Russian Journal of Organic Chemistry, Vol. 37, No. 7, 2001, pp. 935–939. Translated from Zhurnal Organicheskoi Khimii, Vol. 37, No. 7, 2001, pp. 987–992.

Verification of the Modified Approach to Solvation Effects on the Basis of Extended Data Set: III.* Results of Statistical Data Treatment for Particular Processes**

N. Pal'm and V. Pal'm

Institute of Chemical Physics, Tartu University, ul. Yakobi 2, Tartu, EE 51014 Estonia
e-mail:nataly@chem.ut.ee

Received August 2, 2000

Abstract—Detailed results of statistical data processing are given for 346 processes sensitive to solvent effects using previously proposed set of eight (residual) solvent constant scales as descriptors. Physicochemical sense of these descriptors is discussed. Although the set of descriptors used is almost orthogonal, in some processes with incomplete samples of data the weight contributions of descriptors include "mixed" components, depending on the presence of more than one regressor in the model.

In the preceding communication [1] on this topic we described a procedure for the refinement of the previously proposed minimal set of descriptor scales [2], which is sufficient for satisfactory multilinear description of the remaining columns in a data sample for 359 processes and solvent parameter scales (columns) in 45 solvents. The complete list of processes and solvents and relevant references are given in [1], and the present communication follows the same numbering of processes and solvents. The initial set included 9 descriptors which were mainly parameters of two alternative models, Koppel'–Pal'm [3] and Kamlet–Abboud–Taft [4]: polarity parameter *Y*, polarizability parameter *P*, molar refraction *MR*, squared Hildebrandt solubility parameter δ_H^2, basicity and acidity parameters *B* and *E*, and Kamlet–Abboud–Taft parameters π^*, β, and α, which were subjected to the "refinement" procedure by substracting the contributions depending on scales with lower serial numbers. The above set ensures considerably higher accuracy in the description of solvent effects for 346 processes [5], as compared with the particular models taken separately.

The minimal set of descriptor scales [1] was refined by statistical pocessing of the whole extended data set using the modified descriptor scale selection procedure. In the above 9-scale set the acidity scale *E* was replaced by the initial $E_T(30)$ scale (no. 24). The main factor governing selection of the initial set was to attain sufficiently low interrelation between the initial descriptor scales [2]. However, the final sample of residual scales in almost any case was orthogonal; therefore, we tested as additional descriptors some residual scales corresponding to fairly large data series which are poorly described by the given set of descriptors. In addition, specific attention was given to the significance of particular descriptors. For this purpose, we took into account the effect of additional scales, which can be simulated by samples of random numbers. We also improved the procedures for excluding insignificant descriptors and those responsible for inadmissibly high level of the "pumping-over" effect [2] in the data treatment for particular processes and the procedure for excluding strongly deviating points. We have found that for the series of processes under study 8 residual solvent constant scales are significant (i.e., they include almost no cross contributions); these scales are listed below (in parentheses are given the corresponding scale serial numbers): $Y(1)$, $P(2)$, $\delta^2_{H\,rsd}(7)$, $B_{rsd}(8)$, $E_T(30)_{rsd}(24)$, $\pi^*_{rsd}(10)$, $\beta_{rsd}(11)$, and $\alpha_{rsd}(12)$ (the index "rsd" refers to descriptor obtained by the refinement procedure). After exclusion from each series of no more than three strongly deviating points, the overall number of excluded points for 346 processes was 194 (or 2.85% of the total number of points). For 331 process (95.7% of the overall number; RPDS set), the multiple correlation coefficient *R* was greater than 0.95 (the average value was 0.979; 0.973 for the complete set).

* For communication II, see [1].

