Document public

Revue des méthodes de décomposition des chroniques de qualité des eaux souterraines

Rapport final

BRGM/RP-65479-FR
Janvier 2016

ONEMA
Office national de l'eau et des milieux aquatiques

Géosciences pour une Terre durable
brgm

Document public

Revue des méthodes de décomposition des chroniques de qualité des eaux souterraines

Rapport final

BRGM/RP-65479-FR
Janvier 2016

Étude réalisée dans le cadre des opérations de Service public du BRGM fiche 14OBS0830

N. Croiset, B. Lopez, N. Surdyk

Vérificateur :	**Approbateur :**
Nom : Jean-Jacques Seguin	Nom : L. Chery
Fonction : Ingénieur	Fonction: Responsable d'unité
Date :	Date : 20/07/2016
Signature **:**	Signature **:**

Le système de management de la qualité et de l'environnement est certifié par AFNOR selon les normes ISO 9001 et ISO 14001.

ONEMA
Office national de l'eau et des milieux aquatiques

brgm
Géosciences pour une Terre durable

RESUME

Les concentrations en éléments chimiques dans les eaux souterraines évoluent selon trois facteurs principaux : (i) l'évolution temporelle de l'émission de l'élément considéré vers le milieu souterrain ; (ii) l'évolution des conditions hydroclimatique et hydrogéochimique dans lesquelles il évolue; (iii) ses caractéristiques intrinsèques et les performances atteintes pour son analyse. Des études antérieures sur l'analyse des chroniques d'évolution des concentrations en nitrate et en pesticides dans les eaux souterraines ont montré que ces signaux peuvent être « modélisés » mathématiquement selon différentes composantes qui expliquent chacune une partie du signal total. Ces composantes dessinent globalement 3 grands types de comportements :

- l'évolution plus ou moins erratique de la chimie des eaux = évolution considérée comme aléatoire ;
- l'évolution des concentrations structurée dans le temps suivant des grandes tendances linéaires ou non-linéaires = évolution non-stationnaire, par exemple une évolution tendancielle ;
- l'évolution des concentrations structurée dans le temps suivant des cycles périodiques (cycle annuel et/ou pluriannuel) = évolution cyclique.

Chaque type de comportement pouvant être expliqué par des facteurs anthropiques ou naturels spécifiques, il apparait pertinent d'extraire les différentes composantes des chroniques et ainsi estimer sur quelle part de l'évolution de la qualité des eaux souterraines le gestionnaire pourra agir par la mise en œuvre de plans de gestion.

Des méthodes mathématiques et statistiques de décomposition des signaux temporels ont alors été recherchées en bibliographie environnementale. Parmi les méthodes étudiées, seules celles pouvant être « facilement » appliquées sur les chroniques d'évolution de la qualité des eaux souterraines ont été détaillées et critiquées. Elles ont été classées en deux grandes catégories : (i) les méthodes de régression et de stationnarisation des séries temporelles et (ii) les méthodes d'identification des variations cycliques. Chaque méthode nécessite plus ou moins de prétraitements des chroniques pour être appliquée de manière optimale. Elles apparaissent aussi plus ou moins puissantes en fonction des conditions d'application. Des critères de facilité d'utilisation et de puissance[1] vis-à-vis des conditions initiales ont permis de recommander l'utilisation de certaines méthodes spécifiques pour décomposer les chroniques d'évolution de la qualité des eaux souterraines.

Ainsi, en complément des tests de Mann-Kendall et Mann-Kendall modifié déjà recommandés et implémentés dans l'outil HYPE pour l'identification des tendances monotones, l'algorithme LOWESS apparait pertinent pour la régression de tendances complexes non monotones et la stationnarisation des chroniques d'évolution de la qualité des eaux souterraines. Des tests sur des données réelles ont révélé la puissance de cette méthode, notamment lorsqu'elle est couplée à l'utilisation du test de Mann-Kendall. Le paramétrage de la fenêtre de lissage, spécifique à chaque chronique analysée et à l'objectif recherché, limite néanmoins l'automatisation complète d'une telle méthode et seule une semi-automatisation peut être envisagée.

Pour l'identification des variations cycliques, le périodogramme de Lomb-Scargle est apparu approprié pour une application au domaine de la qualité des eaux souterraines. Cette méthode peut en effet être appliquée sur des données non régulièrement espacées et sa puissance

[1] La puissance d'un test statistique est son aptitude à mettre en évidence un effet si celui-ci existe.

apparait tout à fait convaincante, notamment vis-à-vis des faibles fréquences de prélèvements. Un arbre décisionnel précisant la ou les méthodes à appliquer en fonction de la composante à extraire a été construit afin d'aider les opérateurs à appliquer une procédure complète et robuste de décomposition des chroniques d'évolution de la qualité des eaux souterraines.

Mots-clés : Analyse statistique ; Analyse du signal ; Qualité eau ; Eau souterraine ; Tendance ; DCE

Couverture géographique : Nationale
Niveau de lecture : Expert

CORRESPONDANTS ONEMA : STAUB Pierre-François

En bibliographie, ce rapport sera cité de la façon suivante :

Croiset N., Lopez B., Surdyk N. (2016) **–** Revue des méthodes de décomposition des chroniques de qualité des eaux souterraines. BRGM/RP-65479-FR, 56 p., 22 ill.

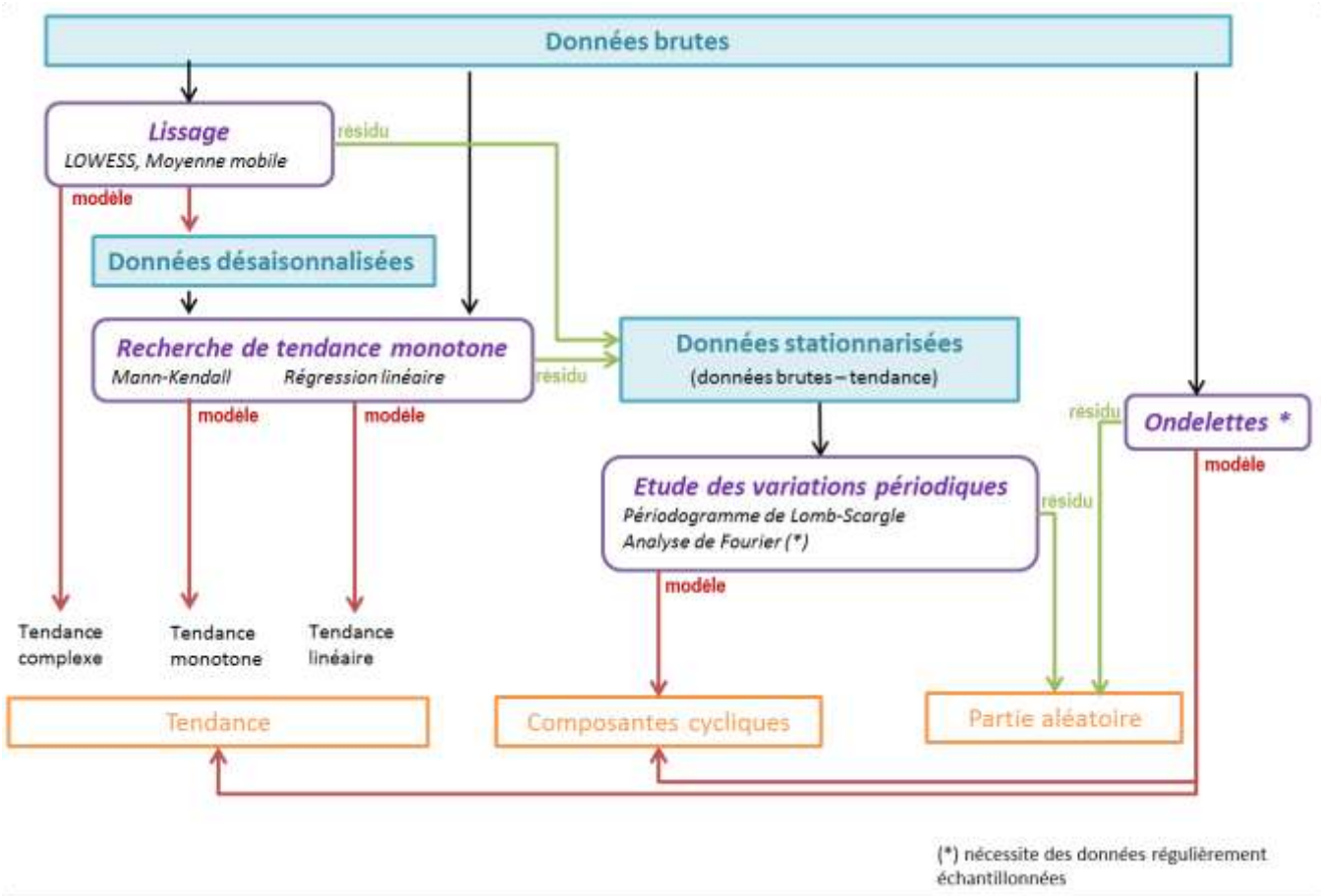
Synthèse pour l'action opérationnelle

Le présent rapport fait état des recherches effectuées par le BRGM sur les méthodes de décomposition des chroniques d'évolution de la qualité des eaux souterraines. Ces travaux ont été réalisés dans le cadre de la convention entre l'ONEMA et le BRGM sur la période 2013-2015. Ces actions s'inscrivent pleinement dans le cadre de la Directive 2000/60/CE et la Directive eaux souterraines 2006/118/CE qui imposent notamment aux Etats Membres d'identifier les tendances et les inversions de tendances d'évolution de la qualité des eaux souterraines. L'objectif final pour les gestionnaires est d'inverser les tendances significatives identifiées à la hausse durable par la mise en œuvre de plans de gestion de la ressource et des pressions qui s'y excercent.

A fin d'aider les gestionnaires dans ces travaux, un outil d'identification des tendances et des inversions de tendances significatives – HYPE – avait précédemment été développé conjointement avec l'ONEMA et mis à disposition des gestionnaires en 2013. Les travaux détaillés dans le rapport permettent d'aller plus loin dans l'étude des chroniques de la qualité des eaux souterraines isues des programmes de surveillance réglementaire. Il s'agit, par l'intermédiaire d'une revue des méthodes statistiques existantes, d'aider à décomposer les séries temporelles afin d'identifier et extraire la part du signal correspondant à l'incertitude analytique et d'échantillonnage, celle correspondant aux variations environnementales et celle attribuable à l'évolution des activités anthropogéniques. Ces informations prennent une importance toute particulière lors de la définition des plans de mesure car elles permettent de prédire sur quelle part du signal il serait possible d'agir à court terme et donc quel impact sur la qualité de ressource peut être espéré suite à des actions menées sur les activités anthropogéniques en entrée des systèmes hydrogéologiques.

Les méthodes de décomposition/déconvolution des séries temporelles ont été recherchées dans la littérature du domaine de l'environnement en général. Elles ont été triées en fonction de leur capacité et leur facilité à être appliquées sur les chroniques issues des programmes de suivis de la qualité des eaux souterraines aux caractéristiques spécifiques. Elles ont été classées en deux grandes catégories : (i) les méthodes de régression et de stationnarisation des séries temporelles et (ii) les méthodes d'identification des variations cycliques.

Les méthodes pré-sélectionnées nécessitent plus ou moins de prétraitements pour être appliquées et montrent des puissances variables. Ces critères de facilité d'utilisation et de puissance vis-à-vis des conditions initiales ont permis de recommander l'utilisation de certaines méthodes spécifiques pour décomposer les chroniques d'évolution de la qualité des eaux souterraines. Un arbre décisionnel précisant la ou les méthodes à appliquer en fonction de la composante à extraire a été construit afin d'aider les opérateurs à appliquer une procédure complète et robuste de décomposition des chroniques d'évolution de la qualité des eaux souterraines. L'arbre présenté ci-dessous liste les méthodes à appliquer dans l'ordre défini par une lecture de haut en bas. L'application de la procédure complète permet d'extraire 3 composantes principales : la composante tendancielle (monotone, linéaire ou complexe), les cycles périodiques et la composante aléatoire du signal.



Arbre décisionnel pour la décomposition des chroniques de qualité des eaux souterraines selon les méthodes statistiques recommandées par Croiset et al. 2016.

Les formulations des méthodes recommandées dans l'arbre décisionnel sont détaillées dans le rapport ainsi que les avantages et les inconvénients de chacune. La procédure complète de décomposition des chroniques nécessitant l'expertise de l'opérateur à différentes étapes de son application, il n'est pas possible de l'automatiser intégralement. On pourra en revanche se poser la question de l'automatisation de certaines méthodes intermédiaires comme le lissage de LOWESS, pratique pour l'identification de tendances complexes. Cette méthode pourrait, à terme, intégrer l'outil HYPE dédié à l'identification des tendances d'évolution de la qualité des eaux souterraines.

Sommaire

Liste des illustrations

1. Introduction

1.1. CONTEXTE DE L'ACTION

Cette action fait suite à plusieurs années d'études sur l'identification et la caractérisation des tendances menées dans le cadre de la convention ONEMA-BRGM. En 2011, Lopez et Leynet (RP-59515-FR) ont réalisé une revue bibliographique des méthodes statistiques sur l'identification des tendances d'évolution des contaminants dans les eaux souterraines. Ce travail a permis en 2013 la réalisation de fiches synthétisant les tendances d'évolution des nitrates pour toutes les masses d'eau souterraine et le développement de l'outil HYPE qui permet d'identifier les tendances de manière automatique (Lopez et al, 2013, RP-61855-FR ; Croiset et Lopez, 2013, RP-63066-FR). En 2014, le travail a été poursuivi en testant les capacités des réseaux de contrôle de surveillance et opérationnels de la qualité des eaux souterraines actuels (RCS et RCO respectivement) à fournir des données permettant d'identifier les tendances d'évolution des contaminants (Lopez et al., 2014, RP-63845-FR).

