REVUE DE STATISTIQUE APPLIQUÉE

J.-P. GAUCHI

Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation

Revue de statistique appliquée, tome 43, n° 1 (1995), p. 65-89

<http://www.numdam.org/item?id=RSA_1995__43_1_65_0>

Rev. Statistique Appliquée, 1995, **XLIII** (1), 65-89

UTILISATION DE LA RÉGRESSION PLS POUR L'ANALYSE DES PLANS D'EXPÉRIENCES EN CHIMIE DE FORMULATION

J.-P. Gauchi

Rhône-Poulenc, Centre de Recherches d'Aubervilliers, France

RÉSUMÉ

En Chimie, les plans d'expériences non orthogonaux sont fréquents dans les problèmes de mélange-formulation. Dans ce cadre, la planification expérimentale présente des difficultés particulières tant au stade de la construction du plan d'expériences proprement dit, qu'au stade de l'analyse des résultats. En effet, la construction du plan d'expériences est rendue délicate par l'existence de nombreuses contraintes sur le domaine expérimental qui font perdre toute symétrie à celui-ci. Une conséquence importante est la perte d'orthogonalité de la matrice du modèle. Par ailleurs, la modélisation des réponses en fonction des variables explicatives qui sont souvent les proportions des constituants du mélange n'est pas aisée si on désire conserver un sens interprétable (au plan de la chimie) aux coefficients du modèle. Sans insister ici sur le premier stade, cet article montre que, pour l'analyse des résultats, la régression PLS est très supérieure à la régression linéaire multiple, celle-ci étant totalement inadaptée dans un tel contexte où les variables explicatives sont obligatoirement très corrélées. Comparativement à cette dernière, la régression PLS est mise en œuvre ici sur un exemple réel de formulation de mortiers autolissants pour sols.

Mots-clés : *chimie de formulation, mélanges, plans d'expériences, régression PLS.*

SUMMARY

In chemistry, non-orthogonal experimental designs are common in mixture-formulation problems. In this framework experimental designing presents particular difficulties, both in constructing the actual experimental designs and in analyzing the results. In fact, the construction of experimental design is made tricky by the existence of numerous constraints on the experimental domain that make it totally asymmetric. An important consequence is the loss of orthogonality of the model's matrix. Moreover, it is not easy to modelize the responses in terms of the explanatory variables, which often correspond to the proportions of the mixture's constituents, if the coefficients of the model are to remain (chemically) interpretable. Without insisting here on the first stage, this article shows that, for the analysis of the results, the PLS regression is greatly superior to the multiple linear regression, as the latter is not all apt in a context where the explanatory variables are necessarily closely correlated. The two kinds of regression are here compared, using an authentic example of formulation for self-smoothing floor mortars.

Keywords : *formulation chemistry, mixtures, experimental design, PLS regression.*

1. Introduction

Dans les sciences expérimentales, et notamment en Chimie, le plan d'expériences le plus répandu est le plan factoriel, complet ou fractionnaire. Nous renvoyons le lecteur à l'ouvrage de BOX, HUNTER et HUNTER [1] pour une excellente introduction sur le sujet. On sait que la caractéristique essentielle de ce plan est sa structure orthogonale, c'est-à-dire que les corrélations deux à deux entre les variables explicatives et/ou les interactions prises en compte est nulle (cette définition pour variables continues trouve son équivalent pour des variables qualitatives). L'avantage de cette orthogonalité est qu'elle permet ultérieurement le calcul d'un modèle de régression pour lequel l'estimation de chacun des coefficients traduira exactement l'effet marginal de la variable explicative correspondante. Cependant, pour certains types de problèmes, par exemple les problèmes de mélange rencontrés en Chimie de formulation, il n'est pas possible d'utiliser un tel type de plan d'expériences de par la nature fondamentalement corrélée du système à étudier. Ici, les variables explicatives sont les proportions des constituants du mélange. Même pour les cas les plus simples, ceux où la seule contrainte (évidente) est que la somme des proportions soit égale à 1, les plans d'expériences proposés par SCHEFFE [8] conduisent à des corrélations élevées. Par exemple, le plan du Tableau 1 conduit à un coefficient de corrélation de $-0{,}5$ entre les 3 variables X_1, X_2 et X_3. On trouvera dans l'ouvrage spécialisé de CORNELL [4] de nombreux plans pour problèmes de mélange.

TABLEAU 1
Un plan d'expériences réseau simplex de Scheffé

EXP	X_1	X_2	X_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1/2	1/2	0
5	1/2	0	1/2
6	0	1/2	1/2
7	1/3	1/3	1/3

Le but de cet article n'est pas la construction de plans d'expériences visant à minimiser de telles corrélations mais plutôt de montrer l'avantage à utiliser la régression PLS pour analyser les résultats issus des plans d'expériences fortement non orthogonaux (à cause des nombreuses contraintes) rencontrés dans ce type de problème. Toutefois, pour apprécier l'intérêt de la méthode, il nous semble judicieux, au préalable, à partir d'un exemple réel, d'examiner le domaine expérimental, la forme du modèle postulé par le chimiste et leurs conséquences sur la structure de corrélation de la matrice du modèle.

L'article est constitué de deux parties :

– dans une première partie on exposera un problème réel de formulation avec ses contraintes, on décrira le domaine expérimental, le plan d'expériences et le modèle postulé,

– dans une deuxième partie on fera l'analyse des résultats en comparant les régressions simple et multiple à la régression PLS.

2. Première partie : postulation d'un modèle et construction du plan d'expériences

2.1. Présentation du problème expérimental

Remarque préliminaire :

Compte-tenu du haut niveau de confidentialité du sujet de recherche dont relève l'exemple étudié ici, les noms chimiques exacts ne peuvent pas être communiqués. Nous assurons le lecteur que ceci ne le gênera aucunement dans la compréhension de l'article.

Le problème à traiter est de mettre en évidence l'influence de la variation des quantités relatives des constituants d'un mélange sur plusieurs réponses physico-chimiques en vue d'optimiser ultérieurement la composition du mélange. L'optimisation proprement dite ne sera pas traitée dans cet article. Le mélange étudié est un mortier autolissant pour sols à huit constituants qui seront nommés ici X_j pour j variant de 1 à 8. Les proportions des constituants X_j seront notées x_j. Pour que le mélange de ces huit constituants ait une chance d'aboutir à un mortier réel il est obligatoire d'imposer certaines contraintes sur les proportions x_j. Dans le cas présent, l'homme de l'art est capable de définir les 10 contraintes suivantes :

a) $$0,08 \leqslant \frac{x_2}{x_1} \leqslant 0,20$$

b) $$0,42 \leqslant \frac{(x_1 + x_2)}{(x_7 + x_8)} \leqslant 0,68$$

c) $$0,55 \leqslant \frac{x_5}{(x_1 + x_2)} \leqslant 0,65$$

d) $$0,01 \leqslant \frac{(x_3 + x_4)}{(x_1 + x_2)} \leqslant 0,03$$

e) $$0,01 \leqslant \frac{x_6}{(x_1 + x_2 + x_3 + x_4 + x_7 + x_8)} \leqslant 0,05$$

f) $$0,20 \leqslant \frac{x_5}{(x_1 + x_2 + x_3 + x_4 + x_6 + x_7 + x_8)} \leqslant 0,40$$

g) $$0,20 \leqslant \frac{(x_1 + x_2)}{(x_1 + x_2 + x_3 + x_4 + x_6 + x_7 + x_8)} \leqslant 0,50$$

h) $$0,30 \leqslant \frac{x_6}{(x_1 + x_2 + x_3 + x_4 + x_6 + x_7 + x_8)} \leqslant 0,80$$

i) $$\frac{x_7}{x_8} = 3$$

j) $$\sum_{j=1}^{8} x_j = 1$$

2.2. *Description du domaine expérimental*

Ce domaine s'inscrit dans un espace à 8 dimensions. Les axes de cet espace sont les proportions x_j des 8 constituants. Les 2 contraintes i/ et j/ limitent à 6 dimensions le domaine expérimental. Compte-tenu de la forme linéaire des contraintes, celui-ci est un polyèdre convexe. Pour le décrire de façon correcte, il nous faut calculer au moins les coordonnées des sommets et celles des milieux d'arêtes dans l'espace considéré. Nous verrons plus loin pourquoi ces coordonnées suffisent. Notons $w = x_7 + x_8$ pour simplifier le problème. On doit calculer les coordonnées du polyèdre $[X_1X_2X_4X_5X_6W]$. Le programme informatique CANDI [2] nous permet un tel calcul (c'est le problème classique du simplexe en programmation linéaire) après saisie des contraintes relationnelles mises auparavant sous la forme d'inéquations :

