Scientific Paper

Retro-Regression – A Way to Resolve Multivariate Regression Ambiguities

Milan Randić[a,]* and Matevž Pompe[b]

[a] National Institute of Chemistry, Ljubljana, Slovenia & 3225 Kingman Rd. Ames, IA 50014, USA.
E-mail: mrandic@msn.com
[b] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia

Received 15-03-2005

Abstract

Stepwise multivariate regression analysis (MRA) is one of the oldest data reduction techniques. When optimal descriptors are selected at each step in a stepwise regression in MRA descriptors that have appeared in earlier steps may disappear and novel combinations of new descriptors may arise. This introduces difficulties in interpretation of the regression equations, because it is not possible to construct orthogonalized descriptors using stepwise regression. We outline a procedure to resolve the difficulties arising in selection of optimal descriptors in multivariate regression analysis which allows construction of orthogonal descriptors to accompany the regression obtained in the final step.

Key words: retro-regression, QSAR, boiling points

Introduction

Modeling structure-property relationship has remained one of central topics in the study of variations of molecular properties with changes in the size and the shape of molecular structure. Most physico-chemical molecular properties are expressed by single number but molecular structure cannot be simply numericaly characterized. One way out of this dilemma of how to relate structrure to its properties is to describe structure by a set of structural invariants, which also mostly are represented by numbers. In this way we are in position to relate one set of numbers (e. g., selected physico-chemical properties) with another set of numbers (invariants which represent mathematical properties of chemical structure). Hence, thus structure-property relationship is transformed into a property-property relationship in which set of physico-chemical properties is related to mathematical properties of the same structures. The simplest such relationship is simple regression analysis in which single molecular property is related to single mathematical property. If the resulting simple regression is of relatively high quality one may say that the particular mathematical property that was used as molecular descriptor *parallels* to a great extent physico-chemical property considered. For example, the normal boiling points of octanes give good simple regression when molecular carbon skeletons of octane isomers are characterized by the connectivity index χ,[1] which is bond additive quantity and gives to different CC bonds different relative weight, depending on the character of carbon actoms involved in CC bond. For instance, CC bonds between primary and secondary carbon atoms are given more weight than CC bonds between primary and tertiary carbon atoms or CC bonds between two secondary carbon atoms.

The difficulties arise from the fact that there are very few simple regressions that offer satisfactory correlations, and that therefore one has to resort to multivariate regression analysis. These difficulties are two-fold: (1) Regression analysis using more than single descriptors are burdened with well known unstabilities of the regression equations, the coefficient of which can change dramatically and unpredictably when additional molecular descriptors are used; and (2) Interpretation of the resulting regression model is at best ambiguous, often meaningless – which does not help in a refinement of molecular models. One of conceptually simple approach is known as the stepwise regression analysis in which one considers a pool of N descriptors and selects the best as the first molecular descriptor for the considered structure-property relationship. One then continues keeping this descriptor and considers which of the remaining $N-1$ descriptors will lead to smallest standard error for the regression considered. This process, known also as "greedy" algorithm, continues till one find the set of best k descriptors – but it is also known that the set of descriptors so arrived need not be optimal. When one allows possibility to consider different descriptors at different steps in the regression then there are two kinds of difficulties in MRA applications. With each step in a regression different set of descriptors may emerge as the best choice. However, even if one keeps all descriptors found in previous

steps, their contribution to the regression, reflected in the magnitudes of the corresponding coefficients, may change dramatically from one step to another. While the quality of regression and its predictive power is not affected necessarily by the above chaotic behavior of the regression equations, interpretation of the regressions equations become impossible. Modeling of the structure - property - activity relationship in such situation may become meaningless.

The problem of choosing molecular descriptors for MRA remains a critical step in most structure-property-activity applications. A success or a failure of a regression analysis may depend critically on use of suitable molecular descriptors. With hundreds of descriptors being available often one resort to statistical methods for selecting best descriptors, hoping that in such an approach one would not miss potentially useful descriptors. There are pitfalls even in such approaches, because typically descriptors that show low correlations are discarded, yet when such descriptors are combined they may lead to a satisfactory result. This has been illustrated for molar refraction and molar volumes of octane isomers when simple connectivity indices $^1\chi$ and $^2\chi$ show no significant correlation when considered individually, but combined give a regression with the correlation coefficient over 0.97.[2] Thus if one is to discard the connectivity indices $^1\chi$ and $^2\chi$ as unsuitable single-variable descriptors one would also discard any of their linear combinations, including also the orthogonalized descriptor $^2\chi*$, which Xu has shown to lead to very satisfactory single variable descriptor for molar refraction.[3]

Another ambiguity of MRA is that sometimes different sets of descriptors offer correlations of a similar quality. It is very difficult to find any reasonable physico-chemical explanation why different sets of descriptors may emerge as the best choice in various applications of MRA. A mathematical explanation is clear: the best set having $n+1$ descriptor will outperform all sets having n descriptors, several of which may be comparable as judged by pertinent statistical parameters. In exhaustive searches for the best n descriptors often we find combinations of descriptors that have not been used in previous steps. Addition of new descriptors to the best set of n descriptors restricts combinatorial selection to N-n, where N is the cardinality of the pool of descriptors. Therefore not only that it is possible, but it is likely that combinatorial search of $n+1$ descriptors from a pool of N descriptors will result in descriptors that have not been adopted in previous steps. Recent work of Trinajstić and Lučić gives ample illustrations of such situation.[4,5]

Is there any kind of reasonable physico-chemical explanation for the mentioned fluctuation of optimally selected descriptors in stepwise MRA? The purpose of this article is to outline a procedure that will resolve the ambiguities arising in the selection of optimal descriptors in MRA during the stepwise selection of descriptors.