L'identification des tendances ne peut toutefois constituer une fin en soi. Cet exercice nécessite d'être poursuivi par l'identification des phénomènes qui guident les évolutions temporelles de la chimie des eaux souterraines. Connaître les facteurs qui expliquent les tendances permet en effet d'identifier les leviers d'actions possibles pour inverser les hausses et ainsi définir les actions de gestion les plus pertinentes à mettre en œuvre. L'identification des facteurs explicatifs des tendances observées est souvent réalisée à dire d'expert en croisant de multiples informations sur les contextes hydroclimatiques, les fonctionnements hydrogéochimiques des systèmes aquifères et les pressions anthropiques notamment. Ces opérations souvent complexes peuvent être facilitées par le développement d'outils d'aide à l'analyse des signaux environnementaux qui portent l'information à extraire pour comprendre les évolutions temporelles de la chimie des eaux souterraines.

1.2. OBJECTIF DE L'ETUDE

L'objectif de l'étude est d'identifier des méthodes qui permettent de décomposer les séries temporelles, afin d'extraire les structures sous-jacentes aux chroniques d'évolution de la chimie des eaux souterraines (cyclicités, tendances, ruptures et partie aléatoire liée aux incertitudes d'échantillonnage et d'analyse). Ces « sous-signaux » pourraient ensuite être comparés individuellement à d'autres données de contexte (pluie, piézométrie, occupation du sol, historique d'application d'une substance…) afin d'aider à reconnaître les phénomènes qui influencent les différentes composantes du signal global. Cette connaissance plus précise des structures qui composent la chronique globale et des phénomènes qui guident l'évolution de chacune d'entre elles permettrait d'affiner les prises de mesures en ciblant les phénomènes responsables des évolutions observées, identifiant ceux sur lesquels il est possible d'agir et en estimant les effets potentiels des changements imposés.

Des méthodes de décomposition des séries temporelles ont ainsi été recherchées dans la littérature du domaine de l'environnement en général. Les méthodes présentées dans ce rapport sont celles qui apparaissent comme pertinentes pour une application aux chroniques d'évolution de la qualité des eaux souterraines – dont les caractéristiques sont rappelées en première partie du document. Ces méthodes permettent, soit d'extraire directement une ou plusieurs composantes des chroniques d'évolution de la chimie des eaux souterraines à expliquer par des phénomènes externes, soit de transformer les chroniques brutes dans le but d'appliquer des méthodes statistiques sensibles à la non-stationnarité ou à l'autocorrélation des

données. Des exemples d'application sont donnés pour les méthodes plus spécifiquement recommandées.

2. Les chroniques d'évolution de la qualité des eaux souterraines

2.1. CARACTERISTIQUES DES DONNEES DE QUALITE DES EAUX SOUTERRAINES

Une chronique d'évolution (ou série temporelle) de la qualité des eaux souterraines correspond à la suite chronologique des résultats des analyses chimiques réalisées sur des échantillons prélevés en un même point. La chronique est constituée de valeurs discrètes de concentration en un élément chimique auxquelles sont associées des dates de prélèvement. La chronique supporte l'information à partir de laquelle on estime l'évolution générale de l'état chimique des eaux souterraines. Elle est caractérisée par une longueur (date de début et date de fin) et une fréquence de prélèvement (nombre de prélèvement par unité de temps). Les chroniques rencontrées dans le domaine de la qualité des eaux souterraines présentent souvent des caractéristiques restreignant grandement l'utilisation des méthodes statistiques conventionnelles. Ces caractéristiques sont décrites dans les paragraphes 2.1.1 à 2.1.5.

2.1.1. Données non régulièrement espacées

Les chroniques de qualité des eaux souterraines ne contiennent que très rarement des données parfaitement régulièrement espacées. Les pas de temps entre deux observations (eg. prélèvements) sont en général très variables. Ceci peut poser problème pour l'application de certains tests statistiques, notamment pour les calculs d'autocorrélation et de corrélation croisée (qui se base sur la comparaison d'une chronique avec elle–même ou avec une autre chronique, décalée d'un pas de temps donné) ou pour les analyses spectrales de type Fourier.

Une solution fréquemment trouvée dans la littérature pour traiter des données non régulièrement espacées est de ré-échantillonner le jeu de données en interpolant avec des pas de temps réguliers. Même si cette approche est souvent utilisée, il est recommandé de l'éviter autant que possible car elle peut introduire des biais importants dans les tests statistiques. Plusieurs autres méthodes, dépendant du test statistique utilisé, sont présentées dans ce travail pour traiter les données avec un pas d'échantillonnage variable.

2.1.2. Présence de données censurées

Pour bon nombre de paramètres chimiques, les concentrations mesurées dans les eaux souterraines sont souvent inférieures aux limites de quantification. La présence de ces données « censurées » dans une série temporelle rend difficile l'évaluation des paramètres statistiques conventionnels (moyenne, pente…) et peut conduire à des erreurs importantes.

Ce problème peut être contourné en faisant appel aux méthodes statistiques non paramétriques, qui sont relativement insensibles à une censure, même sévère, des données.

2.1.3. Chroniques non stationnaires

Les chroniques de qualité des eaux souterraines ne sont pas toujours stationnaires (au sens stationnaire d'ordre 2, cf. 2.3). En effet, elles peuvent comporter une tendance à long terme ou être affectées de variations cycliques de période proche de la longueur totale de la chronique.

La non-stationnarité d'une chronique peut empêcher que certains tests statistiques fournissent des résultats corrects : par exemple, la présence d'une tendance à long terme dans une chronique affectera les résultats des tests d'analyse spectrale qui permettent d'étudier les variations cycliques. Dans ce cas, on va chercher, après avoir étudié les tendances, à stationnariser la chronique.

2.1.4. Données autocorrélées

Les chroniques de qualité des eaux souterraines peuvent présenter des données autocorrélées, notamment lorsque la fréquence d'échantillonnage est élevée. L'autocorrélation peut être due à une tendance dans la chronique temporelle ou à la « mémoire » des processus hydrogéologiques. L'autocorrélation des données peut biaiser les résultats de certains tests statistiques.

Pour étudier l'autocorrélation liée à l'inertie d'un système, on cherchera dans un premier temps à supprimer l'autocorrélation liée à la présence d'une tendance. Si au contraire, on veut qualifier une tendance, il est nécessaire de prendre en compte l'effet de l'autocorrélation des données liée à l'inertie du système.

Il est également important d'estimer l'autocorrélation d'une série temporelle car certaines méthodes statistiques ne s'appliquent pas ou demandent des ajustements si les données sont autocorrélées.

2.1.5. Données non normalement distribuées

Très souvent les chroniques de qualité des eaux souterraines sont composées de données non normalement distribuées. Cela est le cas notamment, des chroniques comportant des données inférieures à la limite de quantification (« données censurées ») ou des valeurs extrêmes. Certains tests statistiques ne peuvent pas être appliqués, ou sont très peu puissants si les données ne sont pas normalement distribuées.

Les tests de normalité permettent de déterminer si les données peuvent être modélisées par une loi normale.

2.2. COMPOSANTES DES CHRONIQUES D'EVOLUTION DE LA QUALITE DES EAUX SOUTERRAINES

Des études antérieures menées par le BRGM et les Agences de l'eau Loire-Bretagne et Seine-Normandie sur les tendances d'évolution des nitrates et des produits phytosanitaires dans les eaux souterraines (Baran et al., 2009 ; Lopez et al., 2012) ont montré, qu'en plus des évolutions plus ou moins erratiques de la chimie des eaux souterraines, dans certains contextes particuliers, les concentrations en certains éléments, comme les nitrates et certains pesticides par exemple, pouvaient évoluer selon des cycles périodiques. Ces cycles s'établissent généralement sur des périodes annuelles et/ou pluriannuelles (de 6 à 12 ans) qui suivent les cycles hydroclimatiques. Une analyse plus poussée de ces chroniques via des méthodes géostatistiques a montré la possibilité de décrire la distribution des données d'évolution des concentrations dans les eaux souterraines par des variogrammes théoriques à cycles plus ou moins longs (Lopez et al., 2015). Ces modèles mathématiques d'évolutions théoriques de la chimie des eaux souterraines dessinent globalement 3 grands types de comportements :

- l'évolution plus ou moins erratique de la chimie des eaux = évolution aléatoire (au sens où l'on ne peut pas connaître les déterminants) ;

- L'évolution des concentrations structurée dans le temps suivant des grandes tendances plus ou moins linéaires = évolution non-stationnaire ou évolution tendancielle;
- l'évolution des concentrations structurée dans le temps suivant des cycles périodiques = évolutions à cycles annuels, à cycles pluriannuels et à double cycles annuels et pluriannuels.

Or on estime que l'évolution temporelle de la chimie des eaux souterraines est fonction de trois facteurs principaux : (i) les caractéristiques intrinsèques de l'élément considéré et les performances atteintes pour son analyse ; (ii) l'évolution temporelle de son émission vers le milieu souterrain; (iii) les contextes hydroclimatique et hydrogéologique dans lesquels elle évolue. Ces facteurs impactent de manière plus ou moins spécifique une ou plusieurs des trois composantes précédemment citées.

Le premier type de comportement – l'évolution monotone – correspond plutôt aux éléments dont les concentrations dans le milieu souterrain sont peu affectées par les évolutions hydroclimatiques c'est-à-dire par les cycles de recharge et de vidange des nappes. Ceci peut être dû soit aux caractéristiques intrinsèques de la molécule dont la réactivité dans le milieu (adsorption et dégradation essentiellement) affecte plus les concentrations que les changement dans le milieu dans lequel elle est dissoute, soit aux périodes et aux volumes des apports de la molécule dans le milieu qui guident de manière prépondérante l'évolution des concentrations dans l'aquifère. Entrent dans ces deux cas de figure les micropolluants organiques actifs du type pesticides et médicaments, les éléments minéraux aux fortes capacités d'adsorption et tous les contaminants émis de manière ponctuelle (dans le temps et dans l'espace) comme les polluants industriels par exemple. Les concentrations en ces éléments évoluent en effet généralement de manière assez rapide et brutale dans les aquifères en formant des signaux aléatoires et non structurés dans le temps.

Le deuxième type de comportement – l'évolution non stationnaire ou tendancielle – correspond généralement à une évolution non naturelle de la chimie des eaux souterraines. Ce comportement peut être dû, soit à une évolution de l'émission des éléments chimiques en entrée des systèmes aquifères, soit à des changements graduels des conditions physi-chimiques qui règnent dans l'aquifère. Ces changements sont souvent induits par des phénomènes externes généralement anthropiques (augmentation des pompages ou acidification des sols par exemple). Les éléments qui répondent à ce type de comportement ont donc soit une origine anthropique, soit sont impactés par des activités anthropiques. Il est donc très pertinent d'identifier la composante tendancielle des chroniques d'évolution de la qualité des eaux souterraines car c'est généralement sur cette partie de l'évolution qu'il sera possible d'intervenir par des changements de pratiques.

Le second type de comportement – l'évolution cyclique périodique – se rapporte aux éléments dont les concentrations dans les eaux souterraines évoluent suivant les fluctuations hydroclimatiques (alternance de phases de fortes pluies efficaces et de périodes sèches) et/ou hydrodynamiques de l'aquifère (hausses et baisses des niveaux piézométriques). Ce type de comportement chimique nécessite que l'élément considéré soit stable en milieu aqueux (ne s'adsorbe et/ou ne se dégrade pas ou très peu) ou bien qu'il soit présent dans le milieu en quantité suffisante pour que les phénomènes d'atténuations naturelles aient peu d'impact sur les valeurs de concentrations. C'est par exemple notoirement le cas des nitrates d'origine agricole, polluant diffus émis depuis plusieurs décennies dans l'environnement et dont les concentrations dans les eaux souterraines, même si des phénomènes de dénitrification peuvent exister, sont généralement guidées par les fluctuations des niveaux des nappes et des débits des sources. Ce cas de figure concerne aussi généralement les paramètres physico-chimiques des nappes comme la conductivité par exemple. On distingue alors :

- les aquifères très réactifs aux variations saisonnières des pluies efficaces qui, lorsqu'ils sont contaminés par des éléments stables dans l'eau, induisent des variations saisonnières (cycles annuels) des concentrations. C'est généralement le cas des aquifères à nappe libre tels que les aquifères karstiques ou les aquifères à faible épaisseur de zone non saturée (ZNS) par exemple ;
- les aquifères dits inertiels, peu réactifs aux variations saisonnières de recharge mais sensibles aux grandes oscillations climatiques globales comme la NAO (North Atlantic Oscillation). L'aquifère de Beauce est un bon exemple d'aquifère à comportement temporel inertiel.

Les évolutions des concentrations en eaux souterraines peuvent alors être corrélées aux fluctuations du niveau des nappes ou bien, à l'inverse, anticorrélées. En effet, si la nappe est moins concentrée en contaminant que le sol ou la ZNS, une arrivée dans l'aquifère d'eaux récentes ayant lessivé le sol (cas des karsts notamment) ou une remontée de nappe dans la ZNS provoquera une augmentation des concentrations mesurées dans les eaux souterraines. Ce schéma correspond à une corrélation positive entre les évolutions du niveau des nappes et les concentrations en eaux souterraines. A l'inverse, si la nappe est plus concentrée en contaminants que le sol ou la ZNS, une arrivée d'eaux récentes provoquera une dilution et donc une baisse des concentrations, anticorrélées avec les fluctuations piézométriques ou les débits des sources.

