VARIABILITE ET SOURCES DE VARIABILITE A DIFFERENTES ECHELLES ANALYSE DE VARIANCE DE MODELES A EFFETS FIXES OU ALEATOIRES

LALOE F.

RESUME

L'identification de sources de variations à divers niveaux d'observation pose une difficulté due à la diffusion de l'hétérogénéité à travers ces niveaux.

Nous présentons le problème à partir de l'analyse de variance de modèle à effets fixes ou aléatoires, et en nous appuyant sur des des exemples issus d'enquêtes de santé et de rendement de pêche.

Nous discutons des conséquences en matière de cartographie ou d'analyse de séries chronologiques.

INTRODUCTION

La collecte d'information à divers échelles d'observation est monnaie courante. Ainsi dans le domaine de l'échantillonnage, les méthodes de sous échantillonnage font l'objet de chapitres entiers dans les manuels de référence (par exemple Cochran, 1977).

Dans le cas de la pêche artisanale au Sénégal, la collecte d'une partie de l'information repose sur un plan à trois niveaux, avec la sélection:

- de poissons dans des pirogues,
- de pirogues au cours de jours d'enquête,
- de jours d'enquêtes au cours de quinzaines.

Dans le cas d'une enquête sur la couverture vaccinale des enfants de Pikine (périphérie de Dakar), les observations concernaient:

- des enfants dans des îlots,
- des îlots dans des quartiers,
- des quartiers dans des sous zones de la ville.

On désire en général analyser la variabilité aux divers niveaux, en tentant de répondre à des questions qui peuvent être formulées de diverses manières. On peut par exemple se demander si la couverture vaccinale peut varier d'une sous-zone à l'autre, ou bien si certaines caractéristiques des sous-zones sont sources de variabilité dans la couverture vaccinale. Les mêmes questions peuvent être posées pour ce qui concerne les pêches réalisées au cours de quinzaines différentes.

Ces questions sont différentes. Il peut y avoir des différences entre les taux de vaccination selon les sous-zones ou entre les pêches selon les quinzaines sans que les sous-zones ou les quinzaines constituent en elles mêmes des sources, mais tout simplement parce que ces sources peuvent se trouver à des échelles plus fines, les quartiers ou les jours de pêche par exemple.

Pour répondre à ces questions différentes, il existe des approches différentes par analyse de variance de modèles à "effets fixes" pour ce qui concerne la recherche de variabilité, ou de modèles à "effets aléatoires" pour ce qui concerne la recherche des sources de variabilité.

I ANALYSE DE LA VARIANCE DE MODELES HIERARCHIQUES

Nous nous placerons ici dans le cas de deux sources de variation, avec des effectifs équilibrés, conduisant à des formulations permettant d'exposer les différentes approches de la façon la plus claire possible. On trouvera une présentation claire et rigoureuse dans le livre de P. Dagnélie (1975).

I-1 Le modèle à effets fixes

En nous inspirant de l'exemple halieutique évoqué en introduction, on peut construire le modèle théorique suivant:

$$Y_{ijk} = m + ai + b_{ij} + E_{ijk},$$

Où Y_{ijk} est par exemple la variable aléatoire dont une réalisation est le poids y_{ijk} de poissons dans la $k^{ième}$ pirogue de retour de pêche le $j^{ième}$ jour d'enquêtes de la ;^{ième} quinzaine. Il y a p quinzaines (i=1...p), q jours d'enquêtes dans chaque quinzaine (j=1...q pour tout i) et n pirogues font l'objet d'observations lors de chacun de ces jours (k=1...n pour chaque les combinaison i et j). Les réalisations (observations) s'écrivent quant à elles:

$$y_{ijk} = m + a_i + b_{ij} + e_{ijk},$$

en imposant les contraintes:

$$\sum_i a_i = 0 \text{ et } \sum_j b_{ij} = 0 \text{ pour tout } i,$$

les estimations par les moindres carrés des paramètres du modèles sont les suivantes:

$$\hat{m}=y..., \quad \hat{a}_i=(y_{i}..-y...) \quad et \quad \hat{b}_{ij}=(y_{ij}.-y_{i}..)$$

