JOURNAL OF PHYSICAL ORGANIC CHEMISTRY
J. Phys. Org. Chem. 2004; **17:** 876–889
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/poc.754

Modification of data processing and interpretation of results related to the use of multiparameter correlation analysis: introduction of additional characteristics and criteria. Part 1—Application to the treatment of solvent effects

V. Palm,* N. Palm and T. Tenno

Institute of Chemical Physics, University of Tartu, Jakobi 2, 51014 Tartu, Estonia

Received 30 July 2003; revised 16 November 2003; accepted 5 January 2004

epoc

ABSTRACT: Several problems with multilinear regression analysis in chemistry are discussed. These problems are connected with the establishment of scales and their inclusion in the model of significant and, as usually desired, physically or chemically meaningful descriptors. The detection of abundant and interrelated descriptors and of the presence and the mode of elimination of strongly deviating data are discussed, as well as the estimation of absent values for incomplete (considering a set of rows treated) descriptor scales. They were studied using an incomplete data matrix for 359 solvent-dependent processes and solvent parameter scales for 45 solvents. Some non-traditional problems, related to the formalism of application of the basic principles of correlation analysis, are discussed and corresponding characteristics and criteria specified. A quantitative measure of the orthogonality of correlation matrices was introduced. As a result of the realization of this approach, it was found that for multiparameter correlation analyses of the solvent-dependent processes under consideration, seven residual (i.e. free from mutual contributions) descriptors are significant and sufficient. Copyright © 2004 John Wiley & Sons, Ltd.
Additional material for this paper is available in Wiley Interscience

KEYWORDS: correlation analysis; multilinear regression; quantitative description of solvent effects; solvent parameter scales

INTRODUCTION

After discovering the existence of linear free-energy relationships (LFERs), it was soon realized that in the general case they are not always sufficient for the quantitative description of the total effect, which some variable factors (substituent, solvent, etc.) exert on the free energy of the processes considered.[1] In accordance with corresponding interpretative (physical and chemical) considerations, the variation of a single factor is usually connected with a concomitant change of different other properties related to it. In the case of substituent effects, inductive, polar and non-polar resonance, different steric effects, etc., can contribute. For medium effects, the non-specific and specific contributions are distinguishable, each being constituted from several qualitatively different terms. Therefore, if even only one influential factor is varied, the corresponding mathematical model is represented by single-parameter linearity only in specific restricted cases. Generally, multiparameter linear models have to be introduced and for each property related to the variable factor a specific descriptor scale has to be used. The situation becomes even more complicated when the simultaneous influence of two or more variable factors is considered.[1,2]

The data processing problems are considered as the main topic of this paper. Attention is devoted to the proof of the real presence of different terms and the possible ways of the maximum preservation (or assignment) of their essential physico-chemical meaning. The goal is not to abandon the basic principles of 'correlation analysis' (CA) used in chemistry, but some non-traditional problems related to the formalism of their application are discussed and the corresponding characteristics and criteria are specified.

Attention is paid to the methods of reinvestigation of the current state of approach, making use of as numerous as possible sets of sufficiently long data series (in the ideal case, all relevant data exist). The data set is a compilation of different solvent-dependent processes carried out in individual solvents, which has already

**Correspondence to:* V. Palm, Institute of Chemical Physics, University of Tartu, Jakobi 2, 51014 Tartu, Estonia.
E-mail: vpalm@ut.ee
Contract/grant sponsor: Estonian Science Foundation; *Contract/grant number:* 4590.

been used by us for a more detailed analysis.[3–5] Processes available for at least 13 solvents were selected. The whole data matrix consists of 45 solvent rows and 359 columns related to 12 potential descriptors (solvent parameter scales) and 347 series of solvent-dependent data.

The detailed list of these data has been reported earlier[3] and is presented in the Supplementary Material (Tables S1 and S2) available in Wiley Interscience. In this paper, this compilation will be cited as the Extended Data Set for Solvent Effects (EDSSE).

This example demonstrates that the use of correlation equations in chemistry suffers from incomplete data compilations. Therefore, approaches (e.g. factor analysis) that require that a complete data matrix should be available are not of straightforward use. Moreover, the missing data can be encountered not only for the processes to be described but also for the potential descriptor scales defined preliminarily or derived from the whole data compilation. For the EDSSE, this shortage is moderate but this depends on the list of rows (solvents) selected. Nevertheless, a definite procedure for the determination of additional (secondary) values of the elements of descriptors has to be defined. Although in real practice their experimental evaluation is of primary importance, from the point of view of data processing one can restrict the treatment to the data already existing. However, it would go too far to state that the majority of problems encountered in the course of data processing would be avoided or easily solved if a suitable selection of additional experimental data were to made available.

GENERAL THEORETICAL BASIS

The first step of the processing of data for a particular solvent-dependent process, making use of some preliminarily selected set of possible descriptor scales, is the detection of a subset of the significant ones. The result of this procedure depends on the value of the risk level (percentage point) that has to be specified preliminarily. Two equivalent approaches, the Student and *F*-tests, are usable in principle. Practically, the former has an advantage with respect to the amount of calculations. There are several methods for the selection of the most preferred set of descriptors. However, in the general case of a multiparameter treatment, only the scan of all possible regressions can lead to the best model (with maximum value of the multiple correlation coefficient R).[6] This method reduces to the detection of the best, with respect to the R value, subset of significant descriptors. Additionally, some other restrictions can be introduced, if desired.

One of the goals of the multiparameter approach is the detection of comparable contributions made by different descriptors. The values of the corresponding natural regression coefficients are of little value for this purpose owing to the usual incompatibility in the scaling of descriptors. The use of scaled terms (by square roots of dispersions for the descriptors and the response columns—the quantities to be correlated) also appears to be unsuitable. The solution is to use the weight contributions of the particular descriptors in the determination coefficient equal to the square of the (non-corrected) multiple correlation coefficients, R^2:[6]

$$R_j^2 = \sum_{l=1}^{N} W_{jl} + \varepsilon_j^2; \quad W_{jl} = R_{jl} X_{0jl} \tag{1}$$

where W_{jl} is the weight contribution by the lth descriptor for the jth response column, ε_j^2 is the residue of the description, R_{jl} represents the coupling correlation coefficient between the jth response column and the lth descriptor and X_{0jl} is the corresponding scaled regression coefficient.

A multiparameter linear approach causes several additional difficulties compared with a single-parameter approach. The main reason is the presence of interrelations between the descriptor scales, that is, their non-orthogonality.[7] Although a complete (without missing values) set of descriptors for some selection of rows can be converted into an equivalent (nearly) orthogonal set, for any corresponding sub-selection of rows the degree of this orthogonality becomes different. The orthogonality for some subsets of rows and descriptors decreases when the rank of the correlation matrix rises (more descriptors are involved) and along with the decrease in the number of rows involved.

If some set of descriptors for a given selection of rows for some variable factor is considered, it would not be, as a rule, an orthogonal one. This means that different descriptor scales contain some common contributions—they usually do not represent pure measures of the definite properties related to the corresponding factor. To overcome this complication, it would be desirable to 'purify' these scales by removing the contributions already reflected by other scales.[7] It is reasonable to order the descriptor scales in the sequence of the decrease in their assumed 'fundamentality'. The term 'fundamental' is used here in two senses. The first (essential) one is related to the situation when a given descriptor scale is related to some definite physical or chemical property of the solvent (dipolarity, polarizability, acidity or basicity) or substituent (some of the electronic or steric effects). For the second, an operational definition can be assigned. When solvent properties are considered, the 'fundamental' scale of the solvent property should be established using measurements related to the pure solvents only. From this point of view, the dipolarity scale defined via the values of dielectric permittivity is considered as a 'fundamental' one but that derived on the basis of spectral data for dissolved species does not satisfy this criterion. It is assumed that the consideration of a less fundamental descriptor as a contribution to the more fundamental one would have little sense. At the same

time, one cannot exclude the possibility that several equally fundamental descriptor scales may be presented.