** This study was financially supported by the Estonian Science Foundation (grant no. 3029).

The present communication reports on more detailed description and analysis of the results of statistical data treatment for 346 interrelated processes listed in [1] with the aid of the proposed 8-scale descriptor set. The complete tabulated results, including free terms A_0, coefficients C_k and weight contributions W_k^0 (k is the descriptor number) to the dispersion of the corresponding terms in the correlation equations, the number of statistical degrees of freedom, multiple correlation coefficients R corrected for the number of degrees of freedom, determination coefficients R^* [1] which characterize the dispersion, and squared normalized standard deviations S_0^2, are available from the authors.* Standard deviations are also given for the free terms and coefficients. The processes under study, solvent numbers for the excluded points, and numbers of solvents involved in the treatment are listed in [1].

Physicochemical interpretation of descriptors. Most initial solvent scales (constituting the scale set under study), from which residual descriptors were derived, were modified by the refinement procedure through subtraction of contributions depending on the scales with lower serial numbers. Therefore, a question arises so as to how their physicochemical sense should be interpreted. Two descriptors taken as basic ones from the classical electrostatic theory (polarity Y and polarizability P) are almost orthogonal; therefore, the polarizability parameter P was not subjected to the refinement procedure. The δ_H^2 scale (squared Hildebrandt solubility parameter), which characterize the Gibbs energy for formation of voids in bulk solvent, depends only on Y (R = 0.54) which describes 30.6% of the dispersion of δ_H^2. The basicity scale B also depends on Y (R = 0.25) but insignificantly: the corresponding dispersion fraction is as low as 8.6%. Subtraction of this contribution is rather symbolic. However, for $E_T(30)$ 82.6% of the dispersion is described by the preceding descriptors. The respective residual scale, $E_T(30)_{rsd}$, can be regarded as a version of the acidity scale E corrected for the contribution of $\delta_{H\,rsd}^2$ [6]. The first five (residual) descriptors originating from the modified Koppel'–Pal'm model [3] describe 85.1% of the average dispersion for processes constituting RPDS (after exclusion of points). We may presume with certainty that their physical sense is consistent with the initial scale interpretation.

For most interrelated processes (hereinafter only processes included in RPDS are meant) which are expected to be influenced by solvent bacicity the B_{rsd} parameter is characterized by a considerable weight contribution of W_8^0 to the dispersion. Large W_8^0 values are observed for all the declared basicity parameters (in parentheses is given the number of the process in the summary table of the results of statistical treatment): Putela parameter *BP* (19), 0.75; Gutmann donor number *DN* (22), 0.87; general basicity parameter β_1 (149), 0.52; basicity parameter of a solute β_2^H (150), 0.63; Drago parameters C_B (319), 0.90, and E_B (320), 0.70; $-DH^0(BF_3)$ (322), 0.87; for processes proposed as mild basicity scales [7]: *Cu* λ_{max} (139), 0.67; B_{soft} (197), 0.86; donor power parameter D_s (205), 0.67; Nagai Lewis basicity parameter (273), 0.52; enthalpy of complex formation of iodine with solvents in inert solvents (317), 0.97; μ (333), 0.31. Exceptions are only Svan parameter *BA* (14) and parameters *G* proposed by Allerhand and Schleyer, 43 and 56. The same also applies to a number of other processes which are sensitive to solvent nucleophilicity, e.g., infrared spectra of substituted phenols (181–185), ^{19}F chemical shifts of *p*-fluorophenol (243), ^1H chemical shifts of chloroform (263), proton activity from mass spectrometry (326), 0.80, enthalpy of dissolution of SnI_4 in nonpolar solvents (332), 0.75, and standard Gibbs energies of cation transfer from water to a given solvent (334–338). Thus the parameter B_{rsd} ensures satisfactory description of most processes in which both Brønsted and Lewis basicities are important. Moreover, B_{rsd} allows description to be performed of not only processes sensitive to hard basicity but also those depending on soft basicity. This leads us to conclude that there is no need of introducing additional soft basicity parameters of solvents.