On ill-behavior or regression equations in stepwise MRA

The instability of coefficient of regression equations toward inclusion/exclusion of new descriptors is well known. In Table 1 we illustrate the case on the boiling points of 100 alcohols considered in a study by Kier and Hall.[6] Using CODESSA, a software developed by Katritzky, Lobanov and Karelson,[7] we obtained the regression equations shown in Table 1. The degree of freedom for the evaluation of standard error of estimates was calculated by substracting the number of parameters of the model from the total number of structures present in the data set. The regression models were obtained by step-wise selection procedure from 56 calculated topological and structural descriptors. The molecular desciptors were calculated from the HyperChem™ file format. We can see that at each successive step the coefficients already established for various descriptors change drastically. It is not possible therefore to attribute relative weight to different descriptors when interpreting the result, because their magnitudes depend on the presence of other descriptors. In other words, all descriptors are mutually interrelated, which is the source of a pronounced instability of the regression equations. If we could use descriptors that are not inter-related this problem would disappear. This is precisely what can be accomplished by orthogonalization of molecular descriptors.

Consider a stepwise regression (equations 1–3):

$$\text{PROPERTY} = c_{11}\, d_1 + \text{const.}_1 \quad (1)$$

$$\text{PROPERTY} = c_{12}\, d_1 + c_{22}\, d_2 + \text{const.}_2 \quad (2)$$

$$\text{PROPERTY} = c_{13}\, d_1 + c_{23}\, d_2 + c_{33}\, d_3 + \text{const.}_3 \quad (3)$$

where c_{mn} are the coefficients of m-th descriptor d_m in the n-th step of the stepwise regression. The chaotic behavior of regression coefficients is reflected in the fact that $c_{11} \neq c_{12} \neq c_{13} \neq \ldots$; $c_{22} \neq c_{23} \neq \ldots$; etc. If instead of descriptors d_1, d_2, d_3... we use orthogonalized descriptors d_1^*, d_2^*, d_3^*... the above set of stepwise regression equation become:

$$\text{PROPERTY} = c_{11}\, d_1^* + \text{const.} \quad (4)$$

$$\text{PROPERTY} = c_{11}\, d_1^* + c_{22}\, d_2^* + \text{const.} \quad (5)$$

$$\text{PROPERTY} = c_{11}\, d_1^* + c_{22}\, d_2^* + c_{33}\, d_3^* + \text{const.} \quad (6)$$

in which none the coefficients associated with each descriptor change from one step to the another.

Table 1. The stepwise regression equations for the boiling points of alcohols (n = 100) obtained by CODESSA using the greedy algorithm (that is keeping at each step all the descriptors selected in the previous steps).

	Descriptor	Coefficients	Standard error	r	s (°C)	F
0	Constant	22.3	3.9	0.9246	8.48	1201
1	Randic index (order 1)	35.4	1.0			
0	Constant	37.7	3.4	0.9585	6.32	1128
1	Randic index (order 1)	71.6	4.1			
2	Molecular weight	–1.27	0.14			
0	Constant	75.6	4.6	0.9793	4.49	1513
1	Randic index (order 1)	146.2	8.1			
2	Molecular weight	–4.6	0.35			
3	Kier & Hall index (order 2)	28.4	2.9			
0	Constant	102.5	6.0	0.9848	3.87	1540
1	Randic index (order 1)	180.1	9.1			
2	Molecular weight	–6.32	0.42			
3	Kier & Hall index (order 2)	40.5	3.2			
4	Kier & Hall index (order 3)	11.1	1.9			
0	Constant	156.9	10.3	0.9892	3.29	1708
1	Randic index (order 1)	298.2	0.92			
2	Molecular weight	–11.6	0.94			
3	Kier & Hall index (order 2)	41.7	2.8			
4	Kier & Hall index (order 3)	24.3	2.7			
5	Randic index (order 2)	34.4	5.7			

How to arrive at orthogonal set from an initial set of descriptors has been already outlined.[8–13] It was interesting to observe that the coefficients appearing in the orthogonalized set of equations derived for descriptors d_1, d_2, d_3,... are precisely the "diagonal" coefficients of the "unstable" stepwise regressions equations! Hence, one can construct the "stable" stepwise regression equations from the set of equations using non-orthogonalized descriptors without actually constructing descriptors d_k^*. Below we show the same stepwise MRA for boiling points of alcohols as described by orthogonalized descriptors (which are marked by an asterisk).