After this kind of ordering of the potential descriptors, the regression parameters of the each subsequent one, starting with the scale in the second position, from preceding (residual) scales have to be calculated and their significant contributions subtracted from this scale. If after completion of this procedure a significant residual scale remains, this may be considered as a conventional 'pure' measure of some property related to the varying factor. This property may even not be the one claimed for the initial scale. For instance, if the starting descriptor has been introduced as a combined effect of the dipolarity and acidity of the solvent, and these properties are represented by the preceding scales, the resulting residual scale has nothing to do with these properties. If the last one appears to be significant, it represents the manifestation of some additional solvent property of known or unknown nature. The set of residual descriptors obtained should preferably be used for the correlation of the response columns of the data matrix.[7]

It has already been stated that, if any subset of rows is separated from the set of orthogonal data columns, then their orthogonality disappears. The same happens if some additional rows are added. Therefore, strictly, the orthogonality can only be related to some definite data matrix with all positions filled. As this situation represents a rare exception, the non-orthogonality of the descriptor columns for particular processes to be described should be recognized as practically inescapable. Therefore, one needs a measure(s) of the degree of non-orthogonality and the rules of usage of criteria for the detection of its unacceptable level. If the descriptor scales are defined rigidly, the only possibility for a rise in the level of orthogonality is a reduction in the number of descriptors involved in the multilinear model. This classical procedure of exclusion of the statistically insignificant descriptors making use of the Fisher test or the Student criterion on some risk (significance) level should be supplemented with one devoted to the keeping of the non-orthogonality at the conventionally acceptable level.[7]

For a pair of columns the level of non-orthogonality is reflected by the corresponding coupling correlation coefficient. For a more numerous set of columns, the situation becomes rather complicated. Although, for the completely orthogonal compilation, the determinant of the correlation matrix equals unity and for completely interrelated one this value becomes zero, this characteristic depends on the rank of the matrix. Therefore, it cannot be used as a sole measure of the level of orthogonality. Nevertheless, the conventional characteristic can be defined if the matrix, with all non-diagonal elements being equal, has the same determinant value as the one to be characterized.[8] Then, the corresponding effective value R_{eff} of the correlation coefficient for this model matrix of the same rank can be accepted as a measure of the non-orthogonality for the matrix under consideration. For the correlation matrix related to two data columns, this effective value is represented by corresponding coupling correlation coefficient. The value of R_{eff} enables one to use the comparable approach to the matrices of different ranks. It allows the introduction of the measures of non-orthogonality (N_{onorth}) and orthogonality (O_{rth}) for the given set of descriptors as follows:

$$N_{\mathrm{onorth}} = {R_{\mathrm{eff}}}^2 \tag{2}$$

and

$$O_{\mathrm{rth}} = 1 - {R_{\mathrm{eff}}}^2 \tag{3}$$

It could be said that the value of O_{rth} is not a less important characteristic of the quality of description than the value of R^2. The product of both of them can be introduced as a combined characteristic G_{d} of the quality (goodness) of description:

$$G_{\mathrm{d}} = R^2 \times O_{\mathrm{rth}} = R^2 \times (1 - {R_{\mathrm{eff}}}^2) \tag{4}$$

Besides R^2, the value R^{*2} of the square of the correlation coefficient, corrected taking into account the number of statistical degrees of freedom, is usable as a stricter characteristic of the precision of description ($R^{*2} = 1 - {S_0}^2$, where S_0 is the scaled standard deviation).

However, ${R_{\mathrm{eff}}}^2$ does not allow the reflection of the role of a particular descriptor in the total manifestation of the non-orthogonality. There are two different modes of appearance of this role. The first has been named the 'over-pumping' effect (OE)[7]—the additional vagueness of the scaled coefficients at the expense of each other. It is observable as too high values of the scaled standard deviations of these coefficients not caused by the random errors in the response values. The sum of the squares of scaled standard deviations S_{X0jl} of the scaled coefficients is comparable to the square of total value S_{0j} of the scaled standard deviation for the corresponding response column. Owing to the random deviations for its elements, the following relation holds:[7]

$$\sum_{l=1}^{N} {S_{X0jl}}^2 \leqslant {S_{0j}}^2 \tag{5}$$

This relation does not hold if the sum on the left-hand side is increased at the expense of the additional indefiniteness of some coefficients caused by the OE. The magnitude Q_j of this effect can be represented by the expression[8]

$$Q_j = \sum_{l=1}^{N} {S_{X0jl}}^2 / {S_{0j}}^2 \tag{6}$$

where Q_j equals the spur (trace) of a rearranged correlation matrix that is entirely related only to the diagonal elements of the rearranged correlation matrix. The

relation $Q_j > 1$ indicates the existence of the considerable OE.

Another approach to the characterization of the nonorthogonality is possible.[5] For the fully orthogonal set of descriptors their weight contributions are equal to the squares of the coupling correlation coefficients between response and corresponding descriptor columns: $W_{jl} = R_{jl}^2$. Equation (1) and the expression for the scaled coefficients of the regression:

$$X_{0jl} = \sum_k R_{jk}R_{jlk}^{-1} = R_{jl}R_{jll}^{-1} + \sum_{k\neq l} R_{jk}R_{jlk}^{-1}$$

where R_{jlk}^{-1} denotes the corresponding element of a rearranged correlation matrix and k is the index of the descriptor, enable one to derive an expression for the scaled (by R_{jl}^2) weight W_{0jl} value:[5]

$$W_{0j,l} = W_{jl}/R_{jl}^2 = 1 + TMT_{jl} = 1 + DMT_{jl} + NMT_{jl}, \tag{7}$$

where

$$TMT_{jl} = \text{total 'mixed' term} = DMT_{jl} + NMT_{jl} \tag{8}$$

$$DMT_{jl} = \text{diagonal 'mixed' term} = R_{jll}^{-1} - 1 \tag{9}$$

$$NMT_{jl} = \text{non-diagonal 'mixed' term} = \sum_{k\neq l} R_{jk}R_{jlk}^{-1}/R_{jl} \tag{10}$$

Owing to the scaling, both DMT_{jl} and NMT_{jl} represent the fractions of weight contributions related to the whole set of descriptors involved, the orthogonal part being equal to unity. Therefore, if $W_{0j,l} = 1$, the solution is fully orthogonal with respect to the lth descriptor. If the absolute value of the sum of 'mixed' terms exceeds unity, the corresponding values of the regression coefficient and weight do not reflect mainly the contribution of the corresponding descriptor. Its presence (although statistically significant) in the solution obtained has no direct interpretative value. To put it in other words, the presence of such terms in the correlation equation is the cause of the formal effect of the rise in the precision of description. However, it may have nothing to do with the real contribution of the corresponding descriptor.

The total value of the 'mixed' term TMT_{jl} equals to the sum of these two terms cited. If they have opposite signs, their influence can be in some extent compensated and one may speak about the hidden part of the 'mixing' effect.

The magnitude of the 'mixed' part of weight can be used as an additional criterion of selection of the set of really significant and acceptable descriptors.

If a numerous set of data series is considered simultaneously, one meets a situation where the accuracy of description varies largely depending on the data series. Therefore, the total result should represent the distribution of these series over different precisions of description, characterized by the scaled standard deviation S_0 or by the corrected multiple correlation coefficient R^*. The simplest way to reflect this distribution is the introduction of a conventional level of accuracy ($S_{0\text{crt}}$ or R_{crt}^*) to distinguish between the sets of comparably precisely and roughly described data series. The numbers of series belonging to each one of these sets and the corresponding root mean square values $\bar{S}_0$ or $\bar{R}^*$ can then be considered as accuracy measures for the total set of response columns.[7] These characteristics can be useful when the results for different approaches are compared or the series belonging to a single one of these sets have to be selected for a separate treatment.

The next problem is the definition of the procedure for the elimination of strongly deviating data points from the response columns. The classical way of exclusion from the treatment of significantly deviating values makes use of the Student criterion on the significance level accepted, but this is not the sole or most reasonable approach. In the case of a numerous data set columns, the accuracy of description varies widely for the different data series. The use of the Student criterion is directly dependent on the precision of description. The higher is the last one, the less are the scaled (by the square root of dispersion) deviations D_{0ji} (i is the index of the data point) of the eliminated points. This makes the results of this procedure non-comparable for different response columns. It is obvious that elimination of points from the data series with a high precision of description has little sense. In contrast, for the series with low precision of description, it is highly desirable that a considerable rise of the (multiple) correlation coefficient R_j^* would be the result. Therefore, the alternative procedure,[4] in which during each attempt the maximally deviating point is eliminated from the data column with $R_j^* < R_{\text{crt}}^*$, would be reasonable. Simultaneously, the maximal yet acceptable number of points eliminated from a single data series M_{exmax} and/or the minimal value for the degrees of freedom $H_{\min}$ can be stated.[4]

For the additional characterization of the results for some data matrix, the prognostic ability of the model based on the set of descriptors used would be reasonable to test. The use of arbitrarily selected subsets of rows of the single data series for the prediction of the remaining values[9] is complicated when a large set of data series with many lacking values has to be treated. Alternatively, a procedure based on the result obtained after the exclusion from this set of a definite number M_{rnd} of randomly selected points belonging to several data series can be introduced. The solution obtained is used for the further calculation of predicted values for this selection of excluded points. To characterize the prognostic ability independently from the particular selection of these excluded M_{rnd} points, the procedure has to be repeated

several times and the averaged results with the range of their uncertainty can be calculated.

