Classification et découpage de l'aquifère rhénan dans la zone d'étude par secteurs homogènes ou ayant un comportement hydrogéologique identique

M Laurencelle[1], M. Ohmer.[2], M. Lincker[3], L. Vaute[4], S. Schomburgk[1], Giuglaris E.[3]

[1]*BRGM, 3 av. C. Guillemin, 45060 Orleans, France*

[2] *KIT, Adenauerring 20b 76131 Karlsruhe, Allemagne*

[3]BRGM, *rue du Pont-du-Péage Bâtiment H1, 67118 Geispolsheim, France*
[3]BRGM, 1 rue Jean Zay, 54500 Vandœuvre-lès-Nancy, *France*

Sous-action « Classification et découpage de l'aquifère rhénan dans la zone d'étude par secteurs homogènes ou ayant un comportement hydrogéologique identique. » (A3.4), réalisé dans le cadre de l'action « Caractérisation de l'évolution des nappes, en fonction de divers facteurs (climatiques / anthropiques) » du projet Interreg GRoundwater EvoluTions and resilience of Associated biodiversity - Upper Rhine (GRETA).

1 Objectif, contexte et méthodes

Le projet Interreg GRETA (Groundwater Evolutions and Resilience of Associated Biodiversity – Upper Rhine) est un projet franco-allemand qui vise à évaluer les effets du changement climatique sur la nappe de l'aquifère du Rhin Supérieur. Face à des pressions climatiques et anthropiques accrues sur l'environnement, ce projet vise à apporter des connaissances clés sur l'impact du changement climatique sur la ressource en eau et les conséquences sur les écosystèmes associés. Pour ce faire, et grâce aux nombreuses données de suivi piézométrique dont bénéficie l'aquifère rhénan, un diagnostic complet du fonctionnement historique de l'aquifère sera réalisé, constituant ainsi un socle de connaissances sur lequel les travaux consécutifs d'évaluation des impacts du changement climatique pourront s'appuyer. Dans ce cadre, une sectorisation de l'aquifère a été réalisée afin de permettre l'identification de zones présentant des comportement et/ou des caractéristiques de fonctionnement similaires. Les travaux de sectorisation conduits sont décrits dans la présente note.

L'aquifère alluvial du Rhin supérieur est situé de part et d'autre de la frontière franco-allemande, accompagnant le fleuve dont il tire son nom. La partie sud du fossé du Rhin supérieur, de Bâle (Suisse) à Karlsruhe (Allemagne) constitue la zone étudiée. Elle est délimitée à l'ouest par les Vosges (France) et à l'est par les montagnes de la Forêt-Noire (Allemagne). Cette zone correspond également au périmètre du modèle numérique LOGAR [14]. Ce système aquifère peut atteindre plus de 200 m d'épaisseur en son centre, celle-ci diminuant vers les limites de la plaine alluviale. L'écoulement des eaux souterraines est orienté du Sud vers le Nord, parallèlement au Rhin dans la partie centrale, et sur les bordures des contreforts en direction SW-NE (France) et SE-NW (Allemagne).

Cet aquifère constitue l'une des plus importantes ressources en eau d'Europe et le volume total d'eau contenu dans la zone d'étude de l'aquifère est estimé entre 65 et 80 milliards de m3.

Sa recharge est assurée par les précipitations locales, les infiltrations du Rhin ainsi que par les rivières prenant leurs sources dans les Vosges et la Forêt Noire, en proportion variable selon les secteurs.

L'aquifère rhénan concentre des enjeux de différente nature puisque sa ressource est exploitée pour de multiples usages anthropiques par des prélèvements en forages : usages industriels, alimentation en eau potable ou encore usages agricoles. Par ailleurs, la nappe de l'aquifère rhénan joue un rôle prépondérant d'alimentation en eau de zones humides et est fortement liée au réseau hydrographique de surface. Les nombreux cours d'eau phréatiques et semi-phréatiques témoignent des apports hydriques depuis la nappe, qui jouent un rôle prépondérant dans leur alimentation.

Le projet franco-allemand Interreg GRETA, dans lequel s'inscrivent les travaux présentés ci-dessous, vise à apporter de nouvelles connaissances sur l'impact du changement climatique sur l'aquifère rhénan ainsi que sur l'effet de ces évolutions piézométriques sur les écosystèmes liés.

Des travaux de regroupement (ou clustering) des piézomètres ont été opérés afin d'intégrer dans l'analyse la variabilité spatiale des comportements de l'aquifère, liée à la fois à la variabilité spatiale des conditions d'alimentations, aux paramètres hydrodynamiques du système aquifère et aux usages locaux de la ressource. Ces travaux de regroupement ont pour objectif de permettre l'identification de groupes de piézomètres témoignant de comportements homogènes au sein de l'aquifère rhénan. Pour ce faire, trois méthodes ont été mises en œuvre dans le cadre de GRETA : (i) regroupement basé sur la corrélation entre les chroniques piézométriques, (ii) regroupement basé sur des caractéristiques numériques décrivant la dynamique des chroniques piézométriques, (iii) regroupement basé sur des caractéristiques hydrodynamiques décrivant la physique du comportement de la nappe (aussi appelés "indicateurs" dans la suite de ce rapport), calculés à partir des chroniques piézométriques.

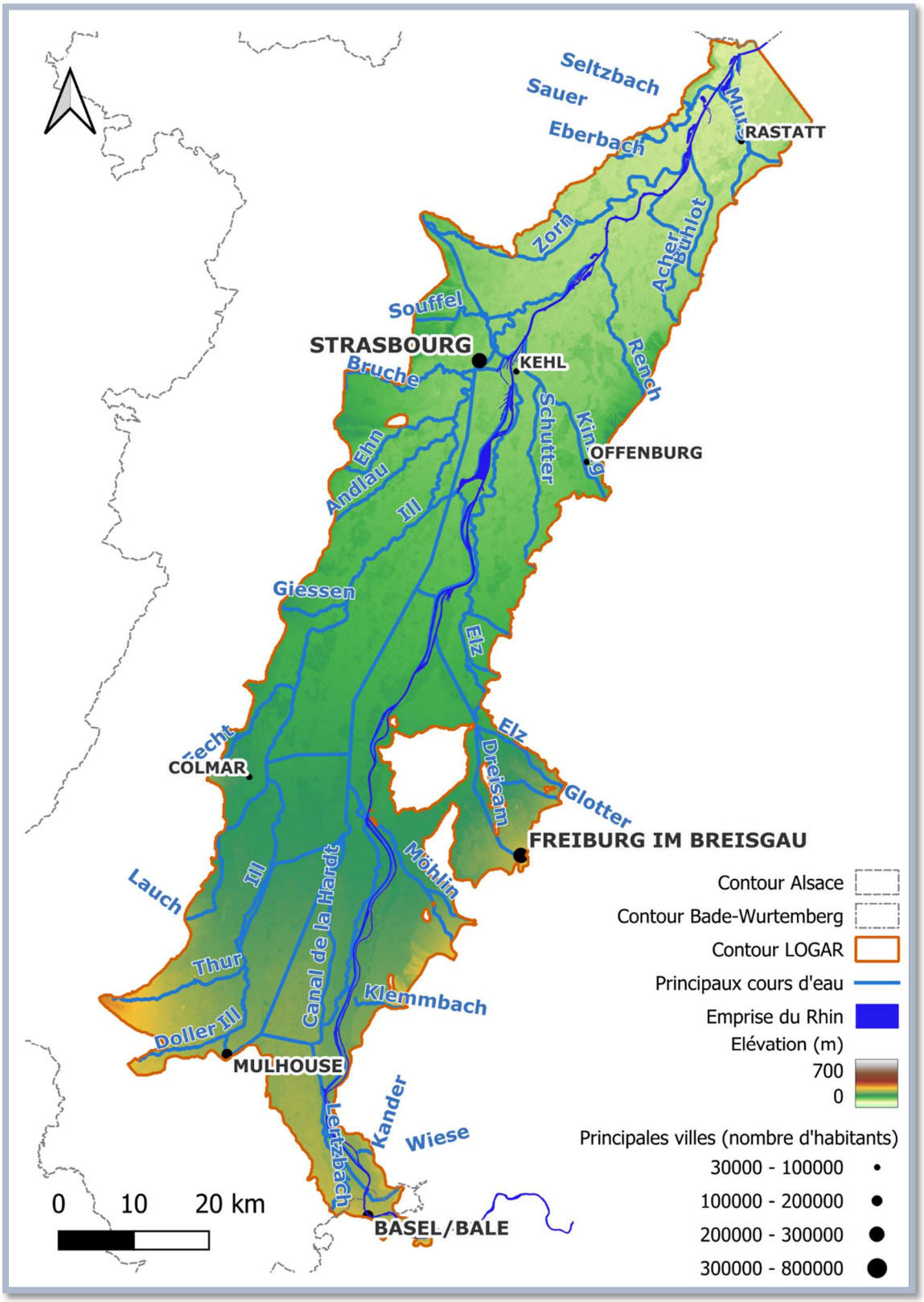


Figure 1 Carte de la zone d'étude : délimitation de l'aquifère dans la zone étudiée et contexte hydrologique

2 DONNEES DISPONIBLES

Les chroniques de piézomètres surveillant l'aquifère rhénan ont été fournies par l'APRONA (Association pour la Protection de la Nappe d'Alsace) pour les ouvrages français et par le LUBW (Landestalt für Umwelt, Messungen and Naturschutz Baden-Württemberg : office de l'environnement du Land allemand du Bade-Wurtemberg) pour les chroniques du réseau de suivi piézométrique allemand. Quelques autres chroniques supplémentaires ont été extraites de la base ADES pour le côté français, par la suite (voir plus bas).

Le jeu de données piézométriques initial est ainsi composé de 1888 chroniques piézométriques dont 1375 du côté allemand et 513 du côté français. 9 chroniques sur 10 suivent la nappe au moins jusqu'à l'année 2023, et environ 2/3 des chroniques débutent en 1980 ou avant. C'est sur cette base qu'a été établie une « période ciblée » (pour les travaux de regroupements de chroniques piézométriques) s'étendant des années 1980 à 2023 inclusivement.

Pour le regroupement basé sur la corrélation entre les chroniques piézométriques, 1244 chroniques ont été utilisées comme jeu de données principal de départ (pour la première session de clustering de ce type) parmi l'ensemble des 1888 chroniques disponibles. Ce nombre réduit de piézomètres a été obtenu en appliquant un critère de disponibilité suffisante en données sur la période ciblée (janvier 1980 à janvier 2024) : seules les chroniques (préalablement agrégées en niveaux piézométriques moyens mensuels) ayant une valeur mensuelle calculable disponible (c.-à-d. au moins une observation piézométrique dans le mois) pour au moins 2/3 des mois de la période ciblée, ont été retenues.

Dans une seconde phase, une fois un premier regroupement réalisé à partir de cette sélection principale de chroniques « longues », une sélection complémentaire a été définie en assouplissant le critère pour n'exiger des valeurs mensuelles que pour au moins 1/5 des mois de la même période ciblée. Cette seconde sélection, intégrant des chroniques « moins longues », a ainsi permis de considérer +364 autres chroniques (toujours parmi l'ensemble des 1888 chroniques disponibles de départ), portant à 1608 le nombre de chroniques finalement considérées dans les traitements de regroupements basés sur la corrélation. Et dans la dernière phase lors de la synthèse des trois approches (i-iii), 14 chroniques supplémentaires ont été intégrées à la liste finale, pour considérer quelques chroniques valorisées par les autres approches de regroupement ou dans le travail parallèle de calcul d'indicateurs hydrogéologiques (Action Greta 3.3., analyse de l'évolution historique du niveau de la nappe : calcul des tendances et des ruptures dans les séries piézométriques).

Le nombre de chroniques finalement retenues dans le regroupement final (cf. 5 SYNTHESE DES RESULTATS : Sectorisation de l'aquifère rhénan en grands ensembles) est ainsi de 1622 chroniques, réparties comme suit : 243 chroniques issues du jeu de données fourni par l'APRONA + 27 chroniques complémentaires extraites d'ADES[1], soit un sous-total de 270 chroniques pour le côté français ; et 1352 chroniques fournies par le LUBW pour le côté allemand.

Pour le regroupement basé sur les caractéristiques numériques décrivant la dynamique des chroniques, 1 633 piézomètres de la vallée du Rhin supérieur (Allemagne et France) ont été compilés à partir de 1913 (les 1622 chroniques finalement retenues plus 11 chroniques

[1] S'il y a peu de chroniques piézométriques qui proviennent directement d'ADES pour ce travail, c'est parce qu'il a été choisi de prioriser, pour un piézomètre français donné, les données provenant d'APRONA afin de garantir leur complétude.

uniquement utilisées pour cette méthode). Les stations de mesure dont la durée d'enregistrement était inférieure à 10 ans ont été exclues.

Pour le regroupement basé sur les indicateurs hydrodynamiques calculés aux piézomètres, le nombre de chroniques utilisées correspond au nombre de chroniques présentant des données suffisantes au calcul des indicateurs. Ainsi, les chroniques de 971 piézomètres ont été utilisées dans ce test de regroupement.

La carte ci-dessous (Figure 2) présente la répartition des piézomètres correspondant aux 1622 chroniques finalement considérées (tout ou en partie) dans les travaux de regroupement réalisés par les trois méthodes.

Cette carte met en évidence les densités de points contrastées entre les deux pays, considérablement plus dense en Allemagne qu'en France d'une part, et selon la proximité du Rhin d'autre part, avec des densités de points moyennes de l'ordre de 0,1 point par km^2 du côté français versus 0,7 point par km^2 du côté allemand ; soit un réseau 8 fois plus dense pour la rive droite (est) du Rhin[2]. Mais si on se concentre sur les terrains à ≤ 5km du linéaire hydrographique du Rhin, la densité de points sur la rive est (côté allemand) s'avère presque deux fois plus forte relativement à la densité moyenne sur la zone allemande, (1,3 point par km^2) tandis qu'elle demeure d'environ 0,1 point par km^2 sur la rive ouest (côté français).

Cet aspect sur la densité des points a son importance dans l'interprétation des résultats de regroupement. La répartition des piézomètres considérés par chaque méthode est présentée plus bas, dans les sections de ce rapport qui y sont dédiées.

[2] Densités estimées à partir des 1622 points de localisation des chroniques piézométriques retenues pour les traitements de regroupements basés sur la corrélation, en considérant le polygone de délimitation du modèle LOGAR, qui contient la vaste majorité de ces points (afin d'éviter de sous-estimer la densité en incluant des surfaces éloignées des points d'intérêt. Seulement 34 points (hors LOGAR) ont été ignorés.

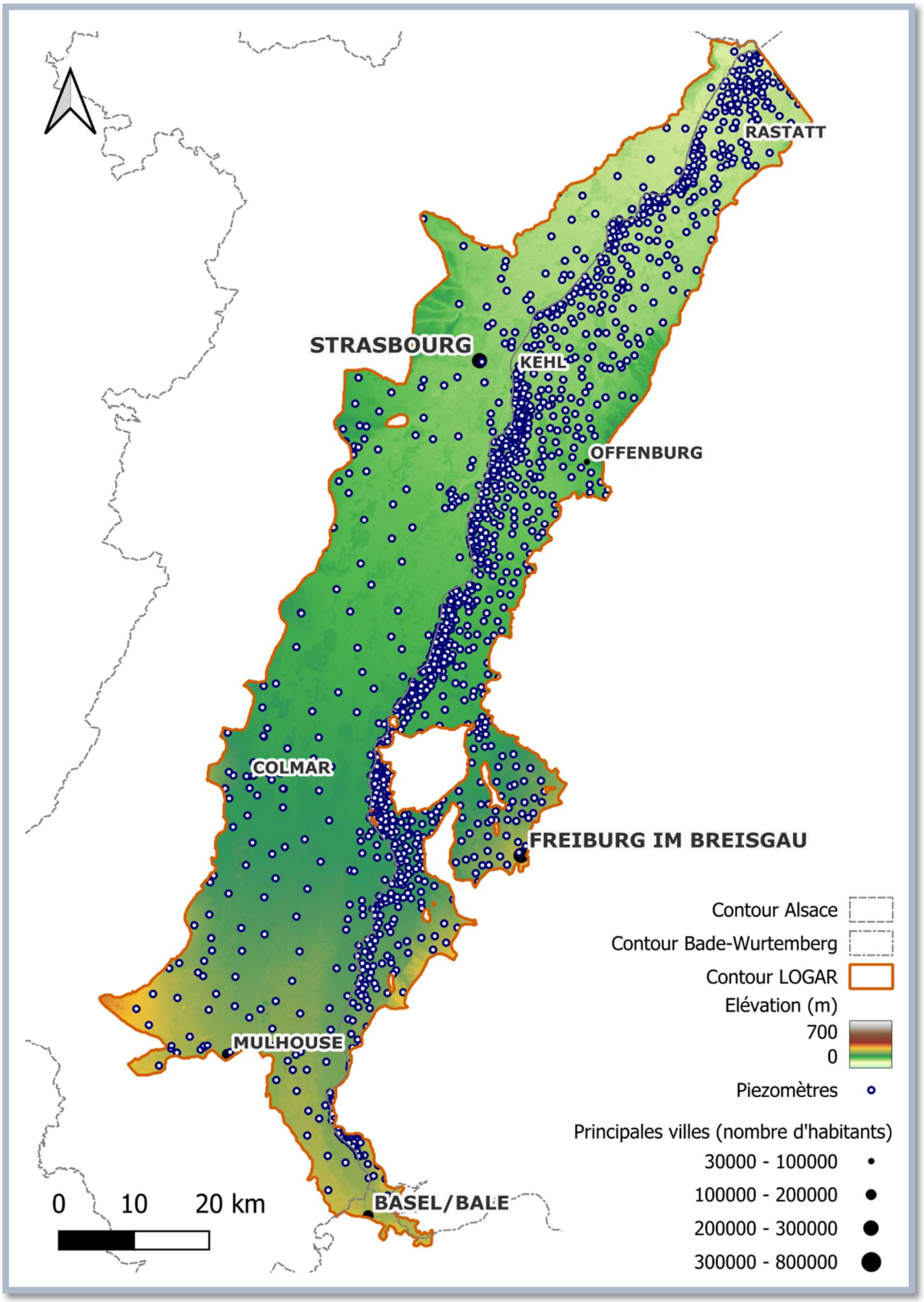


Figure 2 Localisation des points des 1622 chroniques piézométriques considérées

3 METHODES

3.1 Regroupement basé sur la corrélation entre les chroniques piézométriques

La première approche utilisée pour regrouper les piézomètres se concentre sur la comparaison des séries de niveaux piézométriques, sans considérer d'indicateurs calculés à partir de celles-ci ni autres informations contextuelles, du moins en ce qui concerne les étapes de traitements statistiques. Ce n'est que lors des itérations d'expertise manuelle des regroupements calculés automatiquement, que diverses informations pertinentes sont considérées en complément afin d'ajuster les résultats.

Une approche de clustering par algorithme des k-médoïdes a été choisie pour cela, en raison de ses avantages, notamment :

- Le clustering par k-médoïdes apparaît préférable au regroupement hiérarchique, par son principe fondamental qui cherche à identifier la série « médoïde » la plus représentative de chacun des groupes, plutôt que de créer un arbre hiérarchique complexe délicat à diviser en groupes ;
- L'algorithme des k-médoïdes est préféré à celui des k-means en raison de sa moindre sensibilité face aux individus extrêmes et aussi parce qu'il utilise des individus réels, au lieu de points fictifs de coordonnées moyennes, comme représentants des groupes ;
- L'algorithme des k-médoïdes produit des résultats stables d'une exécution à la suivante, contrairement à celui des k-means qui, à cause de la nature aléatoire des configurations initiales testées durant sa phase d'optimisation, propose des regroupements qui peuvent changer notamment en termes de numérotation des groupes. Or, la reproductibilité des résultats est une caractéristique avantageuse en pratique.
- La fonction `pam()` — de l'anglais *Partitioning Around Medoids (PAM)* — disponible dans le langage R pour appliquer cet algorithme accepte en entrée une matrice de dissimilarité (ou autrement dit de « distances ») ce qui permet de baser le regroupement sur une mesure de dissimilarité au choix, non limitée à la distance euclidienne classique comme c'est le cas dans l'algorithme k-means par exemple.

Ce dernier point est justement un avantage clé de l'algorithme des k-médoïdes : il a permis de d'utiliser une mesure de « distance » entre chroniques basée sur leur corrélation. Or, cette mesure de distance est intéressante car elle se base sur une **notion intuitive** de mesure de **ressemblance des fluctuations** entre deux séries temporelles x(t) et y(t) telle qu'on aurait tendance à l'évaluer manuellement par la préparation d'un graphique XY, l'ajustement d'une droite de régression linéaire et enfin l'extraction des résultats de l'ajustement (en particulier le coefficient r).

De plus, les **distances de type corrélation** ne dépendent ni de l'écart de position moyenne entre x et y ni de l'amplitude globale de x ou y. En effet, un écart de position moyenne impactera l'ordonnée à l'origine de la droite de régression linéaire tandis qu'un contraste d'amplitude impactera sa pente, mais ni l'un ni l'autre n'impactera le coefficient de corrélation r en tant que tel. Le calcul ne requiert donc pas obligatoirement de normalisation préalable des données.

La distance de type corrélation la plus classique (et attendue intuitivement) est basée sur le bien connu « coefficient de corrélation de Pearson » (dont le symbole est : *r* minuscule) et la formule pour calculer cette distance entre les séries temporelles x et y se résume tout simplement par :

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - r$$

La **matrice de distances** requise par la fonction `pam()` est préparée comme suit :

1. Préparation de séries temporelles à pas de temps mensuel, en calculant le niveau piézométrique moyen pour chaque mois avec ≥ 1 donnée ;
2. Réduction des séries aux temps (dates) inclus dans la période ciblée (de janvier 1980 à janvier 2024) afin de concentrer la suite des traitements sur cette période récente la plus fournie en données ;
3. Standardisation[3] des séries mensuelles (facultative à ce stade : faite afin de faciliter la superposition visuelle de plusieurs courbes dans les graphiques générés plus tard) ;
4. Création d'une matrice de m lignes (pas de temps) * n colonnes (séries) par fusion de toutes les séries mensuelles alignées temporellement → « matrice des séries » ;
5. Calcul d'une « matrice de distances » à partir de la matrice des séries, en utilisant des fonctions matricielles permettant un calcul rapide malgré le grand nombre de paires d'individus (et donc de distances) à évaluer (>1 000 000). Cette matrice de distances peut être obtenue en calculant d'abord la « matrice de corrélations » R, puis en appliquant la formule présentée plus haut mais ici en contexte matriciel : D = 1 – R. La matrice D ainsi obtenue a n lignes * n colonnes (où n = le nombre de séries) et rassemble les distances calculées entre toutes les paires de séries. Elle est symétrique : D[i, j] = D[j, i] ; où i et j sont des nombres entiers positifs ≤ n.

Une fois la matrice de distances D préparée, la fonction de clustering `pam()` est appelée en utilisant cette matrice comme principal argument d'entrée, et un nombre de clusters *k* spécifié manuellement. La fonction renvoie comme résultats essentiels la liste des *k* individus retenus comme médoïdes optimaux par l'algorithme, ainsi qu'une liste avec l'identifiant du cluster attribué à chacune des *n* séries piézométriques.

Quelques remarques sur cet algorithme « non supervisé » de clustering par k-médoïdes :

- L'algorithme par k-médoïdes, puisqu'il se base sur la comparaison des données entre paires d'individus, ici des séries temporelles, requiert des données disponibles à des mois communs. Il n'est donc pas possible, avec l'algorithme par k-médoïdes en tant que tel, de regrouper des séries sans période de suivi concomitante. C'est pour cette raison qu'un critère de disponibilité minimale des données ≥ 2/3 des mois de la période ciblée a été utilisé pour constituer le jeu de données principal pour lancer le premier appel de la fonction de clustering pam(). En effet, en exigeant une couverture temporelle ≥ p % des mois de la période ciblée, on s'assure que toutes les paires de séries aient ≥ (2p – 100) % de mois en commun (soit a minima 1/3 de mois en commun avec le critère utilisé ici).

L'ajout ultérieur de séries plus courtes dans les clusters préalablement formés est possible, mais il ne se fait plus par un simple appel de la fonction `pam()`. Plutôt, chaque série courte est comparée aux médoïdes déjà définis et rattachée au cluster pour lequel la corrélation au

[3] Une transformation en séries centrées réduites robuste basée sur la médiane et l'écart absolu médian (plutôt que la moyenne et l'écart type) est utilisée car moins sensible aux éventuelles valeurs extrêmes / aberrantes qui n'auraient pas été détectées et retirées.

médoïde est maximale, à condition que cette corrélation soit assez forte. Si les corrélations entre la série et les médoïdes sont toutes trop faibles, la série est soit définitivement écartée, soit placée dans des groupes créés expressément pour rassembler les cas singuliers voire anormaux.

Un module pour le post-traitement des résultats de clusterings et toujours basé sur la corrélation entre séries, a été développé durant le projet. Ce module a permis de retravailler les clusters… :

- tantôt afin de trouver une meilleure série médoïde (plus longue et continue tout en demeurant bien centrale dans le cluster) ;
- tantôt afin de définir un nouveau cluster à partir d'une série choisie à dire d'expert en raison de sa signature assez distincte par rapport à celle des médoïdes existants, à la fois en termes de dynamique et de répartition spatiale ;
- tantôt afin d'écarter des séries peu corrélées avec les autres (montrant une évolution piézométrique singulière, voire anormale) ;
- mais avant tout pour trouver à quel médoïde existant et donc à quel cluster chacune des séries ajoutées ressemble le plus.

Pour aider à attribuer un cluster à une série sans appeler le clustering par k-médoïdes, le module de post-traitement prépare une **matrice des corrélations** (*r*) entre les séries (lignes) et les médoïdes existants (colonnes). Et ici, contrairement à l'algorithme PAM, ce post-traitement n'exige pas que le tableau soit rempli : des vides (*r* = NA : non calculable) sont tolérés. La matrice est ensuite parcourue afin d'identifier, pour chaque série, quel est le cluster offrant la meilleure corrélation, c.-à-d. le *r* maximal parmi les *r* calculables (non NA) et ≥ (+)0.6. Ce cluster placé en tête de la liste décroissante des coefficients de corrélation *r* calculés pour la série, est appelé « **premier voisin** » dans ce module. C'est donc en récupérant cette information sortante, qu'on peut facilement attribuer une première proposition (provisoire) de cluster aux séries plus courtes sans devoir utiliser clustering par k-médoïdes.

Une remarque sur la **distinction entre le cluster attribué par l'algorithme PAM versus celui proposé via le « premier voisin » fourni par le module de post-traitement**. L'algorithme PAM optimise le partitionnement de manière à maximiser la différence entre clusters tout en minimisant la variabilité au sein de chaque cluster. Il y a donc une notion de compromis dans cet algorithme, ce qui fait que certaines séries peuvent être placées dans un cluster alors qu'elles ont une corrélation plus élevée avec le médoïde d'un ou plusieurs autres clusters. Si on souhaite conserver au moins partiellement les résultats de ces compromis faits par l'algorithme PAM lors des phases ultérieures de post-traitement de ses résultats, il faut donc éviter d'utiliser systématiquement les indications fournies via le « premier voisin » (par simples corrélations calculées entre la série et les médoïdes) pour remplacer toutes les attributions de clusters. Il est en effet préférable, en général, de retenir en priorité le cluster attribué par l'algorithme PAM.

