Maturation de codes scientifiques R de traitement de données liées à l'Eau au BRGM (initiative MATUREAU)

Marc LAURENCELLE* Théophile LOHIER†

Résumé

Plusieurs outils et fonctions de traitement de données scientifiques de la thématique Eau ont été développés au cours des dernières années au Bureau de recherches géologiques et minières (BRGM). La démarche MATUREAU initiée au BRGM en 2024 vise à finaliser une sélection d'outils et packages internes existants, en travaillant principalement sur : i) la modularisation des codes (le plus possible en fonctions) ; ii) la création de packages ; iii) la documentation, incluant la proposition d'exemples, de vignettes et de workflows exécutables ; iv) la simplification du mode opératoire. Le tout afin de rendre ces outils plus accessibles, d'encourager la poursuite de leur développement et maintenance d'une manière mieux structurée, et de rendre possible leur éventuelle diffusion à l'externe (en open source lorsqu'approprié). Cette présentation vise à décrire la démarche en cours, en l'illustrant par quelques outils en cours de maturation. Un outil de plus en plus utilisé au BRGM pour explorer la variabilité spatiale et temporelle de la dynamique des nappes d'eau souterraine par le clustering des chroniques piézométriques (séries temporelles de niveaux d'eau) en fonction de la similarité de leur signal, sera présenté plus en détails. Des éléments liés aux algorithmes et packages utilisés dans les outils développés seront également mentionnés afin d'alimenter la discussion et les échanges. Nous souhaitons ainsi ressortir de ces Rencontres R avec plein d'idées et pistes d'améliorations de la part des participants (aspects calculatoires, gestion des données, rendu graphique, création de packages, etc.) tout en intégrant un réseau de scientifiques particulièrement passionnés par le langage R.

Mots-clefs : Maturation – Traitement de données – Outil – Package – Eau

Développement

De nombreux outils et fonctions de traitement de données scientifiques de la thématique Eau ont été développés au cours des dernières années au BRGM. Le Bureau de recherches géologiques et minières (BRGM) est l'établissement public de référence dans les applications des sciences de la Terre pour gérer les ressources et les risques du sol et du sous-sol dans une perspective de développement durable. Ces codes, écrits principalement en langage R, constituent tantôt une collection de fonctions (ex. dédiées à l'analyse ou au prétraitement de séries temporelles ou plus spécifiquement de chroniques de niveaux d'eau souterraine ou de débits de cours d'eau) ; ou tantôt un outil pouvant être exécuté en chargeant un script principal dans l'environnement R, en modifiant manuellement les configurations d'options inscrites dans le script. **Le problème actuel** : peu de ces outils ont atteint un degré de maturité suffisant pour que l'outil soit facilement maniable par un utilisateur autre que le développeur lui-même, ou pour une diffusion de l'outil à l'externe.

L'objectif de la démarche MATUREAU, initiée au BRGM en 2024, est double : il s'agit d'une part de simplifier la mise en œuvre de chaînes de traitements complexes et éprouvées, sur de nouvelles

* BRGM, DEPA/EVE, Orléans, m.laurencelle@brgm.fr

† BRGM, DNG/TIA, Orléans, t.lohier@brgm.fr

sources de données ; et d'autre part de faciliter l'évolution des fonctionnalités existantes, le développement de nouvelles fonctionnalités, et leur intégration dans ces chaînes de traitements. La mise en œuvre est simplifiée à travers le packaging des principales fonctionnalités, l'implémentation de chaînes de traitements de référence, la standardisation des entrées et sorties, et la documentation du tout notamment à l'aide de vignettes. Un travail de structuration, d'harmonisation et de documentation des codes décrivant les principales fonctionnalités facilite l'implémentation de ces chaînes de traitements de référence. Une attention particulière est portée à la modularité afin de permettre aux développeurs actuels et futurs d'intégrer des méthodes alternatives dans les workflows. Enfin, des tests fonctionnels sont mis en place en s'appuyant sur ces workflows de référence afin de faciliter la maintenance et la mise à niveau des packages. **Plusieurs outils ou traitements** ont été identifiés pour une maturation en 2024. En voici les principaux :

