The Three-parameter Nieuwdorp Equation is the Optimal Linear Free Energy Relationship for the Prediction of Missing Data

C. Leo de Ligny *
Laboratory for Analytical Chemistry, University of Utrecht, Croesestraat 77A, 3522 AD Utrecht, The Netherlands
Hans C. van Houwelingen
Institute for Mathematical Statistics, University of Utrecht, Budapestlaan 6, 3584 CD Utrecht, The Netherlands

It has been shown that the linear free energy relationship proposed by Nieuwdorp (1979) gives a better fit to available data on substituent effects than the relationships proposed by Taft (1958), Swain (1983), and Yukawa–Tsuno (1980). One reason is that the Nieuwdorp relationship contains three parameters, while the Taft relationship has only two. It is shown here that the Nieuwdorp relationship gives the best prediction of missing data on substituent effects, and that nothing can be gained by adding a fourth ρ parameter. This is demonstrated with three sets of data: (1) the set of 76 series of data on 17 substituents from which Nieuwdorp derived the values of his σ_I, σ_R, and σ_E variables; (2) a set of 164 series of data on 14 substituents for which the fit by the Nieuwdorp relationship was not satisfactory; and (3) a set of 28 series of data on 14 substituents, selected with the view to provide a large variance that cannot be explained by the Nieuwdorp relationship.

In a previous paper [1] the ability of three recently proposed linear free energy relationships (LFERs) [2–4] and of the Taft [5] LFER to estimate data on substituent effects was compared. The four LFERs are shown in Table 1.

The symbol K denotes an equilibrium or reaction rate constant; the subscripts X and H refer to the substituted and the unsubstituted compound, respectively; ρ, and f, r, and h are reaction-dependent parameters; and σ, F, and R are variables that depend on the nature of the substituent.

The number of parameters in the Taft equation is less than that in the other equations, but in a way this is compensated for by the fact that four sets of σ_R variables are given by Taft,[5] depending on the type of reaction that is investigated. In the Swain [2] equation, the effect of a substituent is characterised by only two variables, F and R. In the Yukawa–Tsuno [3] and the Nieuwdorp [4] equations, three variables are used. The values of the variables that are given by Yukawa and Tsuno are subject to a constraint: σ_π^- is zero for electron-donating substituents and σ_π^+ is zero for electron-withdrawing substituents. The values of the variables given by Nieuwdorp are not subject to any constraint. They were derived by the statistical technique of factor analysis, from data on substituent effects in 76 reaction series. In this way the experimental errors in the data were averaged out.

These four regression equations were applied [1] to 209 reaction series for which it had been reported that the classic Taft equation is not satisfactory. The criterion used to judge the ability of the equations to fit the data was the ratio of the residual standard deviation of the regression over the total standard deviation of the data. A value of 0.1 or smaller for this ratio was considered to be satisfactory. It means that at least 90% of the variation in the data can be explained by the regression equation. This criterion has been advocated by Ehrenson [6] and by Exner.[7] Table 2 gives the results.

It is not surprising that the scope of application of the four free-energy relationships increases in the order seen in Table 2. It is obvious that their ability to fit data on substituent effects increases when the number of parameters increases, and when the number of variables that are used to characterise the behaviour of substituents increases, and when there are no constraints on the values of these variables.

Table 1. Three recently proposed LFERs, and the Taft LFER

Author	$\log(K_X/K_H) =$	Parameters
Taft (1958)	$\rho_I\sigma_I + \rho_R\sigma_R$	ρ_I, ρ_R
Swain (1983)	$fF + rR + h$	f, r, h
Yukawa–Tsuno (1980)	$\rho_I\sigma_I + \rho_\pi^+\sigma_\pi^+ + \rho_\pi^-\sigma_\pi^-$	ρ_I, ρ_π^+, ρ_π^-
Nieuwdorp (1979)	$\rho_I\sigma_I + \rho_R^0\sigma_R^0 + \rho_E\sigma_E$	ρ_I, ρ_R^0, ρ_E

