JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

MATHIAS DRTON
JEAN-MARC AZAÏS

Analyse de la variance non-équirépétée hiérarchique : comparaison de cinq logiciels

Journal de la société française de statistique, tome 140, n° 1 (1999), p. 23-40

<http://www.numdam.org/item?id=JSFS_1999__140_1_23_0>

ANALYSE DE LA VARIANCE NON-ÉQUIRÉPÉTÉE HIÉRARCHIQUE : COMPARAISON DE CINQ LOGICIELS

Mathias DRTON, Jean-Marc AZAÏS *

RÉSUMÉ

Dans le cas de l'analyse de la variance à plusieurs facteurs hiérarchisés, l'article montre la diversité des sommes de carrés calculées par des logiciels comme BMDP, MINITAB, SAS, S-PLUS et SPSS, et donc le grand soin qu'il faut porter à leur utilisation. Nous explicitons les hypothèses testées, principalement en ce qui concerne les effets principaux, et nous étudions leur interprétation pratique. Nous illustrons notre étude sur deux exemples à deux ou trois facteurs respectivement.

Mots clés — Analyse de la variance, hiérarchisation, logiciels statistiques, tests d'effets principaux.

ABSTRACT

This article discusses for ANOVA with multi-way-nesting the different ways in which sums of squares are calculated by various software packages, such as BMDP, MINITAB, SAS, S-PLUS and SPSS, so that the author is adviced to be very careful. We find the explicit form of the tested hypotheses, especially for main effects and we show their practical relevance We specifically investigate two examples, one two-way and one three-way.

Keywords and phrases — Unbalanced analysis of variance, nesting, statistical software, tests of main effects.

Classification AMS . 62J10, 62F03.

1. INTRODUCTION

Chaque logiciel statistique moderne propose des outils pour une analyse de la variance dont l'utilisation est devenue standard et dont la théorie est bien explorée. Cependant des données non-équirépétées nécessitent une interprétation prudente des sorties des logiciels. Leurs nombreuses statistiques

* Laboratoire de Statistique et Probabilités, UMR CNRS C5583, Université Paul Sabatier Toulouse, 118, route de Narbonne, F31062 Toulouse Cedex 04.
e-mail drton@math.uni-augsburg.de, azais@cict.fr

de Fisher servent à tester des hypothèses qui ne sont pas toujours celles qui intéressent.

Ce travail veut montrer les problèmes d'une application irréfléchie des logiciels BMDP, MINITAB, SAS, S-PLUS et SPSS à des données hiérarchiques non-équirépétées. Une étude précédente [1] avait exploré le cas du modèle à deux facteurs croisés. Dans la présente étude, nous nous intéressons tout d'abord au cas de deux facteurs hiérarchisés (paragraphe 2), avant d'introduire un troisième facteur en croisement (paragraphe 3). Nous considérons le cas le plus complexe de la hiérarchisation non-uniforme dans le sens où le nombre de niveaux du facteur hiérarchisé n'est pas constant.

Considérons une expérience comprenant deux facteurs A et B. Il est bien connu que deux relations sont possibles entre ces deux facteurs. Le cas le plus courant est celui où les deux facteurs sont croisés : chacun des niveaux de chacun des facteurs possède un sens propre indépendamment des niveaux de l'autre facteur ; l'expérience explore, de manière plus ou moins complète, l'ensemble des combinaisons des deux facteurs.

Mais il existe un cas où, par exemple, les niveaux du facteur B n'ont de sens que lorsque l'on connaît le niveau correspondant du facteur A. Nous parlerons alors du facteur B hiérarchisé au facteur A ou encore du facteur A hiérarchisant le facteur B.

Il existe deux façons de numéroter les niveaux d'un facteur hiérarchisé. La plus courante consiste à numéroter les niveaux de B à l'intérieur des niveaux de A. C'est celle que nous adopterons. Dans l'exemple que nous utiliserons A est le facteur « marque de voiture », et B est le facteur « modèle de voiture » à l'intérieur de la marque. Par exemple, le modèle 1 de marque 1 n'a aucun point commun avec le modèle 1 de la marque 2.

2. MODÈLE HIÉRARCHIQUE À DEUX FACTEURS

2.1. Le modèle et les données

Supposons que nous voulions comparer la consommation d'essence Y des voitures de $a = 2$ marques (A) en étudiant b_i modèle (B) de la marque i. Le facteur modèle n'ayant aucun sens sans être associé à une des deux marques, il est hiérarchisé au facteur marque de voiture d'où le modèle statistique :

$$Y_{ijk} = \mu + A_i + AB_{ij} + E_{ijk}, \tag{1}$$

$i = 1, \ldots, a$; $j = 1, \ldots, b_i$; $k = 1, \ldots, n_{ij}$. Les termes d'erreur E_{ijk} de moyenne 0 et variance σ^2 sont supposés non-corrélés. On fait $n = \sum_{i=1}^{a} \sum_{j=1}^{b_i} n_{ij}$ observations.

Les logiciels se basent sur ce modèle (1) qui est surparamétré. En constraste, nous exprimerons les hypothèses intéressantes dans les termes du modèle (2)

qui est celui des moyennes des cellules et régulier :

$$Y_{ijk} = \theta_{ij} + E_{ijk}; \tag{2}$$

θ_{ij} est la consommation théorique du j-ième modèle de la marque i.