The analogy with the use of $S_{0\mathrm{crt}}$ for setting up a borderline between relatively more precisely and roughly described data series can be introduced, and this criterion can be used for distinguishing between corresponding selections of points. Instead of the S_{0j} values, the scaled deviations D_{0ji} for particular points will be considered and two selections of them defined. The points with values of $D_{0ji} < S_{0\mathrm{crt}}$ belong to the set of more precisely described ones. The relation $D_{0ji} > S_{0\mathrm{crt}}$ specifies the set of roughly described points. The fractions of the more precisely described points can be detected for processed and prognosticated selections, and also the corresponding mean values $\bar{D}_0$ of the scaled deviations. These characteristics enable one to compare the precisions of the description of prognosticated and processed points. Their dependence on the fraction of points excluded for the sake of the subsequent prognosis can serve as characteristics of the stability of the prognostic ability in the course of the increase in the ratio of the points prognosticated over those used for the preliminary specification of the model.

The specification of an adequate number and the selection of the descriptor columns is a separate problem. For a large set of response columns, the statistical significance of some additional potential descriptor columns for a considerable part of them does not automatically mean that this descriptor belongs to the set of the necessary and sufficient ones. Actually, the introduction of an arbitrary additional descriptor column always results in its statistical significance for some number of response columns, even in the case of columns constituted from random numbers.[4] Therefore, a trivial conclusion can be derived that such an introduction of the additional descriptors will result in a significantly larger rise of the $\bar{R}^*$ value than that caused by introduction of a column of random values. This must be taken into account when the real statistical significance of a descriptor column is detected and the preliminarily corrected statistical characteristics (mean values of S_0 and the total number of the statistical degrees of freedom for the whole set of responses) have to be used for the F-test. The descriptor under test should be moved to the very end of their sequence and the contributions of all preceding ones will be subtracted from it to obtain the corresponding 'pure' residual scale.[4]

The correction terms were calculated as the mean values for a set of runs with different vectors of random values, which are substituted for the descriptor tested. After introduction of the corrections, the remaining values of the scaled standard deviation S_{0j} and the number of the degrees of freedom can be used for the F-test of statistical significance of the descriptor under consideration. It is trivial that some differences can be observed between the results for the data processing runs with different independent versions of this random vector. Therefore, the mean results of the respective data processing are accompanied by some degree of uncertainty reflected by corresponding standard deviations. This uncertainty is transmitted to the corrected F-value. Therefore, the last one is accompanied by the relevant standard deviation and the range of uncertainty of the F-value will also be considered. In our first attempt[4] to use this approach, the descriptor was accepted as a significant one if the corresponding F-value exceeded the value of the percentage point $F(\alpha_\mathrm{r}, \nu_1, \nu_2)$ of the F-distribution (α_r is a risk level and ν_1 and ν_2 are the numbers of the degrees of freedom), at $\alpha_\mathrm{r} = 0.10$ or 0.05 and the lower limit $F - SD_F$ exceeded the value of $F(0.10, \nu_1, \nu_2)$.

In addition, there exists another criterion for the selection of the given descriptor as a significant one. It is obvious that if some descriptor is fully described by the linear combination of a set of other ones, it should be excluded. This means that the corresponding residual column is formed by values indistinguishable from zero. In practice, it would be observed as an insignificant value of ${\varepsilon_j}^2$ in Eqn (1) applied to the dependence of a given descriptor on the set of preceding (residual) ones. To accept the zero hypothesis for the non-described part of a descriptor scale, the relation

$$ {\varepsilon_j}^2 < STD(H,T)SD(\varepsilon^2) = ST_{\mathrm{CR}} \qquad (11) $$

shall be satisfied, where by $STD(H,T)$ the Student criterion for $H = N_{<j} - 1$ degrees of freedom ($N_{<j}$ is the total number of descriptor scales with sequence numbers less than j) and by $SD(\varepsilon_{kj}^2)$ the standard deviation of ε^2 are denoted:

$$ SD(\varepsilon^2) = \sum_{k=1}^{j-1} SD(W_{kj}) = \sum_{k=1}^{j-1} SD(X_{0kj})R_{kj} \qquad (12) $$

where $SD(X_{0k,j})$ is the standard deviation of the corresponding scaled coefficient.

If the potential descriptor columns contain missing positions, the first step of the data processing is the calculation of the (mean) values for corresponding elements of respective descriptor columns making use of the existing data for the response columns. This procedure can be considered as the generalization of the traditional use of the secondary standard processes for the estimation of the missing values of substituent constants.[7] The procedure itself is reduced to the mean square solution of the (system of) equation(s) for a given data row with missing positions for descriptor columns (the regression of the row elements of known values with the regression coefficients for the columns involved considered as descriptors). If several such positions for a single row are presented, the missing values can be calculated by solving the respective system of equations for corresponding row. If possible, it is reasonable to select several subsets of descriptors to be used for the calculation of corresponding different sets of missing values. Further, it would be

highly reasonable to use for calculation the data from the set of more precisely described response columns. Restrictions may be set up to select the scales with high enough positive weight values and the acceptable level of mixed parts of scaled weights. The reliability of the result for a given row is characterized by the (multiple) correlation coefficient and the standard deviations for the missing values calculated, as well as by the numbers of data selected and those excluded as strongly deviating ones.

An analogous procedure enables one to attempt (iterative) re-estimation of all or of some part of the available values of the descriptors.[7]

TECHNIQUE OF DATA PROCESSING

The computer program SMIRC ('Selection of a Set of Minimally Interrelated Columns') created by us and reflected in our previous publications[3–5,7] was modified to permit the execution of the new visions of the data processing reported in this work. The MS Fortran Powerstation Development System for Windows 4.0 (1995) and the corresponding version of Fortran were used. The automation of as many data processing routes as possible was performed and several additional separate modules were written to allow the automatic inspection, extraction and secondary treatment of the different sets of results obtained in course of the execution of the main procedure used. The majority of the specific conditions and criteria can be specified by inserting corresponding initial data, without the necessity to correct the text of the program used.

A more detailed description of the modifications introduced into the algorithm of the multiparameter regression analysis and the software for its realization will be published separately. The process of adjusting the related programs for use by an ordinary user is in progress. After completion, this software will be prepared for distribution.

RESULTS AND DISCUSSION

Application of the approach described to solvent effects making use of EDSSE

The data compilation EDSSE contains 32 solvent parameter scales used earlier or presumably usable for a single or multiparameter description of solvent effects. Additionally, the data for 327 solvent-dependent processes (see Supplementary Material, Table S2) covering 45 solvents (see Table 1) were involved.[4] In our first publication on this topic,[7] by means of the first version of procedure SMIRC, we carried out a statistical treatment of matrices containing 32 solvent parameter scales. Our goal was an estimation of the minimum descriptor set with an acceptably low degree of mutual interrelation that would be sufficient for satisfactory multilinear description of the remaining columns included in the set under investigation. A sampling of nine descriptors was selected. In previous papers,[4,5] the composition of this set of descriptors was re-estimated by the application of a modified selection procedure to the whole compilation EDSSE. As initial approximation, the mentioned set of nine descriptors was used. Several supplementary descriptors—some additional solvent parameter scales and solvent-dependent processes—were tested. As the main criterion, the statistical significance of particular descriptors was used. Amongst the 12 potential descriptors, four have been found statistically insignificant and eight solvent scales remained in the final set of descriptors, significant and sufficient for statistical description of 347 solvent-dependent processes.[4,5] These are the scales that constitute the four-parameter model suggested by Koppel and Palm[10] and those included in the equation of Kamlet, Abboud and Taft.[11–14] The first set is represented by dipolarity $Y(1)$ and polarizability $P(2)$ scales as defined by Kirkwood[15] and the solvent Lewis basicity $B(7)$[16] and acidity $E(8)$[17] scales (numbers in parentheses correspond to the indexes of processes (rows) in Table S2). The Kamlet–Abboud–Taft scales were interpreted by the authors as the combined solvent dipolarity–polarizability $\pi^*(9)$,[11,12] hydrogen-bond acceptor (HBA) basicity $\beta(10)$[13] and hydrogen-bond donor (HBD) acidity $\alpha(11)$[14] scales. Additionally, the square of the Hildebrand solubility parameter, $\delta_H{}^2(12)$, is involved.[18,19]

These 347 different data series, considered as a set of responses, are represented by 128 series of data for UV–visible spectra, 49 for IR, 47 for NMR and four for ESR spectra. In addition to these spectral data, 36 series of kinetic data ($\log k$) and 45 of the equilibrium data ($\log K$) are involved. The last set includes seven series of the enthalpies of solution, nine series of the ion-transfer enthalpies and six series related to the distribution between the liquid and gas phases. For this data matrix, 6801 values (44.7%) are available and 8423 (55.3%) are missing. Only for 25 processes are data for 30 or more solvents available; 69 processes are represented in less than 15 solvents and 171 in more than 17 solvents. One can see that even in this relatively favorable case, the fraction of missing data is rather impressive.

