Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 3–15
Statistique Canada

Inférences au niveau unitaire à partir de données agrégées

D.G. STEEL, D. HOLT et M. TRANMER[1]

RÉSUMÉ

Les données ne sont souvent disponibles que sous la forme de moyennes de groupes ou de régions. Or, on sait pertinemment qu'une analyse statistique articulée sur les données de ce genre aboutit fréquemment à des résultats très différents de ceux obtenus lorsqu'on analyse les données correspondantes sur les individus ou les ménages. Croire que les résultats d'une analyse de niveau régional sont applicables au niveau individuel, c'est risquer de commettre l'erreur écologique. Les effets de l'agrégation ou les effets écologiques résultent en partie du fait qu'une région n'est pas constituée d'un assemblage aléatoire d'êtres humains ou de ménages. Les paramètres socioéconomiques varient considérablement d'une région à l'autre. On doit intégrer la structure de la population au modèle statistique utilisé pour l'analyse si on veut bien saisir les conséquences de l'agrégation. Les auteurs proposent un modèle général simple pour y parvenir et décrivent l'effet de ce modèle sur l'estimation des moyennes et des matrices des covariances de la population. Par ailleurs, ils montrent comment obtenir une estimation non biaisée des paramètres au niveau de l'individu à partir des données agrégées, de façon à éviter l'erreur écologique. Les méthodes qu'ils préconisent supposent l'identification des «variables de groupement» qui caractérisent le processus qui a mené à la structure de la population ou, du moins, les différences entre régions. On doit pour cela trouver une estimation de la matrice des covariances pour les variables d'agrégation, au niveau unitaire, d'une source quelconque. L'analyse des données du recensement de 1991 du Royaume-Uni a permis d'identifier les principales variables d'agrégation et de mesurer l'efficacité des méthodes de correction envisagées pour estimer les matrices des covariances et les coefficients de corrélation. Les résultats de ces travaux concourent à l'élaboration d'une stratégie pour l'analyse des données agrégées.

MOTS CLÉS: Agrégation; erreur écologique; groupement; sélection; composantes de la variance.

1. INTRODUCTION

Les chercheurs qui souhaitent étudier les relations au niveau de l'individu éprouvent souvent des difficultés parce qu'ils doivent utiliser des données agrégées, par exemple une moyenne ou un total régional. Idéalement, on devrait se servir des données au niveau de l'unité recueillies lors du sondage ou du recensement, mais ces données sont inaccessibles parce qu'il faut en préserver la confidentialité ou parce qu'elles ne viennent pas d'une enquête ou d'un recensement récents. Les régimes administratifs nous procurent des renseignements sur diverses variables, par exemple le chômage, la santé et la morbidité. Malheureusement, ces données sont fournies habituellement pour des agrégats, notamment par région, toujours pour des raisons de confidentialité. Le recensement fournit lui aussi des données régionales. C'est pourquoi les recherches sociales et épidémiologiques portent encore couramment sur des données collectives.

Soit une population au sein de laquelle chaque sujet est associé à un vecteur de variables intéressantes, dont la répartition a pour moyenne μ_y et pour matrice des covariances Σ_{yy}. Nous aimerions savoir quels liens existent entre ces variables, comme les représentent les corrélations, les coefficients de régression et les grandes composantes qu'on peut tirer de la matrice des covariances Σ_{yy}, sujet principal de nos inférences. Ces variables pourraient comprendre, par exemple, une batterie de tests d'accomplissement dans le cadre d'une étude sur l'éducation, l'incidence d'une maladie donnée et une suite de variables explicatives dans le contexte d'une étude épidémiologique, voire une série de mesures de privation dans l'optique d'une étude sociologique. Nous supposons ici qu'on ne peut obtenir les données au niveau de l'individu. Néanmoins, la région peut être subdivisée en plus petits ensembles comme des districts de recensement (DR) et pour chacun de ces sous-ensembles g ou pour un échantillon de régions, on peut calculer le vecteur des valeurs moyennes $\bar{y}_g$ pour les variables étudiées de l'échantillon n_g d'où sont extraites ces dernières.

Le but de l'analyse est d'obtenir une matrice des covariances Σ_{yy} couvrant ces petites régions. L'inférence n'est pas conditionnelle à l'appartenance à une petite région, mais se rapporte à la distribution marginale entre ces régions. La situation est différente avec l'estimation d'une petite région. Dans ce cas, l'inférence vise la distribution conditionnelle dans la région en question. Pareil objectif est tout à fait légitime, mais diffère de celui qui nous intéresse. Il se peut qu'on puisse recourir aux mêmes modèles, mais l'inférence n'a pas la même visée. Notre formulation nous permet néanmoins d'ajouter les variables spécifiques à certains groupes aux variables intéressantes, s'il y a lieu. Ainsi, si on associe à chaque individu un ensemble de moyennes DR pour la région où il se trouve, il est alors possible de les inclure au vecteur $\boldsymbol{y}$ qui nous intéresse. Pareille approche se prête aux analyses de régression qui utilisent les moyennes des petites régions comme variables explicatives.

[1] D.G. Steel, Department of Applied Statistics, University of Wollongong, NSW 2522, Australia; D. Holt et M. Tranmer, Department of Social Statistics, University of Southampton, S017 1BJ, United Kingdom.

Les premiers travaux sur l'analyse des données agrégées remontent à Gehlke et Biehl (1934). D'éminents chercheurs comme Yule et Kendall (1950), Robinson (1950), Blalock (1964), Openshaw et Taylor (1979) et, plus récemment, Arbia (1989) ont sensiblement contribué à l'étude du problème. Le fait que les unités régionales utilisées présentent rarement une importance particulière (elles sont construites pour des raisons d'économie, de facilité ou de commodité administrative) soulève aussi des difficultés. Par ailleurs, les résultats de l'analyse effectuée au niveau du groupe dépendent de l'échelle des unités, soit de leur taille moyenne et du jeu de limites retenu. Plusieurs études empiriques ont fait ressortir ces aspects, notamment celles de Clark et Avery (1976), de Perle (1977), d'Openshaw (1984) et de Fotheringham et Wong (1991). Les travaux n'ont malheureusement pas débouché sur une théorie permettant une application générale, ni sur des méthodes pratiques en vertu desquelles on pourrait tirer des inférences fiables de niveau unitaire à partir des résultats des analyses effectuées au niveau du groupe.

On attribue les effets de l'agrégation au fait que la population des unités géographiques ne s'est pas constituée au hasard. En règle générale, les individus d'une même région ont tendance à se ressembler davantage, car ils n'ont pas choisi de vivre là au hasard ou parce qu'ils s'y trouvent consécutivement à l'action de forces analogues, voire parce qu'ils ont des échanges mutuels. Il existe donc des différences socioéconomiques entre les régions, différences qui se confondent avec les effets individuels dans l'analyse statistique des données agrégées sur les régions concernées. Nous proposons d'abord un modèle général simple qui s'efforce d'intégrer ces effets, puis nous examinons les conséquences de son utilisation et les implications d'une telle méthode pour l'analyse au niveau régional. Par ailleurs, nous suggérons des méthodes qui, dans certaines circonstances, donneront une estimation non biaisée des paramètres de niveau individuel à partir des données agrégées, donc permettront d'éviter l'erreur écologique. Ces méthodes font intervenir des variables auxiliaires pour lesquelles certaines sources donnent une matrice des covariances de niveau unitaire pour l'échantillon. Cette approche a été appliquée aux données du recensement de 1991 du Royaume-Uni et une stratégie a été développée en vue de l'analyse des données agrégées.

2. MODÈLES POUR LES EFFETS RÉGIONAUX

Soit une population de N sujets ayant chacun un vecteur y de caractéristiques à étudier. La population se compose de M groupes et la variable aléatoire c_i indique la région à laquelle la i-ème unité de population appartient. On dénombre N_g individus dans la g-ième région.

Supposons que μ_y et Σ_{yy} soient les paramètres de la superpopulation. Compte tenu de ces prémisses, on obtient la théorie statistique qui suit. Quelques aspects relatifs au plan d'échantillonnage seront toutefois examinés à la fin de la partie 2.

Nous partons de l'hypothèse qu'il existe un ensemble de données d'échantillon s de taille n et que ces données individuelles ont été agrégées pour former un ensemble de m moyennes régionales utilisables pour l'analyse. Il est donc possible de calculer les statistiques suivantes au niveau régional:

moyenne de la g-ième région:

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \in g,s} y_i \tag{2.1}$$

moyenne totale de l'échantillon:

$$\bar{y} = \frac{1}{n} \sum_{g \in s} n_g \bar{y}_g = \frac{1}{n} \sum_{i \in s} y_i \tag{2.2}$$

matrice des covariances de l'échantillon au niveau régional:

$$\bar{S}_{yy} = \frac{1}{m-1} \sum_{g \in s} n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'. \tag{2.3}$$

On peut définir des statistiques analogues au niveau unitaire, mais l'analyste ne pourra s'en servir. Par exemple, $S_{yy} = 1/(n-1) \sum_{i \in s} (y_i - \bar{y})(y_i - \bar{y})'$ est la matrice des covariances de l'échantillon au niveau unitaire.