Si on veut tester l'influence du facteur hiérarchisé B il n'y a qu'une seule façon de procéder. L'hypothèse d'intérêt est que la réponse moyenne ne dépend que du facteur hiérarchisant A, c'est-à-dire que

$$\theta_{i1} = \theta_{i2} = \ldots = \theta_{ib_I}, \quad \forall i = 1, \ldots, a. \tag{3}$$

Par contre, pour le facteur hiérarchisant A, deux définitions au moins de la nullité de l'effet de ce facteur sont raisonnables et elles correspondent aux hypothèses

$$\frac{1}{b_i} \cdot \sum_{j=1}^{b_i} \theta_{ij} = \text{const.} \tag{4}$$

et

$$\frac{1}{n_{i+}} \cdot \sum_{j=1}^{b_i} \theta_{ij} \cdot n_{ij} = \text{const., où } n_{i+} := \sum_{j=1}^{b_i} n_{ij}. \tag{5}$$

On peut remarquer que les différentes hypothèses que nous avons exprimées dans le modèle (2) peuvent également s'exprimer dans le modèle (1) par le choix d'un système de contraintes.

La complexité du modèle (1) dépend étroitement du degré d'équilibre des données. On considère en général quatre niveaux de complexité croissante :

1) hiérarchisation uniforme $b_i = \text{const.}$ avec équirépétition $n_{ij} = \text{const.}$; dans ce cas toutes les décompositions sont identiques et, par conséquent, tous les logiciels donnent la même analyse;

2) hiérarchisation non-uniforme $b_i \neq \text{const.}$ avec équirépétition $n_{ij} = \text{const.}$,

3) hiérarchisation uniforme $b_i = \text{const.}$ avec non-équirépétition $n_{ij} \neq \text{const.}$,

4) hiérarchisation non-uniforme $b_i \neq \text{const.}$ et non-équirépétition $n_{ij} \neq \text{const.}$

Par souci de concision, nous nous intéresserons au dernier cas qui renferme toute la difficulté. Un lecteur intéressé par les cas intermédiaires peut consulter [3].

Les deux hypothèses (4) et (5) peuvent être intéressantes. Toutes deux comparent la consommation des voitures des deux marques à l'aide d'un modèle de voiture moyen. Pour l'hypothèse (4), on prend la moyenne sur toute la gamme. Si l'on choisit dans (5) les effectifs n_{ij} proportionnellement aux ventes alors on compare la consommation moyenne de toutes les voitures de marque i existantes. Par conséquent, le test de (4) pourrait être interprété comme un test de comparaison des aptitudes des ingénieurs des différentes firmes à construire des véhicules économiques. Tandis que le test de (5) (avec les effectifs n_{ij} proportionnels aux ventes) pourrait être interprété comme une comparaison des conséquences sur l'environnement.

A cause de la diversité des modèles (grand, petit, sportif,...), il est raisonnable que l'effet du facteur hiérarchisé soit significatif. Cette significativité rend impossible une stratégie de regroupement, c'est-à-dire le test de l'effet marque dans un modèle sans le facteur modèle. Néanmoins, l'étude du facteur principal hiérarchisant est l'objectif de l'analyse (comparaison des marques). Notre premier exemple, pour l'étude des logiciels dans le cas de modèles à deux facteurs, est présenté dans le tableau 1. Le tableau 2 présente une variante minime où les niveaux du second facteur ont été renumérotés.

TABLEAU 1 — Nos données non-équirépétées non-uniformément hiérarchiques

A	B	Y	n_{ij}
1	1	6.1	$n_{11} = 2$
1	1	5.9	
1	2	7.6	$n_{12} = 3$
1	2	7.8	
1	2	7.9	
1	3	11.7	$n_{13} = 3$
1	3	12.3	
1	3	11.8	
2	1	7.4	$n_{21} = 2$
2	1	7.9	
2	2	10.5	$n_{22} = 2$
2	2	10.7	
$n = 12,\ b_1 = 3,\ b_2 = 2$			

TABLEAU 2. — Changement des noms des niveaux du facteur hiérarchisé

A	B	Y	n_{ij}
1	1	6.1	$n_{11} = 2$
1	1	5.9	
1	2	7.6	$n_{12} = 3$
1	2	7.8	
1	2	7.9	
1	3	11.7	$n_{13} = 3$
1	3	12.3	
1	3	11.8	
2	2	7.4	$n_{22} = 2$
2	2	7.9	
2	3	10.5	$n_{23} = 2$
2	3	10.7	
$n = 12,\ b_1 = 3,\ b_2 = 2$			

2.2. Résultats des logiciels

2.2.1. *SAS Release 6.11 pour Unix*

Dans le cas non-équirépété, il faut utiliser le programme GLM qui est décrit dans le manuel [6] et l'article [1] :

```
data sasuser.voitures;
infile 'data.dat';
input A B Y;
run;
proc glm;
class A B;
model Y=A B(A)/ ss1 ss2 ss2 ss4;
run;
```

Les sommes de carrés dans la sortie de SAS se présentent comme suit :

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	1	0.1504167	0.1504167	2.52	0.1567
B(A)	3	56.9779167	18.9926389	317.81	0.0001

Source	DF	Type II SS	Mean Square	F Value	Pr > F
A	1	0.1504167	0.1504167	2.52	0.1567
B(A)	3	56.9779167	18.9926389	317.81	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	0.8211585	0.8211585	13.74	0.0076
B(A)	3	56.9779167	18.9926389	317.81	0.0001

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
A	1	0.8211585	0.8211585	13.74	0.0076
B(A)	3	56.9779167	18.9926389	317.81	0.0001

L'analyse avec GLM ne pose aucun problème. Les sommes de carrés de type I et II coïncident et permettent de tester l'hypothèse (5) basée sur les effectifs et les sommes identiques de type III et IV testent l'hypothèse (4) qui ne dépend que des nombres de niveaux des facteurs.