$$BP = 22.383 + 35.406\, d_1^* \quad (7)$$

$$BP = 22.383 + 35.406\, d_1^* - 1.2758\, d_2^* \quad (8)$$

$$BP = 22.383 + 35.406\, d_1^* - 1.2758\, d_2^* + 28.374\, d_3^* \quad (9)$$

$$BP = 22.383 + 35.406\, d_1^* - 1.2758\, d_2^* + 28.374\, d_3^* + 11.057\, d_4^* \quad (10)$$

$$BP = 22.383 + 35.406\, d_1^* - 1.2758\, d_2^* + 28.374\, d_3^* + 11.057\, d_4^* + 34.438\, d_5^* \quad (11)$$

Thus, for example d_2^* of Table 1 (MW*) is that part of MW (molecular weight, d_2) that does not correlate with $^1\chi$ (the first order connectivity index, d_1). The orthogonal descriptor MW* is the residual of a simple regression of MW against $^1\chi$, since, by definition, a residual is that part of a descriptor that does not correlate with the quantity considered.

There are other benefits of the orthogonalization procedure: at each successive step the standard deviations of the coefficients themselves decrease, while the opposite is typically the case with the stepwise regressions using non-orthogonal descriptors. In addition, it can be shown that the construction of orthogonal descriptors as the residuals of successive correlations between descriptors is mathematically equivalent to Gram-Schmidt orthogonalization of vectors d_k^*. In view of all mentioned it is somewhat surprising that these elegant results that resurrected MRA are yet not sufficiently appreciated, although the number of users of orthogonal descriptors in MRA increases steadily.[14–19]

Search for optimal descriptors

Hundreds of molecular descriptors available for use in the MRA have also found use in the Principal Component Analysis (PCA),[20] the Pattern Recognition,[21] the Artificial Neural Networks (ANN),[22,23] and other data reduction techniques. Recently Lahana and coworkers[24] have shown use of topological indices in computer-assisted drug design based on screening of combinatorial library having some 280,000 virtual compounds. Most of molecular descriptors can readily be calculated using available software packages, like CODESSA,[7] POLLY,[25] MOLCONN,[26] E-CALC,[27] GRAPH III,[28] DRAGON,[29] MOLGEN-QSPR,[30,31] PRECLAV,[32,33] and others.[34] In Table 2 we illustrate the stepwise regression equations using CODESSA for the boiling points of 100 alcohols having from two to ten carbon atoms.

Table 2. The stepwise regression equations for the boiling points of alcohols (n = 100) obtained by CODESSA using an exhaustive search of optimal descriptors (that is at each step selecting the best combination of n descriptors).

		Descriptor	Coefficients	Standard error	r	s (°C)	F
0	D_0	Constant	22.4	4.0	0.9246	8.480	1201
1	D_1	Randic index (order 1)	35.4	1.0			
0	D_0	Constant	24.1	2.8	0.9716	5.232	1658
1	D_1	Information content (order 1)	2.07	0.075			
2	D_2	Kier flexibility index	9.71	0.36			
0	D_0	Constant	29.5	3.2	0.9741	5.024	1202
1	D_1	Information content (order 1)	1.77	0.12			
2	D_2	Kier flexibility index	9.78	0.34			
3	D_3	Kier & Hall index (order 3)	5.39	1.8			
0	D_0	Constant	46.6	2.7	0.9802	4.056	1177
1	D_1	Randic index (order 1)	68.0	4.1			
2	D_2	Randic index (order 2)	–15.6	3.6			
3	D_3	Kier & Hall index (order 2)	30.6	3.3			
4	D_4	Kier & Hall index (order 0)	–29.5	3.9			
0	D_0	Constant	156.9	10.3	0.9891	3.289	1708
1	D_1	Randic index (order 1)	298.2	20.9			
2	D_2	Randic index (order 2)	34.4	5.7			
3	D_3	Kier & Hall index (order 2)	41.7	2.8			
4	D_4	Molecular weight	–11.6	0.94			
5	D_5	Kier & Hall index (order 3)	24.3	2.7			

Table 3. A selection of MRA results as reported in the literature for the boiling points of alcohols.

N	Descriptors	r	s (°C)	F
n = 58	weighted paths p_1, p_2	0.9938	4.039	2193
n = 58	weighted paths p_1, p_2, p_3	0.9943	3.891	1578
n = 100	flexible connectivity	0.9915	4.018	5691
n = 63	connectivity index $^1\chi$	0.970	9.35	
n = 63	$^1\chi$ and c_{OH}	0.982	6.47	
n = 37	valence connectivity $^1\chi^v$	0.9555	10.3	
n = 37	$^1\chi^v$ and $^1\chi$	0.9925	4.4	
n = 123	2 descriptors	0.975	4.05	
	3 descriptors	0.982	3.49	
	4 descriptors	0.989	2.79	
	5 descriptors	0.990	2.66	
	6 descriptors	0.991	2.46	
	7 descriptors	0.993	2.24	
n = 85	CODESSA: 3 descriptors	0.9728	14.0	