Dans la mise en œuvre pour GRETA, à l'issue de la première itération de post-traitement des résultats du PAM, le cluster proposé par l'information du « premier voisin » n'a donc été attribué qu'aux 378 séries plus courtes nouvellement ajoutées à la liste des individus à regrouper (soit les cas sans attribution de cluster par l'algorithme PAM). Les autres chroniques (1244) conservent leur cluster préalablement assigné.

En pratique, l'approche mise en œuvre pour produire les résultats basés sur la corrélation entre les chroniques est séquentielle et récursive, **avec comme objectif de former des clusters explicables en termes hydrogéologiques.** On choisit volontairement un faible nombre de clusters (*k*) au départ pour établir les principales différences dans la dynamique des nappes. Les

séries suffisamment longues du jeu de données sont considérées. Une première itération de clustering par PAM est appliquée. On regroupe les clusters ayant des explications communes ou similaires. On tente ensuite de nouvelles itérations de clusterings à l'intérieur des groupes formés jusqu'à ce qu'il ne soit plus possible de trouver des explications pour justifier les sous-groupes suggérés par la nouvelle itération. Une fois les principaux clusters (et leurs médoïdes) définis par ces itérations, les résultats de ces dernières sont fusionnés et ensuite retravaillés avec le module de post-traitement, là aussi en plusieurs itérations. La mise en œuvre plus précise de cette approche est décrite plus bas dans la présentation des résultats.

3.2 Regroupement basé sur les caractéristiques dynamiques des chroniques piézométriques

Tableau 1 Aperçu des caractéristiques (dynamic features) utilisées pour caractériser les chroniques piézométriques. Les caractéristiques couvrent différents aspects de la variabilité temporelle, de la saisonnalité, des événements extrêmes et des particularités structurelles des piézomètres

Nom de la caractéristique (Abr.)	**Objectif/Description**
Range Ratio (RR)	Détection des signaux superposés à longue période, également sensible aux valeurs aberrantes, calculée comme le rapport entre l'amplitude moyenne annuelle et l'amplitude totale [13]
Skewness (Skew)	Limitation, inhomogénéités, valeurs aberrantes, asymétrie de la distribution de probabilité.
Annual Periodicity (P52)	Intensité du cycle annuel, calculée par corrélation (Pearson) de la périodicité annuelle moyenne (52 semaines) avec la série chronologique complète [13]
SDdiff	Variabilité brusque (« flashiness »), fréquence et rapidité des variations à court terme, calculées comme l'écart-type des premières dérivées de la série [13].
Longest Recession (LRec)	Phases de longue baisse, mesurée comme la plus longue séquence sans remontée du niveau des eaux souterraines [13]
Jumps	Inhomogénéités ou ruptures (changements structurels), reflétant partiellement aussi la variabilité, calculées comme le maximum absolu et standardisé de la différence de moyenne entre deux années consécutives [13].
Seasonal Behaviour (SB)	Position du maximum dans le cycle annuel, comparée à la saisonnalité moyenne attendue (minimum en septembre, maximum en mars) [13].
Median (Med01)	Indice de limitation, médiane après mise à l'échelle sur [0,1], mesure statistique standard dérivée de [4].
High Pulse Duration (HPD)	Durée moyenne des niveaux des eaux souterraines dépassant le 80e centile de non-dépassement, pour plus de détails, voir [11], adapté de [4]

L'objectif de l'approche de regroupement basé sur les caractéristiques décrivant la dynamique des chroniques piézométriques (Dynamic Feature Clustering) est de regrouper de manière robuste les séries piézométriques en fonction de leurs propriétés dynamiques, afin d'identifier des évolutions représentatives pour la modélisation, les prévisions et une meilleure compréhension du système, indépendamment de la longueur des données ou des éventuelles lacunes dans les séries piézométriques.

Le choix s’est porté sur le regroupement par caractéristiques dynamiques car cette méthode permet le traitement de données de différentes périodes et longueurs, avec de données manquantes ou bruitées. Contrairement aux approches de regroupement directes, cette méthode permet d'utiliser des ensembles de données incomplets. Ainsi, les stations de mesure récemment mises en place et celles qui n'ont pas été utilisées depuis longtemps ont pu être incluses dans l'analyse.

Pour l'étude, un ensemble complet de données comprenant 1 633 piézomètres avec des données hebdomadaires (à minima) dans le fossé rhénan supérieur (Allemagne et France) a été compilé à partir de 1913. Les stations de mesure dont la durée d'enregistrement était inférieure à 10 ans (74 stations) n’ont pas été retenues.

L'étape centrale de la préparation des données consiste à transformer les séries chronologiques en un ensemble de caractéristiques descriptives (appelées « features ») qui permettent de caractériser les aspects essentiels de la dynamique. Ces caractéristiques sont conceptuellement proches des signatures hydrologiques, largement utilisées en hydrologie de surface. Cependant, comme les hydrogrammes des eaux souterraines présentent des caractéristiques différentes de celles des écoulements de surface, des caractéristiques appropriées ont été rigoureusement sélectionnées, adaptées ou nouvellement développées.

À partir d'une liste initiale de plus de 50 caractéristiques potentielles, 13 caractéristiques ont finalement été retenues. La sélection s'est basée sur une combinaison de tests de plausibilité visuelle (« Visual Skill Test »), d'analyses de robustesse concernant la qualité des données (par exemple, lacunes, bruit, longueur) et d'analyses de corrélation afin d'identifier et réduire les redondances. Les caractéristiques utilisées comprennent des mesures statistiques classiques (par exemple, asymétrie, écart type), des paramètres dynamiques (par exemple, indices de variabilité rapide, durée des phases de vidange), ainsi que des caractéristiques liées à la périodicité, telles que l’intensité de la composante annuelle ou le degré de saisonnalité.

Les caractéristiques ont été normalisées dans l’intervalle [0,1] afin de permettre une mise à l'échelle uniforme pour le clustering. Une fois ces étapes terminées, chaque chronique piézométrique se présente sous la forme d'un vecteur multidimensionnel qui décrit ses caractéristiques dynamiques sous une forme compacte. Ces vecteurs de caractéristiques constituent la base de la méthode de clustering détaillée ici.

Une procédure itérative a été mise en œuvre afin de déterminer le nombre optimal de clusters. Différents algorithmes de clustering (KMeans, Birch, Agglomerative Clustering, Spectral Clustering) ont été testés pour une plage définie de nombres de clusters possibles (par exemple, de 2 à 14). Les résultats ont ensuite été évalués à l'aide de plusieurs indices de validation interne, notamment l'indice de silhouette, l'indice de Calinski-Harabasz, l'indice de Dunn et l'indice de Ratkowski-Lance. L'évaluation a été effectuée par un classement des scores, permettant une sélection objective du nombre optimal de clusters. Le regroupement final a été effectué avec l’algorithme KMeans. Outre l'appartenance de chaque chronique à un cluster, la distance euclidienne de chaque point de mesure a été calculé par rapport au centroïde du cluster qui lui a été attribué. Cette valeur calculée a ensuite été utilisée pour évaluer l'homogénéité intra-cluster et pour la pondération visuelle.

3.3 Regroupement basé sur les indicateurs hydrodynamiques calculés au piézomètre et algorithmes de science des données (data science)

Cette méthode se base sur l'utilisation :

- d'un ensemble d'indicateurs hydrodynamiques de la nappe, calculés pour chaque piézomètre lorsque la chronique piézométrique correspondante le permet,
- de variables géographiques : l'épaisseur de la zone non saturée, les données de couverture et d'utilisation des sols issue de la campagne 2018 du programme CORINE (Coordination of Information on the Environment, Corine Land Cover), le réseau hydrographique de surface (cours d'eau recensés dans le cadre de la Directive Cadre sur l'Eau (DCE),
- d'un enchaînement d'algorithmes de science des données.

Données d'entrée propre à la méthode

Les calculs et la description des indicateurs produits sont disponibles dans le rapport BRGM/RP-74683-FR [2]. Les indicateurs suivants ont été calculés sur les chroniques piézométriques le permettant :

- Les tendances significatives (test de Mann-Kendall modifié ; [3, 5, 6]) et les signes des tendances sur les variables suivantes : niveaux moyens, minimum et maximum, recharge apparente et durée de la vidange apparente, à partir du signe de la pente de Sen [11].
- Les ruptures d'homogénéité significative et dates associées (test statistique de Pettitt; [8]) sur les mêmes cinq variables mentionnées ci-dessus.
- Les dates médianes d'occurrence et les écarts interquartiles des basses eaux. Ces deux indicateurs traduisent la saisonnalité des minima piézométriques et leur variabilité.
- Les jours pour lesquels est observée une variation intra-journalière des niveaux piézométriques plus importante que la variation entre deux jours consécutifs.
- Les énergies des différentes composantes fréquentielles du signal correspondant au "poids" de différentes gammes de fréquences dans le signal piézométrique. Les gammes pour lesquels l'indicateur est calculé sont : < 1 an, 1-5 ans, 5-12 ans, 12-24 ans et > 24 ans. Ces gammes de fréquence ont en fait un « sens climatique » puisqu'elles sont induites par des processus climatiques différents [1].
- Dans le but également de synthétiser l'information sur les contributions des différentes gammes de fréquences dans le signal piézométrique, un indicateur d'énergie a été développé et construit à partir :
    - D'une moyenne pondérée des pourcentages d'énergie par gamme de fréquences pour la série piézométrique analysée (pondération par des coefficients croissants entre la gamme la plus haute fréquence (< 1 an) et la plus basse fréquence (> 24 ans)) ;
    - D'une moyenne pondérée pour un cas fictif où 100% de l'énergie serait sur la gamme la plus basse fréquence (> 24 ans) et qui servira de valeur référence afin d'obtenir un indicateur dont les valeurs oscillent entre 0 et 1 ;

Un ratio entre la moyenne pondérée des pourcentages d'énergie pour la série piézométrique analysée et la moyenne pondérée du cas fictif est ensuite réalisé. C'est ce ratio qui constitue l'indicateur et ses valeurs s'étendent entre 0 et 1. Il permet en une valeur de résumer l'information de l'analyse spectrale et donc le caractère inertiel ou réactif de la nappe.

- Les coefficients de corrélation entre les chroniques piézométriques et les pluies efficaces pour informer sur la force de la relation entre la piézométrie et la climatologie locale.
- Le temps de ½ tarissement de l'exponentielle utilisée pour calculer la pluie efficace moyenne optimale. Ce temps caractéristique est assimilable au temps de ½ tarissement de la nappe, et donc informatif sur l'inertie de la nappe captée due aux propriétés d'écoulement et de stockage de l'aquifère.
- Le décalage temporel entre le signal climatique et le signal hydrogéologique utilisée pour calculer la pluie efficace moyenne optimale. Ce décalage est un informatif sur la rapidité ou non de l'atteinte de la nappe par les pluies efficaces contribuant à l'infiltration.
- Les coefficients de corrélation et le décalage temporel permettant d'obtenir la meilleure corrélation entre série piézométrique et série de débit du Rhin (total et de base).

- A ces indicateurs ont été ajoutées 4 variables géographiques relatives au réseau hydrographique de surface : le nombre de cours d'eau et le linéaire des cours d'eau se trouvant dans des zones circulaires de 300 m et 500 m de diamètre centrées sur chaque piézomètre. Ces 4 nouvelles variables ont pour but de permettre la prise en compte du réseau hydrographique dans les tests de regroupement, de manière extrêmement schématique toutefois car la nature précise de l'influence sur un piézomètre des cours d'eau proches n'est pas connue.

Description de la séquence algorithmique utilisée

La chaîne de traitement décrite ci-dessous vise à regrouper automatiquement des points de mesure hydrogéologiques (forages, piézomètres, sources, etc.) tout en mettant en lumière les variables qui gouvernent le plus ces regroupements. Elle combine trois familles d'algorithmes : réduction de dimension, estimation d'importance des variables via un algorithme de prédiction par gradient boosting, et clustering fondé sur la densité. L'enchaînement des étapes et des différents algorithmes est représenté de manière graphique sur les Figure 3 et Figure 4 .

a. Pré-traitement des variables

La procédure commence par une standardisation systématique des variables quantitatives (centrage-réduction) et par un encodage numérique cohérent des variables qualitatives (one-hot encoding avec gestion des poids). Cette étape garantit que toutes les grandeurs sont comparables et qu'aucune unité ou plage de valeurs ne domine artificiellement les calculs ultérieurs.

b. Première réduction de dimension avec UMAP

Le jeu initial, souvent très hétérogène, est projeté dans un espace à trois composantes au moyen de l'algorithme UMAP (Uniform Manifold Approximation & Projection) (arXiv). UMAP préserve la structure globale du nuage tout en offrant une représentation compacte que l'on pourra visualiser pour contrôler l'allure des groupes naissants.

c. Hiérarchisation des variables par CatBoost

Sur ces trois composantes, on entraîne ensuite l'algorithme de prédiction CatBoost, un gradient boosting conçu pour traiter correctement les variables catégorielles (arXiv). Pour chaque composante UMAP, le modèle fournit un score d'importance pour chacune des variables d'origine : plus le score est élevé, plus la variable a contribué à la projection. Les variables dont l'importance moyenne est négligeable sont alors écartées, ce qui réduit le jeu initial à un sous-ensemble explicatif plus compact.

d. Deuxième passe UMAP + CatBoost pour affiner l'interprétation

Le jeu filtré passe à nouveau dans UMAP afin de générer trois composantes finales centrées sur les variables réellement déterminantes. Un second CatBoost est ajusté, non plus pour filtrer, mais pour expliquer ces nouvelles composantes et fournir une lecture plus claire des mécanismes de regroupement.

e. Clustering avec HDBSCAN

Les points ainsi décrits en 3-D sont regroupés par HDBSCAN, une version hiérarchique et densité-dépendante de DBSCAN (joss.theoj.org). L'algorithme détecte automatiquement les clusters de forme libre tout en laissant certains points non assignés (« bruit »). Trois hyper-paramètres pilotent son comportement :

1. min_cluster_size définit la taille minimale d'un groupe ;
2. min_samples règle la sévérité vis-à-vis des valeurs aberrantes ;
3. cluster_selection_epsilon permet de fusionner ou de séparer les petits ensembles voisins.

Une recherche systématique sur grille explore de multiples combinaisons de ces paramètres afin de trouver, pour chaque jeu de variables, la configuration la plus équilibrée.

f. Évaluation de la qualité des regroupements

Chaque solution est évaluée sous trois angles :

Volet d'évaluation	**Outils mis en œuvre**	**Objectif**
Qualité interne	- **Silhouette** [10] - **DBCV** (Density-Based Clustering Validation Index) [7]	Mesurer simultanément la compacité interne et la séparation entre clusters, en tenant compte de la densité quand c'est pertinent.
Inspection visuelle	- Nuage UMAP 3-D interactif-Projection **RADVIZ** 2-D	Vérifier la cohérence géométrique globale et repérer les possibles mélanges ou outliers.
Analyse descriptive	Statistiques (médianes, IQR, corrélations) **par cluster**	Comprendre la signature hydrogéologique de chaque groupe et mettre en évidence les variables discriminantes.

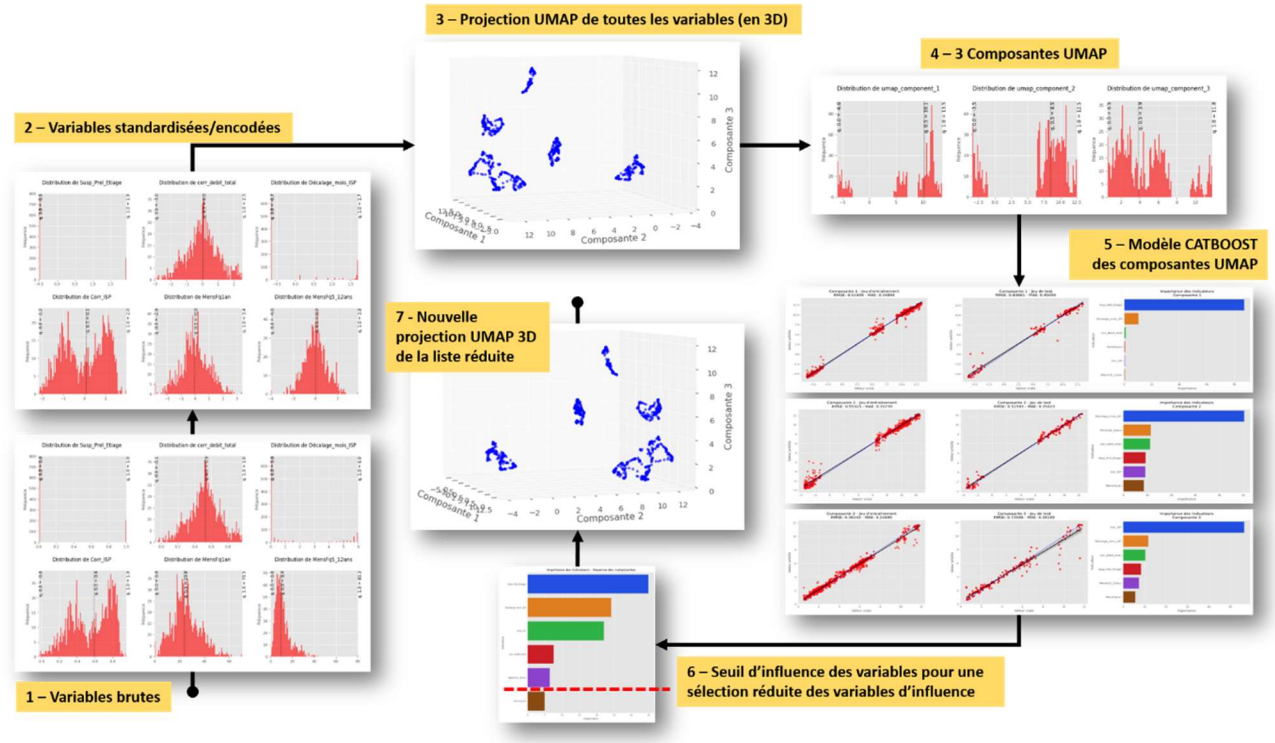


Figure 3 Description de la séquence algorithmique pour la méthode de regroupement à partir d'indicateurs hydrodynamiques et d'algorithmes de science de la donnée, partie 1/2

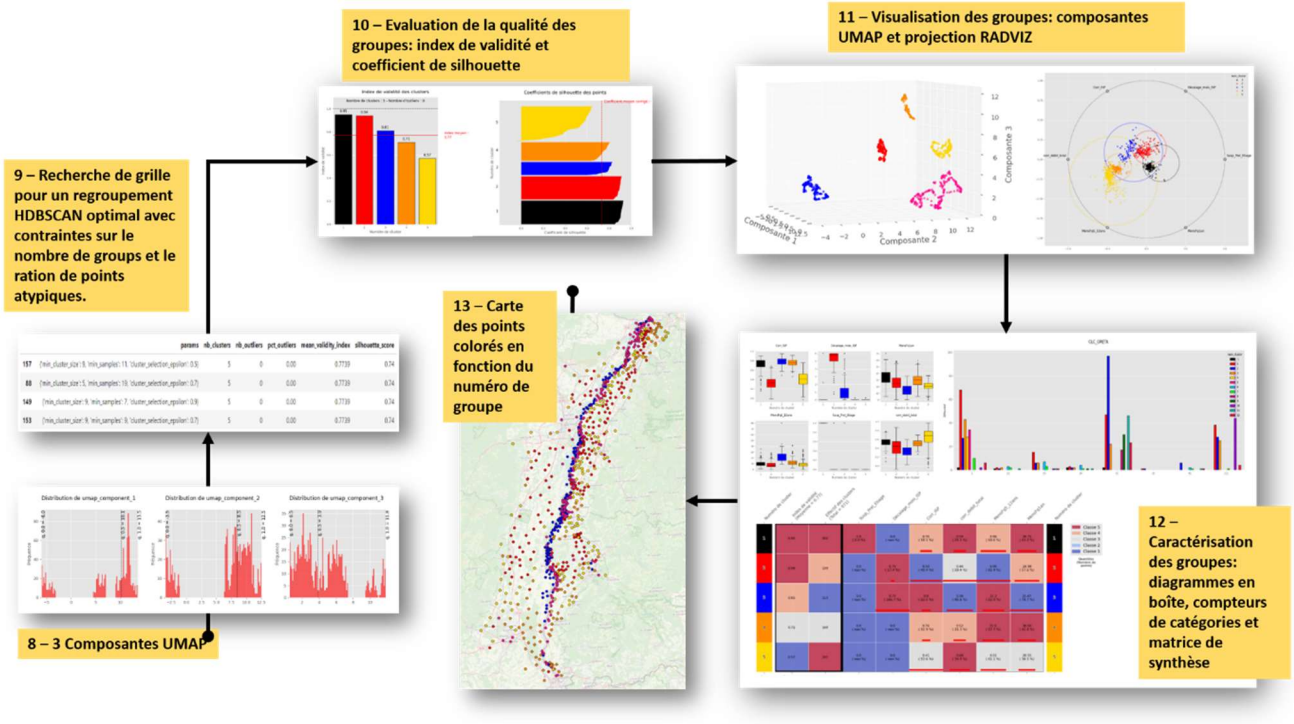


Figure 4 Description de la séquence algorithmique pour la méthode de regroupement à partir d'indicateurs hydrodynamiques et d'algorithmes de science de la donnée, partie 2/2

g. Sélection et interprétation finales

La solution retenue est celle qui maximise à la fois les indices Silhouette et DBCV tout en conservant le taux de points classés le plus élevé possible. On obtient ainsi un découpage robuste, limité en points aberrants, et facilement interprétable en termes de processus hydrogéologiques).

En résumé

1. **UMAP** fournit une réduction de dimension rapide et fidèle, essentielle pour la visualisation et l'entrée de HDBSCAN.
2. **CatBoost** sert de passerelle entre les variables brutes et l'espace réduit : il **quantifie objectivement** l'influence de chaque variable.
3. **HDBSCAN** détecte des groupes sans imposer de nombre prédéfini et laisse la liberté de considérer certains points comme du bruit.
4. Le **bouclage UMAP → CatBoost → filtrage** épure progressivement l'information, ce qui facilite l'interprétation métier.
5. La **combinaison des indices Silhouette et DBCV** garantit que la solution finale est non seulement compacte et bien séparée, mais aussi cohérente avec la distribution réelle des densités.

Ainsi, cette chaîne aboutit à des **groupes physiquement pertinents** tout en restant adossée à des méthodes d'intelligence artificielle éprouvées.

4 RESULTATS par méthode

4.1 Regroupement par corrélation

L'approche séquentielle et récursive décrite plus haut (cf. section sur les Méthodes) a été mise en œuvre en suivant la démarche ci-dessous afin de produire les résultats basés sur la corrélation entre les chroniques :

D'abord pour la production d'un « Clustering Expert version 1 » (en 2024) par PAM sans utilisation d'un module de post-traitement (alors inexistant) :

1. Sélection des 1244 séries du jeu de données de départ ayant assez de données
2. Premier clustering de l'ensemble des points sélectionnés, par k-médoïdes, avec k = 6 : on obtient 3 clusters de points proches du Rhin et 3 autres clusters de points majoritairement à distance du Rhin. On définit ainsi **deux grands groupes explicables** : Rhin (influence forte du Rhin) vs Non-Rhin (influence faible à nulle du Rhin).
3. Ce faisant, on en profite pour écarter temporairement 40 points dont la localisation n'est pas cohérente avec l'explication donnée à leur cluster. Les points éloignés du Rhin et pourtant inclus dans le groupe « Rhin » ou, à l'inverse, les points très proches du Rhin mais inclus dans le groupe « Non-Rhin », sont ainsi mis à l'écart pour le moment → **groupe « Entre-deux »** ('MISC_in_betweens').
4. **Sous influence proximale du Rhin (A)** : Un clustering de la sous-sélection des 545 points du groupe « Rhin » (famille 'A') est ensuite réalisé (k = 6). Mais avec 6 clusters, l'interprétation de la distribution spatiale n'est pas aisée. Seul un de ces 6 clusters se distingue nettement du reste : un cluster de points localisés en amont du barrage agricole de Kehl-Strasbourg, dont la dynamique piézométrique particulière s'expliquerait par les effets de ce barrage → **groupe « Rhin Kehl »** ('RR_A6'). Les points des 5 autres clusters sont mélangés dans un **groupe « Rhin autres »** ('RR_Aothers').
5. **Hors influence proximale du Rhin (B)** : Un clustering de la sous-sélection des 659 points du groupe « Non-Rhin » (famille 'B') est ensuite effectué (k = 6). Des explications hydrogéologiques sont trouvées pour 3 des 6 clusters alors obtenus (**B2**, **B5** et **B6** ; soit 222 points). Les 437 points des 3 autres clusters (B1, B3 et B4) sont mélangés pour constituer une sous-sélection « Bbis ».
6. On effectue une autre itération de clustering par k-médoïdes (là encore avec k = 6) à partir de cette sous-sélection de points (famille 'Bbis'). Seul un des clusters ('**Bbis2**') paraît explicable, par une influence estivale et une distribution spatiale surtout établie en Allemagne. Les points des 5 autres clusters (Bbis1 et Bbis3 à Bbis6) sont donc mélangés pour former un **groupe « Non-Rhin autres »** ('PA_Bothers').
7. C'est ainsi que le Clustering Expert version 1, composé de 1244 chroniques, a été produit.

Puis, en 2025, ces premiers résultats intermédiaires ont été améliorés, bonifiés, afin :

- D'intégrer des séries piézométriques jusque-là ignorées car plus courtes ;
- De gérer des cas particuliers jugés mal classés d'un point vue expert basé sur une analyse visuelle de la série par rapport aux autres membres du cluster attribué, sur les indicateurs calculés par l'outil, sur la localisation du point par rapport aux autres en tenant compte des clusters attribués aux points, etc. ;

- D'optimiser manuellement le choix du médoïde de certains clusters, notamment afin de s'assurer que le médoïde soit une série temporelle la plus représentative mais aussi la plus complète que possible, pour le cluster.

L'intégration de 378 chroniques piézométriques plus courtes au jeu de données traité pour la production du « Clustering Expert version 2 » a été rendue possible en considérant un critère de remplissage des séries moins exigeant (remplissage ≥ 1/5 des mois dans la période ciblée 1980-2024) (cf. section sur les Données disponibles).

La démarche de production du « Clustering Expert version 2 » (en 2025) se résume comme suit :

1. **Post-traitement** d'un tableau des séries avec leur cluster attribué par le Clustering Expert version 1 (ainsi que l'identification des médoïdes), pour obtenir de premiers résultats et tester l'outil de post-traitement.
2. Ajout au tableau des 378 identifiants des séries « plus courtes », en leur attribuant un cluster fictif temporaire au nom quelconque ('ADDING_less_dense').
3. Nouvelle exécution du module de post-traitement. On obtient entre autres résultats les identifiants des clusters voisins (c.-à-d. les mieux corrélés à la série). Le « premier voisin » est utilisé pour attribuer vraiment un cluster à chacun des séries « plus courtes ».
4. Nouvelle exécution du module de post-traitement. On explore alors plus finement les résultats pour repérer les séries trop dissemblables (séries singulières, sans lien explicatif ni corrélation, pour la plupart). Celles-ci sont déplacées dans un nouveau groupe très hétérogène d'individus « aberrants » ('**MISC_outliers**')[4] créé afin de retirer ces séries des groupes principaux (Figure 5).
5. Quelques itérations supplémentaires de modifications au tableau d'entrée listant les séries, leur groupe associé et le médoïde désigné pour chaque groupe, afin d'améliorer encore marginalement les résultats : déplacement de séries dont l'appartenance aux familles Rhin ou Non-Rhin est peu claire → vers le groupe des « Entre-deux » entre Rhin et Non-Rhin ('**MISC_in_betweens**') ; choix d'une meilleure série médoïde, plus longue ; etc.
6. Dernière exécution du module de post-traitement afin de produire les résultats finaux du « Clustering Expert version 2 » (tableaux, graphiques, fichiers SIG, …) et cartographie de ceux-ci.

