Classification de séries temporelles hétérogènes pour le suivi de l'état des cours d'eau

Agnès Braud, Pierre Gançarski, Corinne Grac, Agnès Herrmann, Florence Le Ber, Harrison Vernier

Classification de séries temporelles hétérogènes pour le suivi de l'état des cours d'eau

Agnès Braud*, Pierre Gançarski*, Corinne Grac**, Agnès Herrmann***, Florence Le Ber*, Harrison Vernier*

*ICube, Université de Strasbourg, CNRS, ENGEES, F 67000 Strasbourg
{agnes.braud, gancarski, florence.le-ber}@unistra.fr
**LIVE, Université de Strasbourg, CNRS & ENGEES, F 67000 Strasbourg
corinne.grac@engees.unistra.fr
***LHYGES, Université de Strasbourg, CNRS, ENGEES, F 67000 Strasbourg
agnes.herrmann@engees.unistra.fr

Résumé. Dans cet article, nous présentons le processus collaboratif mis en place entre des thématiciens hydroécologues et des informaticiens. Il s'agissait d'adapter une méthode de clustering à l'analyse de séquences temporelles constituées d'une suite de mesures physico-chimiques effectuées sur des cours d'eau. Les données sont caractérisées par la variabilité de l'échantillonnage et le grand nombre de paramètres diversement suivis, ce qui génère hétérogénéité et incomplétude. Une sélection a été opérée pour construire un jeu de données réduit, comportant environ 300 séquences, sur lesquelles nous avons appliqué une méthode de clustering spécifique aux données temporelles. Nous illustrons et commentons les résultats obtenus au travers de visualisations adaptées.

1 Introduction

Le projet ANR FRESQUEAU qui s'est déroulé de 2011 à 2015 a permis de construire une base de données hydrobiologiques importante sur deux grands bassins hydrographiques, correspondant aux districts Rhin-Meuse et Rhône-Méditerranée-Corse (Bimonte et al., 2015). Cette base contient en particulier des données collectées par les agences de l'eau dans le cadre du réseau de suivi créé pour évaluer l'état écologique de masses d'eau, en accord avec la directive cadre européenne sur l'eau (DCE) (The European Parliament and the Council, 2000). La DCE requiert le bon état (écologique et chimique) des masses d'eau à court (2021) et moyen terme (2027). Elle définit une masse d'eau comme *« un élément discret et significatif d'eau de surface tel que [...] un fleuve, une rivière ou un canal, un segment d'un fleuve, d'une rivière ou d'un canal [...] »* et l'état écologique comme l'expression de la qualité de la structure et du fonctionnement d'un écosystème aquatique continental (DCE, art. 2), dont la physico-chimie de l'eau est une des composantes majeure. Le principal objectif de FRESQUEAU était de proposer des méthodes permettant de rendre compte de cette qualité à partir des données de suivi des cours d'eau. Les méthodes proposées dans ce contexte, fondées principalement sur la recherche de motifs séquentiels (Fabrègue et al., 2014) et l'analyse relationnelle de concepts

(Dolques et al., 2016) permettent de traiter des données discrétisées ou catégorielles. Les résultats obtenus ont montré la validité et la pertinence de ces approches en apportant aux experts hydroécologues des informations utiles à la compréhension des phénomènes réels impliqués dans les variations de la qualité des eaux.

Suite à ce projet, la base FRESQUEAU a été complétée à l'échelle nationale pour la période 2007-2013 : elle contient les données issues de tous les sites d'échantillonnage (dénommés *stations* dans la suite) du réseau de contrôle et de surveillance (RCS) des cours d'eau français. Néanmoins, malgré les avancées méthodologiques et thématiques du projet FRESQUEAU, une question reste en suspens. Elle concerne les apports potentiels de méthodes traitant directement les données numériques originelles. En effet, ces données, formant des séries temporelles constituées chacune d'une suite de mesures physico-chimiques réalisées sur une station, représentent plusieurs millions de mesures, fortement hétérogènes tant en nature et qualité de la mesure qu'en fréquence d'acquisition. La forte évolution temporelle de ces données est connue mais complexe car liée à de multiples mécanismes à la fois naturels, tels que la pluie et les variations saisonnières, et anthropiques, tels que les pratiques agricoles. De plus ces mécanismes peuvent être modifiés par le changement climatique. Si cette évolution conditionne les recommandations de fréquence de suivis mensuels ou bimestriels des stations du RCS, elle n'est pas prise en compte dans l'exploitation de ces données, qui sont agrégées annuellement par paramètre, puis par groupe de paramètres, et enfin discrétisées en cinq niveaux de qualité (Ministère de la Transition Ecologique et Solidaire, 2019). Ainsi, après plusieurs simplifications, l'expert obtient les classes de stations similaires, mais ces classes ne tiennent pas compte des évolutions temporelles des paramètres considérés.