The boiling points of alcohols were studied in the past and can be used to test properties of molecular descriptors for use in QSPR and QSAR (quantitative structure-property relationship and quantitative structure-activity relationship, respectively). Kier and Hall[5] studied the same 100 alcohols that we are re-examining here, however they combined them with 125 alkanes into a singe set of 245 structures. They obtained, by using five electrotopological descriptors in their model, the standard error s = 8.00 °C (the regression coefficient r = 0.97 and the Fisher ratio F = 755). As we see from the bottom part of Table 2 the CODESSA offers when five descriptors are used (but for alcohols only) an impressive standard error s = 3.29 °C. Other studies on alcohols are summarized in Table 3.[35]

On ordering of descriptors

Choice of descriptors may depend on the model considered. For example, Kier and Hall selected as descriptors atomic contributions of the five atomic kinds: CH_3, CH_2, CH, C, and OH.[35] Similarly, in a study of enthalpic properties of alkanes Garbalena and Herndon[36] considered (in addition to atomic contributions) also bond type contribution (CH, CH_2), (C, CH_2), (CH, CH), and so on. Molecular connectivity indices,[1,37] which represent weighted path contributions, similarly allow an ordering of descriptors depending on their relative magnitudes. For these examples the "problem of descriptor ordering" may have been in part "solved," because the selected models themselves dictate the choice of descriptors and their sequencing.

The problem of ordering of descriptors can be viewed as a sub problem of the more serious problem: the selection of molecular descriptors. Orhtogonalization process implies an ordering of descriptors, whether one considers an orthogonalization of vectors in Linear Algebra, an orthogonalization of basis functions in quantum chemistry, or a stepwise regression in MRA. In the case of the connectivity indices one can naturally order these descriptors. The same is true also for half a dozen family of molecular descriptors which are based on paths,[38–39] weighted paths,[40–41] or paths and walks,[43] where the length of path or walks allow "natural" order for descriptors. Descriptors confined to atomic groups lack the inherent hierarchical ordering and the flexibility of truncation

of number of descriptors in stepwise regression. The same is true for the traditional descriptors of Hansch approach to QSAR,[44] because there is no natural way to decide and give preference to descriptors like log P or Hammet's sigma, etc., that may have similar importance and have different physico-chemical origin. The same problem may also face topological indices. This is the case with Bonchev's "overall connectivity",[45] which is based on all subgraphs of a molecular graph. There is no "natural" order that reflects structural aspects of molecular graphs for all subgraphs. Subgraphs can be ordered by size, but subgraph of the same size can only be ordered lexicographically, as already proposed for alkanes by Gordon and Kennedy.[46] Lexicographic ordering does not necessarily reflect structural character of compounds considered. Be as it may, once descriptors are ordered by whatsoever criteria, they can be orthogonalized and as a consequence a set of *stable* regression equations are derived. A change of ordering will alter resulting equations and it is up to a researcher to justify the particular ordering adopted.

The more serious problem arises when in a stepwise regression at different steps different descriptors arise. Here descriptors at one step may differ from descriptors in previous steps. This does not offer an ordering scheme of descriptors as has been the case with stepwise regression using greedy algorithm, which retains previously obtained descriptors and add the best new additional descriptor from the available pool. The difficulty here is not in ordering descriptors but in not having consistent set of descriptors at various stages of development of MRA model.

On construction of a consistent set of descriptors

Models in which optimal descriptors are selected based on prescribed tolerance of some statistical parameters as the criteria for inclusion/exclusion in MRA, are characterized by series of difficulties. This appears to be the case with CODESSA and other statistical packages based on screening of large pool of descriptors. Their stepwise regressions are characterized by an apparent chaotic occurrence of descriptors at each step of the regression. This has been illustrated in Table 2. As we see from Table 2 the best single descriptor, Randic index (order 1), which is the label for the first order connectivity index ${}^{1}\chi$ in the CODESSA Manual,[7] is replaced in the second step by the Information content (order 1), and the Kier's flexibility index. Few steps later, both, the Information content (order 1) and the Kier flexibility index, are no longer among the optimal descriptors.

Despite that this chaotic behavior of MRA in stepwise regressions has been amply demonstrated, and has been well known for a long time, apparently the problem has not received due attention. Some attention was given to the problem associated with selection of descriptors in CODESSA by extending an exhaustive search for the best combinations of descriptors. Thus Lučić and Trinajstić[47] have presented a new method for the selection of descriptors in the best possible MRA by taking advantage of an orthogonalization procedure for descriptors. Their procedure allows fast calculation of the statistical parameters, which in turn makes possible to exhaustively screen all combinations of five (sometimes even more) descriptor in a set of 100 descriptors supplied by CODESSA. In another study Lučić and coworkers[14,48–49] examined over 10^{10} combinations of five descriptors from a pool of almost 300 descriptors suggesting extension of such approach to incorporate additional descriptors when needed.