2.1 Groupement aléatoire

Bien que les groupes géographiques soient rarement le fait du hasard, pareille hypothèse constitue un bon point de départ quand on s'intéresse à l'analyse écologique. Si les groupes se forment de façon aléatoire, maintes analyses effectuées au niveau du groupe seront valables, même si leur utilité est réduite. Steel et Holt (1995) étudient les propriétés de diverses statistiques comme la moyenne, la variance et les coefficients de régression et de corrélation dans une situation de ce genre. Quand les groupes se constituent au hasard, c'est-à-dire $y \perp c$, alors

$$E[\bar{y}_g \mid s,c] = \mu_y \tag{2.4}$$

$$V(\bar{y}_g \mid s,c) = \frac{1}{n_g} \Sigma_{yy}. \tag{2.5}$$

Les propriétés fondamentales des statistiques au niveau unitaire et au niveau du groupe sont faciles à déduire:

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s,c) = \mathbf{0} \quad g \neq h \tag{2.6}$$

$$E[\bar{y} \mid s,c] = \mu_y \tag{2.7}$$

$$E[S_{yy} \mid s,c] = \Sigma_{yy} \tag{2.8}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy}. \tag{2.9}$$

Ces propriétés s'appliquent si on peut négliger l'échantillonnage étant donné les paramètres discriminants du groupe, bref le plan d'échantillonnage peut dépendre des groupes mais pas de y ou d'une variable quelconque associée à y, sous réserve de c. Par exemple, on pourrait utiliser le recensement ou un échantillon aléatoire simple de groupes et d'unités appartenant à ces groupes.

On peut recourir à des statistiques non pondérées au niveau du groupe en posant $n_g = 1$ dans les équations (2.2) et (2.3). On obtient ainsi des estimateurs inefficaces. Le degré d'inefficacité dépendra de la distribution des groupes d'échantillon des groupes. La pondération selon les tailles d'échantillon des groupes est importante. Cela fait, on peut procéder aux inférences habituelles en apportant les corrections appropriées aux degrés de liberté. La variabilité dépend du nombre de régions plutôt que du nombre d'observations et on adapte les intervalles de confiance et les épreuves en conséquence.

2.2 Modèle des composantes de la variance

Une façon simple d'illustrer la corrélation positive entre les membres d'un groupe normalement observée dans une population consiste à utiliser un modèle des composantes de la variance, ce qui, dans le cas d'une analyse à plusieurs variables, correspond à

$$y_i = \mu_y + v_g + \epsilon_i \qquad i \in g$$

où v_g et ϵ_i sont des composantes aléatoires indépendantes au niveau du groupe et au niveau de l'individu, respectivement, avec une espérance mathématique nulle, $V(\epsilon_i \mid c) = \Sigma_{\epsilon\epsilon}$ et $V(v_g \mid c) = \Delta_{yy}$.

Modèle A:

$$E[y_i \mid c] = \mu_y \tag{2.10}$$

$$V(y_i \mid c) = \Sigma_{\epsilon\epsilon} + \Delta_{yy} = \Sigma_{yy} \tag{2.11}$$

$$\begin{aligned} \mathrm{Cov}(y_i, y_j \mid c) &= \Delta_{yy} \quad \text{si} \quad c_i = c_j \quad i \neq j \\ &= \mathbf{0} \text{ sinon.} \end{aligned} \tag{2.12}$$

La notation $V(\cdot \mid c)$ signifie que la matrice des covariances dépend de l'étiquette c du groupe, donc détermine l'appartenance à un groupe commun. On estime cependant qu'elle est inconditionnelle pour les effets aléatoires au niveau du groupe. Par conséquent, $V(y_i \mid c)$ inclut la variance totale, à la fois pour la matrice des covariances de groupe $\Sigma_{\epsilon\epsilon}$ et la matrice des covariances de groupe Δ_{yy}.

On obtient facilement les propriétés des moyennes de l'échantillon au niveau du groupe à partir du modèle A si on peut négliger l'échantillonnage étant donné c,

$$E[\bar{y}_g \mid s, c] = \mu_y \tag{2.13}$$

$$V(\bar{y}_g \mid s, c) = \frac{1}{n_g}(\Sigma_{yy} + (n_g - 1)\Delta_{yy}) \tag{2.14}$$

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s, c) = \mathbf{0} \quad g \neq h. \tag{2.15}$$

Les propriétés des statistiques au niveau unitaire et au niveau du groupe sont les suivantes:

$$E[\bar{y} \mid s, c] = \mu_y \tag{2.16}$$

$$E[S_{yy} \mid s, c] = \Sigma_{yy} - \frac{\bar{n}^0 - 1}{n - 1}\Delta_{yy} \tag{2.17}$$

$$E[\bar{S}_{yy} \mid s, c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy} \tag{2.18}$$

où $\bar{n} = n/m$, $\bar{n}^0 = 1/n \sum_{g \in s} n_g^2 = \bar{n}(1 + C_n^2)$, $\bar{n}^* = \bar{n}(1 - C_n^2/(m-1))$ et $C_n^2 = 1/m \sum_{g \in s}(n_g - \bar{n})^2/\bar{n}^2$ est le carré du coefficient de variation des tailles de groupe de l'échantillon. Notons que le coefficient de Δ_{yy} est $0(m^{-1})$ dans (2.17) et $0(\bar{n})$ dans (2.18), ce qui montre bien comment une petite erreur d'analyse au niveau unitaire prend de l'ampleur après agrégation. Nous approfondirons ces résultats à la partie 2.4.

2.3 Modèles de groupement

Les chercheurs qui s'intéressent à l'analyse écologique ont échafaudé des modèles qui tiennent compte du mécanisme de création des groupes. Ils supposent qu'un processus de synthèse attribue les unités à tel ou tel groupe, selon un vecteur de variables de groupement z_i, de façon stochastique ou déterministe. Blalock (1964) recourt implicitement à cette approche dans son analyse, tandis qu'Hannan et Burstein (1974), Litchman (1974), Langbein et Litchman (1978), Smith (1977) et Blalock (1979, 1985) s'en servent de façon explicite. Steel (1985) parle de modèles de groupement, car il suppose que les groupes naissent au terme d'un processus quelconque qui fait intervenir les variables au niveau des relations à l'étude. La formation du groupe est perçue comme une distorsion, aussi les relations intéressantes sont-elles définies auparavant. On précise souvent dans la discussion sur les modèles contextuels que les effets contextuels apparents peuvent résulter de tels facteurs. La version multivariée du modèle est la suivante:

Modèle B:

$$E[y_i \mid z, c] = \mu_{y.z} + \beta'_{yz} z_i \tag{2.19}$$

$$V(y_i \mid z, c) = \Sigma_{yy.z} \tag{2.20}$$

$$\mathrm{Cov}(y_i, y_j \mid z, c) = \mathbf{0} \quad i \neq j. \tag{2.21}$$

Dans ce modèle, l'espérance conditionnelle de y_i dépend seulement de la valeur des variables auxiliaires de la i-ième unité et est indépendante du groupe à laquelle l'unité appartient ou de la valeur des variables auxiliaires des autres unités de la population. La covariance conditionnelle entre deux unités quelconques est zéro. Ce modèle est valable pour les modèles de groupement dans lesquels le mécanisme de

formation du groupe se caractérise par les variables auxiliaires z_i. Les variables auxiliaires peuvent être considérées comme des variables déterminant à quel groupe appartient telle ou telle unité. Sous un angle plus général, les variables auxiliaires correspondent aux principales variables de niveau individuel qui ne sont pas distribuées au hasard ni entre les groupes à cause des processus de sélection ou de migration auxquels la population est assujettie. On peut aussi inclure au modèle les variables contextuelles sous la forme de variables auxiliaires de valeur identique pour toutes les unités du groupe.

Si le vecteur des variables auxiliaires a une distribution marginale dont la moyenne est μ_z et la matrice des covariances, Σ_{zz} la moyenne marginale et la matrice des covariances de y seront respectivement $\mu_y = \mu_{y.z} + \beta'_{yz}\,\mu_z$ et $\Sigma_{yy} = \Sigma_{yy.z} + \beta'_{yz}\,\Sigma_{zz}\,\beta_{yz}$. Les propriétés des moyennes de l'échantillon au niveau du groupe sont faciles à tirer du modèle B:

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.22}$$

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}\,\Sigma_{yy.z} \tag{2.23}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = \mathbf{0} \qquad g \neq h. \tag{2.24}$$

Les statistiques au niveau du groupe auront donc les propriétés qui suivent:

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.25}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.26}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.27}$$

où S_{zz} et $\bar{S}_{zz}$ sont définis de façon analogue à S_{yy} et $\bar{S}_{yy}$ de l'équation (2.3) et de la phrase suivant cette dernière.

