Prédiction spatio-temporelle par équations différentielles partielles stochastiques.

Thomas Romary, Nicolas Desassis, Xavier Freulon, Denis Allard

Contexte et enjeux

Dans un contexte de transition écologique il est crucial de disposer d'outils d'analyse et de prédiction de l'évolution des milieux naturels et des variables climatiques pour la prise de décision et la gestion des mesures d'atténuation ou d'adaptation. De nombreux domaines des sciences environnementales cherchent à prédire dans l'espace-temps une variable d'intérêt à partir d'observations en certains points d'un domaine d'étude (spatio-temporel) et de variables explicatives (appelées covariables) connues exhaustivement. Par ailleurs l'explosion informatique et les progrès technologiques des instruments de mesure nous ont fait passer de la gestion de la rareté des données à la gestion de leur abondance. Les méthodes numériques doivent être repensées pour traiter de façon efficace ces jeux de données de très grande taille.
Pour cela on dispose de deux approches complémentaires :

1. Les approches classiques de régression (par exemple Random Forest) cherchent à prédire la variable d'intérêt en apprenant des relations, parfois complexes, entre les covariables et la variable d'intérêt. Dans un contexte spatio-temporel, les coordonnées des points de mesure sont alors généralement vues comme une covariable supplémentaire.
2. La géostatistique et les statistiques spatio-temporelles modélisent la structure de dépendance spatio-temporelle de la variable d'intérêt (ou d'un résidu de celle-ci après avoir tenu compte de la tendance pilotée par les covariables) afin de prédire la variable d'intérêt en un lieu et/ou un temps où elle n'a pas été mesurée ; en outre elle fournit une quantification de l'incertitude liée à cette prédiction.

Cependant, les statistiques spatio-temporelles se sont longtemps limitées à l'hypothèse d'une structure stationnaire dans l'espace-temps. Les enjeux de la thèse sont donc de tirer partie de la richesse des jeux de données actuels, ce qui permet de relaxer cette hypothèse de stationnarité et ainsi améliorer la qualité des prédictions des méthodes géostatistiques.

Verrous scientifiques

Dans un cadre non-stationnaire, de nombreuses approches ont été développés pour modéliser ces variations spatiales de structure, cf. Fouedjio (2017) ou Schmidt (2020) pour une revue. L'approche SPDE (Stochastic partial differential equations, Lindgren et al., 2011) permet d'incorporer facilement ces non-stationnarités en faisant varier dans l'espace et dans le temps les coefficients d'un opérateur différentiel.

C'est sur cette approche SPDE (Lindgren, 2011) que nous proposons de nous appuyer pour parvenir à des méthodes de prédiction spatio-temporelles efficaces dans un cadre non-stationnaire. Les travaux engagés dans l'équipe Géostatistique, dont certains en collaboration avec l'unité Biostatistique et Processus Spatiaux (BioSP), de l'INRAE (Avignon), sont en pointe dans le domaine.

Dans le cadre spatial, des avancées mathématiques et algorithmiques majeures (Carrizo et al, 2018 ; Pereira & Desassis, 2018 ; Pereira & Desassis, 2019) ont été accomplies, permettant de traiter de façon efficace des jeux de données de très grande taille. Par ailleurs la thèse de Ricardo Carrizo-Vergara (2018) a permis de définir de nouveaux modèles spatio-temporels dans ce cadre, incorporant les processus physiques liés aux phénomènes étudiés (convection, diffusion,...).

Nous savons actuellement simuler ces modèles mais les problèmes liés à l'inférence et au conditionnement par les données observées restent entiers.

L'objectif de ce projet de thèse est donc de proposer des méthodes efficaces pour l'inférence et la prédiction dans un cadre spatio-temporel, non stationnaire, basé sur l'approche SPDE.

Démarche et méthode

L'adaptation d'algorithmes d'assimilation de données est une première piste identifiée pour le conditionnement et l'inférence de modèles spatio-temporels (cf. Katzfuss et al. (2019), Chau et al. (2018)). Ces approches pourraient permettre de réduire la complexité, notamment liée à la quantité de données, en décomposant le problème par conditionnement séquentiel. Il sera néanmoins nécessaire d'adapter ces méthodes au cadre SPDE.