An efficient screening of all combinations of five descriptors, by being exhaustive, will clearly lead to the best combination for considered descriptors. In CODESSA the screening is limited only to combinations of ten best single descriptors found in the first step. Thus CODESSA may be locked in a local minimum (for the standard error) and may miss some combinations of *two* and *more* descriptors that do not involve initial best descriptors. However, regardless whether optimal descriptors describe local minima or an absolute minimum, the chaotic character of stepwise regressions using statistical screening of descriptors, as illustrated in Table 2, persists. The selection of descriptors remains sensitive to inclusion or exclusion of one or more descriptors. The question therefore remains open: How one decides what are the best descriptors for a particular property? Should one include all, or should we discard the initial descriptors, despite that sometimes they can account for most of the variability of the property? These are the question that we want to address here.

Structure-property subspace

Most MRA are reporting goodness of the fit by listing calculated properties and the residuals without giving much attention to the *regression equations*. Occasionally individual descriptors are highlighted and their occurrence in the regression is tried to be rationalized based on the model considered. The instability of the regression equations and the chaotic occurrence of different descriptors at different level of MRA are not the only "culprit" for a lack of attempts to interpret results of QSPR and QSAR. Additional difficulty comes from the lack of simple structural interpretation of many molecular descriptors used in construction of MRA. For example, an interpretation of the well known and still widely used Wiener number, W, proposed by Wiener over 50 years ago[50] is still open.

Mathematically W is well defined, and can be readily calculated either by an algorithm proposed originally by Wiener, or by summing the entries above the main diagonal in a graph distance matrix,[51] or as the sum of weighted paths in a molecule, when the weights of the path increase with the length.[52] But what W represents? The situation is further complicated by more recent finding that in the case of isomers, W is collinear with a "reversed W," an index computed by giving the largest weight to atoms at the smallest separations.[52] Balaban and Ivanciuc[53] have considered a similar "reversed" order of path weights.

A promising direction in resolving some ambiguities concerning structural interpretation of topological indices has recently been outlined.[54–58] It was suggested that all structural descriptors (particularly topological indices) should be expressed in terms of few elementary structural invariants, such as bonds, and paths and walks of increasing length. It remains to be seen how will this help interpretation of descriptors used in MRA, but before one can even attempt to interpret descriptors one has to resolve the ambiguities arising from the chaotic occurrence of descriptors in stepwise regressions.

The first important step is to recognize that it is not the individual descriptors that occur in a regression equation that are important as is the *totality* of descriptors occurring in the regression equation. In other words, what is important is the *subspace* spanned by the *set of the descriptors* occurring in the equation. This has been overlooked for too long. Too often a few dominant descriptors are viewed as important and a model is discussed having them in mind. However, molecular descriptors occurring in regression equations define a structure-subspace, which apparently captures satisfactorily the dominant structure-property relationships.

Orthogonalization procedure will introduce some "order" in that subspace, in that it offers a basis that spans the subspace. Othogonalization does not change the subspace -- it is still the same subspace! The problem with the stepwise regression is that often at different steps one "jump" from one structure-property subspace to another structure-property subspace. The two subspaces may have a major part of space in common, but that need not be apparent because different descriptors may have been used. The situation is illustrated on the case of alkane heat of atomization (ΔH_a) and alkane heats of formation (ΔH_f). Using several connectivity terms Kier and Hall[59] reported the following regressions for ΔH_a and ΔH_f respectively:

$$\Delta H_a = 286.15n - 12.08\ {}^1\chi + 0.92\ {}^4\chi + 1.50\ {}^5\chi - 2.44\ {}^5\chi_C + 0.86\ {}^4\chi_{PC} - 0.50\ {}^5\chi_{PC} - 1.42\ {}^6\chi_{PC} + 114.65 \quad (12)$$

$$\Delta H_f = 1.15\ {}^1\chi - 2.25\ {}^2\chi + 7.63\ {}^3\chi - 12.02\ {}^4\chi_C - 1.72\ {}^5\chi_C + 0.89\ {}^4\chi_{PC} - 1.46\ {}^5\chi_{PC} - 0.28 \quad (13)$$

Observe that two regressions employ several different descriptors. Therefore it is not suspected that the two quantities may belong to the same structure space and be closely related. Even a direct correlation between ΔH_a and ΔH_f does not indicate intimate relationship:

$$\Delta H_a = 53.260\ \Delta H_f - 413.353 \quad (14)$$

with r = 0.9818, s = 78.91 and F = 1070, where r, s, and F are the correlation coefficient, the standard error and the Fisher ratio, respectively.

If, however, one considers separately isomers of octanes, rather than collecting information on all alkanes into a single correlation, one finds that ΔH_a and ΔH_f are more then closely related: They are collinear! For heptanes and octanes one finds: $\Delta H_a = \Delta H_f + 1751.14$ and $\Delta H_a = \Delta H_f + 2308.12$ respectively. The correlations for isomers of heptane and octane only differ in the constant term. Hence, clearly for alkanes the structure-property subspace for both properties ΔH_a and ΔH_f is the same, but the fact was obscured by the appearance of different combination of molecular descriptors when ΔH_a and ΔH_f are considered for all alkanes.