Compared with our previous publications, a change of the sequence of descriptors has been made: the parameter $\delta_H{}^2$ was located at the last position (12), instead of (7). This is reasoned by acceptance of the preference of the principle of the physical nature for the possible interdependence of descriptor scales to the principle of the formally intrinsic approach to their definition. In this context, it would be natural that $\delta_H{}^2$ is dependent on the HBA basicity and HBD acidity of the solvents, rather than vice versa. Indeed, the free energy of formation of a cavity in the bulk solvent depends on its surface tension, for example, in the case of HBD/HBA solvents, obviously related to the formation of hydrogen bonds,

Table 1. Results of two different evaluations of averaged scaled solvent acidity parameter $E(8)$ values: E_{rsd} (initial approximation – original $E_T(30)$) and E_{itr} (initial approximation – E_{init}) calculated proceeding from the data for processes precisely ($R^* > 0.95$) described by the Koppel–Palm equation; scaling points: 0.0 for *n*-heptane (1) and 1.0 for water (23)

I[a]	Solvent	$E_{rsd} \pm SD$[b]	N_J[c]	N_{ef}[d]	E_{init}	$E_{itr} \pm SD$[b]	N_J[c]	N_{ef}[d]
1	*n*-Heptane	0.000	—	—	0.00	0.0	—	—
2	Cyclohexane	0.000	—	—	0.01	0.000 ± 0.007	—	—
3	Benzene	0.131 ± 0.000	185	77	0.01	0.155 ± 0.000	153	24
4	Toluene	0.271 ± 0.002	125	44	0.01	0.122 ± 0.002	101	10
5	Chlorobenzene	0.000 ± 0.001	106	24	0.01	0.000 ± 0.013	—	—
6	CCl_4	0.036 ± 0.000	155	70	0.01	0.000 ± 0.001	—	—
7	CH_2Cl_2	0.141 ± 0.001	161	76	0.10	0.167 ± 0.001	145	28
8	$CHCl_3$	0.267 ± 0.002	144	62	0.10	0.293 ± 0.002	129	23
9	$ClCH_2CH_2Cl$	0.000 ± 0.002	128	60	0.10	0.089 ± 0.001	113	21
10	$Cl_2C{=}CHCl$	0.025 ± 0.003	38	20	0.10	0.119 ± 0.005	36	7
11	Formamide	0.569 ± 0.001	58	23	0.50	0.499 ± 0.001	57	17
12	Fluorobenzene	0.005 ± 0.001	35	13	0.01	0.070 ± 0.001	22	4
13	DMF	0.000 ± 0.001	170	83	0.01	0.021 ± 0.000	151	33
14	CH_3CONMe_2	0.000 ± 0.002	92	44	0.01	0.009 ± 0.000	86	19
15	Cyclohexanone	0.116 ± 0.006	53	20	0.10	0.169 ± 0.003	47	4
16	HMPA	0.000 ± 0.003	70	35	0.01	0.000 ± 0.014	—	—
17	Acetone	0.000 ± 0.001	173	77	0.10	0.025 ± 0.000	159	29
18	Acetonitrile	0.089 ± 0.002	173	70	0.10	0.109 ± 0.000	144	29
19	Benzonitrile	0.000 ± 0.003	96	36	0.01	0.021 ± 0.002	75	5
20	MeOH	0.556 ± 0.002	168	71	0.50	0.572 ± 0.003	165	38
21	EtOH	0.448 ± 0.002	152	70	0.50	0.484 ± 0.001	151	44
22	*n*-BuOH	0.399 ± 0.003	112	55	0.50	0.448 ± 0.000	112	32
23	H_2O	1.0	—	—	1.00	1.0	—	—
24	*i*-PrOH	0.383 ± 0.007	122	59	0.50	0.402 ± 0.003	124	34
25	$HOCH_2CH_2OH$	0.691 ± 0.005	80	35	0.50	0.667 ± 0.005	73	15
26	Et_2O	0.000 ± 0.000	156	68	0.01	0.000 ± 0.004	—	—
27	$MeOCH_2CH_2OMe$	0.142 ± 0.002	25	12	0.01	0.000 ± 0.016	—	—
28	THF	0.000 ± 0.001	148	68	0.01	0.000 ± 0.005	—	—
29	1,4-Dioxane	0.177 ± 0.000	170	68	0.01	0.201 ± 0.000	137	32
30	Triethylamine	0.000 ± 0.011	—	—	0.01	0.000 ± 0.003	—	—
31	Pyridine	0.094 ± 0.001	—	—	0.01	0.073 ± 0.001	112	19
32	4-Methylpyridine	—	—	—	—	—	—	—
33	Acetophenone	0.000 ± 0.004	42	21	0.10	0.000 ± 0.011	—	—
34	Nitromethane	0.081 ± 0.004	123	39	0.10	0.104 ± 0.001	108	11
35	Acetic acid	0.987 ± 0.008	50	20	1.00	0.948 ± 0.005	47	14
36	DMSO	0.023 ± 0.001	172	83	0.10	0.041 ± 0.000	153	41
37	*t*-BuOH	0.330 ± 0.004	89	37	0.50	0.343 ± 0.003	90	11
38	Ethyl acetate	0.054 ± 0.001	131	66	0.01	0.072 ± 0.001	100	29
39	Methyl acetate	0.000 ± 0.03	43	14	0.01	0.000 ± 0.002	—	—
40	*n*-Bu_2O	0.000 ± 0.003	73	22	0.01	0.251 ± 0.005	50	6
41	*i*-Pr_2O	0.000 ± 0.013	—	—	0.01	0.000 ± 0.003	—	—
42	Anisole	0.000 ± 0.010	—	—	0.01	0.000 ± 0.017	—	—
43	Aniline	0.302 ± 0.005	24	10	0.50	0.354 ± 0.008	22	2
44	Nitrobenzene	0.000 ± 0.005	108	33	0.01	0.000 ± 0.009	—	—
45	CS_2	0.009 ± 0.001	56	19	0.01	0.122 ± 0.002	49	4

[a] Index of solvents.
[b] Standard deviations.
[c] Total numbers of *E*-dependent processes for given solvents.
[d] Numbers of precisely described dependent processes ($R^* > 0.95$) with sufficient contributions of *E* for a given solvents.

which are in turn determined by the acidic and basic behavior of solvents.

Before the modified treatment of data was performed, it was found reasonable to check once more the solvent acidity scale *E*. This scale was defined proceeding from $E_T(30)$ values[2] by means of the subtraction of the contributions described by the dipolarity *Y* and polarizability *P*. The residual *E*-scale obtained[1] was defined via the fraction not described by preceding descriptors. This fraction has a moderate value of 0.350 (in weight units). Therefore, one can suspect a considerable contribution of noise in the *E*-values obtained. From the point of view of the interpretation of residual *E*-values, one would expect that they equal zero for non-acidic solvents. If it is assumed that the hydrogen-bond acidity is the case, one may suspect that weak CH-acids may not be able to form hydrogen bonds. It would be desirable to derive their values in the framework of a more general approach,

especially in cases when non-zero values for solvents with presumably low or absent acidity appear. This the case for benzene, toluene, 1,4-dioxane and CS_2. It would be unreasonable to expect the presence of hydrogen-bond acidity for these compounds.

The calculations were accomplished proceeding from three different initial approximations. One of them was the original $E_T(30)$ values. Another approach was related to the rough qualitative estimation of the relative acidities assumed. For presumably inert solvents the value 0.01, for CH-acids 0.5 and for water 1.0 were assigned (see Table 1). The third version of the initial approximation was obtained by assigning zero values to all inert solvents. In all cases, the scaling based on the points for n-heptane (0.0) and water (1.0) was applied. In the case of the first two versions, for no solvent except *n*-heptane was a lack of acidity assumed from the very beginning. Additionally, for all three versions, it was assumed that negative values for the residual acidity have no sense. Therefore, a lower limit of zero was defined for these values. If in the initial approximation or for some intermediate steps of the iterative procedure for some solvent a zero value had been assigned, its (further) recalculation was avoided. Using this approach, in the case of the first two versions, zero values for the residual acidity may appear as a specific result of the data processing only. For the third version, all the initial zero values were retained during the whole procedure of data processing.

For the calculations, only the data for processes which are more precisely ($R^* > 0.95$) described by the four-parameter equation (descriptors *Y*, *P* and *B* and the scale assumed to be related to acidity—the one to be re-calculated) and due to the high enough weight contribution of the scale subjected to re-estimation were used. A threshold of 20% of the mean weight contribution of the acidity scale over all processes involved was defined. For all solvents the correlation coefficients for the corresponding data row processed are nearly 0.99 or exceed this value. The correlation coefficients between the initial approximations of *E* and the primary (non-residual) results are 0.961, 0.947, and 0.978, respectively.

