Proposition de stage de Master 2 en Statistique Appliquée (2023-2024)

Périodisation archéologique à l'aide d'une méthode de classification ascendante hiérarchique par compromis : Etude de robustesse et généralisation

CV, lettre de motivation et coordonnées d'au moins une référence à adresser à : lise.bellanger@univ-nantes.fr

Contexte

Le développement des approches analytiques, automatisées et statistiques pour l'étude des **données en archéologie** a commencé dès les années 1970, mais surtout dans les années 1990 avec le développement et la généralisation de la micro-informatique. Parallèlement, l'accroissement exponentiel des fouilles, notamment avec le développement de l'archéologie préventive, a entrainé la production de données de plus en plus massives, réclamant le développement de méthodes d'analyses adaptées.

Objectifs

Le Laboratoire de Mathématiques Jean Leray de Nantes (LMJL, UMR 6629 CNRS, Nantes Université) et le Laboratoire Archéologie et Territoires (CITERES-LAT UMR 7324 CNRS, Université de Tours) ont engagé depuis de nombreuses année une **collaboration de recherche interdisciplinaire** qui a donné lieu à de nombreuses publications ; mais aussi au développement d'un package R (SPARTAAS ; Coulon, Bellanger, Husi 2023) ; et à des applications web (R Shiny). Un algorithme de **Classification Ascendante Hiérarchique (CAH) par compromis** a été développé, à l'origine pour faciliter l'établissement de la chronologie et de la périodisation des sites archéologiques ; mais il a depuis aussi été mis en œuvre avec succès dans le cadre d'un projet de recherche sur la détection des troubles de la marche chez des patients atteints de Sclérose en Plaques (Drouin et al. 2022). Il permet de déterminer une partition compromis entre deux sources d'information (par exemple : céramique et stratigraphique) (Bellanger, Coulon, Husi 2021a et 2021b) et est implémenté dans le package R SPARTAAS. Cependant, dans l'approche actuelle seules deux sources d'information peuvent être prises en compte, ce qui reste trop restrictif par rapport à la réalité archéologique et aux nombreux types de données mobilisables ; d'où l'importance de la généraliser à une nombre de sources supérieur.

Les missions du stagiaire seront les suivantes :

(i) **étudier la robustesse de la méthode CAH par compromis** proposée en se confrontant à d'autres méthodes de classification non supervisée de la littérature (e.g. Hulot et al. (2020) à l'aide de données archéologiques et de données simulées,

(ii) **étendre cette approche à plus de deux sources**.

Différents corpus de données archéologiques, provenant de l'étude de la céramique - source matérielle omniprésente en archéologie - pourront être mobilisés (fouilles archéologiques réalisées à Angkor Thom, capitale de l'empire khmer (9e et le 15e s.) ou dans le bassin de la Loire Moyenne).

Mots clés : classification non supervisée, classification ascendante hiérarchique, apprentissage semi-supervisé, compromis, données multi-vues, archéologie, périodisation.

Références :

Aggarwal C., Reddy C. 2014 - Data Clustering: Algorithms and Applications. Boca Raton: Chapman and Hall/CRC.

Bellanger L., Coulon A., Husi P. 2021a - PerioClust: a new Hierarchical agglomerative clustering method including temporal or spatial ordering contraints, *in :* Chatzipantelis Th. *et al.* (ed.), *Data Analysis and Rationality in a Complex World*, XXIII, Springer.

Coulon A., Bellanger L., Husi P. 2023 - SPARTAAS: Statistical Pattern Recognition and dating using Archaeological Artefacts assemblageS. R package version 1.2.1, https://CRAN.R-project.org/package=SPARTAAS.

Drouin P., Stamm A., Chevreuil L., Graillot V., Barbin L., Gourraud P.-A., Laplaud D.-A., Bellanger L. (2022) Semi-supervised clustering of quaternion time series: application to gait analysis in multiple sclerosis using motion sensor data . *Statistics in Medicine*, 1-24.

Husi P. (dir.) 2022 – *La céramique médiévale et moderne du bassin de la Loire moyenne, chrono-typologie et transformation des aires culturelles dans la longue durée (6e-19e s.),* Tours, FERACF, URL : https://ceramedvaldeloire.huma-num.fr/editions/suppl79racf2022

Hulot A., Chiquet J., Jaffrézic F. et al. 2020 - Fast tree aggregation for consensus hierarchical clustering. BMC Bioinformatics 21, 120.

Profil du candidat

- Bac +5 : Master 2 ou Ecole d'Ingénieur en statistique ;
- Connaissances approfondies des méthodes de machine learning et de Science des données ;
- Maîtrise avancée du langage de programmation R ainsi que des outils R Markdown, Shiny, Quarto et Latex pour la création de documents, interfaces et rapports ;
- Capacités rédactionnelles et lecture d'articles scientifiques en anglais ;
- Autonomie, rigueur, grande curiosité, intérêt pour le travail interdisciplinaire, capacités de vulgarisation.

Précisions pratiques

- *Lieu :* Laboratoire de Mathématiques Jean Leray Nantes et CITERES-LAT Tours
- *Gratification* : environ 600 €/mois ; Durée du stage : 5 à 6 mois (date de démarrage en 2024 à discuter) ;
- *Encadrement* : Lise Bellanger (lise.bellanger@univ-nantes.fr, Univ Nantes), Arthur Coulon et Philippe Husi CITERES-LAT(philippe.husi@univ-tours.fr)