2.4 Modèle combiné

Jusqu'à présent, on peut dire que les deux modèles examinés expliquent les effets de groupe chacun de leur côté. Il est possible de les combiner en un modèle plus réaliste qui intégrera les deux effets de groupe et les composantes résiduelles de la variance:

Modèle C:

$$E[y_i \mid z,c] = \mu_{y.z} + \beta'_{yz}\,z_i \tag{2.28}$$

$$V(y_i \mid z,c) = \Sigma_{yy.z} \tag{2.29}$$

$$\begin{aligned}\mathrm{Cov}(y_i,y_j \mid z,c) &= \Delta_{yy.z} \quad \text{si} \quad c_i = c_j \quad i \neq j \\ &= 0 \quad \text{sinon}.\end{aligned} \tag{2.30}$$

Le nouveau modèle accepte les mécanismes de création des groupes que caractérisent les variables auxiliaires z_i. On y retrouve aussi les corrélations résiduelles à l'intérieur d'un groupe qui résultent des effets aléatoires attribués aux variables aléatoires inconnues du niveau du groupe, après exclusion des variables de groupement.

Voici les propriétés des moyennes de l'échantillon au niveau du groupe provenant du modèle C, lorsqu'on néglige l'échantillonnage, étant donné (z,c),

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.31}$$

et

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}\,(\Sigma_{yy.z} + (n_g - 1)\Delta_{yy.z}) \tag{2.32}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = \mathbf{0} \qquad g \neq h \tag{2.33}$$

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.34}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} - \frac{\bar{n}^0 - 1}{n - 1}\,\Delta_{yy.z} \tag{2.35}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z}. \tag{2.36}$$

Les équations (2.17) et (2.18) montrent comment l'agrégation amplifie les effets aléatoires au niveau du groupe dans le modèle des composantes de la variance A. Dans l'équation (2.17), le coefficient de Δ_{yy} est $0(m^{-1})$, tandis que dans (2.18), il correspond à $0(\bar{n})$. Avec le modèle combiné C, les équations (2.35) et (2.36) montrent comment l'inclusion des variables de groupement permet de scinder le biais en deux éléments additifs: le premier associé aux variables de groupement, aux liens de ces dernières avec les variables d'intérêt et à leur effet cumulatif, et le second faisant intervenir $\Delta_{yy.z}$, soit les composantes résiduelles de la variance, après prise en compte des variables de groupement. Notons que les coefficients de $\Delta_{yy.z}$ dans les équations (2.35) et (2.36) restent $0(m^{-1})$ et $0(\bar{n})$ respectivement, comme aux équations (2.17) et (2.18). Les composantes résiduelles de la variance devraient néanmoins être plus faibles, en général. L'équation (2.29) part de l'hypothèse que la variance résiduelle de c est constante.

L'hypothèse que l'échantillonnage est négligeable pour (z,c) signifie que le plan d'échantillonnage peut dépendre des variables auxiliaires et des paramètres discriminants du groupe. On peut donc effectuer une stratification selon z et procéder à un échantillonnage en grappes ou à plusieurs degrés en fonction des groupes.

La matrice $\bar{S}_{yy}$ pondérée au niveau du groupe permet d'estimer Σ_{yy}. Le premier terme de biais de (2.36) vient de l'effet des variables de groupement. Il est égal à zéro si $\beta_{yz} = \mathbf{0}$, ou presque égal à zéro si $\bar{S}_{zz} \doteqdot \Sigma_{zz}$. La condition $\beta_{yz} = \mathbf{0}$ est ferme et signifie que les variables intéressantes n'ont aucun lien avec les variables de groupement. L'effet de l'agrégation sur la covariance de l'échantillon pour deux variables quelconques dépend des liens des deux variables en question avec les variables du groupement z_i. On devrait s'attendre à ce que les effets cumulatifs soient plus importants quand elles sont plus étroitement liées aux variables de groupement. En raison de la condition $\bar{S}_{zz} \doteqdot \Sigma_{zz}$ les variables z échappent aux effets de la sélection et de l'agrégation. Néanmoins, il est peu probable que ces conditions s'appliquent dans la réalité, de sorte que beaucoup de variables auront un biais. L'erreur d'échantillonnage et de groupement attribuable aux variables auxiliaires est mesurée par $S_{zz} - \Sigma_{zz}$ pour l'estimateur de niveau unitaire, et par $\bar{S}_{zz} - \Sigma_{zz}$ pour l'estimateur au niveau du groupe. L'expression $\bar{S}_{zz} - \Sigma_{zz}$ traduit l'effet net de l'échantillonnage et de l'agrégation sur les variables auxiliaires.

Le deuxième terme de biais (2.36) est nul si $\Delta_{yy.z} = \mathbf{0}$. Il n'existe donc aucune corrélation résiduelle à l'intérieur du groupe pour les variables y, sous réserve des variables de groupement. La chose est peu probable en pratique, mais il est souhaitable d'identifier les variables de groupement qui intègrent le plus possible les effets cumulatifs en rendant le terme résiduel le plus faible possible.

Les effets du groupement et de l'échantillonnage selon z et l'effet attribuable à la corrélation résiduelle à l'intérieur du groupe s'additionnent; c'est ce qui se produit avec les types de corrélations intérieures au groupe plus complexes, pourvu que la linéarité du modèle soit préservée. Si z suit un simple modèle des composantes de la variable comme le modèle A,

$$E[\bar{S}_{zz} \mid s,c] = \Sigma_{zz} + (\bar{n}^* - 1)\Delta_{zz}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy} + (\bar{n}^* - 1)\beta'_{yz}\,\Delta_{zz}\,\beta_{yz} + \Delta_{yy.z} \tag{2.37}$$

et les covariances des variables intéressantes à l'intérieur du groupe seront constituées de l'élément résultant des covariances des variables auxiliaires à l'intérieur du groupe et des composantes résiduelles. Le côté droit de l'équation (2.37) est tiré de (2.18) puisque z suit sans conditions le modèle des composantes de la variance, si y en fait autant. Le modèle fondamental a pour but de trouver les variables auxiliaires qui permettront de réduire au maximum les covariances résiduelles ou conditionnelles $\Delta_{yy.z}$ à l'intérieur d'un groupe, ou de les éliminer, dans le meilleur des cas.

2.5 Correction des effets cumulatifs

Peu de propositions utiles ont été avancées sur la manière dont les analyses de niveau régional pourraient être corrigées afin de donner une estimation raisonnable des relations au niveau unitaire. Duncan et Davis (1953) ont examiné la fourchette éventuelle des coefficients de corrélation au moyen d'un tableau carré (2×2) dont les marges étaient connues. Les limites résultantes sont souvent trop importantes pour présenter une utilité pratique. De son côté, Goodman (1959) a énoncé les conditions précises dans lesquelles l'analyse écologique pourrait servir à tirer des inférences sur les relations de niveau individuel avec un modèle de régression. Langbein et Litchman (1978) ont examiné quelques méthodes applicables aux groupements articulés sur les variables dépendantes, quand on connaît les variances de niveau unitaire des variables dépendantes et indépendantes du modèle de régression. Malheureusement, aucune de ces approches n'est assez générale pour permettre la résolution du problème.

Si on examine le biais pour $\bar{S}_{yy}$ dans (2.36), on constate qu'en ajoutant $\beta'_{yz}(\Sigma_{zz} - \bar{S}_{zz})\beta_{yz}$ à $\bar{S}_{yy}$, le terme de biais attribuable aux variables de groupement disparaît. L'équation (2.31) implique que

$$E[\bar{B}_{yz} \mid s,z,c] = \beta_{yz} \tag{2.38}$$

où $\bar{B}_{yz} = \bar{S}_{zz}^{-1}\,\bar{S}_{zy}$.

Si on disposait de la matrice des covariances S_{zzs_0} de z, pour l'échantillon unitaire s_0 tiré de m_0 groupes, l'estimateur corrigé

$$\hat{\Sigma}_{yy}(z) = \bar{S}_{yy} + \bar{B}'_{yz}(S_{zzs_0} - \bar{S}_{zz})\bar{B}_{yz} \tag{2.39}$$

éliminerait le biais d'agrégation résultant des variables de groupement z, pourvu que S_{zzs_0} s'approche de Σ_{zz}. Il se peut que la source de S_{zzs_0} soit largement indépendante des données utilisées dans $\bar{S}_{yy}$ et $\bar{B}_{yz}$. Steel (1985) montre que l'estimateur corrigé (2.39) peut correspondre à l'estimateur du maximum de vraisemblance de Σ_{yy} (après substitution de $m - 1$ par m, *etc.*, comme on le fait d'habitude). Si la normalité de la répartition de (y,z) s'applique, s_0 devient un simple échantillon aléatoire de la population et $\Delta_{yy.z} = \mathbf{0}$. L'estimateur corrigé correspond à l'ajustement de Pearson (1903) que Holt, Smith et Winter (1980) envisagent dans leur analyse de régression et que Smith et Holmes (1989) utilisent dans leur analyse à plusieurs variables. Dans les deux cas, on corrige les statistiques obtenues à partir des données de niveau unitaire issues d'un échantillon prélevé en fonction des variables auxiliaires. En ce qui nous concerne, la correction est appliquée aux statistiques tirées des moyennes des régions, et les variables auxiliaires utilisées pour effectuer la correction comprennent les variables de groupement ainsi que les variables du plan d'échantillonnage. L'estimateur corrigé de μ_y est

$$\hat{\mu}_y(z) = \bar{y} + \bar{B}'_{yz}(\bar{z}_{s_0} - \bar{z}) \tag{2.40}$$

où $\bar{z}_{s_0}$ est la moyenne de s_0.