Ces différentes composantes qui structurent les chroniques temporelles d'évolution de la qualité des eaux souterraines : les tendances à long terme, les variations cycliques et les variations aléatoires, méritent d'être extraites et analysées individuellement afin d'affiner la prévision de la contamination des eaux souterraines, anticiper les problèmes et, lorsque cela est possible, mettre en place des mesures appropriées.

2.3. DEFINITIONS

- Dans la suite de ce travail, on représentera **une série temporelle** de N données de qualité des eaux souterraines par $x(n)$ où $n = 0,1,2,\dots,N$

- Pour l'étude, **la tendance** est définie comme l'évolution générale non périodique d'une série temporelle sur une certaine période. La définition d'une tendance dépend fortement de la longueur de la série temporelle : ce qu'on peut identifier comme une tendance sur une série assez courte peut être vue comme une partie d'évolution cyclique si l'on regarde la série sur une période plus longue.

- Une tendance simplement croissante ou décroissante est dite **monotone**. Dans le cas contraire la tendance est qualifiée de **complexe**. Lorsque cela est possible, on essaiera de décrire la tendance d'une façon déterministe par une fonction mathématique, la plus simple étant l'évolution **linéaire**.

- Les **résidus** sont, pour un modèle, les différences entre les valeurs modélisées et les valeurs observées ; ils correspondent à la partie de l'information non expliquée par le modèle.

- Une fonction aléatoire est dite **stationnaire d'ordre 2** si et seulement si ses moments d'ordre 1 (moyenne) et d'ordre 2 (autocovariance) sont invariants par translation dans le temps.
  Dans les séries de qualité des eaux souterraines, la non-stationnarité peut être due à la présence de tendances à moyen ou long terme ou à des cycles de très basse

fréquence. Pour certains tests statistiques, un traitement préalable est nécessaire pour soustraire les grandes tendances et donc satisfaire la condition de stationnarité. Cette opération est délicate : un mauvais choix de la méthode de filtrage peut aboutir à un signal ayant des caractéristiques différentes de celles du signal d'origine.

- Un processus aléatoire est dit **ergodique** si l'on considère que l'on peut déterminer ses propriétés statistiques à partir d'une seule réalisation du processus (c'est-à-dire à partir de la série que l'on étudie).

- La **régularité** signifie l'équidistance des données. De nombreuses méthodes statistiques nécessitent des données régulièrement espacées. Cette condition n'est quasiment jamais remplie pour les chroniques de qualité des eaux souterraines.

- La fonction d'**autocorrélation** permet de détecter des régularités ou des profils répétés dans un signal.

- **La transformée de Fourier** est une fonction qui permet l'analyse en fréquence d'une série non périodique.

3. Régressions et Stationnarisation d'une série

La recherche de tendance dans une chronique de qualité des eaux souterraines a fait l'objet d'un rapport (Lopez et Leynet, 2011) faisant l'inventaire des méthodes statistiques existantes qui peuvent être appliquées sur les données de qualité des eaux souterraines. Un outil d'identification des tendances d'évolution des éléments chimiques dans les eaux souterraines – HYPE – a été développé suite à ce travail (Croiset et Lopez, 2013) et est disponible en téléchargement en accompagnement de sa notice d'utilisation à l'adresse : http://infoterre.brgm.fr/rapports/RP-63066-FR.pdf. Dans l'optique de décomposer les chroniques de qualité des eaux souterraines, d'autres méthodes ont été recherchées en complément de celles implémentées dans l'outil HYPE. Ces méthodes permettent notamment d'analyser les tendances complexes (non monotones).

3.1. OBJECTIFS

La recherche de tendances dans une chronique est nécessaire pour évaluer l'état chimique des masses d'eau. Le sens et la valeur de la pente sont des résultats directement utilisés pour le calcul de l'état chimique. Il peut toutefois être intéressant de rechercher les tendances dans une chronique afin de satisfaire un autre objectif : extraire les tendances à long terme d'une série temporelle afin de la stationnariser. En effet, la stationnarité d'une chronique est une condition requise pour l'application de plusieurs méthodes statistiques de décomposition d'une chronique, notamment le calcul de l'autocorrélation des données ou l'analyse spectrale (basée sur la transformée de Fourier de la série) qui permettent d'identifier les composantes cycliques. La mise en œuvre de ces méthodes nécessitent d'identifier et de supprimer d'éventuelles tendances avant de pouvoir décomposer les chroniques analysées. Le travail est alors mené sur les résidus stationnaires de la chronique brute non stationnaire.

La recherche de tendance, ou l'analyse de la stationnarité, est donc la première étape de l'étude d'une série de qualité des eaux souterraines. Il existe de nombreux tests statistiques qui permettent d'étudier la stationnarité d'une chronique[2] : Mann-Kendall, test KPSS, test de Leybourne et McCabe, test de Phillips et Peron, test de Dickey et Fuller, test de Dickey et Fuller augmenté (=test ADF)..

Dans le cas où une évolution monotone est attendue, il est proposé d'appliquer une régression linéaire ou une régression de Sen accompagnée d'un test de stationnarité de Mann-Kendall. Ces méthodes, déjà présentées dans les rapports précédents (Lopez et al., 2013; Croiset et Lopez, 2013), sont rappelées dans le paragraphe 3.3.

Si les tendances d'évolution attendues sont plus complexes (non-monotones), il est proposé d'utiliser, soit des méthodes non paramétriques comme les régressions locales ou les lissages selon les algorithmes de LOESS et LOWESS (Cleveland, 1979), soit des méthodes paramétriques comme le lissage par splines. La mise en œuvre, les avantages et inconvénients de ces différentes méthodes sont détaillés dans le paragraphe 3.4.

[2] Par exemple : Mann-Kendall, ou encore test KPSS, test de Dickey et Fuller, test de Dickey et Fuller augmenté (=test ADF) (le lecteur pourra se reporter à l'article de Kwiatkowski et al (1992) pour une description détaillée de ces derniers tests.

3.2. REGRESSION LINEAIRE

Ce test ne permet de détecter qu'une tendance d'évolution linéaire dans une chronique temporelle. Il a l'avantage d'être très puissant même avec peu de données. Il est néanmoins sensible aux valeurs extrêmes. La normalité des données n'est pas une condition nécessaire à l'application d'une régression linéaire. Cependant, la normalité des résidus de la régression doit être vérifiée. Dans le cas où les résidus ne sont pas normalement distribués, le test de régression linéaire sera peu puissant (l'hypothèse d'une tendance sera écartée injustement) et les intervalles de confiance qui peuvent être calculés peuvent être surestimés.

3.3. TEST DE STATIONNARITE DE MANN-KENDALL ET PENTE DE SEN

3.3.1. Test de Mann-Kendall simple

Le test de Mann-Kendall permet d'apprécier la non-stationnarité d'une chronique. La pente de Sen, calculée comme la médiane de toutes les pentes calculées entre chaque paire de points, est souvent associé à ce test. Il convient d'être prudent lorsque l'on utilise les résultats de calcul de cette pente linéaire car le test de Mann-Kendall ne préjuge aucunement de la linéarité de la tendance d'évolution dans une chronique temporelle.

3.3.2. Modification du test de Mann-Kendall pour la prise en compte de l'autocorrélation

Comme présenté notamment par Yue et Wang (2004), le test de Mann-Kendall, s'il est appliqué sur des données présentant une autocorrélation significative, peut être biaisé (Yue et Wang, 2004). Lorsqu'aucune tendance n'existe dans une chronique temporelle, la présence d'une autocorrélation positive augmente la possibilité de rejeter l'hypothèse nulle d'absence de tendance alors qu'elle est vraie. Cependant, pour une série comportant une tendance linéaire et également une autocorrélation, la prise en compte de l'autocorrélation résulte en la sous-estimation de la significativité d'une tendance existante.

Hamed et Rao (1998) ont proposé une méthode pour prendre en compte l'effet de l'autocorrélation : Le principe consiste à calculer un nombre n* d'observations supposées indépendantes, ce qui revient à substituer à la série initiale de n valeurs autocorrélées une série de n* valeurs indépendantes, n*< n (nombre équivalent d'observations indépendantes). Ce nombre n* est ensuite introduit dans le test classique de Mann-Kendall au lieu du nombre n.

Les tests de Mann-Kendall et Mann-Kendall modifiés sont compilés dans l'outil HYPE.

3.4. METHODES DE REGRESSION LOCALE

Les méthodes de régression locale sont utilisées dans le cas de tendances complexes d'évolution des concentrations. Elles permettent de stationnariser une série chronologique en calculant le résidu entre la chronique brute et sa régression. Elles sont de plus très utiles pour comparer entre eux plusieurs gros jeux de données.

Dans les méthodes de régression locale, chaque valeur de la courbe de lissage est calculée à partir des valeurs des points voisins. L'inconvénient des procédures de lissage local est que les courbes lissées n'ont pas de formules analytiques. Par contre, elles permettent de définir un intervalle de confiance pour la valeur prédite.

Parmi les différentes méthodes de lissage locales relevées en bibliographie, les moyennes mobiles, les algorithmes de LOESS et de LOWESS et le lissage par spline sont apparus comme les plus aptes à répondre au problème posé de la décomposition des chroniques de la qualité des eaux souterraines. Ces différentes méthodes sont présentées dans les paragraphes suivants.

3.4.1. Moyennes mobiles

Il y a plusieurs manières de calculer des moyennes mobiles en considérant q le nombre de valeurs prises en compte de part et d'autre du point étudié :.

- **Moyenne mobile symétrique définie par la formule suivante :**

$y(i) = \frac{1}{2q+1}\sum_{j=-q}^{q} x(i+j)$, $i > q$

- **Moyenne mobile pondérée qui utilise des coefficients pour donner un poids distinct à chaque valeur utilisée dans le calcul. Les poids décroissent linéairement avec le temps.**

$y(i) = \frac{1}{2q+1}\sum_{j=-q}^{q} \propto_j x(i+j)$

Où $\propto_j = 1 - \frac{|j|}{q+1}$

- **Moyenne mobile récursive (ou lissage exponentiel)**

qui utilise une pondération des termes qui décroît exponentiellement. La constante de lissage $\propto$ contrôle le degré de décroissance des poids applicables à chaque observation participant à la moyenne.

Pour $0 \leq \propto \leq 1$, on définit la moyenne mobile exponentielle par :

$y(i) = \propto x(i) + (1-\propto) y(i-1)$ pour $i > 1$ et $y(1) = x(1)$

Les moyennes mobiles ont l'avantage de la simplicité de mise en œuvre mais présentent l'inconvénient majeur d'être sensibles aux valeurs extrêmes. D'autre part, leur utilisation sur les bords (en début et fin de chronique) pose problème.. De plus, le choix de la valeur q dépend d'un compromis : une valeur de q trop faible ne permet pas d'extraire la tendance du résidu alors qu'une valeur trop élevée rend mal compte des évolutions de la tendance.

- **Cas des données non régulières**

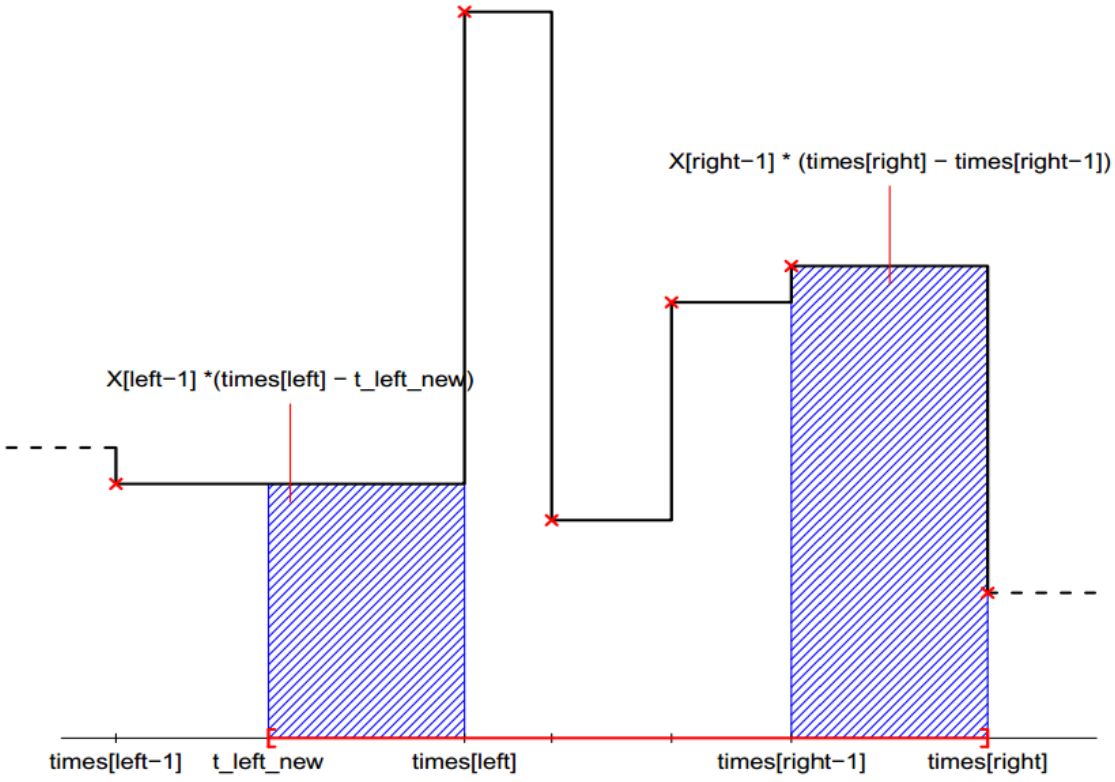
Dans le cas des de données non régulièrement espacées, il n'est pas possible d'utiliser directement les formules de moyennes mobiles exposées précédemment. Eckner (2012) propose plusieurs méthodes alternatives pour calculer les moyennes mobiles de données irrégulières pour lesquelles le paramètre qui délimite le voisinage du point utilisé pour les calculs n'est plus le nombre de point pris en compte mais la longueur de la fenêtre, appelé τ ci-après :

- Moyenne géométrique des valeurs dans la fenêtre.
- Moyenne des valeurs dans la fenêtre pondérées en fonction du laps de temps sur laquelle la valeur reste stable (Illustration 1 a).