– d'après a) on a :

$$0 \leqslant -0,08x_1 + x_2$$
$$0 \leqslant +0,20x_1 - x_2$$

– d'après b) on a :

$$0 \leqslant +x_1 + x_2 - 0,42w$$
$$0 \leqslant -x_1 - x_2 + 0,68w$$

– d'après c) on a :

$$0 \leqslant -0,55x_1 - 0,55x_2 + x_5$$
$$0 \leqslant +0,65x_1 + 0,65x_2 - x_5$$

– d'après d) on a :

$$0 \leqslant -0,01x_1 - 0,01x_2 + x_3 + x_4$$
$$0 \leqslant +0,03x_1 + 0,03x_2 - x_3 - x_4$$

– d'après e) on a :

$$0 \leqslant -0,01x_1 - 0,01x_2 - 0,01x_3 - 0,01x_4 + x_6 - 0,01w$$
$$0 \leqslant +0,05x_1 + 0,05x_2 + 0,05x_3 + 0,05x_4 - x_6 + 0,05w$$

– d'après f) on a :

$$0 \leqslant -0,2x_1 - 0,2x_2 - 0,2x_3 - 0,2x_4 + x_5 - 0,2x_6 - 0,2w$$
$$0 \leqslant +0,4x_1 + 0,4x_2 + 0,4x_3 + 0,4x_4 - x_5 + 0,4x_6 + 0,4w$$

– d'après g) on a :

$$0 \leqslant +0,8x_1 + 0,8x_2 - 0,2x_3 - 0,2x_4 - 0,2x_6 - 0,2w$$
$$0 \leqslant -0,5x_1 - 0,5x_2 + 0,5x_3 + 0,5x_4 + 0,5x_6 + 0,5w$$

– d'après h) on a :

$$0 \leqslant -0,3x_1 - 0,3x_2 - 0,3x_3 - 0,3x_4 + 0,7x_6 - 0,3w$$
$$0 \leqslant +0,8x_1 + 0,8x_2 + 0,8x_3 + 0,8x_4 - 0,2x_6 + 0,8w$$

Le programme CANDI donne les coordonnées des 60 sommets et des 162 milieux d'arêtes, soit un total de 222 points dans l'espace considéré, décrivant la frontière du domaine expérimental.

2.3. Postulation d'un modèle

La forme du modèle postulé dépend de l'objectif du chercheur. Le plan d'expériences dépendra lui-même de la forme du modèle.

La mise au point d'une formulation optimisée ne peut se faire en une seule étape. Lors d'une première étape, un souhait naturel est de repérer les variables explicatives significativement influentes, soit par leurs effets principaux, soit par leurs interactions (celles-ci traduisant en Chimie les synergies et/ou les antagonismes).

Dans l'exemple présent, le modèle de régression suivant est postulé pour chacune des réponses à étudier :

$$y = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \sum_{k=1}^{6} \sum_{j>k}^{7} \beta_{kj} X_k X_j + \varepsilon$$

soit un total de $1 + 8 + 21 = 30$ termes dans le modèle. Les interactions avec la variable X_8 ne sont pas présentes, n'ayant pas de raison d'être, sur le plan de la chimie.

2.4. Construction du plan d'expériences

Pour peu que la matrice du modèle postulé soit de rang plein, on sait qu'il est facile de générer un plan d'expériences D-optimal discret (à N points-supports) à partir d'un réseau d'expériences candidates donné, voir par exemple FEDOROV [5] pour la théorie, MATHIEU [6] pour des exemples d'application et la procédure OPTEX du logiciel SAS/QC [7] pour construire les plans. Ici, le réseau candidat serait constitué des 222 points calculés précédemment. On sait aussi que trouver un tel plan, en cas de postulation d'un modèle de régression linéaire, revient à chercher un plan d'expériences de volume maximal dans l'espace du modèle. Ceçi explique *a posteriori* l'absence de points intérieurs au domaine expérimental, les sommets et les milieux d'arêtes se situant évidemment sur la frontière.

Néanmoins, dans l'exemple exposé ici, on a préféré une méthode de construction différente, en fait une méthode de classification non hiérarchique, basée sur le critère inertiel de Ward. Sans pour autant expliquer dans le détail ce choix, on peut indiquer entre autres raisons, pour ce problème précis, une grande difficulté calculatoire lors du déroulement des algorithmes de la procédure OPTEX, les étapes intermédiaires revenant à tenter d'inverser des matrices quasi-singulières. La procédure FASTCLUS

de SAS fournit une partition à nombre de classes fixé à l'avance. Après plusieurs tentatives, le choix final a porté sur une partition à 30 classes. Les expériences sont représentées par les centres de gravité des 30 classes. Cinq répétitions de la même expérience ont été prévues au centre du domaine expérimental en vue de fournir quelques degrés de liberté pour l'analyse ultérieure des résultats et tester la reproductibilité des expériences. Le plan d'expériences était donc constitué d'un total de 35 expériences respectant le nombre maximal autorisé de 40, dû aux contraintes budgétaires. Exprimé dans les proportions de chaque constituant (multipliées par 1000 car le chimiste a pour coutume de raisonner en parts pour 1000 g), le plan d'expériences figure dans le Tableau 2.

Remarque :

La somme des parts pour chaque expérience a été normalisée exactement à 1000 car, en pratique, on trouve des sommes très légèrement différentes de 1000 compte-tenu des erreurs d'arrondi lors des calculs antérieurs.

La matrice des corrélations deux à deux entre l'ensemble des variables explicatives (8 effets principaux et 21 interactions) a été calculée et un extrait figure dans le Tableau 3 : de nombreuses corrélations très fortes apparaissent, souvent supérieures à 0,90.

On peut résumer cette forte multicolinéarité par le calcul des facteurs d'inflation de la variance des estimateurs, voir par exemple TOMASSONE [11], qui apparaissent au Tableau 4. L'idéal (toutes les corrélations deux à deux égales à zéro) correspond à des facteurs d'inflation tous égaux à 1.

TABLEAU 2
Le plan d'expériences, les proportions ont été multipliées par 1000.

Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	303.091	24.1993	0.00000	3.29990	179.995	8.0998	360.989	120.330
2	281.889	38.5985	4.29983	1.06996	192.092	10.5996	353.586	117.862
3	277.064	34.7955	0.00000	6.85911	183.776	38.8949	343.955	114.652
4	276.383	32.4331	4.14423	2.04208	190.795	38.5393	341.749	113.916
5	253.813	39.5020	3.04015	1.00005	173.809	19.3410	382.119	127.373
6	243.566	34.3952	2.79961	3.43952	166.677	27.5961	391.145	130.382
7	277.070	34.7942	0.00000	6.85925	183.780	38.8957	343.942	114.654
8	294.891	28.2992	1.85994	1.83994	182.295	10.2997	360.389	120.130
9	303.021	24.2017	0.00000	3.30023	180.013	8.1006	361.025	120.342
10	275.792	41.9987	1.99994	1.19996	190.594	11.8996	357.389	119.130
11	269.381	36.8974	2.99979	1.03993	197.786	12.5991	359.475	119.825
12	283.614	28.0014	9.35047	0.00000	186.609	13.2007	359.418	119.806
13	266.681	53.2963	9.59933	0.00000	183.887	15.8989	352.975	117.658
14	269.787	39.3980	4.54977	2.09990	182.491	26.0987	356.682	118.894
15	290.326	23.2021	0.80007	2.34021	172.416	39.4035	353.632	117.877
16	285.097	25.9997	0.93999	7.79992	180.198	39.0396	345.697	115.232
17	265.487	40.3980	0.95995	5.14974	184.391	38.8381	348.583	116.194
18	261.871	41.4954	2.33974	2.89968	187.279	38.6957	349.062	116.354
19	274.381	29.7979	6.69953	0.00000	183.587	38.8973	349.976	116.659
20	258.974	51.7948	5.19948	0.00000	181.082	38.9961	347.965	115.988
21	257.905	37.2007	2.96006	1.10002	170.903	18.8504	383.308	127.769
22	238.207	29.8009	2.60008	0.80002	166.705	14.1404	410.812	136.937
23	235.983	33.3977	5.65960	0.39997	166.688	15.2389	406.972	135.657
24	247.773	36.5960	5.03945	1.79980	166.682	27.0970	386.258	128.753
25	266.573	36.4964	0.00000	2.99970	166.683	39.6960	365.663	121.888
26	264.792	24.1993	1.19996	5.69983	166.695	39.6988	373.289	124.430
27	235.684	29.9979	0.99993	4.29970	166.688	39.6972	391.973	130.658
28	239.583	47.8966	0.49997	4.19971	166.688	39.6972	376.074	125.358
29	233.091	46.5981	2.29991	4.83981	166.693	39.6984	380.085	126.695
30	241.378	34.9969	5.15954	0.39996	166.685	39.6964	383.765	127.922
31	261.826	35.8036	3.40034	2.00020	177.818	26.3026	369.637	123.212
32	261.826	35.8036	3.40034	2.00020	177.818	26.3026	369.637	123.212
33	261.826	35.8036	3.40034	2.00020	177.818	26.3026	369.637	123.212
34	261.826	35.8036	3.40034	2.00020	177.818	26.3026	369.637	123.212
35	261.826	35.8036	3.40034	2.00020	177.818	26.3026	369.637	123.212

TABLEAU 3

Un extrait de la matrice des corrélations des variables explicatives (effets principaux et interactions)

EFFETS	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	−0.447	−0.171	0.113	0.627	−0.308	−0.724	−0.724
X_2	−0.447	1.000	0.305	−0.218	0.097	0.122	−0.069	−0.069
X_3	−0.171	0.305	1.000	−0.724	0.173	−0.242	0.056	0.056
X_4	0.113	−0.218	−0.724	1.000	−0.117	0.494	−0.236	−0.236
X_5	0.627	0.097	0.173	−0.117	1.000	−0.284	−0.768	−0.768
X_6	−0.308	0.122	−0.242	0.494	−0.284	1.000	−0.203	−0.203
X_7	−0.724	−0.069	0.056	−0.236	−0.768	−0.203	1.000	1.000
X_8	−0.724	−0.069	0.056	−0.236	−0.768	−0.203	1.000	1.000
$X_1 \cdot X_2$	−0.158	0.951	0.297	−0.217	0.342	0.024	−0.325	−0.325
$X_1 \cdot X_3$	−0.107	0.284	0.996	−0.714	0.229	−0.252	−0.004	−0.004
$X_1 \cdot X_4$	0.203	−0.265	−0.728	0.994	−0.064	0.453	−0.291	−0.291
$X_1 \cdot X_5$	0.932	−0.246	−0.029	0.018	0.866	−0.337	−0.816	−0.816
$X_1 \cdot X_6$	−0.170	0.053	−0.260	0.517	−0.191	0.986	−0.319	−0.319
$X_1 \cdot X_7$	0.755	−0.699	−0.189	−0.072	0.174	−0.644	−0.097	−0.097
$X_2 \cdot X_3$	−0.186	0.515	0.938	−0.669	0.181	−0.219	−0.010	−0.010
$X_2 \cdot X_4$	−0.051	0.024	−0.679	0.943	−0.122	0.540	−0.202	−0.202
$X_2 \cdot X_5$	−0.281	0.971	0.333	−0.240	0.325	0.048	−0.241	−0.242
$X_2 \cdot X_6$	−0.447	0.534	−0.123	0.334	−0.219	0.880	−0.186	−0.186
$X_2 \cdot X_7$	−0.606	0.973	0.308	−0.261	−0.081	0.081	0.156	0.156
$X_3 \cdot X_4$	−0.336	0.073	−0.013	0.200	−0.149	0.208	0.179	0.179
$X_3 \cdot X_5$	−0.126	0.306	0.997	−0.720	0.231	−0.251	0.002	0.002
$X_3 \cdot X_6$	−0.265	0.282	0.783	−0.558	0.038	0.222	−0.034	−0.034
$X_3 \cdot X_7$	−0.215	0.293	0.997	−0.731	0.123	−0.256	0.121	0.121
$X_4 \cdot X_5$	0.150	−0.218	−0.720	0.997	−0.063	0.479	−0.278	−0.278
$X_4 \cdot X_6$	−0.035	−0.102	−0.600	0.956	−0.153	0.664	−0.235	−0.235
$X_4 \cdot X_7$	0.072	−0.216	−0.729	0.997	−0.164	0.496	−0.187	−0.187
$X_5 \cdot X_6$	−0.244	0.132	−0.224	0.491	−0.176	0.992	−0.295	−0.295
$X_5 \cdot X_7$	−0.078	0.045	0.340	−0.523	0.407	−0.733	0.270	0.270
$X_6 \cdot X_7$	−0.395	0.122	−0.240	0.476	−0.378	0.992	−0.094	−0.094

TABLEAU 4
Facteurs d'inflation de la variance des estimateurs de β

EFFETS	Fact. d'inflation	EFFETS	Fact. d'inflation
X_1	248418	X_2X_3	5632
X_2	1846933	X_2X_4	44761
X_3	3503900	X_2X_5	293507
X_4	50862274	X_2X_6	14412
X_5	553971	X_2X_7	553249
X_6	2089490	X_3X_4	20
X_7	∞	X_3X_5	231844
X_8	∞	X_3X_6	3166
X_1X_2	41488	X_3X_7	811287
X_1X_3	265236	X_4X_5	2058135
X_1X_4	3680816	X_4X_6	113394
X_1X_5	1067414	X_4X_7	11355086
X_1X_6	72771	X_5X_6	91296
X_1X_7	33953	X_5X_7	80362
		X_6X_7	568144

3. Deuxième partie : l'analyse des résultats

Sept réponses ont été mesurées :

Y_1 : étalement à 5 minutes

Y_2 : étalement à 10 minutes

Y_3 : étalement à 15 minutes

Y_4 : étalement à 20 minutes

Y_5 : densité à 3 minutes

Y_6 : viscosité à 2 minutes

Y_7 : viscosité à 18 minutes

Leurs valeurs figurent dans le Tableau 5.