L'objectif du projet ADQEAU (2019-2021) est de proposer une méthode d'analyse de ces séries temporelles numériques qui permette de mettre en évidence les évolutions de la qualité des eaux en réponse à des pressions. Néanmoins, la classification de données complexes, qu'elle soit supervisée ou non, est un processus long. En effet, pour que les algorithmes de classification soient efficaces et rapides, il est indispensable que la phase de préparation de données soit menée avec rigueur et soin. Ainsi, dans une première phase, en plus du choix de la méthode de classification et de son paramétrage, un effort important dans ce projet a porté sur la sélection et la mise en forme des données. Afin de pallier le manque de formalisation opérable de la connaissance du domaine, une méthode de clustering a été proposée dans une deuxième phase permettant ainsi de valider une approche non supervisée. Enfin, dans une troisième phase, encore en cours, un mécanisme d'interaction avec l'experte sera mis en œuvre. Celle-ci peut piloter le processus non supervisé par l'introduction de ses connaissances au fur et à mesure des besoins de l'analyse. Ainsi l'effort, souvent chronophage, de traduction *a priori* de son expertise en connaissances opérables devrait être fortement allégé. Par ailleurs, la nature même de la science des données mise en œuvre dans le projet ADQEAU impose une collaboration forte entre les producteurs/consommateurs de données, ici les hydroécologues, experts du domaine d'application, et les informaticiens, spécialistes du traitement de données, afin de co-construire une méthodologie d'analyse et les outils associés, permettant de réduire fortement le fossé entre les données brutes, les clusters construits et les intuitions de l'expert, comme par exemple les classes thématiques potentielles.

L'article est organisé comme suit. La section 2 justifie la méthode de clustering choisie et le processus collaboratif mis en place entre hydroécologues et informaticiens. La section 3 décrit les données disponibles et la procédure de sélection et de prétraitement mise en œuvre. La

section 4 présente les résultats obtenus pour deux groupes de paramètres aux caractéristiques différentes. Différentes visualisations sont exploitées pour faciliter l'analyse. Enfin, la section 5 conclut et dresse quelques perspectives méthodologiques et thématiques.

2 Cadre méthodologique

2.1 Problématique

L'analyse de données temporelles en hydrobiologie est actuellement un processus très souvent manuel nécessitant d'une part, un investissement important de l'expert et d'autre part, une connaissance approfondie des phénomènes sous-jacents à ces données. Cette analyse consiste en une étude de critères globaux établis par station. Ainsi, localement, les mesures de chaque paramètre sont agrégées annuellement en percentile, moyenne ou concentration maximale, puis synthétisées par groupe de paramètres cohérents, et enfin discrétisées en cinq niveaux de qualité de très bon à mauvais, notés de 1 à 5 (Ministère de la Transition Ecologique et Solidaire, 2019). Les regroupements de stations sont ensuite réalisés à partir de ces données doublement agrégées et discrétisées causant une perte de l'information temporelle. De fait, cette perte interdit une analyse fine de la dynamique pluriannuelle des stations pour mettre en évidence des particularités ou au contraire des similitudes dans les évolutions temporelles.

L'objectif de ce projet est donc de proposer aux experts une méthode d'analyse originale fondée sur l'extraction de groupes de stations permettant ainsi une analyse globale des évolutions internes à chaque groupe, mais aussi une étude comparative entre ces groupes. Dans ce contexte, l'utilisation de méthodes de clustering de séquences temporelles semble tout à fait pertinente et naturelle. Ces méthodes nécessitent néanmoins une interaction forte avec l'expert, permettant une sélection et une préparation pertinente des données en regard des résultats et de leur interprétation.

2.2 Classification non supervisée de séries temporelles

La volonté de détecter, analyser et classifier les améliorations ou dégradations – lentes ou abruptes – qui affectent la qualité des cours d'eau, nécessite le développement de méthodes innovantes d'analyse et d'interprétation en rupture avec les méthodes du domaine. En effet, les méthodes d'apprentissage supervisé classiques ou même récentes, telles que l'apprentissage profond, font l'hypothèse que les classes recherchées sont parfaitement connues et définies et que l'expert du domaine est capable de fournir un jeu de données d'apprentissage suffisant aussi bien en nombre d'exemples qu'en qualité de ceux-ci. De fait, les données d'apprentissage doivent décrire de manière suffisante et complète les classes auxquelles elles sont rattachées. Or, dans le cas des données de suivi des cours d'eau, cette hypothèse n'est pas réaliste. Les hydroécologues sont capables d'évaluer une station à une date donnée, selon un cadre prédéfini, tel que la grille SEQ-eau. Cependant ils n'ont pas de catégories permettant de qualifier l'évolution d'une station à partir des données temporelles disponibles. Pour contourner ce problème, des approches non supervisées, de type clustering, qui ne nécessitent que très peu d'information préalable, sont classiquement utilisées. Ainsi, des travaux antérieurs ont montré la capacité des méthodes de clustering basées sur DTW (Dynamic Time Warping) et DBA (Dynamic Barycentric Averaging) à extraire des regroupements de données (Petitjean et al., 2011).