Table 2. Results of regression analysis of 209 series of data by three recently proposed LFERs, and the Taft LFER

Author	Number of series for which the fit is satisfactory	Number of parameters	Number of variables	Constraints on the values of the variables
Taft (1958)	0	2	2	No
Swain (1983)	15	3	2	No
Yukawa–Tsuno (1980)	27	3	3	Yes
Nieuwdorp (1979)	45	3	3	No

No satisfactory fit by any equation: 164 series
Number of investigated series: 209

The next questions are, whether this result implies that the Nieuwdorp equation is also the best of the considered LFERs for the prediction of missing data, and if so, whether a still better LFER can be obtained by introducing a fourth ρσ term. These questions are answered here.

Theory.—How can the ability of a statistical model to predict missing data be judged? The equations that we must consider to answer this question are shown in Table 3.

The relationship between a set of data, the true statistical model, and the experimental error is given by equation (1) (Table 3), where y represents the data, M_{true} the true statistical model, and ε the experimental error. Usually, instead of the true model a simplified model is used. For instance, in the case of regression analysis, variables of minor importance are omitted. So, instead of equation (1), we must use equation (2). For the

Table 3. Some equations pertaining to the ability of a statistical model to predict missing data

$$y = M_{true} + \varepsilon \quad (1)$$

$$y = M_{simple} + M_{rest} + \varepsilon \quad (2)$$

$$\sigma^2(y - M_{simple}) = \sigma^2(M_{rest}) + \sigma^2(\varepsilon) \equiv \sigma^2 \quad (3)$$

$$y_1 \cdots y_n \rightarrow \hat{M}_{simple} \rightarrow y_{n+1} \cdots y_{n+m}$$

$$y - \hat{M}_{simple} = (y - M_{simple}) - (\hat{M}_{simple} - M_{simple}) \quad (4)$$

$$\sigma^2_{pred.} \equiv \mathrm{Var}\,(y - \hat{M}_{simple}) = \sigma^2(y - M_{simple}) + \sigma^2(\hat{M}_{simple} - M_{simple}) \quad (5)$$

$$\sigma^2_{pred.} = \sigma^2 + \sigma^2(\hat{M}_{simple} - M_{simple}) \quad (6)$$

$$\sigma^2_{pred.} = \sigma^2\left(1 + \frac{p}{n - 1 - p}\right) \quad (7)$$

variance of the data around the simplified model, equation (3) holds.

Now, suppose that the parameters of the simplified model are estimated by some statistical technique from a sample $y_1 \cdots y_n$, and the estimated model is applied to predict future observations $y_{n+1} \cdots y_{n+m}$. Suppose, further, that some of these observations are now actually made. Then the difference between such an observation and its prediction based on other observations is given by the left-hand side of equation (4), which can also be formulated as shown in the right-hand side. As the estimated model is independent of the data y in equation (4), we get for the variances equation (5): the variance of the left-hand side is equal to the sum of the variances of the two terms in the right-hand side. The combination of equations (3) and (5) then yields equation (6).

If the simplified model gets less simple, that means, if the number of parameters increases, the value of $\sigma^2(M_{rest})$ will decrease and thus the total variance σ^2 will decrease until the minimum value $\sigma^2(\varepsilon)$ is reached. The effect of increasing the number of parameters on the last term of equation (6) is less clear. It is an increasing function of both σ^2 and the number of parameters. This can be understood when it is realised that the existence of σ^2 is the very reason why the simplified model cannot be determined exactly. Furthermore, putting additional parameters in this model means putting in the additional random error of their estimates. To give an example: for regression analysis on large data sets and random missing data, equation (6) takes the form of equation (7), where p is the number of parameters and n is the number of data.[8]

So if we increase p, we might expect that $\sigma^2_{pred.}$ will decrease first, because σ^2 decreases, but that later $\sigma^2_{pred.}$ will increase because σ^2 stabilises and p keeps growing. Hence, there will be an optimum value of p which yields the minimum value of $\sigma^2_{pred.}$.

