match no. 24 pp. 119 179 1989

CORRELATION ANALYSIS IN THE STUDY OF SUBSTITUENT EFFECTS

T.M. Krygowski[a] and R.I. Zalewski[b]

[a]Institute of Chemistry, Warsaw University, Warsaw, ul. Pasteura 1, Poland

[b]Division of General Chemistry, Academy of Economics, Poznań, ul. Marchlewskiego 146, Poland

(Received: May 1987)

SUMMARY

The foundation of similarity models with one or more independent variables are discussed. Various substituent parameters and separation of inductive and resonance effects are described. Classical steric effects of substituents and their effect on the geometry of molecules (in organic crystals) are reviewed. Both experimental and theoretical approaches are discussed.

1. INTRODUCTION

In statistics the term correlation analysis denotes such an analysis that describes the strength of mutual interrelations between two (or more) sets of random variables. The same notion used in organic chemistry means something quite different. Correlation analysis in organic chemistry [1,2] [abbreviation CAOC] - since 1979 every three years a European conference has been organized under this title: 1979 Assisi, 1982 Hull, 1985 Louvain la Neuve; the next will be held in Poland (Poznań in 1988) deals with construction and application of simple similarity models in order to explain and/or describe:

a) the mechanism of chemical reactions [3],

b) the influence of variation in the structure of substrates or attacking reagent [1-5], or the nature of environment (exemplified by solvent [5-7]

on logarithms of rate or equilibrium constants or physicochemical properties of chemical species organized in reaction series.

Chemical similarity models use multiple regression analysis equation in the form

$$\delta Q_i = \sum_{j=1}^{N} \alpha_j A_{i,j} + B_i \qquad (1)$$

δQ_i describes experimental chemical reactivities or physicochemical properties of chemical species i=1,2,..n; j=1,2,..N. Explanatory parameters $A_{i,j}$ describe quantitatively j-th independent ways (or factor, or mechanism) of interactions for a set of n species (i=1,2...n) in reaction or process under study. It is assumed that those parameters are mutually uncorrelated. Regression coefficient α_j expresses sensitivity of δQ_i variation on the changes of j-th mechanism. The intercept B_i of regression should be close to zero since operator δ applied in (1) is the Leffler and Grunwald [7] one. This is a difference of property Q for i-th chemical species and some other species accepted as a standard or reference. In this way some part of possible systematic error may be diminished. Additionally, the multiparameter solution of the chemical problem could be supported by calculation of the per cent contribution of each mechanism by the use of expression (2)

$$\% \text{ j-th mechanism} = \frac{100\alpha_j}{\sum_{j=1}^{N} \alpha_j} \qquad (2)$$

Prior to such calculations, all α_j values must be normalized to give them all equal weights. This is necessary because of various units for explanatory parameters $A_{i,j}$ which yield different magnitudes for α_j. Any common variance contributed in

e.g. $A_{i,j=1}$ and $A_{i,j=2}$ results in incorrect values of α_j (j=1,2) in spite of normalization and may lead to false conclusions. Some authors [8,9] prefer other ways of demonstrating various blends of two mechanisms k and l contributing to the total substituent effects, $\varepsilon = \frac{\alpha_k}{\alpha_l}$. If only two factors are being considered, ε and %-contribution calculated by expression (2) are proportional.

2. CONSTRUCTION OF SIMILARITY MODELS

In chemistry it is usually possible to vary one structural or environmental variable at a time for a set of chemical species (i= 1,2,... n). This set of chemical species is often named a reaction series. The dependent variable in the reaction series (experimental chemical or physicochemical property) is affected by one or two major independent ways or modes or mechanisms of action. It often happens that the relative intensity of two simultaneously operated ways is nearly constant for all chemical species in the reaction series, and for a large variety of chemical processes. Then the problem under study simplifies and resembles one way problem. Such a situation seems to take place for *meta*-substituted benzene derivatives. The single parameter representation of eq.(1), i.e. a linear regression is for these reaction series most often sufficent to decribe substituent effects. That is, the blend of resonance and inductive (or delocalized and localized) contributions to the overall substituent effect is relatively constant for a full range of structural variation and for most chemical processes.

The situation is much more complex for the case of a reaction series in which the blend of two independent ways of action is not constant within the frame of the reaction series, and is still dependent on the chemical process.

For these cases, two - or more - parameter regressions (1) have to be used in order to explain (inperpret) the dependence of δQ - values for the given reaction series on the mechanism of interactions between the reaction (process) site and substituents. This case is realized for *para*-disubstituted derivatives of benzene and to a greater extent for *ortho*-disubstituted species. For the first case - for various reaction series - depending on the nature of chemical process, the blend of resonance/induction effects varies from ~40% (acid/base equilibria of *para*-substituted phenylacetic acids) to ~60% (acid/base equilibria of *para*-substituted anilines)‡. Moreover, depending on the nature of the reaction site (electron repelling or accepting) a through-conjugation effect acts in different directions and needs different scales of substituent constants, σ_p^- and σ_p^+, respectively. Nevertheless, it is possible to treat *para*-substituted reaction series in two different ways:

(i) using for the reaction series in question a linear model with substituent constants maintaining the same variability of blend of two independent mechanisms as that used for the reference reaction series, i.e. that used to define substituent constants, and

(ii) using two (or more) parameter regressions which allows to estimate per cent contributions of these independent mechanisms in the substituent effect measured in the given process, δQ .

‡ Percentage calculated by the use of σ_I and σ_R and eq.(2)

In the case of *ortho*-substituted species neither (i) nor (ii) is effective: (i) is not valid since the blend of resonance-inductive contributions varies from compound to compound. This does not allow to apply any simple model of similarity; (ii) is better suited for such a case although steric effects may interfere in this case too.

The fundamental problem in applying the similarity model is to define a reference reaction of known mechanism and complexity. This reaction will be used to define independent variable or substituent constants. The natural requirements for choosing such a reference reaction are as follows:

(i) high accuracy and precision of measurements,

(ii) measurements should be available for a wide range of chemical species $i = 1,2...n$.

(iii) the chemical reactivity parameter or physicochemical property should be highly sensitive to the change of structure (i.e. to the nature of substituent),

(iv) the mechanism of interaction of the reaction (process) site with substituent in the reference reaction series should be as well known as possible. Otherwise, no useful information for interpretation of δQ_i is available even if regression (1) holds. The only predictive value of eq.(1) is in such a case a result of our treatment.

The set of chemical species (i.e. reference reaction series) which follows all above mentioned conditions in a reference process may be used to estimate parameters $A_{i,j}$ of eq.(1) and then to apply them in order to explain and/or predict δQ_i - values of the studied reaction series.

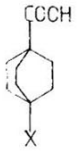
Two historical examples will explain the practical way of similarity model construction.

Hammett [10] applied as a reference reaction the acid-base dissociation of *meta* and *para*-substituted benzoic acids in water at 25°C. This reaction meets all four natural requirements stated in this section. This means (i) high accuracy of potentiometric pK_a data for *para* and *meta* substituents such as strong electron-donors (*OH*, *OPh*), various *alkyls*, all *halogens*, and electron-acceptors ($COCH_3$, *CN*, NO_2) as well as (ii) relatively high sensitivity of pK_a upon the change of substituent. The pK_a data are spread in a range of more than 1 pK_a unit. By definition only one way of action of substituent on the reaction site is accepted (j=1). Benzoic acid is used as a reference (zero - condition) compound to set substituent constant $\sigma \equiv \log K(X)/K(H)$. The sensitivity of the reaction towards various substituents is set as standard (unit condition), and ρ=1. Hence $A_{i,j}=A_i=\sigma_i=\sigma_m(X)$ or $\sigma_p(X)$, where X is a substituent in *meta* or *para*-position to the carboxyl group in benzoic acid derivatives. These explanatory parameters, or substituent constants, were used in the equation now known as Hammett's equation:

$$\delta Q(X) = \rho\sigma(X) + \text{const} \qquad (3)$$

$\delta Q(X)$ are chemical [10] or physicochemical [8] properties of *meta* or *para*-substituted derivatives of benzene or even more complex systems [8] in which *meta* and *para*-like positions may be fixed or this kind of mechanism of interaction (transmission of electrical effect) may be assumed. Eq.(3) does not apply to *ortho* substituted systems in which the nature of substituent effect on reactivity depends to a much greater degree on the type of reaction (or attacking reagent), nature of environment and solvent effect [9].

The second example is taken from a work by Roberts and Moreland [11] who have measured acid/base dissociation of 4-substituted bicyclo-[2.2.2]-octanecarboxylic acids. There is no interaction between substituent and carboxylic group possible through Π-electrons. It was accepted [11] that only inductive and/or field effects may operate in this structure and log K_i values measured in 50% aqueous ethanol at 25°C were accepted as a good inductive/field substituent parameter σ_I, and are often used in eq.(1) when separation of inductive and resonance effects in the substituent effect is a purpose of study.



The difference between these two examples is significant. In the first case the substituent constants σ reflect the blend of two independent mechanisms of electron interaction between substituent and reaction site. This blend is not constant within the series of *meta* and *para*-substituted derivatives of benzene. Nevertheless, these parameters are successfully used to describe substituent effect for those reaction series in which this blend changes similarly as in *meta* and *para*-substituted benzoic acids. The other example presents the reaction with a single mechanism of interaction and this scale of substituent constants may be used either as σ_I, the component in eq.(1), or as similarity parameter for a reaction series with inductive effect only. Then the linear regression version of eq.(1) is sufficient. When a new process or reaction, or a group of related reactions fails to fit in with explanatory parameter $A_{i,j}$, one has at least three choices:

(1) one can use a different set of explanatory parameters,

(2) one can use two (or even more) explanatory parameters, $A_{i,1}$ and $A_{i,2}$ or,

(3) one can introduce no new parameters, but instead use the deviation from the line to suggest special effects, or a more complex process (e.g. two independent ways instead of one, or change of relative intensity of two simultaneously operated ways).

The first choice has been most attractive and of course gives better fits because one is free to change "constants" at will, for each process under study. The multiplicity of substituent parameters reflects the progress or propagation of experimental works but has several disadvantages. The better fit is now less noteworthy. It allows a researcher to accept a subjective and arbitrary selection of parameters after the data are known. The predictive power of such a treatment is very limited. Of course one can feel free to propose new similarity models, however it must be very well done and supported by extensive experimental or theoretical work and in agreement with all natural requirements. Examples will be discussed in part 3.

The second choice offers a possibility to build up a similarity model of greater complexity with two or even more independent modes of action. Then equation (1) with $j=1,2,...n$ must be used and the respective substituent parameters (independent variables) have to be applied.

It is necessary to mention now the alternative procedure. When two ways of action are sufficient, it becomes possible to extract them by a factor analysis [12] of the data set (various reactions or processes * different substituents). In the first,

"abstract" solution of such an analysis, the best least-squares fit between the observed and calculated data (with two factors) is obtained, but the factors are not unique. The following step is the transformation of this "abstract" solution into a practical real solution, in which real factors have a simple, clear physical meaning or interpretation at the molecular level. In order that this transformation be performed, two critical assumptions are required [13]. Correlation coefficients (fits) do not depend on the choice of these critical assumptions. Application of this method is much more difficult when three or more modes of action are involved, because the number of critical conditions to be identified and justified increases dramatically. Therefore it has been left unapplied. Finally, the third choice is to use deviations from linear plots (in magnitude and sign) as an indication of break point in the reaction mechanism or of solvent complexation.

