JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ISABELLA ANNESI-MAESANO
DAVID MOREAU
MICHEL CHAVANCE

Effets de la population atmosphérique particulaire sur la santé respiratoire dans le cas d'observations non indépendantes. L'étude des 6 villes

Journal de la société française de statistique, tome 145, n° 3 (2004), p. 59-67

<http://www.numdam.org/item?id=JSFS_2004__145_3_59_0>

EFFETS DE LA POPULATION ATMOSPHÉRIQUE PARTICULAIRE SUR LA SANTÉ RESPIRATOIRE DANS LE CAS D'OBSERVATIONS NON INDÉPENDANTES L'ÉTUDE DES 6 VILLES

Isabella ANNESI-MAESANO [1], David MOREAU[1] et Michel CHAVANCE [2]

RÉSUMÉ

Nous avons utilisé un modèle logistique marginal, qui permet de traiter des observations corrélées, afin d'explorer les effets de la pollution atmosphérique urbaine sur un indicateur de santé allergique et respiratoire infantile, soit le bronchospasme à l'effort (BE), indicateur d'asthme à l'effort, dans un échantillon de 6 839 enfants fréquentant 108 écoles primaires ayant participé à l'étude française des 6 villes. Dans cette étude, les enfants se retrouvent regroupés en 3 niveaux hiérarchisés (élèves d'une même classe, classes d'une même école, écoles d'une même ville) vis-à-vis des mesures de pollution atmosphérique réalisées. Le BE était significativement lié aux particules de taille inférieure à 2.5 μm ($PM_{2.5}$). Selon les types de structure des matrices des corrélations entre les variables qui ont été considérés, l'application du modèle marginal a montré des différences au niveau des valeurs des coefficients de régression et de la variance de ceux-ci, ce qui a des conséquences sur l'estimation des intervalles de confiance et donc sur la conduite des tests statistiques.

ABSTRACT

We used the logistic marginal model, which allows analysing correlated data (non-independent observations), in order to model the effects of air pollution on exercise-induced bronchial hyper-responsiveness, a marker of exercise-induced asthma (EIB), in a sample of 6,839 children attending 108 primary schools and who participated in the French Six Cities Study. With respect to air pollution exposure as assessed in the study, children were grouped in 3 different hierarchical levels (within a same classroom, within classrooms in the same school, within schools in the same city). Statistical analysis showed that EIB was significantly related to particulate matter with a diameter < 2.5 μm ($PM_{2.5}$). According to the structure of the correlation matrix, the marginal approach lead to variations in regression coefficients and variances respectively, which is needed to be known in order to appropriately estimate confidence interval and determine statistical significance.

1. Équipe d'Épidémiologie des maladies allergiques et respiratoires, INSERM U472, 16 avenue Paul Vaillant-Couturier F94807 Villejuif cedex
annesi@vjf.inserm.fr

2. Équipe de Biostatistique, INSERM U472

1. Introduction

Les études sur les relations entre environnement et santé font souvent intervenir des structures d'échantillonnage impliquant des observations non indépendantes. C'est le cas dans les enquêtes longitudinales où les observations successives sur un même sujet se ressemblent plus qu'elles ne ressemblent à celles provenant d'un sujet différent. C'est le cas aussi dans les enquêtes transversales où les sujets sont regroupés en niveaux hiérarchisés (par exemple élèves d'une même classe, classes d'une même école, écoles d'une même ville...). Les observations d'exposition effectuées dans une même unité d'un niveau donné, par exemple une école, se ressemblent alors plus qu'elles ne ressemblent à celles d'une autre unité du même niveau. Dans les deux cas, la matrice des variances-covariances et la matrice des corrélations des observations présentent une structure bloc-diagonale et cette structure se retrouve au niveau des mesures d'exposition si leur échantillonnage est calqué sur celui des mesures de l'indicateur de santé pris en compte. Ces deux types de corrélations ont évidemment des conséquences au niveau de l'analyse. Dans un modèle linéaire, une régression logistique, ou plus généralement un modèle linéaire généralisé, le fait d'ignorer les corrélations entre les réponses modélisées (les mesures de l'événement de santé) n'entraîne pas de biais sur l'estimation des paramètres, mais fausse l'estimation de la variance de leurs estimateurs et donc le calcul des intervalles de confiance ou la conduite des tests (Diggle, 1988; Molenberghs and Lesaffre, 1999).