(les indices sur lesquels on effectue des moyennes sont remplacés par des points, $Y_{i}..$ étant par exemple la moyenne de toutes les observations réalisées lors de la ième quinzaine). On a l'équation d'analyse de la variance suivante:

$$\sum_{i,j,k} (y_{ijk}-y...)^2 =$$

$$qn\sum_i (y_{i}..-y...)^2 + n\sum_{ij} (y_{ij}.-y_{i}..)^2 + \sum_{ijk} (y_{ijk}-y_{ij}.)^2 \qquad soit$$

$$\sum_{ijk} (y_{ijk}-y...)^2 = SCE_a + SCE_{b(a)} + SCE_r,$$

ou en termes de carrés moyens:

$$\sum_{ijk} (y_{ijk}-y...)^2 = (p-1)CM_a + p(q-1)CM_{b(a)} + pq(n-1)CM_r$$

Si les variables E_{ijk} suivent des lois normales centrées indépendante de même variance σ_r^2, on peut tester les hypothèses suivantes:

H_o; $b_{ij} = 0$ pour tout i et j (égalité des moyennes de tous les jours d'une même quinzaine),

H_o'; $a_i = 0$ pour tout i (égalité des moyennes de toutes les quinzaines).

Sous l'hypothèse H_o' la valeur $\frac{CM_a}{CM_r}$ est une réalisation d'une variable aléatoire $F_{p-1,pq(n-1)}$.

Sous l'hypothèse H_o, la valeur $\frac{CM_{b(a)}}{CM_r}$ est une réalisation d'une variable $F_{p(q-1),pq(n-1)}$.

I-2 Le modèle à effets aléatoires

En poursuivant avec le même exemple, nous pouvons écrire le modèle sous une forme traduisant l'origine des sources de variabilité. Le modèle théorique devient:

$$Y_{ijk} = m + A_i + B_{ij} + E_{ijk},$$

où A_i et B_{ij} et E_{ijk} sont cette fois des variables aléatoires centrées indépendantes de variances respectives σ_a^2, σ_b^2 et σ_r^2. Y_{ijk} devient alors une variable aléatoire de moyenne m et de variance $\sigma_a^2+\sigma_b^2+\sigma_r^2$. Lorsque les lois des variables A, B et E sont normales, les trois termes SCE_a, $SCE_{b(a)}$ et SCE_r sommés dans l'équation de l'analyse de la variance sont des réalisations de lois proportionnelles à des khi^2:

$$SCE_a : (qn\sigma_a^2+n\sigma_b^2+\sigma_r^2)khi^2_{p-1}$$

$$SCE_{b(a)} : (n\sigma_b^2+\sigma_r^2)khi^2_{p(q-1)}$$

$$SCEr : (\sigma_r^2)khi^2_{pq(n-1)}.$$

Les espérances des carrés moyens CM_a, $CM_{b(a)}$ et CM_r sont respectivement égales à $qn\sigma_a^2+n\sigma_b^2+\sigma_r^2$, $n\sigma_b^2+\sigma_r^2$ et σ_r^2.

Le test de l'hypothèse H_o" ($\sigma_a^2 = 0$, absence de source de variation au niveau des quinzaines) s'effectue donc maintenant à l'aide de $\frac{CM_a}{CM_{b(a)}}$ qui est une réalisation d'une loi de fisher à p-1 et p(q-1) degrés de liberté multipliée par la valeur $\frac{(qn\sigma_a^2+n\sigma_b^2+\sigma_r^2)}{(n\sigma_b^2+\sigma_r^2)}$ qui est supérieure à 1 dès que σ_a^2 est strictement positif.

L'hypothèse de nullité de σ_b^2 est quant à elle semblable à l'hypothèse H_o du modèle à effet fixe ($b_{ij} = 0$ pour tout i et j) et peut être testée en remarquant que $\frac{CM_{b(a)}}{CM_r}$ est une réalisation d'une loi de Fisher à p(q-1) et pq(n-1) degrés de liberté multipliée par valeur $\frac{n\sigma_b^2+\sigma_r^2}{\sigma_r^2}$ qui est égale à 1 si la variance σ_b^2 est nulle.

Si l'hypothèse H_o s'exprime de façon semblable dans les deux cas, les hypothèses H_o' et H_o" sont différentes et il est possible (et relativement fréquent) que l'égalité des moyennes au niveau supérieur (quinzaines ici) soit rejetée alors que celle de nullité de la variance σ_a^2 ne le soit pas. Ce résultat est normal et provient du fait qu'une source de variabilité à une échelle donnée peut se traduire par une variabilité à un niveau supérieur. Si par exemple il n'y a pas de variabilité au niveau des quinzaines (pouvant provenir par exemple d'une saisonnalité), le fait que des jours peuvent être meilleurs que d'autres fera que des quinzaines seront différentes les unes des autres parce que ne bénéficiant pas du même nombre de bons ou de mauvais jours. Ceci permet de comprendre pourquoi le carré moyen apparaissant au dénominateur du test de l'hypothèse H_o' (quinzaines semblables) fait référence à la variabilité intra jour alors que celui du test de l'hypothèse H_o" (absence de source de variation au niveau de la quinzaine) fait référence à la variabilité intra et inter jours. Lorsque le test de l'hypothèse H_o conduit à la rejeter, on peut

en fait que peu d'intérêt, puisque la question n'est plus de savoir s'il existe des différences entre quinzaines, auxquelles on s'attend, mais de rechercher si ces différences sont accrues par l'existence de sources de variation s'exprimant à ce niveau.