All three possiblities were explored by various authors with varying interest. In the following part they are discussed in detail.

3. THE ORIGINAL HAMMETT σ's AND THEIR MODIFICATIONS

The quantitative measure of substituent effect in *meta*- and *para*-substituted benzenes has been defined from the simple linear correlation (regression). The substituent scale to be identified and justified requires one trivial reference (zero) condition and one standard (unit) condition. Acid-base dissociation constants of *meta* and *para*-substituted benzoic acids in water at 25^{o}C were chosen as the unit condition. This set up ρ=1, by definition.

(4)

The pK_a of benzoic acid in this reaction was chosen as reference – zero condition. Then

$$\sigma_{m,p} = \log \frac{K_{m,p}(X)}{K(H)} \qquad (5)$$

Hammett substituent constants σ were estimated and defined [10] as a measure of substituent effects independent of reaction, medium and temperature.

These are primary substituent constants. In order to include substituents whose benzoic acids derivatives were insolube in water, other solvents were used to derive secondary substituent constants [14]. These secondary σ values were calculated from the equation

$$\sigma_{m,p} = (\log \frac{K_{m,p}(X)}{K(H)}) \; / \; \rho \qquad (6)$$

in which ρ is the reaction constant for reaction (4) in a given solvent.

In many cases the authors used another reaction as the unit condition for getting σ's which were unavailable by the use of Hammett reaction (4) [15].

Extensive lists of σ values were published in numerous compilations [2,16,17,18]. Jaffé [8] and others [19] applied statistical smoothing in order to obtain "mean" values for σ_x. However, this multiplicity of substituent constants has several disadvantages. A better fit is less noteworthy, a user can subjectively or arbitrarily choose between scales and improve the

fit of experimental data to the regression line. It is no longer clear how many different kinds of mechanisms are involved.

On this ground we present our conviction that it is most reasonable to use primary substituent constants. If they are not available we recommend the use of secondary substituent constants which were established for the same reference reaction in different solvents, provided the precision of the regression is high.

Both sets of substituent constants are known for a large number of substituents and cover all most important and differentiated chemical and physical properties. Some constants were redetermined by McDaniel and Brown [14] with higher precision. Many other ones were redetermined by secondary standard method by Exner and co-workers [20].

The deviations from linearity then mean that there is discontinuity in the mechanism, e.g. a change in the electric nature of the substituent effects [21], in the rate-determining step [22] or in solvent complexation [23]. The sign and magnitude of such deviations are highly informative, frequently used for interpretation and pushes to further exploration.

For example, taking thoroughly into account Hammett's reference reaction (4) it is immediately clear that substituent effects measured by σ values are estimated from the process in which the substituent itself as well as the reaction site on both sides of eq. (4) are dependent on the solvent, and moreover they are not free of through-resonance effects. The carboxylic group is solvated in a different manner than COO^-. In turn, the substituent effect on both sides is differently modi-

fied by solvent effect. In addition, it should be mentioned that for electron-donor substituents in *para* position the substituent effect on the left-hand side is connected with an increase of charge density at *COOH* which is an electron-accepting group. Conversely, for electron-accepting *para*-substituents a decrease of δ-charge at COO^- is observed due to through-conjugation. In consequence, the measured K(X)-values contain both effects influenced additionally by hydration effects of those two groups. As a result, if one analyses how the $\rho\sigma$ equation (3) is fulfilled in other media two conclusions may be drawn [24]:

(i) regression coefficient ρ, called reaction constant, increases with a decrease of solvent polarity (e.g. ρ=1 in water, 1.72 in *EtOH*) and (ii) a decrease of goodness of fit is observed while going from water to water/organic solvent media for acid/ /base equilibrium (4).

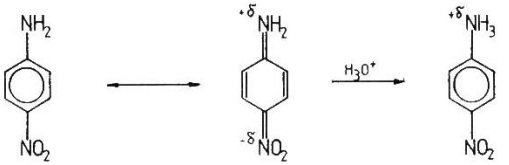
Systematic deviations from equation (3) for pK_a of *para*-substituted anilines and phenols found by Hammett himself [3] were explained by a strong through-resonance effect. This finding was an inspiration for further and deeper analysis of substituent effects, carried out by Wepster et al. [25]. They concluded that one scale of substituent constants cannot serve (accomodate) equally well for various reaction series differing in blend of resonance and inductive contributions. Hence, two additional scales of substituent constants were introduced.

4. THE σ^+ AND σ^- SUBSTITUENT SCALES.

These scales were introduced for a reaction series in which the reaction site Y is either Π-electron attracting (σ^+) and *para*-substituents are Π-electron donors, or Π-electron donating (σ^-) and *para*-substituents are Π-electron acceptors. For such systems the substituent effect is enhanced by an additional effect named: through-resonance, through-conjugation, intramolecular charge-transfer or Π-electron cooperative effect. A system in which these effects are present to a greater degree than in the reference reaction for Hammett's σ (4) needs a new set of substituent constants, which contain an increased contribution of resonance effect.

Analyses of the geometry of systems with a strong through--resonance effect support the classical view of interactions involved in these systems.

For electron donating reaction site Y, substituent constants σ^-, necessary for electron accepting substituents have to account for the intramolecular charge transfer. This is exemplified by appropriate canonical structures: B and Q of *para*-nitroaniline:



(Q) (B) (7)

As a result of protonation, the Π-electron cooperative effect (expressed by high contribution of Q-structure) is distinctly decreased. Comparison of %Q in these two species calculated di-

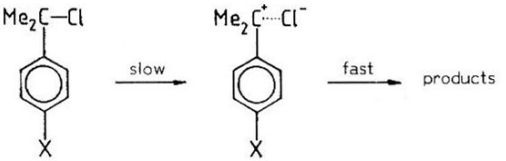
rectly from the geometry of *para*-nitroaniline [26] and *para*-dinitrobenzene which is electronically similar to *para*-nitroanilinum cation by use of the HOSE - method [27] yields 40.4 and 26.5% respectively. Obviously, for electron-accepting substituents X, the basicity of NH_2 - group decreases, i.e. the acidity of a conjugated acid NH_3^+ group increases, resulting in an enhancement of the measured pK_a-values. The *para*-X-aniline/anilinum ion acid-base equilibria are chosen as a reference reaction for σ^- substituent scale. For *meta*-substituted compounds eq.(3) holds with original Hammett σ_m for *meta* substituents. ρ for these systems is found to be 2.77, when deviating points on this graph are drawn for *para*-accepting substituents by δlog K(X)- value one can calculate $\sigma^- = \frac{\delta \log K(X)}{2.77}$.

The following explains the difference between Hammett's σ_p and σ^- substituent constants. For electron accepting substituents, σ^- are strongly influenced by a large contribution of structure Q, in aniline eq.(7). However, in its conjugate acid, anilinum ion, the contribution of structure Q is very low. In contrast, the influence of structure Q on both sides of eq.(4) is moderate: on the left side by electron donating substituents and on the right side by electron accepting substituents. It may be illustrated by %Q structure calculated by the HOSE model [27]. From the geometry of *para*-N,N-dimethylaminobenzoic acid Q=40.46% for electron donating substituents and Q=28.68% for electron accepting *nitro* group [29]. For *para*-nitrobenzoate anion [29] %Q is 30.1, but the COO^- group is involved in strong interactions with water molecules in crystal.

The picture presented above is in a good agreement with the classical interpretation of this effect. No doubt one observes really great changes in the geometry of benzene ring reflected in great differences in %Q-values for the systems with through--resonance and without it. A modern approach based upon *ab-initio* STO-3G calculations leads to a small difference between the energies of interactions for e.g. *para*-nitroaniline (2.2 kcal/ /mol) and *meta*-nitroaniline (0.0 kcal/mol) if one takes into account homodesmic reactions [30]

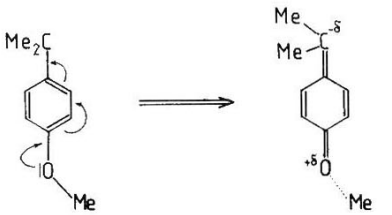
This discrepancy between structural picture (geometries) and energy view is as yet not fully understood.

Brown and Okamoto [31] account for the opposite effect in σ^+ substituent scale. The values of σ^+ were based on solvolysis of *para*-substituted *t*-cumyl chlorides in a 90% acetone-water mixture at 25°C. *Meta* substituents and those *para* which are electron accepting in character allow to draw a straight line with ρ=-4.54 .



(8)

In the transition state, the rate determining step resembles the carbonium ion intermediate and the substituents acting as electron donors influence its stability, as shown below



For points deriving from the Hammett line by δlog k one calculate $\sigma^+ = \delta\log k/(-4.54)$. An extensive list of σ^+ and σ^- values is given in a compilation by Exner [16], while Hansch and Leo [17] and also Hoefnagel and Wepster [32] reexamined many of σ^+ values and introduced some new values. Recently Binev et al. have introduced an extensive list of σ^+ parameters for anionic substituents [33].

5. SIMPLE OR MULTIPLE REGRESSION ?

The general equation (1) offers the possiblity of analysing substituent effects on the reaction center via simple or multiple regression, provided independent parameters are accessible. In our case independent parameters are substituent constants referring to a particular mode of substituent effect transmission. The regression coefficients in simple regression, also called reaction constants, ρ, cover roughly the range 0±4. A reaction which is facilitated by reducing the electron density at the reaction centre has a positive value of ρ, and one facilitated by increasing the electron density at the reaction

centre has a negative value. The standard reaction, dissociation of benzoic acids in water at 25°C has ρ=1.00 by definition [10].

More regression coefficients are calculated in a multiple regression model. Each describes the intensity of substituent action on the reaction center through a particular mechanism of action. The per cent value of the particular mechanism in complex phenomena can be calculated by applying eq.(2).

Nowadays, calculation of regression parameters is a trivial task. The least-squares method is used in programmable clacula-tors or various computers. The main problem for a researcher is to decide:

(i) how to use eq.(1),

(ii) how to choose explanatory parameters,

(iii) whether or not multiple regression is superior to simple regression, and

(iv) how to decide about goodness of fit.

Equation (1) may be used in different ways according to a philosophy of research. The most elegant way is to use simple regression and study the deviation of points, if any. In this approach one avoids any uncertainty which may be connected with intercorrelations between explanatory parameters in multiple regressions. Systematic deviation of points from the straight line means that interaction between substituent and reaction center varies while going from reference to the studied reaction. A distance between the experimental point and the straight line may be a measure of the given effect. For example, in the dissociation of *para*-nitroaniline

$\Delta\log K=(\log K_{experimental} - \log K_{predicted})$ is 0.49 . Hence, $-\delta\Delta G = 2.3RT\ \Delta\log K\rho = 0.68$ kcal/mol * 2.77 = 1.88 kcal/mol = = 7.97 kJ/mol. It is an energy of through resonance interaction of *nitro* group in $PhNH_2/PhNH_2^+$ system as compared with the reference reaction $PhCOOH/PHCOO^-$. Then one can use another well established and widely accepted scale of substituent parameters, describing different reference reactions (e.g. σ^- or σ^+).

