Modéliser le pollen piégé au sol en fonction de la végétation simulée par LPJ-GUESS : Un modèle hiérarchique des processus intégrant sur-dispersion et zéros structurels

Vincent Garreta, Frédéric Mortier, Joël Chadoeuf

Modéliser le pollen piégé au sol en fonction de la végétation simulée par LPJ-GUESS :

Un modèle hiérarchique des processus intégrant sur-dispersion et zéros structurels

Vincent Garreta [a] & Frédéric Mortier [b] & Joël Chadœuf [c]

([a]) *CEREGE UMR6635, Université Aix-Marseille, Europôle de l'Arbois, 13545 Aix en Provence*
([b]) *CIRAD, Campus international de Baillarguet, 34398 Montpellier Cedex 5*
([c]) *INRA, Biométrie, domaine Saint Paul, 84914 Avignon Cedex 9*

Résumé En paléoclimatologie, la reconstruction du climat passé est obtenue par l'utilisation d'un modèle statistique de la relation entre le pollen contenu dans les sédiments et le climat. Ces modèles sont calibrés sur les données modernes. Introduire dans le modèle statistique un modèle physique de végétation devrait améliorer la prédiction des climats anciens. Néanmoins, cela nécessite de modéliser la distribution des pollens en fonction de la végétation. Deux difficultés doivent être prises en compte. D'une part, le modèle doit être ajusté sur une grande quantité de données de façon à intégrer une large gamme de climats. D'autre part les données de pollen disponibles présentent une forte sur-dispersion.

Dans le cadre des modèles hiérarchiques, nous développons un modèle bayésien qui inclut les principaux processus en jeu : production, dispersion spatiale, accumulation et échantillonnage du pollen mais aussi l'erreur induite par le modèle de végétation. Les processus d'accumulation et d'échantillonnage sont représentés par un modèle Multinomial-Poisson. Ce modèle permet de prendre en compte une sur-dispersion et la présence de zéros structurels. La dispersion est modélisée par un noyau Gaussien et l'erreur du modèle de végétation est décrite par un mélange de lois Gamma.

Nous montrons que les paramètres du modèle Multinomial-Poisson sont estimables et le vérifions en pratique par simulation. L'inférence sur une partie des données réelles confirme une sur-dispersion importante des données de pollen. La forme du modèle permet de paralléliser l'algorithme d'ajustement, ce qui permettra de travailler sur le jeu de données complet.

Abstract In palaeoclimatology, climate reconstruction is obtained using a statistical model for the relation between pollen contained in the sediments and climate. These statistical models are calibrated using a modern dataset. The use of a physical vegetation

model in the statistical model should improve past climate prediction. This purpose implies the modelling of pollen distribution as a function of simulated vegetation. Two difficulties must be considered. First, the model have to allow for inference on massive dataset since it must fit for a large climate range. Second, pollen data present an over-dispersion which requires modelling.

In the hierarchical model framework, we develop a Bayesian model which includes main processes: production, spatial dispersion, accumulation and sampling of the pollen. We model accumulation and sampling processes using a Poisson-Multinomial model. This model allows for over-dispersion and structural zeros (null Multinomial probabilities). Dispersion is modelled using a Gaussian kernel. Vegetation model's error is described using a Gamma mixed model.

We demonstrate that Multinomial-Poisson parameters can be estimated and use this result on several simulated dataset. Inference using a part of the whole dataset indicates a strong over-dispersion of pollen data. Model shape allows to parallelise the inference algorithm. Thus, we will manage to run inference on the whole dataset.

Mots-clés Loi Multinomiale-Poisson, Sur-dispersion, zéros structurels, Pollen, Modèle de végétation, Paléoclimatologie

Introduction

Les paléoclimatologues se servent du lien entre le climat et le pollen accumulé dans des pièges naturels comme les lacs afin de reconstruire le climat. Ils utilisent pour cela des modèles statistiques décrivant une relation -directe- entre climat et pollen. Ces modèles sont calibrés sur les données modernes (climat et pollen connus). Ils sont ensuite utilisés sur des enregistrements de pollen ancien pour reconstruire le climat passé. Parallèlement, la recherche en écologie produit des modèles de végétation. Ces modèle physiques simulent une production de végétation pour un climat donné. En utilisant un modèle de végétation tel que LPJ-GUESS (Smith et al (2001)) pour simuler la végétation à partir du climat, la modélisation statistique du lien climat-pollen se réduit à celle du lien végétation-pollen. Cette modélisation devrait améliorer la prédiction des climats anciens. D'une part, en représentant les processus qui lient la végétation au pollen et d'autre part en prenant en compte l'aspect spatial du jeu de données modernes.