2.2.2. *BMDP Release 7.1 pour Unix*

BMDP offre deux programmes pour l'analyse de la variance dans le cas d'effets fixes. Le programme 2V n'est pas capable de traiter une hiérarchisation (voir [2], page 522). Nous appliquerons 4V (voir [2], page 1259) qui permet d'étudier des structures hiérarchiques. Ce programme base son analyse sur des pondérations des cellules. L'utilisateur peut facilement définir une pondération individuelle mais deux sont proposées par défaut : la constante (`equal`) et celle avec les effectifs (`sizes`), d'où les routines :

```
PROGRAM INSTRUCTIONS
/input var=3.
format=free.
file='data.dat'.
/variable names=a,b,y.
/between factors= a,b.
codes(1)=1,2.
codes(2)=1 to 3.
/weights between=equal.
/end
ANALYSIS INSTRUCTIONS
analysis proc=struc.
bform='a+a.b'./#
```

```
PROGRAM INSTRUCTIONS
/input var=3.
format=free.
file='data.dat'.
/variable names=a,b,y.
/between factors= a,b.
codes(1)=1,2.
codes(2)=1 to 3.
/weights between=sizes.
/end
ANALYSIS INSTRUCTIONS
analysis proc=struc.
bform='a+a.b'./#
```

Par elles, BMDP met à la disposition les tests des hypothèses intéressantes. Plus précisément, la pondération constante sert à vérifier (4), et celle par les effectifs aboutit au test de (5). Nous ne présentons pas les résultats numériques car ils sont identiques aux résultats correspondants de SAS.

2.2.3. *S-PLUS Version 3.4 Release 1 pour Unix*

Le livre [8] introduit l'utilisation de S-PLUS. Après la création de la structure des données et du modèle :

```
data1 <- read.table("data.dat",header=T)
data <- data.frame(A=factor(data1$A),B=factor(data1$B),Y=data1$Y$)
data.aov <- aov(Y~A+A:B, data)
```

la commande `anova(data.aov)` est l'outil standard pour effectuer une analyse de la variance. Cette analyse est la séquentielle du type I de SAS et mène à l'hypothèse (5).

Dans une analyse séquentielle, le test d'un effet est le test ajusté pour tous les effets précédents dans la déclaration du modèle. Dans un modèle avec les facteurs $A_1, \ldots, A_n$, on teste A_p à travers l'hypothèse nulle : « la projection du vecteur des réponses moyennes $E[Y]$ sur l'orthogonal de $< A_1, \ldots, A_{p-1} >$ dans $< A_1, \ldots, A_p >$ est nulle ». Ici, $< A_1, \ldots, A_p >$ désigne l'espace des réponses moyennes si on adapte le modèle avec les facteurs $A_1, \ldots, A_p$. Dans notre exemple, on teste le facteur marque en comparant le modèle à un facteur marque avec celui paramétré seulement par la moyenne générale.

En outre, la commande `drop1(data.aov,scope=data ;aov$call)` permet de générer les sommes de carrés de type III. La démarche de `drop1` se base sur des réductions de la somme de carrés du modèle. La réduction correspondant au type III se fonde sur la comparaison du modèle complet avec celui qui est privé du facteur en considération. Cela veut dire que le test de l'effet A_p est le test de « la projection de $E[Y]$ sur l'orthogonal de $< A_1, \ldots, A_{p-1}, A_{p+1}, \ldots, A_n >$ dans $< A_1, \ldots, A_n >$ est nulle ». Par exemple, on teste le facteur marque en comparant le modèle qui inclut les facteurs marque et modèle avec celui n'incluant que le facteur modèle. Mais cette description du type III par réduction n'est possible que si l'on impose au modèle singulier (1) un système de contraintes de régularité (voir [7]).

Ici, c'est la non-uniformité de la hiérarchisation qui fait découvrir les faiblesses de `drop1`. Nous citons les résultats de S-PLUS qui se distinguent de ceux de type III de SAS :

	Df	Sum of Sq	RSS	F Value	Pr(F)
<none>			0.41833		
A	0	0.00000	0.41833		
A:B	3	56.97792	57.39625	317.8051	7.674036e-08

Nous voyons que S-PLUS ne peut définir l'hypothèse associée à A puisque le nombre de degrés de liberté est nul. Il ne permet pas de tester l'hypothèse (4) dans le cadre de hiérarchisation non uniforme.

2.2.4. *SPSS Release 6.1 pour Unix*

Pour appliquer SPSS dans un cas hiérarchique, il faut éditer le code produit par SPSS à la main ce qui est décrit dans le guide [5] :

```
GET FILE=
  'data.sav'.
EXECUTE .
MANOVA
  y BY a(1 2) b(1 3)
  /NOPRINT PARAM(ESTIM)
  /METHOD=UNIQUE
  /ERROR WITHIN
  /DESIGN = a, b WITHIN a .
```

```
MANOVA
  y BY a(1 2) b(1 3)
  /NOPRINT PARAM(ESTIM)
  /METHOD=SEQUENTIAL
  /ERROR WITHIN
  /DESIGN = a, b WITHIN a .
```

SPSS propose deux types de sommes de carrés dont le type séquentiel correspond au type I de SAS. Pour son type unique, SPSS avertit :

The hypotheses tested may not the hypotheses of interest. Different reorderings of the model or data, or different contrasts may result in different UNIQUE sums-of-squares.