Large values of W_9^0 in $E_T(30)_{rsd}$ were also found for all declared acidity parameters: Svan parameter *AC* (13), 0.34; Putela parameter *PA* (15); Svoboda parameter *AP* (18); Gutmann acceptor number *AN* (23), 0.38; and Lewis acidity parameter E_B^N (88), 0.25. The $E_T(30)_{rsd}$ scale is also essential for ^{15}N chemical shifts of benzonitrile (261), 0.32; rate constants of solvolysis of *tert*-butyl chloride at various temperatures (270, 271), 0.16 and 0.17, and of methoxy-2-methyl-2-phenylpropyl *p*-toluenesulfonate (272), 0.3, which are characterized by considerable change of electrophilic solvation of the basic center during activation; and standard Gibbs energies of anion transfer (339–342). Therefore, this descriptor can be regarded as that characterizing acid properties of a solvent, including both hard (hydrogen) and soft components. The standard Gibbs energies of transfer of Cl^-, Br^-, and I^- anions from water to a given

* The summary table containing the results of statistical treatment is also available from http://www.chem.ut.ee/tktool/files/Palm1.zip.

solvent (340–342) can be described satisfactorily without involving additional soft acidity parameter.

Obviously, the last three residual descriptors cannot be interpreted in the same manner as initial dipolarity–polarizability, π^*-basicity (with respect to hydrogen bond) β, and acidity (with respect to hydrogen bond) α, for the corresponding constituents have already been included in the contributions of the first five descriptors substracted therefrom [1]. The weight contributions of *Y* and *P* to π^* are 0.569 and 0.211, respectively. The weight contribution of B_{rsd} to β is 0.444, and that of $E_T(30)_{rsd}$ to α is 0.414. The parameter π^*_{rsd} describes 8.6% of the average dispersion for processes constituting RPDS, β_{rsd} describes only 3.5%, and α_{rsd}, 3.1%.

In our first communication on this topic [2] we presumed that $\pi^*_{rsd}(10)$ can be regarded as a measure of the residual ability to solvate cations. Our assumption was based on the fact that the *BA* scale (14) proposed by Svan and co-workers strongly depends on π^*_{rsd} (W^0_{10} = 0.27) but does not depend on B_{rsd}. Analysis of processes from the extended sample shows that considerable contributions of π^*_{rsd} are observed for all (except 2) 13 Menshutkin reactions included in RPDS; for processes nos. 32, 275, 276, 277, 278, 279, 280, 288, 289, 293, and 302 [1] the weight contributions of π_{rsd} are, respectively, 0.344, 0.156, 0.511, 0.495, 0.472, 0.489, 0.471, 0.267, 0.268, 0.160, and 0.187 (for process no. 290 W_{10} = 0.09; however, this process is not taken into account since its "mixed" component is equal to 6.1).

A similar pattern is observed for the IR spectra of a series of carbopnyl compounds: nos. 172, 173, 174, 175, 209, 210, 211, 212, 213, and 214. The weight contributions of π^*_{rsd} are, respectively, 0.533, 0.519, 0.706, 0.138, 0.226, 0.438, 0.254, 0.404, 0.404, and 0.507. However, the corresponding contributions are insignificant, e.g., for the standard Gibbs energies of cation transfer from water into a given solvent (334–338). We can conclude that π^*_{rsd} is more likely to reflect the ability for solvation of dipoles (dipolar solvation).

Insofar as the average contributions of the last two parameters, β_{rsd} and α_{rsd}, to description of interrelated processes are small, it is difficult to assign them any physicichemical sense. It is reasonable to assume that they could reflect residual soft basicity and acidity, respectively. A number of processes proposed as soft basicity scales are related to B_{rsd} rather than to β_{rsd} (see above); only for the Nagai Lewis basicity parameter (273) the weight contribution of β_{rsd} is equal to 0.45. Also, we cannot draw any definite conclusions on α_{rsd}.