  $y(i) = \frac{1}{\tau}\sum_p x(j) * \big(t(j) - t(j-1)\big)$, où p est tel que $t(i) - \tau \leq t(p) \leq t(i) + \tau$

- Moyenne des valeurs dans la fenêtre en interpolant linéairement entre deux valeurs (Illustration 1 b)

a)



b)

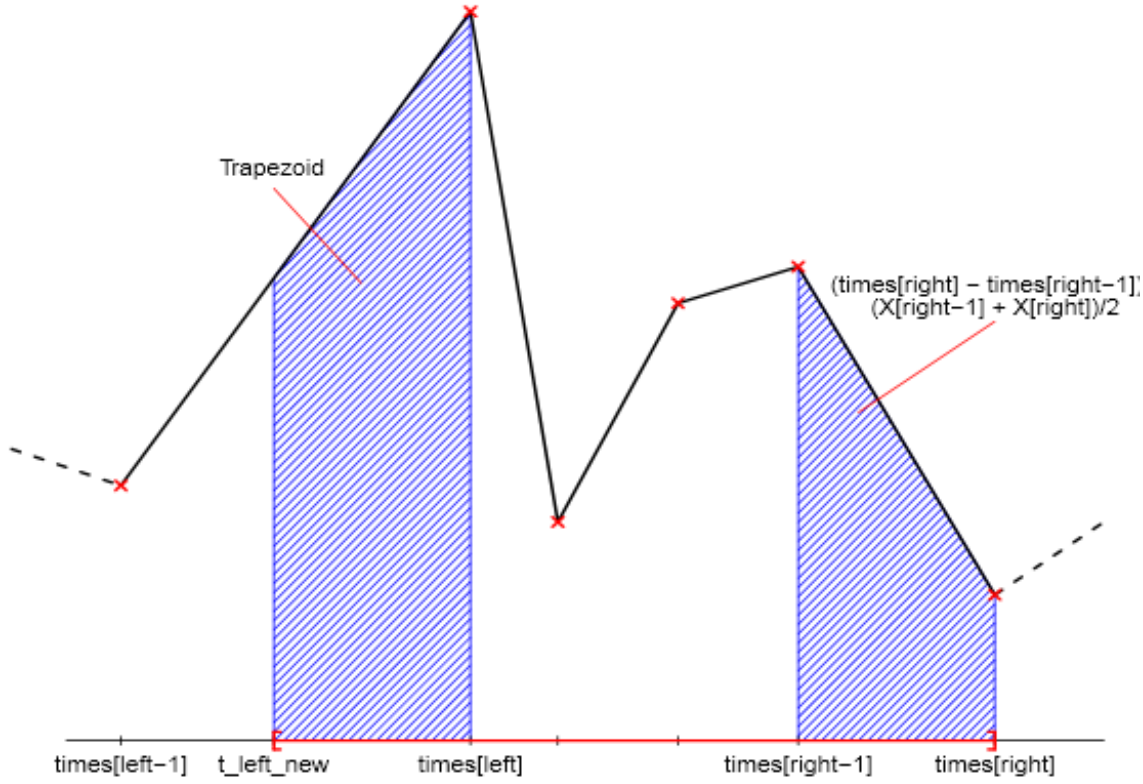


Illustration 1 : Illustration des différentes méthodes de calcul d'une moyenne mobile d'une série non régulière. Moyenne pondérée (a) et interpolation linéaire (b).

Ces méthodes modifiées pour le calcul des moyennes mobiles apparaissent adaptées aux caractéristiques des chroniques en eaux souterraines à étudier.

3.4.2. Algorithmes LOESS et LOWESS

- **Cas général**

Le lissage par les méthodes LOWESS ou LOESS (Locally Weighted Regression and Smoothing Scatter Plot) ont été développées par Cleveland (1979). Ces méthodes consistent à déterminer, pour chaque point de la chronique, les coefficients d'un polynôme de faible degré pour effectuer la régression d'un sous-ensemble des données, puis à calculer la valeur de ce polynôme pour le point considéré. Les coefficients du polynôme sont calculés à l'aide de la méthode des moindres carrés pondérés, qui donne plus de poids aux points proches du point dont la réponse est estimée et moins de poids aux points les plus éloignés. La fonction de pondération classiquement utilisée est une fonction cubique pondérée. Pour définir cette fonction de pondération au point x_0, on définit les coefficients de pondération suivants pour les observations aux différents temps x_i. En définissant $z_i = \frac{x_i - x_0}{h}$ avec h la demi-largeur de la fenêtre de lissage :

$w(z_i) = (1 - |z_i|^3)^3$ pour $|z_i| < 1$

$w(z_i) = 0$ pour $|z_i| \geq 1$

Cette méthode de lissage peut être appliquée sur des données non régulièrement échantillonnées.

De nombreux éléments sont paramétrables : le degré du polynôme (en général 1 : procédure LOWESS ou 2 : procédure LOESS), les coefficients de pondération, la taille de la fenêtre de lissage. Pour choisir la taille du voisinage qui doit être pris en compte autour de chaque point, il y a plusieurs stratégies. En représentant graphiquement les résidus, on peut choisir la largeur la plus importante qui permet que très peu d'informations (cyclicité notamment) se retrouvent dans les résidus. Il existe également des méthodes automatiques qui consistent à minimiser le critère $\log(\hat{\sigma}^2) + \psi(L)$ où $\hat{\sigma}^2$ est la somme moyenne des résidus au carré et $\psi(L)$ est une fonction de pénalité qui décroît quand le lissage augmente.

- **Procédure robuste en présence de valeurs extrêmes**

Si la série contient des valeurs extrêmes, les valeurs lissées peuvent être influencées par ces valeurs et ne plus refléter le comportement du nuage de points principal. La procédure peut être modifiée pour être plus robuste vis-à-vis de la présence d'un petit nombre de valeurs extrêmes.

Cette modification consiste à calculer les résidus de la procédure de lissage puis calculer les poids robustes pour chaque point de la fenêtre de lissage par la formule suivante :

$$wrob(z_i) = \begin{cases} \left(1 - \left(\frac{r_i}{6MAD}\right)^2\right)^2 & , \quad |r_i| < 6MAD \\ 0\,, & |r_i| \geq 6MAD \end{cases}$$

où r_i est le résidu au point i et MAD est la médiane des résidus $MAD = median(|r|)$

Pour les points avec un résidu important, le poids robuste est donc plus faible que pour les points avec un résidu faible. On peut ensuite réitérer la procédure de lissage, mais en utilisant

cette fois les poids robustes calculés à l'étape précédente. Cette procédure peut être répétée plusieurs fois.

Cette méthode est illustrée sur l'Illustration 2 pour un jeu de données fictif contenant une unique valeur extrême et sur l'Illustration 3 pour un une chronique réelle d'évolution des concentrations en atrazine dans l'aquifère de la craie dans le département du Nord Pas-De-Calais.

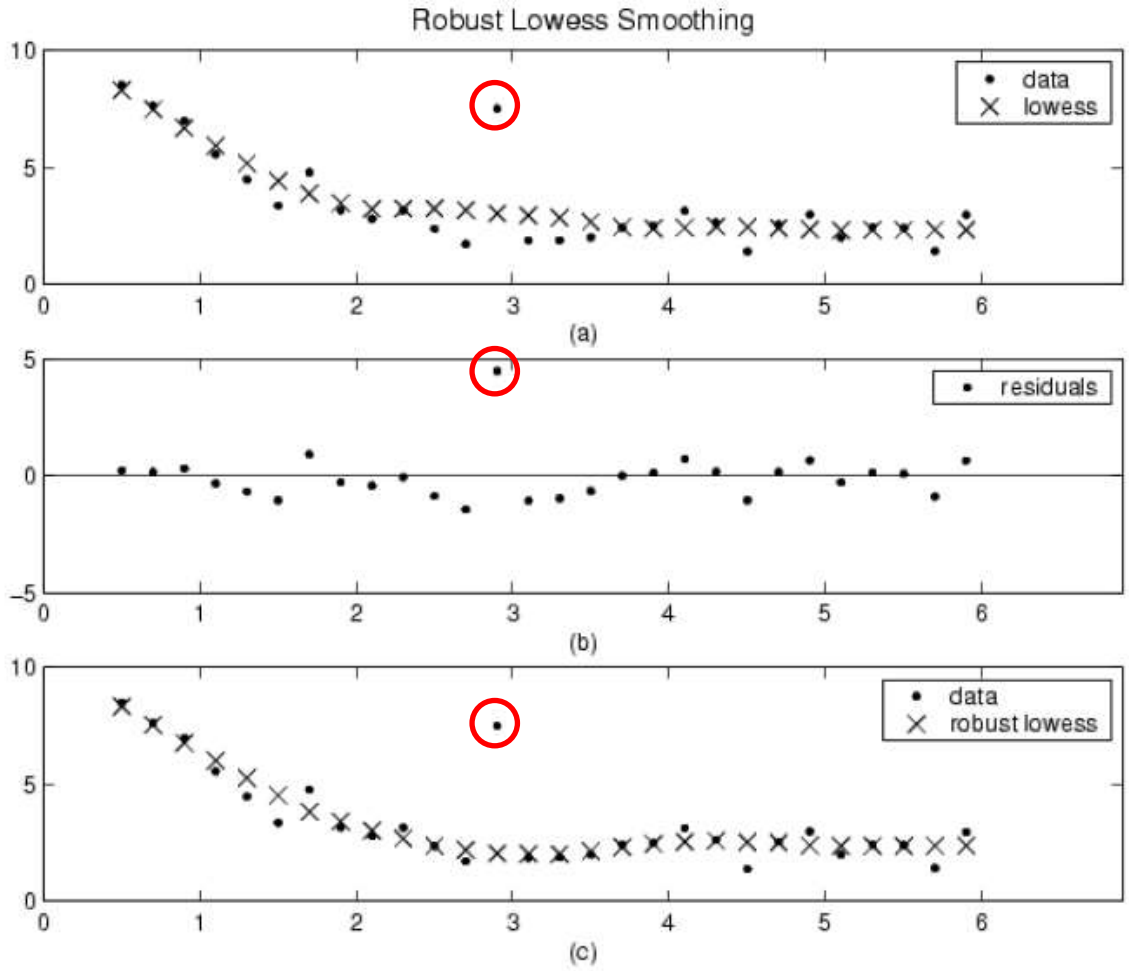


Illustration 2 : Illustration de la procédure LOWESS robuste sur un jeu de données fictif contenant une valeur extrême repérée par un cercle rouge (source : documentation MATLAB)

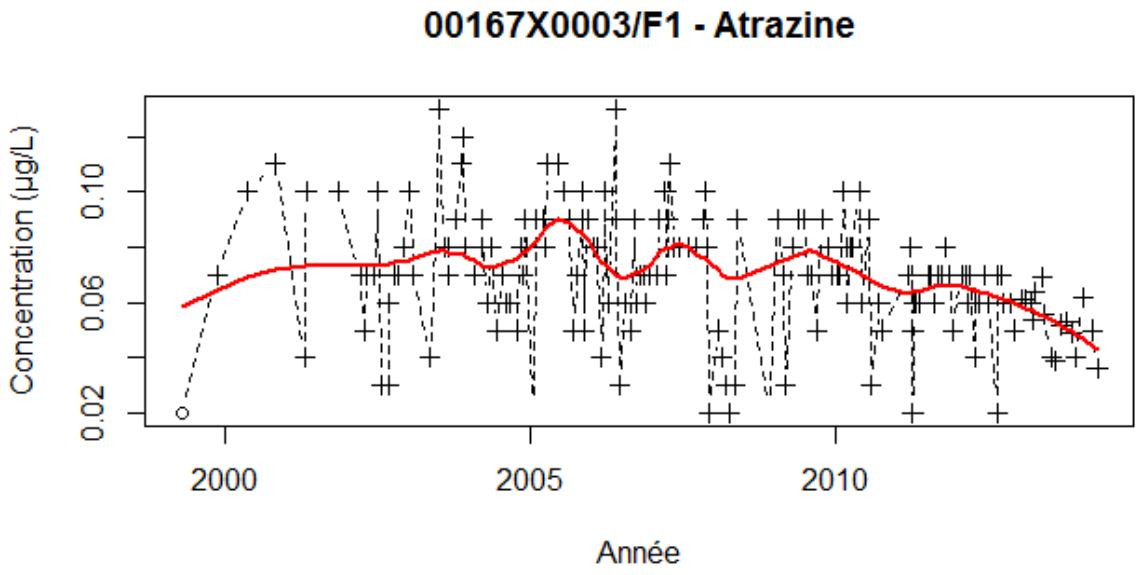


Illustration 3 : Application de la procédure de lissage de LOWESS sur une chronique réelle d'évolution des concentrations en atrazine dans la craie au point 00167X0003/F1 Airon-Saint-Vaast dans le département du Pas-de-Calais (62).