(1) Tout d'abord, un outil d'analyse de « chroniques piézométriques » (séries temporelles de niveaux d'eau souterraine) incluant un module de **clustering de ces séries temporelles**. Cet outil est régulièrement sollicité par les collègues en interne, et intéresserait sans doute aussi plusieurs hydrogéologues en dehors du BRGM. Cet outil permet d'abord d'extraire et prétraiter les données de niveaux d'eau de N piézomètres issues d'ADES (la base de données nationale française sur les eaux souterraines, gérée par le BRGM) et/ou d'autres sources de données (ex. fichiers locaux) au besoin. Ensuite, le regroupement des chroniques de niveaux d'eau (individus) se fait avec la fonction `pam` de la librairie `cluster` qui met en œuvre l'algorithme des *k*-médoïdes, préféré par rapport aux *k*-means car jugé plus robuste face aux bruits et aux valeurs aberrantes. La dissimilarité entre individus est évaluée via deux mesures de distances : i) une distance « statistique » basée sur la corrélation entre séries A et B : $d = 1 - r(A,B)$; ii) une distance « géographique » (optionnelle) entre les points A et B. L'outil permet de tester différentes combinaisons de paramètres, dont une prise en compte ou non de la distance géographique (en plus de la distance statistique toujours considérée). L'utilisateur peut choisir le nombre de clusters, le type de corrélation calculée (linéaire classique de Pearson ; de rang, non paramétrique, de Kendall), etc. L'outil exporte un tableau de résultats (avec le numéro de cluster attribué à chaque individu et plusieurs indicateurs renseignant sur la qualité du clustering) ainsi que des résultats visuels détaillés (par cluster) et globaux (carte, matrice de graphiques). Les **FIGURES 1 ET 2** apportent un exemple de résultats visuels pouvant être générés par l'outil. En termes de maturation, les aspects à améliorer en priorité dans la structure de cet outil sont sa capacité à s'adapter à différents types de sources de données (API en ligne, base de données locale interne, fichiers spécifiques) sans codage complexe, et un accès simplifié à la configuration des options.

(2) Un autre outil permet de traiter en lot N séries piézométriques afin de calculer divers indicateurs de « Bilans Annuels » sur la nappe (ex. sa recharge apparente) tout en servant aussi à la détection du début d'une Recharge notable de la nappe par analyse de sa Piézométrie (d'où son nom « BARP »). Cet outil inclut un algorithme maison très efficace pour **repérer les Hautes Eaux et Basses Eaux** par année hydrologique adaptative dans des chroniques piézométriques mêmes complexes : des informations clés pour dériver quantité d'indicateurs sur l'évolution historique de la ressource en eau souterraine et révéler les situations les plus anormales, extrêmes (**FIGURE 3**). Un des principaux enjeux pour cet outil est de le rendre plus modulaire afin de permettre aux utilisateurs de calculer les indicateurs à la demande et de faciliter l'intégration de nouveaux indicateurs dans le workflow.

(3) Les fonctions de calcul de l'**Indicateur Piézométriques Standardisé (IPS)** utilisé par le BRGM depuis plus de 10 ans pour décrire l'état quantitatif des nappes d'eau souterraine en France dans les bulletins de situation hydrologique (BSH) publiés chaque mois (**FIGURE 4**) font elles aussi l'objet d'un

important travail de maturation et mise en package afin de centraliser et ainsi faciliter le déploiement et la maintenance de ces codes de calcul et d'assurer la cohérence entre services. **(4)** Des outils de traitement de données liées non pas à la quantité mais à la **qualité de l'eau** sont également considérés : génération de diagrammes de Piper, caractérisation des tendances temporelles par régression multi-segments, … **(5)** Enfin, d'**autres fonctions** de portée plus générale seront packagées dans le cadre de la démarche MATUREAU.

Ces outils reposent sur l'utilisation de **nombreux packages R** publics et pour la plupart bien connus : zoo et lubridate (pour les séries temporelles) ; cluster, fpc, amap, usedist, psych et igraph (distances et clustering) ; segmented (régression multi-segments) ; circular (stat. directionnelles) ; RPostgreSQL ; data.table ; httr (requêtes web) ; etc. D'autres packages seraient peut-être plus intéressants parfois ?

Cette participation aux Rencontres R sera l'occasion d'exposer la démarche MATUREAU en cours, mais surtout de montrer les principaux outils et packages en cours de maturation, pour susciter des échanges inspirants et gagnants-gagnants sur les aspects techniques liés à la démarche et aux outils.

