Clustering of Aryl Carbon-13 Nuclear Magnetic Resonance Substituent Chemical Shifts. A Multivariate Data Analysis using Principal Components

Dan Johnels, Ulf Edlund,* Hans Grahn, Sven Hellberg, Michael Sjöström, and Svante Wold
Department of Organic Chemistry, Research Group for Chemometrics, Umeå University, S-90187 Umeå, Sweden
Sergio Clementi
Department of Chemistry, University of Perugia, Perugia, Italy
William J. Dunn III
Department of Medicinal Chemistry, College of Pharmacy, University of Illinois at the Medical Center, Chicago, Illinois 60680, U.S.A.

A principal component analysis of the ^{13}C substituent-induced chemical shifts of 82 monosubstituted benzenes shows that *ca.* 90% of the substituents belong to one of four clusters, acceptors, alkyls, donors, or halogens. This grouping is confirmed statistically. The extensions of the subclasses are not parallel. It is also shown that the predictive capability of the single-parameter models for each subclass is better than any multiparameter model applied on the whole data set. The observed grouping of substituents provides an explanation to the apparent correlation frequently found between ^{13}C n.m.r. chemical shifts and dual substituent parameters. The ability of the statistical method to discover incorrect shift data is also illustrated.

Since aromatic compounds have a key role as models for the understanding of σ- and π-electron distribution, it was natural that n.m.r. studies of monosubstituted benzenes have become an area of great theoretical interest.[1] A large ^{13}C n.m.r. chemical shift range and the expected correlation between shifts and atomic charges reinforced the parallelism with electronic substituent effects.[2] In this latter field a common approach has been the correlation of ^{13}C shift data with various single or dual substituent parameter σ-scales (SSP or DSP).[2] Most σ-constants are derived from equilibrium or kinetic studies, *i.e.* describing differences between the thermodynamic states or the transition state–ground state. Nevertheless, they are assumed to bear information on ground state electronic perturbation and in a loose, quasi-theoretical way related to electron densities.[2e]

The π-electron density is regarded as the most important factor controlling the chemical shift of carbon atoms in aryl systems.[3] This stems from the proposal that the paramagnetic term in the Karplus–Pople equation is dominant.

Several reports have been presented where an SSP model has been found to be statistically sufficient to interpret certain n.m.r. shifts in aromatic compounds.[4] There has been much criticism of this approach, since the SSP model embodies a fixed ratio between field and resonance effects.[5] It has also been common to perform a DSP analysis, where the resonance parameter is chosen among several resonance scales (σ_R^0, σ_R^+, $\sigma_R^{(BA)}$, and σ_R^-) and only report the results of the best fit.[2g,6] Many additional DSP approaches have appeared, for instance where the resonance part of the model was allowed to be a continuous function of the electron demand on the substituent.[7]

A serious drawback is that only very few reports applying the DSP correlation actually give any statistical support (for example an *F* test) for the increased parameterization. This 'Ockham's Razor' dilemma has caused some debate.[4a,5e,8] Lately a ^{13}C n.m.r. study of remote carbons in styrene systems showed that single component or single substituent constants give the same fit and predictive ability as DSP models.[4f] This study was recently dubiously criticized by a claim that an SSP equation is just a particular case of the general DSP equation. A general statement was made that, even if the statistical fit using the two types of treatment was similar, the DSP method provides important information not obtainable from an SSP treatment.[8]

A DSP model puts greater demands on the choice and spread of substituents compared with simple regression. Several basis substituent sets have been suggested, representing as wide a domain as possible in substituent properties.[1c,9] Moreover, during recent years a refinement of the transmission models of substituent effects has been proposed, where the field effect has been dissected into two components, direct field (F_D) and field-induced π-polarization (F_π).[2a] These field effects are generally considered to be more important than the σ-inductive effect in ^{13}C n.m.r.[1c,10]

In a recent critical survey of remote polar substituent effects on olefinic and aromatic carbons, a significant imprecision in the correlation with various presumed measures of field effects was noted.[2e] As mentioned, there is no stringent theoretical foundation for the common belief that reactivity and chemical shifts should respond equally to substituent-induced, especially field-induced, charge perturbations. In fact, although shifts and reactivities are influenced by similar field effects, they may fail to give a precise linear correlation with each other due to a different dependence on distance between the studied position and the point dipole. It was suggested that a more sophisticated approach is necessary.[2e]

Applications of three-parameter models, with an additional semi-empirical term *Q*, are also becoming more frequent.[11] This extra parameter is introduced to get acceptable fits of ^{13}C shifts close to the substituent in, *e.g.*, the *ipso*- and *ortho*-positions.

There are certain problems associated with the use of fixed multiparameter equations.[12] First, using multiple regression one assumes that all the substituent constants (independent variables) are exactly known and completely relevant to the shift data currently considered. Secondly, multiple regression methods need almost orthogonal substituent scales if the regression coefficients are to be precisely interpretable and to have predictive relevance. The third problem is the heavy dependence of the result upon the number and spread of points in the data set and on the selection of σ-scales.

In addition, all statistical models are based on the assumption that the analysed data are homogeneous. Hence, before applying any general statistical model to a given data set one should check if the data are clustered or not.

The possibility of a clustering of substituent effects has been a matter of rather limited interest so far. The small number of 'non-n.m.r.' based σ-values in the scales used is perhaps one