Nous avons étudié la relation entre l'exposition à la pollution atmosphérique urbaine et la présence d'un bronchospasme à l'effort, indicateur d'asthme à l'effort, en tenant compte des corrélations induites par la structure d'échantillonnage de l'Étude des 6 villes, où les mesures de santé et d'exposition étaient hiérarchisées en trois niveaux (classe, école, ville).

2. Modèles considérant la non indépendance des observations

Pour prendre en compte la non indépendance des observations, deux types de modèles sont classiquement utilisés : les modèles mixtes et les modèles marginaux.

Les modèles mixtes supposent que les corrélations proviennent de la présence de coefficients à niveau aléatoire communs à l'ensemble des observations corrélées. On suppose pour simplifier que la structure hiérarchique des observations comprend deux niveaux (par exemple, des sujets regroupés dans des villes). Un modèle linéaire généralisé mixte s'écrit :

$$\begin{aligned} g\{E[Y_{ij}|b_i]\} &= X_{ij}\beta^* + Z_{ij}b_i \\ Y_{ij} &= E[Y_{ij}|b_i] + \varepsilon_{ij} \end{aligned} \tag{1}$$

où g est une fonction de lien monotone, Y_{ij} est la réponse du sujet i dans le groupe j, X_{ij} le vecteur des variables explicatives à effet fixe qui peuvent

caractériser le groupe j ou être spécifiques du sujet i, β^* le vecteur des effets fixes à estimer, Z_{ij} le vecteur des variables explicatives à effet aléatoire (le plus souvent un sous ensemble des variables à effet fixe), b_i le vecteur des effets aléatoires spécifiques au sujet i, enfin les erreurs résiduelles ε_{ij} sont indépendantes, conditionnellement à b_i. Dans le modèle linéaire ces erreurs sont gaussiennes et la fonction de lien est l'identité, alors que dans le modèle logistique les erreurs sont binomiales et le lien logit. Le modèle suppose généralement que les b_i suivent, indépendamment des erreurs résiduelles, des distributions gaussiennes $\mathcal{N}(0, \Sigma_b)$. Le vecteur des espérances est selon ce modèle $E(Y) = X\beta^*$, et la matrice des variances-covariances a une structure bloc-diagonale $\Sigma = \text{Diag}(Z_i \Sigma_b Z_i') + \text{Diag}\{\text{Var}[Y_{ij}|b_i]\}$, où Z_i représente la matrice des Z_{ij} du sujet i $(j = 1, n_i)$.

Les modèles marginaux modélisent séparément l'espérance $E[Y]$ et la matrice des variances-covariances Σ. Pour un modèle linéaire, on a :

$$\begin{aligned} Y &= E[Y] + \varepsilon = X\beta + \varepsilon \\ \Sigma &= \text{Var}[Y] = f(\alpha) \end{aligned} \tag{2}$$

où le vecteur des effets fixes β a la même interprétation et la même valeur que β^* dans le modèle mixte et où α est un vecteur de paramètres permettant de définir une matrice de covariances marginales de structure donnée. La seule différence avec le modèle mixte concerne donc la modélisation de Σ. Dans (1), elle se déduit de Σ_b et Z_i, ce qui autorise une interprétation en termes de composantes de variance, mais impose des contraintes. Alors que (2) autorise une modélisation plus souple (en augmentant le nombre de paramètres, il est possible d'ajuster parfaitement le modèle aux variances et covariances observées) mais dans une optique essentiellement descriptive.

Il existe cependant une différence importante entre les deux approches quand on sort du cadre du modèle linéaire pour envisager des modèles linéaires généralisés utilisant une autre fonction de lien que l'identité (Breslow and Clayton, 1993) car alors le vecteur des effets fixes n'a pas la même signification, et donc pas la même valeur, dans la formulation mixte et dans la formulation marginale. Le modèle devient

$$\begin{aligned} Y &= E[Y] + \varepsilon \\ g\{E[Y]\} &= X\beta \\ \text{Var}[Y] &= \Sigma = f(\alpha) \end{aligned} \tag{3}$$

On montre, par exemple, que pour le modèle logistique (observations binomiales et lien logit) le vecteur β^* du modèle mixte $\text{logit}(E[Y|b]) = X\beta^* + Zb$ et le vecteur β du modèle marginal $\text{logit}(E[Y]) = X\beta$ vérifient pour chacune de leurs composantes k : $|\beta_k^*| < |\beta_k|$ (Breslow and Clayton, 1993).