La Figure 5 ci-dessous illustre avec un exemple de série « aberrante » (point « 166/023-9 ») les étapes suivies pour déplacer ces séries dans le groupe dédié 'MISC_outliers'. La série a d'abord été repérée dans un graphique montrant la composition du cluster attribué initialement (par PAM) à la série (ici = 'PA_Bbis2'). Un graphique généré spécifiquement pour cette série a confirmé que c'était bien cette série qui déviait largement des autres courbes du groupe, en particulier de son médoïde ! En relançant le post-traitement après avoir déplacé la série dissemblable dans le groupe 'MISC_outliers', on obtient un groupe 'PA_Bbis2' plus homogène, sans déviation extrême. On voit enfin, dans le graphique inférieur droit, que la série du point « 166/023-9 » est à sa place dans ce groupe très hétérogène de séries singulières.

[4] Les séries composant le groupe 'MISC_outliers' ont été repérées principalement en examinant les résultats générés par le module de post-traitement. Une série a été considérée aberrante soit parce que sa courbe déviait trop dans un graphique présentant la composition du cluster lui étant attribué ; soit en raison d'autres particularités, ex. lorsque c'était le seul point associé à tel cluster dans les environs, avec un signal corrélé à aucun de ses voisins.

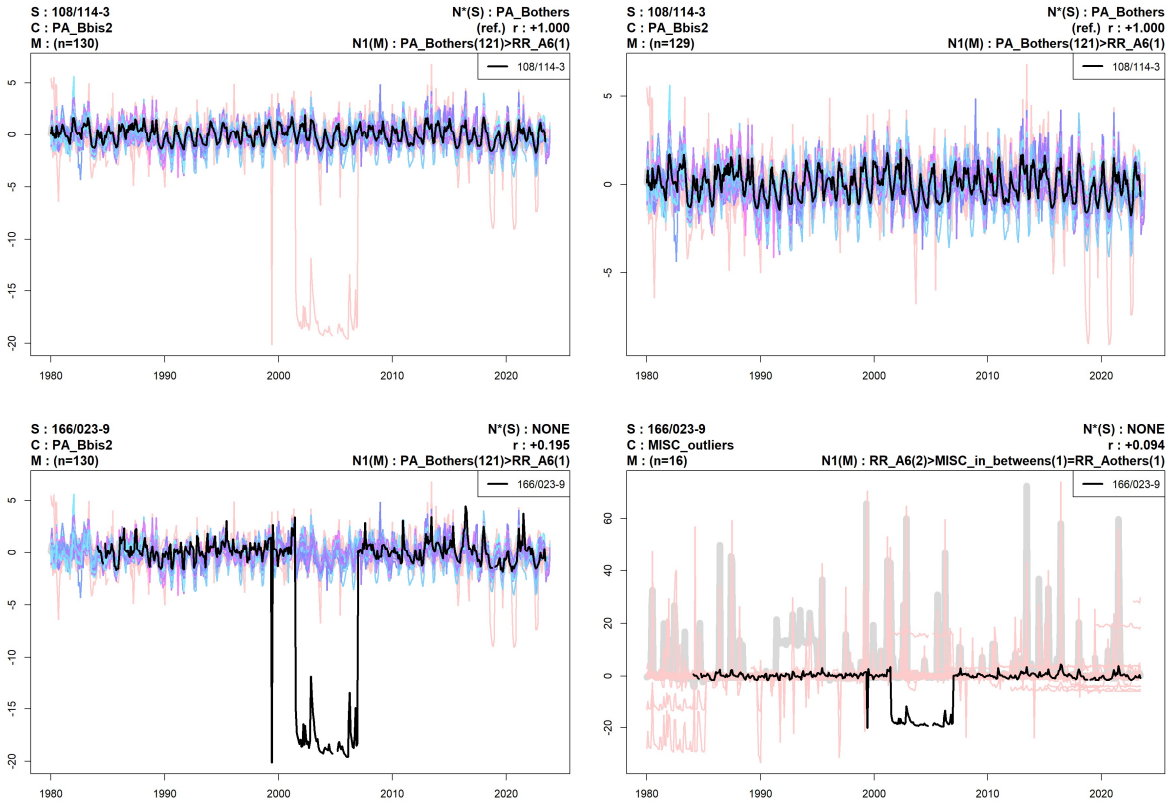


Figure 5 Exemple de série « aberrante » déplacée vers un groupe dédié lors du post-traitement des résultats du regroupement basé sur la corrélation (La courbe noire en avant-plan = le médoïde du groupe. Les autres courbes, en arrière-plan, ont des couleurs aléatoires ; abréviations : cf. Annexe 150)

La carte suivante (Figure 6 et Figure 7) présente une vue cartographique des résultats finaux du « Clustering Expert version 2 » (basé essentiellement sur la corrélation entre les chroniques). On distingue assez nettement la plupart des 7 groupes principaux. La légende (

Tableau 3) qui accompagne la carte décrit chaque groupe avec l'explication hydrogéologique pour chacun d'eux. Une notation de type « CB_# » est utilisée dans la suite de ce rapport pour référer aux clusters basés sur les corrélations (*« Correlation Based »*).

Le Tableau 2 ci-dessous, quant à lui, décrit la cohérence et l'homogénéité du contenu de chaque groupe par des statistiques calculées à partir des coefficients de corrélation (r) entre les N séries du groupe et la série médoïde (de référence) du groupe. On constate ainsi que :

- La plupart des groupes principaux (CB_1 à CB_4 et CB_6) sont composés de séries qui sont presque toutes dans leur « meilleur » groupe d'après les corrélations aux médoïdes (« % *r best* »), soit que : r (série, médoïde du groupe attribué) = MAX(r (série, médoïdes des groupes…)).
- Les deux groupes principaux montrant un moins bon indice de cohérence d'après les r sont les deux plus gros groupes, formés par le mélange des « autres points » d'abord placés dans plusieurs clusters lors des itérations de clustering par l'algorithme PAM : CB_5 ('PA_Bothers') et CB_7 ('RR_Aothers'). Cette hétérogénéité n'est donc pas étonnante.
- De même, l'indicateur « % r very low (r<0.5) » souligne la proportion plus élevée de séries très faiblement corrélées au médoïde dans le cas du groupe CB_7 ('RR_Aothers'). Ce qui n'est pas étonnant non plus, car même si les séries piézométriques influencées par le Rhin se distinguent généralement bien des séries non influencées ou éloignées du Rhin, ce grand groupe de >650 individus rassemble une variété d'expressions locales de ces influences naturelles et/ou anthropiques que peut exercer le Rhin. Néanmoins, comme expliqué plus haut, il n'a pas semblé pertinent de conserver individuellement les 5 autres clusters « sous influence proximale du Rhin » formés à l'itération de clustering des points de la famille 'A', en raison de leur répartition spatiale difficile à expliquer par le contexte.
- Le groupe CB_X0 ('MISC_in_betweens') contient une proportion encore plus élevée (63.5 %) de séries très faiblement corrélées au médoïde du groupe, dont 27.0 % de séries qualifiées de « trop mal corrélées » (r < 0.25). Ceci rappelle simplement que ce groupe est composite et qu'il devra être retravaillé (éclaté et redistribué) vers d'autres groupes plus tard (lors de la « Synthèse des résultats »).
- Enfin, le groupe CB_XX ('MISC_outliers') sert à écarter et signaler quelque 25 séries qui apparaissent aberrantes, afin d'éviter qu'elles ne soient intégrées aux groupes principaux. L'hétérogénéité de ce groupe, soulignée notamment par la forte proportion de séries trop mal corrélées au médoïde (choisi automatiquement en phase de post-traitement pour représenter tant bien que mal ce groupe), est donc un constat cohérent avec la définition même de ce groupe.
- Noter que ces deux groupes de séries à replacer ou écarter (CB_X0 et CB_XX) contiennent relativement peu d'individus (74 + 25 ≈ 100 chroniques piézométriques au total).

Tableau 2 Cohérence et homogénéité du contenu de chaque groupe (CB_1-7, et CB_X) par des statistiques calculées à partir des coefficients de corrélation (r) entre les N séries du groupe et la série médoïde (de référence) du groupe

Group	% r best (= max r)	% r high (r>0.8)	% r high & best	% r low (r<0.8)	% r very low (r<0.5)	% r too bad (r<0.25)	N points TOTAL	n points r very low
CB_1	100.0	61.7	61.7	38.3	4.9	0.0	81	4
CB_2	98.3	53.8	53.8	46.2	5.1	0.0	117	6
CB_3	97.8	52.7	52.7	47.3	3.3	0.0	91	3
CB_4	95.9	71.0	69.4	29.0	3.6	0.0	193	7
CB_5	56.9	49.2	37.2	50.8	4.6	0.3	325	15
CB_6	95.7	54.3	54.3	45.7	8.7	2.2	46	4

CB_7	71.6	12.8	12.7	87.2	34.1	3.8	656	224
CB_X0	56.8	6.8	6.8	93.2	63.5	27.0	74	47
CB_XX	16.0	4.0	4.0	96.0	92.0	84.0	25	23
Globally:	75.6	35.6	33.0	64.4	20.7	4.2	1608	333

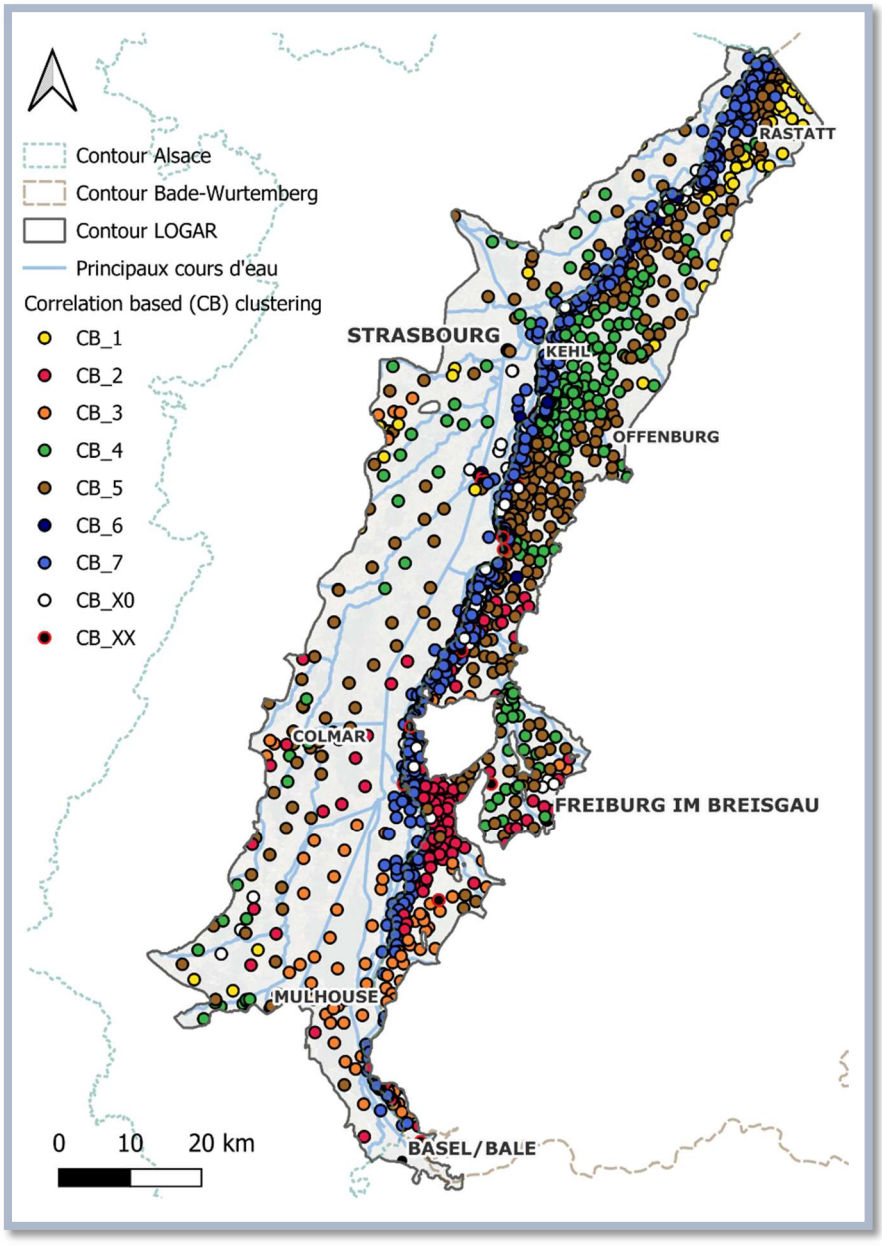


Figure 6 Carte des groupes obtenus par la méthode de regroupement basée sur la corrélation entre les chroniques

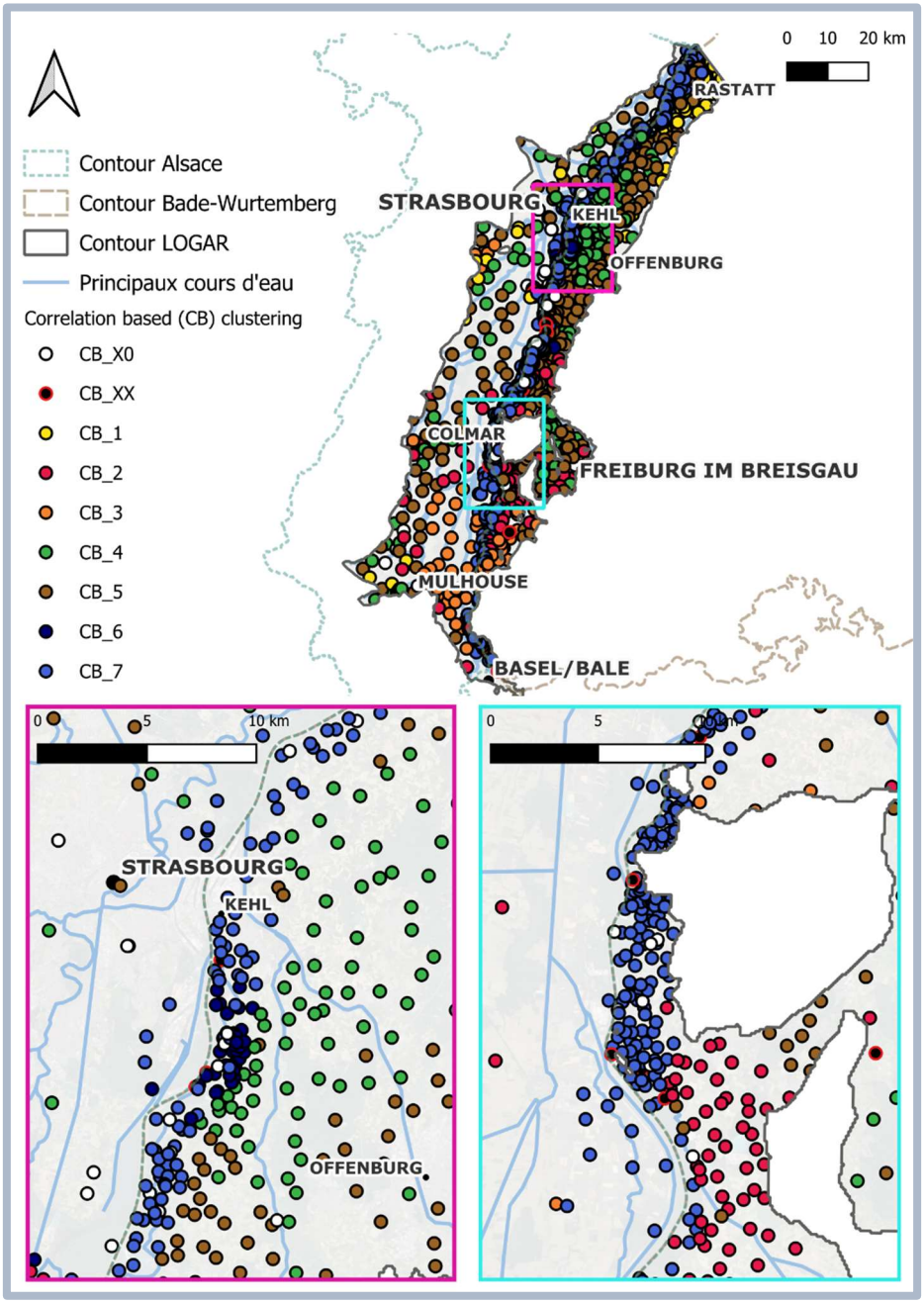


Figure 7 Carte des groupes obtenus par la méthode de regroupement basée sur la corrélation entre les chroniques : zooms sur les secteurs de Kehl et Breisach

Tableau 3 Caractérisation et description des 9 groupes obtenus par la méthode de regroupement basée sur la corrélation entre les chroniques franco-allemands

N° groupe	Caractérisation
CB_1 (PA_B2) n = 81	Zone **inertielle** *principalement* au Nord du secteur allemand (environs de Rastatt – Karlsruhe) avec une **épaisse zone non saturée (ZNS)** presque partout **>5** mètres et par conséquent d'importantes composantes pluriannuelles dans leur dynamique. Explication appuyée par des indicateurs de temps de demi-décroissance (de vidange de l'aquifère) longs ainsi que par des **temps d'arrivée des précipitations importants** aussi.
CB_2 (PA_B5) n = 117	Points localisés majoritairement (~2/3 des points) au Sud d'une ligne Ouest-Est entre Sélestat et Lahr/Schwarzwald. Dont environ la moitié des points concentrés dans une zone relativement étroite de la rive droite du Rhin entre Vieux-Brisach (Breisach am Rhein) et Bad Krozingen. **L'inertie** (**relativement importante** mais **sans délai notable par rapport aux pluies**) semble jouer un rôle important dans l'établissement de ce cluster. **ZNS** là aussi souvent **>5** mètres.
CB_3 (PA_B6) n = 91	Points situés dans le Haut-Rhin, concentrés au Sud d'une ligne Ouest-Est entre Colmar et Fribourg-en-Brisgau. Groupe caractérisé par une **ZNS encore plus épaisse** en général (épaisseur médiane de la ZNS >10 mètres) avec un comportement plus inertiel que CB_2, mais de même ordre que CB_1. Alimentation de l'aquifère par **Sundgau**. Très bonne cohérence avec les longs délais estimés d'arrivée des précipitations.
CB_4 (PA_Bbis2) n = 193	Points situés **dans la plaine du Rhin franco-allemande**, **hors influence significative du Rhin**, très majoritairement en contexte agricole ou proche-urbain ; cohérent avec une dynamique qui apparaît **souvent impactée par des prélèvements estivaux** (anthropiques ou naturels). Faibles épaisseurs de **ZNS <5 mètres** pour >95 % des points du groupe. Concentration principale des points du groupe (>50 %) à l'est de Strasbourg, en Allemagne entre Bühl et Offenburg. Concentration secondaire (~15 %) autour de Fribourg-en-Brisgau jusqu'à Riegel am Kaiserstuhl. Pas de concentration notable des points du côté français (~25 %), dispersés sur toute la longueur nord-sud de la plaine d'Alsace, majoritairement éloigné du Rhin.
CB_5 (PA_Bothers) n = 325	Groupe rassemblant les **autres points de la plaine du Rhin les deux côtés de la frontière,** également **hors influence significative du Rhin** ; sans explication forte pour le distinguer du groupe CB_4 (corrélation entre médoïdes CB_5 et CB_4 : r = +0.88) si ce n'est une **inertie légèrement plus importante** en moyenne et une dynamique plus **rarement impactée** par des **prélèvements** estivaux conséquents. Faibles épaisseurs de **ZNS <5** mètres pour ~85 % des points du groupe. Presqu'aucun point ne se trouve à proximité du Rhin (**en général >1–2 km de part et d'autre du Rhin**).
CB_6 (RR_A6) n = 46	Points **fortement influencés par le Rhin**, plus précisément **impactés par le barrage agricole de Kehl-Strasbourg** au sud-est de la ville. Points concentrés **à l'amont** du barrage, sur la rive droite du Rhin seulement (imperméabilisation anthropique de la rive gauche coté Strasbourg). Evolution temporelle (signature) très particulière de la piézométrie caractérisée par des **niveaux nettement plus bas avant le milieu des années 1980** (hausse soudaine des niveaux vers 1985).

CB_7 (RR_Aothers) n = 656	Groupe rassemblant les **autres points sous forte influence du Rhin** ; sans explication forte pour en distinguer des sous-groupes autres que CB_6. Comportements (évolutions) piézométriques homogènes dans l'ensemble. Différences subtiles entre les sous-groupes qui l'ont composé (clusters de la famille 'A' sauf 'A6' devenu CB_6) mais leur répartition spatiale dispersée est apparue difficilement explicable. Remarque : Il serait **possible, techniquement, d'affiner le découpage** de ce groupe, mais il n'apparaît pas particulièrement utile de le faire d'un point de vue utilitaire pratique.
CB_X0 (MISC_in_ betweens) n = 74	**Points retirés** des groupes principaux à cause d'une **incohérence spatiale** entre leur localisation et leur cluster initialement attribué (lors de la première itération de clustering par corrélation établissant les deux grands groupes Rhin versus Non-Rhin) : soit le point était placé dans un des clusters Rhin alors qu'il était éloigné du Rhin ; soit il était placé dans un des clusters Non-Rhin tout en étant très proche du Rhin. Ces points ne sont pas définitivement écartés, mais plutôt mis de côté, pour une éventuelle réintégration dans les groupes principaux lors de la phase à suivre de « Synthèse des résultats ».
CB_XX (MISC_ outliers) n = 25	Points dont la chronique montre une **évolution piézométrique très singulière voire anormale**. Ce groupe permet d'écarter des séries trop peu corrélées aux médoïdes des groupes principaux CB_1 à CB_7, avec une évolution trop rare dans le jeu de données pour qu'elle ait mené à la formation d'un cluster dédié ; et des séries cassées par une rupture (changement important et soudain) dans l'évolution de leurs niveaux (probablement dû à des erreurs lors du calcul des cotes piézométriques à partir des données de profondeur d'eau). Remarque : Cette liste de points jugés « aberrants » à ce stade est révisée plus tard, lors de la « Synthèse des résultats ».

4.2 Regroupement par caractéristiques dynamiques des chroniques

Au moyen des indices de validation décrits, six clusters ont été identifiés comme regroupement optimal. Ces six clusters peuvent être classés en deux catégories principales en fonction de leurs propriétés dynamiques et de leur localisation spatiale : les clusters à influence dominante du Rhin et les clusters à influence dominée par les précipitations.

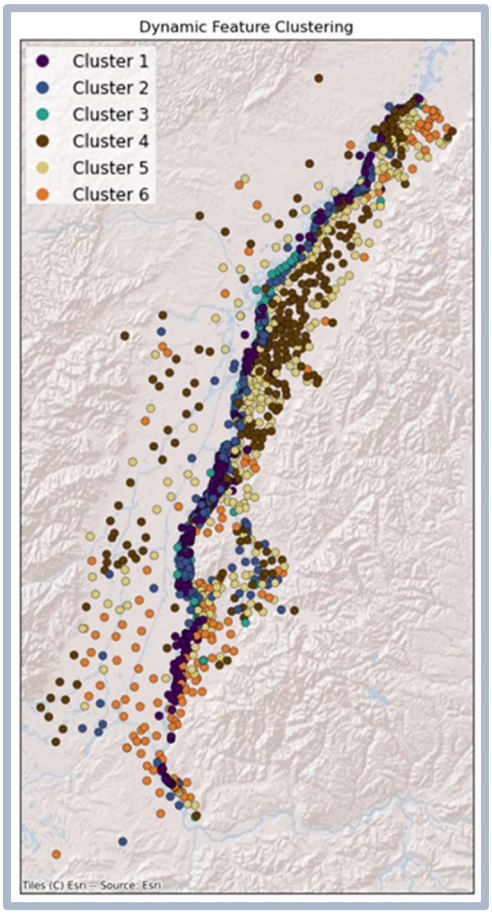


Figure 8 Carte des groupes obtenues par la méthode de regroupement basée sur l'utilisation de caractéristiques dynamique des chroniques

Clusters influencés par le Rhin

Les clusters 1 à 3 présentent un comportement dynamique des eaux souterraines clairement influencé par le Rhin. Leurs piézomètres réagissent fortement aux niveaux d'eau du fleuve, à l'infiltration, au couplage hydraulique et aux interventions anthropiques telles que la gestion des crues. Ils se caractérisent par des réponses rapides aux variations du niveau d'eau du Rhin, des amplitudes prononcées lors des hautes eaux, des cycles saisonniers bien définis et, dans certains cas, des influences techniques liées à des opérations de régulation (inondations contrôlées, effets de retenue, etc).

Les trois clusters présentent une périodicité annuelle élevée (P52) et une bonne concordance avec le cycle hydrologique annuel (SB) – caractéristiques typiques des systèmes nappe-rivière à

dynamique saisonnière. Les courtes phases de vidange de la nappe (LRec) témoignent d'un drainage rapide vers le Rhin, tandis que les phases de hautes eaux souterraines relativement courtes (HPD) indiquent un couplage direct et perméable avec le système fluvial. Ces propriétés hydrologiques distinguent clairement les clusters 1 à 3 des clusters 4 à 6.

Malgré leur influence commune du Rhin, les clusters 1 et 3 se distinguent particulièrement par leur dynamique :

- le cluster 1 représente un système fortement saisonnier, caractérisé par une amplitude relative élevée (RR), des valeurs marquées de P52 et SB, et des phases de vidange courtes. Les réactions rapides et liées aux crues, la courte durée des phases de hautes eaux (HPD) et une valeur médiane intermédiaire indiquent un système actif et perméable, principalement contrôlé par le régime du Rhin.
- Le cluster 3, en revanche, présente un système différé et accumulatif avec une amplitude relative plus faible, des valeurs P52 plus faibles et une longue durée de vidange de la nappe. Les phases de hautes eaux prolongées, la médiane élevée et les fluctuations à court terme (SDdiff) moins importantes indiquent une dynamique des eaux souterraines lente et régulée, typique des zones de rétention, effets de contre-pression ou des zones d'inondations contrôlées.
- Le cluster 2 constitue un type de transition entre la régulation saisonnière du cluster 1 et le comportement caractérisé par une forte capacité de stockage du cluster 3, sans présenter les extrêmes.

