Dyes and Pigments 6 (1985) 189–212

Quantitative Structure–Property Relationships Study of Azo Dyes using Partial Least Squares Analysis in Latent Variables (PLS)

Rosarina Carpignano, Piero Savarino and Ermanno Barni

Istituto di Chimica Organica Industriale, Università di Torino,
C.so M. D'Azeglio 48, 10125 Torino, Italy

and

Sergio Clementi and Gianfranco Giulietti

Dipartimento di Chimica, Università di Perugia,
Via Elce di Sotto 10, 06100 Perugia, Italy

(Received: 26 July, 1984)

SUMMARY

The light, washing, alkali and perspiration fastness of a series of azo dyes are analysed as a function of physicochemical substituent descriptors by the PLS method.

The analysis permits relation of the fastness properties to the structural features and the development of a model which can be used to predict the fastness of new dyes of the same series.

INTRODUCTION

A major aim of dye research is to relate structural changes to technological properties. From this point of view such research requires the same type of approach and the same statistical tools as those used in quantitative structure–activity relationships (QSAR). Preliminary work in this area has recently been published by some of us.[1,2]

Dyes and Pigments 0143-7208/85/$03·30 © Elsevier Applied Science Publishers Ltd, England, 1985. Printed in Great Britain

Quantitative structure–activity relationships are quantitative models which relate the variation of a property (activity) in a series of chemical compounds to the variation in the chemical structure within the compounds of the series. The models used in QSAR range from simple additivity models to complicated models of pattern recognition relating multivariate activity data to multivariate chemical structure data. The primary objective of QSAR is to predict the activity for new compounds and to provide some understanding of the cause of the activity under study. Structure–property studies require therefore appropriate multivariate statistical methods to handle experimental data and structural parameters.[3]

MULTIPLE REGRESSION ANALYSIS (MRA)

The most widely used method of multivariate statistical analysis is that of multiple regression.[4] This model relates the measured values of activity y, for a set of N compounds, to M structure indicator variables x_i, by the linear relationship in eqn (1).

$$y_k = b_0 + \sum_i b_i x_{ik} + e_k \qquad (1)$$

where b_0 and b_i are the regression coefficients and e the residual, for the kth compound of the set. In the simplest case with one x variable, the regression model is a straight line.

The most used regression approaches in QSAR are the Hansch and the Free–Wilson models.

The Hansch multiparametric method[5] correlates the activities (y) of a set of similar compounds, modified by changing substituents, with a set of substituent descriptors (x_i) like electronic (σ), hydrophobic (π), steric and polarizability (E_S, MR) parameters.

The Free–Wilson method[6] is an additive model independent of the physicochemical properties of the substituents. J substituents at G different sites in a parent structure are varied. The activity for the kth compound of the series is formulated as the sum of the contributions of the parent structure b_0 and of the substituents present in the various positions, according to an equation of the form (1), where b_i is the additive contribution of the ith combination of site and substituent and e the residuals. Each variable x_i is specific for one site and one substituent, having the value 1 when the substituent is present at that site, 0 when it is absent.

In the form modified by Fujita and Ban[7] the activity contribution of each substituent is relative to hydrogen whose contribution is defined as equal to zero.

The main limitation of the Free–Wilson model is that the activity predictions cannot be made for a compound containing a combination of substituent and position which was not included as an observation in the original analysis.

The conditions required for applicability of multivariate QSAR have recently been reviewed.[8]

It is noteworthy to observe that all the statistical models are based on the fact that complicated functional relationships between chemical structure and activity can be locally linearized, i.e. approximated by simple mathematical models in limited intervals. Therefore a QSAR study ranging over a wide variety of structures and/or activities must necessarily involve multiple models. In such cases classification methods (pattern recognition) may be needed before a quantitative relationship between chemical structure and activity can be achieved.

COMPARISON OF MULTIVARIATE STATISTICAL METHODS

Multiple regression analysis is aimed at describing a dependent variable y in terms of explanatory variables x. However, the regression methods rely on assumptions that often render them inappropriate for application in chemistry. In fact MRA requires that all the explanatory variables used are independent of each other, exactly known and 100% relevant to the problem.[8,9] Quite often in this area the problem actually consists of finding which variables are relevant to determine the property of compounds. In such cases the application of MRA is to be used with caution. Moreover with MRA one is limited by the number of variables which can be used; this number should be as small as possible to have a high ratio of objects to variables and to diminish the risk of multicollinearity.

When the problem is to detect the relative importance of individual variables in determining the data structure, i.e. the extraction of the chemical information contained in a data set, the most appropriate statistical approach is principal component analysis (PCA),[3,10,11] which seeks for systematic variation in the data table. In this approach no assumption about the relevance of the variables x is required, since their relevance is obtained from the statistical analysis. Furthermore PCA indeed uses collinearities to estimate the statistical parameters.

However, PCA, which has the same mathematical form as MRA, is not aimed at finding out cause–effect relationships, and therefore is not suitable to handle problems of relating properties to chemical structure. This latter problem can be solved by the recently developed method called PLS (partial least squares analysis in latent variables), which is aimed at detecting cause–effect relationships, but does not require any preliminary assumption on the relevance of individual variables since it relies on PCA.[9]

In the latent variables analysis the objective is to find out possible relationships between one or more dependent variables y (dye fastness) and a number of structural descriptors (X-block).

In general this problem is handled by computing PC models for each of the two matrices followed by establishment of any linear relationship between the principal components of these two blocks. Instead of this two-step procedure, it is possible to perform a single analysis accomplishing the two steps simultaneously by using an appropriate algorithm called PLS.[9] This method has already proved to be extremely useful in various branches of chemistry.

PCA and PLS are the statistical approaches used in the SIMCA (soft independent modelling of class analogy)–MACUP (modelling and classification using PLS) method. This article reports a specific application of the SIMCA–MACUP method to relate a number of technological properties of a set of azo dyes to the structure of the dyes.

SIMCA–MACUP METHOD

Assume that we have a data table (matrix X) consisting of M variables observed in N objects. The basis of this method is that the data table can be simplified to describe the main variation by a few components or underlying tendencies of variation. The components constitute a space of lower dimensions than the M-dimensional space spanned by the original

M variables in the raw data. This kind of simplification is obtained by a principal components (PC) model in eqn (2):

$$x_{ik} = \bar{x}_i + \sum_{a=1}^{A} b_{ia} t_{ak} + e_{ik} \tag{2}$$

where x_{ik} is the value of the variable i observed on object k. The parameters of the model, $\bar{x}_i$, b_{ia} and t_{ak} are estimated by least squares as to make the residuals e_{ik} minimal over all objects k and all variables i.

SIMCA has a theoretical foundation similar to that of polynomial models, i.e. on the basis of the Taylor expansion. Thus, provided that the variables characterizing the objects of a class have certain continuity properties and the objects are realizations of a process with limited variability, the data of the objects can be closely approximated by a PC model with a limited number of product terms A.

Geometrically this corresponds to finding an A-dimensional hyperplane in the M-dimensional space which has the closest fit to the data points. In the simplest case, with $A = 1$, the PC model corresponds to a straight line in the measurement space. The value of the parameter $\bar{x}_i$ defines the central point and those of b_{ia} (*loadings*) the direction coefficients of the plane. The values of the parameters t_{ak} (*scores*) define the position on this plane of the kth object.

The data analysis in SIMCA is performed as follows:

(a) the number of the components A required to describe the data of the objects is determined by cross-validation,[12] an objective criterion to evaluate the significance of each added component on the basis of the predictive ability of the model;

(b) the parameters $\bar{x}_i$, b_{ia} and t_{ak} for $a = 1, 2, \ldots, A$ are determined;

(c) the resulting residuals e_{ik} are used to assess the importance of each variable i.