Concrètement, dans nos expériences, nous avons développé, interfacé ou utilisé différents outils de clustering classiques (K-moyennes principalement) et de visualisation. Ainsi, le clustering a été réalisé par les méthodes propres à la plateforme FODOMUST (Gançarski et al., 2020). La sélection des variables et l'affichage des données et résultats sont faits grâce à la bibliothèque Scikit-learn en Python qui a été interfacée avec la plateforme. La mesure de similarité utilisée est basée sur DTW, la moyenne associée étant DBA. Grâce à l'usage de DTW, le problème des données manquantes (séquences de longueurs différentes) est en grande partie résolu (voir section 3.3).

2.3 Le projet ADQEAU : un processus fortement collaboratif

Au vu des spécificités des données de suivi de l'état des cours d'eau et des besoins exprimés dans le projet ADQEAU, il est apparu rapidement que la plateforme FODOMUST n'offrait pas suffisamment de fonctionnalités de visualisation et d'interaction, et nécessitait donc des adaptations. Nous avons alors mis en œuvre un processus de collaboration entre les hydroécologues, experts du domaine d'application, et les informaticiens, spécialistes du traitement de données, afin de réaliser ces adaptations. La figure 1 présente ce processus. L'extraction de jeux de données à partir de bases de données existantes (a) nécessite une connaissance forte sur leurs potentialités. De ce fait cette tâche (1) indispensable et cruciale est faite par les hydroécologues. Les jeux de données (b) extraits décrivent des caractéristiques physico-chimiques des cours d'eau et sont transmis (2) aux spécialistes de traitement de données. Ceux-ci étudient le problème en vue de modifier (3) les algorithmes existants ou d'en développer de nouveaux pour répondre à la demande. Les résultats (c) des traitements (4) sont retournés aux experts pour analyse thématique (5) et éventuelle modification du jeu de données ou énonciation de contraintes pour raffiner les résultats.

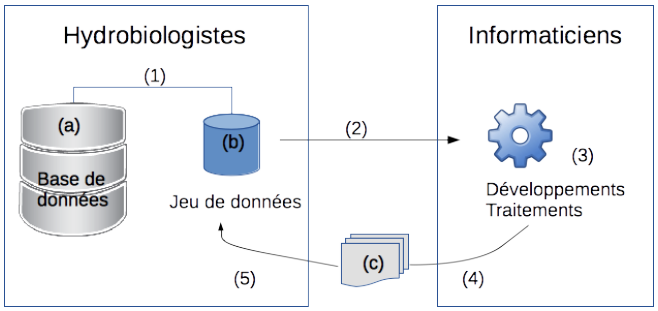

FIG. 1 – *Une collaboration entre informaticiens et hydroécologues*

Grâce à cette collaboration qui s'est avérée très fructueuse tant en développements méthodologiques et technologiques que par une compréhension mutuelle des problématiques de ces deux domaines, nous avons pu obtenir des résultats d'analyse prometteurs.

		NITR (mg/L)	PEST (μ/L)			
Station	**Date**	**NO_3^-**	**Bifenox**	**Aclonifène**	**Pendiméthaline**	**Pendiméthaline**
Station 1	17/01/07	28,8	0,1	0,05	0,05	0,1
	13/02/07	19,2	0,1	0,05	0,05	0,1
	05/03/07	12,7	0,1	0,05	0,05	0,1
	10/04/07	11,4	-	-	-	-
	11/05/07	11,5	-	-	-	-
	05/06/07	10,5	-	-	-	-
	10/07/07	11,5	-	-	-	-
Station 2	17/01/07	22,5	0,1	0,05	0,05	0,1
	13/02/07	14,3	0,1	0,05	0,05	0,1
	05/03/07	10,8	0,1	0,05	0,05	0,1
	10/04/07	7,9	-	-	-	-
	11/05/07	7,4	-	-	-	-
	05/06/07	6,4	-	-	-	-
	10/07/07	7	-	-	-	-
	22/08/07	13,3	-	-	-	-

TAB. 1 – *Exemple de mesures sur deux stations pour l'altération Nitrate (NITR), composée d'un paramètre, et pour 4 des 68 paramètres de l'altération Pesticide (PEST)*

3 Préparation des données

3.1 Les données de suivi de l'état des cours d'eau

Les données du réseau de suivi national qui ont été intégrées dans la base FRESQUEAU couvrent 1781 stations pour la période de 2007 à 2013. Elles représentent des mesures de paramètres physico-chimiques dont le nombre par paramètre est très variable. Chaque mesure est associée à une information temporelle (date du prélèvement) et une information géographique liée à la station (voir tableau 1). Chaque station est elle-même rattachée à un cours d'eau et à une hydro-écorégion, c'est-à-dire une zone homogène pour le contexte physique des cours d'eau (Wasson et al., 2004).

