Stage de Master 2
Modélisation spatiale des écosystèmes microbiens lacustres : étude de la colonne d'eau

Superviseurs et localisations

Pierre Latouche, Louise Alamichel, Laboratoire de Mathématiques Blaise Pascal, LMBP.
David Martin, Laboratoire de Micro-organismes, Génomique et Environnement, LMGE.

Localisation : Campus des Cézeaux, Université de Clermont Auvergne, 63178 Aubière.
Durée : 4 à 6 mois.
Début : Mars 2027.

Contacts

david.martin@uca.fr, pierre.latouche@uca.fr, louise.alamichel@uca.fr

Contexte et motivations écologiques

Au cœur de l'Auvergne, les lacs subissent, de la même manière que les océans, une désoxygénation. Autrement dit, la quantité d'O_2 soluble dans l'eau, disponible au règne du vivant diminue fortement, engendrant un remodelage des écosystèmes macroscopiques mais aussi microscopiques (Jansen et al., 2024). Le règne du vivant microscopique intervient directement dans le captage des gaz à effet de serre atmosphériques (p. ex. CO_2 et CH_4) (Jiao et al., 2010). Or dans le contexte de la désoxygénation, la modification du règne du vivant engendre une dérégulation des flux de ces gaz entre atmosphère et eau. Ce projet s'attache à développer de nouveaux modèles capables de capturer finement les dépendances spatiales de ces écosystèmes aquatiques (Shi et al., 2022; Garcia et al., 2013). En reliant les structures cachées de ces écosystèmes aux mesures environnementales et aux émissions réelles, cette recherche ambitionne, à termes, d'identifier les espèces microbiennes associées aux fluctuations environnementales et à l'émission des gaz. Pour répondre à cette question, l'équipe MEB du LMGE suit l'évolution des espèces microbiennes au sein du lac d'Aydat (Aydat, Auvergne, France) selon : (i) une dimension spatiale : l'étude inclut un séquençage de l'ADN présent dans la colonne d'eau tous les mètres (c.à.d. de 0 à 14 mètres), et (ii) une dimension temporelle : ce processus de séquençage est réalisé toutes les 2 semaines. L'équipe a aussi quantifié (iii) l'expression des ARN messagers liée aux métabolismes anaérobies nous donnant une indication indirecte de la production de méthane et de dioxyde de carbone, et (iv) le flux de ces deux gaz le long de la colonne d'eau et à l'interface atmosphère-lac (c.à.d. ce qui est rejeté versus ce qui est capté). L'acquisition des données qui seront utilisées pour ce projet a été réalisée entre janvier 2025 et janvier 2026, permettant d'embrasser les fluctuations environnementales liées à la saisonnalité.

Projet

Dans ce contexte, nous définirons $\mathbf{Y}_{i,\cdot} \in \mathbb{N}^{N_J}$ comme le vecteur de comptage de toutes les N_J espèces à l'observation $i \in \{1, \ldots, N_I\}$, où N_I est le nombre total d'observations ($N_I = 14$ soit la profondeur maximale du lac en mètres) et N_J est le nombre total d'espèces observées. Ces données sont habituellement normalisées dans la littérature en données compositionnelles, censé limiter les biais d'analyse liés à la profondeur de séquençage (effort d'échantillonnage). Malheureusement, les normalisations usuelles sont mal adaptées à leur nature discrète et sur-dispersée. Le modèle Poisson log-normal (PLN) (Aitchison and Ho, 1989) et sa variante PLN-PCA (Chiquet et al., 2018) sont des modèles à variables latentes qui fournissent à la fois une normalisation adaptée et un cadre d'analyse de ces données (Batardière et al., 2024).

L'enjeu ici est de développer de nouveaux modèles graphiques probabilistes ainsi que les stratégies d'inférence associées, afin de s'adapter aux caractéristiques singulières des données de comptage. Il s'agira dans un premier temps de développer

des méthodologies nouvelles s'appuyant sur des modèles graphiques Gaussiens et modèles pour les données de comptage avec et sans normalisation et/ou transformation, pour prendre en compte la dépendance spatiale présente dans les données. Dans un premier temps, nous ferons le postulat qu'il existe des groupes d'observations similaires. Autrement dit, on s'attend à ce que deux vecteurs de comptage correspondant à des observations proches spatialement soient plus similaires que deux vecteurs de comptage provenant de observations éloignés. Nous associons alors à chaque observation i une variable latente de classe $\mathbf{Z}_i$.

Mots clés du projet. Nouvelles stratégies d'inférence, Modèles à variables latentes en écologie, Statistiques en grande dimension, Données de comptage.

Objectifs

L'objectif de ce stage est de développer un nouvel outil de modélisation de données écologiques prenant en compte la dépendance spatiale des observations. Pendant le stage, il s'agira de :

(i) Réaliser une veille scientifique sur les modèles à variables latentes déployés en écologie spatiale + Définir des postulats et hypothèses écologiques sur lesquelles la proposition du modèle sera construite.

(ii) Modéliser et inférer la dépendance spatiale entre les observations.

(iii) Mener une étude de simulation pour différents scénarios de dépendance spatiale entre les observations et étudier le comportement du modèle proposé.

(iv) Appliquer le modèle aux données réelles et confronter la sortie obtenue, soit la variable de classe latente $\mathbf{Z}_i$, avec les covariables environnementales pour déterminer si l'inférence de $\mathbf{Z}_i$ est cohérente d'un point de vue biologique et écologique.

Profil recherché

Étudiant(e) en Master 2 (ou école d'ingénieurs) en Mathématiques appliquées ou Statistiques appliquées. L'étudiant(e) devra être doté(e) de solides compétences en mathématiques et en analyse de données sous R. Une appétence pour les sciences écologiques et biologiques est un plus.

Références

John Aitchison and C. H. Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4) :643–653, 1989. doi : $10.1093/\mathrm{biomet}/76.4.643$.

Bastien Batardière, Joon Kwon, and Julien Chiquet. pyPLNmodels : A Python package to analyze multivariate high-dimensional count data. *Journal of Open Source Software*, 9(104) :6969, 2024. doi : $10.21105/\mathrm{joss}.06969$.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4) :2674–2698, 2018. doi : $10.1214/18\text{-}\mathrm{AOAS}1177$.

Sarahi L. Garcia, Ivette Salka, Hans-Peter Grossart, and Falk Warnecke. Depth-discrete profiles of bacterial communities reveal pronounced spatio-temporal dynamics related to lake stratification. *Environmental Microbiology Reports*, 5(4) : 549–555, 2013.

Joachim Jansen, Gavin L. Simpson, Gesa A. Weyhenmeyer, Laura H. Härkönen, Andrew M. Paterson, Paul A. del Giorgio, and Yves T. Prairie. Climate-driven deoxygenation of northern lakes. *Nature Climate Change*, 14(8) :832–838, 2024.

Nianzhi Jiao, Gerhard J. Herndl, Dennis A. Hansell, Ronald Benner, Gerhard Kattner, Steven W. Wilhelm, David L. Kirchman, Markus G. Weinbauer, Tingwei Luo, Feng Chen, and Farooq Azam. Microbial production of recalcitrant dissolved organic matter : long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, 8(8) :593–599, 2010.

Jiaxin Shi, Baogang Zhang, Jun Liu, Yun Fang, and Aijie Wang. Spatiotemporal dynamics in microbial communities mediating biogeochemical cycling of nutrients across the xiaowan reservoir in lancang river. *Science of The Total Environment*, 813 : 151862, 2022.