CIMASI' 2000 23-25 Octobre EHTP Casablanca Maroc
Conférence Workshop2

PLANOR : LOGICIEL DE GENERATION AUTOMATIQUE DE PLANS REGULIERS

A. Kobilinsky
Biométrie, BIA – INRA Jouy,
78352 JOUY-EN-JOSAS Cedex
kobi@banian.jouy.inra.fr

1 Introduction

Initialement conçu à la demande d'un groupement de producteurs laitiers (ARILAIT), le logiciel PLANOR pour Windows permet la construction de plans d'expériences fractionnaires avec un ou plusieurs systèmes de blocs. Ces plans, dits *réguliers*, sont définis par des règles algébriques simples et ont des propriétés d'orthogonalité d'où dérivent des propriétés d'optimalité et de robustesse tout à fait remarquables [6]. C'est pourquoi ils sont utilisés avec succès dans de très nombreux cas.

On trouve d'excellents catalogues de tels plans [11] et nombreux sont aujourd'hui les logiciels statistiques qui donnent de tels catalogues sous forme informatisée. Ces catalogues sont cependant difficilement utilisables lorsque la connaissance que l'on a des phénomènes amène une certaine dissymétrie entre les facteurs étudiés. Les règles algébriques de construction doivent alors être adaptées à la situation spécifique créée, ce qui ne peut se faire que par une recherche algorithmique du type de celle utilisée dans PLANOR.

Franklin et Bailey [4] et Franklin [5] décrivent un tel algorithme de recherche. L'algorithme quelque peu différent utilisé par PLANOR est précisé, au travers de petits exemples, dans [10] et décrit de façon plus générale dans [7].

Pour définir la recherche à effectuer, l'utilisateur précise, dans les cas les plus simples, d'une part le modèle factoriel qui donne l'espérance de la réponse en fonction des niveaux des facteurs étudiés, d'autre part la partie de ce modèle qu'il veut estimer. Cas particulier bien connu : le modèle inclue toutes les interactions de deux facteurs mais la partie à estimer est réduite à l'ensemble des effets principaux. Les plans correspondants sont dits de résolution IV. Il ne permettent pas d'estimer les interactions de 2 facteurs, confondues entre elles, mais donnent des estimations non biaisées des effets principaux même en présence de ces interactions. Pour cette raison, ces plans doivent être préférés chaque fois que cela est possible aux plans dits de Plackett et Burman dérivés des matrices d'Hadamard, prévus pour estimer les effets principaux uniquement dans le cadre d'un modèle ne comportant que ces effets principaux.

Il est souvent possible d'obtenir des plans de résolution IV permettant d'estimer, outre les effets principaux, un certain nombre d'interactions de deux facteurs. C'est pour obtenir ce type de plan qu'il est indispensable d'avoir recours à un algorithme tel que celui de PLANOR.

Dans de nombreux cas, la spécification d'une unique paire "*modèle – partie à estimer*" est cependant insuffisante pour spécifier la recherche. Par exemple, si les unités expérimentales sont réparties en un ou plusieurs sytèmes de blocs, la confusion de certains effets factoriels avec les blocs peut être souhaitable ou même indispensable. Ainsi, dans un procédé industriel comportant une cuisson, les unités placées en même temps dans le même four constitue un bloc. Ces unités sont nécessairement cuites à une même température et l'effet principal du facteur température se confond avec des différences entre blocs. Le plan en *split-plot* qui confond également un effet principal avec les blocs, est bien connu en agronomie. Les plans en *confounding* confondent avec les blocs des interactions jugées secondaires pour permettre d'estimer *intra-bloc*, donc avec le maximum de précision possible, les effets jugés importants.

C'est pourquoi PLANOR permet d'introduire plusieurs paires "*modèle – partie à estimer*". Ceci va de pair avec la possibilité également offerte d'introduire des *hiérarchies* entre facteurs, permettant par exemple de contraindre un facteur tel que la température à rester constant pour toutes les unités d'un même bloc.

Une autre caractéristique de PLANOR est de gérer automatiquement des facteurs ayant des nombres de niveaux distincts. Ceci est particulièrement intéressant lorsque les facteurs ont des nombres de niveaux puissances d'un même premier, par exemple pour étudier des mélanges de facteurs à 2 et 4 niveaux.