Nous précisons les sommes de carrés uniques pour notre exemple :

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN CELLS	.42	7	.06		
A	6.57	1	6.57	109.88	.000
B WITHIN A	56.98	3	18.99	317.81	.000

La somme unique de SPSS pour A se distingue des deux sommes de carrés pour les tests de (4) et (5). Par défaut, SPSS utilise des contrastes appelés *deviation*. En jouant sur ces constrastes, nous trouvons que les contrastes *difference* donnent la somme de carrés de type III et le test de (4).

Mais ce ne sont pas toujours les contrastes *difference* qui mènent au type III. Si nous changeons dans notre fichier les noms des niveaux du facteur modèle qui sont hiérarchisés à la deuxième marque de voitures 1 et 2 en 2 et 3, comme fait au tableau 2, ce sont les contrastes *helmert* qui permettent de tester l'hypothèse (4). En particulier, on voit que les sommes de carrés uniques dépendent de la numérotation.

Revenons sur les sommes uniques avec les contrastes par défaut *deviation*. Ces sommes uniques dépendent en fait du type de contraste utilisé dès que la hiérarchisation est non-uniforme et que le facteur hiérarchisé a plus de deux niveaux. Dans notre exemple, nous trouvons que les sommes uniques par défaut servent à tester l'hypothèse

$$1/3 \cdot (\theta_{11} + \theta_{12} + \theta_{13}) = \theta_{22}. \tag{6}$$

Nous ne voyons pas d'intérêt pratique pour cette hypothèse (6) qui exclut les observations pour $(A, B) = (2, 1)$. Mais ce n'est pas toujours la première cellule qui est exclue de l'analyse. Après changement de la numérotation (voir tableau 2). SPSS *unique* avec les contrastes *deviation* teste l'hypothèse qui supprime la cellule $(A, B) = (2, 3)$. Celle-ci s'écrit encore comme (6).

2.2.5. *MINITAB Version 12.2 pour Windows*

MINITAB s'applique facilement en suivant les instructions de son menu. Nous utilisons la routine *Modèle linéaire généralisé* dans le menu *Stat/ANOVA*. La déclaration de notre modèle se fait en écrivant `Y` dans la fenêtre pour les *réponses* de `A B(A)` dans la fenêtre pour le *modèle*. Deux types de sommes de carrés sont proposés sous les *options*. MINITAB explique, par une remarque entre parenthèses, les sommes séquentielles comme type I et les ajustées comme type III. Et, en fait, les séquentielles correspondent au type I de SAS et testent (5) et les ajustées permettent de tester l'hypothèse (4) qui était aussi testée par le type III de SAS.

2.3. Conclusion

Une hiérarchisation non-uniforme non-équirépétée dépasse les capacités de S-PLUS pour Unix[1] et le travail avec SPSS sous Unix[2] demande une adaptation des contrastes qui dépend de la numérotation des niveaux des facteurs. Par contre, nos versions de BMDP, MINITAB et SAS proposent sans intervention de l'utilisateur les tests des hypothèses intéressantes.

3. MODÈLE À TROIS FACTEURS CROISÉS ET HIÉRARCHISÉS

3.1. Le modèle et les données

Après avoir vu les problèmes que pose la hiérarchisation non-uniforme pour certains logiciels dans le cas simple de deux facteurs, nous considérons un modèle comprenant simultanément croisement et hiérarchisation de facteurs.

Reprenons l'exemple des voitures. Nous nous intéressons à l'influence de la vitesse (C) sur la consommation des voitures Y. Ce facteur vitesse C est croisé avec la structure des modèles de voitures B hiérarchisés au facteur marque de voiture A. Nous posons le modèle complet incluant toutes les interactions permises par la relation de hiérarchisation

$$Y_{ijkl} = \mu + A_i + C_k + AB_{ij} + AC_{ik} + ABC_{ijk} + E_{ijkl}, \qquad (7)$$

$i = 1, \ldots, a$; $j = 1, \ldots, b_i$; $k = 1, \ldots, c$; $l = 1, \ldots, n_{ijk}$. Les termes d'erreur E_{ijkl} sont non-corrélés centrés et de variance σ^2. On réalise donc en tout $n = \sum_{i=1}^{a} \sum_{j=1}^{b_i} \sum_{k=1}^{c} n_{ijk}$ observations. Le modèle des moyennes des cellules s'écrit comme

$$Y_{ijkl} = \theta_{ijk} + E_{ijkl}. \qquad (8)$$

1. Les dernières versions de S-PLUS sous Windows proposent le type III de SAS en option.
2. Une version Windows propose également le type III de SAS.

Les modèles statistiques étant plus complexes que ceux du premier exemple, nous procéderons à l'inverse : nous décrirons d'abord les résultats des logiciels puis nous donnerons des outils pour les interpréter.

L'exemple numérique (tableau 3) que nous utiliserons comprend deux marques ($a = 2$), deux vitesses ($c = 2$), trois modèles de la première marque ($b_1 = 3$) et deux dans l'autre ($b_2 = 2$). Cet exemple est extrait de [4]. La structure des facteurs est représentée dans la figure 1.