Le choix entre l'une ou l'autre approche dépend essentiellement de la formulation de la question scientifique étudiée : si l'objectif est de décrire la relation qui existe dans une population entre deux variables, telles qu'une exposition et un indicateur de santé, le modèle marginal donne la réponse. S'il s'agit d'expliquer l'effet d'une variable sur l'autre en éliminant tous les effets de

confusion ou d'effectuer une prédiction concernant un sujet de la population, un modèle mixte est préférable. De façon générale, le modèle mixte correspond davantage à la démarche explicative définie par Lellouch et Schwartz à propos des essais randomisés (voir par exemple Schwartz *et al.*, 1979) et les paramètres qui servent à modéliser la structure des covariances ont une interprétation plus facilement causale que les paramètres marginaux.

Avec l'un et l'autre modèle, quand on s'intéresse avant tout aux moments d'ordre un (mesure de l'effet d'une covariable sur l'espérance, prédiction individuelle...), il est recommandé de calculer les intervalles de confiance ou d'effectuer les tests statistiques en utilisant l'estimateur sandwich de la variance (Liang *et al.*, 1992)

$$\mathrm{Var}(\hat{\beta}) = \sum_i \left\{ (X_i' W_i^{-1} X_i)^{-1} X_i' W_i^{-1} \right\} V_i \left\{ W_i^{-1} X_i (X_i' W_i^{-1} X_i)^{-1'} \right\}$$

de préférence à la formule simplifiée que l'on obtient en supposant $W_i = V_i$, c'est-à-dire que les variances et covariances sont correctement stipulées par le modèle, à savoir

$$\mathrm{Var}(\hat{\beta}) = \sum_i (X_i' W_i X_i)^{-1}$$

Dans ces formules, l'indice i correspond au niveau où l'on a des observations indépendantes (les villes, dans notre exemple, alors que les écoles ou les classes d'une même ville se ressemblent), X_i représente la matrice des covariables de l'ensemble des observations de la ville i, W_i représente la matrice des covariances entre les observations spécifiée par le modèle et V_i la matrice des vraies covariances pour la ville i. Cette matrice qui se trouve au cœur du sandwich, permet d'obtenir un estimateur convergent de la variance, même si la modélisation W n'est pas correcte. On l'estime par la covariance observée des résidus du modèle. Dans la pratique, bien qu'une mauvaise spécification des covariances des observations entraîne une perte d'efficacité théorique, on n'observe guère de différences, en ce qui concerne l'estimateur $\hat{\beta}$ ou sa variance, selon la matrice de travail utilisée, et on peut souvent se contenter de la matrice de travail la plus simple, celle qui suppose l'indépendance des observations.

Un modèle logistique marginal a été utilisé pour étudier les effets des particules atmosphériques fines (c'est-à-dire de diamètre aérodynamique inférieur à 2,5 microns) sur le bronchospasme à l'effort dans un échantillon d'enfants issus de la population générale recrutés dans le cadre de l'étude française des 6 villes.

3. Matériel et méthodes

Population

Dans l'étude des 6 villes (Strasbourg, Reims, Créteil, Bordeaux, Marseille, Clermont-Ferrand), 6 839 enfants (soit 88 % des enfants ayant participé à au moins un des éléments du protocole) âgés de 10 ans en moyenne, ont bénéficié d'un bilan de santé allergique et respiratoire effectué par un médecin et leurs

parents ont répondu à un questionnaire épidémiologique standardisé sur leurs maladies et facteurs de risque.