D'après (2.34) et (2.38), on constate que

$$E[\hat{\mu}_y(z) \mid s,z,s_0,c] = \mu_y + \beta'_{yz}(\bar{z}_{s_0} - \mu_z). \tag{2.41}$$

Par ailleurs, selon Steel (1985), (2.36) et (2.38) impliquent que

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,s_0,c] = \Sigma_{yy} + \beta'_{yz}(S_{zzs_0} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}) \quad (2.42)$$

pourvu que $\mathrm{tr}(\bar{S}_{zz}^{-1} S_{zzs_0})$ et $\bar{n}\,\mathrm{tr}\big((\bar{S}_{zz}^{-1} S_{zzs_0} - I)\bar{S}_{zz}^{-1} \bar{S}_{zz}^{(2)}\big)$ soient bornés, $\bar{S}_{zz}^{(2)}$ étant défini de la même façon que $\bar{S}_{zz}$, sauf que n_g est remplacé par $n_g^2/\bar{n}$.

Quand on compare (2.42) à (2.35), on se rend compte que la composante du biais attribuable aux variables de groupement a été ramenée au biais lié à l'utilisation de S_{yys_0} si elle est disponible. L'estimateur corrige les effets cumulatifs résultant de z et ramène l'effet du plan d'échantillonnage associé à s à celui lié à s_0.

Supposons que le plan d'échantillonnage utilisé pour produire s_0 et les valeurs des variables auxiliaires viennent d'une superpopulation, de telle sorte que

$$E[\bar{z}_{s_0} \mid s_0,c] = \mu_z + 0(m_0^{-1}) \quad (2.43)$$

$$E[S_{zzs_0} \mid s_0,c] = \Sigma_{zz} + 0(m_0^{-1}) \quad (2.44)$$

où m_0 représente le nombre de groupes dans s_0.

Alors,

$$E[\hat{\mu}_y(z) \mid s,s_0,c] = \mu_y + 0(m_0^{-1}) \quad (2.45)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,s_0,c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(\tilde{m}^{-1}) \quad (2.46)$$

où

$$\tilde{m} = \min(m,m_0).$$

Les conditions (2.43) et (2.44) s'appliquent si les valeurs de groupe de la population z viennent d'un modèle dont les composantes de la variance sont similaires à celles du modèle A et si le plan d'échantillonnage de s_0 ne dépend que du groupement, pas des variables auxiliaires. L'échantillonnage aléatoire simple, l'échantillonnage par grappes à probabilité de sélection identique, voire l'échantillonnage à plusieurs degrés respectent cette condition. On pourrait aussi se servir des données du recensement pour que s_0 constitue la population finie dans son ensemble.

Le biais attribuable aux variables de groupement peut donc être corrigé, pourvu qu'on dispose d'une matrice des covariances quelconque pour z au niveau unitaire. La raison pour laquelle on recourt à une telle approche est de créer une situation où les effets prédominants du groupe seraient attribués à la sélectivité ou au groupement, par le truchement des variables de groupement. La correction relative aux variables auxiliaires supprime l'effet de la corrélation apparente à l'intérieur du groupe associée à ces variables. Même corrigé cependant, l'estimateur inclut un élément de biais en raison de $\Delta_{yy.z}$ et si z n'atténue pas assez les corrélations à l'intérieur du groupe, le biais demeure appréciable. Cette approche dépend donc du choix des variables auxiliaires qui réduiront les corrélations à l'intérieur du groupe.

Si le plan d'échantillonnage de s_0 et le modèle de superpopulation de z interdisent l'application de (2.43) et de (2.44), on peut remplacer $\bar{z}_{s_0}$ et S_{zzs_0} par les estimateurs $\hat{\mu}_{zs_0}$ et $\hat{\Sigma}_{zzs_0}$ pour obtenir des estimateurs corrigés $\hat{\mu}_y(z)$ et $\hat{\Sigma}_{yy}(z)$. L'espérance mathématique des estimateurs corrigés est donnée par (2.41) et (2.42), où il suffit de remplacer $\bar{z}_{s_0}$ par $\hat{\mu}_{zs_0}$ et S_{zzs_0} par $\hat{\Sigma}_{zzs_0}$. Plusieurs possibilités peuvent être retenues pour $\hat{\mu}_{zs_0}$ et $\hat{\Sigma}_{zzs_0}$ à partir de l'échantillon s_0. Smith et Holmes (1989) ont examiné toute une gamme d'estimateurs envisageables, en fonction du modèle et du plan d'échantillonnage. Supposons, par exemple, que le plan d'échantillonnage utilisé pour obtenir s_0 comporte une stratification d'après les valeurs du vecteur des variables de taille $\boldsymbol{x}$. Appelons la probabilité d'inclusion de l'unité de population i dans l'échantillon, Π_i. Le poids associé à cette probabilité serait $w_i = (\Pi_i)^{-1}$. L'estimateur de μ_z pondéré pour la probabilité est $\bar{z}_{s_0}^* = \sum_{i \in s_0} w_i z_i$, et celui de Σ_{zz} est $S_{zzs_0} = \sum_{i \in s_0} w_i z_i z_i' - w_0^{-1} \bar{z}_{s_0}^* \bar{z}_{s_0}'^*$ où $w_0 = \sum_{i \in s_0} w_i$.

Les estimateurs corrigés de Pearson pour μ_z et Σ_{zz} sont $\bar{z}_{s_0} + \boldsymbol{B}'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0})$ et $S_{zzs_0} + \boldsymbol{B}'_{zxs_0}(S_{xxu} - S_{xxs_0})\boldsymbol{B}_{zxs_0}$ respectivement. Dans ce cas, $\bar{x}_u$ et S_{xxu} correspondent au vecteur des moyennes et à la matrice des covariances des variables du plan d'échantillonnage de $\boldsymbol{x}$ pour la population finie, et $\boldsymbol{B}_{zxs_0} = S_{xxs_0}^{-1} S_{xzs_0}$.

On pourrait aussi utiliser les estimateurs de Pearson corrigés et pondérés pour la probabilité, à savoir $\bar{z}_{s_0}^* + \boldsymbol{B}'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0}^*)$ et $S_{zzs_0}^* + \boldsymbol{B}'_{zxs_0}(S_{xxu} - S_{xxs_0}^*)\boldsymbol{B}_{zxs_0}^*$.

Ici $\bar{x}_{s_0}^*$ et $S_{xxs_0}^*$ correspondent respectivement à $\bar{z}_{s_0}^*$ et $S_{zzs_0}^*$ et où $\boldsymbol{B}_{zxs_0}^* = S_{xxs_0}^{*-1} S_{xzs_0}^*$. Puisque, jusqu'à présent, l'approche repose essentiellement sur l'utilisation d'un modèle, il serait préférable de prendre les estimateurs μ_y et Σ_{zz}. En pratique néanmoins, les données dont on pourrait se servir pour la correction comprendraient les estimateurs connus de la moyenne et de la covariance, pondérés pour la probabilité et tirés de l'échantillon s_0, qui est indépendant de s. Par conséquent,

$$E_{p_o}[\hat{\mu}_{zs_0} \mid z,c] = \bar{z}_u$$

$$E_{p_o}[\hat{\Sigma}_{zzs_0} \mid z,c] = S_{zzu}$$

où $\bar{z}_u$ et S_{zzu} correspondent au vecteur des moyennes et à la matrice des covariances des variables auxiliaires de la population finie, et où E_{p_o} représente l'espérance mathématique d'un échantillon répété avec le plan d'échantillonnage utilisé pour obtenir s_0, c'est-à-dire la distribution de randomisation. De (2.41) et (2.42), on peut déduire que

$$E[\hat{\mu}_y(z) \mid s,z,c] = \mu_y + \beta_{yz'}(\bar{z}_u - \mu_z)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,c] = \Sigma_{yy} + \beta_{yz'}(S_{zzu} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}).$$

Ces espérances mathématiques sont reprises dans le modèle statistique qui génère les valeurs $\boldsymbol{y}$ et la répartition par randomisation associée à s_0. En réalité $\bar{z}_u$ et S_{zzu} se rapprochent beaucoup de μ_z et Σ_{zz} respectivement.

3. IDENTIFICATION DES VARIABLES DE GROUPEMENT

Dans la partie qui précède, nous avons présenté un ensemble de variables auxiliaires z caractérisant les variations régionales. Il a servi à corriger l'analyse agrégée et à réduire le biais attribuable à l'agrégation. Avec des variables auxiliaires idéales, $\Delta_{yy.z}$ serait égal à zéro et la méthode de correction supprimerait totalement le biais dû à l'agrégation. En pratique cependant, on ignore quelles variables auxiliaires permettraient d'obtenir $\Delta_{yy.z} = \mathbf{0}$. C'est pourquoi nous devons nous en tenir à des variables pour lesquelles on connaît la moyenne régionale, compte tenu des données analysées, et pour lesquelles il est possible d'estimer la matrice des covariances au niveau unitaire $\hat{\Sigma}_{zz}$. On pourra recourir notamment aux données démographiques et aux variables sur l'habitation de base que fournit couramment le recensement. Toutefois, il se peut que ces variables ne caractérisent pas exactement le processus de groupement, donc n'expliquent pas les différences entre régions aussi bien qu'on le souhaiterait.