TABLEAU 3. — L'exemple à trois facteurs croisés et hiérarchisés

A	B	C	n_{ijk}
1	1	1	1
1	1	2	1
1	2	1	2
1	2	2	2
1	3	1	2
1	3	2	2
2	1	1	2
2	1	2	2
2	2	1	2
2	2	2	1
$n = 17$, 10 cellules, $a = 2,\ b_1 = 3,\ b_2 = 2,\ c = 2$			

Notre exemple est formulé avec l'objectif d'étudier l'effet de la vitesse C. Derrière cet intérêt se cache le fait que l'effet C est l'effet qui cause le maximum de confusion dans le sens que les logiciels proposent le plus d'hypothèses différentes.

3.2 Résultats des logiciels sur l'effet principal croisé

Sur l'effet principal C, les cinq logiciels testent huit hypothèses différentes H_q, $q = 1, \ldots, 8$. Ces hypothèses s'écrivent toutes comme une égalité de combinaisons linéaires des paramètres du modèle régulier (8), d'où la forme générale

$$H : \frac{1}{c_{++1}} \cdot \sum_{i,j} c_{ij1} \cdot \theta_{ij1} = \frac{1}{c_{++2}} \cdot \sum_{i,j} c_{ij2} \cdot \theta_{ij2}, \tag{9}$$

avec $c_{++k} := \sum_{i,j} c_{ijk}$. Le tableau 4 décrit les différentes hypothèses envisagées par les différents logiciels en donnant les valeurs des coefficients c_{ij1} en première ligne et des coefficients c_{ij2} en seconde ligne, ainsi que les sommes c_{++k} correspondantes. Le tableau 5 indique quelle hypothèse est utilisée par quel logiciel.

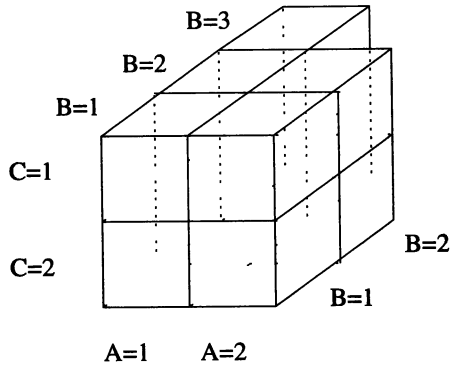

FIG 1. — *Illustration de la structure de l'exemple à trois facteurs*

Afin d'interpréter les hypothèses dans notre contexte, nous rappelons qu'on étudie la consommation d'essence de $n = 17$ voitures et que le niveau de C indique la vitesse à laquelle on conduit la voiture de modèle B de marque A.

Les hypothèses peuvent se classer en deux groupes. Premièrement, les hypothèses H_1, H_2, H_5 et H_6 dépendent des effectifs, secondement, H_3, H_4, H_7 et H_8 sont indépendantes des effectifs. Comme vu au paragraphe 2.2.4, l'hypothèse par défaut de SPSS : H_8 n'est pas intéressante car elle exclut des observations.

Dans le premier groupe, H_1 est l'hypothèse séquentielle. L'hypothèse H_2 est celle du type II de SAS. Nous ne voyons pas l'intérêt de ces deux hypothèses. Par contre, la pondération `sizes` de BMDP utilise les effectifs comme coefficients d'hypothèse et fournit une hypothèse H_5 bien interprétable qui attribue la même importance à chaque voiture observée. Mais la pondération `sizes` génère une non-orthogonalité du plan sous-jacent. Le lecteur intéressé pourra trouver plus de détail dans [3]. Cette non-orthogonalité fait que BMDP propose une hypothèse ajustée qui est H_6. Comme on peut le vérifier sur la sortie de BMDP au paragraphe 3.3.2., cette hypothèse est ajustée pour tous les facteurs à l'exclusion de l'interaction triple. Ceci peut permettre d'apprécier l'influence du facteur C dans un modèle sans cette interaction. Mais dans notre cas cette interaction est bien présente (et même significative) et nous ne voyons pas l'intérêt du test de H_6.

Dans le deuxième groupe, l'équipondération des cellules aboutit à H_4. Elle est fournie par le type IV de SAS et par la pondération `equal` de BMDP. Cette hypothèse apparaît peut-être comme la plus naturelle car elle donne la même importance à tous les modèles de voitures étudiés. L'hypothèse H_7 est basée sur l'importance des marques de voitures. MINITAB la teste par ses sommes ajustées et SPSS permet son test par les sommes uniques après le choix des contrastes *difference*. H_7 donne le même poids aux deux marques puis à chaque modèle. Ce traitement égal des deux marques peut être intéressant si la non-uniformité de la hiérarchisation est artificielle, par exemple s'il y avait des modèles de voiture non disponibles. Une telle irrégularité de l'expérience peut alors être affaiblie en équipondérant les deux marques. En ce qui concerne

l'importance donnée aux deux marques de voitures comme démontré dans [3] les coefficients dans l'hypothèse H_3 se basent sur les nombres des niveaux du facteur B. Plus précisément on trouve que, pour notre modèle (7) avec les nombres de niveaux $a = 2, c = 2$, b_1 et b_2, l'hypothèse H_3 donne le coefficient $\frac{b_1 + 1}{b_1 + b_2 + 2b_1b_2}$ aux cellules avec $A = 2$ et le coefficient $\frac{b_2 + 1}{b_1 + b_2 + 2b_1b_2}$ à celles avec $A = 1$. Ainsi H_3 se trouve en position intermédiaire entre H_4 et H_7 mais plus près de H_7. Hormis ce point, l'intérêt de cette hypothèse n'est pas clair.