The set of residual acidity constants E_{rsd} for the first initial approximation was obtained on the basis of the dependence of re-calculated $E_T(30)$ values on preceding descriptors for presumably inert or low acidity solvents, as was done for the original definition of *E* values.[17] In the case of the second initial approximation, the recalculated acidity values are not significantly dependent on preceding descriptors. Therefore, the recalculated values represent for this case already a sufficiently orthogonal scale, equivalent to the E_{rsd} values. For the third initial approximation, a significant dependence on *Y*, *P* and *B* values appeared and the corresponding contributions were subtracted. One can hardly understand the contribution of basicity *B* to acidity and this may be related to the voluntary assignment of zero values for all presumably inert solvents.

The resulting E_{rsd} values for these three versions are mutually linearly related with the correlation coefficients exceeding 0.968. One can conclude that all of these versions lead eventually to the values of E_{rsd}, which indeed represent a common solvent property. The correlation coefficient with the original *E* values is less than 0.93 for all versions. This is a sign of unsatisfactory agreement.

All three versions considered lead to close results for the multiple regressions with all eight descriptor scales cited above.

The scaled values of E_{rsd} obtained by proceeding from the first two initial approximations are listed in Table 1. The 12 zero values obtained for second (arbitrary) initial approximation are related to non-acidic solvents, which is in accordance with the interpretation of the nature of E_{rsd} values. The low values (<0.1) for some solvents, expected to lack acidity, have comparably smaller (positions 12, 36 and 38) or zero (positions 13, 14 and 19) values for the first version of data processing. One can conclude that for these solvents the zero or very low acidity can be considered as confirmed. For solvents 40 (n-Bu_2O) and 45 (CS_2) the considerably high values are obtained using the second approach. However, for the first one, the corresponding figures are 0.000 and 0.009, respectively. Hence, for 41 solvents from the set treated, the results of the estimation of E_{rsd} values can be considered reasonable.

However, there remain three solvents, benzene, toluene and 1,4-dioxane, characterized in all cases by considerably high values of E_{rsd}, although it is difficult to understand the reason for their acidity. We have no reasonable explanation of the nature of this 'abnormality'.

If a lower limit for E_{rsd} of zero is not introduced, for 12 solvents negative values appear. For the second (arbitrary) initial approximation, the absolute values of all negative values are less than -0.1. Proceeding from $E_T(30)$ (first approximation), absolute values of negative figures are higher than 0.1 for cyclohexanone (-0.12), Et_3N (-0.15) and i-Pr_2O (-0.603). This may be understood in the light of the presence in this case of the additional step of subtraction of the contribution of dipolarity *Y*. As already mentioned above, the primary result obtained from the second version does not depend significantly on preceding descriptors. Hence, one can conclude that the introduction of a lower limit of zero does not essentially influence the results and it may be accepted as a justified procedure.

The conclusion is that all NH- and OH-acids exert considerable acidity. For all but three CH-acids (acetone, acetophenone and methyl acetate), the E_{rsd} values exceed zero, exhibiting low or medium acidity.

When compared with our preceding studies[3–6] devoted to the data processing applied to EDSSE, the re-estimation of the 19 missing values for $\pi^*(9)$, $\beta(10)$, $\alpha(11)$ and $\delta_H^2(12)$ have been modified. The missing values of δ_H^2 were calculated by proceeding from data series described

Table 2. Re-estimated missing values *A* for solvent parameter scales (descriptors) $\pi^*(9)$, $\beta(10)$, and $\delta_H{}^2(12)$

L[a]	I[b]	Solvent	A	SD[c]	N_J[d]	N_{ef}[e]
9	43	Aniline	1.43	0.06	38	8
9	45	CS_2	0.55	0.02	87	23
10	7	CH_2Cl_2	0.20	0.002	209	64
10	11	Formamide	0.49	0.00	25	2
10	12	Fluorobenzene	0.093	0.01	24	3
10	34	Nitromethane	0.17	0.07	70	3
10	35	Acetic acid	0.60	0.0	32	2
10	43	Aniline	0.76	0.02	38	8
10	45	CS_2	0.40	0.04	87	23
11	7	CH_2Cl_2	0.11	0.00	209	64
11	15	Cyclohexanone	(0.007)[f]	0.003	42	6
11	33	Acetophenone	−0.068	0.005	33	6
11	43	Aniline	−0.22	0.06	38	8
11	45	CS_2	(0.01)[f]	0.02	87	23
12	12	Fluorobenzene	21.4	0.42	20	7
12	27	$MeOCH_2CH_2OMe$	60.3	0.95	15	5
12	30	Triethylamine	−180.0	1.5	33	7
12	40	n-Bu_2O	−78.0	1.2	30	7
12	42	Anisole	9.4	0.5	36	5

[a] Index of descriptors from Table S2 in the Supplementary Material.
[b] Index of solvents.
[c] Standard deviations for *A*.
[d] Total numbers of processes dependent on corresponding descriptors for a given solvent.
[e] Numbers of precisely described processes ($R^* > 0.95$) with sufficient contributions of corresponding descriptors for a given solvent.
[f] Values in parentheses are taken as zero.

precisely enough ($R^* > 0.95$) by the five-parameter equation[20] defined by descriptors *Y, P, B, E* and δ_H^2. The use of any other version of E_{rsd} instead of the original *E* does not lead to a significant change of the resulting values of $\delta_H{}^2$. The missing values of π^*, α and β were calculated using the data for series which are described by the Kamlet–Abboud–Taft equation with precision $R^* > 0.95$. For the calculations, only the data for processes represented by high enough weight contributions for the scale subjected to estimation (at least 20% of the mean weight contribution of acidity scale over all processes involved) were used. These calculations were repeated for several risk levels (0.1, 0.05, 0.025 and 0.01) to estimate their stability and mean values. A joint *F*-test of statistical significance, which confirmed the significance of all 19 estimated missing values, was performed. The results are represented in Table 2. One can see that for 14 solvents the restrictions introduced lead to a small number (< 10) of data series usable for calculations. Nevertheless, these averaged results have to be preferred to the estimation based on a single secondary standard process.

With the aim of making a choice between different estimations of E_{rsd}, the statistical treatment of data for 347 solvent-dependent processes was carried out with the set of eight descriptors mentioned above, containing different versions of E_{rsd} instead of original *E*. The results of statistical processing are close enough for different E_{rsd}. The mean number of data series with a higher level of description ($R^* > 0.95$) is 241 ± 3 and the mean root mean square value of multiple correlation coefficient for all processes is 0.944 ± 0.001. Therefore, only the results for the first version of E_{rsd} [$E_T(30)$ as the initial approximation] are discussed below.

For the re-estimated descriptor scales, the procedure of subtracting of the parts described by the preceding more 'fundamental' descriptors was executed. The weight contributions W_{jl} of preceding (residual) descriptors to the description of the initial non-purified ones are given in Table 3. The conclusion is derived that more 'fundamental' descriptors $Y(1)$, $P(2)$ and $B(7)$ and re-estimated $E_{rsd}(8)$ are practically independent of the preceding ones. For the remaining descriptors, $\pi(9)$, $\beta(10)$, $\alpha(11)$ and $\delta_H{}^2(12)$, more than 75% of the dispersion depends on the contributions of preceding descriptors. Moreover, for $\alpha(11)$ the criterion ST_{CR} exceeds the value of the non-described part $\varepsilon_j{}^2$ of dispersion (see Table 3) and the zero hypotheses should be accepted for it. This means that the descriptor $\alpha_{rsd}(11)$ has to be eliminated as exhaustively described by the preceding ones.

For an additional check of this decision, the corrected *F*-test of statistical significance was applied to data averaged over 20 independent sets of randomly assigned figures, substituted for the descriptor scale $\alpha_{rsd}(11)$, which was moved to the last position in the sequence of descriptors. For the run of these random sets, the mean values of weights and the mean number of significant contributions for the scale tested are equal to 0.0086 ± 0.0031 and 57 ± 13 (0.0036 ± 0.0017 and

Table 3. Weight contributions of preceding (residual) descriptors to description of subsequent ones

Subsequent descriptor	Preceding descriptor								ε^2[a]	ST_{CR}[b]
	L	$Y(1)$	$P(2)$	$B(7)$	$E_{rsd}(8)$	$\pi^*_{rsd}(9)$	$\beta_{rsd}(10)$	$\alpha_{rsd}(11)$		
P	2	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.975	0.051
B	7	0.086	0.000	0.000	0.000	0.000	0.000	0.000	0.914	0.092
E	8	0.076	0.094	0.000	0.000	0.000	0.000	0.000	0.830	0.136
π^*	9	0.534	0.184	0.000	0.045	0.000	0.000	0.000	0.238	0.145
β	10	0.145	0.000	0.543	0.000	0.075	0.000	0.000	0.236	0.165
α	11	0.064	0.103	0.000	0.715	0.073	0.000	0.000	0.045	0.091
$\delta_H{}^2$	12	0.248	0.000	0.023	0.393	0.025	0.010	0.000	0.301	0.184

[a] Undescribed parts of dispersions for corresponding descriptors.
[b] Student criteria for accepting the zero hypthesis for the non-described part of a descriptor scale.

