Classification non-supervisée pour l'identification de paysages acoustiques homogènes

Projet de thèse proposé par V. Audigier, F. Bouhadjera et N. Niang

CNAM, Laboratoire CEDRIC, équipe MSDMA

2 rue Conté, 75003 Paris

Keywords : classification non-supervisée, données hétérogènes, données massives, données fonctionnelles, données manquantes, océanographie.

1 Introduction

La prévision océanographique opérationnelle consiste à modéliser le champ spatio-temporel des différents paramètres physiques et chimique de l'océan : température, densité, courants, hauteurs de vagues. Pour le besoin de l'acoustique sous-marine, la grandeur d'intérêt est la vitesse du son dans l'eau, qui se calcule à partir des prévisions de température, de densité et de salinité par une équation d'état. Les grandeurs élémentaires auxquelles s'applique cette recherche sont des profils verticaux de célérité géo-référencés à des profondeurs standards. Ces profils verticaux définissent des guides d'ondes qui régissent la propagation acoustique dans les océans.

Les volumes d'information générés par les modèles océaniques sont trop importants pour être utilisés dans certaines applications contraintes par la transmission de petits volumes de données. On s'intéresse dans cette recherche à obtenir des représentations réduites. On peut poser le problème de représentation géostatistique réduite comme un problème de classification d'une population d'individus (couples latitude $\times$ longitude) à partir d'un tableau de données noté X_{pred}. Chaque individu est défini par un vecteur de scalaires (qui peut être représenté sous la forme d'une trajectoire, voir Figure 1), par une latitude, par une longitude et par un temps t. Plutôt que de transmettre l'ensemble de la prévision profil par profil, on envisage de ne transmettre que certains profils caractéristiques d'un groupe homogène ainsi que leur domaine d'extension géographique et temporelle.

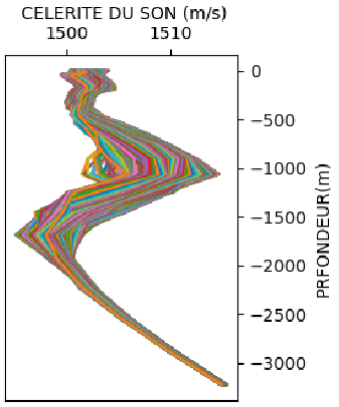

Figure 1: Exemple de profils de célérité.

Ce type de problème peut être résolu par des algorithmes de classification non-supervisée (réseaux de neurones profonds, K-means, cartes topologiques, etc). Cependant, ces approches reposent seulement sur la forme des profils, ce qui ne garantit pas que des profils proches aient un comportement proche du point de vue de la propagation acoustique. Il est possible d'associer aux profils verticaux de célérité des grandeurs acoustiques discriminantes comme par exemple la fréquence de coupure de chenaux, la distance de résurgence des ondes acoustiques, des pertes de propagation en fonction de la distance horizontale,... Il est intuitif de

penser qu'une classification conjointe des profils verticaux à partir de X_{pred} et des variables acoustiques associées, que l'on désigne par Y_{pred}, pourrait améliorer la pertinence de la classification .

Ainsi, le but de cette recherche est de définir des méthodes de représentations réduites, reposant sur des techniques de classification, qui ne soient pas seulement fondées sur des analyses de variables océanographiques mais qui inclut aussi des variables acoustiques associées. Les groupes établis par ces méthodes pourront être interprétés comme des paysages acoustiques homogènes.

2 Objectifs

La thèse proposée traitera du problème de la classification non-supervisée à partir des données de célérité (X_{pred}) et des données acoustiques (Y_{pred}) sous contrainte de continuité spatiale à un temps t fixé. Cette tâche revêt plusieurs aspects méthodologiques complexes en lien avec la nature des données :

1. données de nature hétérogène : les objets X_{pred} et Y_{pred} décrivent tous deux des couples latitude $\times$ longitude, mais ceux-ci sont de dimensions différentes : X_{pred} caractérise les couples par des profils de célérité (qui sont des trajectoires), tandis que Y_{pred} caractérise les couples par des mesures ponctuelles.