Bronchospasme à l'effort

Le bilan médical a inclus entre autres une épreuve de course afin d'identifier l'existence d'un bronchospasme à l'effort, indicateur d'asthme à l'effort. Après avoir effectué 3 mesures du débit expiratoire de pointe (DP), les enfants ont couru 6 minutes (1 minute lentement, 4 minutes de façon soutenue, 1 minute le plus rapidement possible) dans le préau de l'école. Cinq minutes après l'arrêt de la course, les enfants ont réalisé à nouveau 3 mesures du débit de pointe. La diminution relative du débit de pointe après la course (ΔDP) a été ensuite calculée comme suit :

$$\Delta DP = \left[(DP_0 - DP_1)\right]/DP_0\right]$$

où DP_0 et DP_1 étaient donnés chacun par la meilleure des 3 mesures avant (DP_0) et après la course (DP_1) respectivement. Selon la littérature internationale, l'enfant présentait un bronchospasme à l'effort ($BE = 1$) lorsque la diminution relative du débit de pointe était supérieure à 0.10, et n'en présentait pas dans les autres cas ($BE = 0$), soit de façon générale : $BE = 1_{\Delta DP \geqslant 0.1}$.

Particules

Durant le bilan médical, la pollution atmosphérique a été mesurée au niveau des villes, des préaux des écoles fréquentées par les enfants (rapprochant l'exposition aux lieux de vie des enfants) et de leurs classes. Les mesures ont été réalisées en laissant les instruments de mesure (pompes pour les particules fines et capteurs passifs pour les autres) pendant 5 jours, du lundi au vendredi. L'application ici présentée concerne les effets des particules fines ($PM_{2,5}$), dont la mesure a été effectuée à l'aide de pompes avec cyclone et de filtres de collection, qui sont pesés avant et après le recueil. Le cyclone est choisi de façon à sélectionner seulement les particules d'une certaine taille, 2.5 micron dans notre étude. Les sujets étaient classés en exposés ou non exposés selon que la mesure les concernant était au-dessus ou au-dessous de la médiane des concentrations observées. Les élèves d'une même classe, étudiés simultanément, étaient associés à une même mesure de pollution à un niveau donné, en revanche les élèves d'une même école ou d'une même ville pouvaient être caractérisés par des mesures différentes au niveau de l'école ou de la ville si ces mesures avaient été effectuées à des périodes différentes.

4. Application

L'échantillonnage de l'Étude des 6 Villes présente une structure hiérarchique car il est constitué de villes ($i = 1, 6$) dans lesquelles on a tiré un échantillon de 108 écoles ($k = 1, 2, \ldots$) et 287 classes de CM1 et CM2 où les mesures

de santé et de pollution ont été réalisées simultanément, pour un total de 4 787 sujets. Les unités d'observation du niveau le plus bas ne sont donc pas indépendantes et on peut s'attendre à une structure de corrélation par blocs où le bloc correspondant à une ville est lui-même composé de blocs structurés en fonction des écoles où apparaissent les blocs correspondant aux classes d'une même école (Figure 1). Les observations effectuées sur des sujets de villes différentes sont supposées indépendantes.