La procédure de lissage par LOWESS est utilisée dans plusieurs pays européens pour le calcul de tendance dont l'Autriche et la Roumanie. La procédure d'identification des tendances inclut une régression linéaire appliquée en post traitement sur le modèle LOWESS de la série.

3.4.3. Lissage par spline

La méthode du lissage par spline consiste à découper la fonction à ajuster en sous intervalles, puis à ajuster sur chaque sous-intervalle une fonction simple, en veillant à ce que le raccordement aux points de jonction soit cohérent. Le lissage par spline permet d'avoir une formulation analytique de la régression et peut donc être utilisé pour évaluer la fonction à des périodes où la chronique ne dispose pas de données (en prévision ou dans les cas où des données sont manquantes). Les fonctions utilisées sont en général des polynômes, le plus souvent des polynômes de degré 3 (on parle alors de splines cubiques).

L'Illustration 4 montre un exemple d'utilisation de cette technique pour interpoler une série chronologique correspondant à l'évolution d'un niveau piézométrique. Le raccordement de ces arcs est réalisé en imposant aux points de jonction la continuité ainsi que l'égalité des pentes et des courbures. Le découpage permet d'utiliser sur chaque sous-intervalle une fonction sensiblement plus simple que la fonction qu'il aurait fallu ajuster globalement.

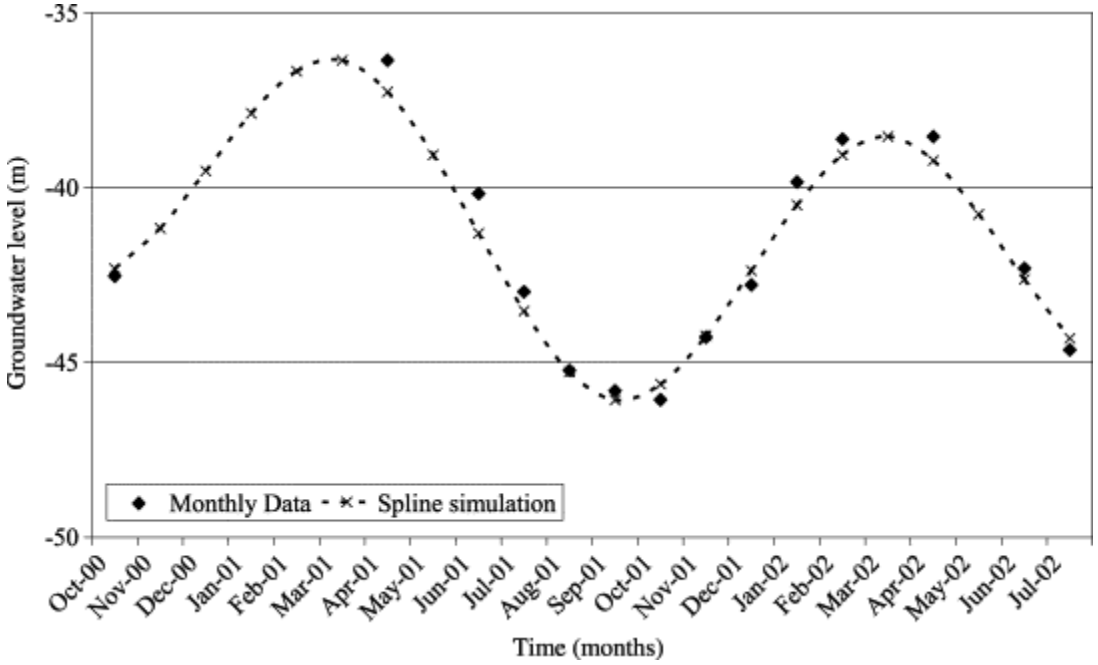


Illustration 4 : Exemple d'interpolation par spline cubique de l'évolution du niveau piézométrique à Messara (Grèce) permettant d'estimer des valeurs manquantes dans la chronique (Daliakopoulos et al., 2005)

3.5. METHODE RECOMMANDEE POUR LA REGRESSION ET LA STATIONNARISATION DES CHRONIQUES

Les méthodes de régression et de stationnarisation des chroniques ont été testées sur des jeux de données réels provenant de la base de données sur les eaux souterraines Ades (http://www.ades.eaufrance.fr/). Seules les chroniques disposant d'au moins 25 données ont été utilisées, et ce quelle que soit la substance considérée.

Afin de ne pas surcharger le présent document, les tests réalisés sur les données réelles ne sont pas présentés dans leur totalité. Seuls les tests sur la méthode de régression locale selon l'algorithme de LOESS et LOWESS sont présentés. Cette méthode apparaît en effet à privilégier dans le cadre de l'analyse des tendances complexes d'évolution des concentrations en eau souterraines. Elle comporte toutefois quelques critères subjectifs paramétrables qu'il convient de tester : le facteur « span » qui définit la taille de la fenêtre de lissage et l'option « Family = symetric » pour minimiser l'effet des valeurs extrêmes.

3.5.1. Le facteur « span »

Le facteur "span" permet de signifier le nombre de données qui seront prises en compte dans le calcul des données lissées. Plus le chiffre du "span" est grand plus le pourcentage de données utilisées autour de la données à prédire sera grand. Le chiffre peut même être supérieur à 1. Dans ce cas, toutes les valeurs de la chronique sont utilisées et un poids plus important est donné aux données les plus proches. Quatre « span » ont été comparés 0,1 ; 0,5 ; 0,9 et 1,5.

- **Comparaison 0,5 et 0,9.**

La première comparaison a consisté à confronter les prédictions d'une fonction LOWESS ayant un "span" de 0.5 et les prédictions d'une fonction LOWESS ayant un "span" de 0,9.

Toutes les chroniques reconstruites par la fonction LOWESS étudiées montrent un lissage cohérent des chroniques de valeurs mesurées. Pour la plupart des chroniques, les résultats des prédictions sont relativement proches pour des fonctions LOWESS ayant un "span" de 0,5 ou 0,9.

L'exemple suivant présente les ajustements réalisés avec la fonction LOWESS en utilisant les deux "span" 0,5 et 0,9 (Illustration 5). Les résultats montrent toutefois que les prédictions réalisées en utilisant le "span" de 0,9 sont plus lisses et moins affectées par la modification des valeurs de la chronique.

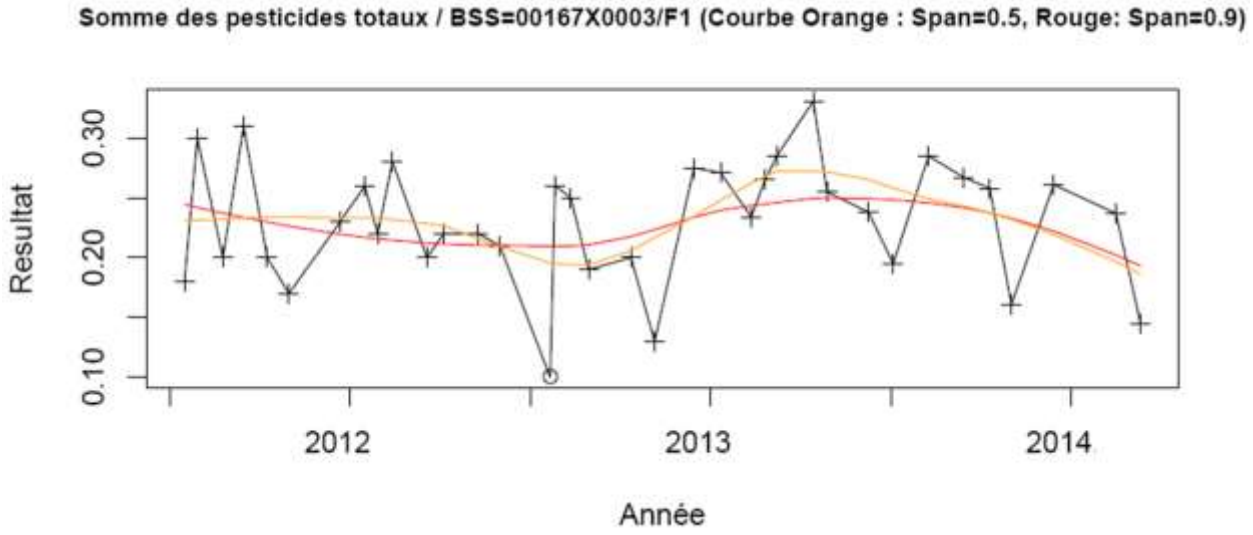


Illustration 5 - Comparaison du lissage de la courbe "somme des pesticides totaux" Orange = Span 0.5. Rouge = Span 0.9

Ainsi, la fonction LOWESS lisse d'autant mieux les données extrêmes que le "span" est important (proche de 1). L'utilisation du "span" de 0,9 permet donc d'obtenir une courbe plus lisse dans le cas de la présence d'une valeur extrême.

Dans l'exemple de l'Illustration 6, la valeur extrême en 2001 a un impact important sur les prédictions réalisées avec un "span" de 0.5. Cet impact est visible durant 5 années environ. L'impact de la valeur extrême est moins important en amplitude sur les prédictions réalisées avec un span de 0.9 mais il s'étend sur une période plus longue de 1993 à 2005 environ. Dans cet exemple, aucun des deux lissages ne semble adapté. La fonction LOWESS, en utilisant ces paramètres, ne peut pas prendre en compte des valeurs extrêmes.

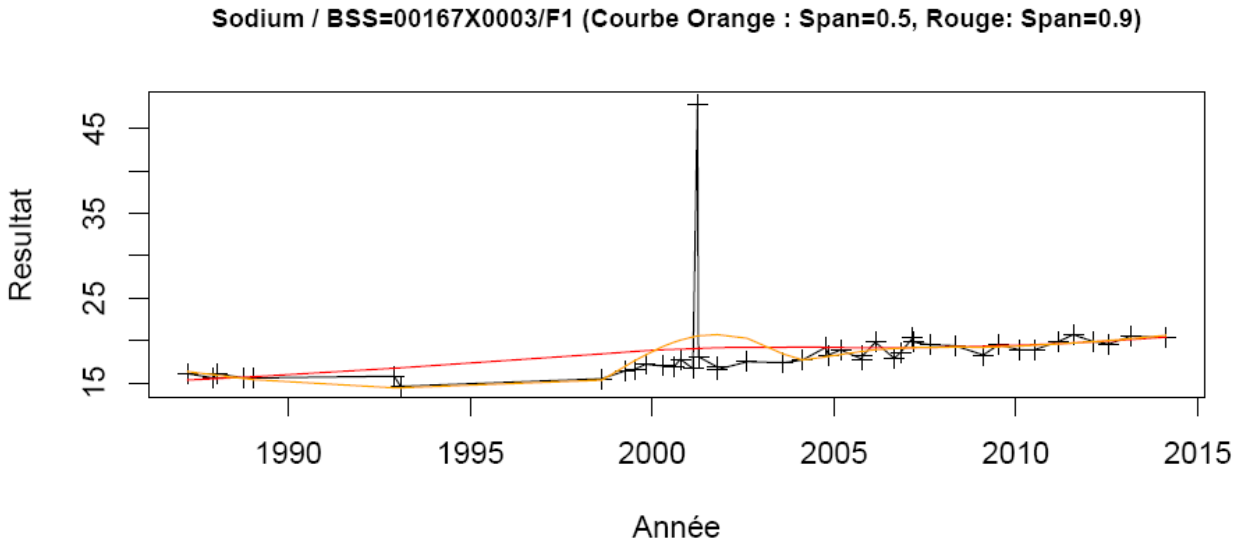


Illustration 6 - Comparaison du lissage de la courbe "Sodium" Orange = Span 0.5. Rouge = Span 0.9

- **Comparaison 0,5 et 0,1.**

La deuxième comparaison a consisté à confronter les prédictions d'une fonction LOWESS ayant un "span" de 0.5 et les prédictions d'une fonction LOWESS ayant un "span" de 0,1.

Dans l'exemple en Illustration 7, le lissage des données du paramètre « somme des pesticides » est réalisé d'une manière cohérente. Ici, la différence entre les prédictions réalisées avec les différents "span" est beaucoup plus visible. En effet, un lissage est opéré lorsque le span de 0,5 est utilisé mais lorsque le span de 0,1 est utilisé, pratiquement aucun lissage n'est exécuté. La courbe des valeurs ajustées par la fonction LOWESS suit pratiquement la courbe des valeurs mesurées. Ce type de résultat est dû au fait qu'un faible nombre de valeurs autour de la valeur prédite est pris en compte dans la méthode.