40 ± 9 for the set with $R^* > 0.95$). For α_{rsd}, the corresponding values are 0.0208 and 128 (0.0156 and 101 for the set with $R^* > 0.95$). These data, together with the corrected mean values of the difference of the degrees of freedom (0.25 ± 0.0025), led to $F = 5.78 \pm 1.32$. The values of the F-criterias are $F(0.10, \nu_1, \nu_2) = 3.52$ and $F(0.05, \nu_1, \nu_2) = 6.08$. Consequently, at the risk level of 0.05, the α-scale has to be recognized as slightly insignificant, but a lower estimation of the F-value (4.46) exceeds the F-criterion for a risk level of 0.1. This situation leaves some ambiguity regarding the conclusion about the significance of this descriptor.

The combination of both criterias leads to the conclusion that the α_{rsd} scale has to be eliminated from the model.

The values of elements of eight initial and seven residual descriptors are listed in Table S3 in the Supplementary Material.

The procedures for multiple linear regressions with the obtained sets of eight initial and seven residual descriptors were executed for each of the 347 dependent columns separately. For selection of the significant descriptors according to the F-test at the risk level of 0.05, the mode of scanning of all possible regressions was used. With the aim of demonstrating the real details of situation, no restrictions were applied for the elimination of the descriptor scales due to the unacceptable magnitude of OE or the appearance of superfluous 'mixed' parts of weight contributions of descriptors. For the criteria mentioned above, the following values were assigned: $R^*_{crt} = 0.95, M_{exmax} = 3$ and $H_{min} = 4$ (for the case when all descriptors from their initial set are significant).

In Table 4, the mean weight contributions $\bar{W}$ of eight initial and seven residual descriptors for the whole selection of solvent-dependent processes are presented. The data for the residual descriptors show the 'true' picture of the relative contributions made by different descriptors—three scales, $Y(1)$, $B(7)$ and $E_{rsd}(8)$, constitute the major part (77%) of the total effect. However, proceeding from the results of the F-test performed in our previous work,[4] the others are unquestionably also significant.

The large data compilations inevitably contain a considerable amount of erroneous values. In the case of the absence of several independent experimental estimations of values used, the reasons for the large differences between experimental values and those calculated making use of the model under consideration could not be explicitly detected. This is related to the problem of the characterization of the predictive power of this model. The applicability of procedure called 'cross-validation' (CV)[21,22] for the characterization of the stability and/or the predictive power of the respective regression equation was tested. The procedure for calculation of the cross-validation scaled standard deviations S_{0cv} and the squared determination coefficients R_{cv}^2 for each data series was slightly modified. As the regression characteristics for the particular data sets are calculated taking into account the numbers of statistical degrees of freedom, this should be done also for the CV characteristics. Therefore, the S_{0cv}^2 and R_{cv}^2 values are calculated as

$$S_{0cv}^2 = \sum \Delta^2 / H \quad \text{and} \quad R_{cv}^2 = 1 - S_{0cv}^2$$

where $\sum \Delta^2$ and H are the sum of the squares of deviations and the number of statistical degrees of freedom, respectively.

Because the relation between S_{0cv}^2 and S_0^2 values is equivalent to that between R_{cv}^2 and R^2 values, only the first pair of these characteristics will be discussed. The difference $D_{cv} = S_0^2 - S_{0cv}^2$ represents the effect of the omission of data rows, one by one, on the magnitude of the sum of scaled squares of residuals. The higher this value (believed to be positive), the lower is the predictive power.

The D_{cv} values for all particular data series and the corresponding mean values over all processes and for more and less precisely described subsets (with $R^* > 0.95$ and $R^* < 0.95$) were calculated. The absolute values of D_{cv} appeared to be very low and for 252 processes negative values were obtained. The average figures obtained are given in Table 5.

Table 4. Mean weight contributions of the initial and residual descriptor scales to description of solvent-dependent processes

		Initial descriptor scales		(Residual) descriptor scales	
L	Descriptor	$\overline{W} \pm SD_w$ [a]	N_{sg} [b]	$\overline{W} \pm SD_w$ [a]	N_{sg} [b]
1	Y	0.080 ± 0.020	94	0.418 ± 0.094	283
2	P	0.050 ± 0.008	171	0.067 ± 0.010	205
7	B	0.128 ± 0.026	148	0.169 ± 0.032	165
8	E	0.048 ± 0.009	69	0.184 ± 0.035	225
9	π^*	0.350 ± 0.080	247	0.091 ± 0.035	227
10	β	0.095 ± 0.019	138	0.021 ± 0.003	117
11	α	0.147 ± 0.030	171	—	
12	δ_H^2	0.103 ± 0.024	145	0.050 ± 0.009	136

[a] Mean weight contributions of descriptors with their standard deviations.
[b] Numbers of dependent processes with significant contributions of corresponding descriptors.

Table 5. Average characteristics of cross-validation procedure

	N [a]	$\bar{S}_0^{2}$ [b]	$\bar{S}_{0cv}^{2}$ [c]	$\lvert\bar{D}_{cv}\rvert$ [d]	$\bar{D}_{cv} \pm SD$ [e]
All series	347	0.3522	0.3539	0.0005	-0.0005 ± 0.0021
$R > 0.95$	205	0.1994	0.1999	0.0002	-0.0002 ± 0.0008
$R < 0.95$	142	0.5163	0.5173	0.0010	-0.0010 ± 0.0023

[a] Numbers of dependent processes of given kind.
[b] Mean scaled standard deviations.
[c] Mean cross-validation scaled standard deviations.
[d] Absolute values of mean characteristics of the omission of data rows $D_{cv} = S_0^2 - S_{0cv}^2$.
[e] Mean values of D_{cv} with their standard deviations.

Table 6. Average results of statistical processing of data for 347 solvent-dependent processes with set of seven (residual) descriptors

	N[b]		$\bar{R}^*$[c]			
K[a]	$R^* > 0.95$	$R^* < 0.95$	$R^* > 0.95$	$R^* < 0.95$	All	M_{excl}[d]
0	205 (59.1%)	142 (40.9%)	0.980	0.860	0.933	0
1	258 (74.4%)	89 (25.6%)	0.978	0.869	0.951	141 (2.1%)
2	292 (84.1%)	55 (15.9%)	0.976	0.865	0.959	229 (3.3%)
3	314 (90.5%)	33 (9.5%)	0.975	0.852	0.964	279 (4.1%)

[a] Index of cycles of elimination of strongly deviating points.
[b] Numbers of dependent processes of given kind.
[c] Root mean square values of corrected multiple correlation coefficients.
[d] Numbers of excluded points.

Proceeding from these results, it is difficult to derive any meaningful conclusion related to $S_{0cv}{}^2$ values. Especially remarkable is the equality (within the range defined by *SD* values) of the mean D_{cv} to zero. This means that at least in the case of the data processed, no additional information can be gained from the cross-validation procedure. Therefore, more explicit ways for the characterization of the effect caused by strongly deviating points and the prognostic ability of the model specified have to be selected. The preferred procedure for the elimination of strongly deviating points has been discussed before.

Average results of statistical processing of data for 347 solvent-dependent processes with set of seven (residual) descriptors before the elimination of strongly deviating points and after three cycles of elimination are given in Table 6. During the cycle, from each roughly described data series with $R^* < 0.95$ a single point representing the maximum deviation was excluded. This procedure led to the exclusion of a maximum of three points from a single series. If the total number of points for a given data series became less than 12, further exclusion of points was avoided.

After exclusion of 279 (4.1%) points the amount of roughly described ($R^* < 0.95$) processes diminished from 142 (40.9%) to 33 (9.5%) and the average precision of the description of the whole set of solvent-dependent processes was characterized by the value $\bar{R}^*_{all} = 0.964$ (0.975 for $R^* > 0.95$). The indexes of excluded points for each data series are reflected in Table S2 in the Supplementary Material. The entire results of statistical treatment of the data for the 347 solvent-dependent processes, containing the intercepts A_0, regression coefficients C_l with their standard deviations, weight contributions W_l of corresponding descriptors with their total 'mixed' terms TMT_l, numbers of statistical degrees of freedom, corrected multiple correlation coefficients R^*, squares of determination coefficients R^2 and squares of scaled standard deviations $S_0{}^2$, reflecting non-described parts of the dispersions of processes, are listed in Table S4 in the Supplementary Material.