The application of eq.(1) as a similarity model will be well supported by preliminary knowledge about the process under study. It needs some fundamental knowledge of various sets of substituent parameters and possible ways of transmission of substituent action.

The researcher must be aware of the fact that dispersion of the data due to substituent effect should be relatively uniform in the whole range. It is recommended to avoid data sets with one cluster of points and a single point far away from the cluster. The distribution of experimental points along substituent parameter axis should be uniform. Ehrenson et al. [19] suggest the following set of substituents: NMe_2, $NHMe$, NH_2, OMe (any two of them), CF_3, $COOR$, CH_3CO, CN, NO_2 (any two of them), H and Me, and two halogens, but not Br and Cl simultaneously. Such a collection of 8-10 *para* and 8-10 *meta* substituents will follow all requirements and will give good statistical evidence.

Multiple regression is an alternative to simple regression. It needs multiple sets of substituent parameters, each describing a single mechanism of substituent interaction on the reaction center. The explanatory parameters $A_{i,j}$ must be independent as far as their physicochemical nature is concerned and

mutually not intercorrelated. If this is strictly obeyed for j=1,2,....N, then eq.(2) may be applied to calculate per cent contributions of each mechanism. If explanatory parameters are not independent the interpretation of results may be false.

In chemistry substituents exert their influence on the character of the reaction center in no more than two or three major physically independent ways. If all mechanisms in the process under examination are identified, and substituent parameters are known, then it is quite easy to calculate regression coefficients in multiple regression. The following procedure [34] may be recommended to solve problems of unknown complexity.

A set of experimental data δQ_i is used in a simple regression model with j=1,2...N various explanatory parameters $A_{i,1}$, $A_{i,2}$ ....$A_{i,N}$. The parameters of linear regression, and correlation coefficients are calculated:

$$\delta Q_i = \alpha_1 A_{i,1} + B_1 \qquad r_1 \qquad (9\text{-}1)$$

$$\delta Q_i = \alpha_2 A_{i,2} + B_2 \qquad r_2 \qquad (9\text{-}2)$$

.....

$$\delta Q_i = \alpha_N A_{i,N} + B_N \qquad r_N \qquad (9\text{-}N)$$

The equation with the highest value of correlation coefficient r_k is selected,

$$\delta Q_i = \alpha_k A_{i,k} + B_k \qquad r_k \qquad (9\text{-}k)$$

and a set of two parameter equations is prepared

$$\delta Q_i = \alpha_k A_{i,k} + \alpha_1 A_{i,1} + B'_1 \qquad (10)$$

$$\delta Q_i = \alpha_k A_{i,k} + \alpha_2 A_{i,2} + B'_2$$

.....

$$\delta Q_i = \alpha_k A_{i,k} + \alpha_{N-1} A_{i,N-1} + B'_{N-1}$$

in which explanatory parameter $A_{i,k}$ is added to all (except 9-k) linear quations (9). The set of equations (10) describes the data set with total correlation coefficients $r_{k,1}$, $r_{k,2}$...$r_{k,N-1}$. Again the equation with the highest correlation coefficient, say $r_{k,l}$ is chosen

$$\delta Q_i = \alpha_k A'_{i,k} + \alpha_l A'_{i,l} + B'_l \qquad r_{k,l} \qquad (11)$$

Now it is necessary to check the hypothesis that the addition of $A_{i,l}$ to eq.(9-k) improves significantly the explained variance. The Fisher-Snedecor distribution is suitable for this purpose (for details see reference [35]). In other words, it is a test to prove whether or not replacement of eq.(9-k) by (11) is statistically justified. Therefore we calculate the Fisher - Snedecor $\mathcal{F}$-value for a situation in which differences are compared between residual variances for regressions (9-k) and (11) respectively, with the residual variance of regression (11) according to formula (12)

$$\mathcal{F}_{\alpha,f_1,f_2} = \frac{\sum_{i=1}^{M}(Q_i - \hat{Q}_i(\text{eq.9-k}))^2 - \sum_{i=1}^{M}(Q_i - \hat{Q}_i(\text{eq.11}))^2}{\sum_{i=1}^{M}(Q_i - \hat{Q}_i(\text{eq.11}))^2 / (M-2)} \qquad (12)$$

$\hat{Q}_i$ (eq.9-k) and $\hat{Q}_i$ (eq.11) stand for estimated values for Q_i predicted by eq.(9-k) and eq.(11), α is the significance level chosen for testing the hypothesis and f_1 and f_2 are degrees of freedom for the expressions in numerator and denominator; $f_1=(M-1)-(M-2)$ whereas $f_2=M-2$.

Equation (12) may be readily transformed into eq.(13) in which instead of residual variances one finds respective corre-

lation coefficients

$$F_{\alpha,f_1,f_2} = \frac{(r^2_{k,l} - r^2_k)^2 \ (M-2)}{(1-r_{k,l})^2} \qquad (13)$$

Usually correlation coefficients r_k, $r_{k,l}$ etc. are either published in papers or easily calculated by standard least-squares regression programs.

In order to decide whether or not significant improvement due to the addition of the second parameter has been achieved, one has to compare the calculated $\mathcal{F}$-value (eq.13) with the value of F_{α,f_1,f_2} taken from the statistical tables [35].

If

$$\mathcal{F}_{calc} > \mathcal{F}_{\alpha,f_1,f_2} \qquad (14)$$

one is allowed to accept the improvement of regression (11) relative to (9-k) at the α significance level. This means that one fails in α*100% of cases on accepting this improvement. If the improvement is significant, one may go further, adding to (11) the next explanatory parameter, say $A_{i,p}$ and following the procedure described above to test if this addition is significant, and continue in the same way until the addition of the next parameter becomes insignificant. For this procedure the estimation of per cent contribution due to $A_{i,k}$, $A_{i,l}$, $A_{i,p}$ etc. is somewhat different. Explanation of the total variance of δQ

$$\mathrm{var}(\delta Q) = \sum_{i=1}^{M} \frac{(\delta Q_i - \overline{\delta Q})}{M-1} \qquad (15)$$

by regression (9-k) is given by $100*r_k^2$. Then explanation due to the next parameter $A_{i,l}$ is given by $100(r_{k,l}^2 - r_k^2)$. For $A_{i,p}$ one has $100*(r_{k,l,p}^2 - r_{k,l}^2)$%. This kind of treatment gives however different results from those by eq.(2) for at least two reasons:

(i) eq.(2) does not take into account mutual correlation between explanatory parameters, whereas procedure [9-15] does

(ii) parameters $A_{i,k}$, $A_{i,l}$, $A_{i,p}$ etc. are not necessarily in exactly the same scale of magnitude. Even for σ_p, σ_I and σ_a it is a difficult task as long as high accuracy is required [36].

Goodness of fit needs some clarification. Jaffé [8] used correlation coefficient

$$r = \sqrt{b_x \, b_{yx}} \qquad (16)$$

to estimate how good is the model in description of the data. His choice has been criticized by many authors [37-39].

Critical remarks against the use of the correlation coefficient as a goodnes of fit parameter are as follows:

(i) r is a proper way to estimate mutual dependence of random variables set {x} and {y}. The calculated value of the correlation coefficient may be compared with tabulated values and hypotheses of dependence may be tested‡.
It seems not to be fully allowed to use the correlation coefficient in the same way for data sets which have been connected *a priori* by some model of similarity. Hence Jaffé [8] introduced an arbitrary scale of goodness of fit:

‡ Cf. the Appendix.

if $r \geq 0.99$ correlation is excellent

if $0.99 > r \geq 0.95$ correlation is satisfactory

if $0.95 > r \geq 0.90$ correlation is fairly good

and so far it is a most frequently used criterion of fit,

(ii) the correlation coefficient <u>does not</u> depend on degrees of freedom,

(iii) the correlation coefficient depends on the slope eq.(16),

(iv) the correlation coefficients are of low significence when two regressions are compared each of a different number of data,

(v) correlation coefficient should not be used for judging whether or not the additional parameter in eq.(1) improves regression significantly (which is often done).

Exner [38] introduced another parameter for goodness of fit

$$\psi = \left[\frac{n \, \Sigma(\hat{y}_i - y_i)^2}{(n-2) \, \Sigma(\bar{y}-y_i)^2} \right]^{1/2} \qquad (17)$$

ψ compares directly the scatter around the regression line, plane or hyperplane with the scatter about the mean value of the data to be explained; $n/(n-2)$ gives some influence of the degrees of freedom on the values of ψ .

The following arbitrary scale was proposed [38]

$\psi < 0.02$ correlation is very good

$0.02 \leq \psi \leq 0.1$ correlation is good

$0.1 < \psi \leq 0.2$ correlation is fairly good

$0.2 < \psi \leq 0.5$ correlation is poor

A rough interpretation of ψ may be as follows [40]: if for instance ψ=0.30 it means that the measured values of δQ are represented by the linear (or other) regression with a standard deviation of about 30% of that obtained by the simple assumption that the substituents have no effect on the reactivity (i.e. $y_i=\bar{y}$).

Other ways of presenting goodness of fit are there proposed by Ehrenson, Taft et al. [19], Swain and Lupton [41], and Koppel and Palm [43]. Their approaches are less accepted and, except by themselves, not often applied in the literature.

Another goodness of fit parameter, easy to estimate during the least-squares procedure, is standard deviation (estimated standard deviations are often denoted as σ). This quantity is a measure of precision with which the model reproduces experimental data. Of course, the criterion of 3σ may be used to indicate the outlier from the reaction series in question, as well as σ values may be used to compare precision of two reaction series.

6. SEPARATION OF INDUCTIVE AND RESONANCE EFFECTS

Hammett σ-constants measure the resultant of inductive and resonance effects of a substituent on the reaction center in reactions where the blend of both effects is nearly constant. Through-conjugation changes this blend and leads to σ^+ and σ^- scales, however these scales describe only some extreme blends. The contribution of the resonance effect of the substituent with respect to its inductive effect must in principle vary continuously as the electron property of the reaction center is

varied. Instead of a pair of discrete sets σ^+ or σ^-, a sliding--scale of substituent constants would be expected. Such a sliding-scale would reduce the value of Hammett equation. Several types of treatment have emerged to improve and rationalize the situation.

a). Definition of inductive parameters, σ_I

Taft and Lewis [43] set up a substituent inductive scale based on reactivity of alicyclic and aliphatic compounds. The substituent scale of inductive effect in aliphatic chains was introduced by Taft [44] for various $-CH_2X$ groups using well-defined conditions. Electron-withdrawing substituents have positive values of σ_I and electron releasing groups, negative ones, in accord with the theory of polar effect. Taft found that the rate or equilibrium constants for various reactions of *RY* could be represented by the equation

$$\log (K/K_o) = \rho^* \sigma^* \qquad (18)$$

Some reactions of *o*-substituted aromatic systems conformed to the above equation, too; σ^* is much less applied than the other parameter, σ_I. This was based on the dissociation reaction in 50% $EtOH$-H_2O at 25°C of 4-substituted bicyclo-[2.2.2]- octane--1-carboxylic acids and reactivity of their esters [11] or *trans*-4-substituted-cyclohexane-1-carboxylic acids [45]. Especially the bicyclooctane moiety provides a good geometrical model for *para*-substituted benzoic acids without the complication by the resonance effect. For that reason σ_I was intensively used to understand substituent effect in aromatic compounds.