Table 1. List of the substituents used in the principal component analysis

No.	Substituent	No.	Substituent	No.	Substituent
1	Me	28	CH_2NMe_2	56	NHMe
2	Et	29	CH_2OH	57	NMe_2
3	$CHMe_2$	30	CHMeOH	58	NHEt
4	CMe_3	31	CMe_2OH	59	NEt_2
5	Bicyclo[2.2.2]octyl	32	CH_2OEt	60	NHCHMeEt
6	CH_2CH_2Me	33	CH_2OCOMe	61	NHPh
7	$CH_2CH_2CH_2Me$	34	CH_2SMe	62	NPh_2
8	$CH_2CH_2CH_2CN$	35	CH_2SPh	63	NHCOMe
9	$CH_2CH_2CH_2OH$	36	CH_2Cl	64	$NHNH_2$
10	$CH_2CH_2CH_2OMe$	37	CH_2Br	65	OH
11	$CH_2CH_2CH_2Br$	38	CH_2I	66	OMe
12	CH_2CH_2Ph	39	Ph	67	OPh
13	CH_2CH_2CN	40	CHO	67 [a]	OPh
14	CH_2CH_2CHO	41	COMe	68	OCOMe
15	CH_2CH_2COOH	42	COEt	69	$OSiMe_3$
16	CH_2CH_2OH	43	$COCHMe_2$	70	F
17	$CH_2CHMeOH$	44	$COCH_2CH_2Me$	71	Cl
17 [a]	$CH_2CHMeOH$	45	$COCH_2CH_2Br$	72	Br
18	CH_2CH_2OMe	46	COPh	73	I
19	CH_2CF_2Me	47	$CONH_2$	74	H
20	CH_2CH_2Br	48	$CONMe_2$	75	CH_2CN
21	$CHBrCH_2Br$	49	COOH	76	CF_3
22	$CH_2CH{=}CH_2$	50	COOMe	77	$C{\equiv}CH$
23	CH_2Ph	51	COOEt	78	$C{\equiv}CMe$
24	CHMePh	52	$COOCH_2CH_2CH_2Me$	79	$CH{=}CH_2$
25	$CHPh_2$	53	COCl	80	$CMe{=}CH_2$
26	CH_2COMe	54	NO_2	81	CN
27	CH_2NH_2	55	NH_2	82	N=NPh

[a] Shift assignments other than in ref. 2*d*, see text.

reason. A prevalent conception is that the transmission of substituent effects is a continuous process with no discontinuous changes in the core region during the perturbation of the system.

One report mentions that *ca.* 50% of 24 common substituents lie within a narrow range, when described by σ_I and σ_R, while the remaining substituents are spread in various directions.[13] Another analysis is based on the assumed existence of spherical clusters among several scales.[14] In a recent multivariate study,[12] 28 common substituents were described by seven substituent descriptors. A strong grouping of the substituents in four classes (acceptors, alkyls, donors, and halogens) was found.

This paper is addressed to the problem of whether the ^{13}C substituent-induced chemical shift (SCS) data of a representative choice of monosubstituted benzenes are homogeneous or clustered. Statistically the presence of clusters can be tested for by estimating the spread within the clusters and comparing it with the distances between the clusters.

In the data analysis we use principal component (PC) models.[15] They can be used to derive independent scales or 'effects'. These models have the same form as the SSP and DSP equations. Contrary to the ordinary linear free energy relationship (LFER) regression type of equation, no *a priori* knowledge of the relevance of any substituent constants is required. The resulting 'effects' will be equivalent to the orthogonal components of the PC model that best fit the experimental data.

Several papers have appeared where PC analysis or the closely related factor analysis has been applied to interpret n.m.r. substituent effects.[4e,5a,16,17]

In short, the advantage of this method, compared with the classical LFER approach, is that no 'fundamental' effects need to be defined in advance. The resulting component values will depend on (a) the problem or classification scheme under study and (b) the compounds chosen as representative of that problem.

Methods

Choice of Data.—The n.m.r. shift data of 82 monosubstituted benzenes were initially chosen from the ^{13}C n.m.r. tabulation given in a recent review.[2d] To get as large a shift matrix as possible we have chosen data measured in both deuteriochloroform and carbon tetrachloride. Some minor differences in solvation can be noted in ^{13}C n.m.r. using these solvents.[2c] These solvent effects, however, in no way affect the conclusions reached in this study. More inert media, such as cyclohexane, would cause solubility problems in some cases and consequently limit the number of possible substituents. Various forms of aggregation might also seriously affect the reliability of shift data in such media. Except for the halogens, only second-row elements were selected as the directly substituted atoms, since the recommended basis substituent sets include only such substituents.[1c,2c]

Some substituents with triple bonds and the styrene derivatives were initially avoided because of an expected anisotropy effect.[2f] The CF_3 compound was also removed from the initial analysis because of earlier reported anomalies.[2f] The classification of these compounds, and some additional substituents, were checked in the final analysis.

In summary, eight compounds altogether were kept out of the initial analysis. A complete list of the substituents is given in Table 1.

Statistical Analysis.—Data analysis of the C(1)–C(4) SCS was performed by the use of principal components. Since a detailed description of the data analysis package SIMCA has been given elsewhere,[15] we will limit the presentation to a summary.

The reported SCS of the monosubstituted benzenes are

first scaled to unit variance (see below) giving x_{ik}. These scaled data form a data matrix X, composed of the elements in equation (1) where $\bar{x}_i$ is the mean of variable i, b_{ia} is the

$$x_{ik} = \bar{x}_i + \sum_{a=1}^{A} b_{ia} \cdot t_{ak} + e_{ik} \qquad (1)$$

loading equivalent to the regression coefficient and t_{ak} constitutes the component value or 'substituent constant' for k. The b_{ia} and t_{ak} values are thus directly derived from the scaled, measured data, by minimizing the sum of the squared residuals, Σe_{ik}^2. The number of product terms, *i.e.* the degree of parameterization, A, is estimated by cross-validation. The average ($\bar{x}_i$) of each descriptor is estimated (constitutes a model with $A = 0$) and the residuals ($x_{ik} - \bar{x}_i$) are calculated. Systematic information in the residuals is then accounted for by adding product terms ($b_{ia} \cdot t_{ak}$) until only non-systematic noise remains.

We have used the so-called modelling power as a measure of the extent of variation in one variable that is accounted for by a given PC model. A variable having a modelling power close to unity participates strongly in the modelling. Scaling is an important concept in multivariate data analysis. A small, but systematic, variation in one variable can be masked by a large variation in another. To avoid this and to be able to compare possibly different class structures, we have used auto-scaled data. The variables are scaled to have the same variance over the whole data set, *i.e.* being equally important in the data analysis.

The fit of the data for a given substituent to a PC model is measured by the ratio between the residual standard deviation for the substituent (S_p) and the total residual standard deviation of all the substituents (S_0). A high S_p compared with S_0 indicates deviant behaviour of that substituent, *i.e.* it is an outlier.