Après avoir présenté la méthode de construction et l'algorithme sur des exemples simples de plans pour facteurs à 2 niveaux, on illustrera la façon d'aborder les mélanges de facteurs à 2 et 4 niveaux, ou 2, 3, 4 et 6 niveaux. On montrera ensuite comment l'introduction de plusieurs paires *modèle – partie à estimer* permet de répartir efficacement en bloc et sous-blocs une fraction de plan 2^{11-5} pour 11 facteurs à 2 niveaux et 64 unités.

2 Méthode de construction dans le cas de facteurs à 2 niveaux

Considérons le plan pour 5 facteurs A, B, C, D, E à 2 niveaux, 8 unités expérimentales, défini à gauche du tableau 1. Les niveaux des facteurs étant codés -1 et 1, ce plan est formé en écrivant d'abord toutes les combinaisons de niveaux des facteurs A, B, C, puis en déduisant les niveaux de D, E par les règles $D = AB$, $E = AC$. On dit que A, B, C sont utilisés comme *facteurs de base* et que D, E sont des *facteurs définis* (ou *dérivés* ou encore *ajoutés*).

Les règles de définition peuvent être résumées par la matrice à droite du tableau 1 donnant les exposants de A, B, C dans les produits définissant D et E. La tranposée de cette matrice est habituellement appelée *matrice clé* [12].

Compte tenu de ce que les facteurs ont les niveaux -1, 1, les égalités $D = AB$, $E = AC$ sont équivalentes aux relations $ABD = 1$, $ACE = 1$. Puisque $A^2 = 1$, le produit de ces égalités donne l'égalité supplémentaire $BCDE = 1$. On a donc

$$1 = ABD = ACE = BCDE. \tag{1}$$

facteurs de base			facteurs définis	
A	B	C	$D=AB$	$E=AC$
-1	-1	-1	1	1
-1	-1	1	1	-1
-1	1	-1	-1	1
-1	1	1	-1	-1
1	-1	-1	-1	-1
1	-1	1	-1	1
1	1	-1	1	-1
1	1	1	1	1

Matrice clé

	D	E
A	1	1
B	1	0
C	0	1

TAB. 1 – *Exemple de fraction régulière* 2^{5-2}

Ces produits de facteurs égaux à 1 sont appelés *produits, mots ou contrastes de définition* de la fraction. En multipliant (1) par les produits A, B, C, AB, AC, BC, ABC entre les facteurs de base, on en déduit les 7 jeux d'égalités suivants :

$$\begin{array}{llll} A & = BD & = CE & = ABCDE, \\ B & = AD & = ABCE & = CDE, \\ C & = ABCD & = AE & = BDE, \\ AB & = D & = BCE & = ACDE, \\ AC & = BCD & = E & = ABDE, \\ BC & = ACD & = ABE & = DE, \\ ABC & = CD & = BE & = ADE, \end{array} \tag{2}$$

Il s'avère [9] que les effets factoriels associés à des produits égaux sont confondus : dans le cas présent on ne peut estimer que la somme de ces effets. Ainsi la première égalité dans (2) montre qu'on ne peut estimer à partir des 8 observations de la réponse que la somme $A+BD+CE+ABCDE$ de l'effet principal A, des interactions de 2 facteurs BD, CE et de l'interaction des 5 facteurs $ABCDE$. Si les interactions sont toutes supposées nulles, le plan permet donc d'estimer l'effet principal A.

Il est ainsi facile de trouver les propriétés du plan à partir des règles le définissant. Mais en pratique, on doit procéder en sens inverse. Le problème est de trouver des règles de définition à partir d'un objectif donné. On se fixe généralement au départ un objectif modeste, impliquant peu de contraintes. Si la recherche trouve des plans ad hoc, soit on étudie les propriétés de ces plans pour choisir le plus adapté, soit on affine la recherche en adoptant un objectif plus ambitieux.