Although the σ_I scale was originally called an inductive parameter, there is much evidence in the literature to the ef-

fect that it is in fact a field parameter [46]. For many years most workers seemed to have assumed that the field effect and σ-inductive effects were linearly related. Consequently, both effects could be described by a single parameter for a simple substituent. This assumption is no longer supported by experiment and theory (as discussed in the last section). Field and σ-inductive effects are in principle not proportional.

b). Definition of resonance parameter, σ_R

Taft and Lewis suggested [43] that the effect of substituents in aromatic compounds should be separable into inductive and resonance contributions:

$$\log (K/K_o) = \rho_I \sigma_I + \rho_R \sigma_R \qquad (19)$$

using a two-parameter equation. The inductive scale σ_I was proposed as described previously, and σ_R, the resonance parameter, was calculated based on the assumption that parameters σ_I operate equally from *meta* and *para* positions. Then

$$\begin{aligned} \sigma_p &= \sigma_I + \sigma_R \\ \sigma_m &= \sigma_I + \alpha\sigma_R \end{aligned} \qquad (20)$$

α being the 'relay coefficient' of magnitude 0.33 . Further attempts at solving the problem was undertaken by Exner [47] who questioned the equal contribution of inductive effect from *meta* and *para* positions and argued in favour of stronger inductive operation from *para* position. This final equation

$$\begin{aligned} \sigma_p &= \lambda\sigma_I + \sigma_R \\ \sigma_m &= \sigma_I + \alpha\sigma_R \end{aligned} \qquad (21)$$

had α=0.33 and λ=1.14 . However, α and λ values lack rigorous justifications. Also the ionization of benzoic acids is not a satisfactory process [25]. To avoid this problem the σ° substi-

tuent scale was established based on ionization of substituted phenylacetic acids, and used to define resonance parameter σ_R° assuming α=0.5 . In addition σ_R^- and σ_R^+ substituent scales were reported based on σ^- and σ^+ scales. Many correlations with eq.(19) have been reported by using σ_I and one of four resonance parameters [48]. Comparison of the different resonance scales shows that the correlation between them is rather limited. This means that they change when the electron--demand of the reaction changes. An extensive review on σ_R^+ scale has been published lately [49].

c). Yukawa-Tsuno equation.

Yukawa and Tsuno [50] have introduced an equation in which deviations from the Hammett equation due to through-resonance effect may be treated by use of the equation

$$\delta \log k = \rho\left[\sigma_{m,p} + r(\sigma^+ - \sigma^-)\right] \qquad (22)$$

For the opposite case, where the reaction site is electron-accepting and the substituents are electron-donating Yoshioka et al. [51] introduced a similar equation

$$\delta \log k = \rho\left[\sigma_{m,p} + r(\sigma^- - \sigma_p)\right] \qquad (23)$$

Both of them reduce to the simple Hammett equation (3) when r=0; it is the case when the through-conjugation effect between substituents *X* and the reaction site *Y* is negligible in comparison with reaction (4).

These two approaches need different sets of substituent constants, and pose different statistical problems. The linear approach (eq.(1) with N=1) may be applied either with substituent constants composed of different "contributing" effects, as e.g. σ_p, σ_m (of Hammett) or substituent constants

describing "purely" one mechanism of interaction e.q. σ_I, σ_R etc. In both cases systematic deviations from a straight line, if not due to experimental errors, may be interpreted.

Molecular systems such as $X\text{-}CH_2\text{-}C_6H_4\text{-}Y$ or $X\text{-}C_6H_4\text{-}CH_2\text{-}Y$ are expected to be unaffected by any through-conjugation effect. Substituent constants σ^n or σ° were defined [25,52] and recently used in a contemporary version of the Yukawa-Tsuno equation

$$\delta Q = \rho \left[\sigma^{n \text{ or } \circ}_{m,p} + r \, (\sigma^{+,-} - \sigma^{n \text{ or } \circ}_{p}) \right] \qquad (24)$$

Differences between σ^n and σ° are practically insignificant [25] and both scales may be used in (24). In equations like (22-24) r is a measure of similarity between a given reaction series (data set δQ) and two reference reaction series:

(i) in which there is no through-conjugation effect; similarity is full if r=0, and

(ii) in which there are through-conjugation effects such as those observed in reactions (7) and (8); similarity is full if r=1 .

In other words, r is a measure of "exaltation" of through--conjugation effect over the $X\text{-}C_6H_4\text{-}CH_2\text{-}Y$ (or $X\text{-}CH_2\text{-}C_6H_4\text{-}Y$) systems. Taft suggested [53] to use eq.(1) with explanatory parameters σ_R and σ_I accounting for resonance and inductive effects, respectively, which in turn determine the overall substituent effect observed in δQ, for the reaction series in which Hammett equation fails.

Evidently, σ_I's are closely related to σ^n or σ^p but they were introduced much earlier by Roberts and Moreland [11]. Then similar systems were employed to define σ_I, including even *meta* and *para*-substituted toluic acids [47]. In the next two decades

the idea of using eq.(1) with two explanatory parameters was developed in various ways. Eherson, Brownlee and Taft [19] re-examined it and suggested a generalized treatment of substituent effects by the use of dual substituent parameter equation. They postulated to use one set of substituent constants accounting for an inductive effect, σ_I, and one out of four sets of substituent constants accounting for different ways of interactions in *para*-X-C_6H_4-Y systems:

- σ_R (BA) for benzoic acid like systems,
- σ_R^- (A) for aniline like system,
- σ_R^+ based on Eaborn's [54,55] protonolysis rates of *para*-substituted phenyltrimethylsilanes,
- σ_R° for a reaction series in which neither electron donating nor electron accepting substituents affect the reaction centre by <u>through-resonance</u> effect; the *F-nmr* shielding effects for *para*-F-C_6H_4-X [56] was used to establish such a scale.

These four scales of σ_R constants and one of σ_I were used by Taft et al. in their extensive analysis of substituted benzene derivatives [19]. It seems, however that when using the above-mentioned scales one cannot avoid a certain inaccuracy. Namly each σ_R-scale (σ_R (BA, A, $^\circ$, and $^+$)) contains some inductive contribution which is then analysed again by a formally independent explanatory parameter σ_I. Then $\lambda=\rho_R/\rho_I$ measures the blend of inductive and mesomeric contributions in the overall substituent effect. Since σ_R are not pure mesomeric constants, the significance of addition of the next explanatory parameter must be tested and per cent of additional explanation of the total variance should be calculated according to equations (9-14). Otherwise an unknown effect of the common variance in σ_I and σ_R may lead to confusing interpretations.

The problem of accuracy and independence of explanatory parameters in eq.(1) has been throughly discussed by Charton [57], who critizised the separation made by Exner [47].

Also the separation proposed by Swain and Lupton [41] met with a strong oposition [58-61]. Swain and Lupton [41] applied factor analysis [12] of experimental data in order to sort out two factors, e.g. inductive and mesomeric effects. The set of experimental data was large, and included many reactions and various substituents. In the first step of factor analysis a good fit between the observed and calculated data has been achieved with two factors. Two critical and four subsidiary conditions were necessary to transform the solution into scales having simple chemical significance, e.g. inductive and mesomeric. The most controversial part of this work [41] is the second critical condition in the set of two:

(i) *trans* 4-substituents in cyclohexane carboxylic acids [45] or 4-substituents in bicyclo-[2.2.2] octane carboxylic acids [41] exert no resonance effect on the carboxylic group and,

(ii) the $(CH_3)_3N^+$ substituent is never more electron donating or more electron attracting than *H* by resonance [41,45].

These two critical conditions accompanied by subsidiary conditions permitted to define the inductive scale $\mathcal{F}$ and the resonance scale $\mathcal{R}$ for more than 40 substituents. $\mathcal{F}$ and $\mathcal{R}$ substituent scales have been quite extensively used, particularly in correlation of spectroscopic and biological data [2,5] with equation

$$\delta Q = f\mathcal{F} + r\mathcal{R} + h \qquad (25)$$

This approach has been seriously criticized [58-61] and is still under dispute [13,62,63].

A similar method has been used to extract three substituent parameters from an experimental body of data including 76 reaction series and 17 substituents [64]. Among the reaction series were various reactions not following the σ_I-σ_R Taft model. In contrast to the Swain-Lupton procedure [41,45], no arbitrary restrictions were imposed. The primary statistical solution with three factors has been converted into real solution by rotation, and three substituent constant sets were derived: σ_I, σ_R and σ_E. σ_E represents a mixture of unknown composition of the substituent effects, σ_I and σ_R are connected to inductive and mesomeric effects. The three parameter equation has been claimed [65] as "optimum linear free-energy relationship" for the prediction of missing data on substituent effects.

In another stream of works [66-67] pattern recognition by means of the disjoint principal components model SIMCA has been applied to the problem of separation of inductive and mesomeric effects of substituents. The resulting σ_I scale [66] corresponds as closely as possible to an inductive effect operative in aromatic reactions. Extraction of σ_I was based on 15 substituents and 46 reactions.

The strength of above approaches lies in the fact that it considers large data sets simultaneously, and resulting substituent constants represent an average of all data. It is possible at any moment to include new data or replace the previous critical conditions by new ones, if such data are available and are better than those originally proposed.

Finally, it should be mentioned that a very important theoretical approach facilitating better understanding of the complex situation in systems with through-conjugation has been proposed by Reynolds et al. [68] who applied *ab initio* (STO 3G or 4-31G) calculations for a series of substituted probe molecules covering a wide range of Π-electron demand. These calculations have revealed a complex pattern of substituent resonance response to a varying electron demand.

7. STERIC EFFECTS

Equation (1) may be used for all physical and chemical properties in any reaction series provided appropriate explanatory parameters are available. As a rule, for *meta* and *para*-substituted derivatives of benzene, and even for more complex systems in which resonance and induction predominate, eq.(1) is used with one of the substituent constant scales presented previously. However, when steric effects may interfere, the situation becomes more complex. The *ortho* substituents cannot be treated in the same way as indicated by Hammett [3] for the rate of hydrolysis in *ortho*-substituted alkyl benzoates plotted against ionization constants of reference acids. Linear regression is unseccessful for *ortho* substituted compounds because of the following factors [9]:

(i) electrical effects which may be resolved into localized (inductive) effect and delocalized (resonance) effect,

(ii) steric effects which are a function of substituent bulk; they may be resolved into: (α) steric hindrance to solvation and may involve the substituent or the reaction site, or both; (β) steric hindrance of the reaction site to be

attacked by the reagent; (γ) steric inhibition of resonance between the aromatic ring and the substituent or the reaction site, or both; and (δ) steric control of the transition state conformation,

(iii) intramolecular secondary bonding forces: (α) hydrogen bonding, (β) Keesom (dipole-dipole), Debye (dipole-induced dipole), London (induced dipole-induced dipole) forces, (γ) charge transfer reactions.