Thus only six of the eight significant (residual) descriptors can be assigned a definite physical sense.

"**Mixed**" **components of weight contributions of descriptors.** It should be emphasized that the weight contributions W^0_k of descriptors to dispersion of correlated quantities have both positive and negative values. Statistical treatment of the results obtained without excluding strongly deviating points showed that the number of negative W^0_k values is 19.5% of the total number of weight contributions of significant descriptors for all interrelated processes; the negative values are more typical of contributions which are smaller than 0.1 in absolute value. For greater contributions, the fraction of negative values sharply decreases. The weight contributions of descriptors are determined by the formula [2]

$$W^0_{kj} = R_{kj} X_{0kj},$$

where R_{kj} is the pair correlation coefficient for the *j*th parameter and *k*th descriptor, and X_{0kj} is the corresponding normalized regression coefficient. The latter may be represented as follows:

$$X_{0kj} = \sum_l R_{lj} R^{-1}_{klj} = R_{kj} R^{-1}_{klj} + \sum_{l \neq k} R_{lj} R^{-1}_{klj},$$

where R^{-1}_{klj} is the corresponding element of the reversed correlation matrix, and *l* is the descriptor index. After substituting, we obtain Eq. (1):

$$W^0_{kj} = (R_{kj})^2 R^{-1}_{klj} + R_{kj} \sum_{l \neq k} R_{lj} R^{-1}_{klj}$$

$$= (R_{kj})^2 + (R_{kj})^2 R^{-1}_{klj} + R_{kj} \sum_{l \neq k} R_{lj} R^{-1}_{klj}. \quad (1)$$

When the set of descriptors is completely orthogonal, all elements of the reversed correlation matrix (except for diagonal) are equal to zero, and the diagonal elements are equal to unity. In this case

$$W^0_{kj} = (R_{kj})^2. \quad (2)$$

This quantity may be only positive. As follows from Eq. (2), when the set of descriptors is incompletely orthogonal, the expression for weight contribution of a given descriptor contains W^0_{kj} and an additional component depending on the other descriptors. This component can be regarded as mixed, and it may have a negative value.

Using the quantity $(R_{kj})^2$ as a base constant of the scale, we can write an expression for the normalized weight contribution of a descriptor:

$$W_{kj}^0/(R_{kj})^2 = 1 + M_{kj},$$

where M_{kj} is the normalized mixed component. Hence

$$M_{kj} = W_{kj}^0/(R_{kj})^2 - 1. \quad (3)$$

The weight contributions of descriptors W_{kj}^0 may be negative only when $M_{kj} < 0$ and $|M_{kj}| > 1$, i.e., it determines the main part of the weight contribution. Therefore, the occurrence of negative weight contributions indicates the determining role of mixed components and incomplete orthogonality of the descriptors used. Although the proposed set of eight residual solvent constant scales is almost orthogonal, their orthogonality is valid only for the complete sample of lines. As concerns incomplete samples typical of most real experimental data series, considerable nonorthogonality is possible. The absolute value of the normalized mixed component $|M_{kj}|$ characterizes the effect of nonorthogonality on the weight contributions. When $|M_{kj}|$ is much greater than unity, most part of the weight contribution is determined by the mixed component. As a result, it is impossible to definitely interpret the corresponding weight contribution and coefficient as parameters characterizing the effect of a given descriptor. The summary table contains both weight contributions W_{kj}^0 and the corresponding M_{kj} values.