The ability of the variable to describe the data structure is called *modelling power* (ψ_i) and is obtained by comparing its residual standard deviation s_i for A dimensions with the corresponding standard deviation of the data for $A = 0$, s_{iy}:

$$\psi_i = 1 - \frac{s_i}{s_{iy}}$$

A value for ψ_i approaching 1 indicates high modelling power while a value near zero indicates low modelling power.

The PLS algorithm is described in detail in ref. 9 and only a brief summary is reported here. In the case when there is just one dependent variable, the method extracts the principal components of the independent variables (descriptors block) (eqn (2)) under the constraint to maximize the relationship with the dependent variable values (eqn (3)):

$$y_k = \bar{y} + \sum_{a=1}^{A} d_a t_{ak} + f_k \tag{3}$$

where d_a represents the proportionality coefficient of the inner relationship and f_k is the residual.

When many dependent variables are available, the analysis results in a description of the X-matrix by one PC-like model (eqn (2)), a description of the Y-matrix by another PC-like model (eqn (4)) and predictive relationships between the latent variables t and u (eqn (5)):

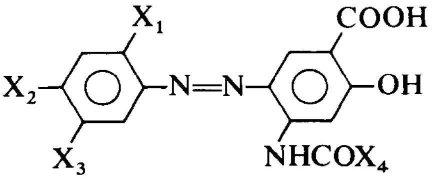
$$y_{ik} = \bar{y}_i + \sum_{a=1}^{A} c_{ja} u_{ak} + f_{jk} \tag{4}$$

$$u_{ak} = d_a t_{ak} + h_{ak} \tag{5}$$

We observe that eqn (4) is identical to eqn (2) where c and u are the loadings and the scores respectively of the Y-block as b and t are for the X-block. In eqn (5) h_{ak} is the residual. The algorithm used in the SIMCA–PLS package is iterative for each dimension as in PCA. It consists of finding the latent variables of the X-matrix, $t(k)$, from start values of $u(k)$ and from the X elements, and then recomputing the latent variables of the Y-matrix, $u(k)$, from the Y elements and the $t(k)$ values, until the process converges.

PRESENT STUDY

The set of dyes under examination consists of 65 members (objects) of general formula I, previously described.[13]



(I)

TABLE 1
Technical Properties of Dyes of General Formula I

Dye		*Fastness values*				
No.	*Symbol*[a]	*Light*	*Washing*		*Alkali*	*Perspiration*
			60°C	*95°C*		
1	HHHT	2·5	3·0	1·0	4·5	3·5
2	HHHP	2·5	2·5	1·0	4·5	3·5
3	HHHE	2·5	3·0	1·5	4·5	3·5
4	HHHU	2·5	4·5	2·5	5·0	3·5
5	HHHQ	2·5	4·0	4·0	5·0	4·5
6	HTHT	2·5	4·0	1·5	5·0	4·0
7	HTHP	2·5	4·0	1·5	4·5	3·5
8	HTHE	2·5	4·5	2·5	4·5	3·5
9	HTHU	2·5	4·5	3·0	5·0	4·0
10	HTHQ	2·5	4·5	4·5	5·0	4·5
11	HKHT	3·5	4·5	1·5	3·5	4·0
12	HKHP	4·0	4·0	1·0	2·5	4·0
13	HKHE	3·5	4·5	1·5	2·5	4·5
14	HKHU	3·5	4·5	3·5	3·5	4·5
15	HKHQ	3·5	5·0	4·5	4·5	5·0
16	HOHT	2·5	4·5	1·0	4·5	3·0
17	HOHP	2·5	4·0	1·5	4·5	3·5
18	HOHE	2·5	4·0	1·5	4·5	4·0
19	HOHU	2·5	4·5	3·0	4·5	4·0
20	HOHQ	2·5	4·5	3·5	5·0	4·5
21	HNHT	4·5	3·5	1·0	2·0	4·0
22	HNHP	4·5	3·5	1·0	1·0	4·0
23	HNHE	4·5	4·0	1·5	3·0	4·0
24	HNHU	3·5	4·5	2·5	2·5	4·5
25	HNHQ	3·5	4·0	3·0	3·5	4·5
26	HXHT	2·5	4·0	1·0	4·0	3·5
27	HXHP	2·5	4·0	1·5	4·5	3·5
28	HXHE	2·5	4·5	2·0	4·5	4·0
29	HXHU	2·5	4·5	3·0	4·5	4·5
30	HXHQ	2·5	5·0	4·5	4·5	4·5
31	HAHT	1·5	5·0	1·5	4·5	3·0
32	HAHP	1·5	4·5	2·0	4·5	2·0
33	HAHE	1·5	4·5	2·0	4·0	2·5
34	HAHU	1·5	4·5	2·5	4·5	3·0
35	HAHQ	2·0	5·0	2·5	5·0	3·5
36	HCHT	4·5	4·0	1·0	3·5	3·5
37	HCHP	4·5	4·0	1·0	4·0	3·5

(continued)

TABLE 1—*contd.*

Dye		*Fastness values*				
No.	*Symbol*[a]	*Light*	*Washing*		*Alkali*	*Perspiration*
			60°C	*95°C*		
38	HCHE	4·5	4·5	1·0	4·0	4·5
39	HCHU	4·5	4·5	1·5	4·5	4·0
40	HCHQ	3·5	4·5	2·0	5·0	4·5
41	HBHT	4·0	4·0	2·0	4·5	2·5
42	HBHP	4·5	4·0	2·5	4·5	2·5
43	HBHE	3·0	4·5	3·5	2·5	3·5
44	HBHU	3·0	4·5	3·0	3·5	3·0
45	HBHQ	3·5	5·0	3·5	4·0	3·5
46	ONHT	3·5	4·5	2·0	2·5	3·5
47	ONHP	5·0	4·0	1·5	3·0	2·5
48	ONHE	4·5	4·5	3·0	2·5	2·0
49	ONHU	4·5	4·0	2·5	2·5	2·5
50	ONHQ	4·0	5·0	2·0	2·5	3·0
51	TNHT	4·5	4·0	1·5	4·5	3·0
52	TNHP	4·5	4·0	1·5	4·5	3·0
53	TNHE	4·5	4·5	1·5	3·0	4·0
54	TNHU	4·5	4·5	2·0	4·5	3·5
55	TNHQ	4·0	4·5	2·0	4·5	3·0
56	XNHT	4·5	3·5	1·5	1·5	1·5
57	XNHP	4·5	3·5	1·5	2·5	1·5
58	XNHE	4·5	3·5	2·5	2·0	2·5
59	XNHU	4·5	4·0	3·0	2·5	2·0
60	XNHQ	3·5	4·5	2·0	2·5	2·5
61	OHNT	1·5	5·0	1·5	4·0	4·0
62	OHNP	3·0	4·5	1·5	4·0	3·5
63	OHNE	1·5	5·0	2·0	4·0	4·0
64	OHNU	3·0	5·0	3·0	4·0	4·5
65	OHNQ	2·5	5·0	2·0	4·5	3·5

[a] The symbol is formed with a letter for each of the four positions: X_1, X_2, X_3, X_4 in general formula I. Each letter represents a substituent, as listed: H = H; O = OCH_3; T = CH_3; X = Cl; K = $COCH_3$; N = NO_2; A = $NHCOCH_3$; C = CN, B = 6-methyl-2-benzothiazolyl; P = C_3H_7; E = C_7H_{15}; U = $C_{11}H_{23}$; Q = $C_{15}H_{31}$.