La thèse de Mike Pereira (2019) a permis le développement de méthodes numériques adaptées au cas spatial, en particulier la possibilité de travailler sur des variétés différentiables et avec des modèles dont la structure de dépendance peut varier spatialement. Le transfert de ces techniques dans un cadre spatio-temporel est une seconde voie à développer. Se pose en particulier la question d'employer alors un maillage dynamique, lorsque les données disponibles sont issues de capteurs mobiles notamment.

Pour la modélisation de la non-stationnarité, les variations de la structure de dépendance peuvent être incorporées comme des fonctions des coordonnées à inférer, comme dans Fulgstad et al. (2015). Elles peuvent également être des fonctions de covariables qui contrôlent la structure de dépendance modélisée par l'opérateur différentiel, comme par exemple la vitesse et la direction du vent dans le cas de la concentration atmosphérique d'un polluant.

Impact et résultats attendus

Un premier cadre d'application concerne la prédiction de la qualité de l'air en incluant des données de micro-capteurs, en lien avec une thèse déjà en cours en partenariat avec l'INERIS. Le cadre d'application de ces travaux ne se limite pas aux données de qualité de l'air. Ce type d'approche peut s'appliquer dans un grand nombre de domaines des géosciences, par exemple le climat, la qualité de l'eau au sein des nappes phréatiques, la quantification de la ressource hydrique, le suivi de données de sol, notamment l'évaluation des stocks de carbone dans les sols. En ce sens, ces travaux s'inscrivent parfaitement dans le cadre du projet de chaire mené avec l'équipe BioSP.

Au-delà des cadres d'application, cette thèse doit déboucher sur un ensemble de résultats mathématiques et de méthodes sous forme algorithmique. Les premiers seront publiés dans les revues scientifiques du domaine. Les algorithmes seront codés et diffusés à travers des librairies et packages ouverts.

Encadrement

La thèse sera dirigée par Denis Allard, Nicolas Desassis et Thomas Romary, en collaboration avec Xavier Freulon et Mike Pereira. L'étudiant sera accueilli dans l'équipe Géostatistique à Fontainebleau, avec des visites fréquentes à BioSP.

Compétences et connaissances requises

De bonnes connaissances en probabilités, statistiques et analyse numérique sont requises, ainsi qu'un intérêt prononcé pour les applications en sciences de l'environnement. Un gout pour la programmation numérique est demandé, et notamment une bonne connaissance des langages C, R et/ou python. La maîtrise de la langue anglaise est aussi nécessaire.

Modalités de candidature

Une lettre de motivation, un descriptif des travaux de stage de Master 2, les résultats d'examen de Master 1 et 2, ainsi que deux lettres de recommandation ou deux référents constitueront les pièces à apporter au dossier de candidature et à envoyer à thomas.romary@mines-paristech.fr

Références

Carrizo-Vergara, R., Allard, D., & Desassis, N. (2018). A general framework for SPDE-based stationary random fields. arXiv preprint arXiv:1806.04999.

Carrizo-Vergara, R., (2018) Development of geostatistical models using stochastic partial differential equations, PhD thesis, December 2018. http://www.theses.fr/2018PSLEM062

Chau, T. T. T., Ailliot, P., Monbet, V., & Tandeo, P. (2018). An efficient particle-based method for maximum likelihood estimation in nonlinear state-space models. arXiv preprint arXiv:1804.07483.

Fouedjio, F. (2017) Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, 31(8). 1887-1906.

Fuglstad G.A. , Lindgren F., Simpson D., and Rue H. (2015) Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, 115-133.

Matthias Katzfuss, Jonathan R. Stroud & Christopher K. Wikle (2019) Ensemble Kalman Methods for High-Dimensional Hierarchical Dynamic Space-Time Models, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1592753

Lindgren F., Rue H., Lindström J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B*, 73(4) :423-498

Pereira, M., Desassis, N. (2018). Finite element approximation of non-Markovian random fields. arXiv preprint arXiv:1811.03004.

Pereira, M., Desassis, N. (2019). Efficient simulation of Gaussian Markov random fields by Chebyshev polynomial approximation. *Spatial Statistics*, 31, 100359.

Mike Pereira. Champs aléatoires généralisés définis sur des variétés riemanniennes : théorie et pratique. PhD thesis, November 2019. http://www.theses.fr/s172855.

Schmidt, A. M., & Guttorp, P. (2020). Flexible spatial covariance functions. Spatial Statistics, https://doi.org/10.1016/j.spasta.2020.100416.