Notre but est alors de construire et d'inférer un modèle paramétrique intuitif -dans le sens de causal- représentant les principaux processus qui relient la végétation (représentée par une simulation de LPJ-GUESS) et le pollen : (a) l'erreur entre la végétation potentielle simulée par LPJ-GUESS et la végétation actuelle, (b) la production d'une quantité absolue de pollen par unité ($kg.an^{-1}.m^{-2}$) de végétation. Elle est linéairement dépendante de la quantité absolue de végétation et propre à chaque espèce (Sugita (2007)), (c) la dispersion du pollen dont la portée est propre à chaque espèce (eg. Sugita (2007)), (d)

l'accumulation du pollen dans différents types de pièges naturels. Ces pièges varient, d'une mousse dans un sous-bois au sédiment d'un lac de très grande surface. L'information sur le type de piège associé à chaque échantillon n'étant pas disponible nous modéliserons ce processus d'intégration plus ou moins locale du pollen "volant" comme une erreur, (e) un processus d'échantillonnage : la reconnaissance de N_i -arbitrairement fixé- grains de pollens par site i.

Dans le cadre Bayésien nous construisons un modèle hiérarchique qui est adapté aux données spatiales actuelles. Chacune de ses couches représente un processus listé conditionnellement à la réalisation du processus le précédent.

Un des principaux problème est la modélisation du pollen. Un échantillon de pollen est un vecteur dont chaque élément est le nombre de grains de pollen d'un type donné (Pin, Chêne, etc). Ces échantillons sont a priori sur-dispersés : un échantillon suit une loi Multinomiale conditionnellement aux fréquence théoriques, mais ces fréquences sont aléatoires (ces aléas provenant des processus d'accumulation). D'autre part, l'existence de limites climatiques par espèce fait qu'une espèce, en un point de l'espace, peut être absente, être absente par effet d'échantillonnage ou être présente. Le modèle Multinomial-Dirichlet (MD) habituellement utilisé n'autorise pas la présence de zéros structurels (fréquences de la loi Multinomiale nulles). Nous proposons un modèle Multinomial-Poisson (MP) qui autorise la présence de tels zéros.

Modèle

On dispose du pollen et de la végétation simulée sur n sites suffisamment denses pour que le pollen produit en un site puisse être retrouvé dans un site voisin. Soit $Y_i = (Y_i^1, ..., Y_i^k)$ le vecteur des comptages du pollen identifié au site i et par taxon $j = 1, ..., k$. $N_i = \Sigma_{j=1}^k Y_i^j$ est le nombre total de pollens au site i. N_i est un choix arbitraire du technicien qui a fait la mesure et représente la qualité de l'échantillon. $\mathrm{NPP}_i = (\mathrm{NPP}_i^1, ..., \mathrm{NPP}_i^k)$ est le vecteur des production primaires nettes simulées par le modèle LPJ-GUESS aux mêmes sites i et pour les mêmes taxons j que le pollen.

Echantillonnage

La distribution des pollens échantillonnés connaissant leurs proportions $p_i = (p_i^1, ..., p_i^k)$ accumulées dans le piège naturel est indépendante entre site i et modélisée par une loi Multinomiale de paramètres (N_i, p_i) : $[Y_i|p_i] = \mathcal{M}(N_i, (p_i^1, ..., p_i^k))$.

Sur-dispersion Poisson

La sur-dispersion par rapport à la loi Multinomiale, due au processus d'accumulation, est modélisée au travers des p_i en définissant des variables aléatoires latentes $X_i = (X_i^1, ..., X_i^k)$ reliées aux p_i par : $p_i^j = X_i^j / \Sigma_j X_i^j$. Les X_i^j suivent des lois de Poisson

$$[X_i^j | b, S_i, S_i^j > 0] = \mathcal{P}(b^j S_i^j)$$
$$[X_i^j | b, S_i, S_i^j = 0] = \delta_{\{X_i^j = 0\}}$$

où $S_i = (S_i^1, ..., S_i^k) \in R^{+k}$ est l'intensité du pollen amené par dispersion spatiale et $b = (b^1, ..., b^k)$ des paramètres positifs décrivant à la fois la sur-dispersion et la production. Nous supposons donc que les X_i^j, conditionnellement à S_i et b sont indépendants.