It will be clear that the ability of a linear substituent free-energy relationship to fit available data is governed by σ^2, but its ability to predict missing data is governed by $\sigma^2_{pred.}$. As the models of Swain, Yukawa–Tsuno, and Nieuwdorp all have three parameters, the last of these three models not only best fits existing data, it also best predicts missing data.

The next question is whether the Nieuwdorp model is also the best possible for the prediction of missing data, or whether a still better model can be obtained by increasing the number of parameters, *i.e.* by adding one or more $\rho\sigma$ terms to the Nieuwdorp model. To compare the merits of LFERs with different numbers of $\rho\sigma$ terms we must consider them as examples of the factor analysis model. In this statistical model it is supposed that both ρ and σ are unknown. From a number of series of data on substituent effects, the various $\rho\sigma$ terms (factors) are extracted successively. Each time a factor is extracted, the values of ρ and σ are estimated with the view to explain as much as possible of the residual variance. Thus, the ability of the factor analysis model to fit available data increases steadily with increasing number of factors, *i.e.* with increasing number of $\rho\sigma$ terms. However, the ability of the model to predict missing data is governed by an equation that is analogous to equation (7) of Table 3, and this ability will go through an optimum with increasing number of parameters, *i.e.* with increasing number of factors. To investigate this point, we used Nieuwdorp's results [4] on factor analysis of a set of 576 data on 76 reaction series and 17 substituents. From the data in his Table 1 (sums of squares of residuals) and the corresponding numbers of degrees of freedom, we calculated the standard deviations of the fit with available data ($\hat{\sigma}$). Then, approximate values of the standard deviations of the prediction of missing data ($\hat{\sigma}_{pred.}$) were calculated from equation (7). The results are shown in Table 4.

Table 4. Standard deviations of the fit to available data ($\hat{\sigma}$), and of the prediction of missing data ($\hat{\sigma}_{pred.}$) of LFERs with different numbers of $\rho\sigma$ terms for a set of 576 data. The symbol p denotes the number of parameters

Model	p	$\hat{\sigma}$	$\hat{\sigma}_{pred.}$
$\sigma_1\sigma_1$	92	0.80	0.87
$\rho_1\sigma_1 + \rho_2\sigma_2$	184	0.23	0.28
$\rho_1\sigma_1 + \rho_2\sigma_2 + \rho_3\sigma_3$	276	0.06	0.08
$\rho_1\sigma_1 + \rho_2\sigma_2 + \rho_3\sigma_3 + \rho_4\sigma_4$	368	0.07	0.12
$\rho_1\sigma_1 + \rho_2\sigma_2 + \rho_3\sigma_3 + \rho_4\sigma_4 + \rho_5\sigma_5$	460	0.07	0.16

Table 4 shows that the values of $\hat{\sigma}$ level off with increasing number of parameters. The limiting value of 0.06—0.07 presumably represents the experimental error $\sigma(\varepsilon)$. The values of $\hat{\sigma}_{pred.}$ show a minimum for the model with 276 parameters and three $\rho\sigma$ terms, *i.e.* for the Nieuwdorp equation. So, with this set of data (that were confined to chemical reactions and equilibria of rigid aliphatic systems and small π-systems) the Nieuwdorp equation is indeed the optimum LFER, both for data fitting and for predicting new data.

However, in the Introduction we mentioned a set of 164 series of data that could not be fitted well by the Nieuwdorp equation. These series contain data on chemical reactions and equilibria, phase equilibria, and physical properties of rigid aliphatic systems, small π-systems, and extended π-systems. The medium used ranged from water, *via* apolar solvents, to the gas phase. We thought that this set of data might be better fitted by a LFER with four $\rho\sigma$ terms, than by the Nieuwdorp relation, and that it thus might be a suitable data set to define a set of σ_4 values.