Results

Initially a data matrix composed of the relative ^{13}C chemical shifts of 74 monosubstituted benzenes was constructed. The shift differences were auto-scaled and a PC analysis of the whole data set was performed. Cross-validation showed that three components were significant and that 85% of the total variance was accounted for by this model. This indicates that either the data has a high information content or that there are strong inhomogeneities in the data. All four variables, C(1)—C(4), are important in the model as indicated by their modelling powers, 0.83, 0.68, 0.92, and 0.85, respectively.

In Figures 1a and b we show two projections of the three-dimensional t-space. Visually, most compounds seem to fall into four clusters, acceptors, alkyls, donors, and halogens. A possible cause for this grouping is that the *ipso*- and *ortho*-positions are strongly affected by steric, anisotropic, bond order, or other neighbouring group effects. Hence, the observed groupings may be caused mainly by the shift behaviour of these positions. Therefore, we weighted down the data of these positions by a factor of ten. The clustering is even more obvious after this weighting. A component plot similar to those above is shown in Figure 2. Thus this result excludes the possibility that the grouping is an artifact due to the inclusion of the *ipso*- and *ortho*-positions in the analysis.

An interesting observation from Figures 1 and 2 is that the separation between the acceptors, alkyls, and donors mainly lies in the first component, t_1. Increased branching at C_α of the alkyls is the major cause for the alkyl cluster extension along the t_2 axis (Figure 1a). A similar 'size' factor is also revealed in Figure 2. In Figures 1a and b the extension of the halogen class does not parallel the components that define the plane connecting the remaining classes. In a separate analysis we also excluded the halogens. As shown in Figure 3 the observed grouping is the same.

Table 2. Results of PC analyses. Residual variances for the different models

Set	N	A [a]	S_0^2	Substituents [b]
Whole	74	0	1.00 (1.02) [c]	1—16, 17a, 18—66, 67a, 68—74
Whole	74	3	0.15 (0.15) [c]	
Acceptors	14	0	0.14	40—53
Alkyls	39	0	0.11	1—16, 17a, 18—39
Alkyls	39	1	0.04	
Donors	13	0	0.28	55—62, 64—66, 67a, 69
Donors	13	1	0.13	
Halogens	4	0	3.09	70—73
Halogens	4	1	0.08	
Pooled subclasses	70	0	0.28	
Pooled subclasses	70	0, 1	0.08	

[a] Number of components used in the modelling. [b] Substituents used in the calculation of the models. [c] Outliers excluded (Nos. 54, 63, 68, and 74).

Next we analysed the individual substituent clusters, by fitting separate PC models to each cluster. The results are reported in Table 2. All clusters are described by separate one-component models ($A = 1$) except for the acceptors, where the mean value ($A = 0$) is adequate. As evident from the b values in Table 3 the 'sensitivity' of the C(1)–C(4) positions varies between the classes, *i.e.* the class extensions are not parallel. To test if it is justified to use local models (separate models for the four subclasses) instead of the single global model, we compared the residual variance of the whole data set [S_0^2 (whole) = 1.02, $A = 0$] with the pooled residual variance of the subclasses [S_0^2 (pooled) = 0.28, $A = 0$] in an approximate F-test. It is possible to make this comparison since the aforementioned scaling was retained [equation (2)].

$$F = S_0^2 \text{ (whole)}/S_0^2 \text{ (pooled)} = 3.64$$
$$F(\text{crit.}) = 1.53 \quad (p = 0.01) \qquad (2)$$

A single model applied to a clustered data set will describe both inter- and intra-class variance. Hence it is informative to find out to what extent the global model of the whole data set describes the intra-class variation. The results of the F-test [equation (2)] shows that the use of subclass models is statistically highly significant, and that the interclass variance accounts for 72% of the total variation. The global three-component model explains 85% of the variance [S_0^2 (whole) = 0.15, $A = 3$], hence this model mainly explains interclass variance. The interclass behaviour, or clustering, is responsible for the gross part of the shift variation, and only a minor part is due to the variance in shifts within the classes (85% − 72% = 13%). The systematic part of the interclass variation