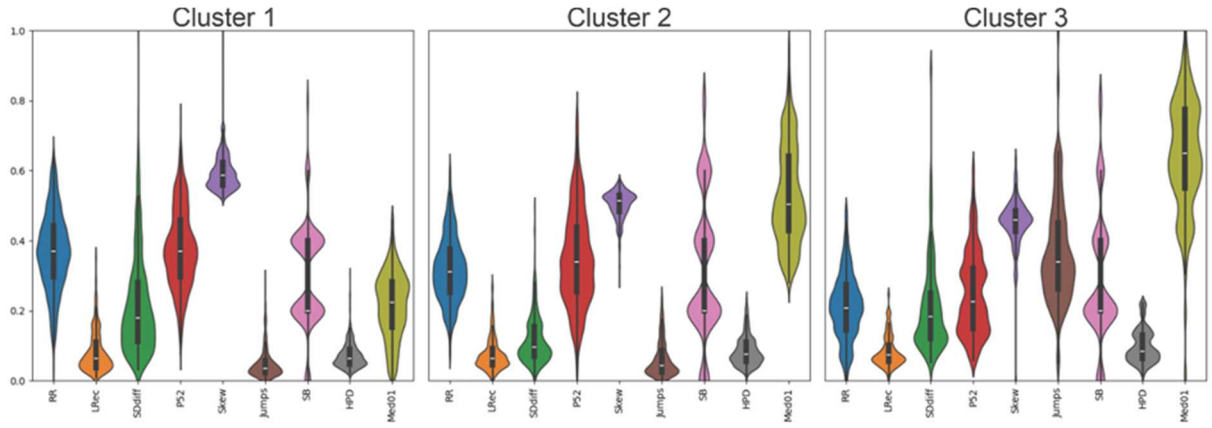


Figure 9 : Violon plots des neuf caractéristiques dynamiques sélectionnées pour les clusters 1 à 3. Les paramètres centraux représentés décrivent : la saisonnalité (P52, SB), la variabilité à court terme (SDdiff), le comportement en période de hautes eaux (HPD) ainsi que la position et la distribution des niveaux piézométriques (médiane, RR, LRec, skew, jumps)

Clusters dominés par les précipitations

Les clusters 4 à 6 représentent des dynamiques des eaux souterraines principalement influencées par les conditions climatiques, les caractéristiques locales de stockage et les contraintes topographiques, à la différence des clusters proches du Rhin qui sont directement connectés au réseau fluvial :

- Le cluster 4 présente le comportement saisonnier le plus marqué, avec une forte périodicité annuelle et une forte concordance avec le cycle hydrologique annuel typique.

La faible variabilité à court terme et les faibles valeurs de sauts indiquent un système stable, à influence climatique ; et non affecté par des perturbations externes majeures.

- Le cluster 6, en revanche, se caractérise par une faible saisonnalité, de longues phases de vidange de la nappe et des niveaux d'eau élevés persistants. Ces caractéristiques traduisent un système à inertie (réponse lente), avec une recharge souterraine retardée, ce qui est typique des zones présentant une forte épaisseur de la zone non saturée ou une faible perméabilité verticale.
- Le cluster 5 occupe une position intermédiaire entre ces deux extrêmes. Avec des caractéristiques modérées, il présente à la fois une régulation saisonnière et des signes inertiels. Il représente donc un régime mixte et équilibré, soumis à des influences mixtes, qui n'est ni très réactif ni totalement inertiel.

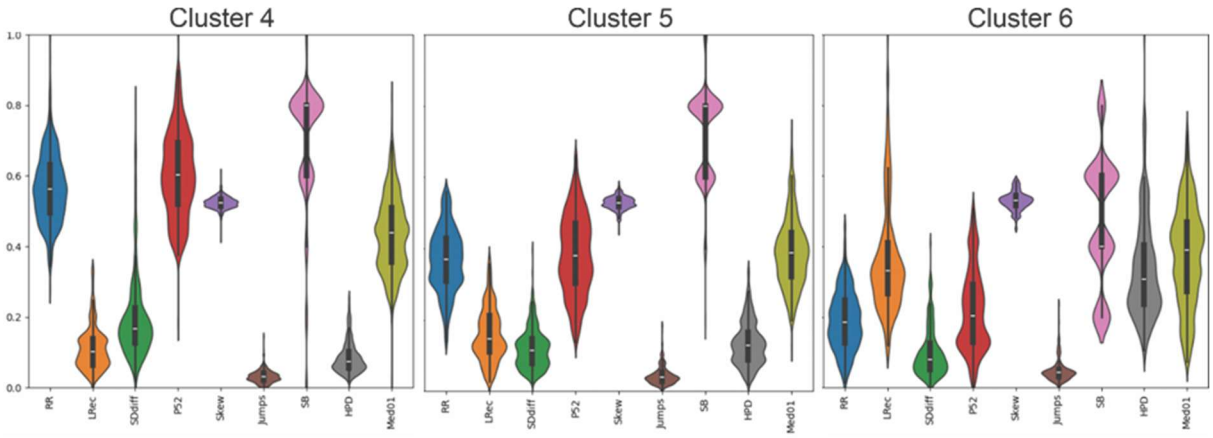


Figure 10 : Graphiques en violon des neuf caractéristiques dynamiques sélectionnées des niveaux des eaux souterraines pour les clusters 4 à 6. Sont représentées les valeurs caractéristiques centrales décrivant la saisonnalité (P52, SB), la variabilité à court terme (SDdiff), le comportement en hautes eaux (HPD), ainsi que la position et la répartition du niveau des eaux souterraines (médiane, RR, LRec, skew, jumps).

4.3 Regroupement par indicateurs hydrodynamiques

Les trois meilleurs tests de regroupement obtenus par cette méthode présentent les valeurs de coefficient de silhouette et d'index de validité suivantes ainsi que le nombre de points atypiques suivants :

N° test	Coefficient de silhouette	Index de validité	Nombre de points atypiques
1	74	77	0
2	67	76	3
3	67	76	7

Le test n°1 est donc retenu. Ce test a permis le regroupement des piézomètres utilisés en 5 groupes (Figure 11). Les 6 variables retenues par la chaîne de traitement sont :

- l'indicateur de précocité des étiages (lié aux prélèvements anthropiques et naturels) (Susp_Prel_Etiage),

- l’indicateur de décalage entre les pluies et la réponse piézométrique (Décalage_mois-ISP),
- la corrélation aux pluies (Corr_ISP),
- la corrélation au débit du Rhin à Maxau (Corr_débit_total),
- l’indicateur de la contribution de la fréquence 5-12 ans au signal piézométrique (MensFq5-12ans),
- l’indicateur de la contribution de la fréquence annuelle au signal piézométrique (MensFq1an).

Parmi les 5 groupes construits à partir des indicateurs hydrodynamiques (HI) identifiés (cf. carte Figure 11), HI_2 et HI_5 se distinguent par leur concentration autour du Rhin, tandis que HI_1, HI_3 et HI_4 couvrent les zones de plaine et de piémont. HI_3 semble également, entre autres, couvrir les zones connues pour leur inertie plus importante (secteur de la Hardt en Alsace et de Rastatt dans le Bade-Wurtemberg).

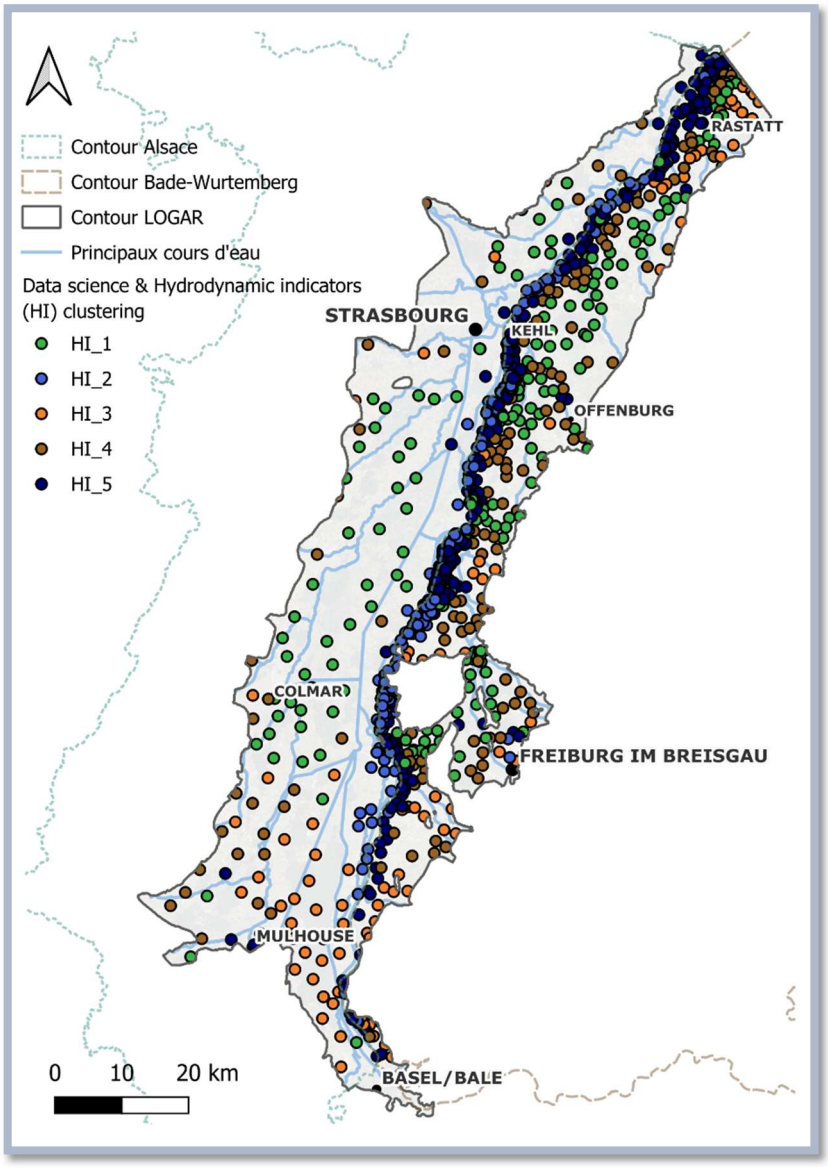


Figure 11 Carte des groupes obtenus par la méthode de regroupement basée sur l'utilisation d'indicateurs hydrodynamiques

Les 5 groupes constitués sont caractérisés par la répartition des valeurs des variables telle que présentée à la Figure 12, avec sa matrice de caractérisation selon 5 classes de quantiles (valeurs sériées (de la plus faible (couleur bleu) à la plus élevée (couleur rouge)). La Figure 13 présente la distribution des variables quantitatives des groupes avec les boites à moustaches.

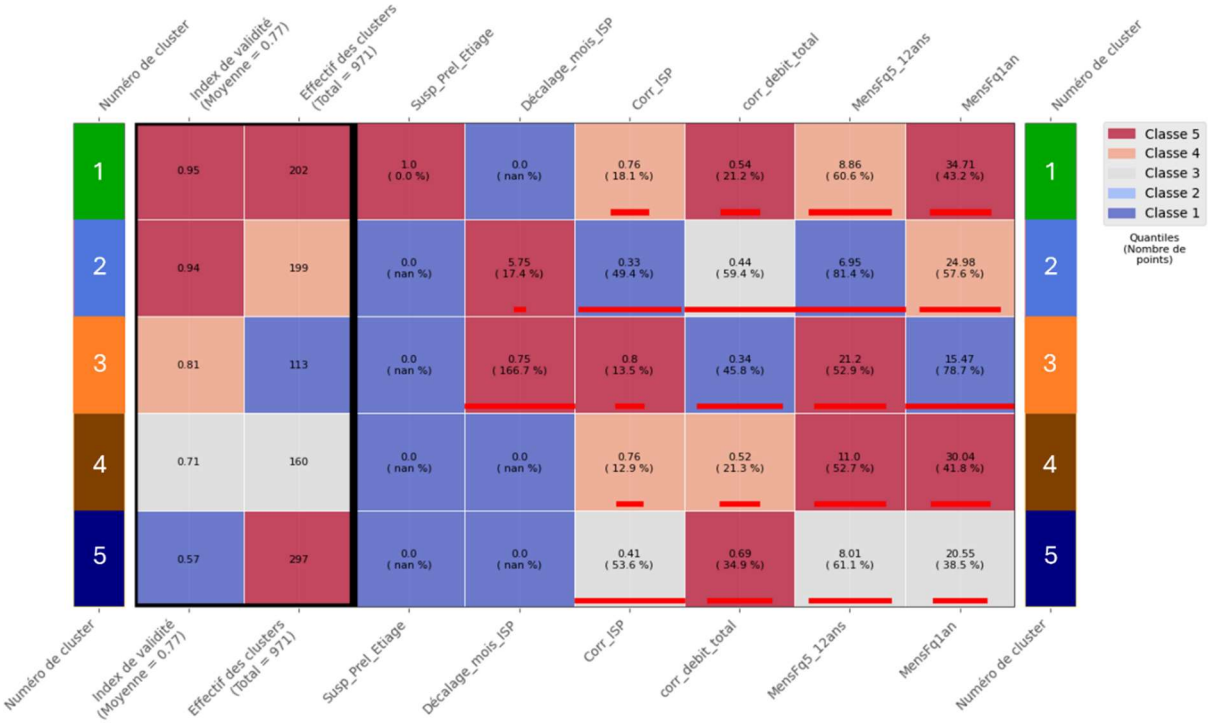


Figure 12 Matrice de caractérisation des groupes obtenus par la méthode de regroupement basée sur les indicateurs hydrodynamiques - Médiane (CVR %). Couleurs des classes : valeur sériée de la plus faible (couleur bleu) à la plus élevée (couleur rouge)

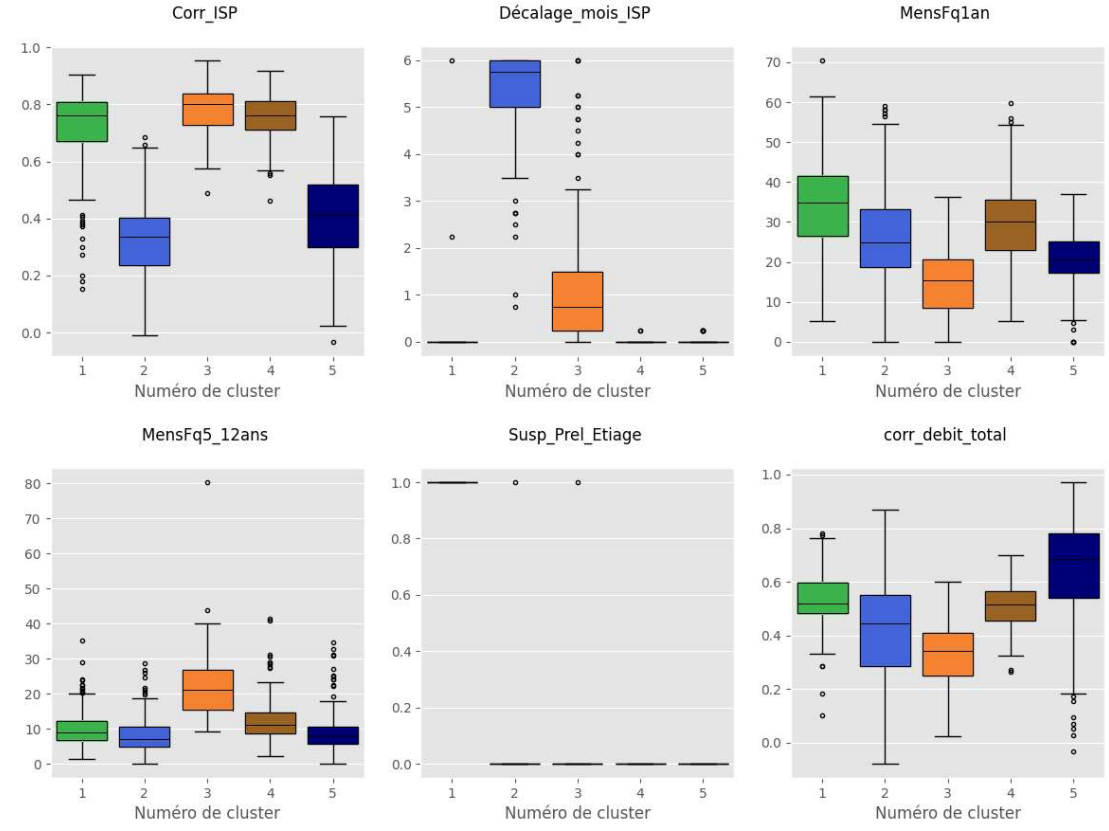


Figure 13 Distribution des variables quantitatives des groupes obtenus par la méthode basée sur les indicateurs hydrodynamiques

La visualisation Radviz des groupes constitués (Figure 14) permet de vérifier la cohérence géométrique des groupes et de constater que ces derniers sont bien distincts et présentent peu de mélange.

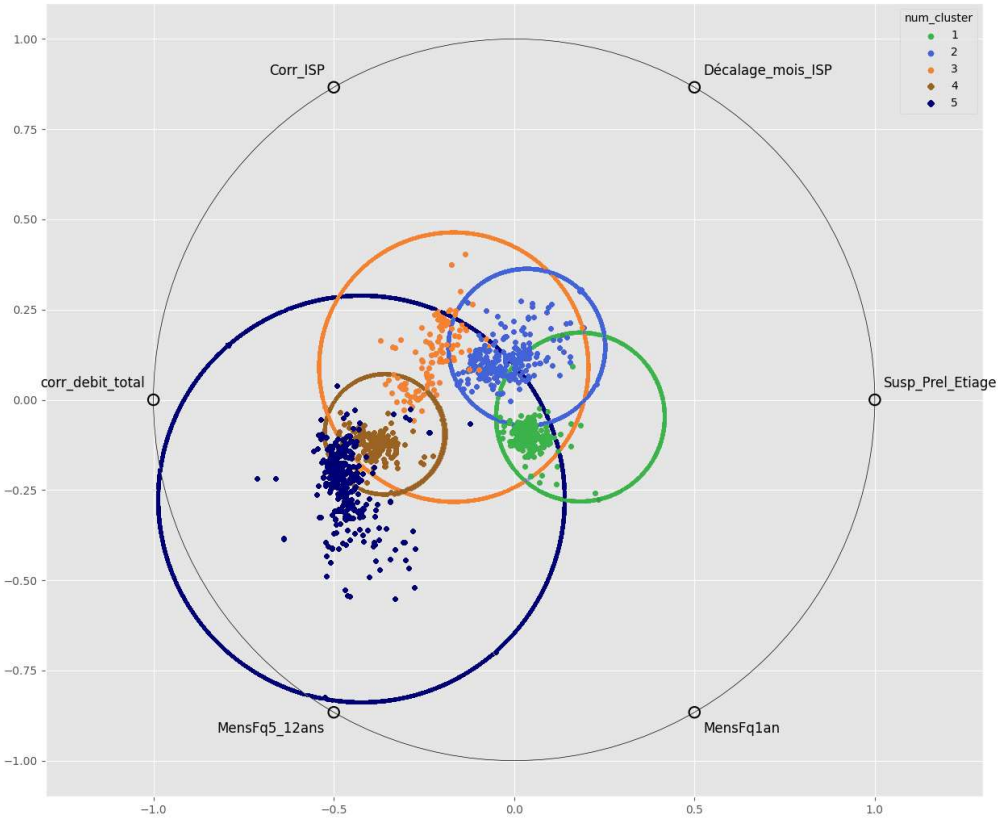


Figure 14 Visualisation Radviz des variables standardisées pour les groupes obtenus par méthode basée sur les indicateurs hydrodynamiques

Les tendances, et notamment les tendances linéaires multisegments calculées par l'outil CensoredStats intégré à Qualistat [11] n'ont pas été intégrées en tant que variables explicatives au regroupement car elles sont une résultante du comportement hydrodynamique de la nappe et de sa réaction aux conditions de recharge et aux prélèvements. Toutefois, les tendances significatives ont été analysées dans le cadre du post-traitement du regroupement, lors de la phase de caractérisation des groupes. Ainsi, l'analyse conjointe des tendances (en italique) avec les caractéristiques des groupes formés permet la caractérisation suivante de chacun des groupes (Tableau 4) :

Tableau 4 Caractérisation et description des 5 groupes obtenus par la méthode de regroupement basée sur les indicateurs hydrodynamiques

N° groupe	Caractérisation
1	Le groupe n°1 présente une **forte corrélation à la pluie**, sans décalage significatif, avec un **cycle annuel dominant**. Ce groupe est **modérément corrélé** aux débits du **Rhin** et montre une **précocité des étiages** importante (lié aux prélèvements, anthropiques ou non). *Les tendances observées dans ce groupe montrent une baisse modérément claire*.

2	Le groupe n°2 présente une **faible corrélation à la pluie**, avec un important décalage d'environ 5 à 6 mois en moyenne. Le **cycle annuel** est dominant. Le groupe est **modérément corrélé** aux débits du **Rhin**.
3	Le groupe n°3 est **fortement corrélé à la pluie**, avec un décalage d'environ 0.5 à 1.5 mois. La contribution de la fréquence pluriannuelle de 5 à12 ans dans le signal piézométrique de ces points est importante. Les piézomètres du groupe sont **peu corrélés** aux débits du **Rhin**. *Une tendance à la baisse est clairement identifiée*.
4	Le groupe n°4 montre une **forte corrélation à la pluie**, sans décalage significatif. Le **cycle annuel est dominant** dans les signaux piézométriques de ce groupe. Ces piézomètres montrent une **corrélation modérée avec** les débits du **Rhin**. *Une tendance à la baisse est modérément claire sur ces points*.
5	Le groupe n°5 montre une **faible corrélation à la pluie**, sans décalage significatif. Le cycle annuel est peu dominant. Les chroniques piézométriques de ce groupe sont toutefois **fortement corrélées** aux débits du **Rhin**.

5 SYNTHESE DES RESULTATS : Sectorisation de l'aquifère rhénan en grands ensembles

5.1 Préparation

Les points (piézomètres) de la zone d'étude ont été regroupés indépendamment selon 3 méthodes différentes, avec :

- un nombre de chroniques piézométriques considérées différent selon les contraintes respectives (certaines séries étant trop courtes ou lacunaires pour permettre le calcul d'indicateurs requis par une méthode de regroupement, par exemple)
- un nombre de regroupements proposés différent :

Regroupement basé sur…	**Nb de points**	**Nb de groupes**
corrélation entre chroniques	1608 / 1608	9 (7 + 2)
caractéristiques dynamiques	1554 / 1559	6
indicateurs hydrodynamiques	965 / 971	5

Remarque : Dans la suite des analyses pour la synthèse des résultats, seuls les points présents dans le regroupement par corrélation entre chroniques seront considérés, d'où les nombres de points indiqués dans ce tableau légèrement inférieurs pour les deux autres méthodes par rapport au nombre de points réellement traités (cf. « Nb points » = [points traités par la méthode et aussi par corrélation…] / [points traités par la méthode]).

L'**objectif** de ce travail de synthèse a été de construire un regroupement qui résume autant que possible l'ensemble de ces résultats, par la comparaison et la combinaison de ces derniers.

Pour ce faire, il a été d'abord vérifié si les groupes formés étaient en partie cohérents entre les 3 méthodes, c'est-à-dire si un groupe d'une méthode donnée partageait une grande majorité de ses points dans seulement 1 groupe de chaque autre méthode avec une faible dispersion dans d'autres groupes. Cependant, cela est rarement le cas : la plupart des groupes d'une méthode ont leurs points dispersés dans plusieurs groupes d'une autre méthode. Autrement dit, une grande diversité de combinaisons de groupes a été observée.

En effet, en listant **toutes les combinaisons possibles** des identifiants de groupe entre les 3 méthodes — en considérant les 907 points inclus dans les trois regroupements mais en ignorant les points associés aux groupes de rebuts provisoires (CB_XX et CB_X0) — 81 combinaisons sont obtenues, dont seulement 21 concernent au moins 10 points tandis que 26 sous-groupes ne concernent que 1 point. Un extrait de cette longue liste de combinaisons est exposé dans le tableau ci-dessous, limité pour l'exemple aux deux groupes sous influence proximale du Rhin (CB_6 et CB_7) d'après l'approche par corrélation entre chroniques (cf. Tableau 5). Un diagramme de Sankey a également été généré afin de visualiser plus efficacement la multiplicité et l'importance relative (fréquence) de ces combinaisons (Figure 15).

Ce premier exercice de comparaison globale des regroupements a permis de discerner une cohérence non pas dans le détail dans des groupes mais dans leur relation au Rhin. En effet, en examinant la liste de combinaisons (Tableau 5) ou le diagramme de Sankey (Figure 15) ci-dessous, une bonne cohérence d'ensemble entre le grand groupe CB_7 ('RR_Aothers') établi par

corrélation, et les groupes DF_1 à DF_3 d'une part, et les groupes HI_2 et HI_5 d'autre part peut être constatée ; tous des groupes de points interprétés comme influencés par le Rhin.

C'est sur la base de ces premiers constats qu'il a été décidé de commencer par regrouper les résultats des 3 méthodes selon une **classification binaire : soit « Rhin », soit « Non Rhin »**. Une **règle de majorité** hybride a été appliquée pour déterminer la classe binaire à attribuer à chaque combinaison. A titre de rappel (depuis les résultats des différentes méthodes) :

- Les groupes liés au Rhin sont : CB_6 et CB_7 ; DF_1 à DF_3 ; HI_2 et HI_5.
- Les groupes non liés au Rhin sont : CB_1 à CB_5 ; DF_4 à DF_6 ; HI_1, HI_3 et HI_4.
- Les groupes sans lien établi à ce stade : CB_X0 et CB_XX.

Si une combinaison est composée d'une majorité de groupes interprétés comme ayant des dynamiques influencées par le Rhin, alors la classe **« Rhin »** lui est attribuée. De même, si une majorité des groupes composant la combinaison sont considérés non influencés par le Rhin, alors la combinaison est classée en **« Non Rhin »**. Différemment, si la combinaison est composée de groupes n'offrant pas de cohérence (en termes d'influence ou non-influence par le Rhin), **une classe « indéterminée » (« ? »)** (nommée 'UNSURE-DISCARD-CHECK' dans les traitements) lui est attribuée à ce stade.

Le bilan de cette étape de classification binaire des résultats combinés est de :

- 749 points (chroniques piézométriques) classés → « Rhin » ;
- 804 points classés → « Non Rhin » ;
- et 69 points indéterminés → « ? ». Remarques : i) Les chroniques de cette classe « ? » sont examinées plus tard, une à la fois à dire d'expert, pour tenter de les réintégrer dans un des groupes de la synthèse. ii) Les combinaisons pour lesquelles le point n'a pas été retenu par toutes les méthodes de regroupement sont quand même classées selon la règle de majorité (ex. 2/2 ou 1/1 méthode(s) ayant attribué à ce point un groupe influencé par le Rhin → classe binaire « influence significative Rhin ; mais 1/2 groupes… → classe indéterminée « ? »).