La base comprend plus de 23 millions de résultats pour 1201 paramètres physico-chimiques. Parmi eux, 2% sont des macro-paramètres (par exemple, pH, azote, matières en suspension) et 98% des micropolluants (par exemple, cuivre, atrazine, ou médicaments). Les analyses des macro-paramètres sont généralement effectuées 12 fois par an, et seulement 6 fois pour les micro-polluants. Toutefois, la complétude des données varie énormément (de 100% à moins de 1%) car, alors que les macro-paramètres sont suivis partout, les micro-polluants sont diversement suivis (voir tableau 1). Les données physico-chimiques sont également accompagnés d'un certain nombre d'informations, qui permettent de les interpréter : unité de mesure, fraction (eau, sédiments ...) sur laquelle la mesure a été réalisée et seuils de quantification et de détection de la mesure. Ces données sont nationales, contrôlées et validées et donc considérées comme fiables.

Finalement, l'analyse de ces mesures se fait habituellement via des grilles qui permettent d'une part de regrouper des paramètres cohérents dans des ensembles appelés "altérations" (par exemple, l'altération nitrates – NITR – est constituée du seul paramètre, nitrates (NO3), alors que l'altération pesticides – PEST – regroupe 68 molécules), d'autre part de les discrétiser au regard de seuils de qualité et enfin de les évaluer. La grille que nous utilisons dans ce projet est la grille SEQ-Eau [1] qui recouvre seulement 206 paramètres physico-chimiques.

1. `http://rhin-meuse.eaufrance.fr/IMG/pdf/grilles-seq-eau-v2.pdf`

	BDD initiale	Données sélectionnées
Période	2007-2013	2007-2013
Nombre d'HER	22	5
Nombre de stations	1 781	303
Intervalle de complétude	1-100%	29-100%
Nombre de paramètres	1201 (206 du SEQ-eau)	303 (150 du SEQ-eau)
Nombre d'enregistrements	23.10^6	$1{,}3.10^3$ (paramètres SEQ-eau)
Nombre d'altérations	16	33 (14 altérations SEQ-eau, PEST divisée en 15 groupes, MPOR divisée en 4 groupes)

TAB. 2 – *Caractéristiques des données de la BDD et de celles sélectionnées*

3.2 Sélection et prétraitement des données

Afin de garantir une explicabilité effective des liens entre les données et les clusters, nous avons sélectionné 303 stations de 5 hydroécorégions (HER) de l'est de la France, pour lesquelles l'expertise connue est suffisamment forte pour mener à bien cette analyse thématique. Une étude statistique des données de ces stations a ensuite été réalisée. L'ensemble des données est divisé en quatre quartiles selon le taux de remplissage de chaque paramètre physico-chimique, c'est-à-dire le pourcentage de couples (station, date) correspondant à une mesure du paramètre par rapport au nombre total de couples. Pour 75% des données, le taux de remplissage est inférieur à 29%, c'est-à-dire que 925 paramètres physico-chimiques sont très peu mesurés. Nous n'avons donc retenu que les 25% des paramètres dont le taux de remplissage est supérieur à ce taux. Du fait de ce seuil relativement bas, les données manquantes peuvent rester nombreuses pour certains des paramètres. Malgré cela, nous avons choisi de ne pas imputer ces données manquantes. En effet, les données physico-chimiques étudiées peuvent présenter des pics occasionnels de valeurs pour des raisons naturelles, comme les pluies, qui lessivent certains éléments (tels que matières en suspension, nitrates), ou liées aux activités humaines (telles que les épandages de pesticides opérés sur les cultures). De fait, forcer un paramètre à une valeur qui pourrait correspondre à un de ces pics ou à l'opposé, calculer une moyenne à partir de deux pics, induirait de l'avis de l'expert, plus de bruit que d'information utile.

Par ailleurs, pour réaliser l'analyse, les paramètres sont regroupés selon les altérations de la grille SEQ-eau. Or, parmi les 206 paramètres considérés dans cette grille, seuls 150 ont un taux de remplissage supérieur à 29%. De plus, deux des seize altérations du SEQ-eau ont un trop grand nombre de paramètres pour que les résultats obtenus soient analysables par l'expert : il s'agit des pesticides (PEST) et des micro-polluants organiques hors pesticides (MPOR), composées de respectivement 68 et 45 paramètres. Aussi avons-nous proposé une subdivision originale de ces paramètres (Ung, 2020) en 15 groupes de PEST, composés de 1 à 5 paramètres, et 4 groupes de MPOR, composés de 2 à 11 paramètres. Une synthèse des données de la BDD initiale et des données sélectionnées est présentée au tableau 2.