II IMPLICATIONS

La distinction entre variation et source de variabilité est tout à fait importante, aussi bien pour la recherche de cadres descriptifs (modèles) susceptibles d'être utilisés pour la synthèse de l'information disponible sur un système, et l'étude de son comportement, que pour ce qui concerne la présentation des résultats, selon une expression cartographique par exemple.

II-1 Les cadres de description

L'exemple halieutique que nous avons présenté est issu d'un problème rencontré dans le cadre d'une recherche descriptive sur les résultats de pêcheurs artisans au Sénégal (Gérard et Greber 1985). Il convenait, pour évaluer les qualités du système de collecte de l'information, de rechercher à quelles fréquences peuvent s'exprimer les diverses sources de variabilité affectant les rendements de pêche. L'analyse a permis de montrer que la variabilité "inter jours-intra quinzaine" est élevée pour les rendements obtenus pour de nombreuses espèces de poissons, mais il était également important de savoir si les différences entre quinzaines peuvent s'expliquer par cette seule variabilité, ou si elles sont amplifiées par d'autres sources de variations s'exprimant à des fréquences plus faibles. Pour cette étude, les déséquilibres du plan rendaient délicate l'approche par modèle aléatoire, mais la mise en évidence de saisonnalités suffisait pour répondre positivement à la question. Ce type de questions est incontournable lorsqu'on désire étendre le sujet, en s'intéressant aux stratégies d'exploitation des unités de pêche par exemple, et à leur impact sur le système d'exploitation en général. En fonction des rapidités selon lesquelles les pêcheurs décident des changements de tactique pouvant impliquer ou non leur migration, les fréquences auxquelles s'expriment les diverses sources de variabilité affectant les rendements prennent des sens différents, et il doit en être tenu compte au niveau de la recherche d'une synthèse générale des données.

II-2 La présentation des résultats

Nous présenterons ici de façon un peu plus complète le traitement d'une partie du jeu de données sur la couverture vaccinale d'enfants de Pikine, dont l'analyse a fait l'objet d'une publication (Laloë et al 1989). Afin de disposer d'un modèle équilibré, nous n'avons conservé

ici que quatre quartiers, pris au hasard, dans chacune des 7 sous-zones où quatre quartiers au moins avaient été visités. Pour chacun des 28 quartiers concernés on dispose de deux observations, proportions d'enfants entre 0 et 4 ans à jour de leurs vaccination dans deux îlots de 14 parcelles.

Le modèle théorique peut s'écrire de deux façons:

$$Y_{ijk} = m + a_i + b_{ij} + E_{ijk} \quad \textit{(modèle à effets fixes)},$$

$$Y_{ijk} = m + A_i + B_{ij} + E_{ijk}, \quad \textit{(modèle à effets aléatoires)}.$$

i est le numéro de la sous-zone (i=1...7),

j est le numéro du quartier dans la sous-zone (j=1...4),

k est le numéro de l'îlot dans le quartier (k=1,2)

La variable E_{ijk} rend compte de la variabilité des proportions dans les quartiers et de l'erreur d'estimation de la proportion réelle d'enfants vaccinés dans l'îlot "ijk". La variance de E_{ijk} (σ_r^2) dépend donc de la proportion et de l'effectif n_{ijk} d'enfants observés. La variabilité modérée de l'ensemble des quantités $\frac{y_{ijk}(1-y_{ijk})}{n_{ijk}}$ nous permet d'admettre un bonne robustesse du modèle et de considérer que les niveaux de rejet des tests présentés sont des approximations satisfaisantes des niveaux réels.