3.1. Les régressions linéaires simples

En vue de la comparaison ultérieure avec la régression multiple et la régression PLS, sont donnés dans le Tableau 6 les coefficients des régressions linéaires simples des sept réponses réduites en fonction des effets principaux et des interactions (exprimés en valeurs centrées-réduites). Seuls figurent les coefficients significativement différents de zéro, en général au seuil de 5% ou moins. Quelques coefficients ont été admis avec des seuils entre 5 et 6% pour éviter de «louper» les effets de moindre importance mais comportant malgré tout une information non négligeable. Nous rappelons en effet qu'ici le contexte est celui d'une recherche expérimentale où l'on

TABLEAU 5
Les valeurs des réponses à modéliser

Obs.	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	110	104	99	94	2.10	17600	26000
2	119	116	109	108	2.05	14800	22800
3	115	114	106	107	2.02	10800	12400
4	112	110	108	108	2.00	20800	29600
5	92	93	89	88	1.98	35200	48000
6	85	80	77	75	2.06	37600	48000
7	55	55	30	30	2.05	72000	84800
8	109	104	96	94	2.07	20000	28000
9	95	91	78	70	2.09	44800	45600
10	104	95	89	80	2.06	16800	32000
11	120	120	118	112	2.04	35012	17600
12	112	109	107	108	2.03	18800	24000
13	98	94	90	91	2.07	24000	41600
14	95	96	93	91	2.08	27200	42000
15	105	107	105	103	2.02	40000	44000
16	99	94	86	85	1.99	26400	46529
17	111	110	108	104	2.01	22000	56000
18	104	103	99	96	2.02	26400	32800
19	80	67	30	30	2.03	45600	64000
20	61	60	30	30	2.05	64000	84000
21	75	72	65	59	2.10	37600	50000
22	68	64	57	60	2.08	32000	48000
23	64	60	59	59	2.09	50000	59200
24	69	68	64	30	2.08	52800	60000
25	90	80	73	67	2.06	36000	59200
26	88	91	90	90	2.01	40000	40000
27	93	90	87	88	2.03	40000	52000
28	91	90	85	83	2.01	41600	56000
29	81	82	75	74	2.03	48000	56000
30	61	61	59	59	2.01	62400	88000
31	93	94	88	85	2.06	36800	48000
32	99	95	94	92	2.06	28000	51200
33	101	96	93	92	2.07	32000	41600
34	97	89	84	79	2.07	34000	49600
35	99	95	87	85	2.06	34400	40000

souhaite d'abord comprendre le phénomène. **Quelque soit le niveau de multicolinéarité, on peut affirmer que les signes de ces coefficients sont justes, ce qui nous autorisera à les prendre comme références ultérieures.**

TABLEAU 6
Coefficients significatifs (×1000) des régressions simples des réponses Y_k réduites en fonction des effets principaux et des interactions des variables explicatives (centrés-réduits)

$\widehat{\beta}$	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	+550	+508	+333	+317		−424	−459
4					−343		
5	+607	+581	+392	+400		−474	−503
6					−608	+370	+465
7	−486	−477					
8	−486	−477					
1.4					−322		
1.5	+635			+389		−494	−531
1.6					−617		+408
1.7	+339				+405	−332	−432
2.4					−353		
2.6					−481	+392	+520
3.6	−323	−338	−379	−357			+372
4.5					−342		
4.6					−494		
4.7					−335		
5.6					−624	+325	+425
5.7							−415
6.7					−587	+408	+497

3.2. Les régressions linéaires multiples

Si on cherche à établir les régressions linéaires multiples des sept réponses réduites en fonction de tous les effets centrés-réduits simultanément présents le logiciel SAS exclut de lui-même les variables X_7 et X_8 (coefficients de régression nuls). Le logiciel trouve en effet que X_7 et X_8 sont également combinaisons linéaires des autres variables (effets principaux et interactions). Les variables X_j restantes sont alors linéairement indépendantes. La régression sur celles-ci conduit aux coefficients de régression du Tableau 7.

On peut également prendre en compte uniquement les variables dont les coefficients étaient significatifs au paragraphe précédent, c'est l'objet du Tableau 8.

Il est clair, au vu de ces deux tableaux que les coefficients obtenus sont totalement incohérents avec ceux des régressions simples, **même le signe est faux**. La raison en est bien connue : la multicolinéarité est très importante comme le mettait en évidence les facteurs d'inflation du Tableau 4. De plus, dans une telle situation, aucune méthode de sélection de variable pas-à-pas ne donne des résultats corrects; nous l'avons vérifié avec les options «forward», «backward», et «stepwise» de SAS. En fait, il est **impossible** de garder toutes les variables explicatives si elles sont linéairement dépendantes en régression multiple ordinaire.

TABLEAU 7

Coefficients (x1000) des régressions multiples des réponses Y_k (réduites) en fonction des effets principaux (centrés-réduits) et des interactions (centrées-réduites) de toutes les variables explicatives

EFFETS	coef de Y_1	coef de Y_2	coef de Y_3	coef de Y_4	coef de Y_5	coef de Y_6	coef de Y_7
X_1	−19173	−26135	−27435	−27173	123778	9806	15713
X_2	16243	−25525	30881	119865	296289	−70464	79081
X_3	46089	−48274	−5239	−93338	−229957	−135647	−52697
X_4	−524770	−830955	−711395	−11548	1658649	−85871	−108269
X_5	−49919	−53580	−35562	−21360	213604	22897	11002
X_6	245364	172290	119719	42254	−349386	−154733	83614
X_1X_2	1535	452	−10648	−25584	−34639	4889	−8459
X_1X_3	4928	25314	269	1143	17816	30495	3894
X_1X_4	137259	215446	178819	−11306	−444375	21408	4631
X_1X_5	60250	69489	53062	37459	−291525	−25238	−24090
X_1X_6	−41582	−20352	−12860	−8358	22963	37856	−12269
X_1X_7	11925	13364	15444	17703	−44389	−6603	−2028
X_2X_3	−177	3982	−935	−711	−1825	7486	−2115
X_2X_4	14444	23489	19067	−1392	−46500	2866	−853
X_2X_5	−5659	16282	−344	−24191	−134317	22501	−31197
X_2X_6	−10238	−2985	−5985	−11928	−14371	13897	−8128
X_2X_7	4738	25335	−6235	−58844	−180293	33833	−38483
X_3X_4	1017	989	1161	751	−880	−883	9
X_3X_5	−21263	5883	11508	53161	67603	26955	21829
X_3X_6	−3700	593	−453	313	4732	6241	2486
X_3X_7	−21107	19813	−922	41051	120128	65869	26604
X_4X_5	113954	179078	163974	31341	−340249	11075	56571
X_4X_6	15211	33437	29471	−1932	−71270	10067	2012
X_4X_7	249979	389841	327989	−4592	−785044	41987	45776
X_5X_6	−43631	−36766	−28471	−1724	57935	15975	−30138
X_5X_7	29470	25575	15878	8450	−81201	−14999	1847
X_6X_7	−123173	−85831	−56071	−13177	183263	77871	−39068

3.3. La régression PLS

Nous renvoyons le lecteur à l'article de M. Tenenhaus [10] pour un exposé théorique complet sur la régression PLS. On se contentera ici de rappeler que cette méthode permet de modéliser la liaison entre un ensemble de variables réponses Y_k (le tableau Y) et un ensemble de variables explicatives X_j (le tableau X). Elle est basée sur des analyses en composantes principales de chaque tableau, sous la contrainte que les composantes principales des X_j soient aussi corrélées que possible aux composantes principales des Y_k. Il est alors possible d'exhiber des équations de régression PLS reliant chaque Y_k aux X_j. Celles-ci conduisent à exprimer les effets marginaux des X_j sur les Y_k et également à faire de la prédiction de réponse pour de

TABLEAU 8

Coefficients (×1000) des régressions multiples des réponses Y_k réduites en fonction des effets principaux (centrés-réduits) et des interactions (centrées-réduites) des variables explicatives jugées significatives lors des régressions simples

EFFETS	coef de Y_1	coef de Y_2	coef de Y_3	coef de Y_4	coef de Y_5	coef de Y_6	coef de Y_7
X_1	−9655	53	−51	1110		−1500	12690
X_4					−6940		
X_5	−1728	526	440	1332		26	−2331
X_6					−17853	7406	18183
X_7	−4405	−47					
X_1X_4					−585		
X_1X_5	4864			−1862		412	−7935
X_1X_6					8106		−3997
X_1X_7	4603				−280	1083	−4997
X_2X_4					−297		
X_2X_6					1338	593	259
X_3X_6	−320	−346	−410	−404			377
X_4X_5					5594		
X_4X_6					−1634		
X_4X_7					3621		
X_5X_6					−989	−2984	−3940
X_5X_7							3864
X_6X_7					9396	−4245	−10054

nouveaux individus. La régression PLS permet de conserver toutes les variables X_j, même si elles sont linéairement dépendantes.