Description of the Mathematical Procedure.—The model considered here is given in equation (8), with $E\varepsilon_{ij} = 0$, and

$$y_{ij} = \sum_{k=1}^{K} a_{ik}b_{jk} + \varepsilon_{ij} \quad (8)$$

$\sigma^2(\varepsilon_{ij}) = \sigma_i^2$ (E = expectation, σ^2 = variance) (i and j denote the series and the substituent, respectively). In order to avoid confusion the symbols ρ and σ have been replaced by a and b in equation (8).

The case $K = 3$ yields the Nieuwdorp equation, $K = 4$ gives the extension with a $\rho_4\sigma_4$ term. In the $K = 3$ case, the a parameters can be obtained by the regression of a row of ys on the bs, *i.e.* by minimising $\Sigma_j(y_{ij} - \Sigma_{k=1}^{3} a_{ik}b_{jk})^2$ as function of a_{i1}, a_{i2}, and a_{i3}, with Σ denoting summation over the indices j for which y_{ij} has been observed. The variance σ_i^2 is estimated using equation (9) where N_i = number of observations in row i.

Table 5. Standard deviations of the fit to available data ($\hat{\sigma}_3$ and $\hat{\sigma}_4$, respectively), and of the prediction of missing data ($\hat{\sigma}_{pred..3}$ and $\hat{\sigma}_{pred..4}$, respectively) of the three-parameter Nieuwdorp equation and its four-parameter extension (log $K_X/K_H = \rho_I\sigma_I + \rho_R^0\sigma_R^0 + \rho_E\sigma_E + \rho_4\sigma_4$), for 28 series of data, taken from Table VIII of ref. 1

Reaction type	Type of data	Substituent position	Number of data	Series [1]	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_{pred..3}$	$\hat{\sigma}_{pred..4}$
σ_R^0	^{13}C N.m.r.	*p*	14	204	2.2	2.2		
σ_R(BA)	Chemical	*m*	5	195	0.19	0.12	0.22	0.29
	Chemical	*p*	8	85	0.025	0.028	0.030	0.062
	Chemical	*p*	8	194	2.0	2.2	2.2	2.6
	Chemical	*p*	6	196	0.49	0.53	0.58	1.2
σ_R^-	Chemical	*m*	6	71	1.0	1.2	1.3	2.7
	Chemical	*m*	6	184	0.14	0.13	0.19	0.18
	^{1}H N.m.r.	*p*	8	118	0.14	0.10	0.17	0.13
	^{1}H N.m.r.	*p*	7	145	0.38	0.36	0.45	0.44
σ_R^+	Chemical	*m*	9	60	0.48	0.44	0.68	0.64
	Chemical	*p*	12	49	0.028	0.012	0.031	0.013
	Chemical	*p*	7	52	0.32	0.27	0.48	0.89
	Chemical	*p*	5	61	0.22	0.10	0.31	0.039
	Chemical	*p*	5	62	0.30	0.018	0.44	0.069
	Chemical	*p*	6	63	0.33	0.40	0.47	0.71
	Chemical	*p*	6	133	0.11	0.041	0.16	0.17
	I.r.	*m*	5	120	2.9	0.89	4.2	3.4
	I.r.	*m*	5	123	3.0	1.7	4.3	6.7
	Ionisation potential	*m*	5	174	0.15	0.21	0.24	0.57
	E.s.r.	*p*	11	48	1.2	0.22	1.2	0.24
	I.r.	*p*	6	121	11.3	5.8	16.1	10.2
	I.r.	*p*	6	124	7.0	6.5	9.9	11.4
	I.r.	*p*	6	127	0.33	0.38	0.47	0.67
Difficult to classify	Chemical		6	55	0.60	0.59	0.89	2.4
	Chemical		8	69	13.6	12.8	20.3	19.5
	Chemical	*p*	7	84	0.12	0.12	0.17	0.21
	U.v.	*p*	8	46	0.24	0.17	0.32	0.24
	U.v.	*p*	8	47	0.033	0.024	0.046	0.035

$$\hat{\sigma}_i^2 = (\Sigma \text{ residuals}^2)/(N_i - 3) \quad (9)$$

Let $\hat{a}_{ik}$ denote the estimate of a_{ik}. The squared prediction error corresponding with a missing y_{ij} is given by $\sigma_i^2 + \sigma^2(\Sigma_{k=1}^3 \hat{a}_{ik}b_{jk})$. This prediction error can be estimated by standard statistical techniques using the estimated covariance matrix of $(\hat{a}_{i1}, \hat{a}_{i2}, \hat{a}_{i3})$. The squared prediction error per row is obtained by averaging the squared prediction of all empty spots.