3.3 Structuration temporelle des données

Comme le montre le tableau 1, les données associées à une station peuvent être appréhendées de deux manières orthogonales :

— structuration horizontale (qualifiée ici de *multidimensionnelle*) : l'évolution d'une station est vue comme la suite d'états composés chacun d'un vecteur de paramètres (ici 4 pour l'altération PEST) à différentes dates (ici 3 dates)
— structuration verticale (qualifiée ici de *monodimensionnelle*) : l'évolution de la station est vue à travers l'évolution de chacun des paramètres, c'est-à-dire par un ensemble de séries temporelles à valeurs numériques pouvant présenter des longueurs différentes (de 3 dates pour chaque pesticide à 8 dates pour NO_3^- sur la station 2).

L'algorithme de clustering mis en œuvre utilise DTW pour évaluer la similarité des séquences temporelles. Or, si DTW est capable de prendre en compte des séquences de longueurs différentes (Tormene et al., 2009), il faut néanmoins pouvoir calculer la distance entre chaque paire d'éléments de deux séquences. Or dans le cas multidimensionnel, il se peut que certains vecteurs d'état soient très clairsemés mais surtout, ils peuvent ne pas se recouvrir (les paramètres renseignés dans un des vecteurs ne sont pas les mêmes que ceux renseignés dans l'autre : $v_1 = (a, \_)$, $v_2 = (\_, b)$) ce qui interdit le calcul d'une distance. Ce cas s'est avéré relativement fréquent bien que non représenté dans le tableau 1.

Par ailleurs, la solution consistant à imputer les valeurs manquantes n'a pas été retenue du fait de la grande variabilité des valeurs (un paramètre peut augmenter brutalement et décroître tout aussi vite, suite à une pluie par exemple). Pour contourner ce problème, il a donc été choisi de travailler sur des séquences monodimensionnelles. Ainsi la distance entre deux stations sera obtenue en moyennant les distances calculées attribut par attribut.

4 Résultats

La méthode utilisée est le clustering collaboratif, une extension du clustering par ensemble, appelé aussi clustering par consensus (Cornuejols et al., 2017). Cela consiste à faire travailler en parallèle plusieurs agents de clustering avec leur propre méthode puis, lorsque chaque agent a établi son résultat, une méthode d'unification combine l'ensemble pour former un nouveau résultat. Lors de cette phase, chaque résultat est remis en cause à partir des informations contenues dans les autres résultats. Plus précisément nous avons utilisé la méthode SAMARAH (Gançarski et Wemmert, 2007) qui présente une architecture originale et générique de collaboration entre des classifieurs (ou agents). Elle est basée sur le principe d'un raffinement mutuel et itératif de plusieurs résultats de clustering jusqu'à obtenir des résultats "similaires" et de qualité. Le système peut être décomposé en trois grandes étapes :

1. Génération des résultats initiaux ;
2. Raffinement collaboratif des différents résultats ;
3. Unification par combinaison, des résultats raffinés.

Dans nos expériences, cette méthode a été configurée avec trois agents Kmeans et les caractéristiques suivantes : *i)* séquences monodimensionnelles (cf. section 3.3) avec distance par moyennage des distances DTW inter-attributs ; *ii)* 12, 14 et 16 graines initialement ; *iii)* 30 itérations à chaque application de Kmeans (initialement et lors du raffinement) ; *iv)* 10 clusters dans le résultat final (choix empirique).

Nous présentons ici les résultats concernant deux de ces expériences : les nitrates (noté NITR dans la suite) et le groupe 1 des pesticides (noté PEST-1 dans la suite), qui regroupe les

pesticides autorisés, solubles, toxiques à très toxiques, à savoir les 4 molécules bifénox, aclonifène, pendiméthaline et tébuconazole (voir le tableau 1). Ces deux altérations se distinguent d'une part, par leur constitution (1 paramètre *versus* 4) et d'autre part par leur complétude : le paramètre NO_3^- est très mesuré tandis que les pesticides le sont beaucoup moins.

Le tableau 3 présente les caractéristiques des classes obtenues pour l'altération NITR. Ces classes sont au nombre de 10. Leur taille varie de 5 à 55 stations avec une variation importante de l'inertie intra-classe, qui peut être forte même pour les petites classes (classe 3). La taille moyenne des séquences est plus homogène (de 39 à 65). Pour l'hydroécologue, il est intéressant de calculer aussi la valeur moyenne de l'altération pour chaque classe ainsi que les valeurs maximale et minimale, qui donnent l'enveloppe des séquences d'une classe.