Tableau 5 : Extrait de la liste des combinaisons de groupes des 3 méthodes pour les groupes CB_6 et CB_7 sous influence du Rhin d'après les résultats du regroupement par corrélation

Identifiant de la combinaison (pseudo-groupe)	Nb points total pour la combinaison	Nb méthodes suggérant une influence du Rhin
CB_6.DF_2.HI_1	1	2/3
CB_6.DF_2.HI_2	1	3/3
CB_6.DF_2.HI_5	7	3/3
CB_6.DF_3.HI_1	1	2/3
CB_6.DF_3.HI_2	5	3/3
CB_6.DF_3.HI_5	4	3/3
CB_6.DF_4.HI_1	2	1/3
CB_6.DF_4.HI_5	5	2/3
CB_6.DF_5.HI_1	3	1/3
CB_6.DF_5.HI_2	1	2/3
CB_6.DF_5.HI_5	2	2/3
CB_7.DF_1.HI_1	4	2/3
CB_7.DF_1.HI_2	19	3/3
CB_7.DF_1.HI_5	80	3/3
CB_7.DF_2.HI_1	1	1/3

CB_7.DF_2.HI_2	107	3/3
CB_7.DF_2.HI_5	63	3/3
CB_7.DF_3.HI_1	1	1/3
CB_7.DF_3.HI_2	32	3/3
CB_7.DF_3.HI_5	26	3/3
CB_7.DF_4.HI_1	1	1/3
CB_7.DF_4.HI_2	3	2/3
CB_7.DF_4.HI_5	19	2/3
CB_7.DF_5.HI_1	6	1/3
CB_7.DF_5.HI_2	1	2/3
CB_7.DF_5.HI_4	3	1/3
CB_7.DF_5.HI_5	19	2/3
CB_7.DF_6.HI_5	5	2/3

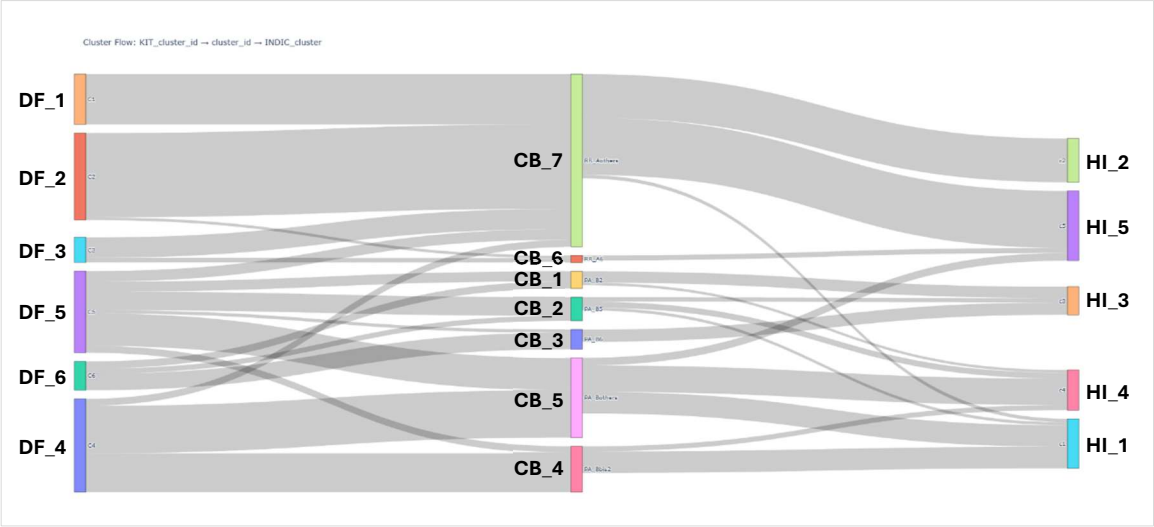


Figure 15 Diagramme de Sankey : croisement entre les trois clusterings et leurs groupes distincts

Une fois cette classification binaire très générale en 2 grandes familles « Rhin » vs « Non-Rhin » effectuée, une suite de modifications a été définie et programmée progressivement, à dire d'expert : tantôt en formulant une règle pour traduire certaines combinaisons ciblées en un groupe plus précis dans l'une des 2 grandes familles ; tantôt en définissant des cas particuliers (liste de points auxquels attribuer tel groupe destination). Seules quelques-unes de ces modifications ont pu être conçues en se basant sur le bilan des combinaisons de groupes fourni par un tableau de fréquence (cf. *« Nb points total pour la combinaison »*,Tableau 5) ou par le diagramme de Sankey (Figure 15Figure 15). La plupart des modifications ont été définies « à dire d'experts », en explorant le regroupement de cette synthèse au fil de sa construction, dans un logiciel de SIG permettant à la fois de localiser les points, de voir le groupe attribué à chacun d'eux et de consulter une vue graphique montrant en premier plan la série piézométrique du point avec en arrière-plan toutes les autres séries du groupe, dont la série médoïde.

A titre d'exemple, voici deux séries piézométriques allemandes dont le groupe attribué a été modifié, précisé, lors de ce travail itératif (Figure 16). Le groupe initialement attribué était 'RHINE' pour ces deux points, indiquant leur appartenance présumée à la grande famille « Rhin ». Comme la chronique du piézomètre 108/064-3 se révèle très similaire au médoïde du cluster **CB_6**

(‘RR_A6’) établi précédemment avec l’approche par corrélation, elle a donc été intégrée à un groupe nommé « RHINE-A6-NEAR-KEHL » dans cette synthèse. La chronique du piézomètre 810/066-0, montre en revanche une évolution singulière de ses niveaux piézométriques, en particulier depuis la fin des années 2010 (on ne retrouve pas de signal similaire dans les points situés à proximité). Ce piézomètre s’avère finalement non corrélé à l’ensemble des médoïdes des groupes progressivement définis dans cette synthèse. D’où son intégration dans un groupe « SINGULAR ».

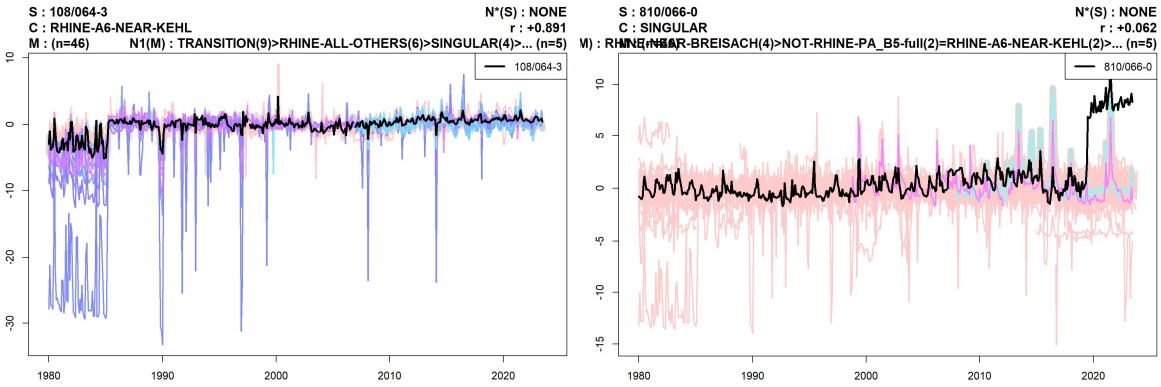


Figure 16 Exemples de deux chroniques piézométriques (normalisées) dont le groupe attribué a été modifié, précisé, au fil de l’exploration SIG et pendant la construction de la synthèse des résultats, (La courbe noire en avant-plan = le médoïde du groupe. Les autres courbes, en arrière-plan, ont des couleurs aléatoires ; abréviations : cf. Annexe 150)

Les conditions définies pour modifier les « groupes de synthèse » attribués des séries jusqu’à aboutir au regroupement final sont détaillées dans le **Erreur ! Source du renvoi introuvable.** ci-dessous. Ce tableau relie le **Groupe initial**ement attribué à chaque série selon la classification binaire « Rhin » (‘RHINE’) / « Non-Rhin » (‘NOT-RHINE’) / « ? » (‘UNSURE…’) ; au **Groupe final**ement attribué dans cette synthèse (C1 à C9) ; en détaillant les différents **Groupes par corrélation** (**CB_…)** obtenus plus tôt. La fréquence de chaque combinaison (colonne « N ») ainsi qu’un indicateur de conservation de la correspondance entre le groupe par corrélation et le groupe final (colonne « Final = **CB ?** ») y figurent aussi. Ce tableau permet de constater que :

- 738 (soit 95 %) des 777 points initialement classés ‘**NOT-RHINE**’ ont un groupe final qui conserve l’explication attribuée précédemment à ces points par corrélation. Ces points « Non-Rhin » cohérents avec les corrélations, aboutissent dans 5 groupes finaux distincts.
- 639 (soit 90 %) des 711 points initialement classés ‘**RHINE**’ ont un groupe final qui conserve l’explication attribuée précédemment à ces points par corrélation. Ces points « Rhin » cohérents avec les corrélations, aboutissent dans 3 groupes finaux distincts.
- Sur les 40 points initialement classés ‘**UNSURE**…’ en raison d’un flou ou d’incohérences constatées dans les résultats des trois méthodes de regroupement : 18 aboutissent dans des groupes sous influence proximale du Rhin (‘RHINE-…’) ; 15 intègrent des groupes hors influence notable du Rhin (‘NOT-RHINE-…’) ; et 7 sont considérés évoquer une dynamique de ‘TRANSITION’ entre « Rhin » et « Non-Rhin ».
- On note seulement une petite minorité de cas où le point est passé d’une grande famille à l’autre entre la classification binaire initiale et le regroupement final : 26 (soit 3 %) des points d’abord classés ‘NOT-RHINE’ sont finalement placés dans des groupes ‘RHINE-…’ et 18 points initialement classés ‘RHINE’ (soit <1 %) sont finalement placés soit dans des groupes ‘NOT-RHINE-…’ (4) ou dans le groupe de ‘TRANSITION’ (14).

- Le groupe final 'RHINE-NEAR-BREISACH', constitué pendant ce travail de synthèse suite au constat d'un signal spécifique discernable concentré dans une zone proche du Rhin, a été créé à partir de points issus majoritairement des groupes **CB_2** ('PA_B5') et **CB_7** ('RR_Aothers'). Cela rappelle qu'il peut y avoir des similitudes entre séries de différents groupes, surtout parmi un sous-ensemble de points voisins.
- La cohérence globale (équivalence) entre les groupes finaux principaux et ceux proposés plus tôt par corrélation est très élevée : 88 % (1337 / 1528). Un constat prévisible puisque le regroupement construit lors de cette synthèse a été largement basé sur les résultats de l'approche de regroupement par corrélation.

Quelques remarques liées à ce tableau :

- Pour limiter la longueur de ce tableau, les 94 points finalement classés dans les groupes 'ANOMALOUS', 'SINGULAR', 'TOO-SHORT-or-MISSING' ou 'DISCARD', y ont été ignorés.
- De même, les groupes obtenus plus tôt par caractéristiques dynamiques (DF_...) ou par indicateurs hydrodynamiques (HI_...) n'ont pas été considérés dans la préparation de ce tableau, car autrement il aurait atteint >200 lignes, le rendant difficilement lisible.
- A titre informatif, les libellés de groupes écrits en rouge mettent en évidence les cas incohérents par rapport au groupe finalement attribué dans la synthèse.

Tableau 6 : Tableau de liaison entre le Groupe initial de chaque série et le Groupe finalement attribué dans cette synthèse (C1 à C9), en détaillant si oui ou non, les binômes de groupes sont corrélés (9 binômes= « TRUE »)

Groupe initial	Groupe final	#	G. par corrélation	#	N	Final = CB ?
NOT-RHINE	NOT-RHINE-ALL-OTHERS	C1	PA_Bothers	CB_5	**303**	TRUE
NOT-RHINE	NOT-RHINE-ALL-OTHERS	C1	PA_Bbis2	CB_4	**184**	TRUE
NOT-RHINE	NOT-RHINE-ALL-OTHERS	C1	MISC_in_betweens	CB_X0	1	FALSE
NOT-RHINE	NOT-RHINE-ALL-OTHERS	C1	RR_Aothers	CB_7	1	FALSE
NOT-RHINE	NOT-RHINE-PA_**B2**	C2	PA_**B2**	CB_1	**77**	TRUE
NOT-RHINE	NOT-RHINE-PA_B2	C2	PA_Bothers	CB_5	2	FALSE
NOT-RHINE	NOT-RHINE-PA_B2	C2	RR_Aothers	CB_7	1	FALSE
NOT-RHINE	NOT-RHINE-PA_**B5**-full	C3	PA_**B5**	CB_2	**76**	TRUE
NOT-RHINE	NOT-RHINE-PA_B5-full	C3	PA_Bothers	CB_5	3	FALSE
NOT-RHINE	NOT-RHINE-PA_B5-full	C3	MISC_in_betweens	CB_X0	1	FALSE
NOT-RHINE	NOT-RHINE-PA_**B6**	C4	PA_**B6**	CB_3	**84**	TRUE
NOT-RHINE	NOT-RHINE-PA_B6	C4	PA_B5	CB_2	4	FALSE
NOT-RHINE	RHINE-A6-NEAR-KEHL	C5	RR_A6	CB_6	5	FALSE
NOT-RHINE	RHINE-A6-NEAR-KEHL	C5	PA_Bbis2	CB_4	1	FALSE
NOT-RHINE	RHINE-A6-NEAR-KEHL	C5	PA_Bothers	CB_5	1	FALSE
NOT-RHINE	RHINE-ALL-OTHERS	C6	MISC_in_betweens	CB_X0	2	FALSE
NOT-RHINE	RHINE-ALL-OTHERS	C6	MISC_outliers	CB_XX	1	FALSE
NOT-RHINE	RHINE-ALL-OTHERS	C6	RR_Aothers	CB_7	1	FALSE
NOT-RHINE	RHINE-NEAR-BREISACH	C7	PA_B5	CB_2	14	FALSE
NOT-RHINE	RHINE-NEAR-BREISACH	C7	RR_Aothers	CB_7	1	FALSE
NOT-RHINE	TRANSITION	C9	RR_Aothers	CB_7	5	FALSE
NOT-RHINE	**TRANSITION**	C9	MISC_**in_betweens**	CB_X0	4	TRUE
NOT-RHINE	TRANSITION	C9	PA_Bothers	CB_5	4	FALSE

Groupe initial	Groupe final	#	G. par corrélation	#	N	Final = CB ?
NOT-RHINE	TRANSITION	C9	PA_Bbis2	CB_4	1	FALSE
RHINE	NOT-RHINE-ALL-OTHERS	C1	RR_Aothers	CB_7	2	FALSE
RHINE	NOT-RHINE-PA_B5-full	C3	RR_Aothers	CB_7	1	FALSE
RHINE	NOT-RHINE-PA_B6	C4	PA_B6	CB_3	1	FALSE
RHINE	RHINE-**A6**-NEAR-KEHL	C5	RR_**A6**	CB_6	**29**	TRUE
RHINE	RHINE-A6-NEAR-KEHL	C5	RR_Aothers	CB_7	7	FALSE
RHINE	RHINE-A6-NEAR-KEHL	C5	MISC_outliers	CB_XX	1	FALSE
RHINE	RHINE-ALL-**OTHERS**	C6	RR_A**others**	CB_7	**579**	TRUE
RHINE	RHINE-ALL-OTHERS	C6	MISC_in_betweens	CB_X0	44	FALSE
RHINE	RHINE-ALL-OTHERS	C6	RR_A6	CB_6	5	FALSE
RHINE	RHINE-ALL-OTHERS	C6	PA_Bothers	CB_5	4	FALSE
RHINE	RHINE-ALL-OTHERS	C6	PA_B5	CB_2	2	FALSE
RHINE	RHINE-NEAR-BREISACH	C7	RR_Aothers	CB_7	19	FALSE
RHINE	RHINE-NEAR-BREISACH	C7	PA_B5	CB_2	1	FALSE
RHINE	RHINE-NEAR-BREISACH	C7	PA_Bothers	CB_5	1	FALSE
RHINE	RHINE-NEAR-BREISACH	C7	MISC_in_betweens	CB_X0	1	FALSE
RHINE	TRANSITION	C9	RR_Aothers	CB_7	14	FALSE
UNSURE…	NOT-RHINE-ALL-OTHERS	C1	RR_Aothers	CB_7	1	FALSE
UNSURE…	NOT-RHINE-ALL-OTHERS	C1	PA_Bothers	CB_5	1	FALSE
UNSURE…	NOT-RHINE-PA_B2	C2	PA_B2	CB_1	1	FALSE
UNSURE…	NOT-RHINE-PA_B5-full	C3	PA_B5	CB_2	6	FALSE
UNSURE…	NOT-RHINE-PA_B6	C4	PA_B6	CB_3	6	FALSE
UNSURE…	RHINE-A6-NEAR-KEHL	C5	RR_Aothers	CB_7	1	FALSE
UNSURE…	RHINE-A6-NEAR-KEHL	C5	RR_A6	CB_6	1	FALSE
UNSURE…	RHINE-ALL-OTHERS	C6	SKIPPED		10	FALSE
UNSURE…	RHINE-ALL-OTHERS	C6	RR_Aothers	CB_7	3	FALSE
UNSURE…	RHINE-ALL-OTHERS	C6	MISC_in_betweens	CB_X0	1	FALSE
UNSURE…	RHINE-NEAR-BREISACH	C7	PA_B5	CB_2	2	FALSE
UNSURE…	TRANSITION	C9	RR_Aothers	CB_7	3	FALSE
UNSURE…	TRANSITION	C9	PA_Bbis2	CB_4	2	FALSE
UNSURE…	**TRANSITION**	C9	MISC_**in_betweens**	CB_X0	1	TRUE
UNSURE…	TRANSITION	C9	RR_A6	CB_6	1	FALSE

5.2 Résultats détaillés

Le regroupement final construit par cette synthèse est ainsi constitué de :

- 9 groupes principaux (C1 à C9) dont 4 groupes « Non-Rhin » (C1 à C4), 3 groupes « Rhin » (C5 à C7), un groupe de transition (C9) et un groupe de dynamiques singulières (C8) ;
- 3 groupes secondaires (**CX**…) que l'on peut généralement ignorer dans la plupart des cas d'utilisation : des séries anormales (CXa), trop courtes (CXs) ou toujours à exclure (CXd) ;

Ainsi, parmi les 1622 séries piézométriques initialement retenus pour la méthode corrélation entre chroniques, 1574 (97 %) peuvent être classées dans les 9 groupes principaux.

Les graphiques illustrant la composition de chacun des groupes de ce regroupement final sont présentés à l'ANNEXE 1 : Composition des groupes finaux.

Le Tableau 7 ci-dessous décrit la cohérence et l'homogénéité du contenu de chaque groupe final par des statistiques calculées à partir des coefficients de corrélation (r) entre les N séries du groupe et la série médoïde (de référence) du groupe. En complément des constats faits plus tôt avec le regroupement par corrélation, on note ici que :

- La plupart des groupes principaux (**C1 à C5, C7 et C9**) sont composés de séries qui sont très majoritairement dans leur « meilleur » groupe d'après les corrélations aux médoïdes. Ces groupes principaux montrent toutefois une part importante (40-50 %) de séries dont la corrélation au médoïde n'est pas élevée, ce qui souligne une variabilité interne.
- Le groupe C6 ('RHINE-ALL-OTHERS') montre un moins bon indice de cohérence mais cela est attendu d'un groupe aussi gros et hétérogène puisqu'il rassemble toutes les « autres » dynamiques estimées influencés par le Rhin. Les raisons pour lesquelles ce groupe n'a pas été subdivisé en petits groupes plus homogènes ont été abordés plus tôt (souhait de pouvoir expliquer chaque groupe). On peut aussi ajouter ici que plus un groupe est gros et diversifié, moins la série retenue comme médoïde pour le groupe est susceptible de très bien corréler avec une majorité des séries membres dudit groupe. D'où le peu de séries (12.6 %) fortement corrélées au médoïde du groupe (C6).
- De même, dans le cas du groupe **C8** ('SINGULAR'), l'hétérogénéité du groupe et le peu de corrélations élevées au médoïde sont des résultats attendus.
- Sans surprise, les 3 groupes secondaires de séries devant généralement être ignorées (**CXa, CXd et CXs**) sont très hétérogènes. Les très mauvaises corrélations au médoïde choisi pour chacun de ces groupes, mettent en évidence leur nature disparate.
  Une majorité des séries placées dans ces groupes secondaires auraient « préféré » aller dans d'autres groupes (% r best ~ 10-15 % dans CXa et CXs). Cela rappelle que la mise à l'écart de ces séries a été faite manuellement (lors d'itérations d'expertise des résultats) et que sans ces interventions manuelles, les groupes finaux principaux auraient été dégradés par ces séries anormales ou trop courtes (qui auraient certes été « mieux » corrélées à ces groupes, mais sans que le choix du groupe soit suffisamment robuste autant en termes de corrélation (r) que de contexte hydrogéologique).

Tableau 7 : Tableau indiquant pour chaque groupe la cohérence et l'homogénéité de son contenu par des statistiques calculées à partir des coefficients de corrélation (r) entre les N séries du groupe et la série médoïde (de référence) du groupe

Group	% r best (= max r)	% r high (r>0.8)	% r high & best	% r low (r<0.8)	% r very low (r<0.5)	% r too bad (r<0.25)	N points TOTAL	N points r very low
C1	84.4	51.1	48.3	48.9	3.0	0.0	493	15
C2	91.4	60.5	59.3	39.5	4.9	0.0	81	4
C3	83.9	59.8	55.2	40.2	2.3	0.0	87	2
C4	88.4	51.6	50.5	48.4	3.2	0.0	95	3
C5	84.8	56.5	56.5	43.5	0.0	0.0	46	0
C6	20.9	12.6	8.9	87.4	37.9	6.6	652	247

C7	61.5	38.5	38.5	61.5	17.9	0.0	39	7
C8	23.9	4.3	4.3	95.7	69.6	37.0	46	32
C9	65.7	25.7	25.7	74.3	20.0	0.0	35	7
CXa	9.1	9.1	9.1	90.9	81.8	63.6	11	9
CXd	50.0	50.0	50.0	50.0	0.0	0.0	2	0
CXs	11.4	2.9	2.9	97.1	77.1	54.3	35	27
Globally:	54.6	33.2	30.5	66.8	21.8	5.3	1622	353

En complément, le Tableau 8 ci-dessous présente les résultats d'une analyse du « 1er voisin » (N1) déterminé pour chaque série lors du post-traitement du regroupement final. Le « 1er voisin » est le groupe (autre que celui retenu) avec lequel la série se corrèle le mieux. Dans la majorité des cas (54.6 %) le groupe attribué à la série est celui offrant la meilleure corrélation (cf. Tableau 7) et le 1er voisin correspond alors au 2e meilleur choix de groupe en termes de coefficient de corrélation r. Mais il y aussi des cas où le 1er voisin offrirait une meilleure corrélation (r + élevé) que le groupe finalement retenu. Dans tous les cas, examiner le 1er voisin le plus fréquent d'un groupe est intéressant et renseigne efficacement sur le degré de ressemblance entre les groupes. On constate ainsi notamment que :

- Une majorité (66 %) des séries du groupe **C1** sont proches du groupe C9 (son voisin le plus fréquent), évoquant une ressemblance tout à fait logique entre 'NOT-RHINE-ALL-OTHERS' et 'TRANSITION'. Réciproquement, près d'un tiers (31%) des séries du groupe **C9** sont proches du groupe C1.
- Les groupes **C2, C3 et C4** se ressemblent, d'après les liens de voisinage recensés ici. Le groupe C3 a 2 voisins de fréquences égales (C1 ou C4) dont C1 rappelle que ce groupe C3 montre une dynamique moins inertielle qui peut effectivement se confondre avec celle du médoïde du gros groupe C1 ('NOT-RHINE-PA_B5-full' ~ 'NOT-RHINE-ALL-OTHERS').
- Le groupe **C6** ('RHINE-ALL-OTHERS') a comme principal voisin le groupe **C8** ('SINGULAR'), un constat peu signifiant qui indique seulement que la série choisie automatiquement comme médoïde pour représenter C8 est série relativement bien corrélée (r > 0.75) aux groupes « Rhin » C6 ou **C7**.
- Les groupes C5 et C8, de tailles modestes, n'ont pas de 1er voisin assez fréquent (<20 %).

Tableau 8 : Tableau indiquant pour chaque groupe les résultats d'une analyse du « 1er voisin » (N1) déterminé pour chaque série lors du post-traitement du regroupement final

Group full name	Group	1st Neighbor (N1)	% N1	N N1	N points TOTAL
NOT-RHINE-ALL-OTHERS	C1	C9	66.33	327	493
NOT-RHINE-PA_B2	C2	C4	41.98	34	81
NOT-RHINE-PA_B5-full	C3	C1 ou C4	34.48	30	87
NOT-RHINE-PA_B6	C4	C3	52.63	50	95
RHINE-A6-NEAR-KEHL	C5	C9	19.57	9	46
RHINE-ALL-OTHERS	C6	C8	36.81	240	652
RHINE-NEAR-BREISACH	C7	C3	38.46	15	39
SINGULAR	C8	C7	10.87	5	46
TRANSITION	C9	C1	31.43	11	35

5.3 Carte des secteurs et interprétation

La **carte** en page suivante (Figure 17 et Figure 18) présente une vue cartographique des résultats de ce travail de synthèse en un « regroupement final ». On y retrouve la plupart des groupes formés plus tôt par corrélation. La **légende** qui accompagne la carte décrit chaque « groupe final » avec l'explication hydrogéologique trouvée pour chacun d'eux (cf. Tableau 9). Cette légende ressemble donc à celle présentée plus tôt pour le regroupement par corrélation.

Par ailleurs, des analyses statistiques d'indicateurs caractérisant les chroniques piézométriques ou le contexte des points de suivi, sont présentées en ANNEXE 2 : Analyses statistiques d'indicateurs du regroupement final (synthèse), à titre complémentaire.

Tableau 9 : Caractérisation et description des neuf principaux groupes (et trois groupes secondaires de rebut)

N° groupe	Caractérisation
C1 (NOT-RHINE-ALL-OTHERS) n = 493	Groupe rassemblant les **autres piézomètres de la Plaine d'Alsace** hors influence significative du Rhin. Faibles épaisseurs de ZNS <5 mètres pour ~85 % des points du groupe. Presqu'aucun point ne se trouve à proximité du Rhin (<1–2 km de part et d'autre du Rhin). Ce groupe contient la majorité des points du groupe CB_4 (soit environ 35% des points de C1), qui apparaissent souvent impactés par les prélèvements estivaux.
C2 (NOT-RHINE-PA_B2) n = 81	Groupe de **piézomètres inertiels** *principalement* concentrés au Nord de la zone d'étude à l'est du Rhin (environs de Rastatt – Karlsruhe) **avec une épaisse zone non saturée (ZNS) presque partout >5 mètres**, d'où les importantes composantes pluriannuelles dans leur dynamique. Explication appuyée par des indicateurs de temps de demi-décroissance (de vidange de l'aquifère) longs ainsi que par des **temps d'arrivée des précipitations importants** aussi.
C3 (NOT-RHINE-PA_B5-full) n = 87	Points localisés majoritairement (pour les 2/3 environ) au Sud d'une ligne W-E entre Sélestat et Lahr/Schwarzwald. Dont environ la moitié des points concentrés dans une zone relativement étroite de la rive droite du Rhin entre Vieux-Brisach (Breisach am Rhein) et Bad Krozingen**. L'inertie (relativement importante mais sans délai notable par rapport aux pluies)** sembler joue un rôle important dans l'établissement de ce cluster. **ZNS là aussi souvent >5** mètres. (Voir aussi le groupe 'C7', ci-dessous)
C4 (NOT-RHINE-PA_B6) n = 95	Points situés dans le Haut-Rhin, concentrés au Sud d'une ligne entre Colmar et Fribourg-en-Brisgau. Groupe caractérisé par une **ZNS encore plus épaisse** en général (épaisseur médiane de la **ZNS >10 mètres**) avec un comportement plus inertiel. Alimentation de l'aquifère par **Sundgau**. Très bonne cohérence avec les **longs délais estimés d'arrivée des précipitations**.
C5 (RHINE-A6-NEAR-KEHL) n = 46	Points **fortement influencés par le Rhin**, plus précisément **impactés par le barrage agricole de Kehl-Strasbourg** au sud-est de la ville. Points concentrés **à l'amont** du barrage, sur la rive droite du Rhin seulement (imperméabilisation anthropique de la rive gauche coté Strasbourg). Evolution temporelle (signature) très particulière de la piézométrie caractérisée par des **niveaux nettement plus bas avant le milieu des années 1980** (hausse soudaine des niveaux vers 1985).