3.3.1. Premier modèle PLS

On réalise les calculs avec le logiciel SIMCA [9] sur PC. Celui-ci propose de sélectionner le nombre de composantes PLS «significatives» par une méthode de validation croisée, basée sur le calcul bien connu du PRESS («prediction error sum of squares»), voir par exemple WOLD [12]. On aboutit ainsi à 2 composantes PLS qui résument 56% de la variance du tableau X et 34% de la variance du tableau Y. Les coefficients des équations de régression PLS établies à partir des réponses réduites et des variables explicatives centrées-réduites figurent dans le Tableau 9 : **les effets sont ainsi comparables entre eux, ce qui est un souhait fondamental du chimiste.**

Considérant uniquement les coefficients déclarés significatifs lors des régressions simples on observe que **le signe des coefficients de régression PLS est juste dans tous les cas et pour l'ensemble des sept réponses, et ceci malgré la singularité du problème**. On remarque aussi que les signes des coefficients peuvent être faux si

TABLEAU 9

Coefficients (x1000) des régressions PLS calculées à partir des effets principaux et interactions centrés-réduits

Les coefficients soulignés ont été jugés significatifs lors des régressions simples

$\widehat{\beta}^{PLS}$	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	+92	+88	+69	+66	+14	−76	−88
2	−27	−25	−20	−19	−15	+25	+30
3	−23	−23	−17	−17	+10	+16	+17
4	+29	+31	+21	+22	−37	−14	−13
5	+115	+113	+86	+83	−14	−88	−99
6	−45	−38	−35	−30	−66	+52	+66
7	−85	−86	−63	−63	+34	+60	+65
8	−85	−86	−63	−63	+34	+60	+65
1 · 2	+3	+4	+2	+3	−11	0	+1
1 · 3	−11	−11	−8	−8	+7	+7	+7
1 · 4	+31	+34	+23	+24	−32	−17	−17
1 · 5	+112	+109	+84	+81	+3	−90	−103
1 · 6	−34	−27	−26	−22	−65	+43	+55
1 · 7	+53	+46	+40	+36	+52	−55	−67
2 · 3	−26	−27	−20	−20	+11	+18	+20
2 · 4	+26	+29	+19	+20	−43	−10	−7
2 · 5	+1	+3	0	+1	−17	+3	+5
2 · 6	−60	−51	−45	−40	−56	+60	+73
2 · 7	−45	−43	−34	−32	−8	+38	+44
3 · 4	+41	+43	+31	+31	−32	−25	−26
3 · 5	−12	−13	−9	−9	+7	+8	+8
3 · 6	−75	−73	−57	−54	−4	+61	+70
3 · 7	−30	−30	−22	−22	+13	+21	+22
4 · 5	+32	+35	+24	+25	−36	−17	−16
4 · 6	+19	+24	+14	+16	−54	−2	+2
4 · 7	+27	+29	+19	+21	−36	−13	−11
5 · 6	−36	−29	−28	−24	−67	+45	+58
5 · 7	+53	+49	+40	+37	+27	−49	−59
6 · 7	−53	−46	−41	−36	−64	+58	+72

les effets ne sont pas significatifs mais que la présence de ces derniers dans le modèle PLS ne gêne pas le calcul des effets significatifs.

SIMCA fournit beaucoup d'autres «sorties», notamment des graphiques «parlants», que nous commenterons à l'occasion du modèle final au paragraphe suivant.

L'analyse peut sembler terminée. Néanmoins, on désire ici que les coefficients traduisent une certaine réalité de l'influence des effets. On peut donc se poser la question naturelle de savoir s'il est bien raisonnable de réfléchir sur les coefficients d'effets aussi fortement corrélés, tout au moins pour certains d'entre eux. Ainsi,

l'effet principal X_1 présente un coefficient de corrélation de 0,93 avec l'interaction $X_1 * X_5$ (voir Tableau 3). Les coefficients de régression PLS respectifs ($\times 1000$) pour Y_1 sont de $+92$ et $+112$ (voir Tableau 9); peut-on affirmer pour autant que le «poids» de l'interaction est plus grand sur les réponses étudiées que le «poids» de l'effet principal? Sûrement pas. En fait, on ne peut conclure qu'en supposant un effet non négligeable de X_1 **ou bien** de $X_1 * X_5$, ou bien encore un effet non négligeable de X_1 **et** un effet non négligeable $X_1 * X_5$. Il est donc nécessaire d'améliorer le modèle pour augmenter l'interprétabilité des coefficients. C'est l'objet du paragraphe suivant.

3.3.2. Modèle PLS final

En vue d'aller au-delà de la simple comparaison de la régression PLS avec les régressions simple et multiple, tentons maintenant d'améliorer le modèle PLS dans l'esprit évoqué plus haut. Un moyen possible est d'essayer de diminuer les corrélations entre les effets avant de calculer un modèle de régression PLS. Pour ce faire, il est bien connu que l'on peut pratiquer un centrage-réduction des X_j avant de construire leurs interactions $X_j * X_k$.

TABLEAU 10
Ecart-types des effets principaux centrés-réduits et des interactions construites après centrage-réduction de ceux-ci

EFFET	EC-TYPE	EFFET	EC-TYPE
X_1	1	X_2X_4	1.09
X_2	1	X_2X_5	0.83
X_3	1	X_2X_6	1.13
X_4	1	X_2X_7	0.83
X_5	1	X_3X_4	0.99
X_6	1	X_3X_5	0.85
X_7	1	X_3X_6	1.08
X_8	1	X_3X_7	0.88
X_1X_2	1.07	X_4X_5	0.87
X_1X_3	0.98	X_4X_6	0.90
X_1X_4	0.90	X_4X_7	1.12
X_1X_5	0.75	X_5X_6	1.03
X_1X_6	1.19	X_5X_7	0.80
X_1X_7	0.92	X_6X_7	0.96
X_2X_3	1.43		

Les corrélations diminuent fortement, voir Tableau 11. Par exemple, la corrélation entre X_1 et $X_1 * X_5$ ne vaut plus maintenant que $-0{,}53$ (à comparer à 0,932 précédemment).Une nouvelle recherche d'effets significatifs par régression simple conduit cette fois à retenir un ensemble d'effets significatifs très différent (Tableau 12) du premier (Tableau 6). Ainsi, l'interaction $X_1 * X_5$ n'est plus sélectionnée.

