zero-order valence connectivity index
khi2: second -order valence connectivity index

kh3: third-order valence connectivity index
p3: number of the paths of length 3
p10: number of the paths of length 10
LUMO: energy of lowest unoccupied molecular orbital
TRE: topological resonance energy per electron
Schr: sum of the π charges in the chromone moiety
Sph: sum of the π charges in the phenyl moiety

All descriptors are non-orthogonal. Our aim was to investigate whether it is possible to find significant differences in case of the above variables selected by multiple linear regression.

The results can be seen on the next Excel table:

	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	2	0	3	0
Loser	0	0	2	3	2	2	4	0	4
No Decision	2	2	6	5	6	4	4	5	4
Rank by	1	2	5	7	6	4	8	3	9
Wins	α (user)	0.05	α (emp.)	0	Crit. Sum	16.15	17		
CondExact		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	2	0	0	0
Loser	0	0	2	2	2	2	3	0	3
No Decision	2	2	6	6	6	4	5	8	5
Rank by	1	2	6	7	5	3	8	4	9
Wins	α (user)	0.05	α (emp.)	0	Crit. Sum	13.3	14		
McNemars		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	0	0	0	0
Loser	0	0	2	2	2	2	2	0	2
No Decision	2	2	6	6	6	6	6	8	6
Rank by	1	2	8	4	5	6	7	3	9
Wins	α (user)	0.05	α (emp.)	0	Crit. Sum	11.4	12		
ChiSquare		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	0	5	0	0	0	0	0	0	0
Loser	1	0	1	1	0	0	1	1	0
No Decision	7	3	7	7	8	8	7	7	8
Rank by	5	1	6	9	2	3	7	8	4
Wins	α (user)	0.05	α (emp.)	0	Crit. Sum	4.75	5		
Williams' t		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	2	0	3	0
Loser	0	0	2	3	2	2	4	0	4
No Decision	2	2	6	5	6	4	4	5	4
Rank by	1	2	5	7	6	4	8	3	9
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	14.25	15		
CondExact		No Differences in y							

	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	2	0	0	0
Loser	0	0	2	2	2	2	3	0	3
No Decision	2	2	6	6	6	4	5	8	5
Rank by	1	2	6	7	5	3	8	4	9
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	11.4	12		
McNemars		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	6	6	0	0	0	0	0	0	0
Loser	0	0	2	2	2	2	2	0	2
No Decision	2	2	6	6	6	6	6	8	6
Rank by	1	2	8	4	5	6	7	3	9
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	11.4	12		
ChiSquare		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
Winner	0	5	0	0	0	0	0	0	0
Loser	1	0	1	1	0	0	1	1	0
No Decision	7	3	7	7	8	8	7	7	8
Rank by	5	1	6	9	2	3	7	8	4
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	4.75	5		
Williams' t		No Differences in y							

	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
pWinner	5.9938173	5.99693	0	0	0	1.974839	0	2.91667	0
pLoser	0	0	1.999233	2.97	2	1.999999	3.95674	0	3.959
No Decision	2	2	6	5	6	4	4	5	4
Rank by	2	1	5	7	6	4	8	3	9
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	14.25	15		
CondExact		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
pWinner	5.9472786	5.974487	0	0	0	1.95074	0	0	0
pLoser	0	0	1.991456	2	2	1.999823	2.94205	0	2.941
No Decision	2	2	6	6	6	4	5	8	5
6	2	1		7	5	3	8	4	9
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	11.4	12		
McNemars		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
pWinner	5.9893621	5.995756	0	0	0	0	0	0	0
pLoser	0	0	1.999	2	2	1.999999	1.99391	0	1.993
No Decision	2	2	6	6	6	6	6	8	6
Rank by	2	1	8	4	5	6	7	3	9
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	11.4	12		
ChiSquare		No Differences in y							
	khi0	khi2	kh3	p3	p10	LUMO	TRE	Schr	Sph
pWinner	0	4.955751	0	0	0	0	0	0	0
pLoser	0.9993909	0	0.997548	1	0	0	0.97355	0.98544	0
No Decision	7	3	7	7	8	8	7	7	8
Rank by	5	1	6	9	2	3	7	8	4

pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	4.75	5		
Williams' t		No Differences in y							

As it can be seen from the table it is relatively easy to differentiate among the "best" variables by PCM.
The various selection criteria provides different results, as expected. The most conservative is the parametric Williams test which select only one descriptor: khi0. The least conservative one is the conditional exact test using simple ordering: khi0, kh2, Schr, and eventually LUMO are selected.

Considering the ordering methods the following conclusions can be drawn:
Simple ordering (SO) is the least conservative i.e. selects the largest number of descriptors. The ordering according to the difference between number of superiority and inferiority (wins minus losses: WL) is the most conservative, i.e. selects the least number of descriptors, whereas the probability weighted ordering is somewhere in between.
Otherwise the selections are stable. The selected variables are subsets of each other.