3.1 Stratégie d'analyse

Dans la pratique, les variables de groupement ne sont pas connues. Il faut donc élaborer une stratégie en vue d'identifier les variables d'ajustement pour lesquelles on possède une estimation de la matrice des covariances au niveau unitaire, et qui expliqueront les effets de groupement. Voici une façon d'y arriver:

1) Identifier un ensemble de variables couvrant le même domaine que les variables auxquelles on s'intéresse, mais pour lesquelles il existe des données de niveau régional et de niveau unitaire pour une période quelconque du passé (un recensement antérieur, par exemple).

2) Ajouter aux variables qui précèdent d'autres variables (variables démographiques et variables sur l'habitation, par exemple) susceptibles de remplacer les variables z qui présentent une étroite association avec les variations régionales. On aura aussi besoin d'une estimation des matrices des covariances de niveau régional et de niveau unitaire pour la même période.

3) Analyser les données pour déterminer les variables qui expliquent le mieux les effets de niveau régional sur les variables d'intérêt. Nous reviendrons plus tard à cette analyse, baptisée analyse des VGC.

4) De (3), trouver la série de variables de correction qu'on pourrait extraire de l'ensemble de données courant et pour lesquelles on pourrait obtenir la matrice des covariances courante de niveau unitaire d'une source quelconque.

5) Il est possible qu'on découvre une estimation de la variance ou de la covariance de niveau unitaire pour certaines variables qui nous intéressent dans les tables existantes, par exemple. Ensuite, il faut encore calculer les effets de l'agrégation $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$ ou $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$.

6) Utiliser les variables identifiées en (4) pour corriger l'analyse agrégée des variables d'intérêt et vérifier les effets de l'agrégation corrigés correspondant à (5), après correction, afin de s'assurer que l'ajustement a bien donné les résultats escomptés.

3.2 Variables de groupement idéales

Nous examinerons d'abord l'ensemble idéal de variables de groupement utilisables pour la correction de façon à déterminer quelle analyse (VGC) convient le mieux aux données agrégées, conformément à la démarche décrite plus haut.

Supposons que pour l'ensemble complet des variables qui nous intéressent, on connaisse la matrice des variances-covariances de niveau régional $\bar{S}_{yy}$ et la matrice des variances-covariances de niveau unitaire S_{yys_1} pour l'échantillon s_1. Bien sûr, si la chose était faisable dans la réalité, le problème d'agrégation ne se poserait pas puisqu'on pourrait écarter $\bar{S}_{yy}$ et simplement prendre S_{yys_1} comme estimation de Σ_{yy}. Néanmoins, il y a trois bonnes raisons pour étudier pareille situation. Tout d'abord, il vaut la peine d'éclaircir la structure du groupement qui établit la relation entre $\bar{S}_{yy}$ et S_{yys_1}. En second lieu, il se pourrait que $\bar{S}_{yy}$ et S_{yys_1} soient connus pour un moment quelconque dans le temps, le jour du recensement par exemple, mais que l'analyse approfondie d'une version plus récente de $\bar{S}_{yy}$ doive reposer sur des données intercensitaires pour lesquelles on ignore S_{yys_1}. Si la structure du groupement persiste dans le temps, comme il est raisonnable de le penser, l'analyse de $\bar{S}_{yy}$ et de S_{yys_1} au jour du recensement pourrait faciliter l'analyse intercensistaire en permettant l'identification des variables clés qui expliquent la majeure partie des effets de l'agrégation. Ces possibilités sont à la base même de la stratégie décrite à la partie 3.1. Troisièmement, si les variables de y couvrent de nombreuses variables socio-économiques et démographiques, comme cela se produit avec le recensement, les variables clés à l'origine des effets de groupement des variables à l'étude pourraient aussi expliquer une bonne partie des effets de groupement des autres variables socio-économiques et démographiques. Soulignons que les deux échantillons s et s_1 peuvent être identiques, sans que cela soit une condition. Ainsi, s pourrait venir d'une source administrative, en fait une sorte de recensement fournissant des données agrégées sur les régions, tandis que s_1 pourrait comprendre les données individuelles d'un sondage sans marqueur géographique.

Pour faciliter l'identification des variables importantes associées au groupement, Steel (1985) suggère de calculer les valeurs propres $\hat{\theta}_1, \ldots, \hat{\theta}_p$ de $S_{yys_1}^{-1}\bar{S}_{yy}$ et la matrice $\hat{D}_y = [\hat{d}_1, \ldots, \hat{d}_p]$ de telle sorte que

$$\hat{D}_y'\bar{S}_{yy}\hat{D}_y = \mathrm{diag}(\hat{\theta}_k) \quad \text{et} \quad \hat{D}_y' S_{yys_1}\hat{D}_y = I.$$

Les variables définies par la transformation

$$\hat{u}_i = \hat{D}_y' y_i$$

ont un ratio maximal pour la variance des groupes par rapport à l'échantillon, une corrélation nulle pour

l'échantillon au niveau unitaire et au niveau de groupe, et une variance égale à 1 pour l'échantillon au niveau unitaire. Ces variables sont baptisées variables de groupement canoniques (VGC) de l'échantillon. Les VGC de l'échantillon présentent les plus fortes corrélations à l'intérieur du groupe. Notons que $\mathrm{tr}(S_{yys_1}^{-1}\bar{S}_{yy}) = \sum_k \hat{\theta}_k$ peut correspondre à l'effet d'agrégation pour des variables multiples.

La matrice $\hat{D}_y$ existe même si S_{yys_1} et $\bar{S}_{yy}$ reposent sur des échantillons différents, tant et aussi longtemps que le premier est défini positif et le second, semi-défini positif. Par ailleurs, la variance des VGC n'est pas négative. Toutefois, si s et s_1 diffèrent, il se pourrait que la variance maximale d'une VGC dépasse $(N-1)/(M-1)$, soit l'effet d'agrégation le plus élevé. Dans ce cas, la VGC comprendrait implicitement une composante négative pour la variance à l'intérieur du groupe. La chose ne présente guère d'importance puisque nous cherchons à identifier les variables de groupement importantes. En principe cependant, la variance fautive pourrait être fixée au maximum théorique. On obtient les VGC de l'échantillon à partir des vecteurs propres de $A_{y\bar{y}} = S_{yys_1}^{-1}\bar{S}_{yy}$. Si s et s_1 sont le même échantillon, $A_{y\bar{y}}$ est le coefficient de régression de l'échantillon pour la régression des moyennes de groupe sur les valeurs de niveau unitaire pour l'échantillon. En fait, les VGC représentent les valeurs canoniques qui relient les données de niveau unitaire de l'échantillon à celles du niveau du groupe, tandis que $\hat{\theta}_k$ sont les corrélations canoniques de l'échantillon.

Une fois les VGC calculées, il est possible d'exprimer l'écart entre la matrice des covariances de l'échantillon au niveau du groupe et au niveau unitaire de la façon suivante:

$$\bar{S}_{yy} - S_{yys_1} = \sum_k (\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

où $\hat{\phi}_k$ correspond au vecteur des covariances de l'échantillon entre la k-ième VGC et les variables originales. On peut donc diviser l'écart entre la matrice des covariances du niveau du groupe et celle du niveau unitaire en k composantes orthogonales, soit une pour chaque VGC.

En ce qui concerne la covariance entre y_{ia} et y_{ib}, l'écart entre la covariance de l'échantillon au niveau du groupe $\bar{s}_{ab}$ et celle du niveau unitaire s_{ab} ($\bar{s}_{ab}$ et s_{ab} étant respectivement des éléments de $\bar{S}_{yy}$ et de S_{yys_1}) est donnée par

$$\bar{s}_{ab} = s_{ab} + (s_{aa}s_{bb})^{1/2}\sum_k (\hat{\theta}_k - 1)\hat{\rho}_{ak}\hat{\rho}_{bk}$$

où $\hat{\rho}_{ak} = \hat{\phi}_{ak}/s_{aa}^{1/2}$ représente la corrélation pour la a-ième variable et la k-ième VGC de l'échantillon.

En utilisant les q premières VGC de l'échantillon pour créer une matrice des variances corrigée au niveau du groupe, c'est-à-dire $\hat{u}_{qi} = \hat{D}_q' y_i$ où $\hat{D}_q = [\hat{d}_1, \ldots, \hat{d}_q]$ servent de variables auxiliaires,

$$\hat{\Sigma}_{yy}(\hat{u}_q) = \bar{S}_{yy} + \bar{B}_{yu_q}'(S_{u_qu_qs_0} - \bar{S}_{u_qu_q})\bar{B}_{yu_q}$$

on élimine les q premiers termes de la décomposition, donc

$$\hat{\Sigma}_{yy}(\hat{u}_q) = S_{yys_1} + \sum_{k=q+1}^{p} (\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

et $\mathrm{tr}(S_{yys_1}^{-1}\hat{\Sigma}_{yy}(\hat{u}_q)) = \sum_{k=q+1}^{p}\hat{\theta}_k$. . En fait, quand on prend les q premières VGC, on obtient la matrice de rang q qui minimise $\| S_{yys_1} - \hat{\Sigma}_{yy}(\hat{u}_q) \|$. Par conséquent, en étudiant les quantités

$$\sum_{k=q+1}^{p}\hat{\theta}_k \quad \text{et} \quad 1 + \sum_{k=q+1}^{p}(\hat{\theta}_k - 1)\hat{\rho}_{ak}^2$$

$$\text{pour} \quad q = 0, \ldots, p-1$$

il est possible d'établir comment les q premières VGC de l'échantillon expliqueront la partie de l'effet d'agrégation global et l'effet d'agrégation individuel de chaque variable.

