

Modèle hiérarchique de processus gaussien des plus proches voisins non stationnaire, multivarié, et non séparable, pour la modélisation des polluants atmosphériques

Sébastien Coube-Sisqueille*

Résumé

Décrire et de prédire l'exposition cumulative aux contaminants atmosphériques malfaisants tels que les oxydes d'azote, l'ozone ou les particules fines est un enjeu sanitaire et social.

Cette présentation propose une structure de modèle bayésien hiérarchique spatio-temporel combinant la non-séparabilité, la multivariabilité, et la nonstationarité. La non-séparabilité permet une description plus fidèle de la répartition des polluants. La multivariabilité permet une interprétation des interactions entre polluants, et augmente la puissance du modèle en "empruntant de la force" non seulement entre voisins spatio-temporels mais aussi entre variables. Enfin, la nonstationarité permet de lier des variables naturelles telles que les saisons ou anthropiques telles les limitations de vitesse à des changements de comportements des observations, tels que les pics de pollution.

Un modèle avec plusieurs variables d'intérêt ainsi qu'une dimension temporelle est un défi computationnel. Une approximation de Vecchia adaptée au "cahier des charges" de non-séparabilité, multivariabilité et nonstationarité est proposée, afin de rendre les calculs faisables pour un grand nombre d'observations. Des algorithmes de Monte-Carlo spécialement adaptés aux modèles spatiaux et aux données de grande taille sont présentés.

La méthode présentée est utilisée sur le territoire de la Communauté Autonome Espagnole d'Euskadi.

Mots-clefs : Statistique bayésienne – Statistique spatiale – Markov chain Monte-Carlo – Pollution – Processus gaussien des plus proches voisins – Approximation de Vecchia – Nonstationarité – Non-séparabilité – Multivariabilité

Développement

Le projet de recherche présenté lors de cette communication vise à modéliser l'exposition des populations de la communauté autonome Espagnole d'Euskadi à différents polluants atmosphériques. Pour ce faire, des modèles hiérarchiques spatio-temporels sont utilisés.

Les modèles spatiaux permettent de rendre compte du fait que les résidus d'un modèle linéaire sont souvent corrélés quand des observations sont faites à des sites géographiques voisins. On écrit alors :

$$y(s) = w(s) + \epsilon(s) + X(s)\beta,$$

s étant un site géographique, $\epsilon(\cdot)$ étant une erreur gaussienne dont les réalisations sont indépendantes, $X(\cdot)$ étant une collection de variables explicatives, et $w(\cdot)$ étant un processus gaussien dont les réalisations sont corrélées quand les sites spatiaux sont proches. La corrélation est le plus souvent calculée en utilisant un kernel défini-positif sur l'espace géographique étudié : $cov(w(s), w(s')) = K(s, s')$.

Lorsque les données sont observées non seulement dans l'espace mais aussi dans le temps, le temps noté t est ajouté comme une dimension supplémentaire. On parle de kernel séparable quand $K((s, t), (s', t')) = K_1(s, s') \times K_2(t, t')$. Cette hypothèse permet de grandes économies de temps de calcul mais elle n'est pas toujours réaliste, et des modèles non-séparables telles que celles de [Gneiting, 2002] peuvent améliorer les performances du modèle. Un développement récent de Allard et al. [2022]

*Basque Center for Applied Mathematics, scoube@bcamath.org

étend ce cadre de non-séparabilité à des fonctions de covariance multivariées. La multivariabilité permet de rendre compte des corrélations entre les variables, dues aux redondances des mesures ou à des relations chimiques entre les différents polluants. Le modèle devient alors, pour k variables :

$$y(s, t, v) = w(s, t, v) + \epsilon(s, t, v) + X(s, t)\beta_v,$$

$v \leq k$ étant un entier indiquant la variable. La fonction de covariance de Allard et al. [2022] s'écrit alors $cov(w(s, t, v), w(s', t', v')) = K(s, t, v, s', t', v')$, $v, v' \leq k$. Le bruit devient gaussien à k variables, $\epsilon(s, t, v = (1, \dots, k)) \stackrel{a\ priori}{\sim} \mathcal{N}(0, \Sigma_\epsilon)$.

Une autre hypothèse qu'il peut être intéressant de relâcher est la stationarité du processus. Par exemple, des variables naturelles telles que la météo, ou anthropiques telles que les limitations de vitesse, peuvent changer les comportements des polluants observées en empêchant ou en favorisant les phénomènes de pics. En utilisant un kernel de corrélation vérifiant $K_0(s, t, v, s, t, v) = 1$, la variance marginale de $w(\cdot, \cdot, \cdot)$ peut dépendre de covariables X_σ via la relation $\sigma(s, t, v) = \exp(X_\sigma(s, t)\beta_\sigma(v))$:

$$cov(w(s, t, v), w(s', t', v')) = \sigma(s, t, v)\sigma(s', t', v')K_0(s, t, v, s', t', v').$$

Enfin, la variance du bruit peut être expliquée par des covariables via le logarithme matriciel $\expm(\cdot)$ d'une matrice symétrique :

$$Var(\epsilon(s, t, v = (1, \dots, k))) = \expm(M(s, t)), \text{ avec } vech(M(s, t)) = X_\epsilon(s, t)\beta_\epsilon.$$

Les processus gaussiens ont comme faiblesse leur coût computationnel cubique dans le nombre d'observation. Pour avoir une méthode faisable, une approximation de Vecchia est utilisée. Cette méthode approche la densité $f(\cdot)$ d'un processus gaussien observé sur $s_1, \dots, s_n$ par

$$f(z(s_1, \dots, s_n)) = f(z(s_1))\prod_{i=2}^n f(s_i | s_1, \dots, s_{i-1}) \approx f(z(s_1))\prod_{i=2}^n f(s_i | pa(s_i)),$$

avec $pa(s_i) \subset (s_1, \dots, s_{i-1})$. Le fait que les sites de mesures soient fixes dans le temps et réalisées à intervalles réguliers est mis à profit pour proposer une approximation de Vecchia itérée, permettant d'avoir un temps de calcul proportionnel au nombre de mesures faites à un moment donné et non au nombre total de mesures, ce qui est très utile quand les observations s'étalent sur plusieurs années. Une observation de la variable v , faite au site s et au moment t , est conditionnée par ses voisins spatiaux et des observations faites à des sites voisins mais à des temps antérieurs :

$$pa(s, t, v) = pa_1(s, v) \times t \cup pa_2(s, v) \times (t - d, \dots, t - 1),$$

où $pa_1(s, v)$ et $pa_2(s, v)$ sont des parents spatiaux de différentes variables, t est la position temporelle, et d est la profondeur temporelle de l'approximation. Cette méthode induit aussi d'autres économies via la structure bloc-symétrique-Toeplitz de la matrice de covariance.

Afin de pouvoir prendre en compte la non-stationarité du modèle tout en gérant les observations manquantes, un modèle avec champ gaussien latent est utilisé. Un algorithme de MCMC utilisant la méthode d'entrelacement de Yu and Meng [2011] est présenté.

Références

- Denis Allard, Lucia Clarotto, and Xavier Emery. Fully nonseparable gneiting covariance functions for multivariate space-time data. 2022.
- Tilman Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458) :590–600, 2002.
- Yaming Yu and Xiao-Li Meng. To center or not to center : That is not the question—an ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3) :531–570, 2011.