Reprenons l'exemple avec les 5 facteurs A, B, C, D, E à 2 niveaux et 8 unités. Un premier objectif peut être d'estimer les 5 effets principaux dans le cadre du modèle additif $A+B+C+D+E$ où toutes les interactions sont nulles. Le plan du tableau 1 répond à cet objectif. On peut se fixer le second objectif plus ambitieux d'estimer ces mêmes effets dans le modèle incluant les interactions de deux facteurs

$$A+B+C+D+E+AB+AC+AD+AE+BC+BD+BE+CD+CE+DE \tag{3}$$

mais on montre que cet objectif ne peut être atteint. Faute de pouvoir introduire toutes les interactions de deux facteurs, on peut alors se fixer le troisième objectif d'introduire un maximum d'interactions pour prendre en compte celles qui sont le plus susceptibles d'exister.

Dans cet exemple élémentaire, il est facile de prouver qu'à une permutation des facteurs près le plan du tableau 1 est l'unique plan répondant au premier objectif. Les égalités (2) montrent alors que ce plan permet d'estimer deux interactions telles que BC, CD si les autres interactions sont supposées nulles. Il montre aussi que les effets principaux restent estimables sans biais dans un modèle incluant les 4 interactions BC, DE, CD, BE.

Lorsqu'il y a d'avantage de facteurs, il est difficile voire impossible de trouver sans programme un plan, ou les plans, répondant aux objectifs fixés. Un algorithme de recherche devient indispensable.

Si tous les facteurs ont 2 niveaux, l'algorithme de Draper et Mitchell [2], [10] permet de trouver l'ensemble des solutions aux permutations près. Cet algorithme, qui ne semble malheureusement pas implémenté aujourd'hui dans les logiciels statistiques, cherche l'ensemble des solutions en travaillant par récurrence sur le nombre des facteurs introduits. Son efficacité vient de ce qu'il élimine à chaque étape les solutions équivalentes à une solution déjà trouvée en ce sens qu'elle s'en déduisent par permutation des facteurs.

L'algorithme de PLANOR est plus rustique en ce sens qu'il ne comporte pas cette phase d'élimination des solutions équivalentes. Mais il fonctionne aussi dans des cas asymétriques ou les facteurs ont des nombres de niveaux différents. Pour préciser son fonctionnement, reprenons l'exemple simple précédent en imaginant qu'on ignore qu'il n'existe qu'une seule solution. Cherchons s'il est possible d'estimer l'ensemble $\mathcal{E} = \{A,B,C,D,E\}$ des effets principaux dans le modèle ayant $\mathcal{M} = \{A,B,C,D,E,AB,CE\}$ comme ensemble de termes. Ce dernier ensemble $\mathcal{M}$ est parfois appelé l'ensemble des termes *non négligeables*, et $\mathcal{E}$ l'ensemble des *termes requis* (requirement set).

Aucun des termes de $\mathcal{E}$ ne doit se confondre avec un terme de $\mathcal{M}$, donc aucun produit entre un terme de $\mathcal{E}$ et de $\mathcal{M}$ ne doit être égal à 1. L'ensemble $\mathcal{N} = \mathcal{E}\mathcal{M}$ de ces produits qui doivent être distincts de 1 est appelé *ensemble inéligible*. On a ici :

$$\mathcal{N} = \{ \underbrace{\mathbf{1},A,\ldots,DE}_{\text{nb. lettres} \leq 2} ,ABC,ABD,ABE,ACE,BCE,CDE\}$$

Pour trouver un plan, on cherche successivement des règles de définition de D, puis E, telles qu'aucun des produits inéligibles ne soit égal à 1. Si la procédure échoue pour E, elle revient en arrière en cherchant un autre choix pour D, d'où le nom de rétrorécurrence (backtrack search) utilisé pour désigner ce type de procédure.

En l'occurrence, si $P = A^a B^b C^c$ est le produit définissant D, on a $DP = 1$.Donc DP ne doit pas appartenir à $\mathcal{N}$ et P ne doit pas appartenir à $D\mathcal{N}$. Comme les termes de $D\mathcal{N} =$ ou n'apparaissent que A, B, C sont $A = D \times AD$, $B = DBD$, $C = DCD$, $AB = D \times ABD$, le produit P définissant D doit être choisi différent de A,B,C,AB.