One can realize that the maximum number of elimination cycles used, and also the minimal yet acceptable number of degrees of freedom are 'reasonably' selected conventional numbers. If the limit of acceptable number of points eliminated from a single data series is removed, the total number of points excluded rises to 330 (4.8%) and the number of roughly described series falls to 15 (4.32%). After practically removing of the lower limit of the degrees of freedom, (a single degree of freedom for a current state of solution is required), altogether 361 (5.3%) points were excluded and for 22 series this figure exceeds 3 (up to 10). After that, the number of more precisely described series rises to 347 (100%) and $\bar{R}^*_{all} = 0.975$. One can consider this result as a limit of the rise of precision of the description of the data matrix processed making use of the cited procedure of elimination of strongly deviating points. As the percentage of excluded points remains low, they may be considered as the erroneous data.

In Table 7, the distribution of the processes according to the precision of description before and after exclusion of no more than three strongly deviating points and corresponding mean values RT_{mean} of the ratio of numbers of statistical degrees of freedom to the numbers of equation parameters are summarized. The latter demonstrate the availability of sufficient numbers of statistical degrees of freedom.

The most general characteristics of the orthogonality and the quality of description are defined by Eqns (2)–(4).

Table 7. Distribution of solvent-dependent processes (%) over the ranges of values of corrected correlation coefficients R^*

	M_{exmax}[a] = 0		M_{exmax}[a] = 3	
Range of R^*	N	RT_{mean}[b]	N	RT_{mean}[b]
> 0.99	16.1	2.2	17.0	2.2
0.98–0.99	16.4	3.5	19.3	3.3
0.97–0.98	13.0	3.4	18.4	3.2
0.95–0.97	13.5	3.4	35.7	3.3
0.90–0.95	18.4	4.0	5.8	3.3
< 0.90	22.5	4.9	3.7	5.2

[a] Maximum number of points excluded for individual processes.
[b] Average ratio of numbers of statistical degrees of freedom to numbers of significant descriptors for individual processes.

Table 8. General characteristics of orthogonality, non-orthogonality and the quality of description

						$\bar{O}_{rth}$[g]			$\bar{G}_d$[h]			$\bar{G}_d^*$[i]		
Set[a]	N_{dsc}[b]	R_{eff}[c]	D_{et}[d]	O_{rth}[e]	N_{onoth}[f]	$R^*>0.95$	$R^*<0.95$	All	$R^*>0.95$	$R^*<0.95$	All	$R^*>0.95$	$R^*<0.95$	All
Initial	8	0.071	0.001	0.503	0.497	0.883	0.874	0.897	0.850	0.722	0.803	0.842	0.684	0.784
Initial	7	0.058	0.028	0.664	0.336	0.709	0.667	0.760	0.646	0.617	0.635	0.640	0.587	0.620
Residual	8	0.021	0.484	0.957	0.043	0.883	0.874	0.897	0.850	0.722	0.803	0.842	0.684	0.784
Residual	7	0.023	0.490	0.945	0.055	0.870	0.851	0.900	0.825	0.773	0.781	0.817	0.676	0.762
Residual (279)[j]	7	0.024	0.490	0.945	0.055	0.866	0.941	0.873	0.832	0.710	0.820	0.822	0.676	0.808
Residual (361)[j]	7	0.024	0.490	0.945	0.055	0.862	—	0.862	0.828	—	0.828	0.818	—	0.818

[a] Set of descriptors.
[b] Total number of descriptors used.
[c] Effective value of the correlation coefficient of correlation matrix.
[d] Value of determinant for corresponding set of descriptors.
[e] Measure of orthogonality for corresponding set of descriptors; see Eqn (3).
[f] Measure of non-orthogonality for corresponding set of descriptors; see Eqn (2).
[g] Mean measure of orthogonality.
[h] Mean measure of quality of description; see Eqn (4).
[i] Mean measure of corrected quality of description.
[j] Number of points excluded.

The values of N_{onorth} and O_{rth} are applicable to the correlation matrices related to the whole set of potential descriptors, and also to the sets of the significant ones for the particular processes. The values of G_d and G_d^* (the asterisk refers to the use of R^* instead of R) can be calculated for particular processes only. The mean values over all processes and for more and less precisely described subsets (with $R^*>0.95$ and $R^*<0.95$) can be calculated for them. The last is true for N_{onorth} and O_{rth} also. Corresponding results are represented in Table 8. The dependence of the characteristics listed on the number of points excluded can also be inspected. The last row in Table 8 is related to the limiting case when all points causing $R^*<0.95$ situations are removed. Therefore, two characteristics, $G_d = 0.828$ and the number of excluded points, equal to 361 (5.27%), characterize the upper limit of the quality of description, reachable for $R_{crit} = 0.95$.

The distribution of solvent-dependent processes over the ranges of spur values Q_j characterizing the extent of OE is shown in Table 9. In comparison with more orthogonal residual descriptors, for initial descriptors a sufficiently higher OE is observed: 19.3% of processes have $Q_j > 1$ for initial descriptors, but only 2.6% for residual ones.

Table 9. Distribution of solvent-dependent processes (%) over intervals of spur values Q_j

Range of Q_j	For initial descriptors			For residual descriptors		
	$R^*>0.95$	$R^*<0.95$	All	$R^*>0.95$	$R^*<0.95$	All
<1.0	72.8	93.8	80.7	95.6	100.0	97.4
1.0–2.0	18.9	5.4	13.9	2.9	0.0	1.7
2.0–3.0	4.6	0.8	3.2	0.5	0.0	0.3
3.0–5.0	2.8	0.0	1.7	1.0	0.0	0.6
5.0–10.0	0.9	0.0	0.6	0.0	0.0	0.0
10.0–50.0	0.0	0.0	0.0	0.0	0.0	0.0

In Table 10, the distribution of the numbers of contributions of significant descriptors for all solvent-dependent processes over the ranges of absolute values of scaled 'mixed' terms $|TMT_{jl}|$ and corresponding mean weights of these descriptors $\overline{W}$ are presented. This distribution is practically independent of the precision of description of the processes. The fraction of contributions of significant descriptors, characterized by $|TMT_{jl}|>1$ is equal to 25.5%. Figure 1 shows the dependence of the mean weight contributions of significant descriptors on the values of their scaled 'mixed' terms. It is obvious that the magnitudes of 'mixed' terms have a tendency to decrease when the mean weight contributions of significant descriptors increase. At the same time,

Table 10. Distribution of the numbers of contributions (%) of significant descriptors for all solvent-dependent processes over the ranges of absolute values of scaled 'mixed' terms $|TMT_{jl}|$ and corresponding mean weights $\bar{W}$

Range of $|TMT_{jl}|$	M_{exmax}[a] = 0		M_{exmax}[a] = 3	
	F[b]	$\bar{W}$	F[b]	$\bar{W}$
0.0–0.5	60.10	0.3700	58.9	0.3767
0.5–1.0	14.9	0.1581	15.6	0.1620
1.0–2.0	7.4	0.0957	8.4	0.0979
2.0–5.0	11.4	0.0420	10.9	0.0499
5.0–10.0	3.5	0.0308	3.6	0.0266
10.0–25.0	1.6	0.0082	1.6	0.0105
25.0–50.0	0.6	0.0040	0.5	0.0093
50.0–100.0	0.3	0.0089	0.2	0.0015
100.0–250.0	0.3	0.0008	0.4	0.0008
>250.0	0.0	0.0000	0.0	0.0000

[a] Maximum number of points excluded for particular processes.
[b] Fractions (%) of the total number of contributions of significant descriptors.

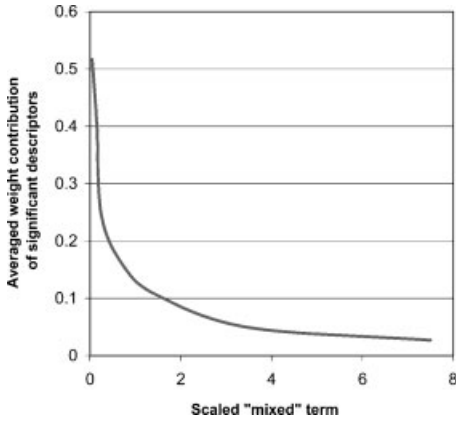

Figure 1. Dependence of the average weight contributions of significant descriptors on the values of their scaled 'mixed' terms

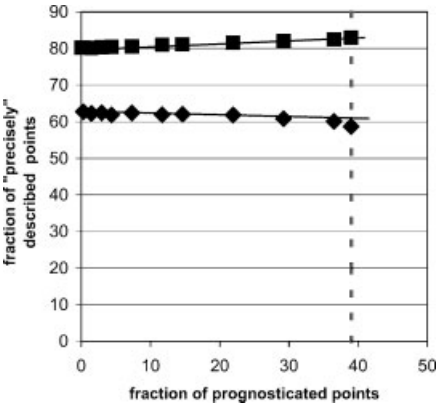

Figure 2. Dependence of the fraction (%) of 'precisely' described points on the fraction of prognosticated points. ■, Fraction of 'precisely' described processed points; ◆, Fraction of 'precisely' described prognosticated points. Maximum possible fraction of prognosticated points is marked by the dashed line

almost all very small weight contributions have values of $|TMT_{jl}| > 1$. The mean value of $|TMT_{jl}|$ is 1.88 for the set of all processes and 1.30 for that of the more precisely described processes. This can be considered as a meaningful warning about the presence of a number of the unacceptably high values of the 'mixed' parts of weights.