Classes	Nombre Stations	Taille séquences	Moy* (CM)	Max* (VR)	Min* (VR)	Inertie intra pondérée**
1	29	65	5,90	15,30	1,00	213
2	28	48	4,36	12,00	0,50	117
3	13	52	7,38	66,10	0,50	109 167
4	5	46	1,12	5,00	0,50	24
5	46	53	15,53	53,00	0,50	21 361
6	9	41	3,71	33,40	0,50	4 386
7	48	46	11,87	42,20	0,50	1 614
8	23	50	25,09	59,00	2,60	228 786
9	55	53	8,04	26,80	0,40	883
10	47	39	2,87	9,00	0,60	9

TAB. 3 – *Caractéristiques des 10 classes obtenues par clustering de l'altération NITR (CM= courbe moyenne ; VR = valeurs réelles mesurées ; * moyenne, maximum et minimum exprimés en mg/L de nitrates ; ** pondération par le nombre de stations)*

4.1 Visualisation et interprétation thématique des résultats

Pour poursuivre l'analyse, l'expert peut évaluer la cohérence des classes en examinant leur dispersion interne. Pour faciliter sa tâche, nous proposons une visualisation *via* un positionnement multidimensionnel ou carte MDS (multidimensional scaling). Cette technique permet de représenter en deux ou trois dimensions les informations d'une matrice de distance (ou dissimilarité) à l'aide de modèles de distances spatiales. Le principe est de préserver les proximités entre objets (les stations) et non leurs valeurs exactes ou relatives. A partir des valeurs p_{ij} de la matrice, on détermine des distances $f(p_{ij})$, la fonction f vérifiant une propriété de monotonie : si $p_{ij} < p_{i'j'}$ alors $f(p_{ij}) \leq f(p_{i'j'})$. Un extrait d'une telle carte est présenté en figure 2 pour l'altération NITR. On y voit toutes les stations de classes dont l'inertie intra est faible (classes 1, 2, 9, 10) tandis que les classes de forte inertie sont très dispersées (hors de l'extrait présenté) : c'est la cas de la classe 3, dont seulement 8 des 13 stations sont visibles sur la carte. L'intérêt de cette carte est aussi de visualiser les stations qui sont en limite de classes et partagent moins de caractéristiques avec les stations centrales.

Cette présentation est insuffisante pour l'hydroécologue qui s'intéresse au comportement commun des stations d'une classe. Nous avons alors proposé de représenter la séquence moyenne de chaque classe pour chaque paramètre. La figure 3 présente par exemple les courbes des

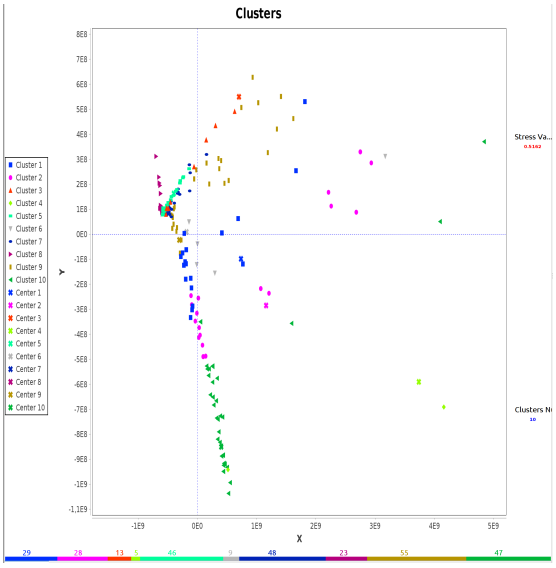

FIG. 2 – *Visualisation 2D de la distribution des classes pour l'altération NITR (extrait centré sur les classes de faible inertie-intra)*

5 classes les plus caractéristiques (au sens où elles ont des profils temporels très différenciés) parmi les 10 classes obtenues pour l'altération NITR. Pour faciliter l'interprétation sont aussi affichés les seuils de qualité du SEQ-eau, qui permettent de visualiser la position des classes en terme de qualité de l'eau. Ainsi la classe 4, en bleu, est toujours très bonne qualité (sous le seuil bleu). La classe 3, en noir, est la plupart du temps sous le seuil de bonne qualité (seuil vert) avec quelques pics dans la gamme de qualité moyenne (seuil jaune) et un pic dans la gamme de qualité médiocre. A l'inverse, la classe 8, en orange, se situe généralement dans la gamme de qualité moyenne et a des pics dans les gammes de qualité médiocre.

Ces courbes permettent également à l'expert de définir le type d'évolution du paramètre : stable (classes 4, en bleu, et 10, en vert), variable avec une certaine périodicité (classes 5, en jaune, et 8, en orange) à très variable parfois de façon erratique (classe 3 en noir). Les autres classes non représentées sur la figure 3 ont des évolutions proches de la classe 10, pour les classes 1 et 2, ou de la classe 5, pour les classes 6, 8 et 9, tout en ayant des valeurs de moins bonne qualité.