C6 (RHINE-ALL-OTHERS) n = 652	Groupe rassemblant les **autres points sous forte influence du Rhin**, car sans explication forte pour distinguer des sous-groupes en son sein. Evolutions piézométriques visuellement similaires dans l'ensemble. Malgré cela, seulement une moitié des points du groupe environ ont une piézométrie bien corrélée (r > +0.6) avec le débit du Rhin ; l'autre moitié pourrait être quand même influencée par le Rhin mais en des zones où ses débits seraient davantage artificiels (effets de seuil en amont des infrastructures hydroélectriques) [2]. Très **faibles épaisseurs de ZNS <5 mètres** pour ~90 % des points du groupe, et même **<2 mètres** pour ~2/3 des points du groupe. Remarque : Il serait possible, techniquement, d'affiner le découpage de ce groupe, mais cela n'apparaît pas particulièrement utile d'un point de vue utilitaire pratique, ces points étant de toute façon très influencés par le Rhin et de ses aménagements.
C7 (RHINE-NEAR-BREISACH) n = 39	Groupe de piézomètres aux comportements relativement homogènes, concentrés dans une petite zone d'environ 40 km^2 sur la rive droite du Rhin, **près de Vieux-Brisach** (Breisach am Rhein). **Faibles épaisseurs de ZNS presque partout <6 mètres**. **Dynamique plus réactive** (moins inertielle) que son groupe voisin 'C3', **sans décalage notable de la réponse aux pluies**.
C8 (SINGULAR) n = 46	Points dont la chronique montre une **évolution piézométrique très singulière voire anormale**. Ce groupe permet d'écarter des séries trop peu corrélées aux médoïdes des groupes principaux CB_1 à CB_7, avec une **évolution trop rare dans le jeu de données** pour qu'elle ait mené à la formation d'un cluster dédié ; et des **séries cassées par une rupture** (changement important et soudain) dans l'évolution de leurs niveaux (probablement dû à des erreurs lors du calcul des cotes piézométriques à partir des données de profondeur d'eau). Remarque : Cette liste de points jugés « aberrants » à ce stade est révisée plus tard, lors de la « Synthèse des résultats ».
C9 (TRANSITION) n = 35	Points retirés des groupes principaux à cause d'une **incohérence spatiale entre leur localisation et leur cluster** initialement attribué (lors de la première itération de clustering par corrélation établissant les deux grands groupes Rhin versus Non-Rhin) : soit le point était placé dans un des clusters Rhin alors qu'il était éloigné du Rhin ; soit il était placé dans un des clusters Non-Rhin tout en étant très proche du Rhin. Ces points ne sont **pas définitivement écartés**, mais plutôt mis de côté, pour une éventuelle réintégration dans les groupes principaux lors de la phase à suivre de « Synthèse des résultats ».
CXa (ANOMALOUS) n = 11	Chroniques jugées « aberrantes », « anormales » ou en tout cas **inexploitables par ces analyses**. Ces chroniques peuvent être marquées par un saut abrupt peu réaliste du niveau piézométrique entre 2 portions, ou par des amplitudes de variations aberrantes (excessives) par rapport à tout son voisinage, etc. Ces points demeurent intéressants à conserver quelque part, en tant que cas particuliers à investiguer éventuellement (pour identifier plus précisément les raisons de leur apparence aberrante).
CXd (DISCARD) n = 2	Chroniques pouvant être **écartées systématiquement, car il y a une autre chronique montrant la même évolution mais avec un suivi plus complet** dans un point très proche, voire au même point XY mais à un autre intervalle de profondeur (cas d'un piézomètre à multiples intervalles crépinés). Si ces 2 chroniques « à écarter » sont présentées ici malgré cela, c'est justement pour signaler que ces 2 'points' sont superflus dans le jeu de données (du LUBW).

CXs (TOO-SHORT-or-MISSING) n = 35	Chroniques qui ont été ajoutées lors des itérations de post-traitement des résultats, mais qui n'ont finalement pas pu être intégrées avec suffisamment de confiance aux groupes définis, car effectivement **trop courtes et/ou avec trop d'interruptions de leur suivi piézométrique** (périodes sans donnée). Ces chroniques n'ont pas montré de corrélation (ni visuelle ni statistique) assez claire et forte avec le groupe suggéré par l'algorithme de post-traitement, pour que ces points soient confirmés dans ces groupes. Ce groupe 'CXs' réunit ainsi des points dont les chroniques s'avèrent inexploitables, non en raison d'un signal aberrant mais plutôt ici parce que trop courtes ou trop lacunaires.

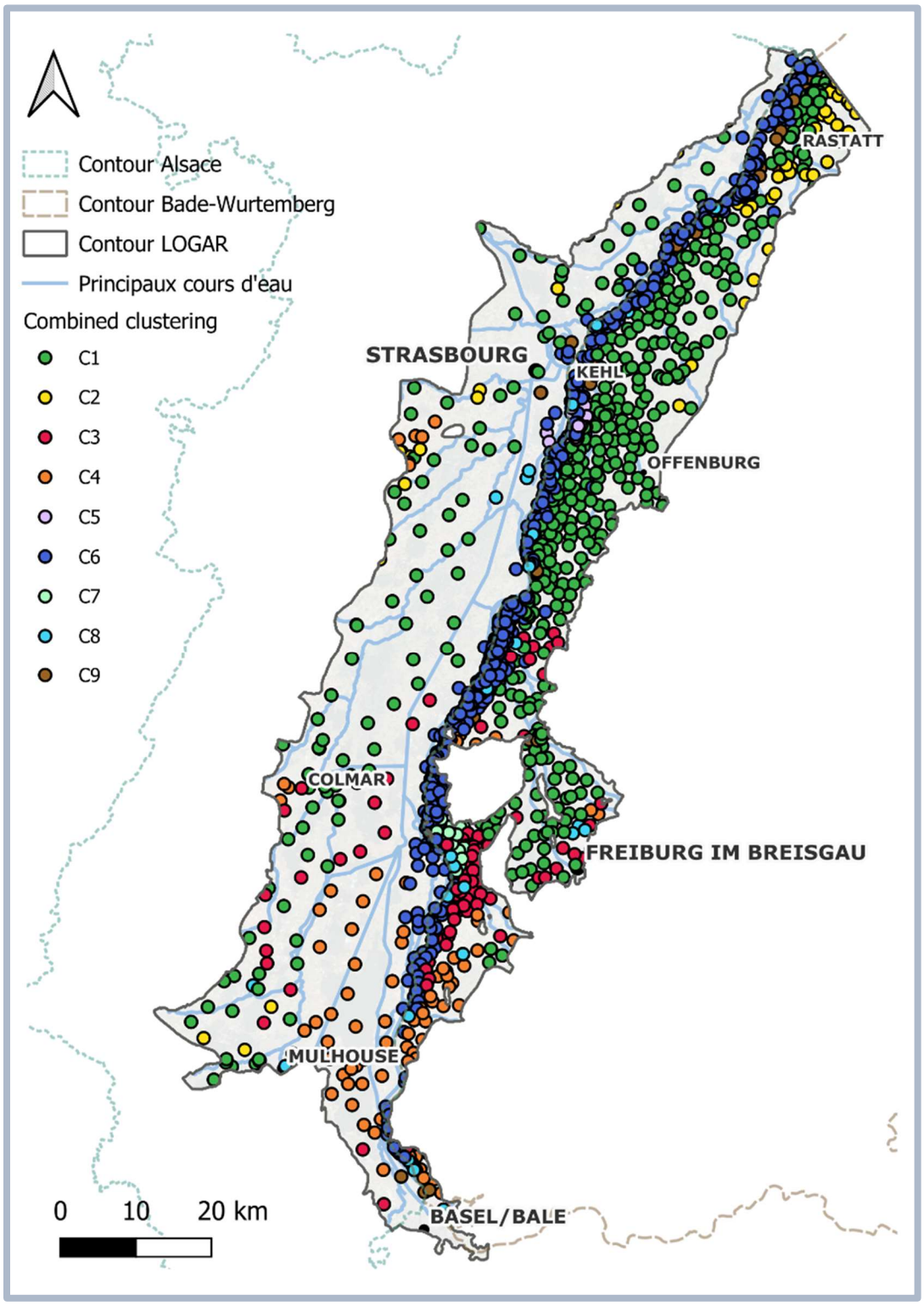


Figure 17 Carte de synthèse des résultats avec le regroupement final.

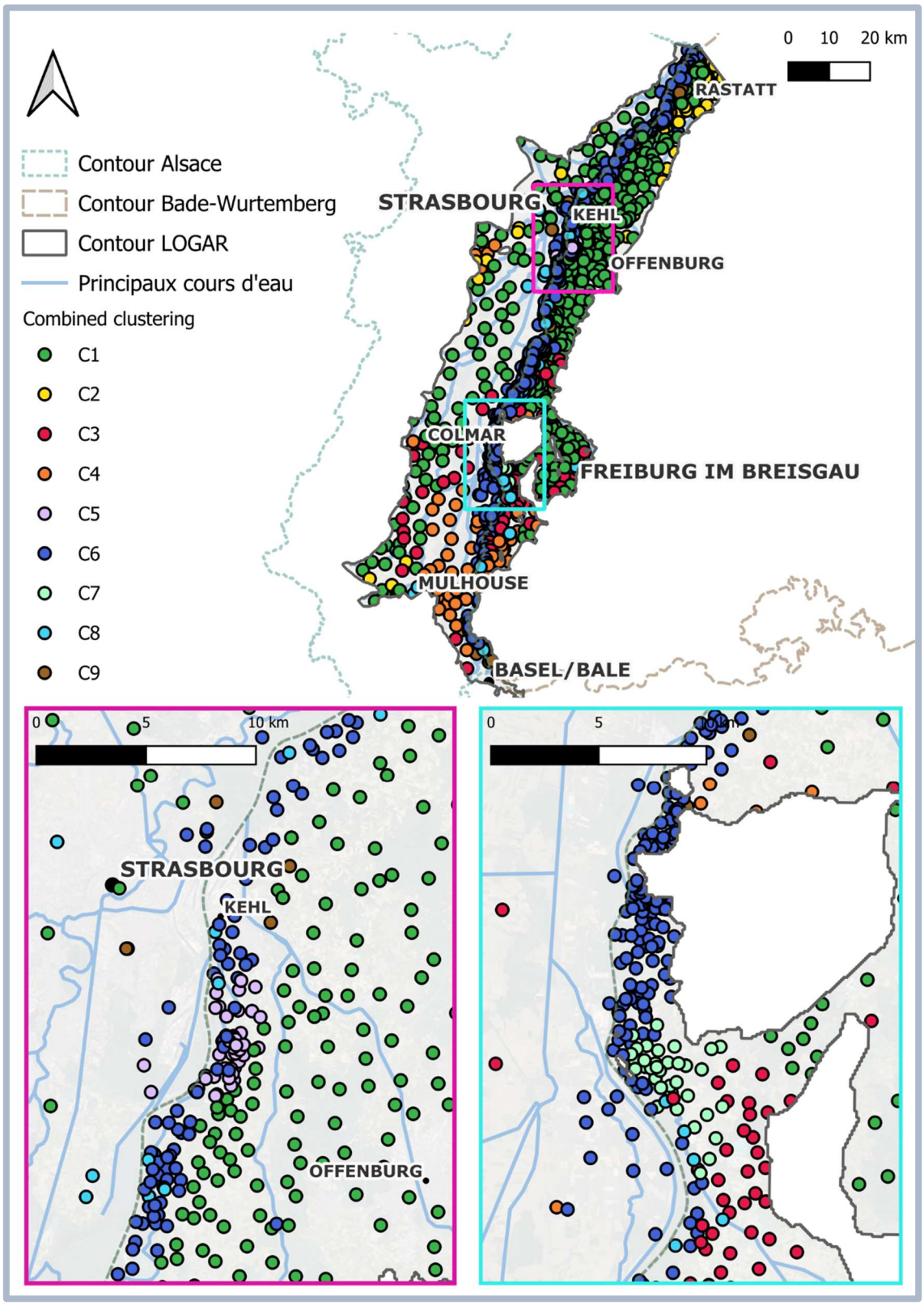
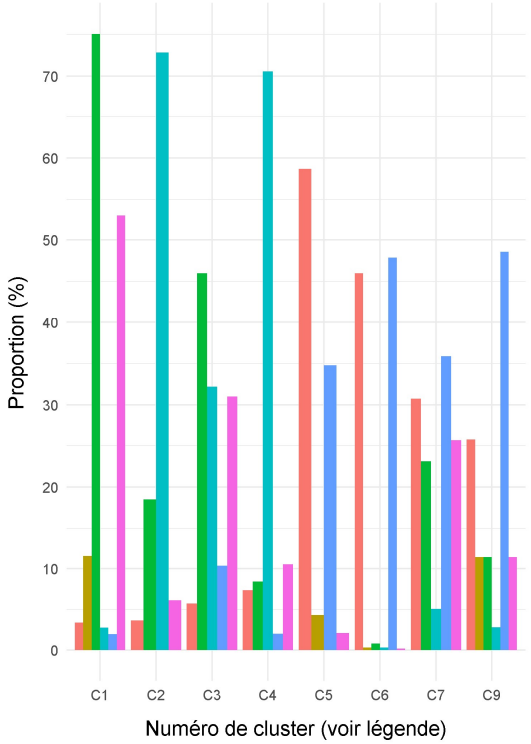


Figure 18 Carte de synthèse des résultats avec le regroupement final : zooms sur les secteurs de Kehl et Breisach

L'importance relative, dans chaque groupe, de l'influence des forçages climatiques (précipitations), du Rhin, ainsi que de l'influence des prélèvements estivaux, a également pu être analysée. La Figure 19 indique la répartition, par groupe, du forçage dominant de chaque point.



N°	ID du cluster	N.tot	N.NAs
C1	NOT-RHINE-ALL-OTHERS	493	25
C2	NOT-RHINE-PA_B2	81	4
C3	NOT-RHINE-PA_B5-full	87	5
C4	NOT-RHINE-PA_B6	95	11
C5	RHINE-A6-NEAR-KEHL	46	1
C6	RHINE-ALL-OTHERS	652	31
C7	RHINE-NEAR-BREISACH	39	2
C8	SINGULAR	46	2
C9	TRANSITION	35	0
CXa	ANOMALOUS	11	0
CXd	DISCARD	2	0
CXs	TOO-SHORT-or-MISSING	35	12

Figure 19 Résumé sur les forçages dominants par groupe : diagramme des proportions d'influence des 9 groupes retenus (C1-C9) selon le climat (précipitations), le Rhin ainsi que l'influence des prélèvements estivaux

5.4 Vers une délimitation géographique de secteurs…

L'objectif est de délimiter des secteurs géographiques avec les différents partenaires franco-allemands, à partir des résultats (groupes de points) obtenus par le travail de synthèse présenté ci-dessus. Ces travaux, en cours, permettent d'ores et déjà de délimiter plusieurs secteurs géographiques :

1) Des secteurs de forçages naturels homogènes ou de contextes particuliers :
   a. Des zones qui pourraient être délimitées par une importante épaisseur de la zone non saturée (ZNS) (cf. groupes C2, C3 et C4) ;
   b. Des zones géologiques / lithologiques particulières, ex. cônes de sédimentation en contrebas des Vosges (« cônes de déjections »), aussi avec une épaisse ZNS (cf. exemple avec le groupe C4 : Figure 20 ci-dessous) ;

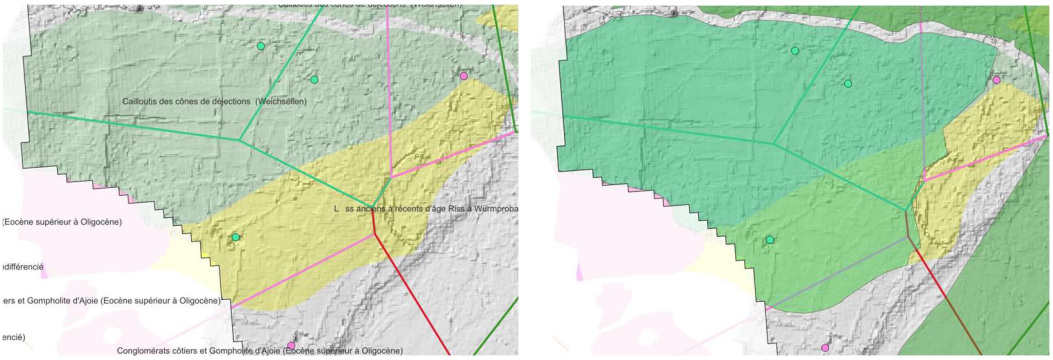


Figure 20 A gauche : Sous-groupe de piézomètres (points verts du groupe C4), leurs polygones de Voronoï (contours verts), et les contours de la carte géologique harmonisée au 1 : 50 000). A droite : Les contours géographiques dessinés (surface verte) à partir de ces trois informations délimitent une zone sous influence des « cailloutis des cônes de déjection » (avec une épaisse ZNS).

   c. Des repères topographiques, ex. la basse vallée du Rhin et ses zones fortement liées à la dynamique du fleuve (groupes C5, C6, C7), cf. l'exemple en Figure 21;

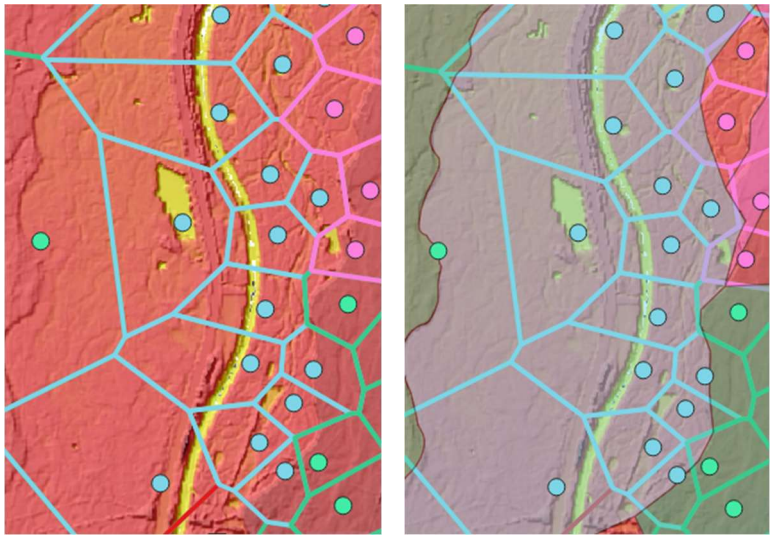


Figure 21 A gauche : Sous-groupe de piézomètres (points bleus du groupe C6), ses polygones de Voronoï (lignes bleues), l'épaisseur de la zone non saturée (rouge foncé = nappe profonde, jaune = nappe très proche de la surface voire cours/plan d'eau). A droite : Les contours géographiques dessinés (surface grise) à partir de ces trois informations délimitent une zone sous forte influence du Rhin avec une ZNS très peu épaisse.

d. Des secteurs de la Plaine d'Alsace hors influence du Rhin (sans corrélation significative avec ses débits) et de faibles épaisseurs de ZNS (cf. groupe C1) ;
e. Des zones avec un délai notable de la réponse aux pluies (réactivité/inertie) ;
f. etc.

2) Des secteurs affectés par des forçages anthropiques :
   a. Des zones situées en aval d'aménagements hydroélectriques du Rhin, soit dans la partie amont (cf. groupes C6 et C7) ;
   b. Des zones possiblement influencées par des prélèvements majoritairement estivaux, qui pourrait être affinées par suite des résultats dans les actions en cours (cf. groupes C1 et C3).

Ces travaux de définition de secteurs géographiques (délimitation par dessin de polygones à dire d'expert sur la base de la synthèse des résultats des clusterings et de plusieurs informations sur le contexte hydrogéologique des points) vont se poursuivre en fin 2025/2026.

6 CONCLUSION

Le travail réalisé dans le cadre de cette sous-action a permis l'identification de neuf groupes principaux dans l'aquifère rhénan, obtenus par une combinaison de trois méthodes de regroupement (basées sur des indicateurs ou sur la corrélation entre les chroniques piézométriques).

Ce travail a également permis de déterminer les types de forçages (influences perceptibles sur la piézométrie) qui caractérisent chacun des groupes identifiés.

De fait, le travail réalisé constitue une base solide pour l'exercice de délimitation géographique de secteurs hydrogéologiques dans la zone d'étude (*finalisation en fin 2025/2026*).

Le travail réalisé a mis en lumière les avantages relatifs de plusieurs méthodes différentes pour analyser puis regrouper des chroniques piézométriques. Les méthodes utilisées se différencient par leurs données d'entrée (chroniques entières, caractéristiques numériques ou indicateurs hydrodynamiques), leurs contraintes (induisant un nombre de chroniques utilisables *in fine* différent selon la méthode), et leur flexibilité (possibilité ou non de rajouter des points *a posteriori*). La nature différente des données utilisées pour procéder aux regroupements souligne le fait que chaque méthode peut être la plus adaptée en fonction de l'objectif recherché (regrouper les chroniques similaires ; celles qui partagent des influences communes ; ou encore, celles qui présentent le plus de variabilité). Dans une approche globale, un exercice de synthèse de l'ensemble des résultats a été conduit.

L'exercice de synthèse des résultats produits par plusieurs approches de regroupement est complexe et nécessite une hiérarchisation des méthodes employées. Une perspective pourrait être de développer une méthode de regroupement hybride combinant explicitement et simultanément des informations de type indicateurs et de type corrélations, avec une pondération qui pourrait être à ajuster en fonction des objectifs pratiques de l'exercice.

De plus, le choix méthodologique a été fait ici de définir de grands groupes de comportements relativement similaires, sans affiner (subdiviser) géographiquement en fonction de variations plus subtiles. Les secteurs (groupes de points) ainsi formés constituent des ensembles « globaux » qui pourraient être subdivisés en plusieurs sous-ensembles, en fonction de différences plus fines des évolutions/variations piézométriques, ou encore par le repérage de comportements particuliers d'échelles plus « locales ». Les secteurs « globaux » présentés ici constituent ainsi un socle solide pour d'éventuels exercices de définition de secteurs « locaux » (plus ciblés géographiquement, cf. chapitre 5.4), par exemple en vue de définir des périmètres de représentativité des modèles hydrogéologiques aux points qui sont développés par ailleurs dans cette étude GRETA.

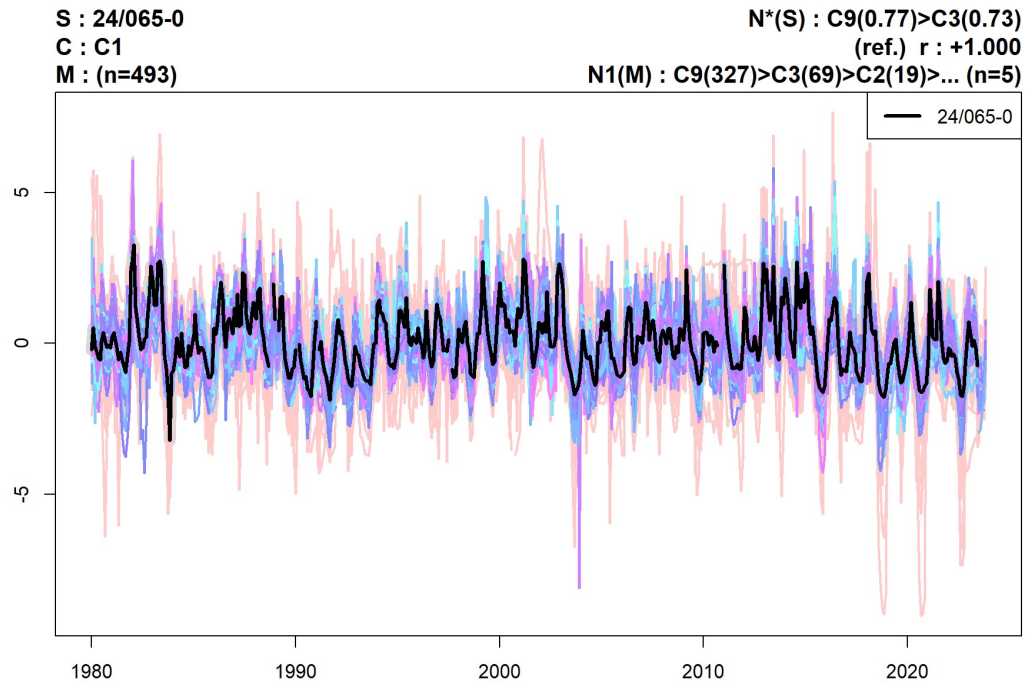
A l'heure de la rédaction de ce document, d'autres travaux sont encore en cours dans le cadre du projet GRETA, notamment les travaux d' « Analyse et d'interprétation sur chaque secteur des relations entre le niveau de la nappe, les rivières et les prélèvements dans l'aquifère rhénan » , ainsi que les travaux de modélisation maillée (LOGAR). Ces derniers pourront compléter cette étude de regroupement des chroniques piézométriques.

1. ANNEXE 1 : Composition des groupes finaux

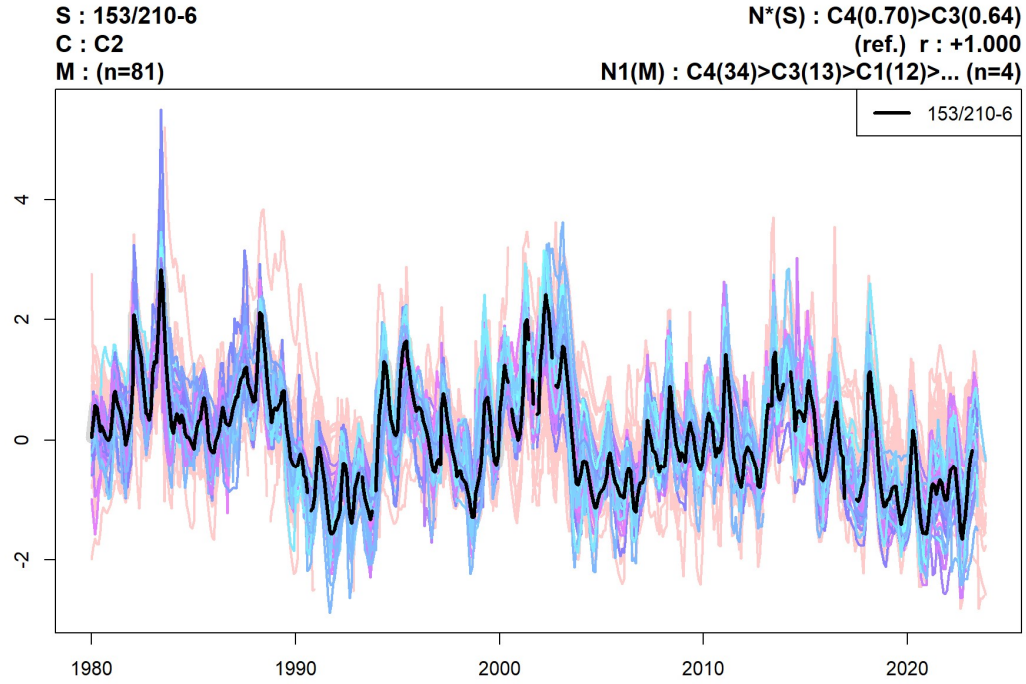
Ci-dessous, un graphique par groupe (*cluster*) montrant les courbes standardisées (centrées et réduites, d'où l'absence d'unité de l'axe Y et la position autour de Y = 0 des courbes) des séries piézométriques des n points composant le groupe. **La courbe noire en avant-plan = le médoïde du groupe. Les autres courbes, en arrière-plan, ont des couleurs aléatoires.** Explications des textes placés en haut de chaque graphique de ce type :

- **S** = identifiant de la série (ici = le médoïde du groupe) ;
- **C** = identifiant court du groupe (*cluster*, d'où 'C') ;
- **M** = membres du groupe (ici cette information se résume à les compter : « (n=493) ») ;
- **N*(S)** = résumé des principaux voisins de la série (id groupe et corrélation au médoïde) ;
- **r** = corrélation de la série par rapport au médoïde du groupe (ici parfaite : r = +1, puisque c'est la série médoïde qui est mise en avant-plan ; d'où l'ajout de la mention « (ref.) ») ;
- **N1(M)** : résumé des plus fréquents « premiers voisins » des membres (id groupe voisin et nombre d'occurrences ; puis mention du nombre de groupes voisins distincts recensés pour ces membres (info utile si tous les voisins ne peuvent pas être affichés : « (n=...) »).

