Ecole Doctorale Carnot-Pasteur

Proposition de sujet de thèse

Intitulé français du sujet de thèse proposé :

Nouveaux modèles statistiques pour données fonctionnelles motivés par des applications en océanographie

Intitulé en anglais du sujet de these proposé :

New statistical models for functional data motivated by applications in oceanography

Unité de recherche : IMB (UMR 5584, Université Bourgogne Europe & CNRS)

Nom, prénom et courriel du directeur (et co-directeur) de thèse :

Directeur : Cardot, Hervé, herve.cardot@ube.fr (Institut de Mathématiques de Bourgogne, Dijon)
Co-directeur : Nerini, David, david.nerini@univ-amu.fr (Institut Méditerranéen d'Océanologie, Marseille)

Domaine scientifique principal de la thèse :

Mathématiques Appliquées, Statistique

Domaine scientifique secondaire de la thèse :

Océanographie, Sciences de l'Environnement

Description du projet scientifique :

L'océan mondial couvre 71% de la surface de la Terre et constitue un milieu difficile d'accès. Les études dans l'océan ouvert sont de plus en plus assistées par des systèmes autonomes capables d'échantillonner l'océan en temps continu et de transmettre des données simultanément. Satellites, animaux équipés, bouées dérivantes se substituent de plus en plus aux grandes campagnes océanographiques, limitées dans le temps et assurées par des navires. L'ensemble de ces nouvelles mesures haute-fréquence, qui peuvent vues comme des données fonctionnelles car échantillonnées le long d'un continuum spatial ou temporel, portent sur un grand nombre de variables et engendrent des volumes considérables de données. Beaucoup de ces données sont sous exploitées et souvent stockées dans l'attente d'être traitées ultérieurement.

Cette thèse porte sur le développement et l'étude des propriétés de nouveaux modèles statistiques pour données fonctionnelles, motivés par des questions de recherche en océanographie. Plusieurs pistes, connectées entre elles, pourront être explorées.
La base de données MEOP (Roquet et. Al., 2014) rassemble depuis plus de 20 ans des données océanographiques échantillonnées dans l'Océan Austral par des éléphants de mer (SES) équipés de balises. L'étude des trajectoires SES est réalisée à l'aide de données GPS, ce qui permet d'évaluer la vitesse de progression de l'animal et de détecter des portions de trajectoires particulières où ces animaux vont évoluer au sein de gyres océaniques pour se nourrir. Pour mettre en relation ces comportements alimentaires avec les données océanographiques, on dispose de champs de courant estimés à l'aide de données satellitaires aux dates correspondants aux passages de l'éléphant de mer. Cependant, ces données satellitaires sont partielles puisqu'échantillonnées dans une zone de l'Antarctique souvent recouverte de nuages. Il faut donc reconstituer le champ de vecteurs vitesse de manière fine à l'aide de méthodes d'approximation. Nous proposons de développer des approches basées sur des modèles nonparamétriques construits via des splines pénalisées par des normes d'opérateurs différentiels qui permettront d'effectuer un compromis entre ajustement aux données et respect des équations de la physique (normes sur la divergence ou le rotationnel du champ de vecteur par exemple). De manière plus ambitieuse, nous souhaiterions pouvoir combiner ces décompositions sur fonctions splines avec les outils récents développés dans le domaine des données fonctionnelles pour la reconstruction de trajectoires observées partiellement.
Une autre question d'intérêt mathématique et océanographique porte sur la modélisation des trajectoires des plongées des éléphants de mer. Les SES plongent continument et profondément le long de leur trajet en mer qui peut durer des mois. Les formes de ces trajectoires de plongée permettent d'étudier leur comportement (chasse, repos, ….) et sont également contraintes par la physique de l'environnement dans lequel ils évoluent. La décomposition (linéaire) de Karhunen-Loève, des grands modes de variabilité de ces trajectoires qui repose sur une approximation « optimale » dans un espace vectoriel de dimension finie, n'est sans doute pas la plus pertinente dans ce contexte. Il s'agira, en vue d'une représentation optimale d'un échantillon de trajectoires, de développer des techniques d'analyse, de type analyse en composante principales, qui respectent ces contraintes physiques locales. On pourra pour cela s'appuyer sur des travaux récents (ACP de densités, recalage de trajectoires, …) qui proposent de quitter le cadre des espaces de Hilbert pour celui des variétés sur lesquelles il s'agira de déterminer la métrique la mieux adaptée.

Références bibliographiques (sélection)

Dai, X., Mueller, H.G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics* **46**, 3334-3361.

Fonvieille, N. et al. (2023). Swimming in an ocean of curves: A functional approach to understanding elephant seal habitat use in the Argentine Basin. *Progress in Oceanography*, 218, p. 103120

Kneip, A., Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *Annals of Statistics,* 48, 1692-1717.

Le Guyader , C., Gout, C., Macé, A-S., Apprato, D. (2013). Gradient field approximation : Application to registration in image processing. *Journal of Computational and Applied Mathematics*. 240, 135-147.

Palummo, A., Arnone, E., Formaggia, L., Sangalli, L-M. (2024). Functional principal component analysis for incomplete space-time data. *Environmental and Ecological Statistics*, 31, 555-582.

Petersen, A. and Mueller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44, 183–218.

Roquet, F., Williams, G., Hindell, M. *et al.* (2014) A Southern Indian Ocean database of hydrographic profiles obtained with instrumented elephant seals. *Sci Data* **1**, 140028.

Srivastava, A. and Klassen, E-P (2016). Functional and shape data analysis. Springer Series in Statistics. Springer-Verlag, New York.

Wu, Y, Huang, C., Srivastava, A. (2024). Shape-based functional data analysis. *TEST*, 33, 1-47.

Connaissances et compétences requises :

Master en modélisation statistique/mathématiques appliquées avec des bases solides en statistique. Programmation avancée R ou Python. Manipulation de données massives (base de données MEOP ou COPERNICUS).