It is obvious that due to such complex interactions possible between *ortho* substituents, linear regression fails. Charton [69] demonstrated that when eq.(1) is applied to reaction rates and equilibria of *ortho*-substituted compounds with σ_I and σ_R substituent scales as explanatory variables, the per cent values of resonance contribution vary from 20 to 52% . For comparison, variance of resonance contribution in various σ_p scales is much smaller and amounts to 50% in σ_p, 59.5% in σ_p^- and 61.5% in σ_p^+ [69]. Per cent of resonance is 24.8% in σ_m. In paper [69] instead of %R the blend is given as $\rho_R/\rho_I=\varepsilon$ and $\%R=(\varepsilon/\varepsilon+1)*100$. If other *ortho*-substituted species are analyzed, %R may range even between 6 and 82% . It was concluded [9] that the enormous range of variation in %R makes it impossible to define a single set of *ortho*-substituent scale.

An important contribution in the field is due to Taft [44, 70] who elaborated quantitatively Ingold's method of polar and steric effects separation in hydrolysis of esters [71]. The work of Taft has been reviewed by Shorter [72,73].

The basic idea of Taft's approach was to separate polar, steric and resonance effects. The polar effect described by polar substituent constant σ^* was defined by eq.(26):

$$\sigma^{*} = \left[\log\left(\frac{k}{k_o}\right)_B - \log\left(\frac{k}{k_o}\right)_A\right] / 2.48 \qquad (26)$$

The rate k refers to the reaction of *R-COOR'* and k_o refers to the reaction of CH_3COOR' as standard. A and B stand for acidic and basic hydrolysis, respectively, carried out with equal *R'*, solvent and temperature. The factor 2.48 puts σ in the same range of magnitude as Hammett σ.

The second term in eq.(26) was named steric substituent constant E_s

$$E_s = \log\left(\frac{k}{k_o}\right)_A \qquad (27)$$

on the grounds of the following assumptions:

(a) the relative changes of the free energy of activation may be treated as a sum of three independent contributions from polar, steric and resonance effects,

$$\delta\Delta G = \alpha_p \delta\Delta G_p + \alpha_s \delta\Delta G_s + \alpha_r \delta\Delta G_r \qquad (28)$$

$$(\alpha_p \neq \alpha_s \neq \alpha_r)$$

(b) the steric and resonance effects are equal in acidic or basic hydrolysis

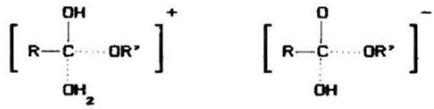
$$\delta\Delta G_s^A = \delta\Delta G_s^B$$

$$\delta\Delta G_r^A = \delta\Delta G_r^B \qquad (29)$$

(c) the polar effects of substituents are considerably stronger in the basic hydrolysis.

Assumption (a) is necessary to carry out any analysis of this kind. Assumption (c) is supported by the magnitude of reaction constant ρ_o. The hydrolysis of substituted benzoates in alkaline media proceeds with ρ in the range 2.2-2.8, whreas in acidic media ρ it is small, in the range 0.2-0.5 . The most controversial is assumption (b). It is believed that the transition states A and B, in acidic and basic media, closely resemble the

following structures:



(30)

(A) (B)

and differ by two protons. Due to the small size of the protons, the difference between steric interactions caused by various substituents, R, should be essentially the same in both acidic (A) and basic (B) transition states. Thus, steric effects diminish in eq.(26), whereas in eq.(27) they are the dominating contribution to variation in $\log(k/k_o)_A$.

At present σ_I is seldom used as a measure of polar effect. The steric constant E_s was modified in part by inclusion of hyperconjugation effect [74,75]. Charton recognized that steric constants may be represented by the Van der Waals size of substituents [58,76]:

$$\theta = r_{VX} - r_{VH} = r_{VX} - 1.20 \qquad (31)$$

where r_{VX} is the Van der Waals radius of the substituent X (in Angströms), and r_{VH} is the corresponding one for hydrogen atom. According to more recent results [76,77], correction for hyperconjugation is not necessary.

In recent years there has been some disputation as concerns the use of various scales of steric effects [40,78,79] and statistical evidence based upon a large experimental material is given by Charton [17,76,80-82]. Additional steric constants have been established [83] for some new systems [84], mainly related to branched alkyl groups [85] and bulk substituents in biologicaly active compounds [86].

8. APPLICATION OF QUANTUM CHEMISTRY TO RATIONALIZE SUBSTITUENT EFFECTS

The substituent scales presented above cover classical approaches to substituent effects using a linear regression model. Now a separate chapter is needed in order to present some new trends, offered by theoretical chemistry.

Quantum chemistry from the very beginning had been used to rationalize various substituent effects. A rapid development of *ab initio* techniques in the last decade has permitted to use them as a tool precise enough to study particular "contributions" operating in the overall substituent effect. An extensive review in this field [86] was supported by earlier contributions [87-89]. The agreement obtained between theoretical and experimental data allows one to relate theoretical transmission models which may not be measurable in practice, to chemical reality.

Substituent electronic effects [90] have their origin in the following factors : the substituent dipole, the electronegativity difference between the substituent and the atom to which it is attached, and charge transfer between the substituent and the group to which it is attached. These factors lead directly to three substituent effects. The field effect [90,63] involving a direct through-space interaction, is the predominant mechanism of transmission where one or more atoms separate the substituent from the center reaction. This effect arises from the dipole moment. Another way of transmission is the progressively diminishing relay of polar effect along a chain of carbon atoms originating from the electronegativity of substi-

tuent. It is called the σ-inductive effect. The third, resonance effect depends on the ability of the substituent to donate or accept Π-charge to/from a conjugated Π-system.

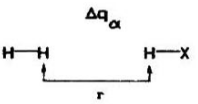
All mentioned effects were defined and characterized by *ab initio* STO-3G level and higher basis calculations.

A). Substituent Field Effect and Parameters

Recently two ways of estimating theoretical scales of field effects have been suggested [87,91]. The first of them is the "theoretical reaction" of equilibrium (32)

(32)

in which any possibility of indirect polarization effects is avoided and the energy of reaction (32) is a direct measure of the field effect. The other way is to calculate the relative polarization of electron population in a hydrogen molecule by an isolated H-X molecule at constant r-distance



(33)

Then, the field substituent constants are defined as (34):

$$\sigma_F = -0.074\ \Delta E \qquad (34a)$$

and

$$\sigma_F = -35.5\ \Delta q_H \qquad (34b)$$

The agreement between these two methods is excellent [91]. When σ_F's (34b), which are less sensitive to polarization effects of such substituents as NH_2 or NMe_2, are plotted against σ_F determined experimentally (σ_I from ref.[36]) the regression is very good [91]

$$\sigma_F(34b) = 0.93\sigma_F(exp) + 0.03 \qquad (35)$$

with correlation coefficient r=0.986 .

Thus, the old σ_I (called by Topsom et al. [92,93] σ_F) are very well supported by the theoretical model and visualize well the mechanism of interaction.

B) Substituent Electronegativity Parameters as a measure of σ-inductive effect

Electronegativity of substituent represents the power of the atom (or a group of atoms constituting the substituent) in a molecule to attract electrons. Mariott et al. [92] have recently suggested to use the charge at *H*-atom in *H*-*X* systems as a quantitative measure of substituent electronegativity. Thus substituent electronegativity constants were defined as:

$$\sigma_\chi = 1 - q(H) \qquad (36)$$

It was shown that σ_χ correlates well with electronegativity of Allred and Rochow [93] and Boyd and Marïeus [94] but rather poorly with σ_F. Thus it is worth mentioning here that σ_χ and σ_F express two independent mechanisms of "inductive" effects which may affect the reaction site without participation of Π-electrons. The transmission of σ-inductive effects is not important beyond the second atom [46,87-89]. Such σ_χ values thus provide a simple and well-defined scale of electronegativity-corresponding to σ-inductive effect, for a wide range of substituents.

C) Theoretical Scale of Substituent Resonance Parameters (σ_p°)

Theoretical scales of resonance effects are more difficult to define since the resonance effects vary according to the Π--electron demand of the substrate to which the substituent is attached [68]. Thus, the Π-electron response of a substituent when attached to an unperturbed benzene ring (σ_R°) may be markedly different from that attached to Π-electron - attracting (σ^-) or Π-electron - donating (σ^+) systems. Substituent X attached to a Π-electron system such as benzene or ethene changes their Π-charge and these changes are the measure of the resonance effect of σ_R° - type:

$$\sigma_R^\circ = a\,\Sigma\Delta q_\Pi + b \qquad (37)$$

a,b - coefficients depend on the level of basis set; for 4-31G, a=0.004 and b=0.075 . σ_R° are experimental values of *IR* intensity for *mono*-substituted benzenes [95] and $\Sigma\Delta q_\Pi$ is a sum of Π-electron changes at all carbon atoms in benzene ring (or ethene skeleton).

D) *ab initio* Interpretation of Hammett's σ_p and σ_m .

From the very beginning quantum chemistry had been applied to interpret Hammett's σ_m and σ_p [96]. Recently Hammett's classical σ_m and σ_p have been successfully treated by triple parameter regression with Mulliken Π- and σ- or total charge densities on *meta* and *para* C- and H- atoms.

$$\sigma_{(m\ or\ p)} = a\Delta q^\Pi + b\Delta q^\sigma + c\Delta q^H + d \qquad (38)$$

The results obtained for 12 *meta* and 10 *para* data points were surprising.

The most effective explanatory parameters for *para* positions are $\Delta q_{tot} = \Delta q^{\sigma} + \Delta q^{\Pi}$ and, to a lesser extent, Δq^{Π} (88.6 and 88.3% of explained variance in σ_p). If both these parameters were used together the percentage of variance explained by this model rose up to 94.5% . Separation of Δq into Δq^{σ} and Δq^{Π} is not effective and hence corroborates with the former finding by Taft [98] and Exner [99] that the ratio of σ to Π contributions for substituent effects in *para* positions is close to 1.

A much more difficult situation is for *meta* position. The most effective single parameter regression for σ_m is that with Δq^{H}; it explains 85.8% of variance. The addition of Δq^{σ} and Δq^{Π} as explanatory parameters ends up with explanation of the total variance in σ_m equal to 96.1% . In this regression, except for Δq^{Π}, the contribution due to σ electrons is three times greater than that for Π-electrons. It is in a good agreement with the finding of Charton [57] where the per cent of σ-electron effect in σ_m was estimated as 72% .