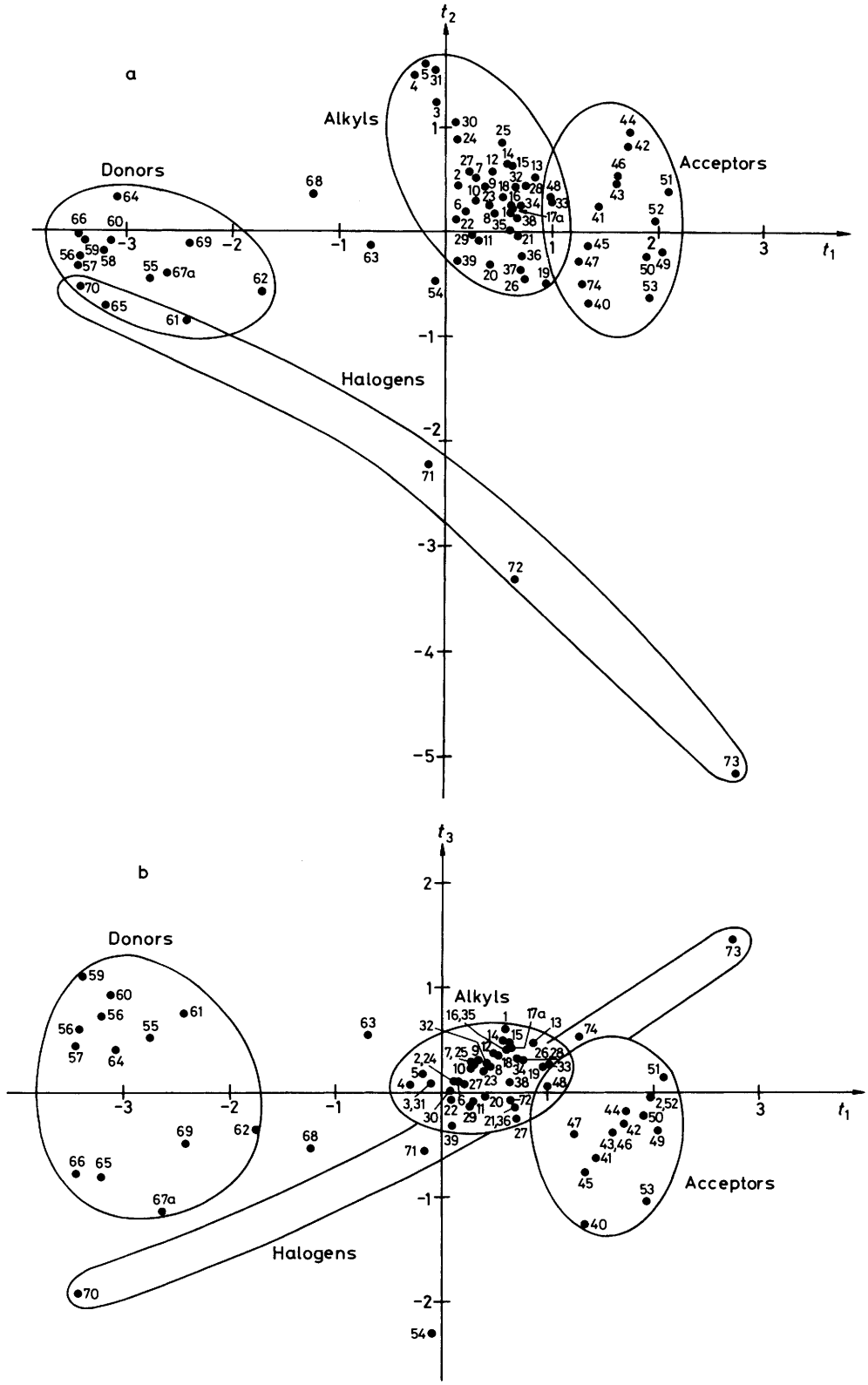

Figure 1. a Plot of the t_1 against the t_2 component for the whole data set model. The separate classes, as given by the individual class models, are surrounded by approximate lines. b Plot of the t_1 against the t_3 component for the whole data set model. Substituent clusters are marked as in Figure 1a

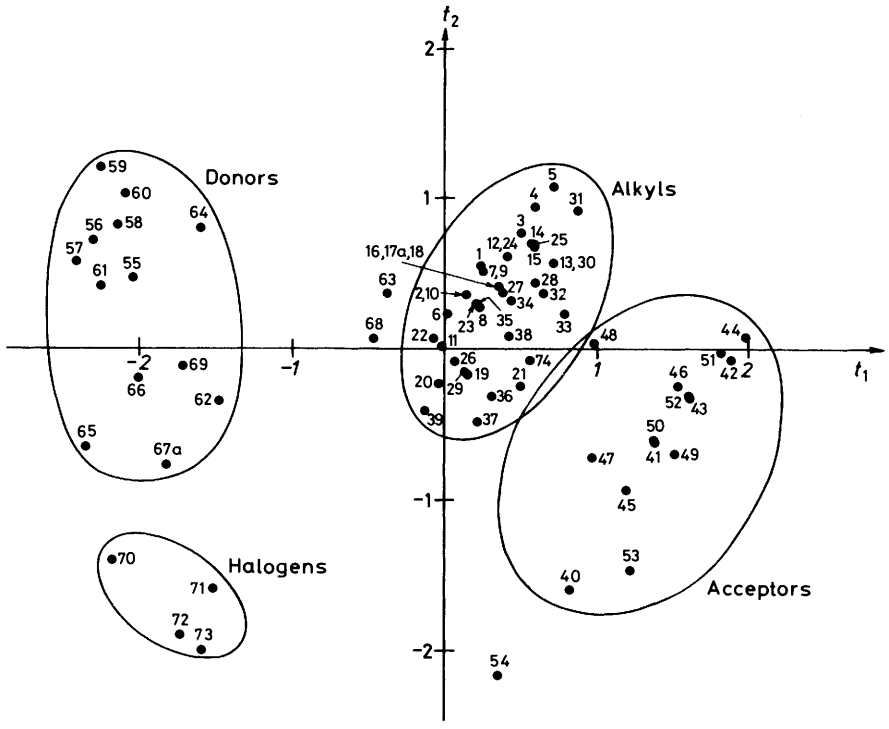

Figure 2. Plot of the t_1 against the t_2 component after multiplying the scaled *ipso*- and *ortho*-^{13}C SCS by a factor of 1/10. The indicated grouping of substituents is marked as in Figures 1a and b