Les valeurs moyennes de proportions d'enfants à jour de leurs vaccinations sont données dans le tableau ci-dessous:

Quartier / sous-zone	1	2	3	4	Moyenne
1	0.432	0.404	0.442	0.700	0.494
2	0.394	0.317	0.308	0.515	0.383
3	0.364	0.216	0.250	0.321	0.288
4	0.274	0.477	0.290	0.288	0.332
5	0.140	0.726	0.444	0.303	0.403
6	0.423	0.283	0.390	0.263	0.340
7	0.575	0.551	0.308	0.446	0.470

Les carrés moyens obtenus sont:

$$CM_a = 0.045 \quad CM_{b(a)} = 0.036 \quad et\ CM_r = 0.016$$

La valeur de CM_r est une estimation de σ_r^2 et, dans le cas particulier du modèle à effets aléatoires, les estimations de σ_a^2 et σ_b^2 peuvent se déduire des espérances des carrés moyens CM_a et $CM_{b(a)}$:

$\hat{\sigma}_b^2 = \frac{1}{2}(CM_{a(b)} - CM_r) = 0.020/2 = 0.01$ correspondant à un écart-type de 0.1

$\hat{\sigma}_a^2 = \frac{1}{8}(CM_a - CM_{b(a)}) = 0.009/8 = 0.001$ correspondant à un écart type de l'ordre de 0.03

La valeur $\frac{CM_{b(a)}}{CM_r} = 2.23$ permet de rejeter l'hypothèse H_o de nullité des b_{ij} ou de nullité de la variance σ_b^2 avec un risque d'erreur inférieur à 5% (le seuil est de 1.96 pour une loi de Fisher à 21 et 28 ddl).

La valeur $\frac{CM_a}{CM_r} = 2.82$ permet de rejeter l'hypothèse H_o' de nullité des a_i avec un risque d'erreur également inférieur à 5% (le seuil est de 2.45 pour une loi de Fisher à 6 et 28 ddl; il existe des différences entre les proportions d'enfants vaccinés selon les sous-zones).

Mais la valeur $\frac{CM_a}{CM_{b(a)}} = 1.26$ ne permet pas de rejeter l'hypothèse H_o'': $\sigma_a^2 = 0$ (il n'existe pas de source de variabilité significative s'exprimant au niveau de la sous-zone; le seuil de rejet au risque 5% est de 2.60 pour une loi de Fisher à 6 et 21 ddl).

Ces conclusions engendrent une difficulté au niveau de l'expression cartographique des résultats. En effet, à partir des différences de couverture vaccinale selon les quartiers, on pourrait rechercher une présentation de ces quartiers, au moyen de couleurs différentes sur une carte. Mais on ne dispose pas de données dans tous les quartiers, et on aimerait pouvoir affecter chaque quartier à une couleur. Une solution pourrait alors consister en un regroupement des quartiers par sous-zone, permettant alors de présenter une information dans

les quartiers non visités. Mais une telle présentation laisse entendre une homogénéité intra sous-zone ou au moins l'existence d'une source de variabilité s'exprimant à l'échelle de la sous-zone.

Lors de l'étude plus complète des données disponibles, nous avons également envisagé une classification des quartiers en fonction de leur proximité avec des dispensaires, mais cette tentative n'a pas non plus donné de résultat positif.

La conclusion générale que nous avons tirée dans ces conditions a été l'impossibilité d'une cartographie de la couverture vaccinale sur l'ensemble de Pikine, à partir des seules informations disponibles à partir de cette enquête.

CONCLUSION

L'analyse de la variabilité est un domaine vaste dans lequel on s'intéresse à un objet qui peut être abordé selon différents points de vue. L'analyse de la variance, sur laquelle nous avons insisté ici n'est heureusement pas le seul outil permettant des analyses, mais les différentes formulations d'hypothèses présentées permettent une mise en évidence de la variété des questions auxquelles on peut tenter de répondre.

Il s'agit là de différentes façons de qualifier l'information dont on dispose, qualification conduisant à des connaissances sur la distribution des données, et donc à des formulations de problématiques qui, en étant des traductions plus fidèles des questions émises par les choses étudiées, peuvent conduire à des définitions plus satisfaisantes de nos objets d'études, et à une meilleure appréhension de leurs formes.

REFERENCES

COCHRAN W.G. 1977. Sampling techniques. Third edition, J.Wiley and sons, 428 p.

DAGNELIE P. 1975. Théorie et méthodes statistiques, vol. 2. Presses agronomiques de Gembloux, 463 p.

GERARD M. et P. GREBER, 1985. Analyse de la pêche artisanale au Cap-Vert. Description et étude critique du système d'enquête. Doc. Sci. Centre Rech. Océano. Dakar Thiaroye, 98, 77 p.

LALOE F., SALEM G. et C. BENARD, 1989. Définition de sous-zones à risques: dimensions géographiques de la couverture sanitaire à Pikine. In Urbanisation et santé dans le tiers monde, G. Salem et E. Jeannée ed. 471-476. Collection Colloques et Séminaires, ORSTOM, Paris.