Table 1 shows distribution of the overall numbers of weight contributions W_k^0 of significant descriptors for all interrelated processes, depending on the range of $|M_{kj}|$ values before and after exclusion of strongly deviating points. The distribution almost does not change in going from RPDS processes to those belonging to the RRDS set (which are poorly described) [1]. The fraction of contributions for which $|M_{kj}| > 1$ is 31%. It is seen that the fraction of mixed components clearly tends to decrease as the average absolute values of weight contributions increase. Although the majority of very small weight contributions are characterized by $|M_{kj}| > 1$, a number of the opposite examples can be given: process no. 58, for α_{rsd}, $W_k^0 = 0.004$ and $M_{kj} = 0.076$; no. 69, for α_{rsd}, $W_k^0 = 0.002$ and $M_{kj} = 0.10$; no. 294, for β_{rsd}, $W_k^0 = 0.003$ and $M_{kj} = -0.17$. We conclude that weight contributions of descriptors cannot be used as discriminating factor for their selection.

Table 1. Distribution of the total numbers of weight contributions W_k^0 of significant descriptors for all interrelated processes, depending on the range of absolute values of normalized mixed components $|M_{kj}|$ before and after excluding strongly deviating points

$\lvert M_{kj}\rvert$ range	Before excluding						After excluding					
	$R > 0.95$ (RPDS)		$R < 0.95$ (RRDS)		all		$R > 0.95$ (RPDS)		$R < 0.95$ (RRDS)		all	
	N[a]	W_k^0[b]	N[a]	W_k^0[b]	N[a]	W_k^0[b]	N[a]	W_k^0[b]	N[a]	W_k^0[b]	N[a]	W_k^0[b]
0.0–0.1	210	0.3942	85	0.3454	295	0.3801	273	0.3858	16	0.3763	289	0.3853
0.1–0.2	148	0.3596	48	0.2314	196	0.3282	201	0.3469	6	0.2352	207	0.3437
0.2–0.5	325	0.2457	107	0.2364	432	0.2434	425	0.2477	15	0.1983	440	0.2460
0.5–1.0	209	0.1405	42	0.2667	251	0.1616	270	0.1567	3	0.2568	273	0.1578
1.0–2.0	179	0.0962	26	0.1322	205	0.1008	211	0.1048	1	0.2337	212	0.1055
2.0–5.0	181	0.0536	29	0.0661	210	0.0553	230	0.0644	3	0.1075	233	0.0650
5.0–10.0	34	0.0203	13	0.0364	47	0.0247	52	0.0276	0	0.0000	52	0.0276
10.0–25.0	20	0.0072	8	0.0376	28	0.0159	29	0.0110	0	0.0000	29	0.0110
25.0-50.0	7	0.0095	3	0.0111	10	0.0100	12	0.0104	0	0.0000	12	0.0104
50.0–100.0	2	0.0030	0	0.0000	2	0.0030	3	0.0057	0	0.0000	3	0.0057
100.0–250.0	1	0.0003	0	0.0000	1	0.0003	1	0.0003	0	0.0000	1	0.0003
250.0–500.0	1	0.0001	1	0.0016	2	0.0009	1	0.0001	0	0.0000	1	0.0001
>500.0	0	0.0000	1	0.0005	1	0.0005	0	0.0000	0	0.0000	0	0.0000
All	1317		363		1680		1708		44		1752	

[a] Number of weight contributions W_k^0 of significant descriptors.

[b] Average absolute value of weight contribution.

Table 2. Total numbers of significant descriptors and numbers of descriptors characterized by absolute values of normalized mixed components $|M_{kj}| > 1$ for the complete set of interrelated processes

D[a]	$\bar{W}_k$[b]	N_{sgn}[c]	$N_{sgn}(|M_{kj}| > 1)$[d]	Fraction of ($|M_{kj}|$),[e] %
Y	0.4642 ± 0.0998	307	19	6.19
P	0.0652 ± 0.0093	250	121	48.40
δ_H^2	0.1099 ± 0.0182	236	126	53.39
B	0.1359 ± 0.0236	176	41	23.30
E	0.0842 ± 0.0128	221	67	30.32
π^*	0.0804 ± 0.0127	221	49	22.17
β	0.0310 ± 0.0043	149	71	47.65
α	0.0293 ± 0.0037	171	51	29.82

[a] Descriptor.
[b] Average effective weight contributions [1].
[c] Overall number of processes with a significant contribution of the given descriptor.
[d] Number of processes with a normalized value of mixed component greater than unity.
[e] Percentage of the latter processes with respect to N_{sgn}.