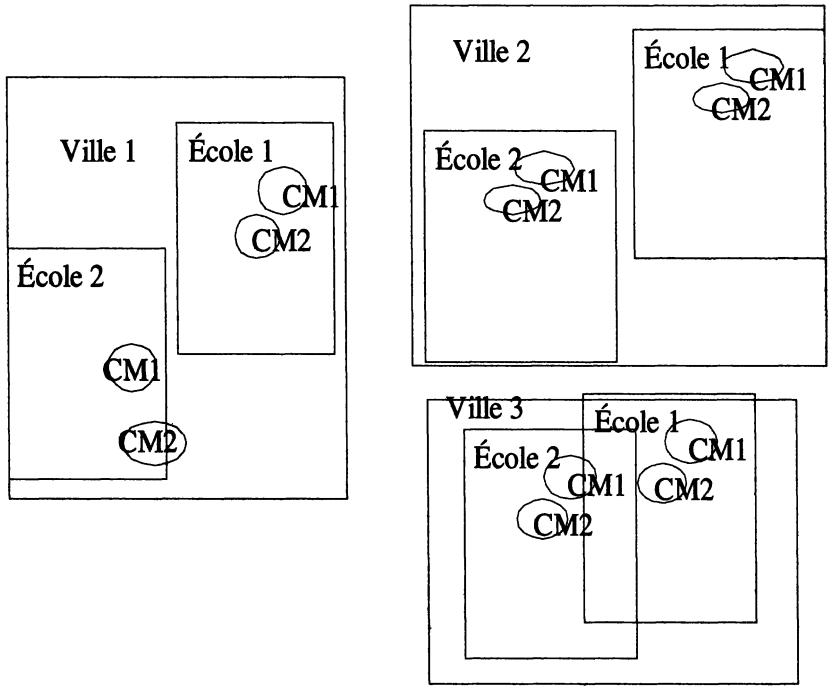

FIG 1. — Hiérarchisation des mesures de pollution : ville, école, classes de CM1 et CM2.

Cette structure hiérarchique se retrouve au niveau des mesures de pollution puisqu'on dispose d'enregistrements au niveau des salles de classe, des préaux et des villes et que, là aussi, on peut s'attendre à une structure de corrélations par blocs.

Indépendamment de la prise en compte des corrélations entre les réponses, les corrélations entre les mesures effectuées à différents niveaux sont susceptibles d'entraîner une augmentation des variances des estimateurs et une perte d'efficacité dans des modèles prenant en compte les effets de la pollution mesurée aux différents niveaux. Au moins dans un premier temps, il nous a donc semblé préférable de modéliser séparément l'effet de la pollution mesurée dans la classe, dans l'école ou dans la ville, ce qui a donné des résultats comparables (Annesi-Maesano *et al.*, 2003). Cependant, la mise en œuvre des GEE (Generalized Estimated Equations) permettant d'appliquer le

modèle linéaire marginal afin de réaliser une analyse plus appropriée doit faire intervenir dans l'estimateur sandwich les covariances des résidus au niveau le plus élevé, c'est-à-dire la ville, quel que soit le niveau de mesure. Les calculs ont été réalisés à l'aide du logiciel SAS en utilisant la procédure PROC GENMOD, qui permet de modéliser la structure de corrélation.

5. Résultats

Le tableau 1 présente les résultats de l'application du modèle linéaire généralisé marginal à l'étude de la relation entre le bronchospasme à l'effort et l'exposition aux particules fines selon 3 matrices de corrélations différentes, dont l'une ne prend pas en compte la non-indépendance des données. La BE étant une variable dichotomique, le modèle se réduit à une régression logistique.

TABLEAU 1. — Application du modèle linéaire marginal à l'étude de la relation entre le bronchospasme à l'effort (variable dichotomique) et l'exposition aux particules fines.

	Mesure dans le préau			**Mesure dans classe**		
	β	**OR**	**Test du chi^2**	β	**OR**	**Test du chi^2**
I	0,33	1,39	10,10	0,48	1,62	21,95
	[0,125-0,528]	[1,13-1,70]		[0,292-0,687]	[1,33-1,99]	
W	0,32	1,38	4,84	0,44	1,55	10,8
	[0,035-0,602]	[1,04-1,83]		[0,179-0,704]	[1,20-2,02]	
S	0,32	1,38	5,34	0,44	1,55	9,67
	[0,048-0,529]	[1,05-1,81]		[0,163-0,719]	[1,18-2,05]	

β : coefficient de régression ; OR : odds-ratio ; [ ] : intervalle de confiance à 95 %

I : estimation initiale sans tenir compte de la non-indépendance des observations

W : modèle utilisant la matrice de corrélation entre les observations afin de tenir compte de la non-indépendance de celles-ci ; la matrice choisie est constante au sein d'un même groupe dans un niveau (école par école)

S : modèle empirique utilisant la matrice de corrélation entre les observations afin de tenir compte de la non-indépendance de celles-ci ; la matrice choisie est calculée à partir des données.

L'exposition aux particules est significativement associée avec la BE. Lorsque la non-indépendance des observations est prise en compte (modèles W et S), la valeur du coefficient β du modèle diminue et sa variance augmente par rapport aux valeurs initialement obtenues (modèle initial (I)), ce qui donne une diminution de la statistique du test. Ceci est observé pour les 2 niveaux d'exposition : école et classe.