The measured properties y ('dependent' or 'effect' variables) are lightfastness, fastness to washing at 60 °C and 95 °C, fastness to alkali and fastness to perspiration.

The objects and their measured properties are listed in Table 1.

DESCRIBING THE STRUCTURE BY A SET OF VARIABLES

A critical factor in structure–activity studies is the appropriate description of the structure of the objects. Various structural descriptors and measured variables have been used and discussed.[14,15]

In a previous reported structure–activity study Dunn and Wold[16] used as descriptor variables physicochemical parameters such as Hammett σ constants, and hydrophobic and steric constants. This approach may be considered an extension of the extrathermodynamic assumptions as used by Hansch and his coworkers. We took an analogous approach; the structure of each object of the set was described by a number of physicochemical parameters for each substituent, as given in Table 2. The

TABLE 2
Parameters for Substituents X_1, X_2, X_3 and X_4

Substituent	σ_p	π	*MR*	v
H	0·00	0·00	1·03	0·00
OCH_3	−0·27[a]	−0·02	7·87	0·36
CH_3	−0·17	0·56	5·65	0·52
Cl	0·23	0·71	6·03	0·55
$COCH_3$	0·50	−0·55	11·18	0·50
NO_2	0·78	−0·28	7·36	1·39
$NHCOCH_3$	0·00	−0·97	14·93	—
CN	0·66	−0·57	6·33	0·40
B[b]	0·29	2·13	38·88	—
$NHCOC_3H_7$[c]	—	0·11	24·23	—
$NHCOC_7H_{15}$[c]	—	2·27	42·83	—
$NHCOC_{11}H_{23}$[c]	—	4·43	61·43	—
$NHCOC_{15}H_{31}$[c]	—	6·59	80·03	—

[a] $\sigma_p^+ = -0{\cdot}78$ is used for OCH_3 at X_1 when a nitro group is present at X_3, to account for through conjugation.

[b] B = 6-methyl-2-benzothiazolyl. The values used are those for 2-benzothiazolyl.

[c] Calculated values.

choice of parameters to be used as descriptors was made by a preliminary study limited to the available substituents on a number of variables describing electronic effects (σ_m, σ_p, σ_I), steric effects (E_S, Charton's v), and hydrophobic and polarizability effects (π, MR), with the aim of excluding those having practically the same information content. As a result we took into account only σ_p, π, MR and v, since they were shown to be the minimal set containing significantly different information. These four variables were used to describe substituents at X_1 and X_2, where a number of different substituents were available.

Since the substituents at X_3 are only H or NO_2 we could not use any figure to indicate the structural change; the presence of two entries only would give enormous importance to this variation. Accordingly we decided to describe the structural feature by modifying the electronic descriptor of the X_1 substituent (methoxy) from σ_p to σ_p^+ because of the presence of through conjugation in such structures. For describing the acylamido groups, where the main feature is the increase in size of the alkyl chain, we used estimates of the π and MR variables, calculated on the basis of systematic increase on increasing the chain length.

PLS ANALYSIS

The MACUP method was applied to find systematic relationships between the fastness properties (65 × 5 Y-matrix) and the block of descriptor variables (65 × 10 X-matrix) by means of equations (2), (4) and (5). The data were first scaled to give them equivalent variance for each x and y variable. This procedure gives equivalent importance to the variables with small variation and to those with large variation, and then prevents masking of the variables with little variation by those with large variation.

The weights w, which are the inverse of the variable standard deviations, are listed in Table 4.

The PLS analysis gives a two-dimensional ($A = 2$) model which accounts for 41 % of the variance of Y (Table 3).

The model on the whole set is aimed just at detecting inhomogeneities in the data and outliers, since the model may be strongly affected by their presence. The results (not shown) indicated that two groups of dyes (objects 41–45: X_2 = methylbenzothiazolyl and objects 61–65: X_1 = methoxy, X_3 = nitro) are strong outliers for the X-block, presumably

TABLE 3

Residual Standard Deviation and Percentage of the Variance Explained after Model Expansion for the Whole Set and the Final Set Respectively

Dimension	*Whole set model*		*Final model*	
	S^a	V (%)^b	S	V (%)
$A = 1$	0·845	29	0·814	~34
$A = 2$	0·766	41	0·740	45
$A = 3$			0·672	55
$A = 4$			0·636	60

[a] Residual standard deviation for the model.
[b] Percentage of the variance explained.

because the descriptors used are not adequate. Furthermore, objects 1–3 are outliers of the Y-block owing to largely low values for the fastness to washing. Accordingly, these 13 dyes were omitted from the final analysis in order to obtain a better model. The PLS computations were then repeated on the remaining set of 52 objects.

The fraction of variance explained by this model is reported in Table 3. The resulting variable parameters (weights, averages, loadings and modelling powers) are listed in Table 4, while the object parameters (scores) are listed in Table 5. The data structure is described by a four-component model. The first dimension explains as much as 34 % of the total variance, the second a further 11 %, the third a further 10 % and the fourth 5 % more up to a global 60 % (Table 3).

The first dimension describes the main variation in fastness to light, alkali and perspiration (high c_1 values in Table 4) and appears to be due to the contributions of variables 2, 3, 4, 5 and 8 (high b_1 values in Table 4), which indicate the relevance of ring substitution. The second dimension takes into account fastness to washing in the property block (c_2) and the chain parameters for the acylamido group in the descriptors block (b_2 of variables 9 and 10). The third dimension explains properties I, II and V in terms of the variables 5, 6 and 7 referring to substitution at the X_2 position, and finally the fourth dimension explains a different combination of all five y variables (see the c_4 values in Table 4) as a function of variables 1, 2, 5, 6 and 7. Accordingly, all descriptors used, except for variable 1, are highly relevant to defining the mathematical model.

TABLE 4
Variable Parameters for the Whole and the Final Data Set

y Variables		*Whole set model*					*Final model*						
No.	*Fastness*	w_j	$\bar{y}_j$	c_{j1}[a]	c_{j2}[a]	ψ_j	w_j	$\bar{y}_j$	c_{j1}[a]	c_{j2}[a]	c_{j3}[a]	c_{j4}[a]	ψ_j
I	Light	0·98	3·22	0·57	0·25	0·30	0·97	3·27	0·55	0·17	0·60	0·49	0·53
II	Washing 60 °C	1·91	8·18	−0·26	0·55	0·15	2·50	10·71	−0·34	0·54	−0·51	0·34	0·29
III	Washing 95 °C	1·07	2·28	−0·32	0·78	0·33	1·05	2·22	−0·30	0·81	−0·04	−0·35	0·34
IV	Alkali	1·00	3·81	−0·53	−0·10	0·23	0·92	3·48	−0·50	0·07	−0·26	−0·54	0·30
V	Perspiration	1·24	4·37	−0·47	0·07	0·17	1·16	4·10	−0·49	0·12	0·56	0·47	0·38

x Variables			*Whole set model*					*Final model*						
No.	*Parameter*		w_i	$\bar{x}_i$	b_{i1}[a]	b_{i2}[a]	ψ_i	w_i	$\bar{x}_i$	b_{i1}[a]	b_{i2}[a]	b_{i3}[a]	b_{i4}[a]	ψ_i
1	σ_p	X_1	4·31	−0·33	0·06	−0·07	0·00	8·23	−0·17	−0·06	−0·14	0·02	−0·31	0·05
2	π		4·27	0·40	0·43	0·19	0·39	3·91	0·47	0·39	0·15	−0·11	−0·39	0·53
3	MR		0·36	1·02	0·33	0·24	0·25	0·39	1·02	0·43	0·25	−0·09	−0·12	0·60
4	v		4·68	0·64	0·43	0·26	0·46	4·49	0·62	0·46	0·24	−0·11	−0·26	0·90
5	σ_p	X_2	2·60	0·87	0·44	0·22	0·44	2·46	0·96	0·40	0·17	0·32	0·38	0·77
6	π		1·33	0·02	−0·15	0·06	0·02	2·09	−0·39	−0·13	−0·04	0·49	−0·60	0·52
7	MR		0·11	1·02	−0·05	0·09	0·00	0·33	2·63	0·00	−0·08	−0·77	0·40	0·64
8	v		1·84	1·32	0·44	0·20	0·49	2·04	1·71	0·43	0·18	0·19	0·06	0·55
9	π	X_4	0·36	0·89	−0·22	0·61	0·74	0·36	0·93	−0·20	0·62	0·03	0·00	0·83
10	MR		0·04	1·86	−0·22	0·61	0·74	0·04	1·89	−0·20	0·62	0·03	0·00	0·83

[a] The PLS solution is found in such a way that $\sum c_j^2 = \sum b_i^2 = 1$.