Sélectionnons $D = ABC$. Le produit Q des lettres A, B, C définissant E doit alors être sélectionné de façon à ce que EQ et $EQABCD$ n'appartiennent pas à $\mathcal{N}$. Ce produit Q doit donc être distinct des produits dans $E\mathcal{N}$ et $EABCD\mathcal{N}$. Les produits de $E\mathcal{N}$ et $EABCD\mathcal{N}$ où n'apparaissent que A, B, C sont respectivement A, B, C, AB, AC, BC et $ABC = EABCD \times DE$, $AB = EABCD \times CDE$. On doit donc sélectionner pour E un produit différent de A, B, C, AB, AC, BC, ABC ce qui n'est pas possible.

La recherche revient donc sur D et sélectionne un des deux autres choix possibles $D = AC$ ou $D = BC$. Supposons que l'on retienne $D = AC$. Alors le produit Q définissant E doit être tel que EQ et $EQACD$ ne soient pas dans $\mathcal{N}$, donc il ne doit appartenir ni à $E\mathcal{N}$ ni à $EACD\mathcal{N}$. Cela exclue d'une part comme auparavant les produits A, B, C, AB, AC, BC, d'autre part les produits $AC = EACDDE$ et $A = EACDCDE$. Il y donc un unique choix possible $E = ABC$.

Si l'on cherche une unique solution, la recherche peut s'arrêter là. Mais si l'on veut l'ensemble des solutions, on continue la recherche en essayant l'autre choix $D = BC$. On trouve également $E = ABC$. Il est cependant clair dans ce cas que la permutation échangeant les facteurs A et B, qui laisse $\mathcal{N}$ invariant, fait passer de l'une des solutions à l'autre. En fait, cette permutation transforme $D = BC$ en $D = AC$. Ces deux choix sont donc équivalents et la recherche avec $D = BC$ s'avère inutile.

PLANOR n'inclue pas à ce jour de module de détection d'équivalence entre choix. Il s'en suit que la recherche de l'ensemble des solutions ne peut se faire que dans les problèmes de très petite taille et que s'il y a des solutions, le résultat obtenu contient beaucoup de solutions équivalentes. D'autre part, hors des problèmes de petite taille, il est pratiquement impossible, car trop long, d'explorer tous les choix possibles pour conclure à l'inexistence d'une solution.

Mais PLANOR donne une possibilité intéressante qui compense partiellement cette absence d'un module permettant de reconnaître les choix équivalents. Il permet de chercher plusieurs solutions en modifiant de façon aléatoire à chaque recherche l'ordre dans lequel sont explorés les différents choix. L'examen des mots de définition de ces solutions et notamment de leurs longueurs permet alors de se faire une idée souvent correcte de ce que sont les différents types de solution.

3 Facteurs avec des nombres de niveaux différents

Quand on construit un plan fractionnaire avec des facteurs qui ont des nombres de niveaux différents, on dit qu'il s'agit d'un plan asymétrique. Lorsque ces nombres de niveaux sont tous puissances d'un même nombre premier, on se ramène facilement au cas symétrique en introduisant des pseudofacteurs.

Exemple. Cas d'un mélange de facteurs à 2 et 4 niveaux.

Si un facteur A a 4 niveaux, on peut le décomposer en deux pseudofacteurs A_1, A_2 à 2 niveaux en utilisant une correspondance entre niveaux de A et paires de niveaux du couple (A_1,A_2) du type indiqué dans le tableau 2. Avec ce type de décomposition, la recherche se ramène à une recherche sur des facteurs à 2 niveaux. Mais le modèle $\mathcal{M}$ et sa partie à estimer $\mathcal{E}$ doivent prendre en compte le fait qu'un produit de deux pseudofacteurs tel que A_1A_2 ne représente pas une interaction mais un effet principal au même titre que A_1 ou A_2, et de même qu'un produit A_1A_2B appartient à l'interaction de 2 facteurs seulement. Contrairement à d'autres logiciels, PLANOR gère automatiquement cette décomposition, ce qui facilite beaucoup l'écriture du modèle et de la partie à estimer.