The model with $K = 4$ is a bit harder to handle. Now, besides the a values, the b_{j4} values are also unknown. The unknown parameters are estimated by minimising equation (10) as a

$$\Sigma_i w_i \Sigma_j (y_{ij} - \Sigma_{k=1}^4 a_{ik}b_{jk})^2 \quad (10)$$

function of the a_{ik} and b_{j4} values under the side restrictions $\Sigma_{j=1}^{14} b_{jk}b_{j4} = 0$ for $k = 1,2,3$ and $\Sigma_{j=1}^{14} b_{j4}^2 = 1$, needed to make the parameters unique. The weight factors w_i are taken as $w_i = 1/c_i^2$ where c_i is the last significant digit of the ith row, in order to make the rows comparable. The minimisation is achieved by an iterative procedure with the following steps: given the b values, determine new a values in the same way as in the case $K = 3$ (this can be done row-wise); given the a values and b_{j1}, b_{j2}, b_{j3} for all j values, determine new b_4 values by simple weighted regression of $y_{ij} - \Sigma_{k=1}^3 a_{ik}b_{jk}$ on a_{i4} with weights w_i.

Starting with a sensible b_4, convergence is achieved quickly. If there were no missing data, the computation could be simplified a lot: a principal components' analysis on the Nieuwdorp residuals would suffice. Unfortunately, the irregular pattern of missing data makes it more complicated. Finally, for the case $K = 4$, prediction errors are determined in the same way as for the case $K = 3$, neglecting the sampling error in the estimate b_{j4}.

In this way, the prediction errors are slightly underestimated. In view of the conclusion of this paper, namely that the model with $K = 4$ is no improvement over the model with $K = 3$, this is no serious problem. Obtaining the correct prediction errors as in ref. 9 would cost a lot of programming effort and computer time.

Data.—The data were selected from the above mentioned set of 164 series of data on 17 substituents that are given in Tables 7 and 8 of ref. 1. The following selection criteria were applied.

(*a*) Series with *o*-substituents were not considered, as it is doubtful whether *o*-effects can be fitted by any LFER.

(*b*) Series with a fixed substituent in the *o*-position with respect to the variable substituent or the reaction centre were not considered. See ref. 4 for the arguments.

(*c*) Only those data were considered that are proportional to (free) energy differences. For other data (*e.g.* absorbances A) it is often not clear whether or not they should be transformed before applying an LFER. For example, in ref. 10 a relationship between A and σ_R is proposed, but in ref. 11 a relationship between $A^{\frac{1}{2}}$ and σ_R.

(*d*) We reasoned that if the addition of a fourth $\rho\sigma$ term to the Nieuwdorp equation should turn out to be a significant improvement, the cause would probably be that it gives a better description of the direct resonance between the substituent and the reaction centre. (In fact, the Yukawa–Tsuno equation contains two terms to take account of direct resonance, whereas the Nieuwdorp equation contains only one.) Therefore, series in which direct resonance does not occur, *i.e.* series of data on reaction rate and equilibrium constants in σ_I and σ_R^0 reactions, were not considered.

(*e*) When series were available at different temperatures or in slightly different solvents, only one was included in the data set, to prevent the inclusion of strongly correlated data.

(*f*) Only those substituents were considered for which at least

ten data were available. This criterion led to the omission of the substituents C_2H_5, C_3H_7, and CF_3.

(*g*) A series should contain data for at least five substituents, at least four of them being strong electron donors or acceptors.

(*h*) Series in which all data are <15 times the last significant digit were omitted, to prevent the introduction of too much experimental error in the data set.