The theory for generalized substituent effects has been reviewed by Taft [100] in a form of three parameter equation:

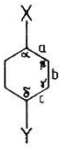
$$\delta\Delta G = \rho_R \sigma_R + \rho_F \sigma_F + \rho_P \sigma_P \qquad (39)$$

in which σ_R, σ_F and σ_P describe : substituent resonance or Π-electron delocalization, substituent field/inductive effect and substituent polarizability effect, respectively. Substituent constants were calculated by the STO-3G minimal basis set method for proton transfer reactions of amines and anilines. A similar general discussion of substituent effects in terms of partial electrical effects was performed by Charton [101].

9. SUBSTITUENT EFFECT ON THE GEOMETRY OF MOLECULES

Due to the enormous development of both computers and X-ray difractometry techniques, a large number of relative by precise data on the geometry of substituted systems appeared in the last decades. Then a few attemps at their rationalization have been published. Hence, because of their importance for a better understanding of the relationships between structure (geometry) and reactivity of the substituted species, the most significant results are shortly presented in this section.

The first complete and still important approach was that presented by Domenicano, Vaciago and Coulson [102,103] who found that in substituted benzene derivatives the substituent affects chiefly the α-angle at *ipso* carbon atom in the ring and both a-bonds. These changes may be rationalized by use of either the Walsh rule [104,105] or VSEPR - model [106]. Then it was found that the α-angle may be linearly related to the electronegativity of X its inductive substituent constant [107]. More recently it has been shown that if a quantity defined as $\Delta=b-a$, (i.e. the difference between b- and a- bond lengths) varies linearly with α for *mono*- and symmetrically *para*-disubstituted benzenes [108]. This is an evident support for using the Walsh rule to interpret the substituent effect in *mono*- and *para*-disubstituted benzene derivatives, provided *para*-disubstituents do not interact by the Π-electron cooperative effect (through-conjugation).



Another important step forward was made by Norrestam and Schepper [109] and Domenicano and Murray-Rust [110]. They have

introduced a concept of angular substituent parameters, $\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$ and $\Delta\delta$, which describe the difference between angles α,β,γ and δ in monosubstituted benzene derivatives and in benzene itself. These quantities permit to predict substituent effects on the geometry (angles) in the ring provided there are no strong steric or Π-electron interactions between the substituents in question. Two lists of angular parameters have been published. One of them is based on *mono-* or weakly interacting *para*-disubstituted benzene derivatives [110] and it is preferentially recommended to study the additivity of substituent effect on the geometry of the substituted species. The other [109] is based on a variety of polysubstituted benzene (and even pyridine) derivatives, including strongly interacting system. These angular substituent parameters may be used to reproduce the geometry of a relatively wide range of substituent species. The problem arises however for substituents which are built up of three atoms bearing Π-electrons - their Π-electron system may either interact with that of benzene ring if the planes of both systems are nearly parallel, or their mutual interactions are hindered, if their planes are far from coplanarity. This problem is not solved in general, but Norrestam and Schepper [101] estimated angular substituent parameters for two kinds of *nitro-* groups: coplanar (or nearly coplanar) and bent (with angles of bend $\varphi \geq 35^{\circ}$). Application of these parameters to interpret intramolecular interactions between the substituents is presented in a few recent papers [26,29,111-113]. Some new parameters have recently been estimated as well, for *acetoxy*-group [29] and for COO^{-}-group [114].

In the case of *para*-substituted benzene derivatives with strongly interacting substituents it was found [26,115,116] that the angles at the substituted carbon (α and δ) do not follow the additivity rule, and often highly surpass the predicted value. Since these angles are expected to depend less on Π-electron than on σ-electron effects it appears that much more sensitive quantities to study the through-conjugation effect are the Δ-values, defined as $\Delta=b-a$ or $\Sigma\Delta=(b-a)+(b-c)$. It has been found recently [108] that the quantity $\Sigma\Delta$ depends linearly on σ^+ values of the substituent being the Π-electron donor, whereas the dependance on σ^- is completely insignificant. This is an important result since it seems to suggest that the Π--electron donating substituents affect the geometry (bond lengths) of the ring much more effectively than the Π-electron accepting ones of the *para*-counter substituent. At present it may be said that geometrical features of the molecules (valence angles and bond lengths) affected by the substituent are related in a regular way to substituent parameters described in principle in terms of molecular reactivity. However, it should be also pointed out that the above-mentioned geometrical parameters may be subject to intermolecular interaction in the crystal lattice [116] and then may be contaminated by an unknown uncertainty of the parameter in question. Hence only the most precise data should be taken for any structure-reactivity analysis and throughly examined from the point of view of close intermolecular contacts which may be a source of deformations [117].

10. CONCLUSIONS AND RECOMMENDATIONS

In our opinion the application of eq.(1) or its modification is most fruitful for two purposes: (a) prediction of data unavailable from direct measurements and (b) interpretation of experimental data for less know processes and reaction series.

For the former purpose (a) a multiparameter version of (1) is most profitable for its low values of estimated standard deviation, i.e. for its relatively high precision of prediction. For the latter purpose (b), eq.(1) should be used either as a linear model of similarity through analysis of deviating points or as a multiparameter version of (1) but with the certainty that no colinearity exists between the explanatory parameters. If this condition is not exactly fulfilled the step with regression as shown by equations (9-16) should be applied. It is important to use high quality data as explanatory parameters. In many compilations one can find many possible parameters for a given variable and manipulation with them may increase the risk of improper interpretation.

REFERENCES

1. N.B. Chapman and J. Shorter (Ed.), Advances in Linear Free Energy Relationships, Plenum Press, New York, 1972
2. N.B. Chapman and J. Shorter (Ed.), Correlation Analysis in Chemistry, Recent Advances; Plenum Press, New York, 1978
3. L.P. Hammett, Physical Organic Chemistry, McGrew-Hill, New York, 1940
4. C.D. Johnson, The Hammett Equation, Cambridge University Press, Cambridge, 1973

5. J. Shorter, Correlation Analysis of Organic Reactivity, Research Studies Press, Plenum Press, London 1982

6. Ch. Reichardt, Solvent Effect in Organic Chemistry, Verlag Chemie, Weinheim 1979

7. J.E. Leffler and E. Grunwald, Rates and Equilibria of Organic Reactions, Wiley, New York, 1963

8. H.H.Jaffé, Chem. Rev., 53, 191 (1953)

9. M. Charton, Prog. Phys. Org. Chem., 8, 235 (1971)

10. L.P. Hammett, Chem. Rev., 53, 175 (1935)

11. J.D. Roberts and W.T. Moreland, J. Am. Chem. Soc., 75, 2167 (1953)

12. E.R. Malinowski and D.G. Hoovery, Factor Analysis in Chemistry, Wiley, New York, 1980

13. C.G. Swain, J. Org. Chem., 49, 2005 (1984)

14. D.H. McDaniel and H.C. Brown, J. Org. Chem., 23, 420 (1958)

15. J.D. Roberts, E.A. McElhill and R. Armstrong, J. Am. Chem. Soc., 71, 2923 (1949)

16. O. Exner, in ref. 2, page 439

17. C. Hansch and A. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology,

18. V.A. Palm, Osnovy Kalitschestvennoy Teorii Organitscheskich Reakcji, Izdatelstvo Kchimia, Leningrad, 1967

19. S. Ehrenson, R.T.C. Brownlee and R.W. Taft, Progr. Phys. Org. Chem., 10, 3 (1973)

20. O. Exner and J. Jonaś, Coll. Czech. Chem. Commun., 27, 2296 (1962)

21. D.N. Kershaw and J.A. Leisten, Proc. Chem. Soc., 1960 84

22. B.M. Anderson and W.P. Jencks, J. Am. Chem. Soc., 82, 1773 (1960)

23. C.D. Johnson, The Hammett Equation, Cambridge University Press, 1973, Ch.2

24. T.M. Krygowski and W.R. Fawcett, Can. J. Chem., 53, 3622 (1975)

25. H. van Bekkum, P.E. Verkade and B.M. Wepster, Rec. Trav. Chim., 78, 815 (1958)

26. M. Colapietro, A. Domenicano, C. Marciante and G. Portalone, Z. Naturfosch., 37B, 1309 (1982)

27. T.M. Krygowski, R. Anulewicz and J, Kruszewski, Acta Cryst., B39, 232 (1983)

28. R. Anulewicz, G. Häfelinger, T.M. Krygowski, C. Ragelman and G. Ritter, Z. Naturforsch., 42b, 917 (1987)

29. I. Turowska-Tyrk, M. Gdaniec, T.M. Krygowski, G. Häfelinger and C. Ritter, J. Mol. Struct., *in press*

30. W.J. Hehre, L. Radom, P.v.R.Schleyer and J. Pople, *Ab initio* Molecular Orbital Theory, Wiley and Sons, London, 1986

31. H.C. Brown and Y. Okamoto, J. Am. Chem. Soc., 80, 4979 (1958)

32. A.J. Hoefnagel and B.M. Wepster, J. Am. Chem. Soc., 95, 5357 (1973)

33. I.G. Binev, R.B. Kuznanova, J. Kaneti and I.N. Juchnowski, J. Chem. Soc. Perkin Trans II, 1533, 1982

34. T.M. Krygowski, J.P. Radomski, A. Rzeszowiak, P.K. Wrona and Ch. Reichardt, Tetrahedron, 37, 119 (1981)

35. G.W. Snedecor and W.G. Cochran, Statistical Methods, Iows State University, Ames, 1975

36. O. Exner, in ref.1, page 36

37. C.K. Hancock, J. Chem. Educ., 42, 608 (1965)

38. O. Exner, Coll. Czech. Chem. Commun., 31, 3222 (1976)

39. W.H. Davis and W.A. Pryor, J. Chem. Educ., 53, 285 (1975)

40. Ref.5, page 217

41. C.G. Swain and E.C. Lupton, J.Am. Chem. Soc., 90, 4328 (1968)

42. A. Koppel and V. Palm, in Ch.5 of ref.1

43. R.W. Taft and I.C. Lewis, Tetrahedron, 5, 210 (1959)

44. R.W. Taft, in "Steric Effects in Organic Chemistry" Ch.13, M.S. Newman Ed. Wiley 1956

45. C.G. Swain, S.H. Unger, W.R. Rosenquist and M.S. Swain, J. Am. Chem. Soc., 105, 492 (1983)

46. W.F. Reynolds, J. Chem. Soc. Perkin Trans II, 985, 1980

47. O. Exner, Coll. Czech. Chem. Commun., 31, 65 (1966)

48. R.T.C. Brownlee, S. Ehrenson and R.W. Taft, Progr. Phys. Org. Chem., 10, 1 (1973)

49. M. Charton, Molecular Structure and Energetics, Vol. 4, Ch.4, Ed. J. Liebman and A. Greenberg, Verlag Chemie, 1987

50. Y. Yukawa and Y. Tsuno, Bull. Chem. Soc. Japan, 32, 971 (1959)

51. M. Yoshioka, K. Hamamoto and T. Kubota, Bull, Chem. Soc. Japan, 35, 1723 (1962)

52. R.W. Taft, S. Ehrenson, I.C. Lewis and R.E. Glick, J. Am. Chem. Soc., 81, 5352 (1959)

53. R.W. Taft, J. Am. Chem. Soc., 79, 1045 (1957)

54. C. Eaborn, J. Chem. Soc., 4858, 1956

55. E.B. Dean and C. Eaborn, J. Chem. Soc., 2299, 1959

56. R.W. Taft et al., J. Am. Chem. Soc., 85, 3146 (1963)

57. M. Charton, Progr. Phys. Org. Chem., 13, 119 (1981)

58. M. Charton, Progr. Phys. Org. Chem., 10, 81 (1973)

59. W.F. Reynolds and R.D. Topsom, J. Org. Chem., 49, 1989 (1984)

60. M. Charton, J. Org. Chem., 49, 1997 (1984)

61. A.J. Hoefnagel, W. Oosterbeck and B.M. Wepster, J. Org. Chem., 49, 1993 (1984)

62. M. Charton, J. Org. Chem., 36, 266 (1971)

63. R.D. Topsom, Progr. Phys. Org. Chem., 12, 1 (1976)

64. G.H.E. Nieuwdrop and C.L. de Ligny, J. Chem. Soc. Perkin Trans II, 537, 1979

65. C.L. de Ligny and H.C. Van Houvelingen, J. Chem. Soc. Perkin Trans II, 559, 1987

66. M. Sjöström and S. Wold, Chemica Scripta, 6, 114 (1974); S. Wold, Chemica Scripta, 6, 97 (1974)

67. S. Alunni, S. Clementi, U. Edlund, D. Johnles, S. Halberg, M. Sjöström and S. Wold, Acta Chem. Scand., B37, 47 (1983)

