Changements climatiques passés reconstruits à partir du pollen : Vers une modélisation statistique basée sur les mécanismes

Vincent Garreta

Changements climatiques passés reconstruits à partir du pollen :

Vers une modélisation statistique basée sur les mécanismes

Vincent Garreta

CEREGE UMR6635, Europôle de l'Arbois, 13545 Aix en Provence

Résumé La reconstruction des changements climatiques survenus dans les derniers milliers d'années est cruciale pour comprendre la dynamique du climat dans des conditions différentes de celles enregistrées de nos jours (e.g insolation, concentration en CO_2). De telles reconstructions peuvent être obtenues en modélisant les relations entre le climat et les pollens retrouvés dans les sédiments lacustres. Ces modèles, appelés Fonction de Transfert (FT), sont des modèles statistiques purement descriptifs qui, malgré la diversité de leur forme, se basent tous sur un même groupe d'hypothèses. Cela réduit la confiance que l'on peut avoir dans des reconstructions uniquement obtenues à l'aide de ces FT. Au contraire, des FT basées sur les mécanismes liant l'environnement à la végétation et au pollen fourniraient des reconstructions basées sur des hypothèses différentes et soutenues par les recherches récentes en écologie. Pour obtenir ce type de FT nous proposons de coupler un modèle mécaniste de végétation et un modèle bayésien hiérarchique. Le cadre bayésien est naturellement approprié pour l'inférence d'un modèle dont la vraisemblance est implicitement définie par un simulateur mécaniste. Cependant, ici, elle est compliquée par les dimensions spatio-temporelles considérées. Nous montrons comment le problème peut être abordé en combinant des algorithmes de Monte Carlo pour un cas réel de reconstruction de l'Holocène en Suède. Ce type de modèle forme la prochaine génération de FT mais nécessite de poursuivre le développement d'outils statistiques pour l'inférence de modèles composites dans la lignée de ce qui est proposé en Calcul Bayésien Approché (ABC) et émulation de modèle.

Mots-clés Modéle bayésien hiérarchique, modèle mécaniste, vraisemblance implicite, inférence par méthodes de Monte Carlo, paléoclimatologie, pollen, modèle de végétation

Abstract Reconstruction of past climate change that occured during the last thousands of years is crucial for understanding climate dynamics under conditions different than those recorded today (e.g insolation, CO_2 concentration). Such reconstructions can be obtained by modelling the relations between climate and pollen sampled in lacustrine sediments. These models, called Transfer Functions (FT), are purely descriptive statistical models. Despite their diversity in shape, they are all based on a same set of hypotheses.

This reduces the confidence we may have in reconstructions only available through these FT. On the contrary, FT based on the processes linking environment, vegetation and pollen would provide reconstructions based on different hypotheses supported by recent research in Ecology. To obtain this type of FT, we propose the coupling of a mechanistic vegetation model and a hierarchical Bayesian model. The Bayesian framework is adapted to the inference of models whose likelihood is implicitly defined by a mechanistic simulator. However, here, it is complicated by the number of spatio-temporal dimensions considered. We show how the problem can be addressed in a first approach by combining Monte Carlo algorithms. This is applied to a real-world reconstruction of Holocene climate in Sweden. This type of model forms the next generation of FT but requires to continue the development of statistical tools for the inference of composite models following what as been proposed in Approximate Bayesian Computation (ABC) and model emulation.

Key-words Hierarchical Bayesian Model, Mechanistic model, Implicit likelihood, inference by Monte Carlo methods, Palaeoclimatology, Pollen, Vegetation model

Introduction

La reconstruction des changements climatiques survenus dans les derniers milliers d'années est cruciale pour comprendre la dynamique du climat dans des conditions qui étaient différentes de celles enregistrées de nos jours (e.g insolation, concentration en CO_2). En particulier, ces reconstructions sont utilisées pour calibrer et valider les modèles climatiques globaux (GCM).