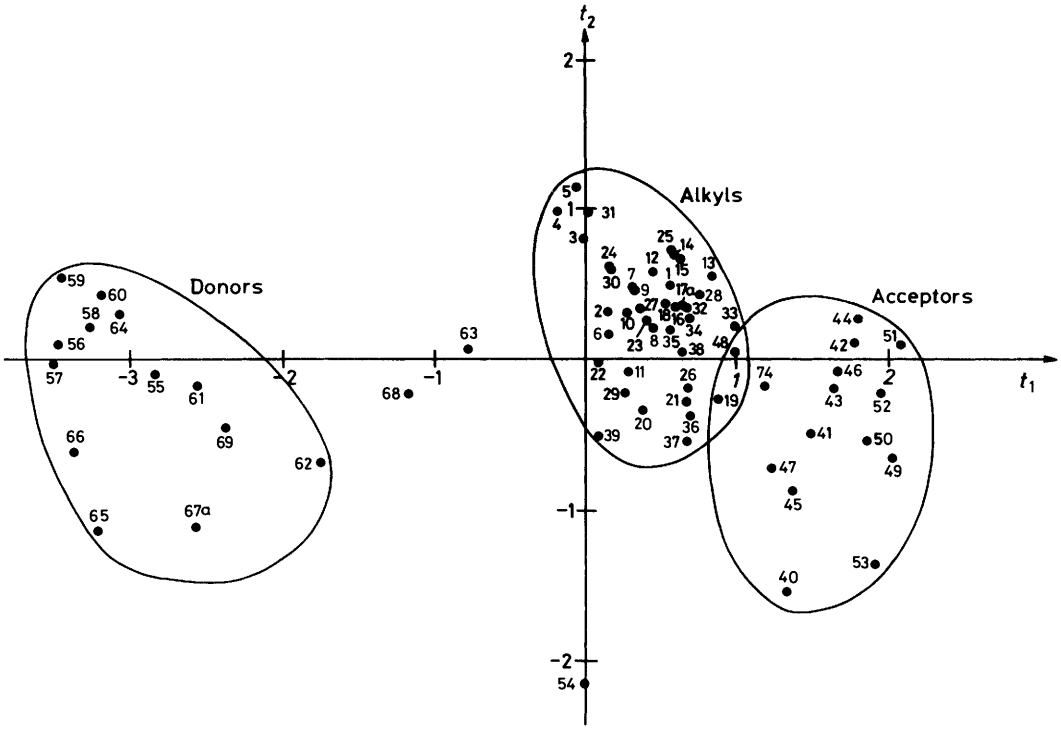

Figure 3. Plot of the t_1 against the t_2 component after excluding the halogens in the whole data set analysis

is accounted for by the local models, as measured by S_0^2 (pooled) = 0.08 (A = 0 and 1). These models give a significantly better description of the ^{13}C shift matrix than the global model (A = 3) according to an approximate F-test [equation (3)]. Consequently, by the use of the less complex

$$F = S_0^2 \text{ (whole)}/S_0^2 \text{ (pooled)} = 1.88$$
$$F\text{(crit.)} = 1.66 \quad (p = 0.01) \qquad (3)$$

models found for the separate classes (A = 0 and 1) we are able to get a better description of the intraclass behaviour

Table 3. a $\bar{x}_i$ and b_{i1} for the global model and for the sub-class models

Set	$\bar{x}_i$ *ipso*	*ortho*	*meta*	*para*	b_{i1} *ipso*	*ortho*	*meta*	*para*
Whole	1.35	−0.47	0.35	−0.44	−0.47	0.59	−0.37	0.55
Acceptors	0.66	0.09	−0.26	0.95				
Alkyls	1.38	−0.06	−0.12	−0.42	0.71	−0.32	−0.59	−0.21
Donors	2.43	−2.33	1.57	−2.09	−0.38	−0.55	−0.40	−0.63
Halogens	0.03	−0.02	2.81	−0.50	0.88	−0.46	−0.01	−0.09

b b_{i2} and b_{i3} for the global model of the whole data set

Set	b_{i2} *ipso*	*ortho*	*meta*	*para*	b_{i3} *ipso*	*ortho*	*meta*	*para*
Whole	0.58	−0.09	−0.81	0.05	−0.52	0.02	−0.41	−0.75

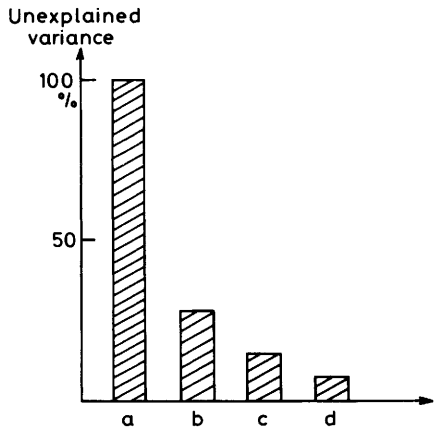

Figure 4. Illustration of the remaining unexplained variances using different class models: a, $A = 0$, whole data set, 100%; b, $A = 0$, pooled subclass models, 28%; c, $A = 3$, whole data set, 15%; d, $A = 0$ and 1, pooled subclass models, 8%

than by the more complex three-component model based on the whole data set. These results are summarized in Figure 4.

This means that any two- or three-parameter model calculated on the whole data set has only a limited predictive ability of the intra-class behaviour.

The validity of the specific clusters is seen from Table 4 where the residual standard deviations (S_p) are given for each compound when (a) fitted to the local class models and (b) when fitted to the three-component global model.

Alkyls.—In the alkyl class, the $CH_2CHMeOH$ compound (17) shows a large residual standard deviation ($S_p = 0.66$) *i.e.* it is an ' outlier '. The ^{13}C spectrum of this compound was re-examined, and the coupled spectrum clearly revealed that the *ortho*- and *meta*-shifts had been interchanged in the earlier report.[2d] After reassignment, it falls well within the confidence limit of the alkyl class ($S_p = 0.17$).

We note that the erroneous data for this compound fit the global model (Table 4), which thus has a less predictive capability. No compounds from the other classes were classified as alkyls. CH_2OCOMe (33) is a borderline substituent with characteristics more similar to these of the acceptors. Due to the anisotropic behaviour of the substituents $CH_2CH{=}CH_2$ (22) and Ph (39), these compounds are both on the borderline, but they are definitely closer to the alkyl class.

Acceptors.—The acceptor class is adequately described by its mean value. A few alkyl substituents are found within or

Table 4. Residual standard deviation for each substituent when (a) fitted to the overall three-component model and (b) when fitted to the separate class models

