MATUREAU : du code monolithique au package pour un clustering sur mesure applicable aux séries temporelles hydrogéologiques

Marc LAURENCELLE* Théophile LOHIER* El Mamoun SQUALLI*
Violaine BAULT*

Résumé

La démarche MATUREAU, lancée en 2024 au BRGM, vise à renforcer l'accessibilité, d'abord interne, à des outils de traitement de données scientifiques liées à l'eau, par un effort de maturation de codes existants. Cette initiative se veut aussi une invitation adressée aux « hydro-développeurs » du BRGM, à adopter de meilleures pratiques de développement, notamment en termes de modularisation et packaging du code source, de standardisation des entrées-sorties, de proposition de chaînes de traitement de référence (workflows spécifiques à un cas d'étude, via des fonctions et/ou vignettes), de documentation, de tests fonctionnels pour assurer la maintenabilité des outils, … Plusieurs outils jugés prioritaires ont ainsi été sélectionnés et maturés, depuis 2024. L'outil présenté à ces Rencontres R est basé sur la notion de « clustering » appliquée à des séries temporelles.

Le clustering par k-médoïdes est apparu comme une méthode statistique particulièrement efficace pour partitionner un grand nombre de chroniques piézométriques (de niveau d'eau souterraine) en un nombre relativement réduit de groupes (clusters), en utilisant une matrice de corrélations (entre une présélection des séries temporelles assez complètes à l'intérieur d'une période d'intérêt) comme mesure de dissimilarité, en entrée de l'algorithme PAM (Maechler et al., 2023).

Dans cette présentation nous détaillerons l'approche suivie et montrerons comment la démarche MATUREAU, à travers la transformation d'un code d'origine monolithique en un package de fonctions faciles à enchaîner, a permis de gagner en interactivité, en reproductibilité et en pouvoir de généralisation.

Mots-clefs (3 à 5) : Statistique – Hydrogéologie – Package – Maturation – Refactoring

Développement

Une première mouture de cet outil a été développée sous la forme d'un script monolithique, pour répondre aux objectifs spécifiques de quelques études scientifiques. Puis, dans le cadre de la démarche MATUREAU, ce script a d'abord été modularisé de manière à séparer le chargement des données, leur préparation, leur traitement par l'algorithme de clustering, et la génération de diagnostics sur la qualité de la classification. Lors de cette phase, une attention particulière a été apportée à la définition d'un format de données standard (des objets en mémoire) pour les données issues des piézomètres. Une structure relationnelle a été adoptée de manière à pouvoir séparer les séries temporelles piézométriques, utilisées lors de la classification, des métadonnées pouvant apporter des détails descriptifs sur le contexte et les conditions dans lesquels elles ont été acquises (localisation du point d'eau, type de matériel, etc.), essentielles pour un diagnostic éclairé. Une fois ce format de référence défini, un ensemble de parsers a été développé de manière à assurer la cohérence entre des données issues de sources hétérogènes, pour construire des objets R de stockage bien structuré de ces données.

* BRGM, Orléans, m.laurencelle@brgm.fr

A l'issue de cette première étape de maturation du code, il s'est avéré que les résultats du clustering, uniquement basés sur les séries temporelles de niveaux piézométrique, étaient perfectibles grâce à l'intégration de connaissances expert. Ces connaissances ne peuvent pas être formalisées à travers des variables directement interprétables par l'algorithme de clustering. La solution proposée repose donc sur des aller-retours entre le processus de classification et l'expert qui peut intervenir (entre itérations de traitement) pour opérer des modifications manuelles des résultats de la classification (ex. créer un nouveau groupe, retirer des membres, changer le groupe attribué à un point en se basant par exemple sur la forte corrélation au médoïde d'un autre groupe). D'autres améliorations ont également été apportées dans l'optique de généralisation des fonctionnalités, notamment la possibilité de préparer des séries temporellement alignées par une agrégation à un pas de temps régulier au choix de l'utilisateur (n mois ou n jours ; était jadis fixé à 1 mois) pour s'adapter à la densité du jeu de données.