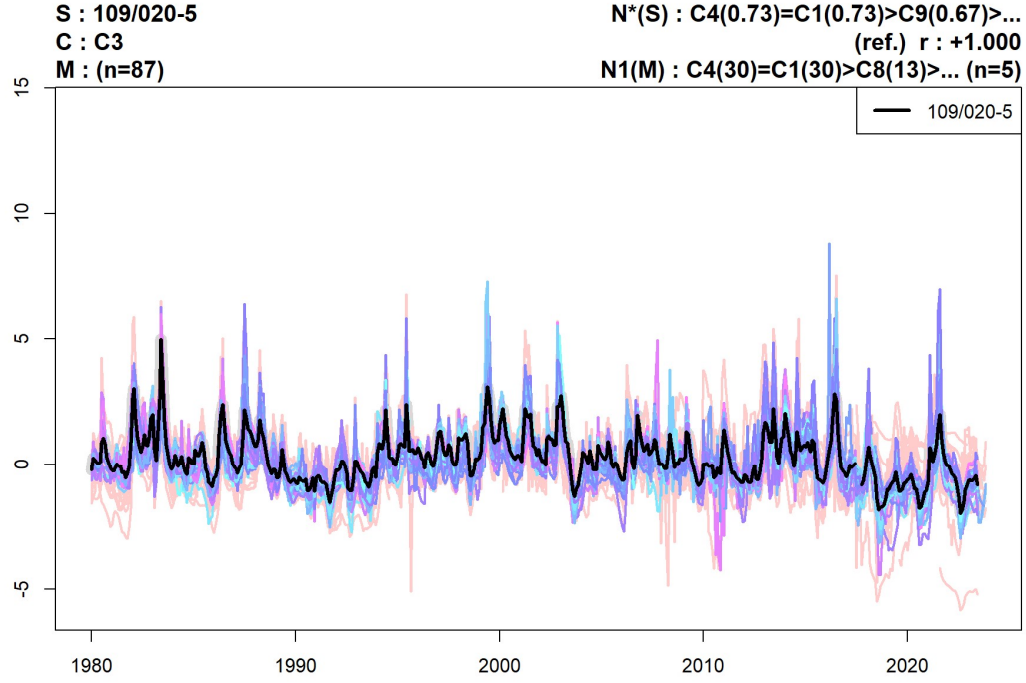
C1 : NOT-RHINE-ALL-OTHERS



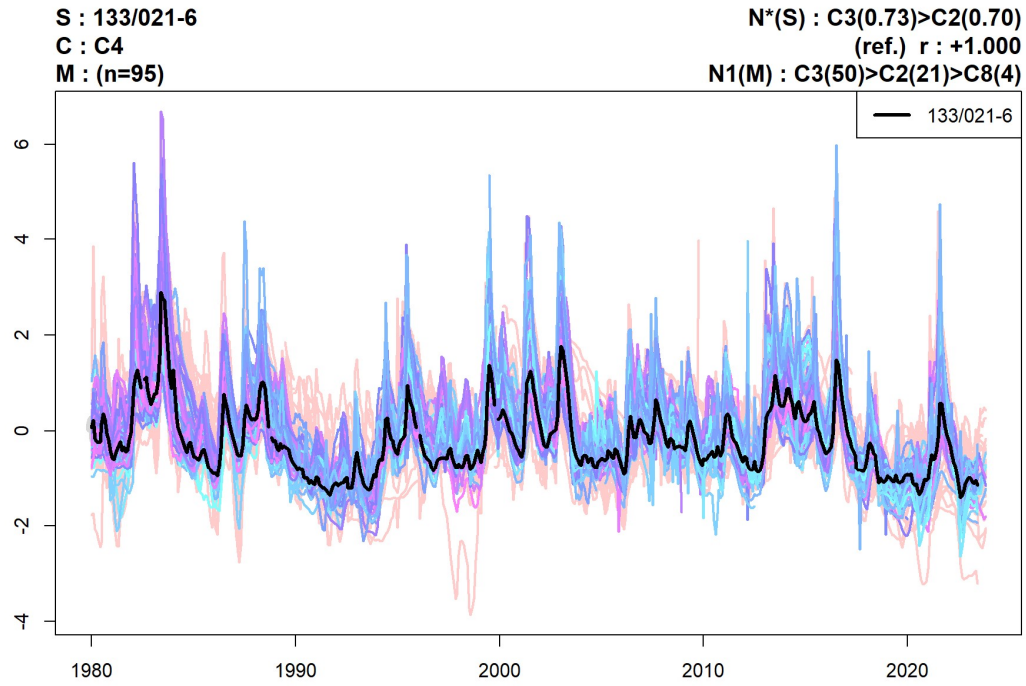
C2 : NOT-RHINE-PA_B2



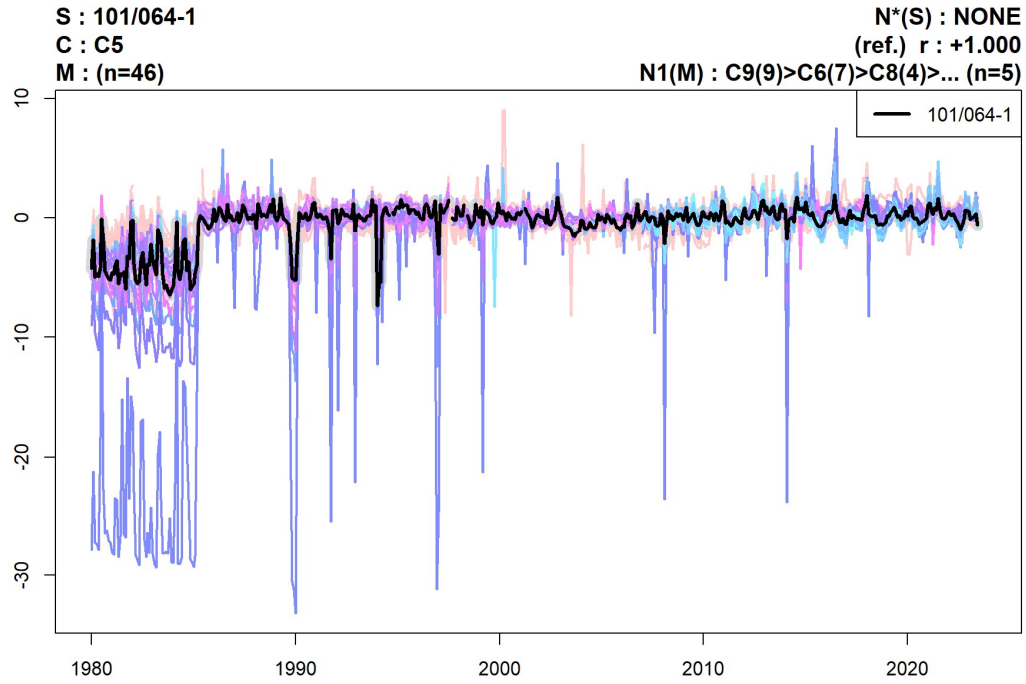
C3 : NOT-RHINE-PA_B5-full



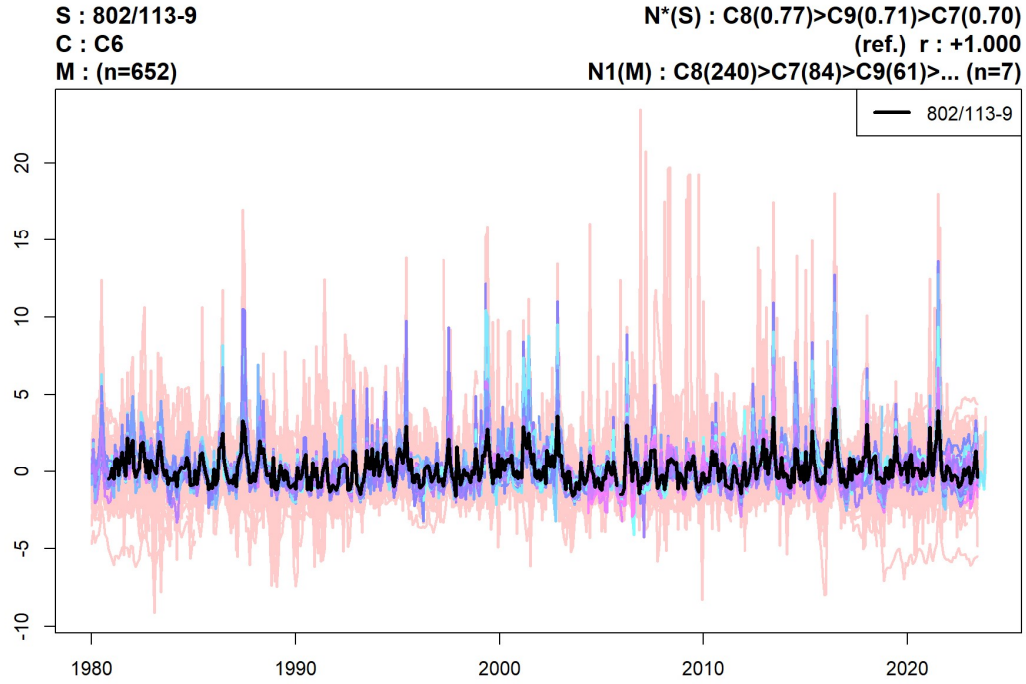
C4 : NOT-RHINE-PA_B6



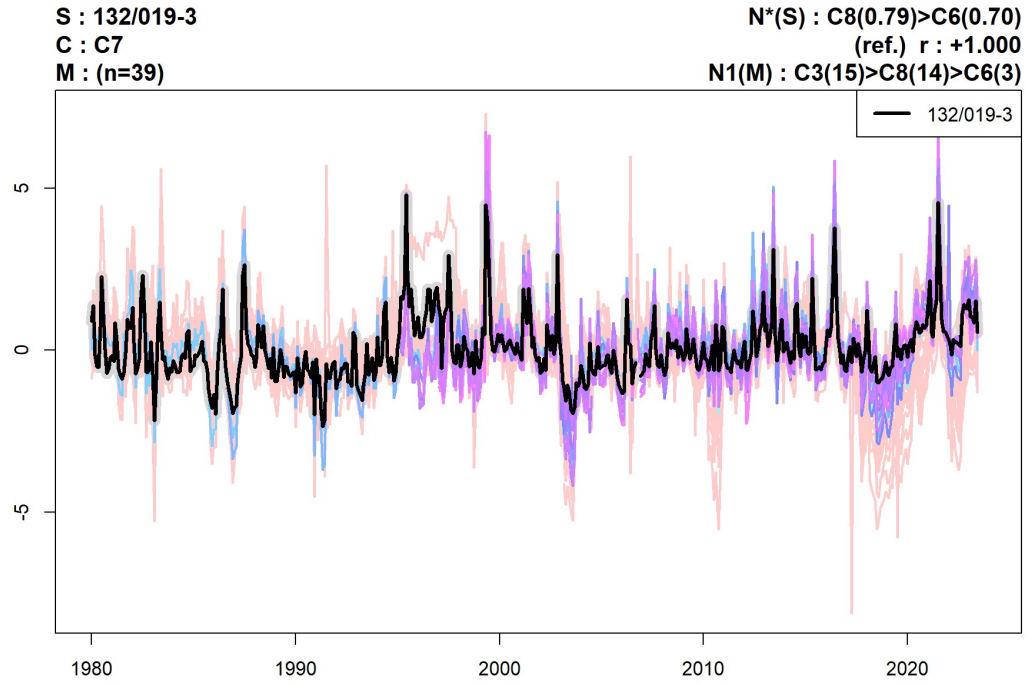
C5 : RHINE-A6-NEAR-KEHL



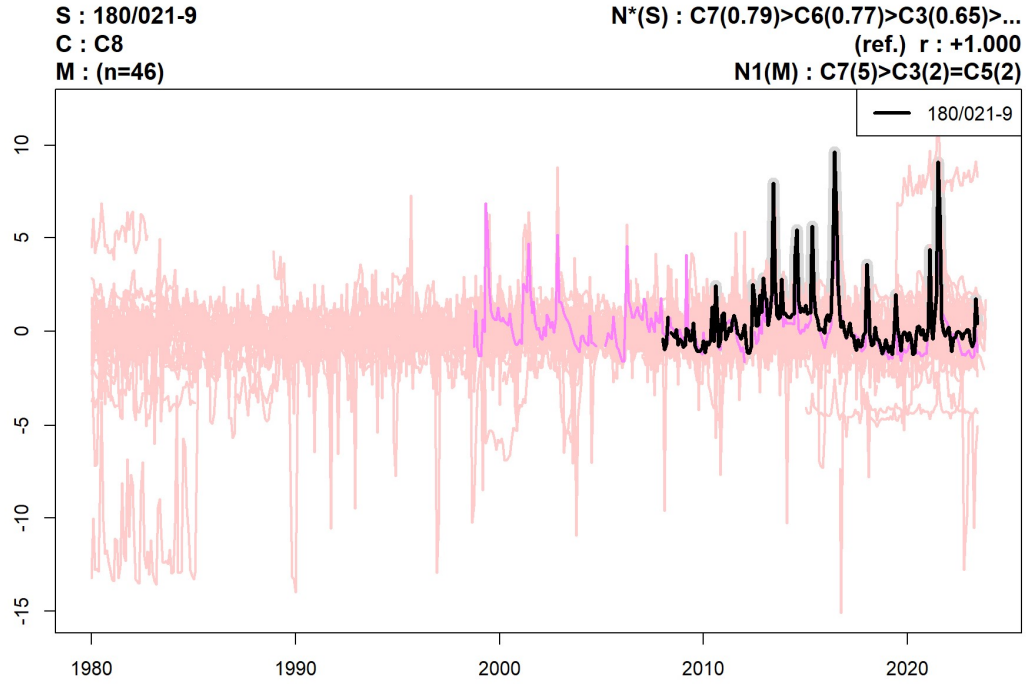
C6 : RHINE-ALL-OTHERS



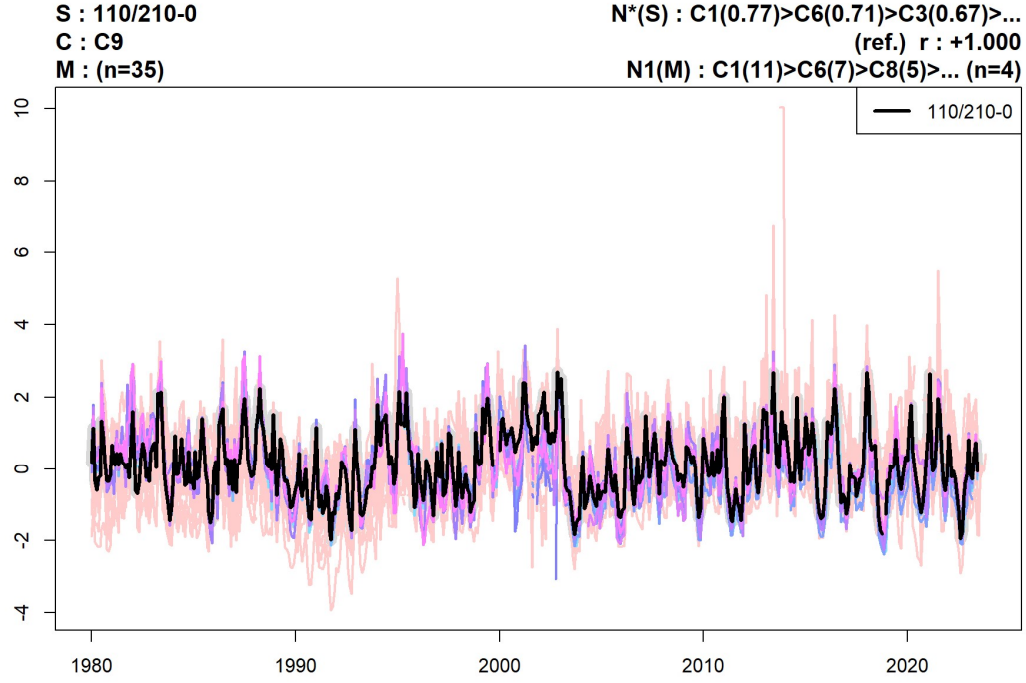
C7 : RHINE-NEAR-BREISACH



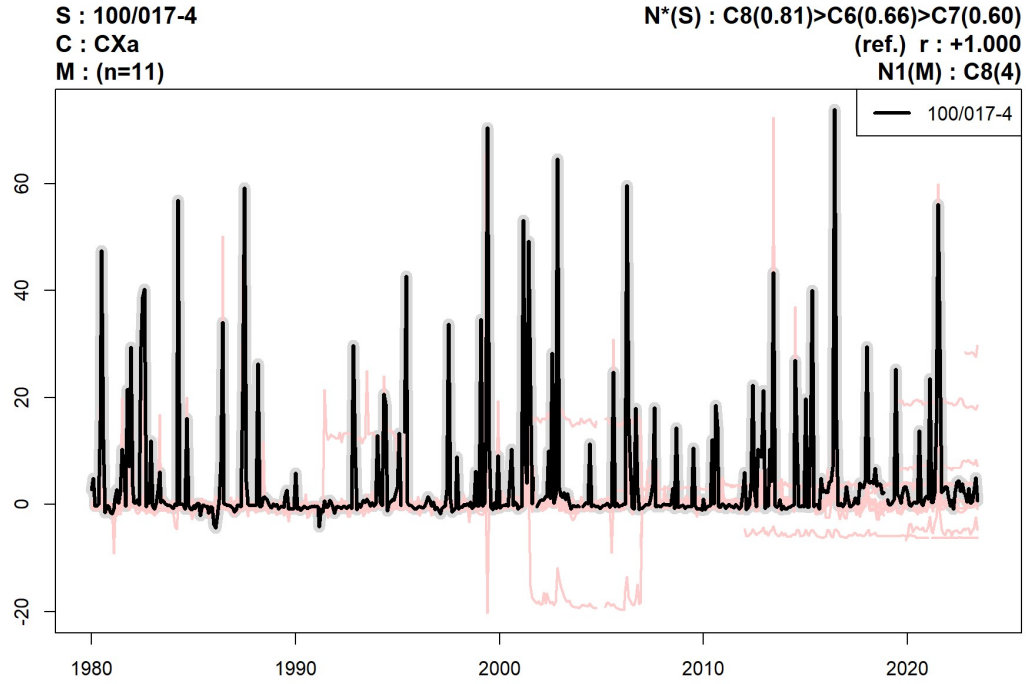
C8 : SINGULAR



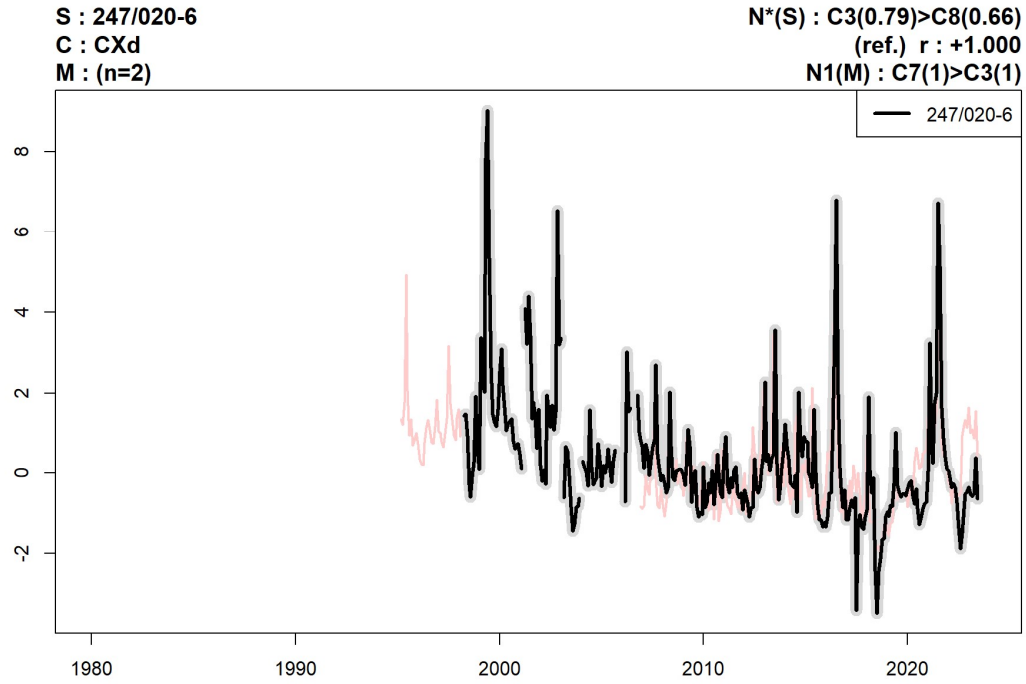
C9 : TRANSITION



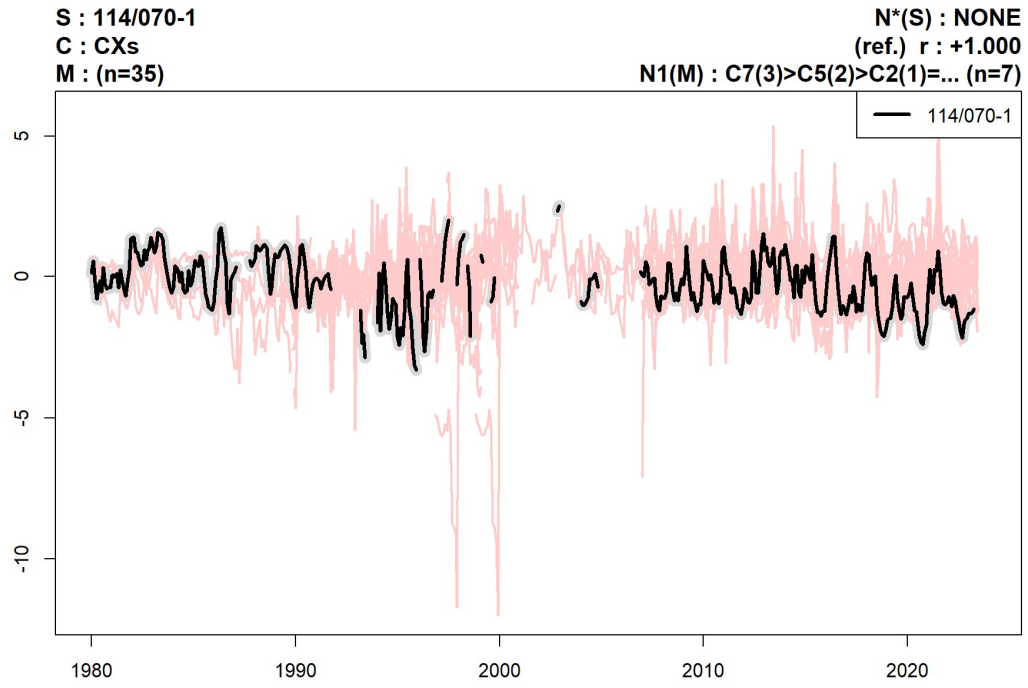
CXa : ANOMALOUS



CXd : DISCARD

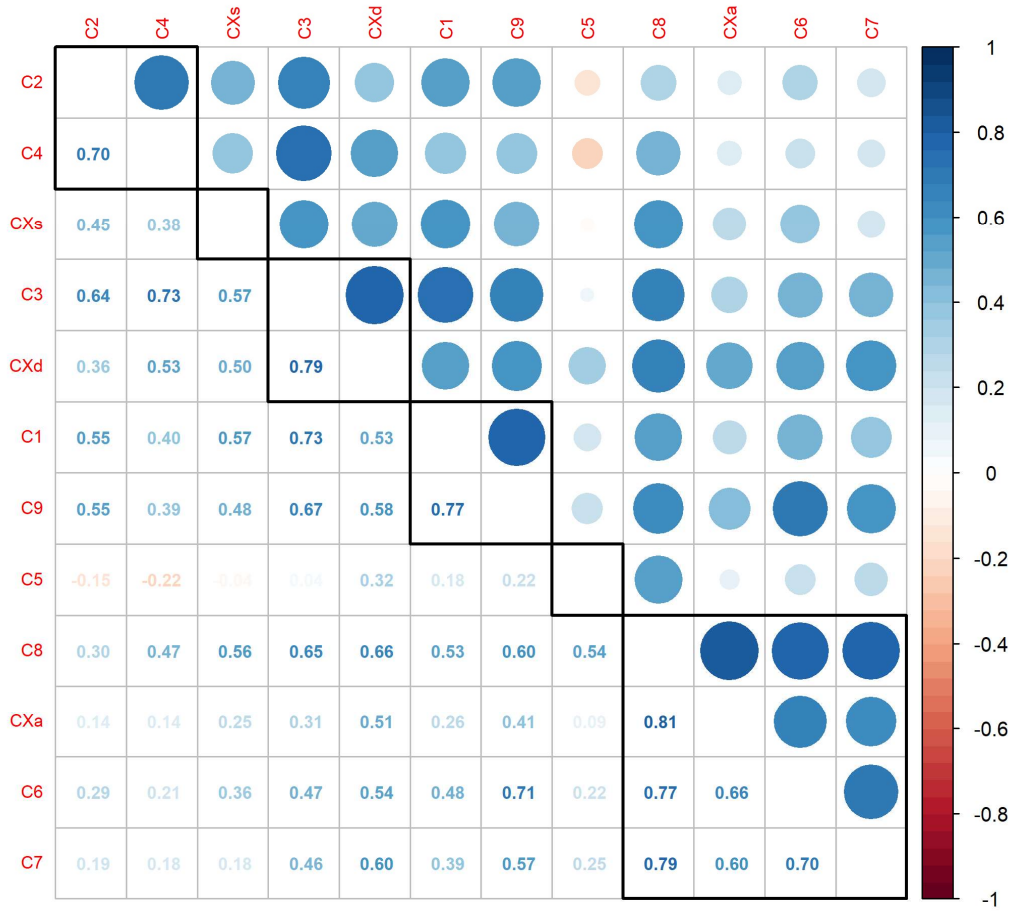


CXs : TOO-SHORT-or-MISSING



2. ANNEXE 2 : Analyses statistiques d'indicateurs du regroupement final (synthèse)

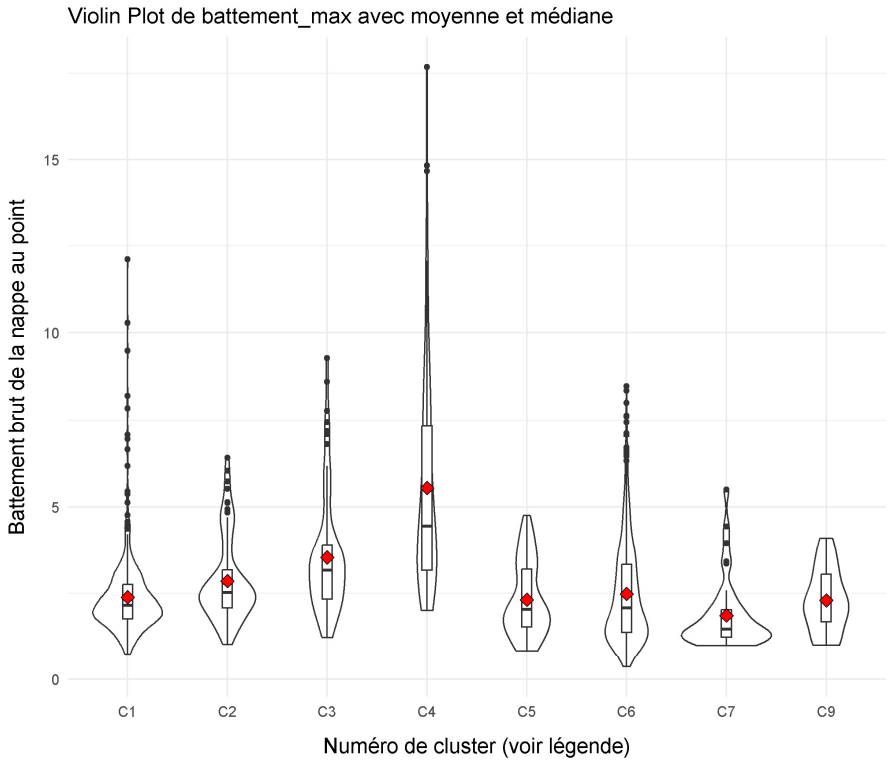
Matrice des corrélations (r) entre les médoïdes et Groupes de groupes bien corrélés



Il est conseillé d'ignorer les groupes secondaires (CX…) sauf si le but est spécifiquement d'identifier d'où les séries de ces groupes peuvent provenir majoritairement (ex. CXd surtout de C3 ?). La partie supérieure droite du diagramme illustre la force des corrélations, dont les coefficients r sont rapportés dans la partie inférieure gauche.