TABLEAU 4. — Description des hypothèses en donnant les coefficients c_{ijk} de (9). Le niveau de AB forme les colonnes et celui de C les lignes.

hypothèse			AB					c_{++k}
		C	11	12	13	21	22	
Séquentiel	H_1	1	7	14	14	12	12	59
		2	7	14	14	16	8	59
SAS type II	H_2	1	3	6	6	6	4	25
		2	3	6	6	6	4	25
SAS type III	H_3	1	3	3	3	4	4	17
		2	3	3	34	4	4	17
Equipondérqtion	H_4	1	1	1	1	1	1	5
		2	1	1	1	1	1	5
Effectifs	H_5	1	1	2	2	2	2	9
		2	1	2	2	2	1	8
Effectifs ajustés	H_6	1	35	70	70	72	48	295
		2	35	70	70	72	48	295
Selon A	H_7	1	2	2	2	3	3	12
		2	2	2	2	3	3	12
SPSS défaut	H_8	1	1	1	1	0	3	6
		2	1	1	1	0	3	6

TABLEAU 5. — Correspondance entre les logiciels et les hypothèses du tableau 4

	Les hypothèses							
Les logiciels	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
SAS	x	x	x	x				
S-PLUS	x							
SPSS	x						x	x
MINITAB	x						x	
BMDP				x	x	x		

TABLEAU 6. — Les observations de l'exemple

A	B	C	Y	A	B	C	Y
1	1	1	54	2	1	1	17
1	1	2	14	2	1	1	12
1	2	1	21	2	1	2	21
1	2	1	17	2	1	2	25
1	2	2	36	2	2	1	15
1	2	2	28	2	2	1	14
1	3	1	24	2	2	2	18
1	3	1	25				
1	3	2	18				
1	3	2	15				
$n = 17$							

3.3. Les codes et les sorties

Pour convaincre le lecteur de l'existence de huit hypothèses différentes, nous considérons des observations concrètes (extraites de [4]) et nous donnons ci-dessous les codes et les sorties des logiciels correspondantes.

3.3.1. *SAS*

Le programme d'appel de SAS après la génération des données est

```
proc glm;
class A B C ;
model Y = A C B(A) A*C B(A)*C / ss1 ss2 ss3 ss4 ;
run;
```

et la sortie regroupe les quatre types de sommes de carrés

Dependent Variable: Y		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	9	1622.00000	180.22222	19.11	0.0004
Error	7	66.00000	9.42857		
Corrected Total	16	1688.00000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	1	248.685714	248.685714	26.38	0.0013
C	1	2.561743	2.561743	0.27	0.6183
B(A)	3	261.502542	87.167514	9.25	0.0079
A*C	1	156.250000	156.250000	16.57	0.0047
B(A)*C	3	953.000000	317.666667	33.69	0.0002

Source	DF	Type II SS	Mean Square	F Value	Pr > F
A	1	251.011347	251.011347	26.62	0.0013
C	1	4.166667	4.166667	0.44	0.5275
B(A)	3	250.266667	83.422222	8.85	0.0088
A*C	1	156.250000	156.250000	16.57	0.0047
B(A)*C	3	953.000000	317.666667	33.69	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	314.285714	314.285714	33.33	0.0007
C	1	42.750000	42.750000	4.53	0.0707
B(A)	3	253.600000	84.533333	8.97	0.0085
A*C	1	291.844156	291.844156	30.95	0.0008
B(A)*C	3	953.000000	317.666667	33.69	0.0002

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
A	1	314.285714	314.285714	33.33	0.0007
C	1	81.384615	81.384615	8.63	0.0218
B(A)	3	253.600000	84.533333	8.97	0.0085
A*C	1	291.844156	291.844156	30.95	0.0008
B(A)*C	3	953.000000	317.666667	33.69	0.0002

On voit immédiatement les quatre sommes différentes pour l'effet C.

3.3.2. *BMDP*

Le code de BMDP doit être adapté au nouveau modèle comme suit

```
/input var=4.
  format=free.
  file='data.txt'.
/variable names=a,b,c,y.
/between factors= a,b,c.
  codes(1)=1,2.
  codes(2)=1 to 3.
  codes(3)=1,2.
/weights between=sizes.
/end
analysis proc-struc.
  bform='a+c+a.b+a.c+a.b.c'./#
```

La sortie pour la pondération `sizes` contient aussi les ajustements proposés à cause de la non-orthogonalité apparaissant avec `sizes` :

EFFECT	VARIATE	STATISTIC	F	DF	P
OVALL: GRAND MEAN					
	y	SS= 8228.000000			
		MS= 8228.000000	872.67	1, 7	0.0000
a	y				
		SS= 248.685714			
		MS= 248.685714	26.38	1, 7	0.0013
c	y				
		SS= 0.236111			
		MS= 0.236111	0.03	1, 7	0.8787
ac	y				
		SS= 167.485876			
		MS= 167.485876	17.76	1, 7	0.0040
ab	y				
		SS= 259.897619			
		MS= 86.632540	9.19	3, 7	0.0080
acb	y				
		SS= 953.000000			
		MS= 317.666667	33.69	3, 7	0.0002
a|c,ac,ab					
	y	SS= 251.618425			
		MS= 251.618425	26.69	1, 7	0.0013
c|a,ac,ab					
	y	SS= 3.489708			
		MS= 3.489708	0.37	1, 7	0.5622
ac|a,c,ab					
	y	SS= 156.250000			
		MS= 156.250000	16.57	1, 7	0.0047

```
 EFFECT  VARIATE       STATISTIC          F      DF          P

OVALL: GRAND MEAN
ab|a,c,ac
        y        SS=     250.266667
                 MS=      83.422222    8.85    3,    7    0.0088
 ERROR  y
                 SS=    66.00000000
                 MS=     9.42857143
```

La pondération constante `equal` donne les résultats du type IV de SAS.