Recently, the specific QSPR approach related to the CODESSA procedure and the principal component (factor) analysis (FA) was applied to a set of 46 solvent property scales defined for 65 solvents.[9,23] By means of the CODESSA procedure, the calculated values were substituted for the vacant positions (missing data) of this data matrix. As result, a data matrix of 40 rows and 40 columns with all positions filled was obtained. This matrix was then processed making use of the FA procedure to determine the proper number of meaningful factors. However, objections can be raised with respect to the use of both of these procedures.

The main problem with the CODESSA procedure is the use of a highly abundant initial number of descriptors (hundreds). Such a selection is inevitably highly interrelated (completely non-orthogonal), which can be proved by treatment of a compilation of random values of that size. For each particular data set subjected to description, a small number of suitable descriptors is selected. Although the descriptors used are called 'theoretical', this only means that their elements could be defined mostly by pure calculations without any essential relation to the nature of the data set correlated. Hence, in the paper cited above,[9] for the correlation of 45 solvent scales altogether 106 different descriptor scales are used; 85 of them are applied only to a single solvent scale, 17 to two and only five to three or more (up to eight) cases. This means that in the majority of cases different sets of descriptors are involved. Therefore, the calculated missing values in the data matrix under consideration many turn out irrelevant.

Although the classical FA is equipped with a number of characteristics, they all have a smooth dependence on the number of factors employed. This problem was discussed in a previous paper[7] and the rise of the scaled standard deviation S_0 caused by the introduction of a subsequent factor was suggested to be a sign of the unacceptability of its introduction. This rise of S_0 is a result of the unacceptable decrease in the number of statistical degrees of freedom as a consequence of the addition of the subsequent factor. Although it is not for general use, this procedure enables one to draw the conclusion that the selection of the number of significant factors[23] is arbitrarily underestimated. A more detailed consideration of related problems is desirable and this will be the subject of a separate paper.

With the procedure considered in the theoretical part of this paper, the prognostic ability of the model, based on seven (residual) descriptors, was tested for the data compilation under investigation. This set contains 6801 available values. If the minimum acceptable value for the degrees of freedom $H_{\min} = 4$ is adopted, a maximum of 2668 (39%) data points can be used for the check of the prognostic ability. For every level of the numbers of points used for the prognosis, 20 parallel runs were executed. Figure 2 shows the dependence of the fraction of 'precisely' described points ($D_{0ji} < 0.3123$) for processed and prognosticated selections of data on the fraction of prognosticated points. The conclusion can be derived that this fraction is practically independent of the fraction of prognosticated points and equals approximately 80% for the processed points and 60% for the prognosticated points. The same is true for the mean scaled standard deviations $\bar{D}_0$ for 'precisely' described points. This figure equals 0.167 ± 0.010 for 100 prognosticated points and 0.174 ± 0.002 for 2668 points. Such a stability of results indicates a satisfactory prognostic ability of the model.

CONCLUSIONS

The results reported in this work confirm that the application of our procedure allows successful concerted multiparameter processing of large data sets. The missing positions of the descriptors can be filled and, when required, some of the initial descriptor scales can be either recalculated or (re)built, taking into account the considerations about the essential physico-chemical meaning of the particular descriptors. The orthogonalization of the correlation matrix via substitution of the residual descriptors for more or less linearly interrelated initial ones was demonstrated. A combined measure of the quality of description was introduced and used. The weights were used for the correct comparison of the contributions of different descriptors to the dispersion of the particular data column. The root mean square values of the scaled weights of particular significant descriptors and the corresponding numbers of data series for which they are significant were used as the characteristics of their relative importance for the description of the whole data matrix processed.

A method of characterization of the checked prognostic ability of the multilinear model was developed and used. The application of this method to a large variety of solvent-dependent data sets showed that the results are highly stable with respect to the increase in the percentage of points transformed into prognosticated ones.

At the same time, even the use of the set of residual descriptor scales does not prevent the appearance of a number of obviously too high contributions of the 'mixed' parts to the values of weight contributions of different descriptors for a number of processes. Even in the case when the upper limit of the quality of description is reached, for 365 (26.7%) weight values amongst a total of 1368, the $|TMT_{jl}|$ values are higher than 1.0. This phenomenon is the inevitable consequence of the presence of a number of missing positions in the matrix of the response columns.

As regards the essential meaning of the seven descriptors remained, the interpretation of *Y*, *P*, *B* and E_{rsd} as measures of dipolarity, polarizability, basicity and acidity of the solvents can be considered as well established. π^*_{rsd} has been considered to be likely a measure of the ability for the dipolar solvation of cationic centers.[5] The physicochemical meaning of β_{rsd} remains uncertain. Nevertheless, despite its low mean weight contribution (0.021 ± 0.003), the presence of this residual descriptor in the general model is once again confirmed.

Acknowledgement

Support of this work by the Estonian Science Foundation, grant No. 4590, is gratefully acknowledged.

REFERENCES

1. Palm VA. *Osnovy Kolichestvennoi Teorii Organicheskikh Reaktsii (Foundations of Quantitative Theory of Organic Reactions)* (2nd edn). Khimiya: Leningrad, 1977; *Grundlagen der Quantitativen Theorie Organischer Reactionen* (translated into German by Heublein G). Akademie Verlag: Berlin, 1971.
2. Reichardt C. *Solvents and Solvent Effects in Organic Chemistry* (3rd edn). Wiley-VCH: Weinheim, 2002.
3. Palm V, Palm N. *Org. React. (Tartu)* 1997; **31**: 141–158; *Chem. Abstr.* 1998; **128**: 47875r.
4. Palm N, Palm V. *Russ. J. Org. Chem.* 2000; **36**: 1075–1104.
5. Palm N, Palm V. *Russ. J. Org. Chem.* 2001; **37**: 935–939.
6. Seber GAF. *Linear Regression Analysis*. Wiley: New York.
7. Palm V, Palm N. *Org. React. (Tartu)* 1993; **28**: 125–150; *Chem. Abstr.* 1994; **121**: 56706u.
8. Bennett CA, Franklin NL. *Statistical Analysis in Chemistry and the Chemical Industry*. Wiley: New York, 1980.
9. Katritzky AR, Tamm T, Wang Y, Sild S, Karelson M. *J. Chem. Inf. Comput. Sci.* 1999; **39**: 684–691.
10. Koppel IA, Palm VA. In *Advances in Linear Free Energy Relationship*, Chapman NB, Shorter J (eds). Plenum Press: London, 1972; 203–280.
11. Kamlet MJ, Abboud JLM, Taft RW. *J. Am. Chem. Soc*. 1977; **99**: 6027–6038.
12. Kamlet MJ, Hall TN, Boykin J, Taft RW. *J. Org. Chem*. 1979; **44**: 2599–2604.
13. Kamlet MJ, Taft RW. *J. Am. Chem. Soc*. 1976; **98**: 377–383.
14. Kamlet MJ, Taft RW. *J. Am. Chem. Soc*. 1976; **98**: 2886–2894.
15. Kirkwood J. *J. Chem. Phys*. 1934; **2**: 351.
16. Koppel IA, Paju AI. *Org. React. (Tartu)* 1974; **11**: 121–136; *Chem. Abstr.* 1975; **82**: 42805q.
17. Koppel IA, Paju AI. *Org. React. (Tartu)* 1974; **11**: 137–140; *Chem. Abstr.* 1975; **82**: 42806r.
18. Herbrandson HF, Neufeld FR. *J. Org. Chem*. 1966; **31**: 1140–1143.
19. Barton AFM. *Chem. Rev*. 1975; **75**: 731–753.
20. Makitra RG, Pirig YN. *Zh. Obshch. Khim*. 1986; **56**: 657–665; *Chem. Abstr.* 1986; **104**: 231587d.
21. Stone M, Jonathan P. *J. Chemom*. 1993; **7**: 455–475.
22. Karelson M. *Molecular Descriptors in QSAR/QSPR*. Wiley–Interscience: New York, 2000; 394.
23. Katritzky AR, Tamm T, Wang Y, Karelson M. *J. Chem. Inf. Comput. Sci.* 1999; **39**: 692–698.