68. W.F. Reynolds, P. Pais, D.W. MacIntyre, R.D. Topsom, S. Marriott, E. von Nagy-Kelsobuki and R.W. Taft, J. Am. Chem. Soc., 105, 378 (1983)

69. M. Charton, J. Org. Chem., 30, 3341 (1965)

70. R.W. Taft, J. Am. Chem. Soc., 74, 3120, 2729 (1952); 75, 4231, 4534, 4538 (1953)

71. C.K. Ingold, J. Chem. Soc., 1032, 1930

72. J. Shorter, Quart. Rev., 24, 433 (1970)

73. J. Shorter, Ch.2 in ref.1

74. C.K. Hancock, E.A. Meyers and B.J. Yager, J. Am. Chem. Soc., 83, 4211 (1961)

75. C.K. Hancock and C.P. Falls, J. Am. Chem. Soc., 83, 4214 (1961)

76. M. Charton, J. Am. Chem. Soc., 97, 1552, (1975); 91, 615 (1969)

77. J.A. MacPhee, A. Panaye and J.E. Dubois, Tetrahedron, 34, 3553 (1978)

78. J.A. MacPhee, A. Panaye and J.E. Dubois, J. Org. Chem. 45, 1164 (1980)

79. D.F. DeTar, J. Org. Chem., 45, 5166 (1980)

80. M. Charton, J. Am. Chem. Soc., 97, 3694, 3691 (1975)

81. M. Charton and B. Charton, J. Am. Chem. Soc., 97, 6472 (1975)

82. M. Charton, J. Org. Chem., 41, 2906 (1976)

83. M. Charton, Topics in Current Chemistry, vol.114, page 57, Springer Verlag, Berlin-Heidelberg, 1986

84. M. Charton, J. Org. Chem., 41, 2217 (1976)

85. M. Charton, J. Chem. Soc. Perkin Trans II, 97, 1983

86. R.D. Topsom, in "Molecular Structure and Energetics" Liebman and Greenberg Ed., vol.1, Chemie Verlag 1985

87. R.D. Topsom, Acc. Chem. Res., 16, 292 (1983)

88. W.J. Hehre, Acc. Chem. Res., 9, 399 (1976)

89. W.J. Hehre, R.F. Steward and J.A. Pople, J. Chem. Phys., 51, 2657 (1969)

90. R.D. Topsom, J. Am. Chem. Soc., 103, 39 (1981)

91. S. Marriott and R.D. Topsom, Tetrahedron Lett., 1485, 1982

92. S. Marriott, W.F. Reynolds, R.W. Taft and R.D. Topsom, J. Org. Chem., 49, 959 (1984)

93. A.L. Allred and E.G. Rochow, J. Inorg. Nucl. Chem., 5, 264 (1958)

94. R.J. Boyd and G.E. Marleus, J. Chem. Phys., 75, 3385 (1981)

95. A.R. Katritzky and R.D. Topsom, Chem. Rev., 77, 639 (1977)

96. H.H. Jaffé, J. Chem. Phys., 20, 279, 778 and 1554 (1952)

97. T.M. Krygowski and G. Häfelinger, J. Chem. Res., 348, 1986

98. R.W. Taft, J. Phys. Chem., 64, 1805 (1960)

99. O. Exner, Coll. Czech. Commun., 31, 65 (1966)

100. R.W. Taft, Progr. Phys. Org. Chem., 14, 247 (1983)

101. M. Charton, Progr. Phys. Org. Chem., 16 (*in press*)

102. A. Domenicano, A. Vaciago and C.A. Coulson, Acta Cryst. B39, 452 (1975)

103. A. Domenicano, A. Vaciago and C.A. Coulson, Acta Cryst., B39, 221 (1975)

104. A.D. Walsh, Discuss. Farad. Soc., 2, 18 (1947)

105. H.A. Bent, Chem, Rev., 61, 275 (1961)

106. R.J. Gillespie and R.S. Nyholm, Quart. Rev., 11, 339 (1957)

107. A. Domenicano, A. Vaciago, P. Marrec, Tetr. Letters, 1029, 1976

108. T.M. Krygowski, J. Chem. Res., 238, 1984

109. R. Norrestan and L. Schepper, Acta Chem. Scand., A35, 91 (1981)

110. A. Domenicano and P. Murray-Rust, Tert. Letters, 2283, 1979

111. T. Więckowski and T.M. Krygowski, Croat. Chim. Acta, 58, 5 (1985)

112. S.J. Grabowski and T.M. Krygowski, Acta Cryst., C41, 1224 (1985)

113. J. Maurin and T.M. Krygowski, Acta Cryst., C41, 64 (1987)

114. A. Roszak, PhD Thesis, Univ. of Poznań, 1986

115. T.M. Krygowski, J. Chem. Res. (S) 120, 1987

116. T.M. Krygowski and I. Turowska-Tyrk, Chem. Phys. Letters 138, 90 (1987)

117. T.M. Krygowski and I. Turowska-Tyrk, *submitted*

APPENDIX

STATISTICAL COMMENTS ON USING CORRELATION AND REGRESSION ANALYSIS

by K. Woźniak[a], T. M. Krygowski[a] and R. I. Zalewski[b]

a) Department of Chemistry, University of Warszawa, 02-093 Warszawa, Pasteura 1, Poland.

b) Department of General Chemistry, Academy of Economics, Poznań, Poland.

It is clear from the preceding review that estimation of goodness of fit applied in CAOC* is most often arbitrary. Moreover, statistics as a tool is not commonly known to chemists involved in CAOC. The aim of this appendix is to apply the statistical point of view to the problems encountered in CAOC. In CAOC papers a technique most often used to optimize the model to the data is the least-squares fitting, and for estimating the quality of fit the following two parameters are used:(a) correlation coefficient R, and (b) estimated standard deviation from the model. This appendix offers a somewhat broader look at the problem of (i) goodness of fit, (ii) significance of correlation between variables in question, (iii) range estimation of correlation coefficient.

QUALITY OF FIT

Let us build up a mathematical model for our experimental data. Let this model be a function of variable X and parameters β, i.e. $Y = f(X,\beta)$. Let X and β be - depending

---

*Correlation Analysis in Organic Chemistry.

on the need of interpretation and context of a research - either scalars or vectors. For n experimental data y_1, y_2,...,y_n we have n explanatory parameters x_1,x_2,...,x_n, i.e. n pairs (x_i,y_i), i=1,2,...,n. Our purpose is to find the best estimates **b** of parameters β of the model in question. It is commonly accepted that the model is the better the lower are the differences $e_i=y_i-f(x_i,b)$. They are called residuals and estimate an error of the model in the sample. There are many possible quantities Q[1] estimating joint error of the model, but the most convenient and most often applied are (1) and (2):

$$Q = \sum_i |y_i-f(x_i,b)|^2 \qquad (1)$$

$$Q = \sum_i w_i |y_i-f(x_i,b)|^2, \qquad (2)$$

where w_i are weights.

Mathematically this is convenient due to the continuity of the derivatives, hence minimization of Q in this case is much facilitated and leads to the least-squares method. In the case of normal distribution of errors ε_i and a linear model $Y=\beta_o+\beta_1 X$, this minimization is equivalent to the most reliable optimization. However, even in the least-squares method, there exists a problem of choice of the function of error. Deviation of the point from the line may be measured in three ways: perpendicularly to X- or Y- axes or perpendicularly to the line. The most often used procedure is the first case, since x-values are asummed to be not biased by any errors and only the error in y is taken into account. If the errors of y_i are known one can use the weighted regression.

MEASURES OF DEPENDENCE

For any two samples of elements characterized by two parameters y and x, for n points (x_i,y_i) (i=1,2,...,n), it is possible and convenient to define a measure of mutual dependence $\delta(X,Y)$. Renyi [1,2] has given conditions to be fulfilled by this kind of measure and the most often used parameter describing this mutual dependence is the correlation coefficient. It has got, however, its pros and cons. One of its shortcomings is that for extremal values, i.e. close to 0 or ±1 its value is not too sensitive for goodness of fit. The correlation coeficient R for a sample (x_i,y_i), i=1,2,...,n is an estimate of the true correlation coeficient ρ of the total population (which in principle is infinite). By definition:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var}(Y)}. \qquad (3)$$

The point estimate for $\rho(X,Y)$ for a given sample (x_i,y_i), i=1,2,...,n is defined as follows:

$$R = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{1/2}}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \qquad (4)$$

It is worth recalling that R^2x100 is a percent of variability of one variable (say Y) explained by the other one (X) and it is often called a determination coefficient. In other words, R^2x100, described % of the var(Y), is explained by the model based on explanatory parameter X. It should be emphasized that correlation coefficient R of the sample (X,Y) is only an estimate of $\rho(X,Y)$ for a total population which is, of course, unknown. Since R(X,Y) is a function of the size of

the sample (and of the sampling process as well) the estimated correlation coefficient R may be considered as a random variable. Its distribution is given by a rather complicated function and hence this distribution is not often applied in practice. Fortunately, for large samples the distribution of R is approximately close to normal $N[\rho,(1-\rho^2)/(n)^{1/2}]$. However, for this approximation, particularly for $|\rho| \gtrsim 1$, very large samples (large n) are required. Much more common in use is an approximation given by Fisher [3,4] :

$$Z = (1/2)\ln\frac{1+R}{1-R} = \operatorname{arctanh} R. \qquad (5)$$

In this case the random variable Z is approximately normally distributed, and the formulas presented below work:

$$EZ = (1/2)\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n+3)}\left(1 - \frac{3-\rho^2}{4(n-3)} + \ldots\right), \qquad (6)$$

$$D^2Z = \frac{1}{n-3}\left(1 - \frac{\rho^2}{2(n-3)} - \frac{2-6\rho^2+3\rho}{6(n-3)^2} + \ldots\right) \qquad (7)$$

where E stands for expectation value of the random variable Z and D^2 stands for the variance of Z. It results from the above formulas that random variable $\frac{Z-E}{DZ}$ is approximately normally distributed N(0,1); this approximation is satisfactory even for $n \geq 20$. Two other measures of dependence (relatively popular, but not used in CAOC papers) are Kendall's coefficient τ and Spearman's coefficient r_s. Both are superior to the correlation coefficient since they work correctly independently of the distribution of X and Y. Hence they are used as a ground for numerous non-parametric tests. Applying these coefficients instead of correlation coefficient ρ is particularly advantageous in those cases when we do not have any information about the distribution of

the random variables, or when due to paucity of data we can not verify the hypothesis about their normality. If we cannot verify this hypothesis, we have to introduce into our model additional assumptions which obviously weakens our model. If observables x_i and y_i in the formula for R (eq.4) are replaced by their ranks we get Spearman's rank coefficient r_s [4].