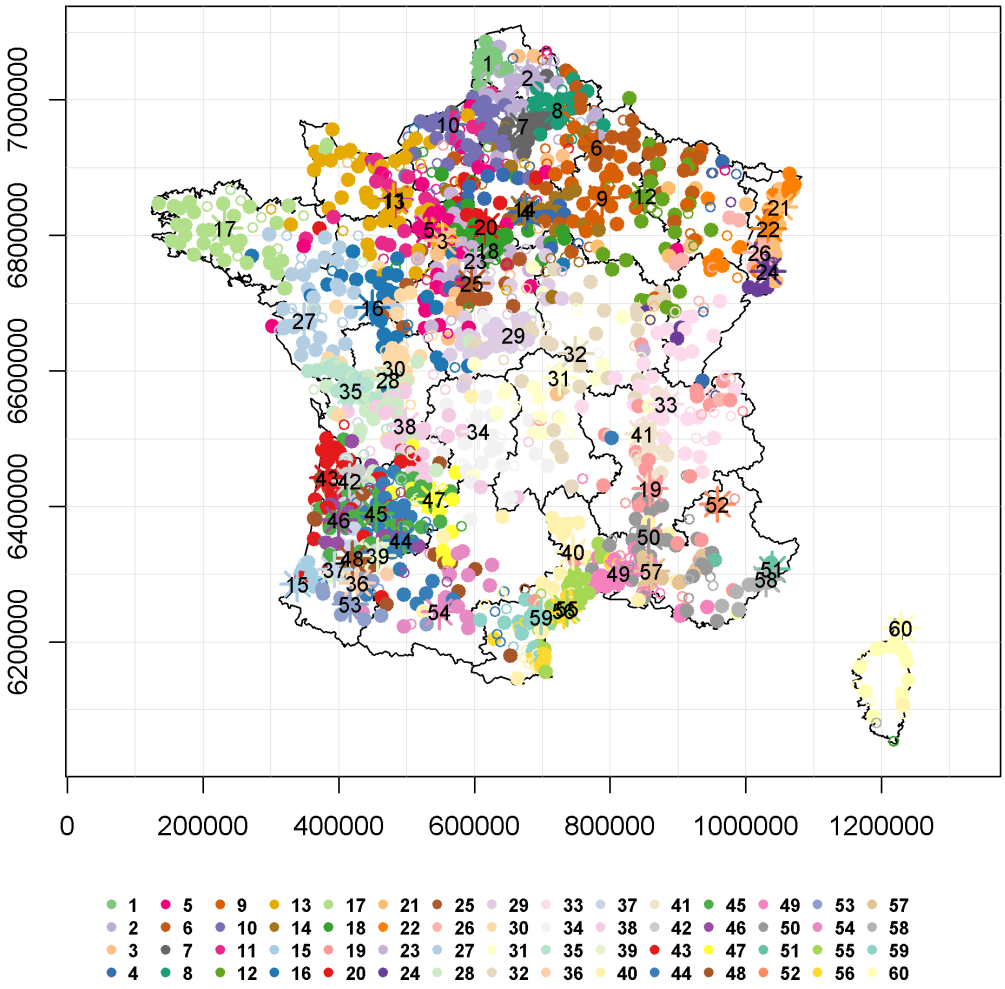

Figure 1 : Exemple d'un clustering de plus de 1 000 chroniques piézométriques à l'échelle de la France métropolitaine : a) vue cartographique globale des clusters de piézomètres formés