There are also a few number of cases for which significant weight contributions of descriptors are concerned with $|M_{kj}|$ values much greater than unity. In 15 cases, weight contributions of drescriptors with $|W_k^0| > 0.2$ (10 of these are negative) are characterized by $|M_{kj}| > 2$, e.g., for process no. 80, $\delta_H{}^2{}_{rsd}$, $W_k^0 = -0.24$ and $M_{kj} = -2.8$; no. 84, $\delta_{H\,rsd}{}^2$, $M_{kj} = -0.34$ and $M_{kj} = -2.3$; no. 165, $\delta_H{}^2{}_{rsd}$, $W_k^0 = 0.70$, $M_{kj} = 3.6$; no. 202, $\delta_H{}^2{}_{rsd}$, $W_k^0 = -0.41$, $M_{kj} = -2.3$; no. 281, $\delta_H{}^2{}_{rsd}$, $W_k^0 = -0.24$, $M_{kj} = -2.8$; no. 323, Y, $W_k^0 = 0.54$, $M_{kj} = 2.4$. When using the maximal allowable value of M_{kj} equal to unity as a fairly rigid arbitrary additional criterion for selection of descriptors, the above processes are described by quite different sets of descriptors; in most cases the correlation quality becomes considerably or much poorer. For example, for process no. 80, the correlation coefficient R changes from 0.960 to 0.921; no. 202, from 0.963 to 0.957; no. 281, from 0.952 to 0.742; no. 323, from 0.995 to 0.848. Therefore, the presence of significant mixed components not only makes it impossible to unambiguously interpret the corresponding weight contributions as parameters characterizing the effect of a particular descriptor but also creates illusions of a better description of interrelated processes.

Nevertheless, significant mixed components in weight contributions do not depreciate physical sense of the above six descriptors, for in all cases the corresponding absolute values of mixed components are fairly small (considerably smaller than 0.5), except for ^{19}F chemical shifts of *p*-fluorophenol (process no. 243; B, $M_{kj} = 1.2$].

Table 2 compares the number of significant descriptors for all interrelated processes characterized by high absolute values of normalized mixed components ($|M_{kj}| > 1$) with the overall number of descriptors. Their fraction, with a single exception, does not exceed 50%. Therefore, for the sample of processes under study the presence of a mixed component is unlikely to affect the number and nature of the required descriptors.

The question so as to whether the quantity M_{kj}, by analogy with SPUR values characterizing the pumping-over effect [2], should be used as an additional criterion for selection of regressors included in the model and which versions of the corresponding procedure can be proposed requires separate study.

REFERENCES

1. Pal'm, N.V. and Pal'm, V.A., *Russ. J. Org. Chem.*, 2000, vol. 36, no. 8, pp. 1075–1104.
2. Palm, V. and Palm, N., *Org. Reactiv.*, 1993, vol. 28, p. 125.
3. Koppel, I.A. and Palm, V.A., *Advances in Linear Free Energy Relationships*, Chapman, N.B. and Shorter, J., Eds., London: Plenum, 1972, chap. V, pp. 203–280.
4. Kamlet, M.J., Abboud, J.-L.M., Abraham, M.H., and Taft, R.W., *J. Org. Chem.*, 1983, vol. 48, p. 2877.
5. Palm, V. and Palm, N., *Org. Reactiv.*, 1997, vol. 31, p. 141.
6. Koppel', I.A. and Payu, A.I., *Reakts. Sposobn. Org. Soedin.*, 1974, vol. 11, p. 137.
7. Sandstrom, M., Persson, I., and Persson, P., *Acta Chem. Scand.*, 1990, vol. 44, p. 653.