TABLE 5
Object Scores for the Final Model

Object No.[a]	t_1	u_1	t_2	u_2	t_3	u_3	t_4	u_4
4	−2·19	−1·32	0·32	−0·11	1·19	−0·52	−0·96	−0·98
5	−2·67	−1·94	1·37	0·50	1·17	0·91	−0·93	−1·31
6	−0·95	−0·87	−2·17	−1·20	0·31	−0·03	−1·51	−0·70
7	−1·20	−0·35	−1·65	−1·37	0·30	−0·13	−1·49	−0·67
8	−1·69	−1·09	−0·61	0·01	0·29	−0·60	−1·46	−0·51
9	−2·18	−1·76	0·43	0·39	0·28	−0·21	−1·43	−0·56
10	−2·67	−2·53	1·47	1·60	0·27	0·26	−1·40	−0·74
11	0·05	−0·06	−2·04	−0·15	−0·47	0·03	1·09	1·02
12	−0·19	1·25	−1·52	−1·30	−0·48	1·32	1·11	1·57
13	−0·68	0·12	−0·47	−0·36	−0·49	0·91	1·14	1·96
14	−1·17	−0·98	0·57	1·26	−0·50	0·79	1·17	0·82
15	−1·66	−2·46	1·61	2·78	−0·51	0·41	1·20	0·76
16	−1·05	−0·33	−2·21	−1·16	−0·72	−1·15	−1·02	0·01
17	−1·30	0·35	−1·69	−1·40	−0·73	−0·11	−1·00	−0·27
18	−1·78	−0·63	−0·64	−1·47	−0·74	0·42	−0·97	0·10
19	−2·27	−1·53	0·40	0·33	−0·75	−0·06	−0·94	0·08
20	−2·76	−2·21	1·44	0·72	−0·76	0·33	−0·91	0·03
21	1·13	2·17	−1·89	−1·53	1·49	1·99	1·44	0·64
22	0·89	2·64	−1·37	−1·66	1·48	2·34	1·46	1·20
23	0·40	1·13	−0·33	−0·59	1·47	1·41	1·49	0·54
24	−0·09	−0·20	0·71	0·67	1·46	0·81	1·52	0·75
25	−0·58	−0·40	1·76	0·34	1·44	1·39	1·55	−0·25
26	−0·54	0·04	−2·10	−1·63	0·92	−0·20	−0·94	−0·60
27	−0·78	−0·35	−1·57	−1·25	0·92	−0·24	−0·93	−0·98
28	−1·27	−1·22	−0·53	−0·22	0·90	−0·35	−0·90	−0·36
29	−1·76	−1·81	0·51	0·56	0·89	0·15	−0·87	−0·34
30	−2·25	−2·72	1·55	2·36	0·88	−0·35	−0·84	−0·37
31	−0·15	−1·45	−2·10	0·04	−2·41	−2·61	0·83	0·20
32	−0·45	−0·62	−1·58	−0·44	−2·43	−2·54	0·43	−0·89
33	−1·05	−0·67	−0·53	−0·57	−2·46	−1·85	0·85	−0·23
34	−1·64	−1·34	0·52	−0·22	−2·50	−1·43	0·86	−0·26
35	−2·24	−2·01	1·57	0·46	−2·54	−1·33	0·87	0·57
36	0·23	1·34	−2·08	−1·10	0·92	0·88	1·12	0·44
37	−0·01	1·11	−1·55	−1·14	0·91	0·87	1·14	0·24
38	−0·50	0·12	−0·51	−0·47	0·90	1·10	1·17	1·33
39	−0·99	0·01	0·53	−0·23	0·89	0·84	1·20	0·72
40	−1·48	−1·20	1·57	−0·01	0·88	0·65	1·23	0·19
46	2·72	0·53	−0·57	0·93	−0·04	−0·65	0·72	0·34
47	2·48	2·25	−0·05	−0·10	−0·04	0·20	0·73	0·07

(*continued*)

TABLE 5—*contd.*

Object	t_1	u_1	t_2	u_2	t_3	u_3	t_4	u_4
48	1·99	1·60	0·99	1·52	−0·05	−0·78	0·76	−0·22
49	1·50	1·89	2·03	0·35	−0·07	0·41	0·79	−0·08
50	1·01	0·66	3·08	1·10	−0·08	−0·58	0·82	1·10
51	3·45	1·00	−0·65	0·26	−0·35	−0·42	−0·47	−0·74
52	3·21	1·00	−0·13	0·19	−0·35	−0·32	−0·45	−0·69
53	2·72	0·71	0·91	0·77	−0·37	0·28	−0·42	1·15
54	2·23	0·14	1·95	1·07	−0·38	−0·22	−0·39	0·04
55	1·74	0·15	3·00	0·77	−0·39	−0·63	−0·36	−0·37
56	3·87	3·66	−1·26	0·68	−0·15	−0·20	−1·53	−0·65
57	3·62	3·20	−0·74	−0·69	−0·16	−0·34	−1·51	−1·09
58	3·13	2·55	0·30	0·13	−0·17	0·60	−1·48	−0·56
59	2·65	2·02	1·34	1·04	−0·18	−0·29	−1·45	−0·74
60	2·16	1·10	2·38	0·62	−0·20	−0·92	−1·42	−0·04

[a] See Table 1.

The model so obtained can be used to predict the five properties. The prediction values and the residuals are listed in Table 6 and it is relevant to observe a fairly good agreement between calculated and experimental values (50–100% of the residuals ≤0·5; 80–100% ≤1·0; see Δ values in Table 6). The possibility of using a single mathematical model to predict simultaneously five different properties seems to be a valuable contribution in the application of structure–properties studies.