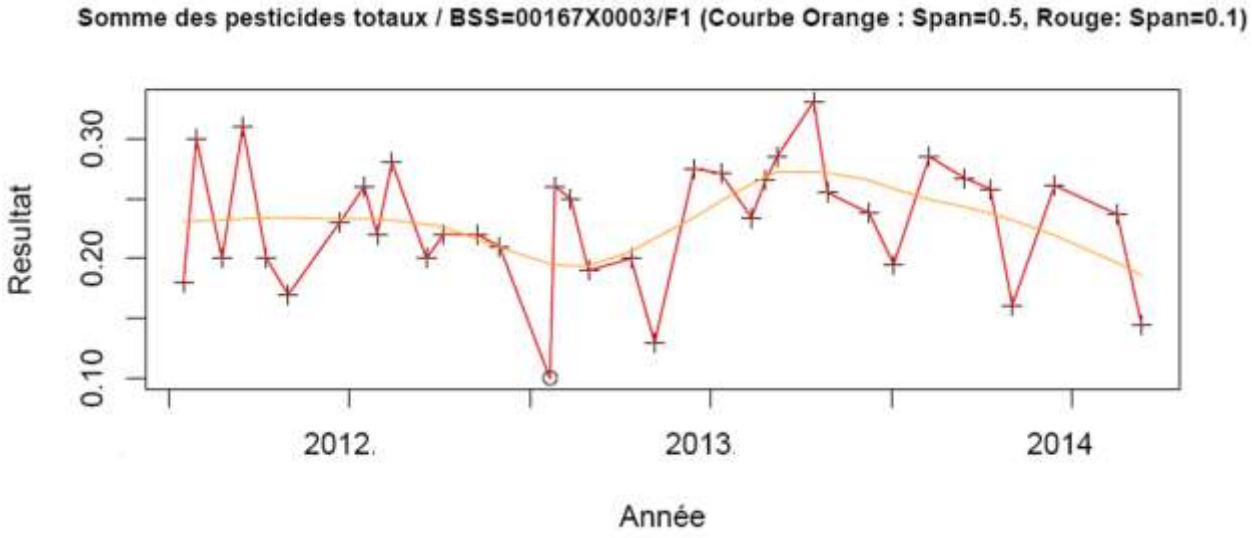


Illustration 7 - Comparaison du lissage de la courbe "somme des pesticides totaux" Orange = Span 0.5. Rouge = Span 0.1

Une différence entre la chronique des valeurs mesurées et celles des valeurs prédites ne devient visible pour une fonction LOWESS ayant un span de 0,1 que lorsque le nombre de valeurs de la chronique devient important. Dans l'exemple de l'Illustration 8 sur l'évolution des concentrations en atrazine, une différence entre les deux courbes d'ajustement est clairement visible. Ce résultat provient du fait que si le nombre de données est très important, 10% des données (span=0,1 signifie que 10% des données au voisinage sont pris en compte) représente quand même un nombre important de données.

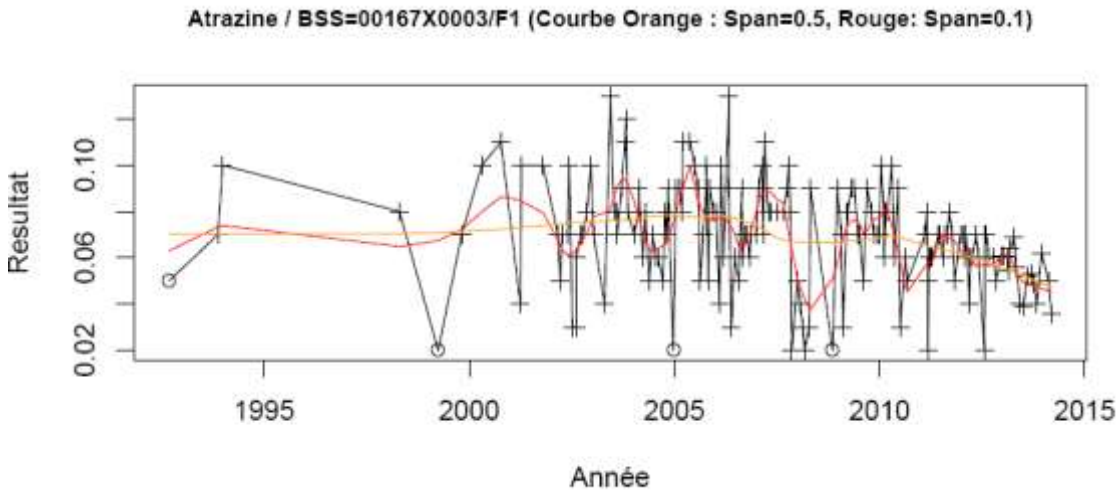


Illustration 8 - Comparaison du lissage de la courbe "Atrazine" Orange = Span 0.5. Rouge = Span 0.1

- **Comparaison 0,5 et 1.5.**

La troisième comparaison a consisté à confronter les prédictions d'une fonction LOWESS ayant un "span" de 0.5 et les prédictions d'une fonction LOWESS ayant un "span" de 1,5.

En augmentant la valeur du paramètre "span" à 1,5, il est possible de lisser de manière importante les données brutes. Dans ces conditions, la courbe lissée forme pratiquement une droite. Ainsi dans l'Illustration 9, la prédiction est proche de la moyenne de la série (0,23 µg/L).

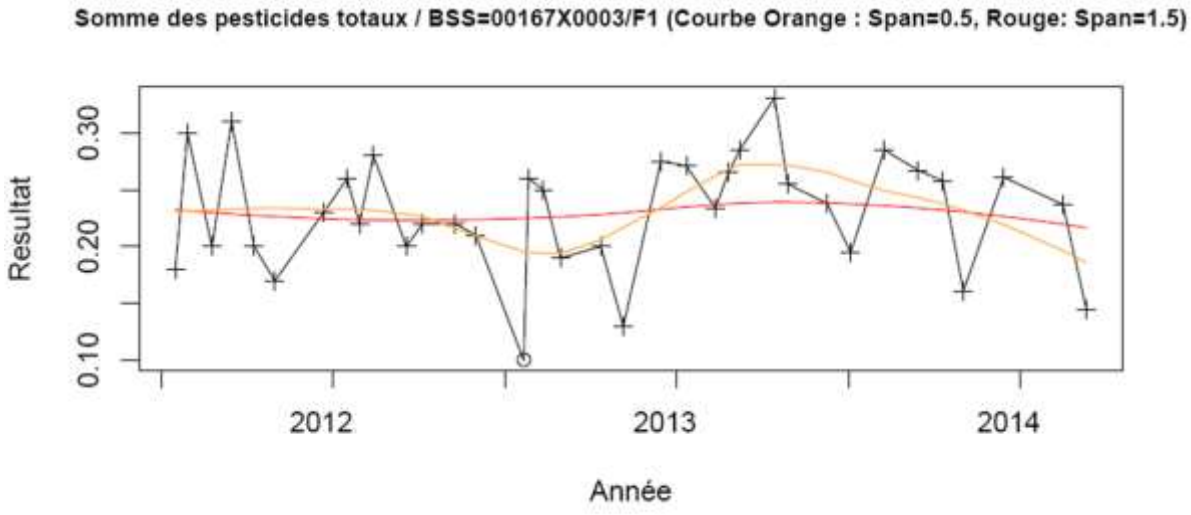


Illustration 9 - Comparaison du lissage de la courbe " somme des pesticides totaux" Orange = Span 0.5. Rouge = Span 1.5

3.5.2. Utilisation de l'Option "Family = symetric »

Cette fonction a pour intérêt de minimiser l'impact des valeurs extrêmes sur le calcul du lissage. Les prédictions d'une fonction LOWESS ayant un "span" de 0.5 utilisant l'option "Family = Symetric" ont ainsi été confrontées avec celles d'une fonction LOWESS ayant un "span" de 0,5 sans cette option (Illustration 10).

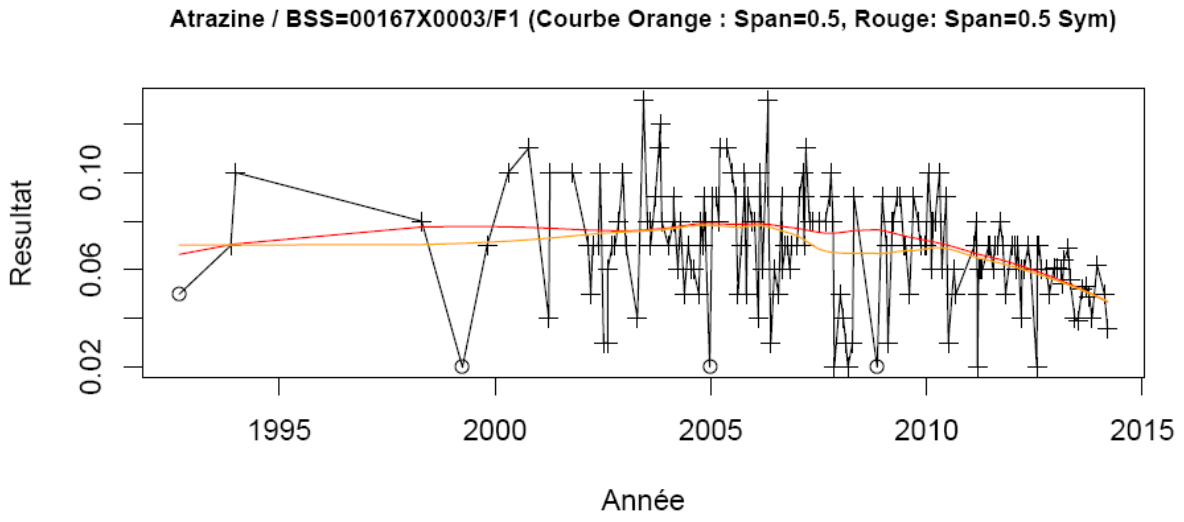


Illustration 10 - Comparaison du lissage de la courbe "Atrazine" Orange = Span 0.5 sans option Rouge = Span 0.5 avec option

L'utilisation de l'option "Family = Symetric" sur la chronique d'évolution des concentrations en atrazine au point 00167X0003/F1 n'apparaît pas particulièrement sensible. En effet, aucune valeur mesurée n'étant particulièrement extrême, l'impact de l'option est donc mineur. Néanmoins, certaines valeurs prédites sont tout de même influencées, montrant que l'impact de la fonction n'est pas nul.

L'Illustration 11 présente l'impact de l'option "Family= symetric " pour le lissage des données de concentration en sodium comportant une valeur extrême déjà testée au paragraphe 3.5.1. Sans l'option, la valeur extrême a une influence sur les prédictions entre les années 2000 et 2005. Une telle influence d'une seule valeur, même mesurée, n'est pas justifiable, surtout au regard des autres valeurs de la chronique. Avec l'option "Family= symetric ", l'impact de la valeur extrême est fortement minimisé puisque son impact sur le lissage est pratiquement invisible, autant en amplitude qu'en durée.

Dans ce cas spécifique, l'utilisation de l'option "Family= symetric " permet de grandement améliorer la prise en compte des valeurs extrêmes. Ainsi pour la prise en compte de ces valeurs, il est recommandé d'utiliser cette option plutôt que d'augmenter la valeur de "span" comme testé dans l'exemple de l'Illustration 6.

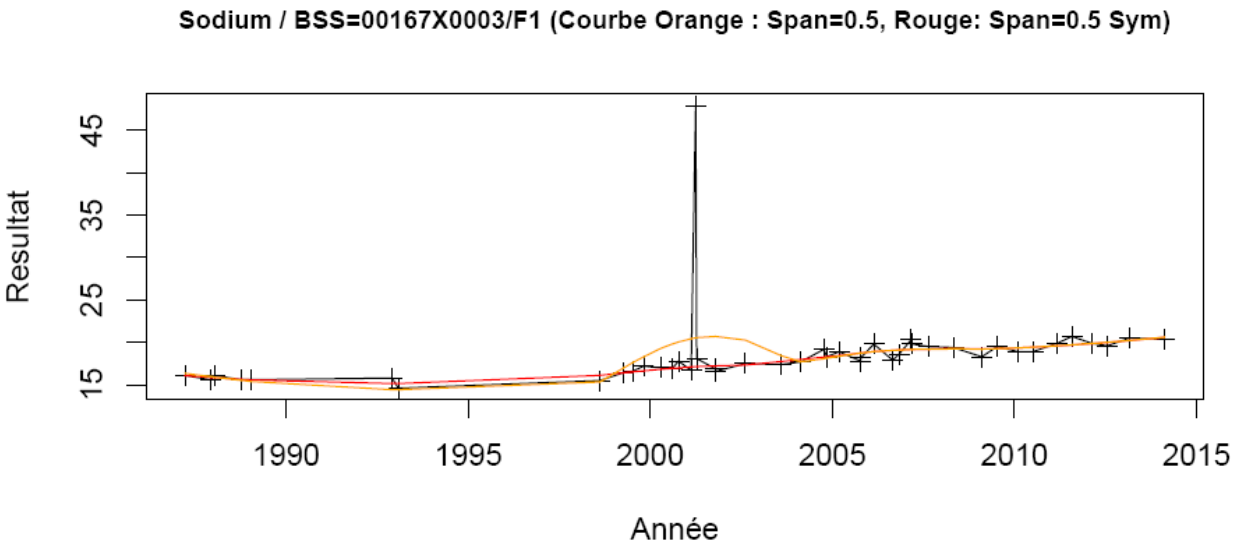


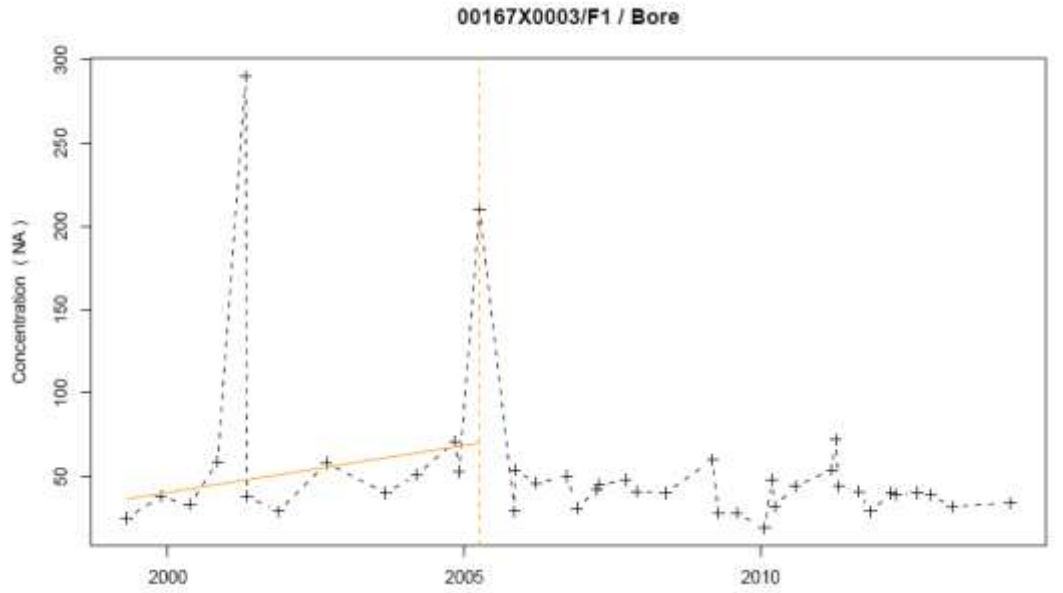
Illustration 11 - Comparaison du lissage de la courbe "sodium" Orange = Span 0.5 sans option Rouge = Span 0.5 avec option

3.5.3. Identification des tendances

La fonction LOWESS permet de produire des courbes lissées, ajustées à un nuage de points de données. La chronique issue de ces lissages est de fait moins affectée par des valeurs extrêmes. En partant de ce constat, il paraît intéressant d'essayer d'identifier des tendances sur les chroniques de données lissées. En effet, les outils de calcul de tendance ne seront pas perturbés par les valeurs extrêmes. Plusieurs types de tendance ont été recherchés. Des tendances sur l'ensemble de la chronique, de type régression linéaire ou de type régression de Sen, ont d'abord été recherchées. Ensuite, des tendances avec une rupture ont aussi été recherchées.