De telles reconstructions peuvent être obtenues en modélisant les relations entre le climat et les pollens retrouvés dans les sédiments lacustres. Ces modèles, appelés Fonction de Transfert (FT), sont des modèles statistiques purement descriptifs du lien entre la composition pollinique et un petit nombre de variables climatiques. Ils sont calibrés sur un jeu de données moderne composé d'échantillons pollinique de surface et d'enregistrements du climat actuel. En utilisant l'une des TF classiques pour reconstruire le climat correspondant à un échantillon pollinique ancien, on est obligé de faire les hypothèse que, (i) ce sont les quelques variables climatiques considérées qui contrôlent seules l'absence, la présence et la productivité des espèces considérées et (ii) la réponse de le végétation au changement climatique est "instantanée". Ces hypothèses sont intrinsèques à l'approche descriptive des FT classiques. Elles semblent trop restrictives car les travaux récent en modélisation de la végétation considèrent des mécanismes (e.g compétition, migration) qui supportent des hypothèses contraires.

La construction de FT entièrement basées sur les mécanismes liant l'environnement à la végétation et au pollen permettrait d'obtenir des reconstructions basées sur des hypothèses différentes et soutenues par les recherches récentes en écologie. Pour obtenir ce type de FT nous proposons de coupler (a) un modèle mécaniste de végétation pour lier le climat à la végétation et (b) un modèle bayésien hiérarchique des processus pour représenter les

mécanismes principaux liant la végétation au pollen piégé dans les sédiments lacustres. Le modèle de végétation considéré, LPJ-GUESS (Smith et al, 2001) est un modèle dynamique simulant une Production Primaire Nette (NPP) aléatoire.

En partant de la NPP de différentes espèces au temps $t-1$, notée NPP_{t-1}, et d'une chronologie liant $t-1$ à t, notée C_t, le modèle simule la NPP au temps t: NPP_t. Suivant Garreta et al (2009a) nous considérons donc LPJ-GUESS comme la distribution suivante

$$p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t) \tag{1}$$

Cette distribution définie par le simulateur mécaniste (i.e implicite) peut être simulée (en faisant tourner le modèle) mais sa valeur en un point ne peut être évaluée car elle est définie au travers d'une chaîne complexe de mécanismes simulatoires. Ce modèle de végétation est couplé à un modèle statistique $p(Y|\mathrm{NPP}, \theta)$ de paramètres θ représentant la chaîne de processus liant le NPP au pollen Y: les perturbation locales de la végétation $\rightarrow$ la production pollinique $\rightarrow$ la dispersion pollinique $\rightarrow$ l'accumulation du pollen dans le lac $\rightarrow$ l'échantillonnage (Garreta et al, 2009b). La structure interne de notre modèle composite est donc

$$p(Y_t, \mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t, \theta) = p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)\, p(Y_t|\mathrm{NPP}_t, \theta) \tag{2}$$

L'inférence pour ce modèle apparaît dans les étapes de calibration des paramètres de la FT en utilisant les données modernes et en reconstruction du climat passé en utilisant les données polliniques anciennes. Le cadre bayésien est naturellement approprié pour l'inférence d'un modèle dont la vraisemblance est implicitement définie par un simulateur mécaniste. En effet, des algorithmes comme l'Importance Sampling (IS, Robert et Casella, 1999) ne requièrent que la capacité de simuler suivant le modèle. Cependant, dans notre cas elle est compliquée par la prise en compte de liaisons spatio-temporelles appliquées à de gros jeux de données. Nous montrons comment le problème peut être abordé en combinant des algorithmes de Monte Carlo pour un cas réel de reconstruction de l'Holocène en Suède.

Calibration d'une FT basée sur les mécanismes

Le processus de calibration est définit, dans le cadre Bayésien, comme l'obtention de la distribution a posteriori des paramètres θ contrôlant le lien climat-pollen sachant les données modernes de climat et pollen. Nous notons la collection des données actuelles de pollen en tous sites $s = 1..N$ $(y_{s=1..N})$ $\mathbf{y}$, de climat $\mathbf{c}$ et de végétation $\mathbf{NPP}$. Cette distribution a posteriori est

$$\begin{aligned} p(\theta|\mathbf{y}, \mathbf{c}) &= \int p(\theta, \mathbf{NPP}|\mathbf{y}, \mathbf{c})\, d\mathbf{NPP} \\ &\propto \int \left(\prod_{s=1}^{n} p_{\mathrm{LPJ}}(\mathrm{NPP}_s|c_s) \right) p(\mathbf{y}|\mathbf{NPP}, \theta)\, p(\theta)\, d\mathbf{NPP} \end{aligned} \tag{3}$$

avec $p_{\mathrm{LPJ}}(\mathrm{NPP}_s|c_s)$ la distribution de la végétation moderne sachant le climat du XXème siècle.