The results can be illustrated by a number of diagrams which make

TABLE 6
Calculated Values of the Properties and Residuals

Object No.	*Light*		*Washing 60°C*		*Washing 95°C*		*Alkali*		*Perspiration*	
	Calcd	Δ[a]	*Calcd*	Δ	Calcd	Δ	*Calcd*	Δ	*Calcd*	Δ
4	2·81	−0·31	4 31	0·19	2·80	−0·30	4·63	0·37	4·39	−0·89
5	2·72	−0·22	4·48	−0·48	3·33	0·67	4·85	0·15	4·59	−0·09
6	2·57	−0·07	3·99	0·01	1·63	−0·13	4·39	0·61	3·50	0·50
7	2·52	−0·02	4·08	−0·08	1·89	−0·39	4·50	0·00	3·60	−0·10
8	2·43	0·07	4·25	0·25	2·42	0·08	4·72	−0·22	3·81	−0·31
9	2·34	0·16	4·42	0·08	2·94	0·06	4·93	0·07	4·01	−0·01
10	2 25	0·25	4·58	−0·08	3·47	1·03	5·15	−0·15	4·21	0·29
11	3·26	0·24	4·18	0·32	1·10	0·40	3·47	0·03	3·45	0·55
12	3·22	0·78	4·27	−0·27	1·37	−0·37	3·58	−1·08	3·55	0·45

TABLE 6—*contd.*

Object No.	*Light*		*Washing 60°C*		*Washing 95°C*		*Alkali*		*Perspiration*	
	Calcd	Δ[a]	*Calcd*	Δ	*Calcd*	Δ	*Calcd*	Δ	*Calcd*	Δ
13	3·13	0·37	4·43	0·07	1·89	−0·39	3·79	−1·29	3·75	0·75
14	3·03	0·47	4·60	−0·10	2·42	1·08	4·01	−0·51	3·96	0·54
15	2·94	0·56	4·77	0·23	2·94	1·56	4·23	0·27	4·16	0·84
16	2·21	0·29	4·17	0·33	1·59	−0·59	4·49	0·01	3·29	−0·29
17	2·17	0·33	4·25	−0·25	1·85	−0·35	4·60	−0·10	3·39	0·11
18	2·08	0·42	4·42	−0·42	2·38	−0·88	4·82	−0·32	3·59	0·41
19	1·98	0·52	4·59	−0·09	2·90	0·10	5·03	−0·53	3·79	0·21
20	1·89	0·61	4·76	−0·26	3·43	0·07	5·25	−0·25	4·00	0·50
21	4·59	−0·09	3·85	−0·35	0·84	0·16	2·60	−0·60	3·84	0·16
22	4·55	−0·05	3·94	−0·44	1·10	−0·10	2·71	−1·71	3·94	0·06
23	4·45	0·05	4·11	−0·11	1·63	−0·13	2·93	0·07	4·15	−0·15
24	4·36	−0·86	4·28	0·22	2·16	0·34	3·15	−0·65	4·35	0·15
25	4·27	−0·77	4·45	−0·45	2·68	0·32	3·37	0·13	4·55	−0·05
26	3·12	0·62	3·91	0·09	1·47	−0·47	3·97	0·03	3·69	−0·19
27	3·08	0·58	4·00	0·00	1·74	−0·24	4·08	0·42	3·79	−0·29
28	2·99	−0·49	4·17	0·33	2·26	−0·26	4·29	0·21	3·99	0·01
29	2·90	−0·40	4·34	0·16	2·79	0·21	4·51	−0·01	4·20	0·30
30	2·80	−0·30	4·51	0·49	3·31	1·19	4·73	−0·23	4·40	0·10
31	2·46	−0·96	4·42	0·08	1·20	0·30	3·89	0·61	2·92	0·08
32	2·66	−1·16	4·46	0·04	1·39	0·61	3·80	0·70	3·02	−1·02
33	2·76	−1·26	4·60	−0·10	1·87	0·13	3·87	0·13	3·22	−0·72
34	2·82	−1·32	4·74	−0·24	2·35	0·15	3·97	0·53	3·42	−0·42
35	2·85	−0·85	4·89	0·11	2·84	−0·34	4·09	0·91	3·62	−0·12
36	3·91	0·59	3·97	0·03	1·01	−0·01	3·13	0·37	3·85	−0·35
37	3·86	0·64	4·06	−0·06	1·27	−0·27	3·24	0·76	3·95	−0·45
38	3·77	0·73	4·23	0·27	1·80	−0·80	3·45	0·55	4·15	0·35
39	3·68	0·82	4·40	0·10	2·33	−0·83	3·67	0·83	4·36	−0·36
40	3·59	−0·09	4·57	−0·07	2·85	−0·85	3·89	1·11	4·56	−0·06
46	4·55	−1·05	4·02	0·48	1·22	0·78	2·54	−0·04	2·83	0·67
47	4·50	0·50	4·10	−0·10	1·49	0·01	2·65	0·35	2·93	−0·43
48	4·41	0·09	4·27	0·23	2·01	0·99	2·87	−0·37	3·13	−1·13
49	4·32	0·18	4·44	−0·44	2·54	−0·04	3·08	−0·58	3·33	−0·83
50	4·23	−0·23	4·61	0·39	3·07	−1·07	3·30	−0·80	3·54	−0·54
51	4·43	0·07	3·91	0·09	1·24	0·26	2·64	1·86	2·29	0·71
52	4·38	0·12	3·99	0·01	1·50	0·00	2·75	1·25	2·39	0·61
53	4·29	0·21	4·16	0·34	2·03	−0·53	2·97	0·03	2·59	1·41
54	4·20	0·30	4·33	0·17	2·55	−0·55	3·18	1·32	2·80	0·70
55	4·10	−0·10	4·50	0·00	3·08	−1·08	3·40	1·10	3·00	0·00
56	4·37	0·13	3·71	−0·21	1·06	0·44	2·70	−1·20	2·00	−0·50
57	4·32	0·18	3·79	−0·29	1·32	0·18	2·81	−0·31	2·10	−0·60
58	4·23	0·27	3·96	−0·46	1·84	0·66	3·03	−1·03	2·30	0·20
59	4·14	0·36	4·13	−0·13	2·37	0·63	3·25	−0·75	2·51	−0·51
60	4·04	−0·54	4·30	0·20	2·90	−0·90	3·47	−0·97	2·71	−0·21

[a] Δ = Observed − calculated value. Observed values are reported in Table 1.

easier the interpretation of the structural features influencing the various fastness properties and indicate the molecular structures better suited to improve the technological properties of the dyes. This discussion is limited to the first two components since they account for the larger degree of explained variance (45 over 60 %) and are easier to understand.

Figure 1 represents the loadings plot of both the *X*- and *Y*-blocks. This plot illustrates the associations between the variables. Variables highly correlated to each other lie either very close ($r \approx 1$) or in a symmetrical position with respect to the centre of the diagram ($r \approx -1$). In the present case we observe that property I (lightfastness) is highly associated with variables 2, 3, 4, 5, 8 (descriptors π, MR, ν of X_1 and σ_p and ν of X_2: fastness increases on increasing each of these variables), whereas properties IV and V (alkali and perspiration fastness, highly associated with each other) depend upon the same parameters but in an opposite way. Accordingly, any improvement in lightfastness will result in a decrease in the other two properties. On the contrary, properties II and III (washing fastness) are almost independent of the other properties and appear to be associated with variables 9 and 10 (descriptors of the X_4

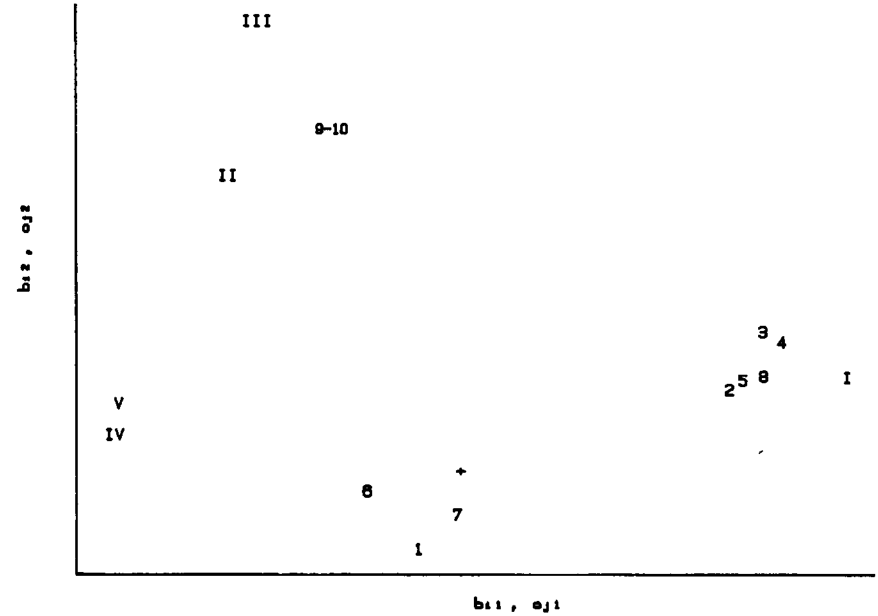


Fig. 1. Loadings plot for the first two components.

position). Consequently, fastness to washing is mainly due to the length of the X_4 chain.