3.3.3. *S-PLUS*

En S-PLUS il faut créer le modèle et la structure des données par

```
data1 <- read.table("data.dat",header=T)
data <- data.frame(A=factor(data1$A),B=factor(data1$B),
                                    C=factor(data1$C),Y=data1$Y)
data.aov <- aov(y~A+C+A:B+A:C+A:B:C, data)
```

Puis la commande `anova(data/aov)` mène au type I de SAS. La deuxième commande d'analyse `drop1(data.aov,scope=data.aov$call)` ne donne pas de résultats.

3.3.4. *SPSS*

Les sommes séquentielles de SPSS reproduisent le type I de SAS mais les sommes uniques avec les contrastes par défaut *deviation* obtenus par

```
GET FILE=
  'data.sav'.
EXECUTE .
MANOVA
  y BY a(1 2) b(1 3) c(1 2)
  /NORPINT PARAM(ESTIM)
  /METHOD=UNIQUE
  /ERROR WITHIN
  /DESIGN = a,c,b WITHIN a,a BY b WITHIN a .
```

aboutissent, en avertissant que les résultats peuvent dépendre des contrastes utilisés (voir la section 2.2.4.), à la sortie

```
Tests of Significance for Y using UNIQUE sums of squares
 Sources of Variation        SS  DF       MS       F  Sig of F

 WITHIN CELLS             66.00   7     9.43
 A                       223.21   1   223.21   23.67       .002
 C                        34.30   1    34.30    3.64       .098
 B WITHIN A              253.60   3    84.53    8.97       .009
 A BY C                  118.30   1   118.30   12.55       .009
 C BY B WITHIN A         953.00   3   317.67   33.69       .000

 (Model)                1622.00   9   180.22   19.11       .000
 (Total)                1688.00  16   105.50
```

Avec les contrastes *difference* réalisés par la commande

```
/CONTRAST (a)=difference /CONTRAST (b)=difference
                         /CONTRAST (c)=difference
```

incluse avant la commande \NOPRINT nous obtenons (avec encore le même avertissement sur le fait que les résultats dépendent des contrastes utilisés)

Tests of Significance for Y using UNIQUE sums of squares

Sources of Variation	SS	DF	MS	F	Sig of F
WITHIN CELLS	66.00	7	9.43		
A	314.29	1	314.29	33.33	.001
C	30.03	1	30.03	3.18	.118
B WITHIN A	253.60	3	84.53	8.97	.009
A BY C	291.84	1	291.84	30.95	.001
C BY B WITHIN A	953.00	3	317.67	33.69	.000
(Model)	1622.00	9	180.22	19.11	.000
(Total)	1688.00	16	105.50		

3.3.5. *MINITAB*

Nous appliquons la routine *Modèle linéaire généralisé* qu'on trouve dans le menu *stat/ANOVA*. Dans notre exemple, nous avons la réponse `Y` et nous posons le modèle `A C B(A) A*C B*C(A)`. Sous *options* nous pouvons choisir le type de sommes de carrés désiré. Les sommes séquentielles correspondent au type I de SAS mais les sommes ajustées qui sont référées entre parenthèses comme type III donnent le même résultat que SPSS *unique* avec les contrastes *difference* :

Analyse de la variance pour Y, en utilisant la SC séquentielle pour les tests

Source	DL	SC Seq	SC Ajust	CM Seq	F	P
A	1	248,69	314,29	248,69	26,38	0,001
C	1	2,56	30,03	2,56	0,27	0,618
B(A)	3	1261,50	253,60	87,17	9,25	0,008
A*C	1	156,25	291,84	156,25	16,57	0,005
C*B(A)	3	953,00	953,00	317,67	33,69	0,000
Erreur	7	66,00	66,00	9,43		
Total	16	1688,00				

Nous voyons donc que la référence à ces sommes comme étant des sommes de type III est trompeuse puisqu'elle peut faire penser qu'elles sont égales aux sommes de type III de SAS.

3.4. Conclusion

Les logiciels proposent trois hypothèses intéressantes. Ce sont : H_5 pondérée par les effectifs, l'équipondération des cellules H_4 et H_7 qui donne le même poids aux deux niveaux de A. À part BMDP, tous les logiciels offrent une analyse séquentielle dont l'utilité est plutôt impénétrable.

Parmi les trois hypothèses raisonnables, SAS ne teste que H_4 et MINITAB seulement H_7. Quant à MINITAB, il faut remarquer que ses sommes ajustées qui testent H_7 sont nommées de façon trompeuse « de type III » : elles n'ont pas de rapport avec le type III de SAS. Les contrastes par défaut de SPSS effectuent des tests dont nous ne voyons pas l'intérêt. Le choix des contrastes convenables permet de vérifier H_7. Mais ce choix est difficile car les contrastes convenables dépendent de la numérotation des niveaux des facteurs.