Retro-regression

We propose in this section an answer to the troublesome situation in which at different steps in stepwise regression different descriptors appears, suggesting each time different structure subspaces. The answer is: Retro-regression,[60] which was introduced for modeling of boiling points of nonanes by using connectivity indices of different orders. Here we will show the aplication of retro-regression on modeling properties of compounds containing heteroatoms. At the same time matematical structural representation of chemical structures was extended to other molecular descriptors. Let us reconsider the results given in Table 2. Let us assume that the regression using five descriptors is the solution that one has selected as optimal. Then the following regression equation describes the boiling points (BP) of alcohols (n=100):

$$BP = 298.27\ {}^1\chi + 34.438\ {}^2\chi + 41.707\ {}^2\chi^v - 11.636\ MW + 24.342\ {}^3\chi^v + 156.90 \quad (15)$$

Here MW stands for molecular weight as a descriptor, ${}^m\chi$ are the connectivity indices and ${}^m\chi^v$ are the valence connectivity indices. From equation (15) we see that the structure-property subspace is defined by the set: {${}^1\chi$, ${}^2\chi$, ${}^2\chi^v$, ${}^3\chi^v$, MW}, (the order of listing elements in a set, of course, is not important).

Consider now the above subset of descriptors as the starting point. We will search for suitable basis of the subspace defined by the five descriptors. This implies to select a particular order in which to consider descriptors

in a stepwise regression. The collection of regression steps shown in Table 1, which have lead to the final regression equation involving five descriptors, should be viewed as a "history" of arriving at the final solution. Once we have arrived at the final regression equation (and associated subspace) we can consider alternative basis for that. Here we advocate a route based on the concept of retro-regression as very plausible and the most natural approach to ordering of descriptors.

The concept of retro-regression,[60] or backward stepwise regression, starts from the final regression equation and its subspace as the solution and searches for an ordering of descriptors that will lead to orthogonal basis for that subspace. Thus, we consider the five descriptors that define the solution set and search for the descriptor that makes the least significant contribution. When identified such descriptor will be eliminated as the least important. The process continues with a search for the next least important descriptor, till all but the last of the descriptors is eliminated in a stepwise fashion.

The least important descriptor is descriptor that is associated with the smallest decrease of the standard errors when removed from the set. In the case of the regression of Table 2 by testing the five possibilities, each time removing one of the final five descriptors, we arrive at regressions based on the following sets of four descriptors: $\{^1\chi, ^2\chi, ^2\chi^v, ^3\chi^v\}$, $\{^1\chi, ^2\chi, ^2\chi^v$, MW$\}$, $\{^1\chi, ^2\chi, ^3\chi^v$, MW$\}$, $\{^1\chi, ^2\chi^v, ^3\chi^v$, MW$\}$, $\{^2\chi, ^2\chi^v, ^3\chi^v$, MW$\}$. We find that Randic index (order 2) makes the smallest contribution to the standard error in the last step of the regression, hence it is discarded. Now we have four descriptors and four possibilities to examine: $\{^1\chi, ^2\chi^v, ^3\chi^v\}$, $\{^1\chi, ^2\chi^v$, MW$\}$, $\{^1\chi, ^3\chi^v$, MW$\}$, and $\{^2\chi^v, ^3\chi^v$, MW$\}$. As we see from Table 4, where the results of retro-regression have been summarized, the next descriptor to be eliminated is Kier and Hall index (order 3). The process is continued and in the next step Kier & Hall index (order 2) is discarded, then the molecular weight as descriptors is discarded to leave Randic index (order 1) as the dominant descriptor of the five considered. If we now compare Table 2 (the "history" of the best regression) and Table 4 (retro-regression) we see that stepwise regressions of both routes end with the same structure-property subspace $\{^1\chi, ^2\chi, ^2\chi^v, ^3\chi^v$, MW$\}$. In this particular illustrations also the both table start with the same best single descriptor, but that need not be the case, and in general it is not the case. Table 4 and Table 1 in fact contain the same information, because we have selected descriptors for Table 1 to correspond to the final five descriptor structure subspace as determined by CODESSA.

Because by retro-regression of Table 4 we arrived at a particular ordering of descriptors we can order the descriptors as follows: $^1\chi$, MW, $^2\chi^v$, $^3\chi^v$, $^2\chi$. The stepwise regressions associated with this order are:

$$BP = 35.406\ ^1\chi + 22.383 \quad (16)$$

$$BP = 71.629\ ^1\chi - 1.2758\ MW + 37.703 \quad (17)$$

$$BP = 146.25\ ^1\chi - 4.6113\ MW + 28.374\ ^2\chi^v + 75.647 \quad (18)$$

$$BP = 180.13\ ^1\chi - 6.3231\ MW + 40.533\ ^2\chi^v + 11.057\ ^3\chi^v + 102.55 \quad (19)$$

$$BP = 298.27\ ^1\chi - 11.636\ MW + 41.707\ ^2\chi^v + 24.342\ ^3\chi^v + 34.438\ ^2\chi + 156.90 \quad (20)$$

Table 4. The Retro-Regression equations for the boiling points of alcohols (n = 100). At each step the least important descriptor is eliminated in a stepwise fashion.