Substituent	Class	S_p (whole)	S_p (I)	S_p (II)	S_p (III)	S_p (IV)
1	II	0.30	0.83	0.26	2.02	1.76
2	II	0.44	0.94	0.19	1.71	1.66
3	II	0.31	1.13	0.11	1.77	1.96
4	II	0.24	1.27	0.13	1.78	2.06
5	II	0.26	1.27	0.04	1.89	2.17
6	II	0.47	0.92	0.23	1.69	1.56
7	II	0.45	0.91	0.13	1.84	1.77
8	II	0.27	0.77	0.05	1.87	1.65
9	II	0.36	0.87	0.10	1.86	1.76
10	II	0.42	0.88	0.14	1.79	1.66
11	II	0.40	0.84	0.21	1.73	1.46
12	II	0.32	0.86	0.13	1.94	1.86
13	II	0.06	0.68	0.32	2.17	1.96
14	II	0.20	0.82	0.23	2.05	1.96
15	II	0.18	0.79	0.24	2.07	1.96
16	II	0.26	0.75	0.15	2.00	1.76
17	O	0.48	1.14	0.66	1.50	0.83
17 [a]	II	0.29	0.75	0.17	2.02	1.76
18	II	0.31	0.78	0.11	1.95	1.76
19	II	0.27	0.68	0.09	2.11	1.46
20	II	0.32	0.78	0.22	1.76	1.35
21	I, II	0.13	0.53	0.28	1.93	1.57
22	II	0.54	0.95	0.33	1.61	1.47
23	II	0.42	0.82	0.11	1.85	1.67
24	II	0.39	1.00	0.10	1.79	1.87
25	II	0.49	0.90	0.23	2.03	2.00
26	II	0.19	0.72	0.10	2.02	1.44
27	II	0.18	0.85	0.09	1.78	1.75
28	II	0.25	0.70	0.21	2.08	1.87
29	II	0.13	0.77	0.26	1.68	1.44
30	II	0.06	0.95	0.21	1.82	1.95
31	II	0.06	1.18	0.20	1.90	2.15
32	II	0.06	0.65	0.23	2.04	1.85
33	I	0.12	0.51	0.35	2.20	1.87
34	II	0.18	0.68	0.15	2.03	1.76
35	II	0.21	0.75	0.13	1.97	1.64
36	I, II	0.11	0.59	0.19	1.91	1.48
37	II	0.20	0.63	0.25	1.88	1.39
38	II	0.13	0.64	0.10	1.95	1.66
39	O	0.25	0.87	0.43	1.54	1.25
40	O	0.17	0.62	0.87	2.14	1.56
41	I	0.40	0.16	0.77	2.33	1.96
42	I	0.53	0.36	0.94	2.62	2.33
43	I	0.53	0.17	0.84	2.50	2.13
44	I	0.60	0.44	0.98	2.67	2.42
45	I	0.54	0.26	0.78	2.26	1.79
46	I	0.18	0.25	0.78	2.47	2.13
47	I	0.61	0.30	0.64	2.26	1.69

Table 4 (*continued*)

Substituent	Class	S_p (whole)	S_p (I)	S_p (II)	S_p (III)	S_p (IV)
48	I	0.42	0.42	0.48	2.19	1.89
49	I	0.57	0.26	0.90	2.68	1.99
50	I	0.55	0.23	0.81	2.62	1.92
51	I	0.69	0.39	1.03	2.84	2.29
52	I	0.63	0.26	0.91	2.71	2.10
53	I	0.27	0.52	0.92	2.49	1.78
54	O	0.40	1.31	1.54	1.45	1.36
55	III	0.38	2.27	1.92	0.22	1.33
56	III	0.50	2.59	2.21	0.16	1.53
57	III	0.31	2.59	2.22	0.22	1.41
58	III	0.54	2.49	2.10	0.03	1.57
59	III	0.59	2.62	2.21	0.07	1.81
60	III	0.53	2.48	2.06	0.09	1.67
61	III	0.07	2.18	1.83	0.56	1.27
62	III	0.27	1.76	1.39	0.55	0.73
63	O	0.69	1.28	0.92	1.34	1.49
64	III	0.75	2.40	1.94	0.38	1.56
65	III, IV	0.08	2.48	2.25	0.40	0.63
66	III	0.03	2.56	2.19	0.53	0.80
67	O	0.28	1.99	0.96	1.32	1.97
67 [a]	III, IV	0.39	2.21	1.95	0.12	0.40
68	O	0.00	1.47	0.98	0.92	1.19
69	III	0.52	2.09	1.66	0.27	0.80
70	IV	0.25	2.68	2.55	0.74	0.19
71	IV	0.57	1.56	1.08	1.65	0.28
72	IV	0.46	1.85	1.12	2.47	0.14
73	IV	0.07	2.86	1.92	4.30	0.19
74	I	0.42	0.59	0.50	2.39	1.65
75	O	0.50	0.85	1.05	2.81	2.47
76	I	0.93	0.56	0.74	2.20	1.51
77	O	0.31	0.91	1.04	3.01	2.06
78	O	0.22	0.86	0.88	2.83	1.95
79	II	0.31	0.63	0.32	1.97	1.74
80	O	0.33	0.80	0.37	2.03	2.05
81	O	0.74	1.43	1.06	3.25	1.17
82	O	0.22	1.21	1.06	1.35	1.48
C.L.		0.76	0.59	0.33	0.61	0.63

I, acceptor; II, alkyl; III, donor; IV, halogen. C.L. class limit (p 0.05).

near the confidence limit of this class [the α-halogenated alkylbenzenes (36)—(38) and, as mentioned, the CH_2OCOMe derivative (33)]. As earlier indicated in Figure 1 the NO_2 substituent (54) is classified as an 'outlier' (Table 4), since its unique behaviour, especially noted for C(1) and C(2), is not represented among the other substituents in the acceptor class. However, the NO_2 derivative is closest to the acceptor class. The CHO substituent (40) is also on the borderline, but as for NO_2, the deviating behaviour is due to the *ipso*- and *ortho*-carbon shift values. As seen in Figure 2 these two compounds lie in the extension of the acceptor class.

Donors.—The donors are described by a one-component model and it is seen that the OCOMe (68), the NHCOMe (63), and the OPh (67) derivatives fall outside this model (Table 4). However, using earlier reported shift data for the OPh compound from this laboratory,[17] it was correctly classified ($S_p = 1.32$ *versus* $S_p = 0.12$). It can also be noted that the major extension of the donor class is caused by change from O-donors to N-donors.

Halogens.—The halogen class includes the four halogens. The class structure intersects the plane defined by the other classes in the proximity of the OPh (67) and the OH (65) positions. One can note that among the halogens, the Cl (71) and F (70) derivatives lie closest to this plane (Figures 1a and b).

Others.—Of the remaining compounds which were not used in the modelling, the CF_3 compound (76) falls well within the acceptor class. The styrene derivatives (79) and (80) are close to the alkyl class. The other compounds, most containing a triple bond, are 'outliers' to all class models. Hydrogen (74) is not well described by any of the separate models (Table 4).