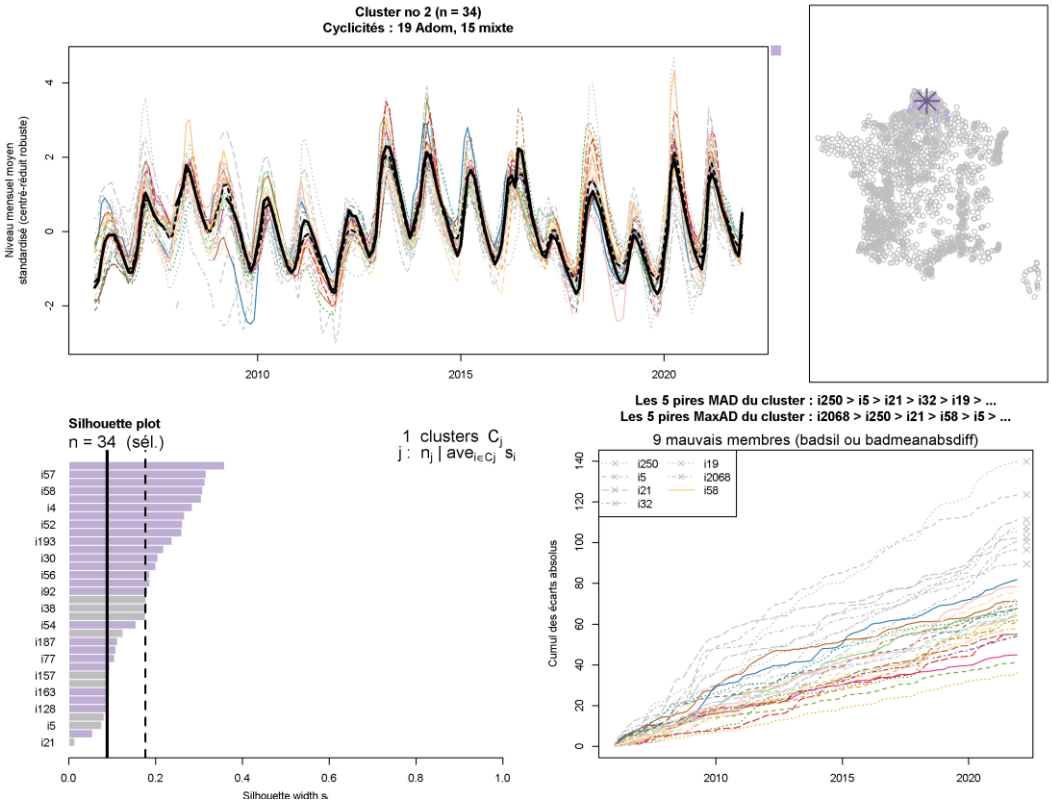

Figure 2 : Exemple d'un clustering… (suite) : b) résultats détaillés du cluster n°2 concentré dans le nord du pays

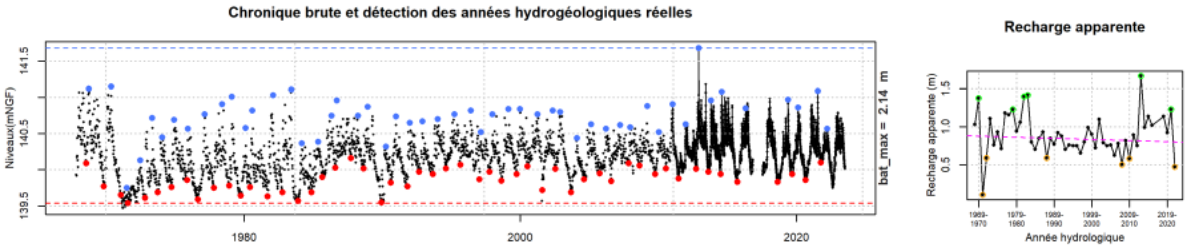

Figure 3 : Exemple de résultats produits par l'outil « BARP » : hautes eaux (en bleu) et basses eaux (en rouge) détectées par année flexible et indicateur annuel de recharge apparente calculé à partir de celles-ci

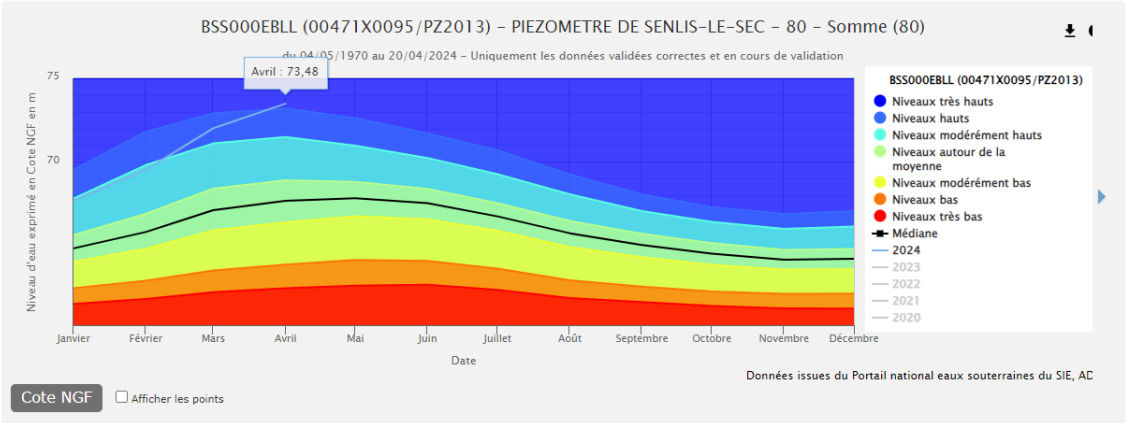

Figure 4 : Exemple de présentation visuelle de l'Indicateur Piézométrique Standardisé (IPS) sur le site web ADES : niveaux seuils mensuels calculés par la méthode (pour un piézomètre donné) afin de délimiter les classes d'IPS et courbe d'évolution récente du niveau piézométrique moyen mensuel au cours des mois de l'année 2024