4.2 Cartographie

L'interprétation approfondie des classes nécessite une projection cartographique, qui permet de visualiser la localisation des classes par rapport aux informations contextuelles, telles que les hydroécorégions (HER), les cours d'eau, mais aussi l'occupation du sol : on utilise la

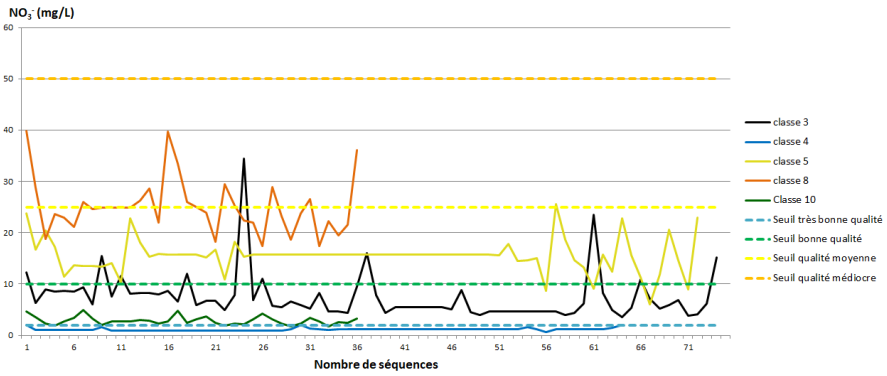

FIG. 3 – *Courbes moyennes pour les cinq classes les plus caractéristiques (NITR)*

BD Carthage [2] et la carte 2018 du programme Corine land Cover [3] qui permet de distinguer zones urbaines, zones agricoles et autres.

Ainsi la carte de la figure 4 (gauche) représente la localisation des stations des 5 classes caractéristiques de l'altération NITR. Elle permet de compléter l'interprétation de l'expert : les classes 4 et 10, où les concentrations moyennes de nitrates sont faibles et évoluent peu, sont localisées dans des zones de montagne (Alpes, Vosges) pas ou peu cultivées. Les stations des classes 5 et 8, où les concentrations sont moyennes à élevées, variables avec une certaine périodicité, sont situées dans des zones cultivées et sont alors affectées par des lessivages d'engrais sur terres agricoles, et/ou situées sur des cours inférieurs de rivières et sont alors affectées par des rejets cumulés de stations d'épuration, qui finissent par enrichir les eaux en nitrates malgré le respect des normes de rejets. La raison de l'existence de chaque classe est identifiable et interprétable par l'expert, mais le clustering lui permet de rapidement différencier les stations. La classe 3 est particulièrement intéressante : elle regroupe des stations globalement de bonne qualité, mais qui subissent des pics ponctuels de pollution aux nitrates, dont l'origine resterait à définir, en étudiant chaque station de cette classe plus précisément.

La carte 4 (droite) présente la localisation des classes de l'altération PEST-1. L'étude de cette carte combinée à celle des courbes moyennes permet à l'expert de comprendre que la majorité des stations ne présentent pas de pollutions avec ces molécules pour 6 des 10 classes obtenues (classes 1, 2, 3, 4, 5 et 10). La distinction entre ces classes apparaît liée à des techniques analytiques différentes (dans le temps et selon les opérateurs). En revanche, 4 classes regroupent des stations situées dans des zones agricoles et polluées par l'une ou l'autre de ces molécules qui sont liées au type de cultures en place. Les stations de la classe 6 sont polluées en tébuconazole, fongicide à large spectre, utilisé en viticulture, mais aussi par les céréaliers et les maraîchers. Les stations de la classe 7 sont polluées en aclonifène, herbicide utilisé sur les cultures de protéagineux. Les stations de la classe 8 sont polluées en aclonifène et tébuconazole. La classe 9 résulte d'un artefact, elle regroupe des stations qui ont été très peu mesurées.

2. https://www.data.gouv.fr/fr/datasets/cours-deau-metropole-2016-bd-carthage/
3. https://www.data.gouv.fr/fr/datasets/corine-land-cover-occupation-des-sols-en-france

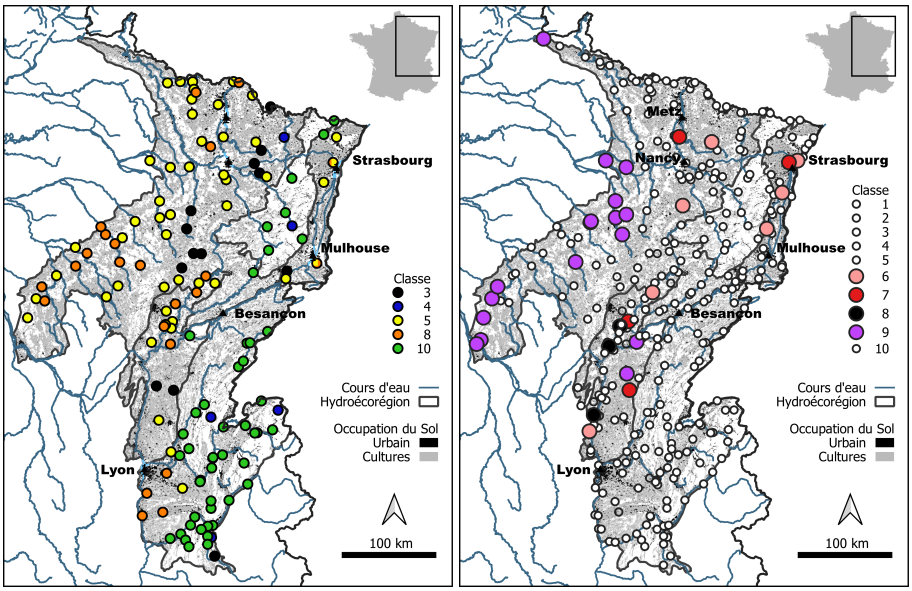

FIG. 4 – *Localisation géographique des classes pour l'altération NITR (à gauche) et PEST-1 (à droite)*