A	A_1	A_2
1	-1	-1
2	-1	1
3	1	-1
4	1	1

TAB. 2 – *Décomposition en pseudofacteurs d'un facteur à 4 niveaux*

Il est aussi possible avec PLANOR d'identifier les niveaux de A aux éléments $\{0,1,2,3\}$ du groupe cyclique $\mathbb{Z}/4\mathbb{Z}$. Cette option a l'inconvénient de restreindre beaucoup les possibilités combinatoires de formation d'un plan approprié. Mais elle peut être utile pour retrouver certains plans cycliques généralisés comme ceux décrits dans [1]. Dans ce cas de figure, l'algorithme utilise la méthode de formation de l'orthogonal d'un sous-groupe donné par [3].

Si les nombres de niveaux sont divisibles par plusieurs nombres premiers distincts, la recherche se complique.

Exemple. Cas d'un mélange de facteurs à 2, 3, 4, 6 niveaux.

Chaque facteur est décomposé en produit de pseudofacteurs à 2 et 3 niveaux. On cherche alors

A) une fraction adaptée aux pseudofacteurs à 2 niveaux
B) puis compte tenu de la fraction trouvée en A), une fraction adaptée aux pseudofacteurs à 3 niveaux.

Si la recherche B) échoue, on revient à A) et on essaye un autre choix. Si on aboutit in fine à une solution, le plan résultant est le "*produit d'Hadamard*" entre la fraction trouvée pour les pseudofacteurs à 2 niveaux et celle trouvée pour les pseudofacteurs à 3 niveaux.

Ces plans produits d'Hadamard ont cependant souvent trop d'unités pour être pratiquement intéressant. En revanche en les juxtaposant, on obtient des plans à "*fractions parallèles*" extrêmement efficaces, comme ceux décrits dans [11]. Mais la procédure de recherche automatique de PLANOR n'est pas adaptée à la construction de telles juxtapositions.

4 Utilisation de plusieurs paires "modèle, partie à estimer"

Dans une étude sur la cohésion du fromage de conté frais, on étudie 11 facteurs de commande du procédé avec 64 unités expérimentales. Ces unités sont des microcontés que l'on fabrique en début de semaine (lundi et mardi) à raison de 4 par jour. L'expérimentation se fait donc sur 8 semaines, avec 8 fabrications par semaine. Les variations saisonnières du lait obligent à prendre en compte les blocs définis par les semaines et il peut être judicieux de considérer aussi les sous-blocs de taille 4 correspondant aux journées de fabrication.

Trois des facteurs étant plus spécialement importants, on cherche d'abord une fraction 2^{11-5} permettant d'estimer dans le modèle $\mathcal{M}$ comportant toutes les interactions de 2 facteurs la partie $\mathcal{E}$ comportant toutes les interactions avec ces trois facteurs principaux notés A, B, C. Cette recherche donne une solution [10].

On introduit alors les blocs BL et sous-blocs SB en rajoutant l'effet sous-bloc SB.BL dans le modèle $\mathcal{M}$. Ceci a pour effet de contraindre les effets de $\mathcal{E}$ à être estimés intra-sous-bloc, donc avec la précision maximum. Mais cette recherche n'aboutit pas. En revanche, si on ne rajoute que l'effet bloc BL dans le modèle $\mathcal{M}$, c'est à dire si on demande seulement l'estimabilité intra-bloc et non intra-sous-bloc des effets de $\mathcal{E}$, on obtient une solution. Il est alors naturel de se demander si on ne peut pas obtenir, en plus de l'estimabilité intra-bloc de tous les effets de $\mathcal{E}$, l'estimabilité intra-sous-bloc des effets principaux. Pour cela, on demande, outre l'estimation de $\mathcal{E}$ dans le modèle $\mathcal{M}$ avec l'effet bloc BL, l'estimation de la sous-partie $\mathcal{E}'$ des effets principaux dans le modèle $\mathcal{M}'$ incluant l'effet sous-bloc BL.SB. Cette recherche aboutit.

