# Recherche de motifs dans des données archéo-environnementales

## Informations générales

### Encadrement

Lionel Tabourier (ComplexNetworks, LIP6, Sorbonne Université/CNRS)
Yannick Miras (Histoire Naturelle de l'Homme Préhistorique, MNHN/UPVD/CNRS)
Jerry Lonlac-Konlac (CERI SN, IMT Nord-Europe, Université Lille Douai)
**Contact :** `lionel.tabourier@lip6.fr`

### Localisation

LIP6, Sorbonne Université, 4 Place Jussieu, 75005 Paris.

### Durée et remuneration

Jusqu'à 6 mois. La rémunération du stage est basée sur les taux standards des stages académiques (4.05 euros/h, 35 h/semaine), plus une subvention pour les repas et le transport.

### Profil

Le stage est destiné aux étudiants de Licence 3 ou Master (préférentiellement Master) ayant une formation en informatique. De bonnes capacités de programmation et d'algorithmique sont nécessaires, ainsi qu'une bonne connaissance d'un langage multi-usage tel que le python. Une formation en fouille de données est très préférable, mais pas absolument nécessaire. Enfin, un goût pour l'interdisciplinarité, en particulier avec les sciences humaines et les sciences de l'environnement est essentiel pour ce stage.

## Description scientifique

### Contexte

On cherche à reconstituer les changements environnementaux passés à l'aide de données d'abondance de pollens, spores fongiques et autres bioindicateurs. En pratique, des carottages sont effectués dans les sédiments de milieux naturels dont les différentes profondeurs correspondent à différentes époques, que l'on évalue à l'aide de méthodes de datation. Les pollens sont comptés par des spécialistes de la discipline afin d'évaluer quelles espèces étaient présentes à quelle époque.

Du point de vue informatique, on peut étudier les motifs de co-évolution des différents pollens afin d'évaluer quelles espèces sont apparues ou disparues simultanément dans l'environnement. Cela permet par exemple d'identifier des marqueurs attestant de la présence humaine dans une région à une époque donnée. Dans ce contexte, on peut chercher à comprendre quels sont les facteurs qui expliquent les modifications de l'environnement : présence humaine, fluctuations climatiques, etc.

## Méthodologie

Le dénombrement de motifs de co-évolution est une méthode relativement peu utilisée en archéologie environnementale, les approches privilégiées reposant davantage sur l'analyse statistique multivariée des marqueurs présents ou non simultanément. Cependant, on pense que la co-évolution des marqueurs peut être une information utile pour identifier des indices secondaires jusqu'alors peu connus. C'est dans cette optique qu'est né le projet Mobipaléo, entre le laboratoire HNHP (MNHN) et le LIMOS (Université Clermont Auvergne). Dans le cadre de ce projet, un algorithme de recherche de motifs de co-évolution a été développé. Basé sur le problème célèbre de la détection d'*itemsets* fréquents [1], cet algorithme a été adapté aux propriétés spécifiques du contexte dynamique que nous considérons ici.

Cependant, cet outil n'a été utilisé pour l'instant que sur un jeu de données issu de carottages lacustres en Auvergne. Il a permi d'identifier des motifs encore inconnus qui seraient des traces indirectes de la présence humaine près de ces lacs [4, 5]. Des questions se posent sur l'utilisation efficace de cet outil : quels motifs spécifiquement viser ? comment l'adapter aux paramètres de différents contextes environnementaux, par exemple quel seuil de variations faut-il envisager [2] ? Pour répondre à ces question, il est nécessaire de l'utiliser sur d'autres jeux de données et de comparer ces résultats à ce que produisent des méthodes d'analyses statistiques multivariées, classiques dans ce domaine (dans la ligne de ce qui est proposé dans [3]).

En parallèle de cet objectif applicatif, le projet contient une composante algorithmique. Il s'agit d'améliorer les performances de la technique de recherche de motifs pour qu'elle puisse passer à l'échelle sur de plus grands jeux de données. Ce type de méthodes sont de complexité pire cas exponentielles en théorie, mais en pratique, on sait qu'elles peuvent être utilisées sur des jeux de données réels de relativement grandes tailles, en fonction des caractéristiques de ces données. C'est pourquoi il est pertinent d'adapter l'algorithme aux données employées et d'étudier expérimentalement les temps d'exécution.

## Objectifs

Les objectifs de ce stage sont les suivants :

— explorer l'algorithme de détection de motifs pour déterminer les ajustements de paramètres pertinents pour l'analyse,
— comparer ces résultats à ceux produits par des méthodes d'analyse statistique multivariée,
— l'adapter pour en améliorer la complexité expérimentale sur des données articifielles et réelles.

# Références


[1] Christian Borgelt. Frequent item set mining. *Wiley interdisciplinary reviews : data mining and knowledge discovery*, 2(6) :437–456, 2012.

[2] Michael Chirmeni Boujike, Jerry Lonlac, Norbert Tsopze, Engelbert Mephu Nguifo, and Laure Pauline Fotso. Grapgt : Gradual patterns with gradualness threshold. *International Journal of General Systems*, pages 1–21, 2023.

[3] Maureen Domche, Jerry Lonlac, Norbert Tsopzé, and Engelbert Mephu Nguifo. Une etude comparative entre motifs graduels et corrélations statistiques. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-40 :333–334, 2024.

[4] Jerry Lonlac, Yannick Miras, Aude Beauger, Vincent Mazenod, Jean-Luc Peiry, and Engelbert Mephu Nguifo. An approach for extracting frequent (closed) gradual patterns under temporal constraint. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2018.

[5] Yannick Miras, Jerry Lonlac, Beauger Aude, Benjamin Legrand, Karen K Serieyssol, Marlène Lavrieux, Paul M Ledger, Engelbert Mephu Nguifo, and Jean-Luc Peiry. Tracking plant, fungal and algal diversity through a data mining approach : towards an improved analysis of holocene lake Aydat (Puy-de-dôme, France) dynamics and ecological legacies. *Revue des Sciences Naturelles d'Auvergne*, 2021.