L'obtention de la distribution a posteriori définie Eq. 3 n'est pas réaliste car l'intégration requise est nécessairement numérique dut à la définition implicite de $p_{\mathrm{LPJ}}(\mathrm{NPP}_s|c_s)$ et se fait sur un espace de dimension $\dim(\mathbf{NPP}) > 15000$ dans notre cas. Le calcul numérique demanderait des centaines de milliers de simulations suivant le modèle LPJ-GUESS ce qui représente un temps de calcul trop important.

Nous proposons de calibrer directement la relation en utilisant un seul jeu $\mathrm{NPP}_{s=1..n}$ (noté **npp**) simulées sous le climat actuel. C'est-à-dire de replacer l'a posteriori précédent par

$$p(\theta|\mathbf{y}, \mathbf{npp}, \mathbf{c}) = p(\theta|\mathbf{y}, \mathbf{npp}) \propto p(\mathbf{y}|\mathbf{npp}, \theta)\, p(\theta) \tag{4}$$

En court-circuitant l'intégration sur LPJ-GUESS nous ignorons la calibration de ses paramètres internes. Cet aspect forme une des perspectives majeures de l'approche et mérite de futurs développements.

La modélisation et l'inférence des différentes couches formant le modèles $p(\mathbf{y}|\mathbf{npp}, \theta)$ sont discutés dans Garreta et al (2009b). L'inférence est obtenue grâce à un algorithme de Monte Carlo par chaîne de Markov qui a été parallélisé sur un calculateur multi-cœur.

Reconstruction du climat à partir d'une séquence pollinique

Nous considérons l'inférence de la végétation et du climat passé à partir du modèle calibré précédemment et d'une séquence de $n+1$ échantillons polliniques. Chaque échantillon est supposé daté sans incertitude et noté Y_t avec t variant de t_0 le point le plus vieux à t_n le point le plus récent. Une reconstruction paléoclimatique bayésienne consiste à obtenir la distribution a posteriori suivante

$$\begin{aligned} &p(\mathrm{NPP}_{t_0:t_n}, C_{t_0:t_n}, \theta_3|y_{t_0:t_n}) \\ &\qquad \propto p(C_{t_0:t_n}|\theta_3)\, p(\theta_3)\, p_{\mathrm{LPJ}}(\mathrm{NPP}_{t_0:t_n}|C_{t_0:t_n}) \prod_{t=t_0}^{t_n} p(y_t|\mathrm{NPP}_t) \end{aligned} \tag{5}$$

A la seconde ligne, le premier terme est la distribution a priori du climat paramétrée par θ_3 et suivie de l'a priori sur θ_3. Le modèle choisit est une marche aléatoire gaussienne dont la variance du saut entre deux temps est proportionnelle à la distance en temps et à θ_3. Le troisième terme est la dynamique de végétation définie par la chaîne de conditionnement (Eq. 1)

$$p(\mathrm{NPP}_{t_0:t_n}|C_{t_0:t_n}) = p_{\mathrm{LPJ}}(\mathrm{NPP}_{t_0}|C_{t_0}) \prod_{t=t_1}^{t_n} p(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_{t-1})$$

Le quatrième terme est le produit des distributions $p(Y|\mathrm{NPP})$ (définies en calibration) appliquées aux échantillons $y_{t_0:t_n}$ et dont le paramètre $\theta = (\theta_1, \theta_2)$ a été sorti par intégration

pour prendre en compte les incertitude de calibration. i.e

$$\begin{aligned} p(y_t|\mathrm{NPP}_t) &= \int p(y_t|\mathrm{NPP}_t,\theta)\, p(\theta|\mathbf{y},\mathbf{npp})\; d\theta \\ &= \int p(y_t|X_t)\, p(X_t|V_t,\theta_2)\, p(V_t|\mathrm{NPP}_t,\theta_1)\, p(\theta|\mathbf{y},\mathbf{npp})\; dV_t\; dX_t\; d\theta \end{aligned} \tag{6}$$

A la première ligne, $p(y_t|\mathrm{NPP}_t,\theta)$ est le modèle utilisé en calibration et $p(\theta|\mathbf{y},\mathbf{npp})$ l'a posteriori de calibration. La seconde ligne explicite les couches cachées définissant le modèle végétation-pollen. Ce modèle est fait de deux couches représentant les variables végétation perturbée, V et pollen accumulé, X (Garreta et al, 2009b).