6. Discussion

L'étude des effets de la pollution atmosphérique représente un exemple approprié de l'utilisation des modèles permettant de considérer la non-indépendance des données; les observations étant répétées au cours du temps et/ou l'estimation de la pollution étant limitée, pour des raisons de faisabilité, à des mesures de population ou de groupe. L'application des modèles marginaux parmi d'autres modèles permet d'améliorer l'estimation des intervalles de confiance de l'association et ainsi la conduite des tests de signification statistique. Cependant, à ce jour, très peu sont les études ayant fait recours à ces modèles pour établir les effets sanitaires de la pollution de l'air. Dans une étude prospective sur l'asthme pédiatrique aux États-Unis, l'application du modèle marginal a montré que les variations à court terme de la pollution atmosphérique particulaire, le fait d'utiliser des traitements contre l'asthme et la sévérité des symptômes d'asthme étaient tous des indicateurs sensibles des effets adverses de la pollution atmosphérique (Delfino *et al.*, 1998). De même, le modèle marginal a été appliqué pour identifier l'association entre l'exposition à la pollution atmosphérique avec la mortalité par asthme dans la ville de Barcelone pendant l'épidémie de la période 1986 et 1989 (Saez *et al.*, 1999). Récemment, dans une étude longitudinale réalisée parmi des travailleurs chinois exposés aux poussières de coton, la relation entre l'exposition aux poussières de coton et aux endotoxines et l'incidence et la persistance des symptômes respiratoires a été mise en évidence en utilisant le modèle marginal (Wang *et al.*, 2003). Nous avons appliqué le modèle marginal à des données transversales afin de tenir compte de façon appropriée des mesures groupées d'exposition à la pollution atmosphérique. D'autres efforts doivent être menés afin de diffuser l'application de ces modèles dans les études épidémiologiques sur les effets de la pollution atmosphérique.

7. Remerciements

Les auteurs remercient les enfants, les parents, les directeurs et les professeurs des écoles ayant participé à l'enquête des 6 villes ainsi que Ginette Debotte qui a coordonné l'enquête sur le terrain.

Références

ANNESI-MAESANO I., MOREAU D., CAILLAUD D., DEBOTTE G., KOPFERSCHMITT C., LAVAUD F., RAHERISON C., TAYTARD A., TUNON de LARA J.M. and CHARPIN D. (2003). Adverse effects of PM2.5 on allergic response. The French Six Cities Study. *Am. J. Respir. Crit. Care Med.* **167** A35.

BRESLOW N.E. and CLAYTON D.G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88** 125-34.

DELFINO R.J., ZEIGER R.S., SELTZER J.M. and STREET D.H. (1998). Symptoms in pediatric asthmatics and air pollution : differences in effects by symptom severity, anti-inflammatory medication use and particulate averaging time. *Environ. Health Perspect.* **106** 51-61.

DIGGLE P. (1988). An approach to the analysis of repeated measures. *Biometrics* **44** 959-71.

LIANG K.Y., ZEGER S.L. and QAQISH B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. R. Stat. Soc.* **B.54** 3-40.

MOLENBERGHS G. and LESAFFRE E. (1999). Marginal modelling of multivariate categorical data. *Stat. Med.* **18** 2237-55.

PRIFTANJI A., STRACHAN D., BURR M., SINAMATI J., SHKURTI A., GRABOCKA E., KAUR B. and FITZPATRICK S. (2001). Asthma and allergy in Albania and the UK. *Lancet* **358** 1426-7.

SAEZ M., TOBIAS A., MUNOZ P. and CAMPBELL M.J. (1999). A GEE moving average analysis of the relationship between air pollution and mortality for asthma in Barcelona. *Spain. Stat. Med.* **30** 2077-86.

SCHWARTZ D., FLAMANT R. et LELLOUCH, J. (1979). *L'essai thérapeutique chez l'homme.* Collection médecine-sciences, Éditions Flammarion, Paris.

WANG X.R., EISEN E.A., ZHANG H.X., SUN B.X., DAI H.L., PAN L.D., WEGMAN D.H., OLENCHOCK S.A. and CHRISTIANI D.C. (2003). Respiratory symptoms and cotton dust exposure; results of a 15 year follow up observation. *Occup. Environ. Med.* **60** 935-41.