Une telle analyse indiquera combien de dimensions sont nécessaires pour expliquer les effets cumulatifs et en supprimer une partie précise. Par ailleurs, en examinant la charge des variables à l'origine des VGC, on devrait être en mesure d'identifier les variables qui «expliqueront» le mieux les effets cumulatifs des autres variables. Les chercheurs devraient s'efforcer d'obtenir les données de niveau unitaire sur ces variables afin de les appliquer à l'estimateur corrigé.

Les résultats qui précèdent ont certaines implications importantes quant à l'usage des données au niveau du groupe complétées par des données de niveau unitaire limitées. En effet, ils autorisent la combinaison des données d'enquête et des données de groupe d'une ou de plusieurs sources, et laissent entrevoir une stratégie d'analyse pour les effets de groupe et les données au niveau du groupe.

4. QUELQUES RÉSULTATS EMPIRIQUES

Nous illustrerons les principes énoncés précédemment par une analyse des données du recensement de la population du Royaume-Uni effectuée en 1991 pour le district local (DL) de Reigate, Banstead et Tandridge. Ce DL compte 188,700 habitants répartis dans 371 DR, ce qui donne une population moyenne de $\bar{n} = 508.6$ par DR. On possède les données de groupe pour les habitants de chaque DR grâce au fichier de données régionales (FDR). Les données de niveau unitaire correspondantes du DL viennent d'un échantillon de 2 pour cent d'enregistrements sur des sujets anonymes. Il est donc impossible d'associer un enregistrement à un DR précis, de telle sorte que $\bar{S}_{yy}$ repose sur les données complètes du fichier se rapportant au DR et que l'estimation de $\bar{S}_{yys_1}$ repose sur l'échantillon de 2 pour cent d'enregistrements anonymes. L'analyse qui suit tient compte de 16 variables de recensement, pour chaque personne.

On s'est servi des données de groupe et des données de niveau unitaire de chaque variable pour calculer l'effet cumulatif $\bar{Q}_a = \bar{s}_{aa}/s_{aas}$. Le paramètre $\delta_{aa} = \Delta_{aa}/\Sigma_{aa}$,

Tableau 1

Effet d'agrégation et corrélation à l'intérieur de la classe pour les variables du recensement dans le DL de Reigate

	Effet d'agrégation	Corrélation
18 à 29 ans	9.20	.016
30 à 44 ans	4.56	.007
45 à 59 ans*	5.97	.010
60 ans et plus*	17.17	.032
Femme	1.08	.000
Personne de couleur*	8.29	.014
Marié	6.24	.010
Maladie débilitante à long terme	7.24	.012
Travailleur à plein temps	8.55	.015
Chômeur	2.27	.003
Autre situation d'emploi	11.19	.020
Soutien de famille né au R.-U.	4.48	.007
Soutien de famille né dans le Nouveau Commonwealth	3.59	.005
Soutien de famille émigré	9.04	.016
≤ 1.5 personne par pièce: densité	27.96	.053
Personnes dans un ménage sans automobile	32.98	.063

* Variable retenue pour la correction.
Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

défini pour les éléments diagonaux appropriés de Δ_{yy} et de Σ_{yy}, correspond à la corrélation de la a-ième variable à l'intérieur du groupe. On peut obtenir l'estimation $\hat{\delta}_{aa}$ de la corrélation à l'intérieur du groupe à partir de (2.18), puisque $\bar{Q}_a = 1 + (\bar{n}^* - 1)\,\hat{\delta}_{aa}$. Les résultats relatifs à ces variables apparaissent au tableau 1. En général, les corrélations à l'intérieur du groupe sont faibles, mais étant donné le nombre d'observations pour chaque DR, les effets cumulatifs peuvent être importants (voir la remarque qui suit l'équation (2.18)).

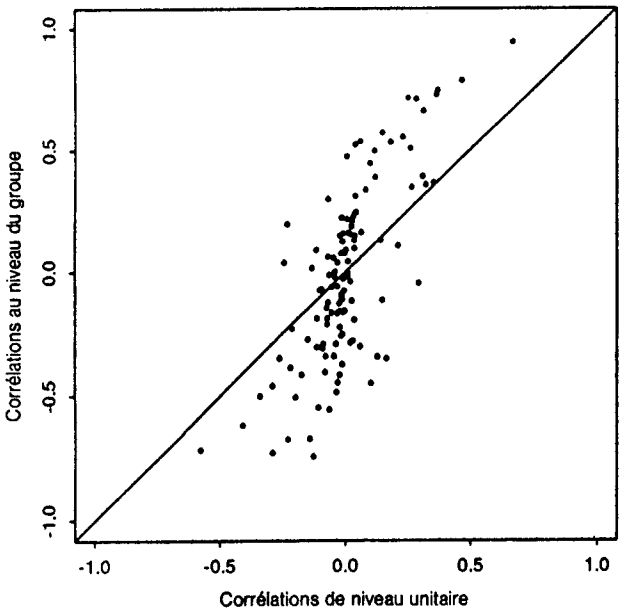

Figure 1a.

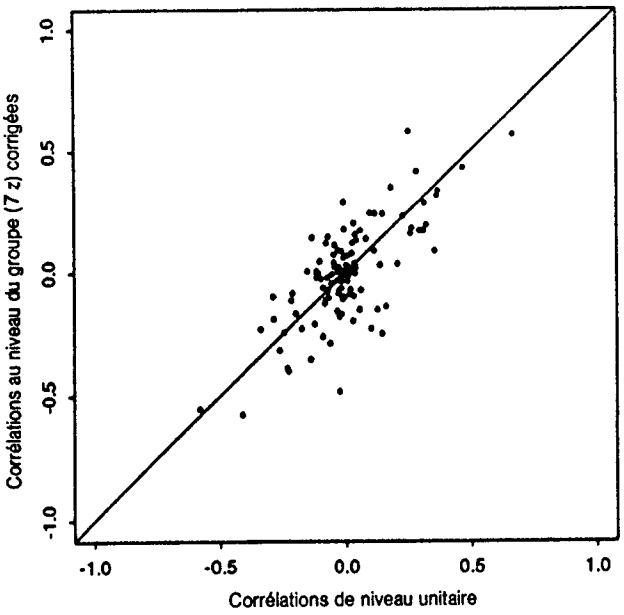

Figure 1b.

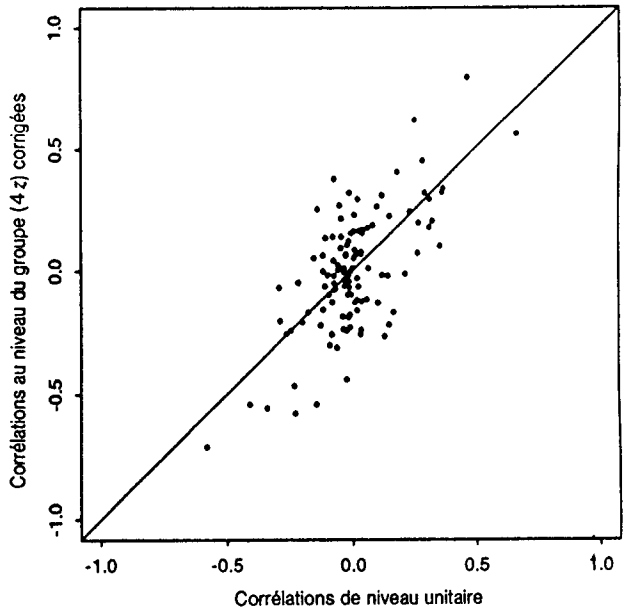

Figure 1c.

La figure 1a montre la courbe de la corrélation au niveau du groupe $\bar{r}_{ab}$ par rapport à la corrélation de niveau individuel r_{ab} pour chaque paire de variables. La forme en S caractéristique de la courbe fait ressortir l'importance des effets dus à l'agrégation. Les petites corrélations de niveau unitaire sont habituellement amplifiées si bien que dans la plupart des cas, $|\,\bar{r}_{ab}\,|$ est beaucoup plus grand que $|\,r_{ab}\,|$.