The application of these selection criteria to the 164 series given in Tables VII and VIII of ref. 1, and the omission of a few series for trivial reasons, resulted in a set of 57 series: 46—49, 51—53, 55, 60—63, 69, 71—73, 76, 77, 83—86, 100, 101, 118, 120, 121, 123, 124, 126, 127, 132, 133, 142—145, 149, 150, 157, 158, 162, 163, 169, 172—174, 178, 183, 184, 193—201, 203, and 204.

Results and Discussion

It appears that in most cases $\hat{\sigma}_{pred.}$ is smaller for the Nieuwdorp equation than for its four-parameter extension. We thought that the cause might be that for a number of series, $\hat{\sigma}$ for the Nieuwdorp equation is rather small, *i.e.* smaller than ten times the last significant digit. For these series, the residuals from regression analysis by the Nieuwdorp equation (from which the fourth $\rho\sigma$ term must be estimated) are for a large part due to experimental error, rather than to the model error of the 'simple' model, the Nieuwdorp equation. As the inclusion of these series might hamper the estimation of the fourth $\rho\sigma$ term, we repeated the procedure for the 28 series of which $\hat{\sigma}$ for the Niewdorp LFER is larger than ten times the last significant digit, *i.e.* the series 46—49, 52, 55, 60—63, 69, 71, 84, 85, 118, 120, 121, 123, 124, 127, 133, 145, 174, 184, 194—196, and 204. The results are shown in Table 5.

It appears that for half of the number of series, $\hat{\sigma}_{pred.,4}$ is larger than $\hat{\sigma}_{pred.,3}$. Therefore, the addition of a fourth $\rho\sigma$ term to the Nieuwdorp equation does not yield better predictions of missing data, even for this set of 28 series which was selected to provide a large variance that can not be explained by the Nieuwdorp equation. (This negative result may be caused partly by the small number of data in some series. If only series with seven or more data are considered, $\hat{\sigma}_{pred.,4}$ is larger than $\sigma_{pred.,3}$ in only four out of the 12 cases. For the series with eight or more data, $\hat{\sigma}_{pred.,4}$ is larger than $\hat{\sigma}_{pred.,3}$ in two of the nine cases.)

Conclusions

For the prediction of missing data on substituent effects, the Nieuwdorp equation is the optimum linear free-energy relationship.

References

1 M. C. Spanjer and C. L. de Ligny, *J. Chem. Res.*, 1986, (*S*) 176; (*M*) 1701.
2 C. G. Swain, S. H. Unger, N. R. Rosenquist, and M. S. Swain, *J. Am. Chem. Soc.*, 1983, **105**, 492.
3 M. Sawada, M. Ichihara, Y. Yukawa, T. Nakachi, and Y. Tsuno, *Bull. Chem. Soc. Jpn.*, 1980, **53**, 2055.
4 G. H. E. Nieuwdorp, C. L. de Ligny, and J. C. van Houwelingen, *J. Chem. Soc., Perkin Trans. 2*, 1979, 537.
5 R. W. Taft and I. C. Lewis, *J. Am. Chem. Soc.*, 1958, **80**, 2436.
6 S. Ehrenson, R. T. C. Brownlee, and R. W. Taft, *Prog. Phys. Org. Chem.*, 1973, **10**, 1.
7 O. Exner, *Collect. Czech. Chem. Commun.*, 1966, **31**, 3222.
8 L. Breiman and D. Freedman, *J. Am. Stat. Ass.*, 1983, **78**, 131.
9 J. C. van Houwelingen, 'Proceedings of the 3rd Prague Symposium on Asymptotic Statistics,' Elsevier, Amsterdam, 1984, 295.
10 A. R. Katritzky, R. F. Pinzelli, M. V. Sinnott, and R. D. Topsom, *J. Am. Chem. Soc.*, 1970, **92**, 6861.
11 R. T. C. Brownlee, A. R. Katritzky, and R. D. Topsom, *J. Am. Chem. Soc.*, 1965, **87**, 3261.

Received 24*th April* 1986; *Paper* 6/784