Finalement, BMDP poursuit conséquemment une stratégie basée sur des pondérations des cellules. Ses deux pondérations par défaut testent H_4 et H_5. Le code de BMDP rend facile l'explicitation d'une pondération individuelle[3] ce qui donne l'accès au test de H_7. Par contre, l'ajustement de BMDP pour sa pondération avec les effectifs `sizes` ne donne pas une hypothèse raisonnable.

3.5. Remarques sur les interactions et le deuxième effet principal

D'abord le test de l'interaction triple ABC est comme toujours unique et il est proposé par toutes les routines présentées. Nous nous concentrons alors sur les effets A, AB et AC.

En ce qui concerne les analyses séquentielle où type II de SAS, comme nous l'avons déjà remarqué dans l'interprétation de l'hypothèse H_6 au paragraphe 3.2., ces hypothèses possèdent un potentiel de prévision sur des sous modèles obtenus en supprimant des interactions comme l'interaction triple en évitant de faire « un pooling » : regroupement de cette interaction avec la résiduelle. Cependant dans cet article nous considérons le cas où le modèle complet est considéré comme vrai, et dans ce modèle beaucoup d'hypothèses ne sont pas interprétables.

Le test séquentiel de A est basé sur une pondération par les effectifs car seul l'ajustement pour la moyenne générale est effectué. Mais pour ce qui concerne les effets AB et AC nous ne voyons pas l'intérêt des hypothèses séquentielles.

Le type II du SAS qui dépend également des effectifs n'aboutit pas à un test intéressant. Pour chacun des effets, BMDP `sizes` propose des tests non ajustés basés sur les effectifs qui sont interprétables. Mais, contrairement à ce qui se passe pour le facteur C, les tests ajustés ne génèrent pas d'hypothèses intéressantes. Le test ajusté pour A est ininterprétable et les tests ajustés de AB et AC correspondent au type II de SAS.

En ce qui concerne les décompositions qui ne dépendent pas des effectifs.

– Pour SPSS *unique* avec ses contrastes par défaut *deviation*, on retrouve pour les effets A, AB et AC le même phénomène que pour C (ou que pour A au paragraphe 2) : l'hypothèse est ininterprétable.

– Pour les types III et IV de SAS, SPSS *unique difference*, MINITAB ajusté et BMDP `equal`, on trouve le même résultat. Tous ces tests ont un sens et peuvent facilement être déduits de l'équipondération des cellules.

3. Par exemple pour H_7 : `/weights between=2,2,2,2,2,2,3,3,3,3.`

– Enfin, la routine `drop1` de S-PLUS donne en plus du test de ABC, celui de AB qui est celui du groupe précédent. Par contre elle ne fournit pas de test de A et AC : la somme des carrés est nulle.

4. CONCLUSION GÉNÉRALE

L'exemple de la partie 3 est particulièrement complexe. Il a pour but de pousser les algorithmes des logiciels dans leurs dernières limites. Fort heureusement de tels modèles ne se rencontrent que peu souvent en pratique.

Pour ce qui concerne le modèle hiérarchique à deux ou trois facteurs, le lecteur a trouvé dans cet article les éléments pour effectuer une analyse qui ait un sens. Il est difficile par contre de dégager des grandes lignes qui puissent s'appliquer à tous les modèles. Par exemple, on montre dans [1] qu'en utilisant SAS avec un modèle à deux facteurs croisés ce sont les sommes de carrés de type III qui sont le plus souvent indiquées et que les sommes de carrés de type IV sont difficilement interprétables. Dans l'exemple de la partie 3, on arrive à la conclusion strictement inverse.

L'utilisateur doit savoir que les modèles d'analyse de la variance à plusieurs facteurs et à données non-équilibrées renferment de nombreux pièges. En dehors des quelques cas simples qui ont été étudiés en détail, l'utilisateur ne peut se reposer de manière aveugle sur le logiciel, il doit au contraire faire l'effort de comprendre de façon précise quelles sont les hypothèses réellement testées.

Remerciements. Les auteurs remercient A. Kobilinsky pour avoir permis le travail sur son exemple dans [4].

BIBLIOGRAPHIE

[1] AZAIS J.-M. (1994). Analyse de variance non orthogonale – L'exemple de SAS/GLM. *Rev. Statistique Appliquée* **XLII** (2) : 27-41.

[2] DIXON W.J. (1992). *BMDP Statistical Software Manual.* Volume 1 & 2. University of California Press, Berkeley.

[3] DRTON M. (1999). Analyse de variance dans des situations hiérarchiques non-équirépétées. *Mémoire de DEA mathématiques appliquées.* Laboratoire de Statistiques et Probabilités. Université Paul Sabatier. Toulouse.

[4] KOBILINSKY A. (1998). Reparametrisation of interest in non uniform factorial designs. *Papier interne du Laboratoire de Biométrie à l'INRA Versailles.*

[5] MARIJA J.N./SPSS Inc. (1994). *SPSS Advanced Statistics 6.1.* SPSS Inc., Chicago.

[6] SAS Institute Inc. (1989). *SAS/STAT User's Guide.* Version 6. Fourth Edition. Volume 1 & 2. SAS Institute Inc., Cary, North Carolina.

[7] SEARLE S.R. (1987). *Linear Models for Unbalanced Data.* John Wiley & Sons, New York.

[8] VENABLES W.N., RIPLEY B.D. (1997). *Modern applied statistics with S-PLUS.* 2nd edition. Springer, New York.