TABLEAU 11

Un extrait de la matrice des corrélations des effets principaux et interactions (celles-ci étant construites après centrage-réduction des effets principaux)

EFFETS	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	−0.447	−0.171	0.113	0.627	−0.308	−0.724	−0.724
X_2	−0.447	1.000	0.305	−0.218	0.097	0.122	−0.069	−0.069
X_3	−0.171	0.305	1.000	−0.724	0.173	−0.242	0.056	0.056
X_4	0.113	−0.218	−0.724	1.000	−0.117	0.494	−0.236	−0.236
X_5	0.627	0.097	0.173	−0.117	1.000	−0.284	−0.768	−0.768
X_6	−0.308	0.122	−0.242	0.494	−0.284	1.000	−0.203	−0.203
X_7	−0.724	−0.069	0.056	−0.236	−0.768	−0.203	1.000	1.000
X_8	−0.724	−0.069	0.056	−0.236	−0.768	−0.203	1.000	1.000
X_1X_2	−0.444	0.194	0.294	−0.224	0.031	0.046	0.244	0.244
X_1X_3	−0.358	0.322	0.391	−0.162	0.062	0.190	0.037	0.037
X_1X_4	0.159	−0.267	−0.177	0.217	−0.028	−0.016	−0.023	−0.023
X_1X_5	−0.526	0.044	0.082	−0.034	−0.350	−0.134	0.595	0.595
X_1X_6	−0.282	0.042	0.156	−0.012	−0.084	0.315	0.070	0.070
X_1X_7	0.436	0.284	0.040	−0.023	0.486	0.091	−0.664	−0.664
X_2X_3	0.220	0.156	0.106	−0.037	0.051	−0.161	−0.170	−0.170
X_2X_4	−0.220	−0.052	−0.048	−0.133	−0.113	−0.009	0.254	0.254
X_2X_5	0.040	−0.019	0.089	−0.148	0.133	−0.090	−0.026	−0.026
X_2X_6	0.044	0.031	−0.203	−0.008	−0.066	−0.183	0.093	0.093
X_2X_7	0.315	−0.399	−0.292	0.334	−0.026	0.127	−0.176	−0.176
X_3X_4	−0.161	−0.053	−0.336	−0.220	−0.169	−0.176	0.349	0.349
X_3X_5	0.072	0.086	0.336	−0.195	0.045	0.039	−0.138	−0.138
X_3X_6	0.171	−0.212	−0.254	−0.160	0.031	−0.159	0.038	0.038
X_3X_7	0.041	−0.275	−0.440	0.391	−0.134	0.047	0.090	0.090
X_4X_5	−0.029	−0.141	−0.192	0.160	−0.163	0.081	0.094	0.094
X_4X_6	−0.016	−0.011	−0.194	0.566	0.078	0.022	−0.054	−0.054
X_4X_7	−0.019	0.248	0.309	−0.422	0.073	−0.043	−0.063	−0.063
X_5X_6	−0.097	−0.072	0.032	0.067	−0.102	0.196	0.031	0.031
X_5X_7	0.554	−0.027	−0.146	0.102	0.219	0.040	−0.526	−0.526
X_6X_7	0.087	0.109	0.043	−0.050	0.033	−0.154	−0.038	−0.038

On ne conservera que ces effets significatifs pour le calcul du second modèle PLS. Malgré tout, pour conserver une certaine réalité chimique, tous les effets principaux seront présents car ceux-ci sont les proportions de constituants obligatoires lors de la formulation du produit. Le chimiste aurait en effet des difficultés à comprendre l'absence dans le modèle de certains constituants de sa formulation qui ne peut exister que si ces derniers sont tous présents ! Naturellement, ceci n'interdit pas de lui préciser ceux qui semblent significativement influer sur ses réponses.

En conséquence, on traite un tableau X modifié, comprenant tous les effets principaux, et les interactions jugées significatives lors des régressions simples soit : $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_1 * X_3, X_1 * X_7, X_4 * X_5, X_5 * X_6, X_5 * X_7, X_6 * X_7$.

TABLEAU 12
Coefficients significatifs (x1000) des régressions simples des réponses (centrées-réduites) en fonction des effets principaux et interactions centrés-réduits construits à partir des variables déjà centrées-réduites

EFFETS	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
X_1	550	508	333			−424	−459
X_4					−344		
X_5	607	581	392	400		−474	−503
X_6					−608	370	465
X_7	−486	−477					
X_8	−486	−477					
X_1X_3					−342		
X_1X_7	432	432					
X_4X_5	−411	−399					
X_5X_6	−316	−322	−330				
X_5X_7	442	428					
X_6X_7			346				

SIMCA explique maintenant 44% de la variance du tableau X et 36% de la variance du tableau Y en 2 composantes PLS, sélectionnées par validation croisée. On trouve Tableau 13 les nouveaux coefficients des équations de régression PLS obtenus à partir des variables brutes. Leurs écart-types sont suffisamment proches de 1 (Tableau 10) pour que les coefficients soient comparables entre eux. Les signes sont encore justes et les coefficients traduisent mieux la réalité des effets. Toutefois, n'oublions pas que la connaissance précise et indépendante des effets aurait exigé la réalisation d'un plan d'expériences orthogonal (impossible ici). L'effet X_7 reste indiscernable de X_8.

Pour aider à l'interprétation chimique, on peut proposer un classement des effets significatifs, dans l'ordre décroissant et en valeur absolue, pour les réponses d'intérêt principal pour le chercheur soit Y_1, Y_5, Y_6, chacune étant de nature différente (étalement, densité, viscosité).

Classement des effets pour la réponse Y_1 :

$$X_5 > X_1 > X_4 * X_5, X_5 * X_6 > X_7, X_8 > X_1 * X_7, X_5 * X_7.$$

Classement des effets pour la réponse Y_5 :

$$X_6 > X_4 > X_1 * X_3.$$

Classement des effets pour la réponse Y_6 :

$$X_6 > X_5, X_1.$$

TABLEAU 13
Coefficients ($\times 1000$) *des équations de régression du modèle PLS final (effets centrés-réduits). Seuls figurent les coefficients sélectionnés par régressions simples*

$\widehat{\beta}^{PLS}$	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	+172	+157	+132			−150	−185
4					−67		
5	+189	+174	+141	+130		−156	−189
6					−371	+242	+328
7	−87	−92					
8	−87	−92					
1 · 3					−36		
1 · 7	+68	+71					
4 · 5	−105	−100					
5 · 6	−94	−90	−67				
5 · 7	+58	+61					
6 · 7			+64				

Examinons les graphiques fournis par SIMCA. Les figures 1 et 2 donnent une idée de la qualité prédictive du modèle PLS. Les expériences sont représentées dans les plans reliant terme à terme les composantes PLS u_1 et u_2 pour Y aux composantes PLS t_1 et t_2 pour X. Plus elles sont proches de la première bissectrice, pour chaque plan, plus le modèle donne des bonnes prédictions pour les réponses. Les prédictions ne sont pas ici de qualité exceptionnelle mais suffisante, relativement à l'importante variabilité expérimentale et à la complexité du problème posé. On peut remarquer de meilleures prédictions dans la partie supérieure des droites que dans leur partie inférieure.

Les autres graphiques qui méritent attention sont ceux des figures 3 et 4, équivalents aux cartes des individus en ACP mais orientés vers l'explication de Y par X ; ils permettent de repérer notamment les groupes d'expériences (observations) et les expériences séparées de ceux-ci.

Ainsi, la figure 3 montre que :

– les expériences 22 et 23 se distinguent des autres si on considère les variables explicatives (Tableau 2)

– le petit groupe [1,2,8,9,10,11,12] se sépare nettement des autres,

– le petit groupe [14,31,32,33,34,35] est au centre du graphique, ce qui est naturel puisque les expériences 31 à 35 représentent des formulations «moyennes», voir Tableau 2.

La figure 4 permet de comparer les ressemblances et dissemblances des expériences mais en considérant les réponses.