Tableau 2

Cinq premières VGC des variables du tableau 1

	VGC1	VGC2	VGC3	VGC4	VGC5
18 à 29 ans	0.4	0.3	0.9	1.1	0.1
30 à 44 ans	0.1	0.5	0.36	1.0	0.2
45 à 59 ans*	−0.1	1.2	−0.2	1.0	0.1
60 ans et plus*	0.3	2.2	−0.5	2.6	0.9
Femme	0.1	0.0	0.0	0.3	0.1
Personne de couleur*	0.5	−0.4	1.4	−1.1	5.2
Marié	−0.2	−0.5	−0.4	−0.8	−0.1
Maladie débilitante à long terme	0.3	0.1	−0.2	0.2	0.3
Travailleur à temps plein	0.7	−0.3	0.2	1.2	0.4
Chômeur	0.7	0.0	−0.1	0.0	−0.4
Autre situation d'emploi	0.1	0.1	0.0	−0.2	−0.1
Soutien de famille né au R.-U.	0.5	−0.1	−1.0	0.4	0.2
Soutien de famille né dans le Nouveau Commonwealth	0.0	−0.1	−0.3	0.1	0.6
Soutien de famille émigré	0.2	0.1	1.4	0.6	−1.3
≤ 0.5 personne par pièce	−1.4	0.3	1.2	−0.7	−0.2
Personnes dans un ménage sans automobile	2.2	0.6	0.8	−1.9	−0.7

* Variable retenue pour la correction.
Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

Connaissant $\bar{S}_{yy}$ et S_{yys_1} on peut analyser les variables de groupement canoniques en vue de comprendre la structure particulière du groupe. Le tableau 2 donne la charge des cinq premières variables de groupement canoniques. Ces cinq variables expliquent 89% de l'effet cumulatif des 16 variables utilisées.

La première VGC influe fortement sur la densité de population dans le logement et sur l'accès à une automobile. On pourrait donc la considérer comme un paramètre socio-économique. La deuxième attribue une forte charge aux variables qui identifient les personnes des deux groupes les plus âgés. La contribution des chefs de famille de couleur à la variable suivante est également remarquable. Comme on pouvait s'y attendre, certaines variables comme la proportion de femmes ne présentent pratiquement aucune corrélation à l'intérieur du groupe, donc n'ont aucun effet cumulatif et n'exercent presque aucune influence sur les VGC. Pareilles variables ne changent pas d'une région à l'autre. Elles n'ont donc habituellement aucune valeur explicative.

Dans la pratique, pareille analyse des VGC serait irréalisable, l'existence même de S_{yy} rendant une analyse agrégée inutile. L'analyse des VGC a néanmoins le mérite d'identifier les variables les plus importantes en raison de la charge qu'elles imposent aux premières VGC.

On sait pertinemment qu'au Royaume-Uni, les variables associées au mode d'habitation (qui ne font pas partie des 16 variables intéressantes) partagent des liens très étroits avec maintes variables liées à la situation socio-économique, au comportement et à la santé. Il y a tout lieu de croire qu'en les utilisant comme variables auxiliaires z pour la correction, on tiendrait compte d'une bonne partie de la première dimension socioéconomique. Elles pourraient donc remplacer les variables illustrant la densité de la population dans le logement et l'accès à une automobile, qu'on estime influer fortement sur la première VGC. L'autre raison pour envisager l'utilisation de ces variables est que si l'analyse actuelle doit illustrer ce qui pourrait se produire dans d'autres situations, les variables de base sur le mode d'occupation et sur le logement seront sans doute plus faciles à obtenir que celles sur la densité de la population dans le logement et l'accès à une automobile pour la correction. Compte tenu des résultats de l'analyse des VGC et puisqu'on souhaite des variables d'ajustement faciles à retrouver, peu importe la situation, nous proposons les sept variables d'ajustement potentielles qui suivent, soit les trois variables d'intérêt identifiées au tableau 1 par un astérisque (45-59 ans, 60 ans +, personnes de couleur) et les quatre variables liées au logement du tableau 3, avec leurs effets cumulatifs et leurs corrélations à l'intérieur de la grappe.

Tableau 3

Effet d'agrégation et corrélation à l'intérieur des variables liées au ménage dans le DL de Reigate

Variable		Effet d'agrégation	Corrélation
Mode d'occupation:	Locataire	133.43	0.261
	Propriétaire occupant	90.83	0.177
Type d'habitation:	Unif./semi/terrasse	90.03	0.175
	Commodités adéquates	59.52	0.113

Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

Nous corrigerons ensuite la matrice des covariances au niveau du groupe des 16 variables originales d'après la matrice des covariances de niveau unitaire pour les 7 variables z (trois variables démographiques de base de la série originale et quatre variables relatives au ménage).

Deux mesures globales déterminent l'efficacité de l'ajustement. La première est donnée par

$$1 - \frac{\operatorname{tr}\left(S_{yys_1}^{-1}\,\hat{\Sigma}_{yy}\,(z)\right) - 1}{\operatorname{tr}\left(S_{yys_1}^{-1}\,\bar{S}_{yy}\right) - 1}$$

qui est la réduction de l'effet agrégé multivarié. La seconde est

$$\frac{\| S_{yys_1} - \bar{S}_{yy} \| - \| S_{yys_1} - \hat{\Sigma}_{yy}(z) \|}{\| S_{yys_1} - \bar{S}_{yy} \|}$$

et indique la réduction de la distance généralisée entre les deux matrices de covariances aux niveaux unitaire et de groupe, avant et après la correction.

Le tableau 4 illustre l'effet de diverses combinaisons de variables sur la correction des résultats de l'analyse agrégée. Les deux variables associées à l'âge présentent manifestement de l'importance (elles expliquent 38% de l'effet d'agrégation multivariée et 53% de l'écart généralisé), mais il en va autant des variables sur le mode

Tableau 4

Combinaison des variables Z	Nombre de variables	Réduction de (%)	
		Effet d'agrégation multivarié	Écart général
60 ans et plus	1	16	24
45-59, 60+	2	38	53
Mode d'occupation	2	30	21
Type d'habitation	2	31	19
45-59, 60+, personne de couleur	3	44	54
45-59, 60+, mode d'occupation	4	57	71
45-59, 60+, type d'habitation	4	57	69
45-59, 60+, mode d'occupation, personne de couleur	5	63	72
45-59, 60+, type d'habitation, personne de couleur	5	62	70
45-59, 60+, type d'habitation, mode d'occupation, personne de couleur	7	68	75

d'occupation ou le type d'habitation. Quand on combine ces dernières aux variables liées à l'âge, la réduction obtenue avec chaque mesure, en pourcentage, se rapproche de la somme des effets des variables prises séparément, signe que l'âge et le mode d'occupation ou le type d'occupation entraînent des corrections distinctes. Comme on peut le constater, les meilleurs résultats proviennent de l'application des 7 variables, puisque ces dernières expliquent respectivement 68% et 75% de la correction des deux mesures de l'agrégation.

Les résultats révèlent que la correction supprime environ 70% des effets cumulatifs. Les figures 2a et 2b montrent l'effet de l'ajustement attribuable aux variables. À la figure 2a, l'axe des ordonnées intègre $|\ \bar{s}_{ab} - s_{abs_1}\ |$ le biais absolu de la covariance de chaque paire de variables au niveau du groupe. De son côté, l'axe des abscisses intègre $|\ \hat{\Sigma}_{ab}(z) - s_{abs_1}\ |$, le biais absolu de l'estimateur corrigé. La variance des variables est représentée par le symbole vide et la covariance, par le symbole plein. Pratiquement toutes les valeurs du graphique révèlent un biais plus faible (souvent beaucoup plus) que le biais original après correction. L'ajustement entraîne une amélioration appréciable dans presque tous les cas. La figure 2b trace les corrélations plutôt que les covariances (les corrélations y_a, y_a ont été omises pour des raisons évidentes). Encore une fois, on note une nette amélioration de l'analyse au niveau du groupe, le biais résiduel étant beaucoup plus faible que le biais original après la correction. Les résultats sont néanmoins inférieurs à ceux obtenus pour les covariances car, dans certains cas, de légers biais ont empiré lors de l'analyse au niveau du groupe. Ici, les corrections s'appliquent à la covariance et aux deux variances utilisées pour chaque coefficient de corrélation. Il y a plus de risques que la correction relative de chaque élément débouche

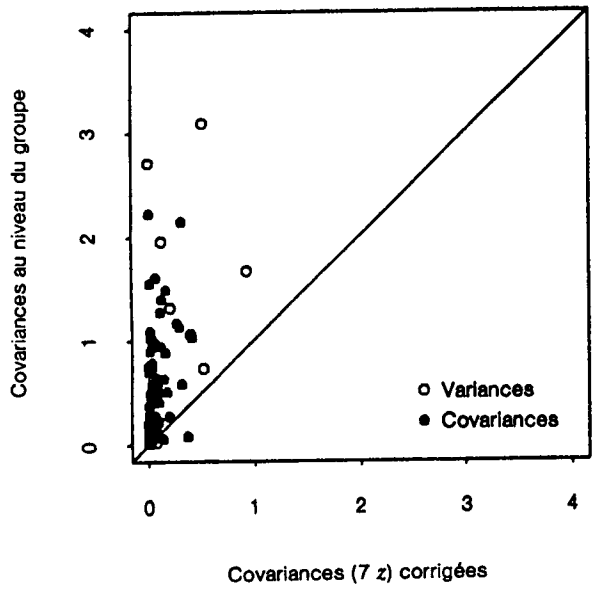

Figure 2a.