RANGE ESTIMATION OF $\rho(X,Y)$

The point estimation R of correlation coefficient does not give any information about the precision of this estimate. This is only possible by use of range estimation of ρ. Any two quantities ϕ_1 and ϕ_2 such that

$$P(\phi_1 < \rho < \phi_2) = 1-\alpha \quad (8)$$

determine a confidence interval (ϕ_1,ϕ_2) for a correlation coefficient ρ at the level of confidence $1-\alpha$. In order to estimate this kind of confidence level for ρ we may use one of two ways:

(i) applying graphic nomograms available in many statistical tables,

(ii) direct application of the above-mentioned approximations of ρ-distribution.

SCHEME FOR PRECEDING

Let us have a known correlation coefficient R. We transform it into Z_R using Fisher's formula. In the first approximation DZ is only a function of the size of the sample n. One can easily estimate confidence interval for Z_R:

$$Z_R - k(\alpha)DZ < Z_R < Z_R + k(\alpha)DZ, \quad (9)$$

where $k(\alpha)$ is a coefficient dependent on significance level α. Then the upper and lower limits of the confidence interval

for Z_R can be transformed into upper and lower limits for confidence interval for R applying the reverse function defining Z_R. We use then:

$$R=\frac{e^{z^2}-1}{e^{z^2}+1} \qquad (10)$$

SIGNIFICANCE OF THE CORRELATION

If the n-element sample is taken from a general population with two-dimensional normal distribution with $\rho=0$,i.e. if X and Y are uncorrelated i.e. independent, thena) a random variable

$$t = \frac{|R|(n-2)^{1/2}}{(1-R^2)^{1/2}} \qquad (11)$$

is characterized by Student's distribution t with n-2 degrees of freedom [1,4],

(b) a random variable R^2 has a standard β-distribution with parameters p=1/2, q=(1/2)(n-2),

(c) statistics $F=\frac{R^2}{1-R^2}(n-2)$ has a Snedecor distribution with 1 and n-2 degrees of freedom.

In general, if in m-dimensional normal distribution the multiple correlation coefficient ρ is zero then the estimated determination coefficient R^2, from a sample built up of n elements (n > m), has a standard β distribution with parameters p=(1/2)(m-1) and q=(1/2)(n-m) and a statistics

$$F=\frac{R^2}{1-R^2}\cdot\frac{n-m}{m-1} \qquad (12)$$

has Snedecor distribution with m-1 and n-m degrees of freedom.This conclusion permits to formulate a null hypothesis:

$$H: \rho=0 \qquad (13)$$

contrary to an alternative hypothesis:

$$K: \rho\neq 0 \ (\text{or } K^{+}: \rho > 0, \text{ or } K^{-}: \rho < 0). \qquad (14)$$

Then in order to determine the significance of correlation in the general population we calculate one of the above-mentioned statistics from the sample (i.e. t, R^2 or F), accept a given level of significance α and find in statistical tables the values of $t_{\alpha,n-2}$, $R^2_{\alpha,p,q}$ or $F_{\alpha,m-1,n-m}$. Having compared the values of statistics obtained from the sample with table-values we may either reject the null hypothesis H or we state that we have no arguments to reject it. If the value of statistics from the sample is greater than that from the table then we reject H and state that correlation is significant at level α. In this sense investigation of the significance of correlation is equivalent to that of the significance of a regression. Nonparametric equivalents of tests described above are Spearman's R_s and Kendall's τ - tests[4].

DISTRIBUTION OF RESIDUALS AS A MEASURE OF ADEQUACY FOR A LINEAR MODEL

An important problem in CAOC is how far a description of the data by the model is adequate. The answer to this question may be given by analysis of residuals. Let E_i be random variables whose realizations are errors of the model ε_i. Assuming that[1]:

i) E_i are random variables with expectation value $E(E_i)=0$ and variances σ^2,

ii) E_i and E_j are uncorrelated, i.e. $cov(E_i,E_j)=0$ for $i\neq j$,

iii) E_i are normally distributed $N(0,\sigma)$ for $i=1,2,\ldots,n$.

we may get information about distributions of estimates b_o and b_1 of the model parameters β_o and β_1. With those assumptions in mind b_o and b_1, obtained by minimization of e_i^2 (e_i is an estimate of an error of the model ε_i), are random variables normally distributed. This results from the consideration that E_i and E_j are independent random variables. After computing the best estimates b_o and b_1 one may ask: are the assumptions (i÷iii) really fulfilled? To verify the shape of distribution it is convenient to use appropriate tests, e.g.: χ^2 or the Kołmogorov one and ,in the case of normal distribution, the Shapiro-Wilk[5] test.However, to verify the assumption about the lack of correlation between E_i and E_j it may be concluded that from the assumption $cov(E_i,E_j)=0$ it results that both $cov(Y,E)=0$ and $cov(X,E)=0$. This means that residuals E should not be correlated with any of the variables X or Y. It may help to decide whether or not the model applied is a proper one, and if not, whether it should be extended. In many cases it is sufficient to make a visual estimation of the dependence e vs. X or e vs. Y. If any regularity is observed in these plots it means that the linear model is too poor and should be extended (into planar or another one).

CORRELATIONS BETWEEN EXPLANATORY VARIABLES IN THE CASE OF MULTIPLE REGRESSION

In many cases linear regression is too weak to explain the total variance of variable Y: then the multiple regression with a few sets of explanatory variables is applied. It is important to comment now that from the statistical point of view there is no distinction between

explanatory variables X_i, i=1,2,....n, and the variable to be explained, Y. Statistics does not investigate any reason-result relations. This distinction is important for cases of similarity models (CAOC-people are of this kind); the less recognized experimental data are to be described by much more known ones. On the basis of strong dependence estimated statistically a physico-chemical conclusion may be drawn. Taking this into account to study residuals we formulate the multiple regression as below:

$$X_1 = \beta_{1,2}X_2 + \ldots + \beta_{1,n}X_n, \qquad (15)$$

where $\beta_{1,k}$ is the regression coefficient for variable X_k. Thus:

$$\varepsilon_1 = X_1 - \beta_{1,2}X_2 - \beta_{1,3}X_3 - \ldots - \beta_{1,n}X_n \qquad (16)$$

is an error of the model for variable X_1. Estimates of $\beta_{i,k}$ are found by the least-squares method. Similarly, as it was mentioned for the linear model, it is possible to show that with the same general assumptions ε_i should not be correlated with any of the variables $X_1, X_2, \ldots, X_n$. A question which often appears in CAOC reasearch is: how far are the "explanatory" variables intercorrelated? To answer it, let us choose any two variables, say X_1 and X_2. If we do not take into account the existence of other variables involved in the multiple regression the measure of correlation between them is a well known correlation coefficient $\rho(X_1,X_2)$. If, however, the variability of X_1 and X_2 is taken into account as certain contributions to the variability of all variables involved in the multiple regression then the measures of these contributions are given by the errors ε_1 end ε_2[1].We have:

$$\varepsilon_1 = X_1 - \gamma_{1,3}X_3 - \gamma_{1,4}X_4 - \ldots - \gamma_{1,n}X_n \qquad (17)$$

$$\varepsilon_2 = X_2 - \gamma_{2,3}X_3 - \gamma_{2,4}X_4 - \ldots - \gamma_{2,n}X_n \qquad (18)$$

where γ are estimates of β. The correlation coefficient between these errors may be taken as a measure of correlation between X_1 and X_2 after elimination of a certain variability due to $X_1, X_2, \ldots, X_n$. This quantity is known as a partial correlation coefficient and is defined as follows:

$$\rho_{1,2} = \frac{E(\varepsilon_1, \varepsilon_2)}{(E(\varepsilon_1^2)E(\varepsilon_2^2))^{1/2}} \qquad (19)$$

However, an error ε_1 may also be written as $\varepsilon_1 = X_1 - \hat{X}_1$, where $\hat{X}_1$ is the best regressional estimate of X_1. The correlation coefficient $\rho_{1(2,\ldots n)}$ between X_1 and $\hat{X}_1$ is called a multiple correlation coefficient and may be used as a measure of linearity of the model used:

$$\rho_{1(2,\ldots n)} = \frac{E(X_1, \hat{X}_1)}{(E(X_1^2)E(\hat{X}_1^2))^{1/2}}. \qquad (20)$$

Evidently, the estimates from the sample of these correlation coefficients may be expressed in a simple way via the estimates of the correlation coefficients from the sample, which are expressed by use of x_i and y_i.

CONCLUSIONS

1) In order to estimate the significance of a given regression it is better to use the tests mentioned above than to use arbitrary scales of R (suggested by e.g. Jaffé[6]) or any other arbitrary procedure.

2) In order to estimate the utility (i.e. predictive power) of the regrresion it is safe to use R^2 - it gives % of the total variance explained by the model -with its range

estimate.

3) In the case of multiparameter regressions used in the model of similarity it is necessary to search indepedency of the explanatory parameter sets.

4) In general it is adventagous to verify the shape of the distribution of the variables taken into consideration: if the distribution is known then we may use parametric tests, and if not - then non-parametric tests should be used to verify the significance of the correlation.

REFERENCES

1. J. B. Czermiński, A. Iwasiewicz, Z. Paszek and A. Sikorski, Statistical Methods in Applied Chemistry,Elsevier, Amsterdam, in print.

2. A. Renyi,On measures of dependence,Acta Math. Acad. Sci. Hung.,10,(1959).

3. R. Zieliński,Tablice statystyczne,PWN,Warszawa(1972).

4. M. R. Spiegel,Statistics,Mc Grew-Hill , New York(1961).

5. M. B. Shapiro, H. J. Chen, A comparative study of various tests for normality,J. Amer. Statistical Association,63(1968).

6. H. H. Jaffé, Chem. Revs., 53, 193(1953)