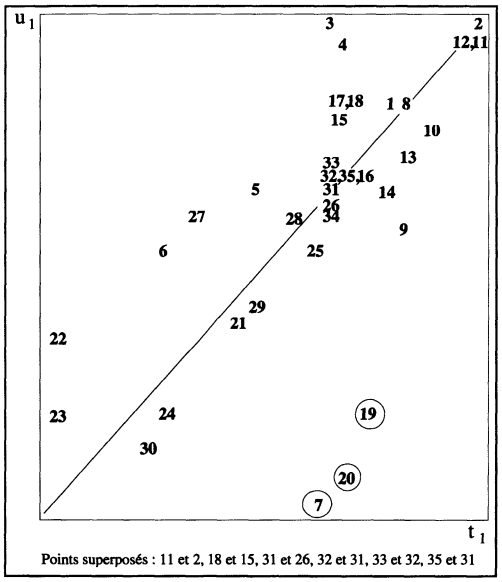

FIGURE 1
Observations dans le plan des composantes PLS t_1 *et* u_1

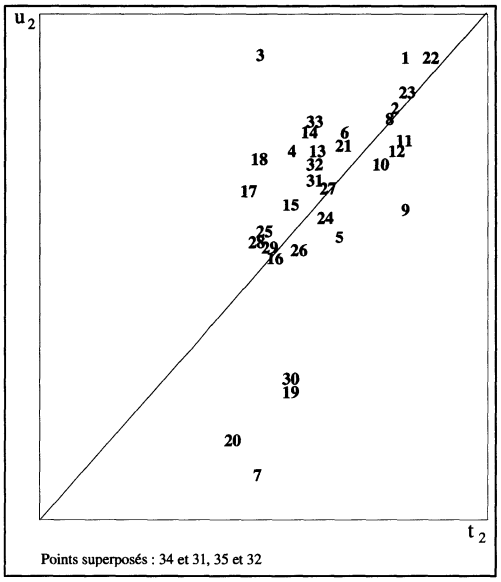

FIGURE 2
Observations dans le plan des composantes PLS t_2 *et* u_2

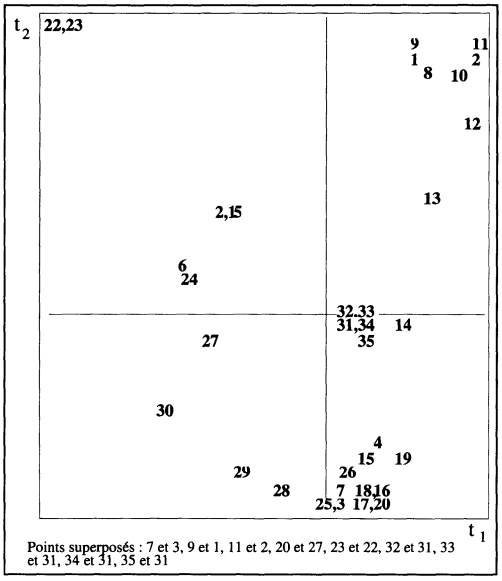

FIGURE 3
Observations dans le plan des composantes PLS t_1 et t_2

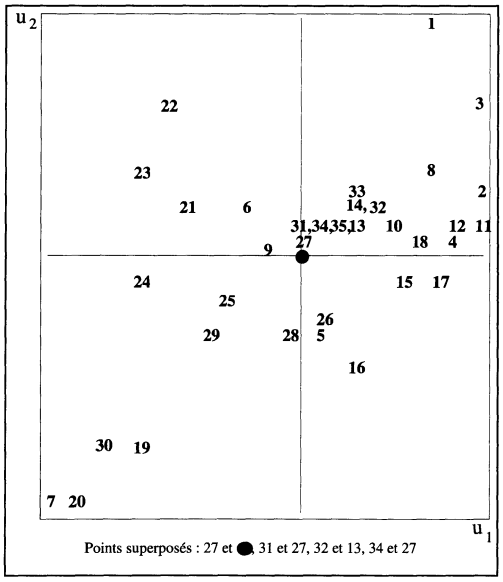

FIGURE 4
Observations dans le plan des composantes PLS u_1 et u_2

La figure 5 présente un graphique particulièrement synthétique. Il est construit à partir des termes suivants :

– w_1 = vecteur des coefficients des variables explicatives dans la construction de t_1,

– w_2 = vecteur des coefficients des variables explicatives résiduelles dans la construction de t_2,

– r_1 = le vecteur des coefficients de régression des réponses sur t_1,

– r_2 = le vecteur des coefficients de régression des réponses résiduelles sur t_2.

Ce graphique $\{(w_1, r_1), (w_2, r_2)\}$ traduit les corrélations des variables explicatives entre elles, des réponses entre elles et des variables explicatives avec les réponses. Les remarques qui s'imposent sont :

– les 4 réponses Y_1, Y_2, Y_3 et Y_4 sont très corrélées entre elles, ce qui est assez naturel d'un point de vue chimique puisqu'elles mesurent toutes un phénomène d'étalement à différents temps,

– même type de remarque pour Y_6 et Y_7 qui mesurent une viscosité du produit à 2 temps différents,

– la réponse Y_5, mesurant la densité, n'est pas du tout corrélée aux autres,

– chaque réponse est graphiquement proche des effets qui l'influencent.

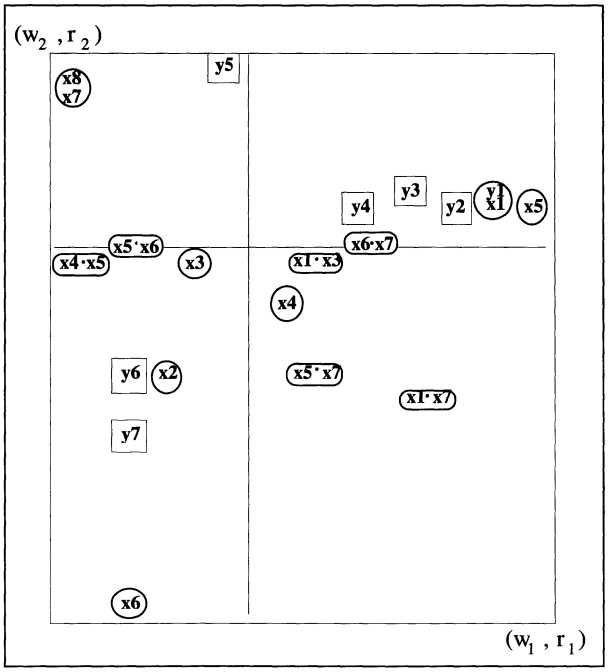

FIGURE 5
Graphique des variables pour le modèle PLS final

Enfin, on peut résumer la reconstitution des données obtenue avec ce modèle PLS en traçant les graphiques des valeurs calculées en fonction des valeurs observées. C'est l'objet, pour les réponses Y_1, Y_5 et Y_6, des figures 6,7 et 8. On peut dire que la réponse $Y1$ est assez bien reconstituée, tandis que les réponses Y_5 et Y_6 le sont moins bien. Cette information, transmise au chercheur, a permis à ce dernier de tenter d'améliorer la qualité de mesurage des réponses Y_5 et Y_6 et d'accorder plus de confiance dans la modélisation de la réponse Y_1.

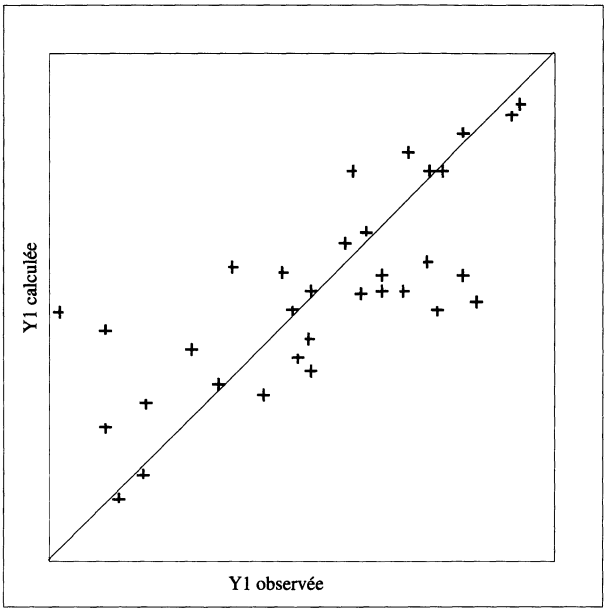

FIGURE 6
Graphique [Y_1 observée, Y_1 calculée]

Pour terminer, on peut signaler que l'on a établi également des modèles PLS avec des tableaux Y partiels, par exemple le tableau Y réduit aux réponses de même nature Y_1, Y_2, Y_3 et Y_4, toutes relatives à la propriété d'étalement. Les coefficients ainsi obtenus sont très proches de ceux du modèle final mais la reconstitution des données est moins bonne. On peut expliquer sommairement ce phénomène par la présence de corrélations non négligeables entre les réponses du tableau Y, ce qui améliore l'information d'une réponse particulière. On peut rapprocher ce comportement de la pratique courante qui consiste à estimer une valeur manquante d'une variable aléatoire donnée à partir d'une autre variable aléatoire quand celles-ci sont fortement corrélées.