2. trajectoires de longueurs différentes : la longueur d'un profil de célérité est étroitement liée à la distance entre la surface et le fond marin. Or, cette distance varie de façon évidente. Ainsi, les trajectoires qu'il s'agit de classifier ne sont pas toutes de la même longueur, ce qui pose des problèmes méthodologiques à résoudre.

3. données massives : le nombre de couples latitude $\times$ longitude est de l'ordre de plusieurs millions. La tâche de classification nécessitera de pouvoir passer à l'échelle.

4. données mixtes : les données acoustiques contenues dans Y_{pred} sont de nature variée, il peut s'agir de données de présence/absence, de valeur numérique, de comptage, etc. Ainsi, la tâche de classification nécessite aussi de tenir compte de cette variété dans la nature des variables.

5. données manquantes : les données acoustiques contenues dans Y_{pred} ne sont pas toutes disponibles. Cela est notamment lié à l'existence ou non de certaines caractéristiques de propagation comme le chenal de surface. Il conviendra de mener une réflexion pour évaluer le traitement adéquat de ces valeurs manquantes.

L'objectif de la thèse est de proposer une méthode de classification prenant en compte ces différentes aspects méthodologiques. On pourra faire appel à des méthodes de classification de données fonctionnelles, des méthodes d'analyse multiblocs, des méthodes d'imputation de données manquantes, etc... Le recours à des distances adaptées pourra permettre de gérer le caractère mixte des données ainsi que les variations dans les longueurs des trajectoires.

3 Calendrier prévisionnel

La thèse pourra débuter dans les deux mois succédant la sélection du candidat. Le travail de thèse se déroulera selon les étapes suivantes

- Étape 1 (9 mois) : exploration des données et étude bibliographique détaillée sur la gestion des données manquantes et la classification de trajectoires.

- Étape 2 (12 mois) : proposition d'une nouvelle méthodologie pour la classification de trajectoires issues des profils de célérité en tenant compte des variables acoustiques de même nature et complètes.

- Étape 3 (9 mois) : extension de la méthodologie proposée au cas de variables acoustiques mixtes et incomplètes

Chacune des propositions fera l'objet d'une publication intégrant une application à des données océanographiques.

Les 6 derniers mois seront consacrés à la rédaction de la thèse.

4 Moyens consacrés

Les moyens tant matériels (informatiques) qu'humains (compétences propres) du laboratoire d'accueil (CEDRIC/CNAM) seront mis à disposition du doctorant.

5 Profil du candidat

Candidat disposant d'un Master 2 ou d'un diplôme d'ingénieur dans le domaine des mathématiques, de la statistique, ou de la science des données. Un bon niveau en analyse des données, en programmation, ainsi que des capacités à rédiger en Français et en Anglais sont attendues. Nécessité d'être présent sur le site.

6 Modalités de candidature

Les dossiers de candidatures devront être composés d'un cv détaillé, présentant l'adéquation du candidat par rapport au sujet, d'une lettre de motivation mettant en évidence les raisons de la candidature, ainsi que des relevés de notes associés au diplôme le plus élevé. Le dossier pourra être accompagné de lettres de recommandation. Ces éléments devront être transmis par mail aux trois adresses suivantes : vincent.audigier@cnam.fr ; feriel.bouhadjera@lecnam.net ; ndeye.niang_keita@cnam.fr

References

[1] Charles Bouveyron, Julien Jacques, Amandine Schmutz, Fanny Simoes, and Silvia Bottini. Co-clustering of multivariate functional data for the analysis of air pollution in the south of france. *The Annals of Applied Statistics*, 16(3):1400–1422, 2022.

[2] Yosra Ben Slimen, Julien Jacques, and Sylvain Allio. Co-clustering for binary and functional data. *Communications in Statistics - Simulation and Computation*, 2020.

[3] Gholamreza Soleimani and Masoud Abessi. Dlcss: A new similarity measure for time series data mining. *Engineering Applications of Artificial Intelligence*, 92:103664, 2020.

[4] Julien Jacques and Cristian Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8:231–255, 2014.

[5] Mimi Zhang and Andrew Parnell. Review of clustering methods for functional data. *ACM Transactions on Knowledge Discovery from Data*, 17(7):1–34, 2023.