L'Illustration 12 présente les tendances détectées sur la chronique des concentrations en bore mesurées au point 00167X0003/F1 puis sur la chronique de données prédites. Sur l'illustration montrant la chronique de données prédites (Illustration 12 b), la chronique de données mesurées a été rappelée à titre de comparaison en gris clair. Cet exemple montre que l'identification des tendances sur la chronique de données prédites plutôt que sur la chronique de données mesurées permet de révéler des tendances non détectées sur les données brutes. Ainsi, dans cet exemple, il est possible de conclure à une diminution de la concentration en utilisant les données prédites alors que les données mesurées ne permettaient d'émettre aucune conclusion.

a)



b)

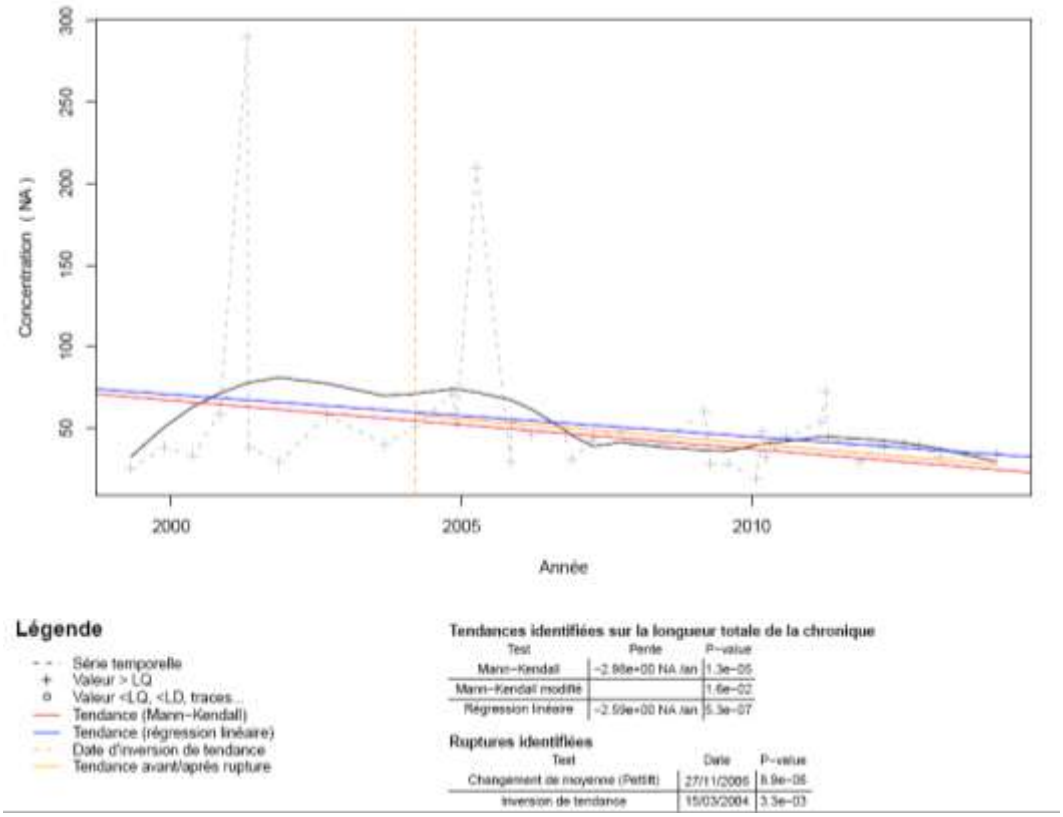
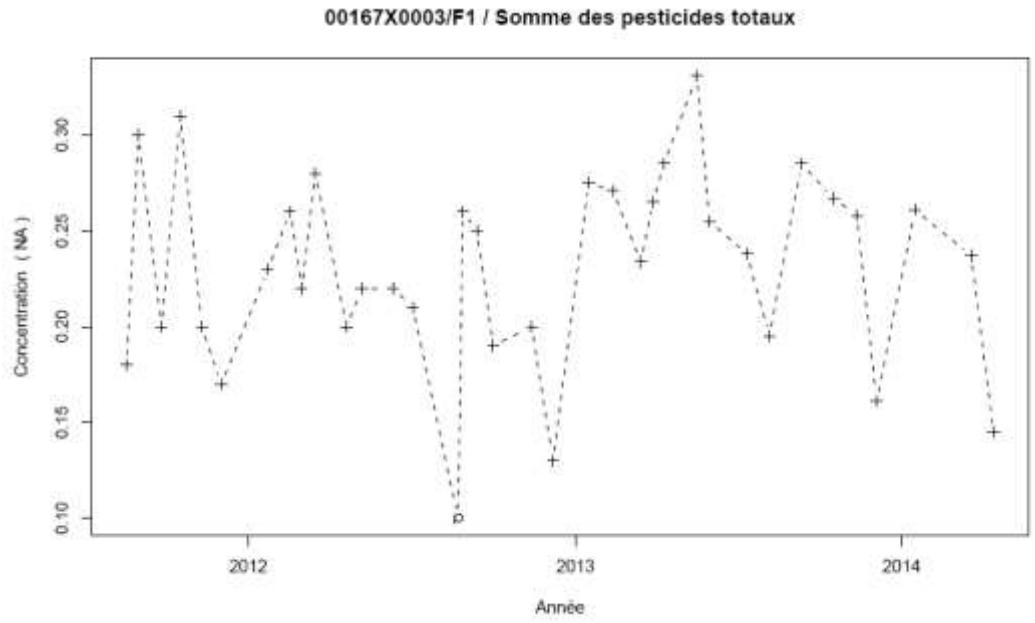


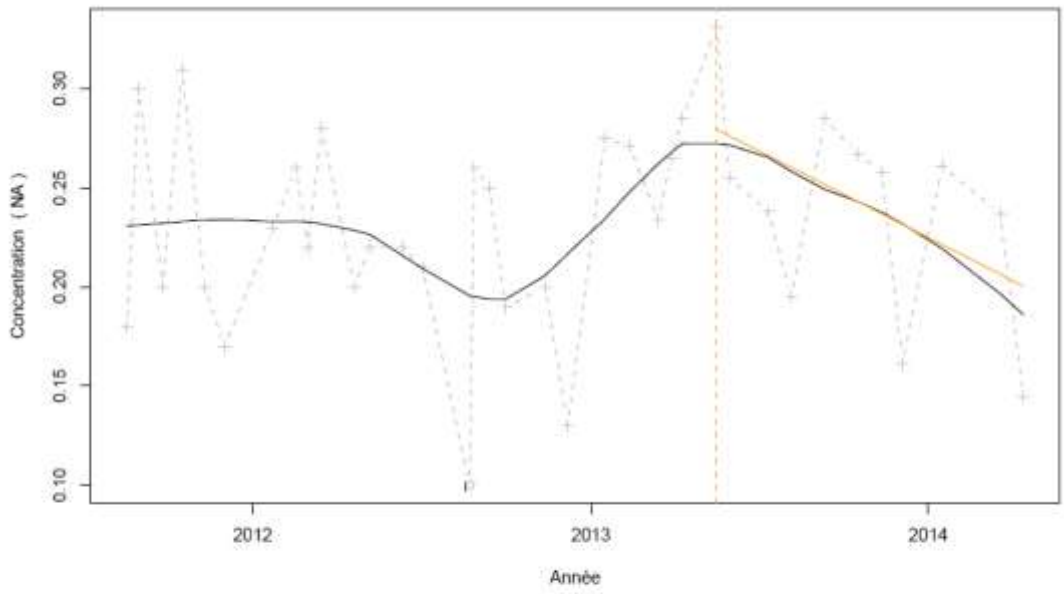
Illustration 12 : Recherche de tendance avec l'outil HYPE sur a) une chronique brute d'évolution des concentrations en bore dans les eaux souterraines (point 00167X0003/F1) et b) sur les mêmes données prédites par lissage LOWESS.

L'Illustration 13 montre un autre exemple de détection de tendance sur la chronique lissée alors qu'elle n'est pas détectable sur la chronique brute.

a)



b)



Légende

- Série temporelle
- Valeur > LQ
- Valeur <LQ, <LD, traces...
- Date d'inversion de tendance
- Tendance avant/après rupture

Tendances identifiées sur la longueur totale de la chronique

Test	Pente	P-value
Mann-Kendall	Aucune tendance significative détectée	6.8e-01
Mann-Kendall modifié	Non effectué (pas assez de données)	
Régression linéaire	Aucune tendance significative détectée	5.4e-01

Ruptures identifiées

Test	Date	P-value
Changement de moyenne (Buishand)	Pas de rupture significative détectée	
Inversion de tendance	17/05/2013	1.9e-04

Illustration 13 : Recherche de tendance avec l'outil HYPE sur a) une chronique brute d'évolution des concentrations en pesticides totaux dans les eaux souterraines (point 00167X0003/F1) et b) sur les mêmes données prédites par lissage LOWESS.

Suite aux tests menés sur des données réelles, l'algorithme de LOWESS apparaît pertinent à la fois pour extraire des tendances complexes des chroniques d'évolution de la qualité des eaux souterraines, mais aussi pour désaisonnaliser les chroniques et les stationnariser (résidus). Les

tests ont aussi montré la nécessité de bien appréhender les paramètres de la fonction LOWESS en fonction des caractéristiques de la chronique analysée et des objectifs à atteindre. Une valeur de span trop importante peut par exemple lisser les données de manière excessive et entraîner la disparition des tendances à extraire. A l'inverse, si l'objectif de l'application de la fonction LOWESS est de désaisonnaliser la chronique brute, un span trop faible peut entraîner un modèle contenant toujours une certaine périodicité dans l'évolution des données. Cet outil puissant ne peut donc vraisemblablement pas être automatisé sans intervention de l'opérateur pour choisir les valeurs des paramètres de la fonction LOWESS à chaque chronique analysée.

4. Identification et quantification des variations cycliques

4.1. OBJECTIF

L'identification d'une composante cyclique, saisonnière ou pluri-annuelle est une étape intéressante pour comprendre les phénomènes qui guident l'évolution d'une chronique temporelle. Il s'agit notamment d'identifier la partie de l'évolution du chimisme de l'eau souterraine qui est vraisemblablement guidée par des phénomènes naturels, hydroclimatiques principalement, sur lesquels il n'est pas possible d'agir.

La cyclicité périodique dans les chroniques engendre de l'autocorrélation des données. Or, à l'instar de la non-stationnarité, l'autocorrélation est un phénomène que l'on cherche à supprimer des chroniques en vue de leur décomposition. L'autocorrélation a en effet tendance à biaiser les tests statistiques à appliquer pour l'analyse des séries chronologiques. Estimer l'autocorrélation des données revient à chercher à savoir si une valeur observée à un temps t dépend de ce qui a été observé dans le passé. L'autocorrélation peut être due à une tendance dans la chronique temporelle ou à la « mémoire » des processus hydrogéologiques.

Pour étudier l'autocorrélation liée à l'inertie d'un système, on cherchera dans un premier temps à supprimer l'autocorrélation liée à la présence d'une tendance, en utilisant l'une des méthodes présentées dans le paragraphe 3. Il sera ensuite nécessaire d'étudiée l'autocorrélation potentiellement induite par la « mémoire » des processus hydrogéologiques ou hydrogéochimiques en appliquant l'une des méthodes présentées aux paragraphes 4.2 et 4.3. Parmi elles, seule l'analyse par ondelettes (cf paragraphe 4.3.3) permet d'étudier conjointement cyclicité et tendances à moyen et long terme. Pour les autres méthodes, il est nécessaire de supprimer dans un premier temps les tendances avant de rechercher la cyclicité.