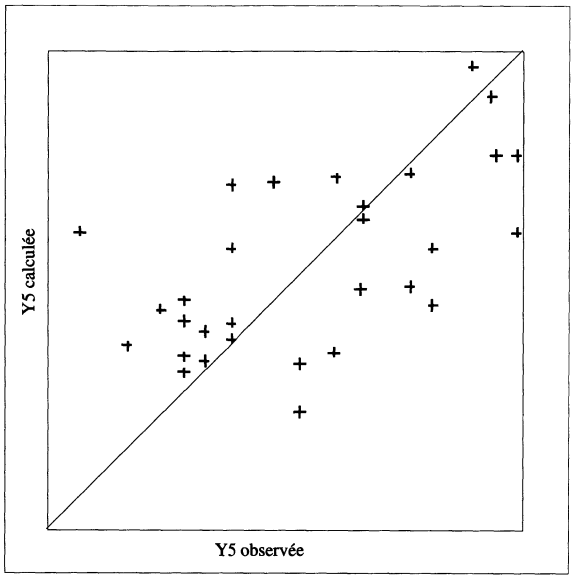

FIGURE 7
Graphique [Y_5 observée, Y_5 calculée]

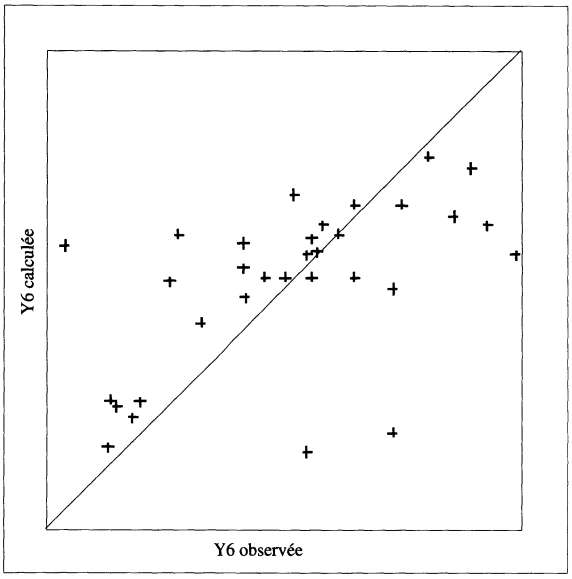

FIGURE 8
Graphique [Y_6 observée, Y_6 calculée]

4. Conclusion

Cet exemple réel nous permet de vérifier l'incapacité de la régression linéaire multiple (moindres carrés ordinaires) pour modéliser un tel problème. En revanche, la régression PLS fournit un modèle raisonable rendant possible une première conclusion sur l'influence des effets. De plus, toutes les réponses peuvent être analysées en même temps et les graphiques sont faciles à interpréter. Toutefois, cet exemple montre aussi le danger d'une utilisation non réfléchie de la méthode, notamment pour des tableaux X très multicolinéaires et même singuliers, phénomène fréquent en Chimie de Formulation. Dans ce cadre, on peut proposer la méthodologie suivante :

1 – faire un centrage-réduction des proportions (effets principaux),

2 – construire les interactions (et éventuellement les effets carrés) à partir de ces proportions centrées- réduites,

3 – établir la matrice des corrélations linéaires entre tous ces effets,

4 – déceler les effets significatifs par régression simple de toutes les réponses en fonction de tous les effets (très rapide avec les logiciels actuels),

5 – calculer un modèle PLS en imposant tous les effets principaux (toutes les proportions doivent être présentes pour constituer une formulation) et en sélectionnant les interactions (et éventuellement les carrés) décelées lors de l'étape précédente,

6 – réfléchir sur les coefficients (par exemple en les comparant entre eux) en examinant simultanément la matrice des corrélations,

7 – étudier la répartition des observations dans les plans (t_i, u_i), les plans (t_i, t_j) et les plans (u_i, u_j),

8 – examiner le graphique $\{(w_1, r_1),\ (w_2, r_2)\}$,

9 – vérifier la qualité de la reconstitution des données en affichant les graphiques réponse observée, réponse calculée.

On peut signaler que l'étape 4 est une étape particulièrement cruciale pour deux raisons :

– on a pu vérifier que les signes des coefficients d'effets non significatifs sont faux si ceux-ci sont présents dans le modèle,

– par simulation on a mis en évidence une sensibilité importante de la méthode à des facteurs de bruit dans le tableau X ; la part d'explication du tableau Y peut chuter alors de façon anormale, même si plusieurs effets fortement explicatifs existent dans le tableau X, voir par exemple CLARK *et al* [3].

Enfin, il faut souligner un effet pervers, évident pour un statisticien, mais pas forcément pour un chimiste. Comme la méthode arrive toujours à calculer un jeu de coefficients, ce dernier peut oublier (naïvement) de réaliser ses expériences selon une bonne planification; il peut être tenté de réaliser ses expériences au hasard, sachant que la régression PLS lui donnera malgré tout un modèle! Nous lui rappellerons alors qu'aucune méthode, aussi excellente soit-elle, ne séparera des effets qui auront varié exactement de la même manière.

En résumé, la méthode de régression PLS s'avère être une méthode efficace en Chimie de Formulation si on applique la méthodologie préconisée plus haut.

Remerciements

Nous tenons à remercier tout particulièrement le Professeur Michel Tenenhaus qui nous a initiés à cette méthode de régression et a corrigé le premier jet de cet article. Nous remercions également Messieurs Charrin et Colombet, ingénieurs chimistes au Centre de Recherches d'Aubervilliers, qui nous ont fait confiance en nous soumettant ce problème réel passionnant, ainsi que Monsieur Chagnon, ingénieur au Service Calcul Scientifique, qui a collaboré à la construction du plan d'expériences.

5. Bibliographie

[1] BOX G.E.P., HUNTER W.G., HUNTER J.S. (1978) : Statistics for experimenters, Wiley.

[2] CANDI (1987) : Programme informatique Rhône-poulenc, J.-P. Gauchi.

[3] CLARK M., CRAMER III R.D. (1993) : The probability of chance correlation using partial least squares (PLS), QSAR, 12, 137-145.

[4] CORNELL J.A. (1981) : Experiments with mixtures, Wiley.

[5] FEDOROV V.V. (1972) : Theory of optimal experiments, Academic Press.

[6] MATHIEU D. (1981) : Contribution de la méthodologie de la recherche expérimentale à l'étude des relations structure-activité. Thèse d'état, Université Aix-Marseille.

[7] SAS/QC : Logiciel de la Société SAS Institute, Cary, USA, Version 6.04.

[8] SCHEFFE H. (1958) : Experiments with mixtures, JRSS, B, vol. 20, p. 344-360.

[9] SIMCA : Logiciel de la Société Umetri, Umea, Suède, Version 4.4

[10] TENENHAUS M. (1993) : La régression PLS. Revue MAD (revue interne Rhône-Poulenc), n°5, p. 5-20.

[11] TOMASSONE R., LESQUOY E., MILLIER C. (1983) : La régression, nouveaux regards sur une ancienne méthode statistique, Masson.

[12] WOLD S. (1978) : Cross-validatory estimation of the number of components in factor and principal components models. Technometrics, vol. 20, n°4, p. 397-405.