Par la suite l'indice est supprimé pour alléger les notations. Le rapport b^j/b^l s'interprète comme la production relative de pollen de l'espèce j par rapport à l'espèce l, $K = \Sigma_j\, b^j S^j$ s'interprète comme l'efficacité moyenne des pièges et enfin, $a^j = b^j S^j / \Sigma_j\, b^j S^j$ comme la proportion théorique de pollen de l'espèce j. On notera que $b^j = K a^j$. De plus, pour un K suffisamment grand et $\forall\, a^j \in [0, 1]$:

$$\mathsf{E}\left[p^j | \Sigma_j\, X^j > 0\right] \to a^j \tag{1}$$

$$\mathsf{Var}\left[p^j | \Sigma_j\, X^j > 0\right] \to a^j(1 - a^j) \left(\sum_{i=1}^{\infty} \frac{K^i}{i!\, i}\right) e^{-K} \approx \frac{a^j(1 - a^j)}{K} \tag{2}$$

Les formules 1 et 2 impliquent que les paramètres a et K sont identifiables, le premier par l'espérance des p^j, le second par leur variance. D'autre part, dans le cadre Bayésien, la formule 2 permet de donner une borne supérieure pour l'a priori sur K et ainsi rendre le modèle identifiable même en absence de sur-dispersion ($b \to \infty$) : la limite à partir de laquelle nous considérons que la sur-dispersion est négligeable.

Dispersion spatiale

Nous modélisons l'intensité de pollen apporté par dispersion spatiale au point i pour l'espèce j selon : $S_i^j = \sum_{k=1}^{n} \alpha_j(d(s_i - s_k)).V_k^j$
où $d(s_i - s_k)$ est la distance euclidienne et $\alpha_j(d(s_i - s_k)) = \exp(-\frac{d(s_i - s_k)^2}{2\gamma_j^2}) / \sum_{k=1}^{n} \exp(-\frac{d(s_i - s_k)^2}{2\gamma_j^2})$ est un noyau Gaussien. γ_j s'interprète comme le paramètre de portée de dispersion de l'espèce j.

Production de pollen

Comme la variabilité de production entre types de végétation est modèlisée au travers des paramètres b^j, la production de pollen est unitaire par unité de végétation.

Erreurs du modèle de végétation

Pour ne pas propager les possibles incohérences de simulation de LPJ-GUESS à tout le modèle hiérarchique nous modélisons de façon aléatoire la végétation contribuant effectivement à la production de pollen. En se basant sur la végétation potentielle (NPP) simulée par LPJ-GUESS nous modélisons V_i^j la végétation de l'espèce j au point i suivant un mélange de loi Gamma et de masse de Dirac

$$\left[V_i^j | \mathrm{NPP}_i^j = 0\right] = q_1^j \delta_0 + (1 - q_1^j)\Gamma(1, 1/\sigma_1^j)$$
$$\left[V_i^j | \mathrm{NPP}_i^j > 0\right] = (1 - q_2^j)\delta_0 + q_2^j \Gamma\left(\frac{(\mathrm{NPP}_i^j)^2}{(\sigma_2^j)^2}, \frac{\mathrm{NPP}_i^j}{(\sigma_2^j)^2}\right)$$

où q_1^j et $q_2^j \in [0.5, 1]$ sont respectivement, la probabilité que l'espèce j soit présente sachant que le modèle ne l'a pas simulé et la probabilité que l'espèce j soit présente sachant que le modèle l'a simulé. σ_1^2 et σ_2^2 sont les variances des lois Gamma quand V_i^j n'est pas nulle.

Premier résultats

Nous avons développé un algorithme de Metropolis-Hasting within Gibbs pour l'inférence du modèle.

Simulations Pour quantifier la qualité de l'estimation nous avons procédé à des tests sur jeu de données simulées. Dans un premier temps nous considérons le modèle sans erreur sur V (V_i =NPP$_i$). Nous travaillons avec $k = 3$ champs en dimension 1. Nous simulons la valeur de chacun des champs en $n = 30$ points aléatoirement répartis sur $[0, 20]$. Les paramètres varient selon : (a) la régularité du champ NPP^j simulé suivant une loi normale seuillée en 0, de moyenne 0 et de variance 1 en considérant deux cas : les valeurs sont indépendantes et les valeurs ont une corrélation spatiale de portée $j/2$, (b) le contraste entre les productions de pollen en considérant un cas peu contrasté $b^1 = K * 0.3 = b^2 = K * 0.3 \approx b^3 = K * 0.4$ et un autre contrasté $b^1 = K * 0.65 \gg b^2 = K * 0.3 \gg b^3 = K * 0.05$, (c) la sur-dispersion : $K = (20, 50, 100, 150, 200)$. Chacune des possibilités est simulée 10 fois.Les intervalles de confiance a posteriori contiennent bien à chaque fois les valeurs simulées de b, X et γ.