Discussion

The present result clearly indicates that the substituent effects as probed by ^{13}C shieldings are strongly clustered and can hardly be seen as a result of a continuous transmission mechanism. *Ca.* 90% of the studied substituents belong to one of the four classes. The halogens are few and the F and Cl compounds are close to the donors. This could be one explanation why the fit to single- or dual-parameter equations is often surprisingly good. It is always possible to describe three clusters reasonably well with a plane. Two components such as t_1 and t_2 or σ_I and σ_R will give a fair description of these three classes. In those cases reported earlier where a one-component model has proved to be adequate, the probe positions have been chosen four bonds or more away from the substituent.[4]

This condition supports the earlier assumption that at least in certain conjugated systems there exists a 'cross-over' point, *i.e.* a distance beyond which a single parameter model is sufficient to interpret the observed SCS.[4a] Thus use of only SCS of remote positions and/or a specific choice of substituents may reduce the correlation model to a single parameter model, *i.e.* the classes are merged on a line.

It is important to note that deviations are frequently observed for the halogen substituents, especially for Br and I, when two-parameter models are used.[2f,5a,18] In a PC analysis of 2-substituted indenes we noted that the Cl and Br compounds deviate from a two-component model.[16c] A similar result was achieved in a study of the substituent effects on ^{13}C and ^{15}N shifts of *para*-phenyl-substituted triazenes.[18] The halogen class was there described by a one-component model while the remaining structures were modelled by a separate two-component model. This result is similar to those of the present report. However, the number of substituents in the triazene study were too few to establish firmly any grouping of substituents. Many 'non'-n.m.r. investigations have also reported anomalous behaviour of the halogens.[19] This peculiarity of the halogens may also be reflected intentionally or unintentionally, in the recommendations given for the most commonly used basis sets for DSP correlations: 'use two halogens (but not both Cl and Br)',[1c] or 'use F, Cl, or Br'.[2c]

In a recent investigation of 28 of the most common substituents, it was shown that a strong grouping into four classes, acceptors, alkyls, donors, and halogens, was obtained if the substituents were characterized by seven common descriptors.[12] Moreover a grouping into these four classes is indicated from a simple plot of $\sigma_I(^{19}F)$ and $\sigma_R(^{19}F)$ although there are only a limited number of substituents in common with the present data set (Figure 5).[20] The mentioned examples could be an indication that this grouping behaviour is prevalent in a variety of systems using various substituent probes.

It has recently been asked [4f] 'if it is sufficient to relate C-13 chemical shifts to semi-empirical parameters such as σ, σ_I, and σ_R,' and we certainly agree that we now have reached a stage where one should try to go beyond such correlations and try to understand the reason for their success. In our opinion the present study gives at least a partial answer to this question. The specific nature of certain

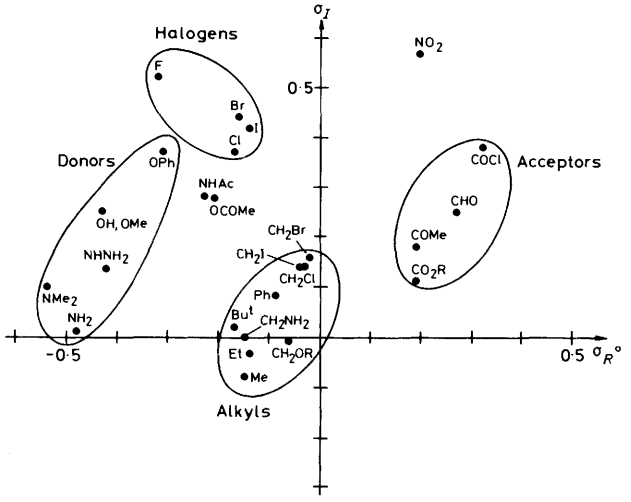

Figure 5. Plot of σ_R^0 (^{19}F) against σ_I (^{19}F) for those substituents which are in common with the analysed benzene ^{13}C SCS data set

substituents, indicated by their groupings, is an explanation for this correlation. Any parameterized model with a potential to describe the relative positions of these three or four classes will give this apparent correlation. But, in order for a fit to be statistically significant, a critical condition that must be fulfilled before applying any multiple regression analysis on a data set is that the used objects are continuously spread in the variable domain.

A natural approach at the present stage would be to test the *intra-class* predictive capability of generally used reactivity models. Unfortunately, an insufficient number of non-n.m.r. σ-values for the compounds obstructs a study of whether the common substituent parameter models have any predictive ability on ^{13}C chemical shifts within the separate classes.

Conclusions

Using a principal component data analysis of the ^{13}C SCS of 82 monosubstituted benzenes we have the following conclusions.

(A) *Ca.* 90% of all substituents belong to one of four clusters, alkyls, donors, acceptors, and halogens. From a statistical point of view this means that in the benzene system the question about substituent effects and their transmission is not a simple one-class problem. The continuity that is associated with the term 'effect' is only valid within the four subclasses.

(B) Simple (zero and one-component) models for the separate classes accommodate better to the shift data than any dual- or triple-component model used on the whole data set.

(C) In this case, incorrect shift data or false shift assignments can be detected by the local class models but not by the global model.

(D) The condition that a majority of substituents falls within three clusters (except heavy halogens) offers a reasonable explanation why the apparent correlation of ^{13}C chemical shifts to common single- or dual-parameter equations is so often observed. However, if the chemical shift data are clustered as for the present benzene derivatives, this type of data analysis is not statistically allowed. A requirement for the use of multiple regression analysis is that the substituents are continuously spread in the variable space.

The strong grouping found in the present study is probably a consequence of the relative strong interactions that exists between the benzene carbons and the substituent. Since this grouping also has been found using substituent descriptors (σs) as independent variables this could mean that the validity of global reactivity models for reactions with strong interaction between the substituent and the reaction centre is in doubt. Models such as the Hammett equation which apply for reactions at remote positions (involving weak interactions) are probably not affected by the present result, but this proposal demands further investigations.

Acknowledgements

Grants from the Swedish Natural Science Research Council and the Swedish Council for Planning and Coordination of Research to U. E. and S. W. are gratefully acknowledged.