Le tableau 3 précise la syntaxe d'introduction dans PLANOR de ces deux paires ($\mathcal{M}$,$\mathcal{E}$), ($\mathcal{M}'$,$\mathcal{E}'$) définissant la recherche effectuée. La notation $P.P$ où P est la somme formelle des 11 facteurs est un raccourci pour noter la somme des 55 produits entre deux des 11 facteurs. Comme chaque modèle est automatiquement complété par les sous termes des termes y figurant, la présence de ces produits dans les modèles $\mathcal{M}$,$\mathcal{M}'$ entraîne celle des effets principaux A à K. Cette complétion n'a pas lieu dans les parties à estimer car dans certains cas, on peut pour éviter de confondre une interaction, confondre les effets principaux des facteurs y figurant. Donc le produit $P(A+B+C)$ dans $\mathcal{E}$ ne comprend que les interactions entre les trois facteurs A, B, C et les autres facteurs. Pour inclure dans $\mathcal{E}$ les effets principaux, on doit soit rajouter le terme P, soit utiliser le produit $P(1+A+B+C)$ où on a rajouté 1 entre parenthèses.

La troisième paire $\mathcal{M}'' = BL.SB$, $\mathcal{E}'' = \emptyset$ permet de rendre inéligible les termes de BL.SB, ce qui évite des confusions entre effets blocs et assure qu'il y a effectivement 16 sous-blocs dans le plan obtenu.

Partie de modèle

$P : A+B+C+D+E+F+G+H+I+J+K$

		Modèle			Partie à estimer
$\mathcal{M}$	:	$P.P+BL$	$\mathcal{E}$	:	$P+P(A+B+C)$
$\mathcal{M}'$	:	$P.P+BL.SB$	$\mathcal{E}'$	:	P
$\mathcal{M}''$	:	$BL.SB$	$\mathcal{E}''$	:	

TAB. 3 – *Recherche avec plusieurs paire (modèle, partie à estimer) dans PLANOR*

Références

[1] Dean A.M. and Lewis S.M. (1980). A unified theory for generalized cyclic designs. J. Statist. Plan. Inference **4**, 13-23.

[2] Draper N.R. and Mitchell T.J. (1967). The construction of saturated 2_R^{k-p} designs. *Ann. Math. Statist.*, **38**, 1110-1126.

[3] El Mossadeq A., Kobilinsky A. et Collombier D. (1985). Construction d'orthogonaux dans les groupes abéliens finis et confusion d'effets dans les plans factoriels. Linear Algebra and its Applications **70**, 303-320.

[4] Franklin M.F., Bailey R.A. (1977). Selection of Defining Contrasts and Confounded Effects in Two-level Experiments. *Appl. Statist.*, **26**, No. 3, 321-326.

[5] Franklin M.F. (1985). Selecting Defining Contrasts and Confounded Effects in p^{n-m} Factorial Experiments. *Technometrics*, **27**, No. 2, 165-172.

[6] Kobilinsky A. (1990). Complex linear models and cyclic designs. *Linear Algebra and its Applications*, **127**, 227-282.

[7] Kobilinsky A. (1994). Automatic generation of asymmetrical regular designs. Document interne Biométrie et IA, INRA Jouy.

[8] Kobilinsky A., Mettler E., Carpentier B. (1995). Construction et analyse d'un plan asymétrique de résolution IV. *Actes 4èmes journées v européennes Agro-Industrie et méthodes statistiques. ASU. DIJON.*

[9] Kobilinsky A. and Monod H. (1995). Juxtaposition of regular factorial designs and the complex linear model. *Scand. J. Statist* **22**, n° 2, 223-254.

[10] Kobilinsky A. (1997). Les plans factoriels. Chap 3, p69–209. In : *Plans d'expériences. Applications à l'entreprise.* Eds : J.J. Droesbeke, J.Fine, G. Saporta. TECHNIP, Paris. 509p.

[11] McLean R.A. and Anderson V.L. (1984). Applied factorial and fractional designs. Marcel Dekker, New-York, 373p.

[12] Patterson H.D., Bailey R.A. (1978). Design keys for factorial experiments. Applied Statistics, **27**, 335-343.

[13] Ripoche A., Kobilinsky A., Vendeuvre J.L. (1997). Vers le contrôle de fabrication du jambon cuit supérieur. *Actes des 5èmes journées européennes Agro-Industrie et méthodes statistiques. ASU, INRA Versailles*