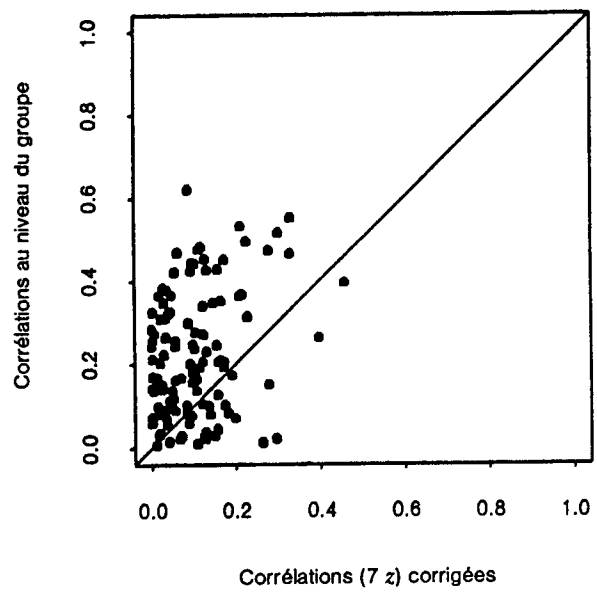

Figure 2b.

sur une moins bonne corrélation que l'originale. La plupart des biais importants relevés au niveau du groupe ont toutefois été atténués.

La figure 1b trace les corrélations corrigées au niveau du groupe $\bar{r}_{ab}(z)$ tirées de $\hat{\Sigma}_{yy(z)}$ par rapport aux corrélations de niveau unitaire. Elle peut être comparée à la figure 1a qui représente les corrélations non corrigées. La courbe en S caractéristique de la figure 1a a été remplacée par une série de points autour de la ligne $\bar{r}_{ab}(z) = r_{ab}$ ce qui est exactement ce qu'on espère obtenir en éliminant le biais agrégé.

Les figures 1b, 2a et 2b révèlent qu'on peut réduire sensiblement l'effet cumulatif en utilisant 4 variables liées à l'habitation et 3 des variables y originales. Les 120 variances et

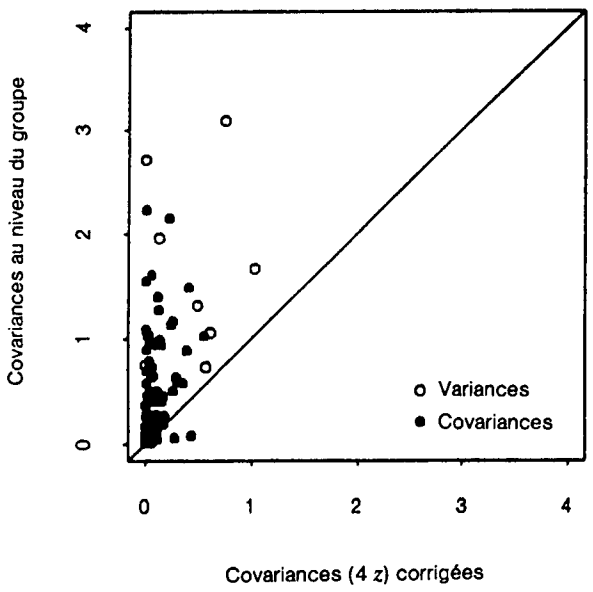

Figure 3a.

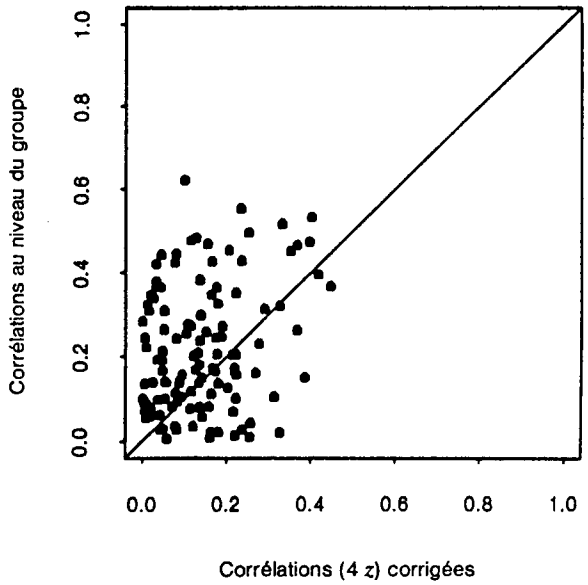

Figure 3b.

covariances de départ de la matrice 16 × 16 doivent donc être corrigées au moyen des 21 variances et covariances des variables z. Pour avoir une idée des résultats qu'on pourrait obtenir avec un minimum d'information, nous avons restreint les variables d'ajustement aux quatre variables liées à l'âge et au mode d'occupation. Le tableau 4 indique que ces variables expliquent respectivement 57% et 71% des deux mesures de l'agrégation. Les figures 3a et 3b donnent le tracé correspondant aux figures 2a et 2b pour ce cas particulier. La figure 1c compare les corrélations corrigées au moyen des quatre variables aux corrélations de niveau individuel. Bien sûr, la correction n'est pas aussi efficace, mais il est encourageant de voir ce qu'on obtient avec si peu de variables. L'écart médian absolu entre $\bar{r}_{ab}$ et r_{ab} (0.186) donne une indication supplémentaire des conséquences de la correction. Après correction au moyen des quatre variables, l'écart n'est plus que de 0.126 (0.090 après correction au moyen de sept variables). Les valeurs médianes correspondantes pour $|\ \bar{s}_{ab} - s_{ab}\ |$ sont respectivement de 0.173, de 0.039 et de 0.017.

5. CONCLUSIONS ET DISCUSSION

Les auteurs proposent un modèle qui décompose en deux éléments le biais observé dans l'analyse au niveau du groupe au moyen des matrices des covariances pour des populations groupées. Le premier élément résulte des variables de groupement et le second, des corrélations résiduelles entre les variables y à l'intérieur du groupe, compte tenu des variables de groupement z. Pareille décomposition nous aide à comprendre l'ampleur des effets cumulatifs. Elle indique aussi comment supprimer le biais attribuable aux variables de groupement quand on dispose de renseignements supplémentaires sur la matrice des covariances de niveau unitaire des variables de groupement.

Beaucoup de données de groupe à divers degrés d'agrégation peuvent être extraites du recensement et de nombreuses autres sources dans maints pays. L'expansion des systèmes d'information géographique accroîtra l'accessibilité de telles données. Il est important d'analyser et de décomposer les effets de groupe. La théorie et la méthode exposées ici procurent le cadre nécessaire à un tel exercice. En identifiant les variables qui expliquent la majeure partie des effets de groupe, donc qui devraient permettre la correction des analyses écologiques, on ouvrira la voie à l'utilisation des données agrégées.

REMERCIEMENTS

Le présent projet de recherche n'aurait pu être mené à bien sans la subvention H507 26 5013 de l'Economic and Social Research Council du Royaume-Uni. Les auteurs remercient sincèrement les examinateurs et le rédacteur adjoint pour leurs commentaires utiles.

BIBLIOGRAPHIE

ARBIA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems.* Dordrecht: Kluman.

BLALOCK, H.M. (1964). *Causal Inference in Nonexperimental Research.* Chapel Hill NC: University of North Carolina Press.

BLALOCK, H.M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-894.

BLALOCK, H.M. (1985). Cross level analysis. Dans *The Collection and Analysis of Community Data*, (Éd. J.B. Casterlin), ISI, World Fertility Survey.

CLARK, W.A.V., et AVERY, K.L. (1976). The effect of data aggregation in statistical analysis. *Geographical Analysis*, 8, 428-438.

DUNCAN, D.P., et DAVIS, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665-666.

FOTHERINGHAM, A.S., et WONG, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning*, A, 23, 1025-1044.

GEHLKE, C.E., et BIEHL, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, Supplement, 169-170.

GOODMAN, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64, 610-625.

HANNAN, M.T., et BUSTEIN, L. (1974). Estimation from grouped observations. *American Sociological Review*, 39, 374-392.

HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, A, 143, 474-87.

HOLT, D., et SCOTT, A.J. (1981). Regression analysis using survey data. *The Statistician*, 30, 169-173.

LICHTMAN, A.J. (1974). Correlation, regression, and the ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.

LANGBEIN, L.I., et LICHTMAN, A.J. (1978). *Ecological Inference*. Thousand Oaks, CA: Sage.

OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning*, A, 6, 17-31.

OPENSHAW, S., et TAYLOR, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Dans *Statistical Applications in the Spatial Sciences*, (Éd. N. Wrigley), 127-144.

PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philological Transactions of the Royal Society*, A, 200, 1-66.

PERLE, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning*, A, 9, 549-558.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2ième éd.). New York: Wiley.

ROBINSON, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.

SMITH, K.W. (1977). Another look at the clustering perspective on aggregation problems. *Sociological Methods and Research*, 5, 289-316.

SMITH, T.M.F., et HOLMES, D. (1989). Multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt et T.M.F. Smith), 165-187.

STEEL, D. (1985). Statistical Analysis of Populations with Group Structure. Unpublished PhD Thesis, Department of Social Statistics, University of Southampton.

STEEL, D., et HOLT, D. (1994). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited. Department of Applied Statistics, University of Wollongong, prétirage 1/94.

STEEL, D., et HOLT, D. (1995). Rules for random aggregation. *Environment and Planning* (à paraître).

YULE, U., et KENDALL, M.S. (1950). *An Introduction to the Theory of Statistics*. Glendale, CA: Griffin.