References

1 (*a*) D. F. Ewing in 'Correlation Analysis in Chemistry,' eds. N. B. Chapman and J. Shorter, Plenum Press, New York, 1980, p. 357; (*b*) M. T. Tribble and J. G. Traynham in 'Advances in Linear Free Energy Relationships,' eds. N. B. Chapman and J. Shorter, Plenum Press, New York, 1972, p. 143; (*c*) R. D. Topsom, *Prog. Phys. Org. Chem.*, 1976, **12**, 1; (*d*) G. L. Nelson and E. A. Williams, *ibid.*, 1976, **12**, 229.
2 (*a*) W. Adcock and T.-C. Khor, *J. Am. Chem. Soc.*, 1978, **100**, 7799; (*b*) T. C. Brownlee and D. J. Craik, *Org. Magn. Reson.*, 1981, **15**, 248; (*c*) J. Bromilow, T. C. Brownlee, V. O. Lopez, and R. W. Taft, *J. Org. Chem.*, 1979, **44**, 4766; (*d*) D. F. Ewing, *Org. Magn. Reson.*, 1979, **12**, 499; (*e*) E. A. Hill and H. E. Guenther, *ibid.*, 1981, **16**, 177; (*f*) G. K. Hamer, I. R. Peat, and W. F. Reynolds, *Can. J. Chem.*, 1973, **51**, 897; (*g*) N. K. Wilson and R. D. Zehr, *J. Org. Chem.*, 1982, **47**, 1184.
3 D. G. Farnum, *Adv. Phys. Org. Chem.*, 1975, **11**, 123.
4 (*a*) P. J. Mitchell and L. Phillips, *J. Chem. Soc., Perkin Trans. 2*, 1974, 109; (*b*) G. R. Wiley and S. I. Miller, *J. Org. Chem.*, 1972, **37**, 767; (*c*) R. E. Bilbo and D. W. Boykin, *J. Chem. Research (S)*, 1980, 332; (*d*) D. A. R. Happer, *Aust. J. Chem.*, 1976, **29**, 2607; (*e*) U. Edlund and S. Wold, *J. Magn. Reson.*, 1980, **37**, 183; (*f*) A. Cornelis, S. Lambert, P. Laszlo, and P. Schaus, *J. Org. Chem.*, 1981, **46**, 2130.
5 (*a*) W. F. Reynolds, P. Dais, D. W. MacIntyre, G. K. Hamer, and I. R. Peat, *J. Magn. Reson.*, 1981, **43**, 81; (*b*) W. F. Reynolds, P. G. Mezey, and G. K. Hamer, *Can. J. Chem.*, 1977, **55**, 522; (*c*) D. F. Ewing and K. J. Toyne, *J. Chem. Soc., Perkin Trans. 2*, 1979, 243; (*d*) M. J. Shapiro, *J. Org. Chem.*, 1978, **43**, 3769; (*e*) R. T. C. Brownlee and D. J. Craik, *J. Chem. Soc., Perkin Trans. 2*, 1981, 760.
6 J. Bromilow, R. T. C. Brownlee, D. J. Craik, P. R. Fiske, E. J. Rowe, and M. Sadek, *J. Chem. Soc., Perkin Trans. 2*, 1981, 753.
7 D. A. R. Happer and G. J. Wright, *J. Chem. Soc., Perkin Trans. 2*, 1979, 694.
8 D. J. Craik, R. T. C. Brownlee, and M. Sadek, *J. Org. Chem.*, 1982, **47**, 657.
9 S. Clementi, *CAOC Newsletter*, 1981, **2**, 13.
10 W. F. Reynolds, *J. Chem. Soc., Perkin Trans. 2*, 1980, 985.
11 (*a*) W. B. Smith and T. W. Proulx, *Org. Magn. Reson.*, 1976, **8**, 567; (*b*) J. Llinares, J.-P. Galy, R. Faure, E.-J. Vincent, and J. Elguero, *Can. J. Chem.*, 1979, **57**, 937.
12 S. Alunni, S. Clementi, U. Edlund, D. Johnels, S. Hellberg, M. Sjöström, and S. Wold, *Acta Chem. Scand.*, in the press.
13 S. Clementi and F. Fringuelli, *Anal. Chim. Acta*, 1978, **103**, 477.
14 C. Hansch, S. H. Unger, and A. B. J. Forsythe, *J. Med. Chem.*, 1973, **16**, 1217.
15 S. Wold, (*a*) *Pattern Recognition*, 1976, **8**, 127; (*b*) *Technometrics*, 1978, **20**, 397.
16 (*a*) M. Sjöström and U. Edlund, *J. Magn. Reson.*, 1977, **25**, 285; (*b*) U. Edlund and Å. Norström, *Org. Magn. Reson.*, 1977, **9**, 196; (*c*) B. Eliasson and U. Edlund, *J. Chem. Soc., Perkin Trans. 2*, 1981, 403; (*d*) G. Musumarra, S. Wold, and S. Gronowitz, *Org. Magn. Reson.*, 1981, **17**, 118.

17 (*a*) P. H. Weiner and E. R. Malinowski, *J. Phys. Chem.*, 1971, **75**, 1207, 3160; (*b*) K. B. Wiberg, W. E. Pratt, and W. F. Bailey, *J. Org. Chem.*, 1980, **45**, 4936; (*c*) W. F. Bailey, E. A. Cioffi, and K. B. Wiberg, *ibid.*, 1981, **46**, 4219.
18 (*a*) W. J. Dunn III, G. K. Lins, T. Manimara, S. Grigoras, U. Edlund, and S. Wold, *Org. Magn. Reson.*, in the press; (*b*) T. A. Holak, S. Sadigh-Esfandiary, F. R. Carter, and D. J. Sardella, *J. Org. Chem.*, 1980, **45**, 2400.
19 P. Politzer and J. W. Timberlake, *J. Org. Chem.*, 1972, **37**, 3557.
20 For a compilation see O. Exner in 'Correlation Analysis in Chemistry: Recent Advances,' eds. J. Shorter and N. B. Chapman, Plenum Press, New York, 1978, p. 439.

Received 23*rd August* 1982; *Paper* 2/1463