	Descriptor	Coefficients	Standard error	r	s (°C)	F
0	Constant	156.9	10.3	0.9892	3.29	1708
1	Randic index (order 1)	298.2	0.92			
2	Molecular weight	–11.6	0.94			
3	Kier & Hall index (order 2)	41.7	2.8			
4	Kier & Hall index (order 3)	24.3	2.7			
5	Randic index (order 2)	34.4	5.7			
0	Constant	102.5	6.0	0.9848	3.87	1540
1	Randic index (order 1)	180.1	9.1			
2	Molecular weight	–6.32	0.42			
3	Kier & Hall index (order 2)	40.5	3.2			
4	Kier & Hall index (order 3)	11.1	1.9			
0	Constant	75.6	4.6	0.9793	4.49	1513
1	Randic index (order 1)	146.2	8.1			
2	Molecular weight	–4.6	0.35			
3	Kier & Hall index (order 2)	28.4	2.9			
0	Constant	37.7	3.4	0.9585	6.32	1128
1	Randic index (order 1)	71.6	4.1			
2	Molecular weight	–1.27	0.14			
0	Constant	22.3	3.9	0.9246	8.48	1201
1	Randic index (order 1)	35.4	1.0			

Observe that the last equation is identical to that selected from Table 2, but now we can, by orthogonalization, constructed regression equation that have no chaotic behavior of descriptors. We have resolved the "problem of inconsistent sets of descriptors" associated with selection of descriptors. We are left with the "problem of instability of equations," that is reflected in variations of the magnitudes of the coefficients for individual descriptors at different steps of the regression. However, as we mentioned in the introduction, this "problem" is no more a problem and can be removed by using orthogonalized descriptors. By using the outlined orthogonalization procedure we arrive at the stepwise orthogonalized regression equations (**7**) – (11) shown before.

Observe an interesting relationship between the "chaotic" equations, (16) – (20) associated with mutually interrelated descriptors, and the "steady" equations, (7) – (11) associated with the orthogonalized descriptors. The coefficients that occur in the orthogonalized equations and show the "stability" have appeared also in the "chaotic" equations but only when the descriptors appear for the first time. Thus, for example, the coefficient –1.2758 (of MW) appeared the first time when MW was added as a descriptor, in equation (17). Again the addition of subsequent descriptors changes this coefficient, because they mutually correlate and thus in part duplicate the same information.

Conclusions

It is well known that stepwise multivariate regression analysis (MRA) has serious deficiencies, which make the interpretation of structure - property - activity relationships very difficult. There are two kinds of difficulties in MRA applications, which become apparent in comparison of stepwise regressions. With each step in a regression *different set* of descriptors may emerge as the best choice and even if one keeps all descriptors found in previous steps, their contribution to the regression, reflected in the magnitudes of the corresponding coefficients, may change dramatically from one step to another.

The main reason for these ambiguities is inter-relation between descriptors used in MRA models. Usually, several descriptors may define the same structural subspace and therefore when they are introduced into the existing MRA model they may change descriptors used, or may change the relative contribution of descriptors already present in the model. As outlined in this paper the first problem can be solved by retro-regression, and the second problem by the orthogonalization of the regression descriptors. All that one has to do is to interpret the descriptors in the *final* MRA equation as descriptors that define the structure-property space. Then applying the Retro-Regression one can order these descriptors and initiate orthogonalization of the descriptors that will result in stable regression equations.

Clearly Retro-Regression, combined with construction of Orthogonalized Regression Equations as outlined above, eliminates the both ambiguities in MRA, those associated with oscillatory behavior of the coefficients of regression equations, and those associated with the occurrence of descriptors not previously encountered in a stepwise regression. Hence, Retro-Regression offers firm foundation for interpretation of the regression equations and discussing MRA models. Finally, we should emphasize that Retro-Regression could be applied to any multivariate regression analysis, whether one considers stepwise regression or not.

Acknowledgements

This work was supported by the Ministry of Education, Science and Sport of the Republic of Slovenia (Grant P1-0017 and P1-0153).

References

1. M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609.
2. M. Randić, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672.
3. L. Xu, W.-J. Zhang, *Analyt. Chim. Acta* **2001**, *446*, 477.
4. B. Lučić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121.
5. B. Lučić, D. Amić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403.
6. L. B. Kier, L. H. Hall, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039.
7. A. R. Katritzky, V. Lobanov, and M. Karelson, CODESSA (COmprehensive DEscriptors for Structural and Statistical Analysis), University of Florida, Gainesville, Fl.
8. M. Randić, *New J. Chem.* **1991**, *15*, 517.
9. M. Randić, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
10. M. Randić, *J. Comput. Chem.* **1993**, *14*, 363.
11. M. Randić, *Int. J. Quant. Chem: Quant. Biol. Symp.* **1994**, *21*, 215.
12. M. Randić, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 311.
13. M. Randić, D. J. Klein, N. Trinajstić, *Int. J. Quant. Chem.* **1997**, *63*, 215.
14. D. Amić, D. Davidović-Amić, A. Jurić, B Lučić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034.
15. B. Lučić, S. Nikolić, N. Trinajstić, D. Juretić, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532.
16. M. Soškić, D. Plavšić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829.
17. L. Pogliani, *Amino Acids* **1994**, *6*, 141.
18. L. Pogliani, *J. Pharm. Sci.* **1992**, *81*, 334.
19. L. Pogliani, *J. Pharm. Sci.* **1992**, *81*, 967.