The inner relationships between the latent variables of the *X*- and *Y*-blocks (*t* and *u*, respectively) for the first and the second dimension are reported in Figs 2 and 3, and allow the best structural features for each considered fastness to be established. Figure 2 indicates therefore the increase in lightfastness—and corresponding decrease in alkali and perspiration fastness—(u_1) as a function of the latent variable t_1, which can be expressed as a linear combination of the X_1 and X_2 descriptors cited. Figure 2(A) reports the code letter for the X_1 substituent as in Table 1, Fig. 2(B) the code letter for the X_2 substituent and Fig. 2(C) the code letter for the X_4 chain.

The plots clearly indicate that the highest lightfastness values result from the presence of a chlorine or methoxy group at X_1, a nitro group at X_2 and a short chain at X_4. However, the analogous plot for the whole set model had indicated that the presence of a nitro group at X_3 was not useful for increasing lightfastness.

Figure 3 illustrates a similar plot for the second significant dimension. In this case the washing fastness (u_2) is related to the second latent variable t_2, which is due almost exclusively to the X_4 descriptors. Consequently individual dyes are reported in Fig. 3 by their X_4 chain code. Clearly the fastness is increased by increasing the chain length.

The length of the X_4 chain therefore favours washing fastness, but decreases lightfastness. As a consequence one is left with the problem of optimizing the structural features in order to achieve the best compromise to fit light and washing fastness at the same time. This information can be obtained by inspection of Fig. 4, where the second latent variable of the *Y*-block (u_2) is plotted against the first latent variable of the same block (u_1). In such a plot the objects lying in the upper right corner are the dyes whose structure is appropriate to obtain satisfactory levels of light and washing fastness at the same time.

Again, Fig. 4(A) indicates dyes by the X_1 substituent codes, Fig. 4(B) by the X_2 substituent codes and Fig. 4(C) by the X_4 chain codes. Clearly, molecular structures having the best compromise between light and washing fastness require a substituent, preferably a methoxy group, at X_1, a nitro group at X_2 and an intermediate chain (C_7–C_{11}) at X_4.

The PLS analysis therefore permits the definition of structural features better suited to enhance technological properties.

In order to confirm the general picture described above and to allow a

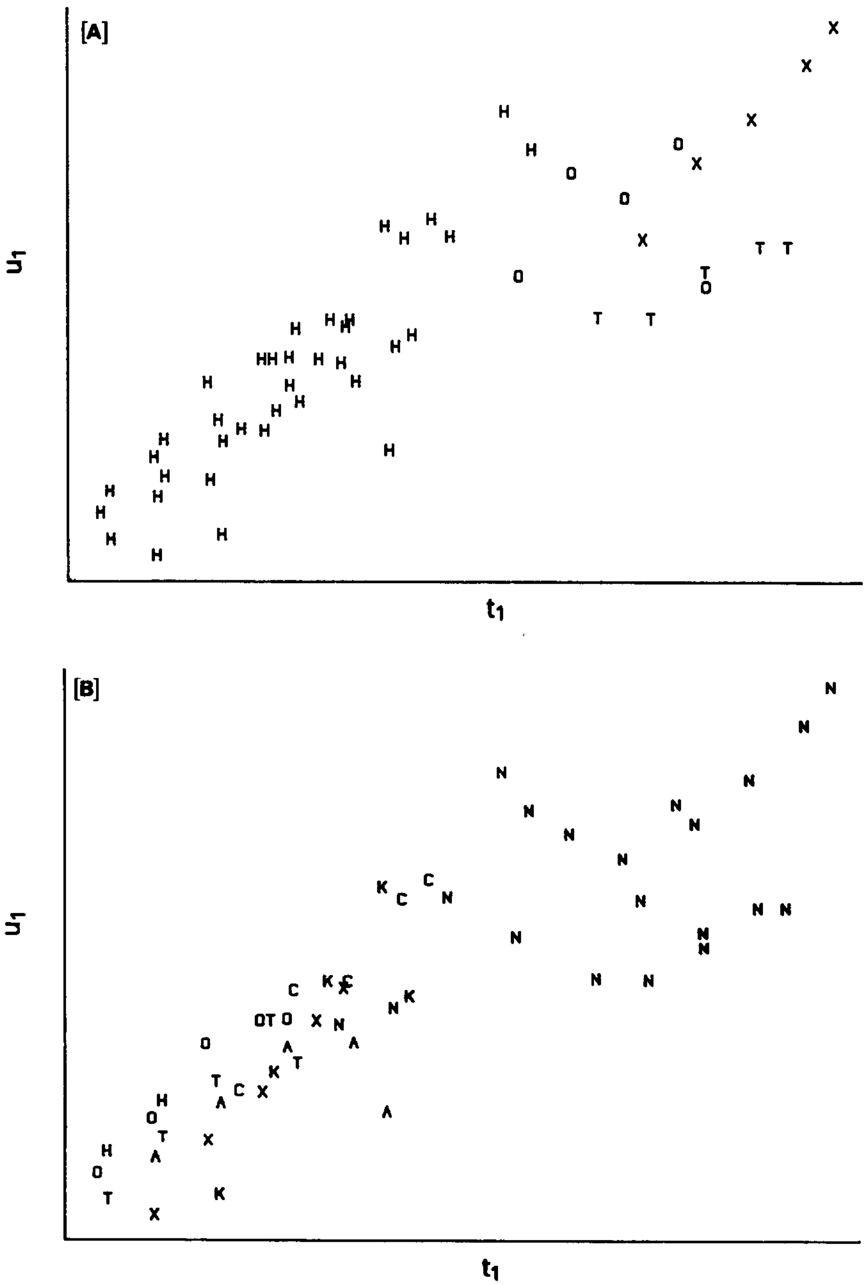


Fig. 2. Plots of the latent variables of the *X* block against the latent variables of the *Y* block for the first dimension. (A) Dyes are indicated by the X_1 substituent codes (see Table 1); (B) dyes are indicated by the X_2 substituent codes; (C) dyes are indicated by the X_4 substituent codes.