Battement de la nappe

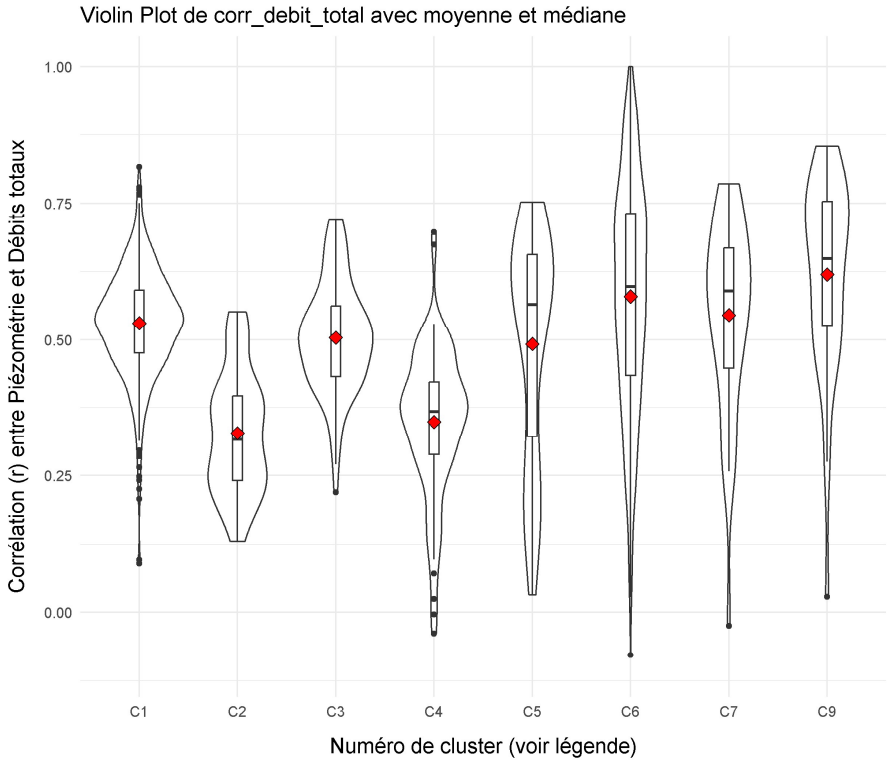
Les battements (amplitudes de variation du niveau piézométrique) sont plus importants dans les groupes « Non-Rhin », en particulier dans le groupe C4…



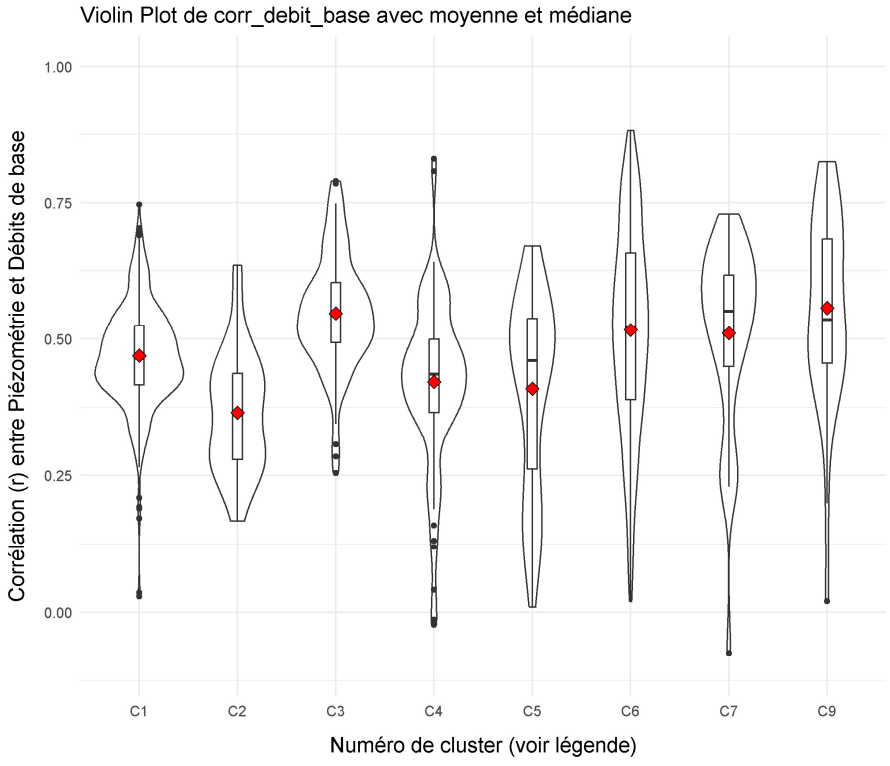
N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35

Corrélation au débit du Rhin

Les groupes C6 et C9 présentent les corrélations les plus fortes, en moyenne, au débit total du Rhin. C'est moins marqué avec le débit de base, où la corrélation médiane des groupes Rhin se confond à celle de certains groupes Non-Rhin.



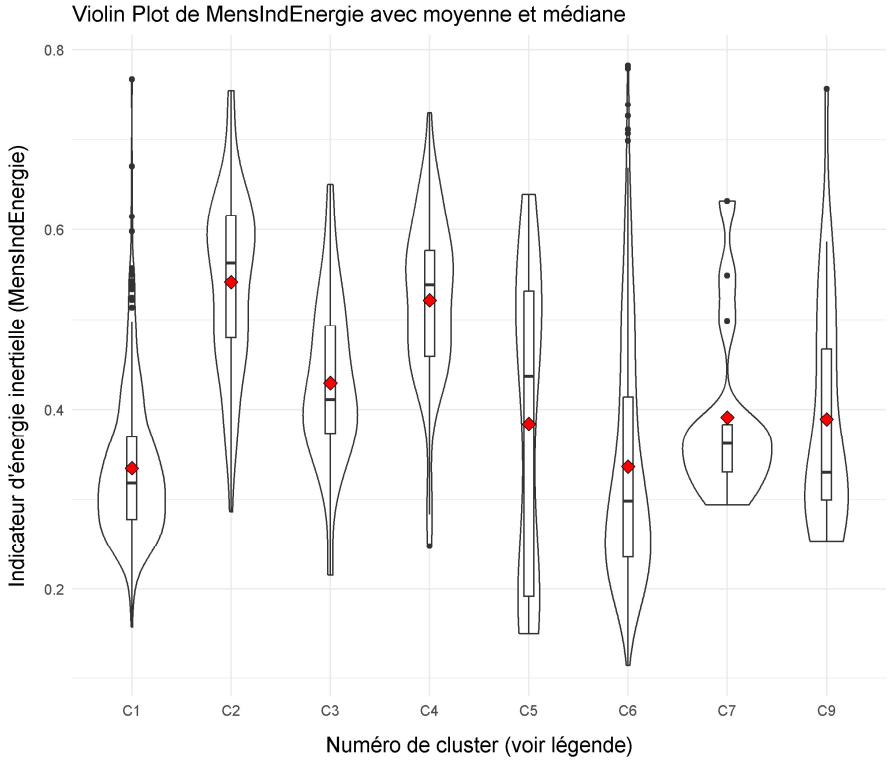
N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35



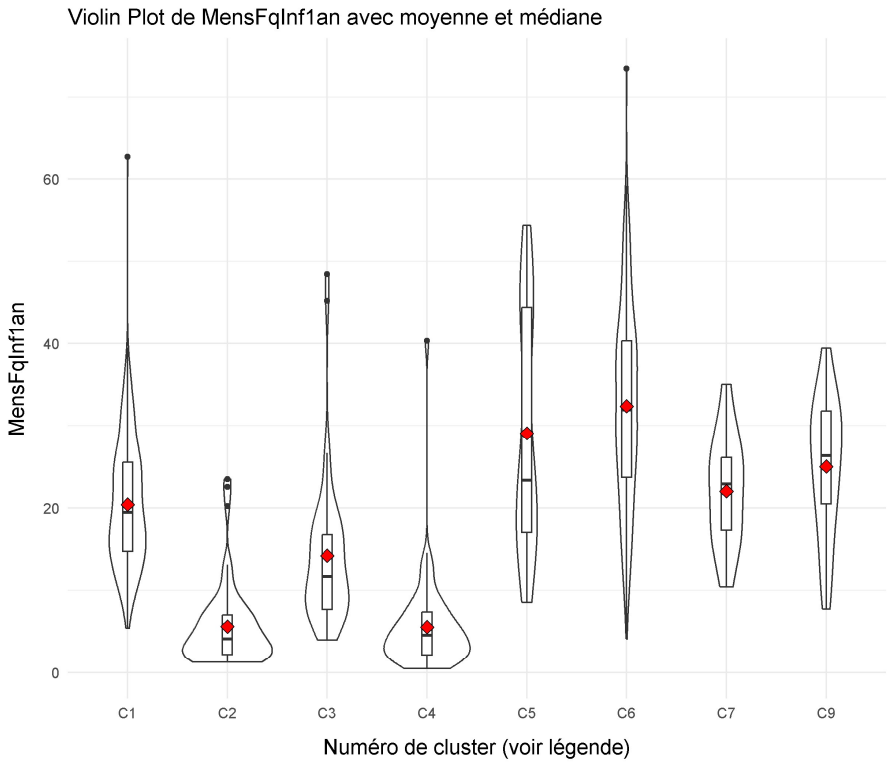
N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35

Décomposition de la série piézométrique : réactivité / inertie

Les groupes présentant en moyenne les dynamiques les moins inertielles sont C1 et C6. Les groupes les plus inertiels sont C2 et C4. Remarque : les résultats ne sont pas fiables pour C5 car la hausse soudaine et pérenne du niveau piézométrique vers 1985 est assimilée par l'algorithme d'analyse à une fluctuation pluriannuelle très lente périodique (alors que c'est un événement unique).

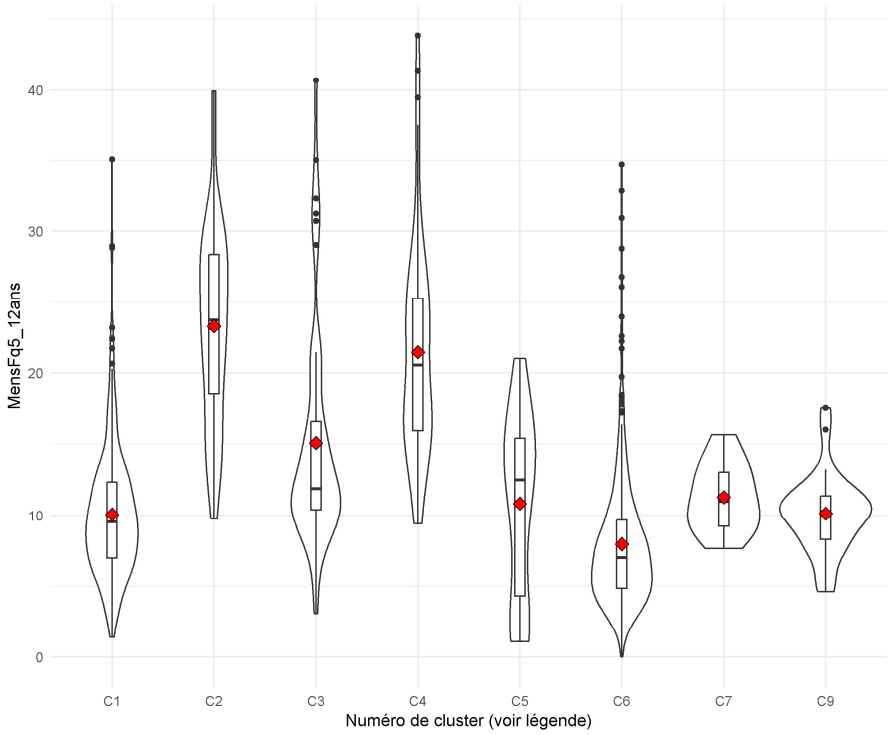


N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35



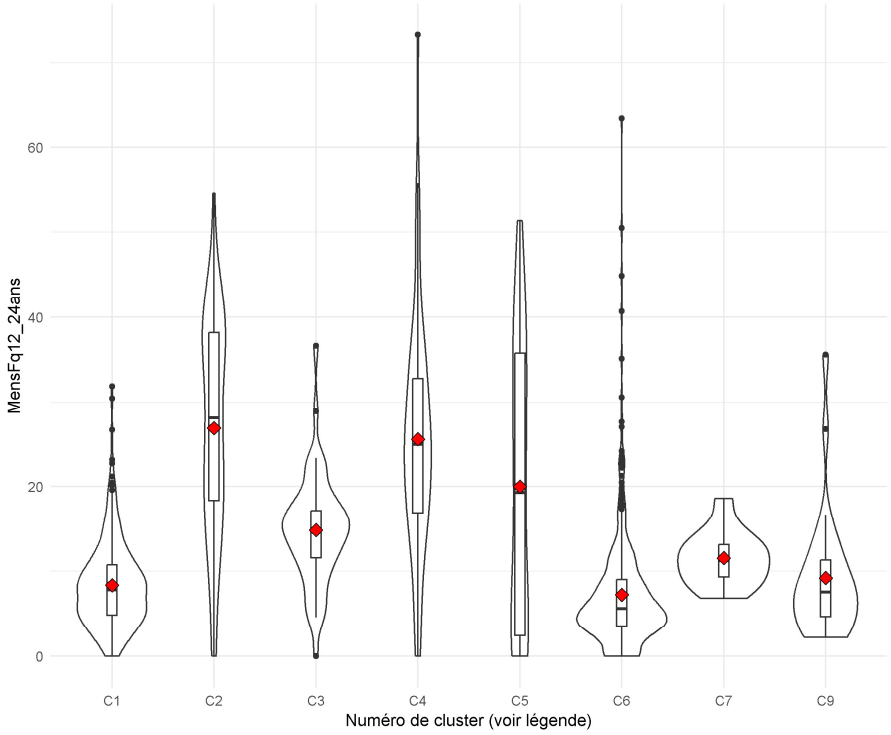
N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35

Violin Plot de MensFq5_12ans avec moyenne et médiane



N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35

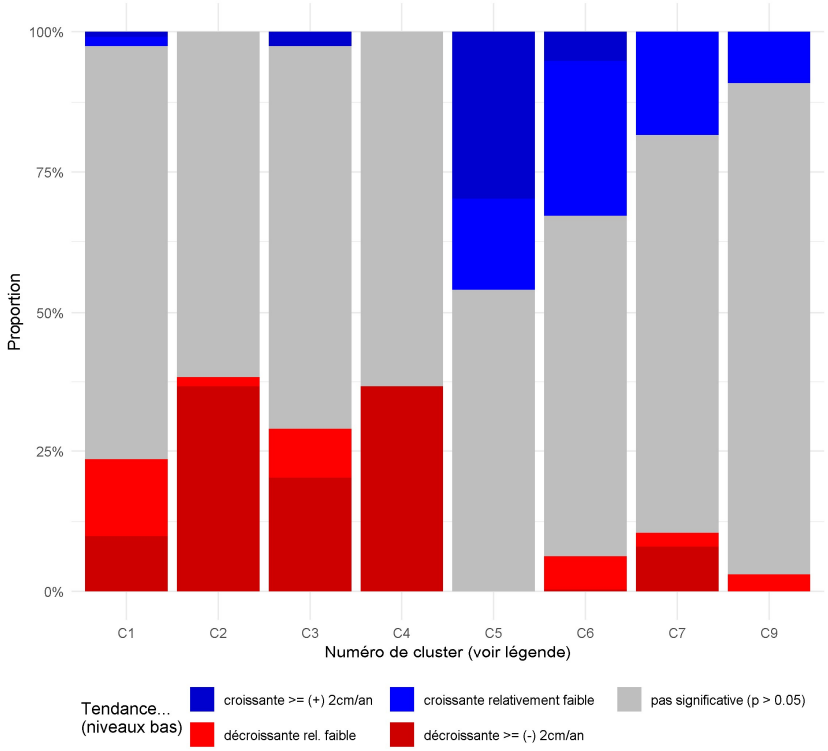
Violin Plot de MensFq12_24ans avec moyenne et médiane



N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35

Tendance d'évolution temporelle des Niveaux bas (étiages) annuels

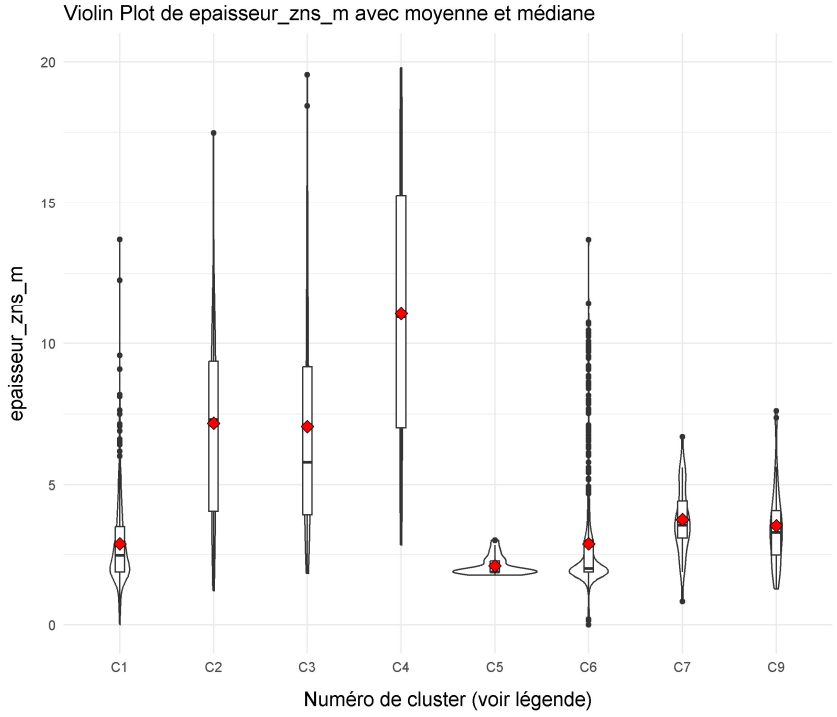
D'après les indicateurs de tendance ayant pu être évalués (soit 1336 des 1528 points classés dans les groupes principaux en excluant C8 'SINGULAR') on voit clairement une opposition de la tendance générale des niveaux piézométriques entre les nappes influencées par le Rhin (plus souvent à la hausse), d'une part, et les autres nappes non influencées par le Rhin (plutôt à la baisse), d'autre part. Noter que la tendance est négligeable, non significative, dans la majorité des cas, peu importe le groupe.



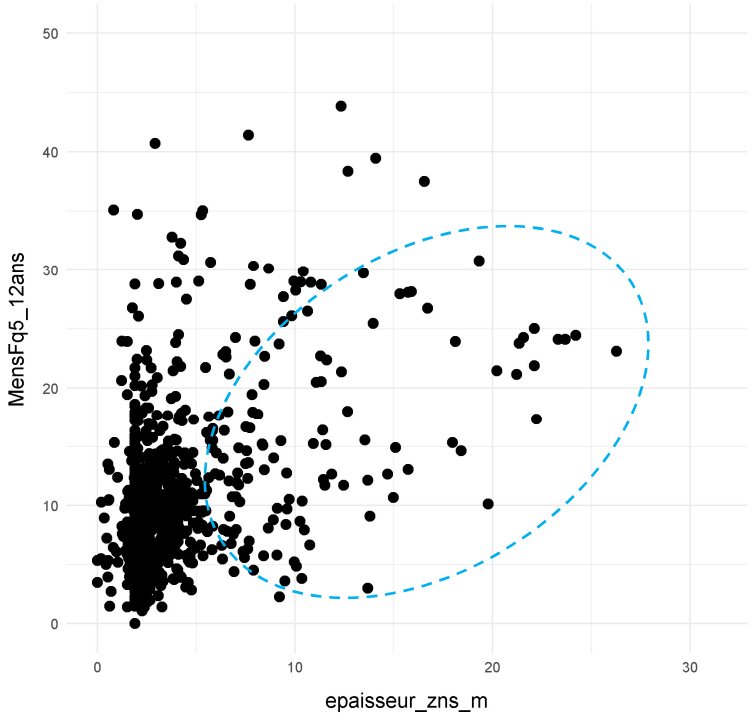
N°	ID du cluster	N.tot	N.NAs
C1	NOT-RHINE-ALL-OTHERS	493	60
C2	NOT-RHINE-PA_B2	81	21
C3	NOT-RHINE-PA_B5-full	87	8
C4	NOT-RHINE-PA_B6	95	16
C5	RHINE-A6-NEAR-KEHL	46	9
C6	RHINE-ALL-OTHERS	652	75
C7	RHINE-NEAR-BREISACH	39	1
C8	SINGULAR	46	5
C9	TRANSITION	35	2
CXa	ANOMALOUS	11	1
CXd	DISCARD	2	0
CXs	TOO-SHORT-or-MISSING	35	25

Epaisseur de la zone non saturée (ZNS)

Les groupes de points des groupes C2, C3 et C4 surtout, sont situés à des endroits où la zone non saturée (ZNS) est d'une épaisseur (>5 mètres) notable et impactante. Cette relation entre épaisseur de ZNS et degré d'inertie (importance des composantes pluriannuelles) se perçoit assez bien dans le graphique bivarié (nuage de points) ci-dessous, bien que la relation ne soit que partielle d'après cette analyse (où les données n'ont pas été filtrées finement).



N°	ID du cluster	N.tot
C1	NOT-RHINE-ALL-OTHERS	493
C2	NOT-RHINE-PA_B2	81
C3	NOT-RHINE-PA_B5-full	87
C4	NOT-RHINE-PA_B6	95
C5	RHINE-A6-NEAR-KEHL	46
C6	RHINE-ALL-OTHERS	652
C7	RHINE-NEAR-BREISACH	39
C8	SINGULAR	46
C9	TRANSITION	35
CXa	ANOMALOUS	11
CXd	DISCARD	2
CXs	TOO-SHORT-or-MISSING	35



Références

[1] Baulon, L., 2023. Déterminisme climatique et hydrogéologique de l'évolution à long terme des niveaux piézométriques. Thèse de doctorat. Normandie Université.

[2] Baulon L., Manceau JC., Giuglaris E. 2025. GRETA - Action 3.3 : Analyse de l'évolution historique du niveau de la nappe rhénane par des indicateurs piézométriques. Rapport final V1. BRGM/RP-74683-FR, 85 p.

[3] Hamed, K.H., Rao, A.R., 1998. A modified Mann–Kendall trend test for autocorrelated data. J. Hydrol., 204, 219–246.

[4] Heudorfer B, Haaf E, Stahl K, Barthel R (2019) Index-based characterization and quantifcation of groundwater dynamics. Water Resour Res 55(7):5575–5592. https://doi.org/10.1029/2018WR024418

[5] Kendall, M.G. 1975. Rank Correlation Methods, 4th edition, Charles Griffin, London.

[6] Mann, H.B. 1945. Non-parametric tests against trend, Econometrica 13:163-171.

[7] Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A., & Sander, J. (2014, April). Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining* (pp. 839-847). Society for Industrial and Applied Mathematics.

[8] Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(2), 126-135.

[9] Richter BD, Baumgartner JV, Powell J, Braun DP (1996) A method for assessing hydrologic alterationwithin ecosystems. Conserv Biol 10(4):1163–1174. https://www.jstor.org/stable/2387152

[10] Rousseeuw, P.J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65 https://doi.org/10.1016/0377-0427(87)90125-7

[11] Sen P.K., (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63, 1379-1389.

[12] Vaute, Laurent ; Laurencelle, Marc (2023) - Typologie des points d'eau pour l'interprétation des tendances d'évolution de la qualité des eaux souterraines du bassin Rhin-Meuse : état de l'art, développements et méthodologie. Rapport final . BRGM/RP-72856-FR, 118 p. infoterre.brgm.fr/rapports/RP-72856-FR.pdf

[13] Wunsch, A., Liesch, T. & Broda, S. (2021)Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles. Water Resource Management 36, 39–54 https://doi.org/10.1007/s11269-021-03006-y

[14] Projet INTERREG IV - « Liaison Opérationnelle pour la Gestion de l'Aquifère Rhénan - LOGAR» : (Région Alsace 2009-2012) : https://www.grandest.fr/preserver-biodiversite/preserver-eau/liaison-operationnelle-gestion-aquifere-rhenan/projet-logar/ Rapport final du projet « Interreg IV – Liaison Opérationnelle pour la Gestion de l'Aquifère Rhénan » 2012