Le modèle général présenté Eq. 5 est connu comme un modèle à espace d'état. Ici, sa particularité réside dans le fait que la transition entre états latents (la végétation) est définie implicitement par LPJ-GUESS. De plus, la contrainte majeure en temps de calcul est donnée par le modèle dont le temps de calcul est proportionnel au temps $t_n - t_0$ donnée par les points de la carotte. Enfin, l'estimation des distributions $p(y_t|\mathrm{NPP}_t)$ demande une intégration sur un nombre non-négligeable de dimensions. Sous ces contraintes, la seule solution trouvée consiste à utiliser un filtre particulaire en deux passages. Le premier passage permet d'obtenir la distribution a posteriori lissée de des paramètres θ_3, le second passage fournit la distribution a posteriori filtrée de la végétation et du climat. Dans l'exposé oral nous illustrons plus en détail la stratégie adoptée.

Application en Suède

La calibration est réalisée sur un jeu de données Européen (voir Garreta et al, 2009). Il comprend N=1301 échantillons polliniques dont les taxons ont été regroupés en $k = 15$ groupes correspondant aux sorties de LPJ-GUESS. Ainsi, une données Y_s est un vecteur multinomial de dimension $k = 15$. L'algorithme de calibration est un algorithme de Metropolis within Gibbs avec adaptation de la variance des loi de proposition dans une étape préliminaire de chauffe. Grâce à la parallélisation de l'algorithme, l'obtention de simulations a posteriori ayant convergées demande seulement quelques jours de calcul.

Nous reconstruisons le climat à partir de 4 carottes prélevées dans des lacs situées au Sud de la Suède. Les sorties qui seront présentées lors de la conférence montrent principalement

- une bonne cohérence avec l'interprétation qualitative que l'on fait des séquences polliniques. Cela démontre la cohérence générale de la méthode.
- une sensibilité forte aux erreurs numériques liées à l'algorithme. Cette sensibilité est sûrement due aux approximations numériques utilisées pour rendre l'inférence faisable.

- une très grande différence entre les reconstructions qui proviennent de sites proches qui devraient montrer des dynamiques comparables. Cela souligne la nécessité d'aller plus loin dans la modélisation, par exemple en incluant de processus spatiaux dans la modélisation de la végétation.

Conclusion

Cette approche est importante en paléoclimatologie pour obtenir des reconstructions indépendantes de celles obtenues par les FT classiques et intégrant la connaissance récente en écologie. Elle pourrait aussi être étendue à des problématiques de reconstruction des dynamiques de végétation passées.

Elle requiert cependant un travail poussé en statistique concernant l'inférence de (ou en présence de) modèles composites, i.e modèles incluant un ou des modèles mécanistes. Pour cela nous nous intéressons à des approches de type ABC et émulation de modèle.

Remerciements

Une partie du travail en statistique est le fruit de collaborations avec Frédéric Mortier (CIRAD Montpellier), Joël Chadœuf (INRA Avignon) et John Haslett (TCD Dublin). Plus largement, je remercie les collègues du laboratoire BioSP de l'INRA d'Avignon, du groupe MIA "Modèle Hiérarchique Spatiaux" et du groupe "Environnement" de la SFdS pour leur confiance et les nombreux échanges que nous avons eut sur différents sujets. Pour la partie appliquée, j'ai été aidé par Simon Brewer (Univ. Wisconsin), Joël Guiot et Christelle Hély (CEREGE) et Paul Miller (Univ. Lund).

Bibliographie

[1] Smith, B., Prentice, I. C. et Sykes, M. T. (2001) Representation of vegetation dynamics in modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Global Ecology & Biogeography* 10:621–637.

[2] Garreta, V., Miller, P. A., Guiot, J., Hély, C., Brewer, S., Sykes, M. T., et Litt, T. (2009a) A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Climate Dynamics* doi: 10.1007/s00382-009-0629-1.

[3] Garreta, V., Mortier, F. et Chadœuf, J. (2009b) 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. 41èmes Journées de Statistique, SFdS, Bordeaux (2009), http://hal.inria.fr/inria-00386776/fr/

[4] Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York.