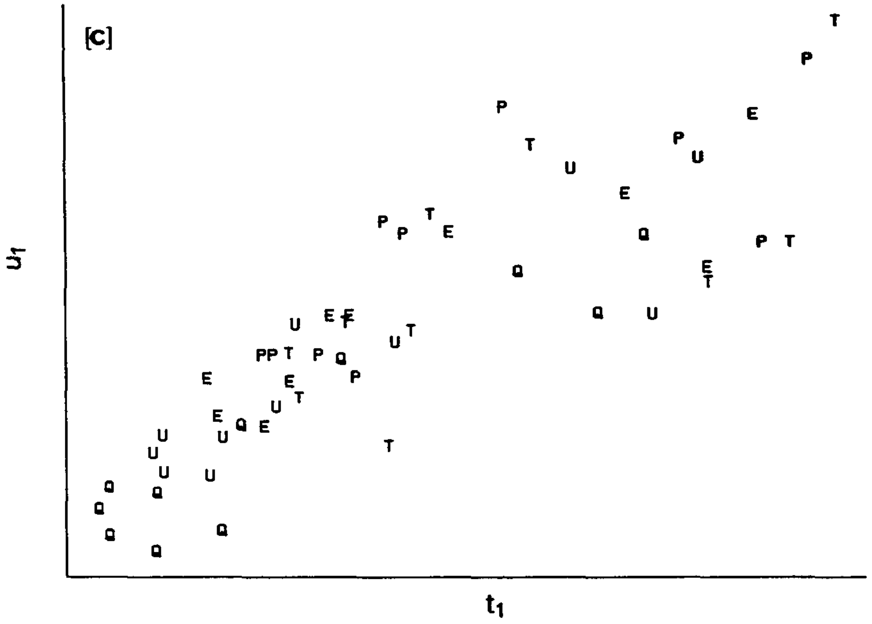


Fig. 2—*contd.*

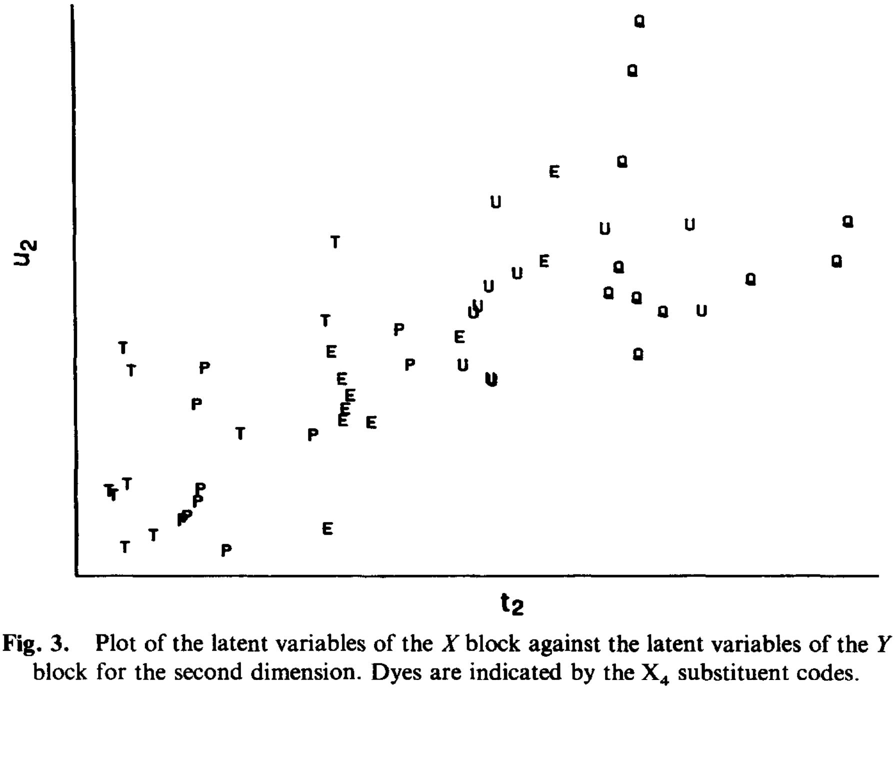


Fig. 3. Plot of the latent variables of the *X* block against the latent variables of the *Y* block for the second dimension. Dyes are indicated by the X_4 substituent codes.

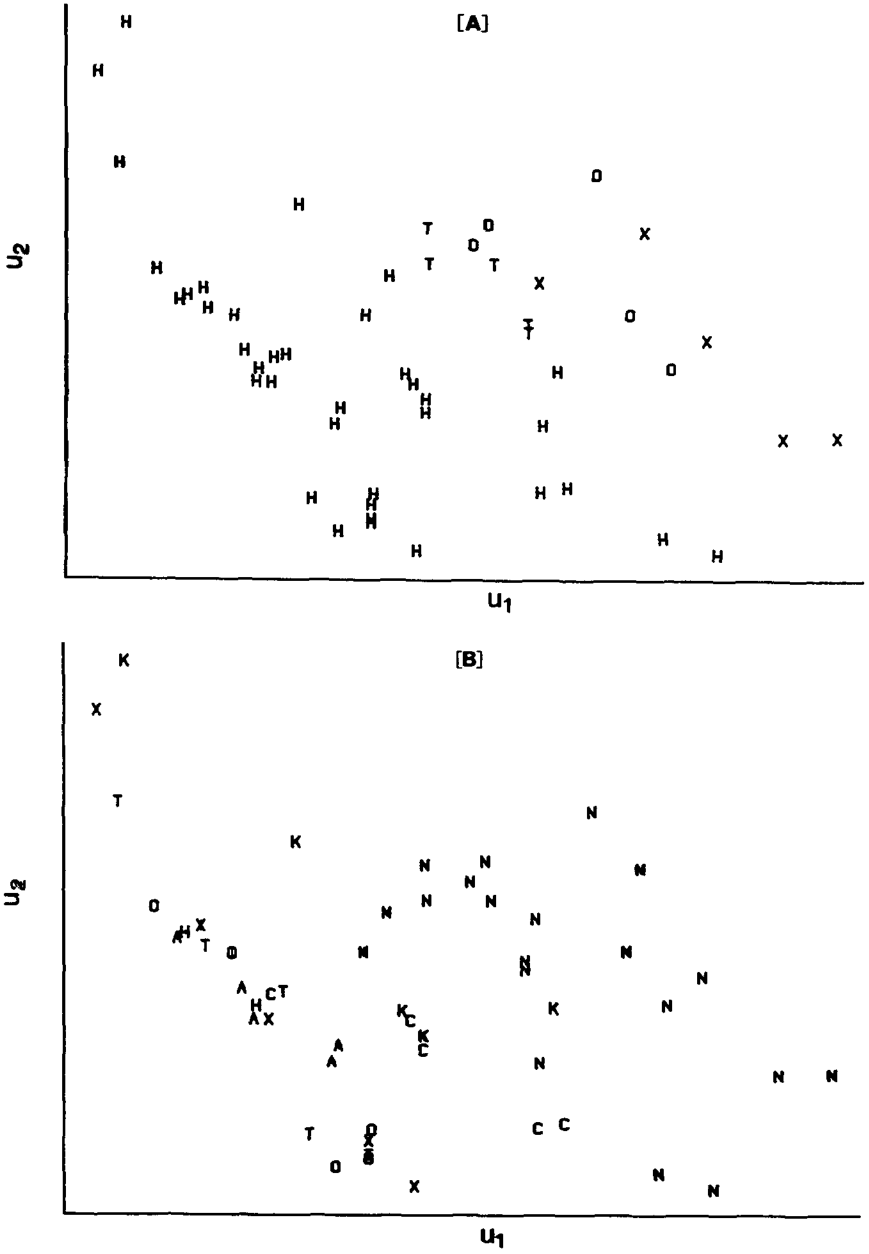


Fig. 4. Plot of the second against the first latent variable of the *Y* block. (A) Dyes are indicated by the X_1 substituent codes; (B) dyes are indicated by the X_2 substituent codes; (C) dyes are indicated by the X_4 substituent codes.