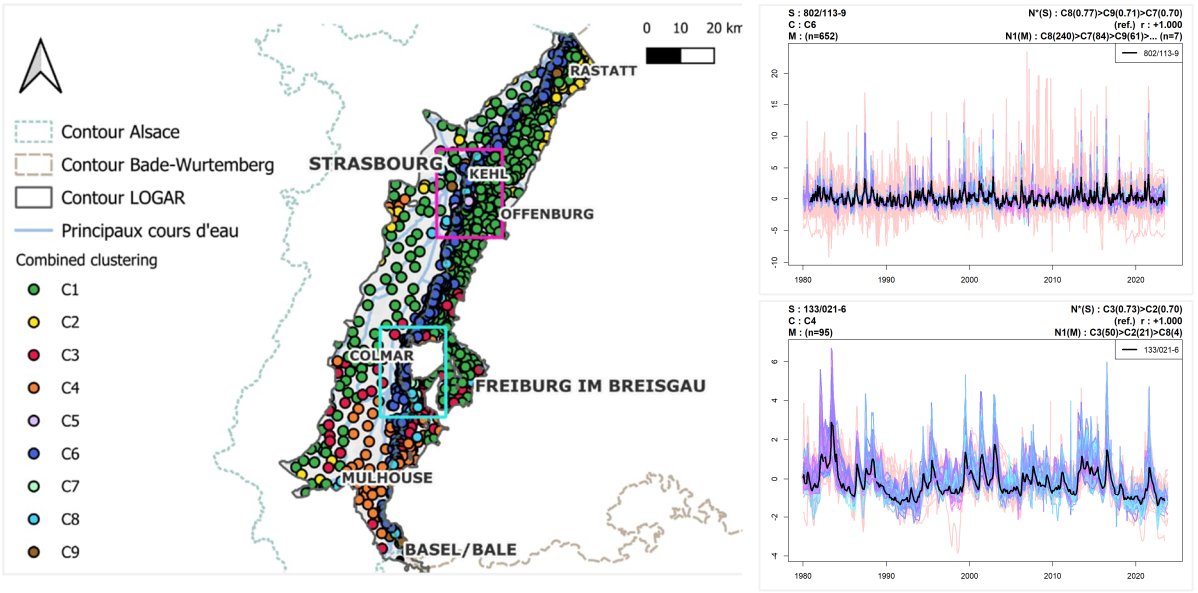

Figure 1 : Exemple de résultats : Carte de synthèse des résultats avec le regroupement final, projet Interreg GRETA (Laurencelle et al., 2026)

En somme, c'est grâce à ce long travail de maturation d'un outil développé au départ sans grand souci de modularisation et de généralisation, bonifié par des développements supplémentaires réalisés en cours de route, qu'on dispose aujourd'hui d'un package-outil de regroupement de points d'eau nettement plus pratique car désormais facile à appliquer à une large diversité de cas d'étude. Enfin, en termes de perspective, on envisage les possibilités de porter l'outil sur une plateforme numérique (d'abord en interne) et de publier ce package (sur le CRAN ou un autre dépôt accessible à tous).

Références

Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik K. (2023). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.6. DOI : https://doi.org/10.32614/CRAN.package.cluster .

Laurencelle M., Ohmer M., Lincker M., Vaute L., Schomburgk S., Giuglaris E. (2026). GRETA - Action 3.4 - Classification et découpage de l'aquifère rhénan dans la zone d'étude par secteurs homogènes ou ayant un comportement hydrogéologique identique. Rapport de synthèse de l'action. URL (site web GRETA) : https://www.upper-rhine-greta.eu/fr/6-actions-du-projet-greta/action-3-caracterisation-modelisation-de-levolution-des-nappes-du-niveau .