5 Conclusion et perspectives

Pour les hydroécologues impliqués dans la projet, l'intérêt premier est de réaliser l'analyse des données en s'extrayant des contraintes de discrétisation, utilisées habituellement dans le cadre de la grille SEQ-Eau. De plus la méthode permet d'exploiter de longues chroniques de données. La contrepartie est la limitation du nombre de paramètres pris en compte dans une classification. Grâce à cette collaboration, et par effet retour aux informaticiens, de nouveaux verrous scientifiques ont été levés et des outils génériques (sélections d'attributs, visualisation temporelle, ...) ont été développés et intégrés dans la plateforme FODOMUST, montrant une nouvelle fois les bénéfices de l'interaction entre science des données et producteurs-analystes de données « réelles ».

La perspective immédiate concerne l'analyse de tous les paramètres physico-chimiques sélectionnés et leur mise en regard de l'état écologique en exploitant les suivis biologiques produits sur les mêmes stations. Ensuite nous mettrons en œuvre des méthodes de clustering sous contraintes qui permettront d'affiner les clusters afin de tendre vers des résultats plus proches des intuitions des experts (c'est-à-dire, des classes thématiques potentielles) et ainsi d'autoriser et simplifier la mise en évidence des types d'évolution de l'état des cours d'eau.

Remerciements. Le projet ADQEAU est financé par le conseil scientifique de l'ENGEES. Plusieurs stagiaires y ont participé : Sylvain Zongo (M2 informatique, IFI Vietnam Nat. University & Université de la Rochelle), Pascal Ung et Gabriel Honda (3ème année ingénieur, ENGEES). Nous remercions Xavier Dolques (ENGEES) pour la préparation des données.

Références

Bimonte, S., K. Boulil, A. Braud, S. Bringay, F. Cernesson, X. Dolques, M. Fabrègue, C. Grac, N. Lalande, F. Le Ber, et M. Teisseire (2015). A decisional system for analysing water quality of watercourses. *RSTI - Ingénierie des Systèmes d'Information 20*(3), 143–167.

Cornuejols, A., C. Wemmert, P. Gançarski, et Y. Bennani (2017). Collaborative clustering : Why, when, what and how. *Information Fusion 39*, 81–95.

Dolques, X., F. Le Ber, M. Huchard, et C. Grac (2016). Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. *International Journal of General Systems 45*(2), 187–210.

Fabrègue, M., A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, et M. Teisseire (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics 24*, 210–221.

Gançarski, P., B. Lafabregue, A.-D. Salaou, et H. Vernier (2020). FoDoMuST - une plateforme de clustering collaboratif sous contraintes incrémental de séries temporelles. In *EGC (RNTI Volume E-36)*, pp. 507–514.

Gançarski, P. et C. Wemmert (2007). Collaborative multi-step mono-level multi-strategy classification. *MTAP 35*(1), 1–27.

Ministère de la Transition Ecologique et Solidaire (2019). Guide technique relatif à l'évaluation de l'état des eaux de surfaces continentales (cours d'eau, canaux, plans d'eau).

Petitjean, F., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognit 44*(3), 678–693.

The European Parliament and the Council (2000). Framework for Community action in the field of water policy. Directive 2000/60/EC.

Tormene, P., T. Giorgino, S. Quaglini, et M. Stefanelli (2009). Matching incomplete time series with dynamic time warping : an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine 45*(1), 11–34.

Ung, P. (2020). Classification et analyse des stations de mesure en rivière à partir des séquences temporelles de certaines altérations physico-chimiques. Mémoire de fin d'étude AgroSup Dijon et ENGEES.

Wasson, J.-G., A. Chandesris, H. Pella, et L. Blanc (2004). Les hydro-écorégions : une approche fonctionnelle de la typologie des rivières pour le directive cadre européenne sur l'eau. *Ingénierie 40*, 3–10.

Summary

This article is about a collaborative process and tools we have built to adapt a clustering method for analysing temporal sequences of physico-chemical measurements done on river streams. These data are characterised by sampling variability and a great number of parameters, that are monitored in different ways. The dataset is thus heterogeneous and incomplete. A subset of about 300 sequences was selected and analysed with a specific clustering method for temporal data. Results are presented and commented, through adapted visualisations.