Application Nous avons réalisé l'inférence du modèle complet sur des données couvrant l'Espagne. Les premiers résultats indiquent une sur-dispersion importante : $K \simeq 50$. Les portées de dispersion (γ_j) sont de l'ordre de la centaine de kilomètres. Ces données sont issues d'un échantillonnage plus vaste (1301 sites et 15 types de pollen) couvrant toute l'Europe mais les temps de calculs ne permettent pas l'inférence sur ce jeu de données complet. Cependant le modèle présente l'avantage d'être parallélisable.

Conclusion et discussion

La construction du modèle Multinomial-Poisson (MP) est similaire à celle du modèle Multinomial-Dirichlet (MD) qui est construit sur la base de lois Gamma : modélisation des probabilités de la loi Multinomiale suivant des rapports $X^j/\Sigma_j X^j$, X^j indépendants. Son avantage est qu'il permet de modéliser à l'aide d'un seul paramètre par champ (b^j) des données contenant des fréquences nulles pour la loi Multinomiale (zéros structurels) en même temps que la sur-dispersion. Les modèles de mélange, classiquement proposés pour la modélisation de zéros comportent plus de paramètres.

Dans le modèle MP la quantité de zéros par champ est dépendante du paramètre de sur-dispersion K. Pour construire un modèle avec zéro-inflation en plus de la sur-dispersion, la loi de Poisson pourrait être remplacée par une loi Binomiale-Négative.

La couche de végétation a un rôle double. D'une part ce rôle est technique : autoriser la présence de végétation là où LPJ-GUESS n'en simule pas. Sinon, il existe des échantillons où la végétation prédite est nulle et où le modèle de dispersion ne permet pas de représenter avec une fréquence suffisante les pollens présents. D'autre part nous faisons l'hypothèse que LPJ-GUESS simule bien les intéractions (dépendances) entre types de végétation et que l'erreur qu'il commet est indépendante entre ces types. Ainsi nous récupérons une dépendance entre types (celle de LPJ-GUESS) et ne propageons pas d'erreurs dans tout le modèle hiérarchique.

L'interêt de cette couche n'est pas tant dans l'identification de ses paramètres que dans ses capacités de prédiction. Nous ne pouvons pas démontrer que tous ses paramètres sont identifiables. Par exemple les lois induites sur S_i^j par dispersion ne sont pas analytiquement calculables. Par contre, la qualité de la prédiction d'un tel modèle dépend de la qualité de la loi a posteriori, les dépendances entre paramètres pouvant se compenser lors de la prédiction. Il est pour cela important de s'assurer de la convergence de l'algorithme de Monte Carlo par chaîne de Markov (MCMC) utilisé.

Dans le contexte de la reconstruction des paléoclimats ce modèle est le premier modèle causal qui intègre une corrélation spatiale des pollens. En proposant une modélisation intuitive -par le formalisme hiérarchique- des processus majeurs reliant climat et pollen il permet une interprétation et donne un sens aux erreurs de reconstruction du climat. Associé au modèle temporel et à l'algorithme de reconstruction par filtrage particulaire de Garreta et al (soumis) il pourrait permettre les premières reconstructions spatio-temporelles du climat.

Bibliographie

[1] Garreta, V., Miller, P., Guiot, J., Hély, C., Brewer, S., Sykes, M.T. et Litt, T. (2008) Method for Climate and Vegetation dynamics reconstruction through the inversion of a vegetation model using pollen data. *Soumis à Climate Dynamics*.
[2] Smith, B., Prentice, I. et Sykes, M. (2001) Representation of vegetation dynamics in modelling of terrestrial ecosystems: comparing two contrasting approaches within european climate space. *Global Ecology & Biogeography*, 10:621–637.
[3] Sugita, S. (2007) Theory of quantitative reconstruction of vegetation i: pollen from large sites reveals regional vegetation composition. *The Holocene* 17(2):229–241.