2.2 Toxicity of Chlorobenzenes

Todeschini et al. proposed new type of descriptors for QSAR studies [8]. The so called WHIM descriptors have several advantages e.g. they are direction and rotation independent. They can be calculated easily and automatically. The large number of WHIM descriptors, however, calls for an effective variable selection method. All of the different selection criteria of PCM was tested using the 40 WHIM descriptors in describing toxicity of chlorobenzenes as measured by algae.

The PCM works well with large number of descriptors.
Although the least conservative tests and ordering methods select more variables than the least conservative ones, the basic selection is always the same as it can be seen from the table:

SO, Wt	Ve	Vs	As	Tp	L2u	Vm	Tm	Am	L2e	Ds	Dv	L2s	E2s	L1u	L2m	E2m																						SO, Wt
WL, Wt	Ve	Vs	As	Tp	L2u	Vm	Tm	Am	L2e	Ds	Dv	L2s	E2s																										WL, Wt
pWpL, Wt	Ve	Vs	As	Tp	L2u	Vm	Tm	Am	L2e	Ds	Dv	L2s	E2s																										pWpL, Wt
SO, MN	Tm	Tp	As	Vm	Vs	Dv	L2e	Am	Ve	L2u	L1u	Ds	L2s	E2s	L2m	E2m	P1s	L1e	Dm	Km	P1p	Ks	P1e	E1m															SO, MN
WL, MN	Tm	Tp	As	Vm	Vs	Dv	L2e	Am	Ve	L2u	L1u	Ds	L2s	E2s	L2m																								WL, MN
pWpL, MN	Tm	Tp	As	Vm	Vs	Dv	L2e	Am	Ve	L2u	L1u	Ds	L2s	E2s	L2m	E2m																							pWpL, MN
SO, CE	Tm	Tp	As	Vm	Vs	Dv	Am	Ve	L2u	L2e	Ds	L1u	L2m	L2s	E2s	E2m	Km	P1s	Ks	L1e	P1p	P1e	E1m	Dm															SO, CE
WL, CE	Tm	Tp	As	Vm	Vs	Dv	Am	Ve	L2u	L2e	Ds	L1u	L2m	L2s	E2s																								WL, CE
pWpL, CE	Tm	Tp	As	Vm	Vs	Dv	Am	Ve	L2e	L2u	L1u	Ds	L2s	E2s	L2m																								pWpL, CE
SO, $\chi 2$	Tm	Tp	Dv	Am	As	L2e	Vm	Ve	Vs	L2s	L2u	L1u	Ds	E2s	L2m	P1s	Km	E2m	L1e	Dm	P1p	P1e	Ks	E1m															SO, $\chi 2$
WL, $\chi 2$	Tm	Tp	Dv	Am	As	L2e	Vm	Ve	Vs	L2s	L2u	L1u	Ds	E2s	L2m	Km	E2m	P1s																					WL, $\chi 2$
pWpL, $\chi 2$	Tm	Tp	Dv	As	Vm	Vs	L2e	Am	Ve	L2u	L1u	L2s	Ds	E2s	L2m	E2m	Km	P1s																					pWpL, $\chi 2$

Notations Wt: Williams test, MN: McNemar test, CE: conditional exact test, χ^2: Chi square test.

2.3 Mutagenecity of aromatic and heteroaromatic amines

Basak et al. collected data for 95 amines from the literature [9]. The mutagenic activities of these compounds were expressed in S. Typhimurium TA98+S9 microsomal preparation as mutation rate in natural logarithm (revertants/nanomole). Hundred and two topological indices were calculated [9]. Topological indices were partitioned into topostructural and topochemical indices. Moreover, quantum chemical parameters were also calculated. The best model developed uses nine descriptors from the three groups. The best descriptors can further be differentiated by PCM.

Simple ordering leaves only one variable out. SIC4 structural information content for 4th order neighborhood of vertices in a hydrogen filled graph seems to be the worse variable related to mutagenecity. Inclusion of two further variables into the model is questionable: The energy of HOMO and dipole moment. On the other hand, Number of paths of length p0, Balaban's index based on distance, j, and heat of formation is always selected as the best variables even with ordering according to difference WL.