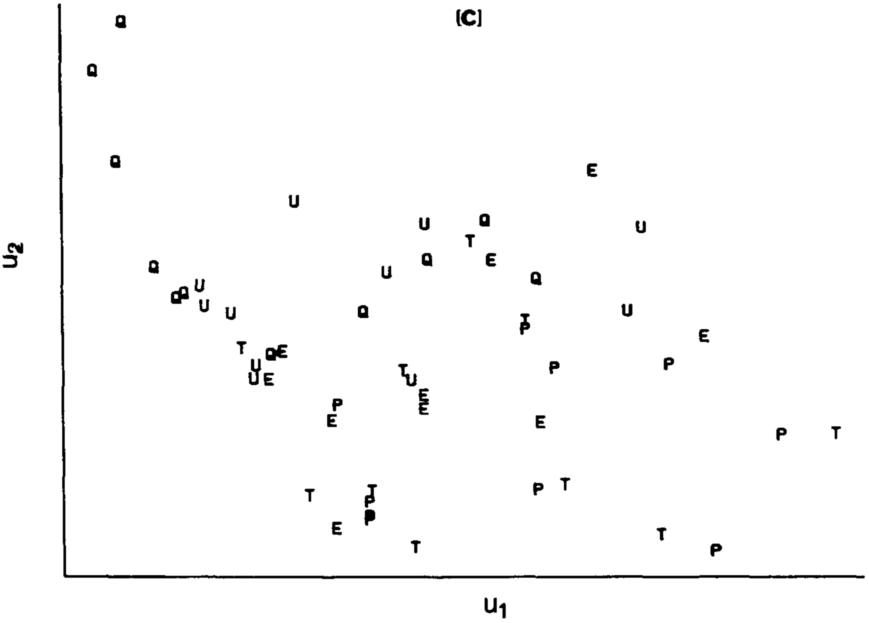


Fig. 4.—*contd.*

deeper consideration of the structural parameters responsible for each individual fastness, it is also possible to apply the PLS analysis for each individual property, by means of eqns (2) and (3). All these analyses were carried out on the same 52 objects previously selected. A further reason for investigating individual properties lies in the better predicting ability of a restricted model compared to that of a general model. This is true either within a group of homogeneous objects in an inhomogeneous whole set or to describe more adequately an individual property among others when they are partly correlated. By far the most important property is lightfastness, which is therefore described in some detail.

The PLS model in this case requires two components (latent variables), the first of which explains 63 % of the property variance and the second a further 18 % up to 81 %. Interestingly, the first latent variable of the X-block involves the same variables (2, 3, 4, 5 and 8: see b values in Table 7) already seen in the global analysis, but the variance explained is much higher and therefore the prediction ability of this model should be higher. Furthermore a second latent variable is significant (i.e. the residuals of the Y-vector still contain systematic information which can be related to the descriptor block) and again picks up variables 2, 4, 5 and 8.

TABLE 7
Percentage of Variance Explained and Variable Loadings for the Individual Property Models Limited to the First Component

	y Variables[a]				
	I	*II*	*III*	*IV*	*V*
V_1 (%)[b]	*63*	*45*	*56*	*47*	*50*
x Variables[a] *No.*			*Loadings* b_{i1}		
1	−0·11	−0·07	0·05	0·06	0·04
2	0·37	−0·39	−0·17	−0·38	−0·41
3	0·46	−0·34	−0·20	−0·44	−0·44
4	0·47	−0·40	−0·20	−0·46	−0·48
5	0·43	−0·32	−0·24	−0·43	−0·36
6	−0·10	−0·04	0·17	0·17	0·17
7	−0·07	0·14	−0·13	−0·03	−0·06
8	0·46	−0·38	−0·27	−0·44	−0·42
9	−0·07	0·39	0·60	0·14	0·17
10	−0·07	0·39	0·60	0·14	0·17

[a] See Table 4.
[b] Percentage of the total variance explained by the model with $A = 1$.

Accordingly this result confirms that lightfastness depends upon the type of ring substitution at X_1 and X_2 and a closer inspection of the results indicates that the second component is required to account for the somewhat peculiar behaviour of dyes substituted at X_2 by an acetylamino or cyano group. Hence, this second component is only useful to model the lightfastness within a group of structures not optimal with respect to the property.

A similar PLS computation can be made on each individual fastness. The results, limited to the first component which is easier to understand and has the maximum technological interest, in terms of fraction of variance explained and loadings of the descriptor variables, are listed in Table 7.

Accordingly, it is confirmed that the washing fastness depends mainly upon the X_4 chain, but also upon some of the parameters of the ring substituents X_1 and X_2. Similarly, the fastness to alkali and perspiration depends negatively upon the non-electronic descriptors of X_1 and the electronic and steric descriptors of X_2.

CONCLUSIONS

The PLS analysis of the fastness properties of the examined dyes as a function of physicochemical descriptors thus permits the fastness properties and structural features to be related and establishes the development of a mathematical model which can be used predictively.

In particular, we have shown that the lightfastness is inversely related to alkali and perspiration fastness and almost independent of washing fastness. The analysis has also indicated that the best structure to meet the light and the washing fastness requirements contains a methoxy group at X_1, a nitro group at X_2 and a C_7–C_{11} chain at X_4.

PLS has thus been shown to be a powerful tool which does not depend, as MRA in traditional QSAR, on the *a priori* selection of descriptors and dimensionality and therefore is particularly appropriate for interpreting the data. Furthermore, establishment of the model permits the prediction of fastness of dyes not yet tested, provided the structural modifications are limited to the positions considered and the descriptors are available or can be reasonably estimated. In this respect PLS potential is far beyond that of the Free–Wilson method.

ACKNOWLEDGEMENTS

Svante Wold is thanked for providing the SIMCA–MACUP package and for stimulating discussion; and the Ministero della Pubblica Istruzione (Italy) is thanked for a research grant.

REFERENCES

1. R. Grecu, M. Pieroni and R. Carpignano, *Dyes and Pigments*, **2**, 305 (1981).
2. R. Carpignano, E. Barni, G. Di Modica, R. Grecu and G. Bottaccio, *Dyes and Pigments*, **4**, 195 (1983).
3. K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate analysis*. London, Academic Press (1979).

4. N. Draper and H. Smith, *Applied regression analysis*, 2nd edn. New York, Wiley (1981).
5. C. H. Hansch, in *Correlation analysis in chemistry—Recent advances*, eds N. B. Chapman and J. Shorter, p. 397. New York, Plenum Press (1978).
6. S. M. Free, Jr and J. W. Wilson, *J. Med. Chem.*, **7**, 395 (1964).
7. T. Fujita and T. Ban, *J. Med. Chem.*, **14**, 148 (1971).
8. S. Wold, *J. Chem. Inf. Comp. Sci.*, **23**, 6 (1983).
9. S. Wold, C. Albano, W. J. Dunn III, K. Esbensen, S. Hellberger, E. Johansson and Sjöström, in *Food research and data analysis*, eds H. Martens and H. Russwurm Jr, p. 147. London, Applied Science (1983).
10. S. Wold, in *Evaluation and optimization of laboratory methods and analytical procedures*, eds D. L. Massart, A. Dijkstra and L. Kaufman, p. 385. Amsterdam, Elsevier (1978).
11. S. Wold and M. Sjöström, in *Chemometrics: Theory and application*, ed. B. R. Kowalski, p. 243. Washington D.C., American Chemical Society (1977).
12. S. Wold, *Technometrics*, **20**, 397 (1978).
13. E. Barni, P. Savarino, G. Di Modica, R. Carpignano, S. S. Papa and G. Giraudo, *Dyes and Pigments*, **5**, 15 (1984).
14. A. J. Stuper, W. E. Brugger and P. C. Jurs, in *Chemometrics: Theory and application*, ed. B. R. Kowalski, p. 165. Washington D.C., American Chemical Society (1977).
15. C. Hansch and A. Leo, *Substituent constants for correlation analysis in chemistry and biology*. New York, Wiley (1979).
16. W. J. Dunn III and S. Wold, *J. Med. Chem.*, **21**, 922 (1978).