20. K. Balasubramanian, S. C. Basak, *Chem. Inf. Comput. Sci.* **1998**, *38*, 367.
21. T. Okuyama, Y. Miyashita, S. Kanaya, H. Katsumi, S. I. Sasaki, M. Randic, *J. Comput. Chem.* **1988**, *9*, 636.
22. S. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, D. Opitz, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885.
23. T. A. Andrea, H. Kalayeh, *J. Med. Chem.* **1991**, *34*, 2824–2836.
24. G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h, R. Buelow. *Nature Biotechn*. **1998**, *16*, 748.
25. S. C. Basak, POLLY (Natural Resources Research Institute, Duluth, University of Minnesota, Duluth, MN).
26. L. H. Hall, (1991) MOLCONN-Z, Hall Associates Consulting, Quincy, MA, see: http://www.eslc.vabiotech.com/molconc/manuals/.
27. L. B. Kier, L. H. Hall, *Molecular Structure Description,* Academic Press, New York 1999.
28. A. Sabljić, D. Horvatić, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 837.
29. R. Todeschini et al.: http://www.talete.mi.it/dragon.htm
30. C. Rucker, M. Meringer, A. Kerber, *J. Chem. Inf. & Mod.* **2005**, *45*, 74–80.
31. C. Rucker, M. Meringer, A. Kerber, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2070–2076.
32. L. Tarko, O. Ivanciuc, *MATCH (Comm. Math. Comp. Chem.)* **2001**, *44*, 201–214.
33. L. Tarko, *ARKIVOC* **2004**, part XIV, 74–82.
34. M. Vedruna, S. Marković, M. Medić-Šarić, N. Trinajstić, *Computers Chem.* **1997**, *21*, 355.
35. M. Randić, S. C. Basak, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261.
36. M. Garbalena, W. C. Herndon, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 37.
37. L. B. Kier, W. J. Murrey, M. Randić, L. H. Hall, *J. Pharm. Sci.* **1976**, *65*, 1226.
38. J. R. Platt, *J. Chem. Phys.* **1947**, *15*, 419.
39. J. R. Platt, *J. Phys. Chem.* **1952**, *56*, 328.
40. L. B. Kier, *Quant. Struct. – Act. Relat.* **1985**, *4*, 109.
41. L. B. Kier, *Quant. Struct. – Act. Relat.* **1986**, *5*, 1.
42. L. B. Kier, *Quant. Struct. – Act. Relat.* **1986**, *5*, 7.
43. M. Randić, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607.
44. C. Hansch, *Acc. Chem. Res.* **2000**, *40*, 934.
45. D. Bonchev, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 934.
46. M. Gordon, J. W. Kennedy, *J. Chem. Soc., Faraday Trans.* **1973**, *69*, 484.
47. B. Lučić, N. Trinajstić, *SAR & QSAR Environ. Res.* **1997**, *7*, 45.
48. B. Lučić, S. Nikolić, N. Trinajstić, A. Jurić, Z. Mihalić, *Croat. Chem. Acta* **1995**, *69*, 417.
49. B. Lučić, N. Trinajstić, S. Sild, M. Karelson, A. R. Katritzky, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610.
50. H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17.
51. Hosoya, *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
52. M. Randić, *New J. Chem.* **1997**, *21*, 945.
53. O. Ivanciuc, T. Ivanciuc, A. T. Balaban, *Models in Chemistry* **2000**, *137*, 57.
54. M. Randić, J. Zupan, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550.
55. M. Randić, J. Zupan, On the structural interpetation of topological indices. In: Topology in Chemistry-Discrete mathematics of Molecules; Rouvray, D. H.; King, R. B., Eds.; Horwood Publ.: Chichester, England, 2002, pp. 249–291.
56. M. Randić, M. Pompe, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575.
57. E. Estrada, *J. Phys. Chem. A.* **2003**, *107*, 7482–7489.
58. E. Estrada, *J. Phys. Chem. A.* **2004**, *108*, 5468–5473.
59. L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research,* Academic Press: New York 1976.
60. M. Randić, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 602.

Povzetek

Stopenjska multivariatna regresijska analiza (MRA) je ena najstarejših tehnik krčenja merskega prostora. Ko izberemo optimalne deskriptorje v neki fazi stopenjske MRA opazimo, da lahko izginejo nekateri deskriptorji, ki so se pojavili v prejšnjih fazah, prav tako se lahko pojavijo linearne kombinacije novih deskriptorjev. Opisane spremembe močno otežijo interpretacijo regresijskih enačb, prav tako je onemogočena konstrukcija ortogonalnih deskriptorjev. V članku smo predlagali postopek, ki rešuje težave pri selekciji optimalnih deskriptorjev pri multivariatni regresijski analizi in omogoča v zaključni fazi konstrukcijo stabilnih multivariatnih regresijskih enačb z uporabo ortogonalnih deskriptorjev.