A Summary is given below:

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7
Wins	α (user)	0.05	α (emp.)	0.03	Crit. Sum	30.4	31		
CondExact		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7
Wins	α (user)	0.05	α (emp.)	0.03	Crit. Sum	30.4	31		
McNemars		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7
Wins	α (user)	0.05	α (emp.)	0.03	Crit. Sum	30.4	31		
ChiSquare		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	4	4	4	4	0	0	0	4	0
Loser	0	0	0	0	5	5	5	0	5
No Decision	4	4	4	4	3	3	3	4	3
Rank by	1	2	3	4	8	6	7	5	9
Wins	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
Williams' t		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
CondExact		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
McNemars		No Differences in ELNR							

	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	5	7	7	4	0	1	2	5	1
Loser	2	0	0	4	8	6	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	3	1	2	5	9	8	6	4	7

Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
ChiSquare		No Differences in ELNR							
	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
Winner	4	4	4	4	0	0	0	4	0
Loser	0	0	0	0	5	5	5	0	5
No Decision	4	4	4	4	3	3	3	4	3
Rank by	1	2	3	4	8	6	7	5	9
Win-Los	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
Williams' t		No Differences in ELNR							
.	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
pWinner	5	6.999729	7	4	0	1	1.991743	5	1
pLoser	1.99999962	0	0	4	8	5.991743	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	4	2	1	5	9	8	6	3	7
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
CondExact		No Differences in ELNR							
	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
pWinner	4.9999997	6.998279	6.99998	4	0	1	1.951585	5	1
pLoser	1.99997772	0	0	4	8	5.951585	5	1.998	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	4	2	1	5	9	8	6	3	7
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
McNemars		No Differences in ELNR							
	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
pWinner	5	6.999525	7	4	0	1	1.98491	5	1
pLoser	1.99999927	0	0	4	8	5.98491	5	2	5
No Decision	1	1	1	0	0	1	1	1	2
Rank by	4	2	1	5	9	8	6	3	7
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
ChiSquare		No Differences in ELNR							
	CHI4PC	p0	j	SIC2	SIC4	CHI5Cb	EHOMO1	Hf	mue
pWinner	3.99870165	3.999849	3.99978	3.98	0	0	0	3.995	0
pLoser	0	0	0	0	4.995444	4.990986	4.990893	0	4.999
No Decision	4	4	4	4	3	3	3	4	3
Rank by	3	1	2	5	8	6	7	4	9
pWin-pLos	α (user)	0.05	α (emp.)	0	Crit. Sum	19	20		
Williams' t		No Differences in ELNR							

Conclusions

Prediction of biological activity related to the environmental sciences can be carried out after proper variable selection.

PCM equipped with suitable selection criteria and ordering methods of variables is a suitable tool for this purpose.

Comparison of selection criteria suggests reliable flexibility inherent of the method as a whole. Similar statement can be formulated for the ordering methods.

Acknowledgement

This work has been supported by OTKA T-29748 and F-025287.

Literature

[1] K. Héberger and H. Fisher: Rate Constants for the Addition of the 2-Hydroxy-2-Propyl Radical to Alkenes in Solution, *International Journal of Chemical Kinetics*, **25,** 913-920 (1993)

[2] K. Héberger and H. Fisher: Rate Constants for the Addition of the 2-Cyano-2-Propyl Radical to Alkenes in Solution, *International Journal of Chemical Kinetics*, **25,** 249-263 (1993)

[3] K. Hébeger and R. Rajkó: Discrimination of Statistically Equivalent Variables in Quantitative Structure - Activity Relationships, Quantitative Structure - Activity Relationships in Environmental Sciences VII. Eds. Fei Chen and G. Schüürmann SETAC Special Publication Series *Proceedings of QSAR'96, 24-28 June 1996* , SETAC Press, Chapter 29, pp. 425-433 (1997)

[4] W.J. Conover: *Practical Nonparametric Statistics* (2^{nd} Ed.) John Wiley & Sons, New York, Chapter 3.5 pp. 130-133 (1980)

[5] R. Rajkó and K. Héberger: Selection criteria based on statistical hypotheses testing for Pair-Correlation Method (PCM) to discriminate between variables, *in preparation.*

[6] E.J. Williams: *Regression Analysis*, John Wiley & Sons, New York (1959)

[7] D. Amic, D. Davidovic-Amic, A. Juric, B Lucic, and N. Trinajstic: Structure – Activity Correlation of Flavone Derivatives for Inhibition of cAMP Phosphodiesterase., *Journal of Chemical Information and Computer Sciences.*, **35**, 1034-1038 (1995)

[8] R. Todeschini, C. Bettiol, G. Giurin, P. Grammatica, P. Miana, and F. Argese: Modeling and Prediction by using WHIM descriptors in WSAR studies, submitochondrial particles (SMP) as toxicity biosensors, *Chemosphere*, **33**, 71-79 (1996)

[9] S. C. Basak, B. D. Gute, and G. D. Grunwald: Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenecity of Chemicals, Quantitative Structure - Activity Relationships in Environmental Sciences VII. Eds. Fei Chen and G. Schüürmann SETAC Special Publication Series *Proceedings of QSAR'96, 24-28 June 1996* , SETAC Press, Chapter 17, pp. 245-261 (1997)