La régularité des prélèvements est un autre élément important à prendre en compte dans le choix de la méthode à appliquer pour étudier l'autocorrélation. Certaines méthodes imposent en effet que les données soient régulièrement espacées dans la chronique (méthodes présentées au paragraphe 4.2). Les chroniques de qualité des eaux souterraines ne satisfaisant que très rarement cette condition, l'application de ces méthodes nécessiteront un prétraitement des données brutes consistant en un ré-échantillonnage (cf. paragraphe 4.3.1).

4.2. SERIES REGULIEREMENT ECHANTILLONNEES

4.2.1. Calcul de l'autocorrélation

Si les données sont régulièrement échantillonnées, l'autocorrélation au rang k d'une série temporelle peut se calculer par la formule suivante (Jenkins et Watts ,1968) :

$r_k = C_k/C_0$, où $C_k = \frac{1}{n}\sum_{i=1}^{n-k}(x_i - \bar{x})\,(x_{i+k} - \bar{x})$

L'autocorrélation au rang 0 est égale à 1. Plus les données sont autocorrélées, plus l'autocorrélation est proche de 1 pour les rangs immédiatement suivants, en particulier au rang 1 (ensuite aux rangs 2 et 3, mais cela peut se poursuivre au-delà, si les données successives de la série sont fortement liées).

La valeur des autocorrélations est ensuite comparée à la valeur limite définie ci-dessous à un seuil de significativité donné :

$r_{lim} = \frac{1}{\sqrt{n}} qnorm\left(1 - \frac{\propto}{2}\right)$, où $qnorm$ est la fonction quantile d'une loi normale centrée réduite $\propto$ est le seuil de significativité.

Si $r_k > r_{lim}$ l'autocorrélation au rand k est considérée comme significative.

Par exemple, Jones et Smart (2005) ont étudié les concentrations en nitrates dans 5 sources karstiques dans le Sud de l'Angleterre (Illustration 14). Leur étude montre que pour les 5 sources, une autocorrélation significative à court terme est observée. Ce phénomène peut être expliqué par la présence d'un stock de nitrates dans le sol qui permet de maintenir des concentrations importantes dans les eaux de recharge.

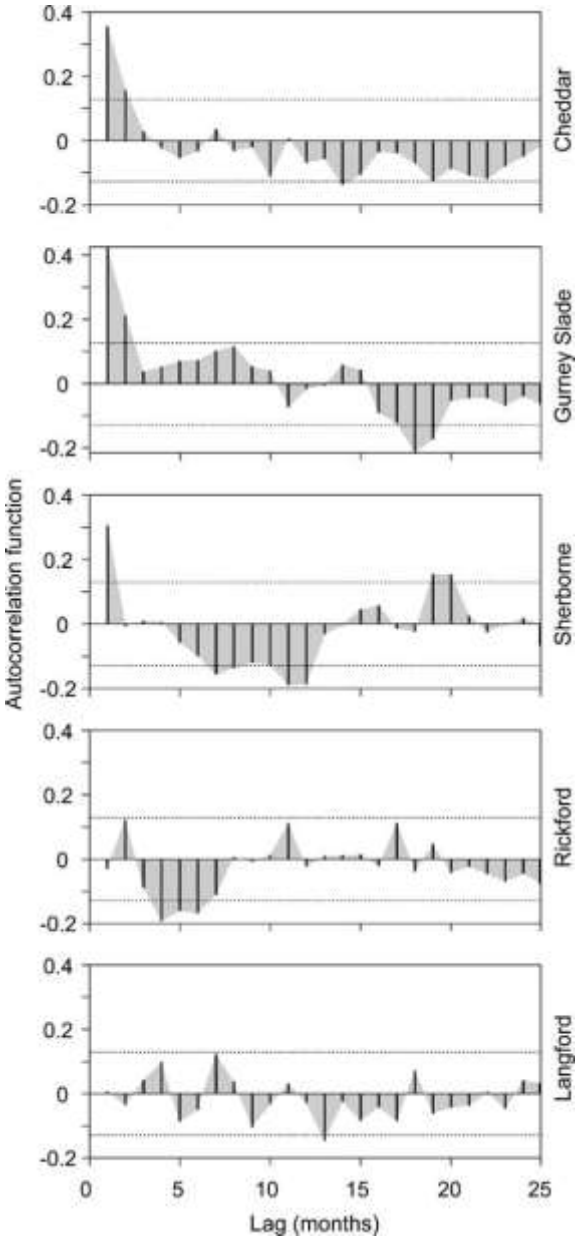


Illustration 14 : Fonctions d'autocorrélation calculée à partir des concentrations en nitrates dans 5 sources, une fois les composantes de tendance et saisonnière enlevées (Jones et Smart, 2005). La droite en pointillés correspond à l'intervalle de confiance à 95%

4.2.2. Analyse spectrale

L'analyse spectrale est un outil classiquement utilisé pour l'étude des séries temporelles. Elle consiste à décomposer un signal cyclique complexe en différents signaux élémentaires cycliques qui la composent par le passage de la série dans l'espace des fréquences (via une transformée de Fourier). On peut alors détecter quelles sont les fréquences qui contribuent le plus à la dynamique de la série.

Le résultat de l''analyse spectrale est un graphique dit spectre des puissances, ou périodogramme. Il correspond à une décomposition harmonique de la variance de la série dans l'espace des fréquences et la DSP est une quantification de la contribution de chaque harmonique à la variabilité de la série.

Le lecteur pourra se reporter à l'Annexe 1 pour une description plus détaillée de la mise en œuvre de cette méthode.

4.2.3. Périodogrammes (Papoulis, 1984, Welch, 1988)

L'analyse spectrale se heurte à deux difficultés importantes : le faible nombre de données et le bruit dont elles sont entachés. Les méthodes de périodogrammes permettent de calculer des estimations du spectre d'une série avec un nombre limité de données.

Le lecteur pourra se reporter à l'Annexe 1 pour la description de la mise en œuvre de ce type de méthode.

4.3. SERIES IRREGULIEREMENT ECHANTILLONNEES

Plusieurs méthodes ont été développées pour calculer l'autocorrélation d'une série avec des pas de temps non réguliers. Rehfeld et al. (2011) ont présenté et comparé différentes méthodes dont le rééchantillonnage des données et une méthode basée sur le périodogramme de Lomb-Scargle.

4.3.1. Rééchantillonnage

Schulz et Stattegger (1997) montrent que l'application des méthodes d'analyses spectrales pressenties pour la décomposition des chroniques d'évolution des concentrations en polluants dans les eaux souterraines sur des chroniques irrégulières en rééchantillonnant les données manquantes par interpolation (que ce soit par interpolation linéaire ou en utilisant des fonctions plus complexes comme des splines cubiques) sous-estime les composantes hautes fréquences du spectre de puissance. Ces méthodes de rééchantillonnage ne sont donc pas conseillées dans le contexte spécifique d'étude des chroniques de mesures effectuées dans les eaux souterraines.

4.3.2. Méthode basée sur le périodogramme de Lomb-Scargle

Scargle (1982,1989) a proposé des modifications de la transformée de Fourier discrète pour que le périodogramme de séries à pas de temps variable ait le même comportement statistique que le périodogramme classique. On appelle le périodogramme ainsi développé, le périodogramme de Lomb et Scargle. D'autres méthodes existent (Ferraz-Mello, 1981 ou Zechmeister and Kürster, 2009).

Le lecteur pourra se reporter à l'Annexe 1 pour la description de la mise en œuvre de cette méthode.

4.3.3. Analyse spectrale de série avec des ondelettes

- **Principe**

Dans l'analyse de Fourier, on décompose une série temporelle en différentes fréquences qui existent sur la totalité de la série. L'analyse spectrale avec des ondelettes permet de prendre en compte un changement dans la décomposition de Fourier en fonction du temps.

La première méthode qui a été proposée pour prendre en compte la non-stationnarité d'une série est la transformée de Fourier avec fenêtre, appelée également méthode de Gabor. Cette méthode n'est pas satisfaisante car la taille de la fenêtre étant fixe, la représentation de signaux ayant des composantes de tailles différentes de celle de la fenêtre est impossible. En raison de ces limites, la transformée en ondelettes a été introduite au début des années 1980 pour traiter des signaux sismiques (Grossmann A., Morlet J., 1984). L'intérêt des ondelettes (il en existe plusieurs types) réside dans la possibilité de suivre l'évolution du signal à la fois en temps et fréquence dans un même diagramme (appelé scalogramme).

L'Illustration 15 présente un exemple simple d'une série temporelle composée dans une première partie d'un signal périodique (T1=0.05) puis à partir d'un temps t=0.5, la période de la série change (T2=0.012). On a représenté 3 graphiques[3] :

- le graphique du haut représente la série ;

- le suivant représente sa transformée de Fourier : on voit apparaître les deux fréquences de la série, mais cette représentation ne permet pas de savoir si les deux fréquences sont simultanées ou différenciées dans le temps ;

- le dernier représente la transformée en ondelettes. Cette représentation permet d'identifier le changement dans la périodicité de la série.

[3] Repris du cours de Valérie Perrier « *Transformée en ondelettes continues. Théorie, applications à l'imagerie médicale* ». Cours de l'Ecole Doctorale, Orsay 2005.

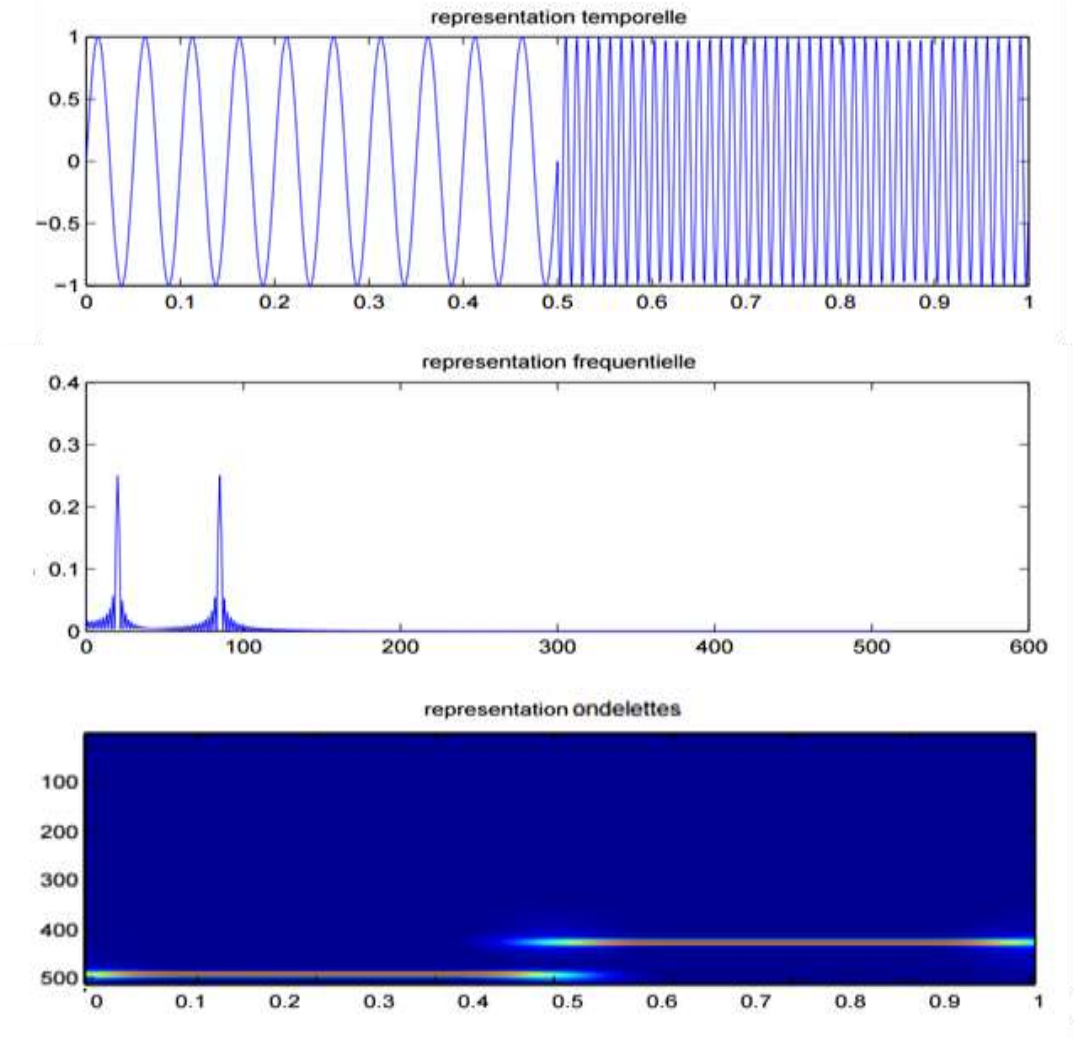


Illustration 15 : Exemples de recherche de périodicité pour une série composée d'une suite de 2 sinusoïdes. Dans la représentation en ondelettes, les couleurs représentent la puissance du spectre, les couleurs rouge représentant les périodes les plus significatives.

Les ondelettes sont utilisées en hydrologie et hydrogéologie. Par exemple, Labat (2000) a étudié des signaux de pluies et de débits de sources karstiques. Etant donné que les sources karstiques ont des propriétés marquées de non linéarité et de non stationnarité, les études spectrales classiques ne permettent pas d'identifier les variations temporelles dans la structure du signal. L'Illustration 16 présente la chronique du débit de la source de Fontestorbes ainsi que le spectre en ondelettes de ce signal. On peut reconnaître une composante annuelle, qui est constante sur toute la durée du signal et une composante pluri-annuelle qui évolue (de 3 ans les premières années, à 7 ans par la suite).

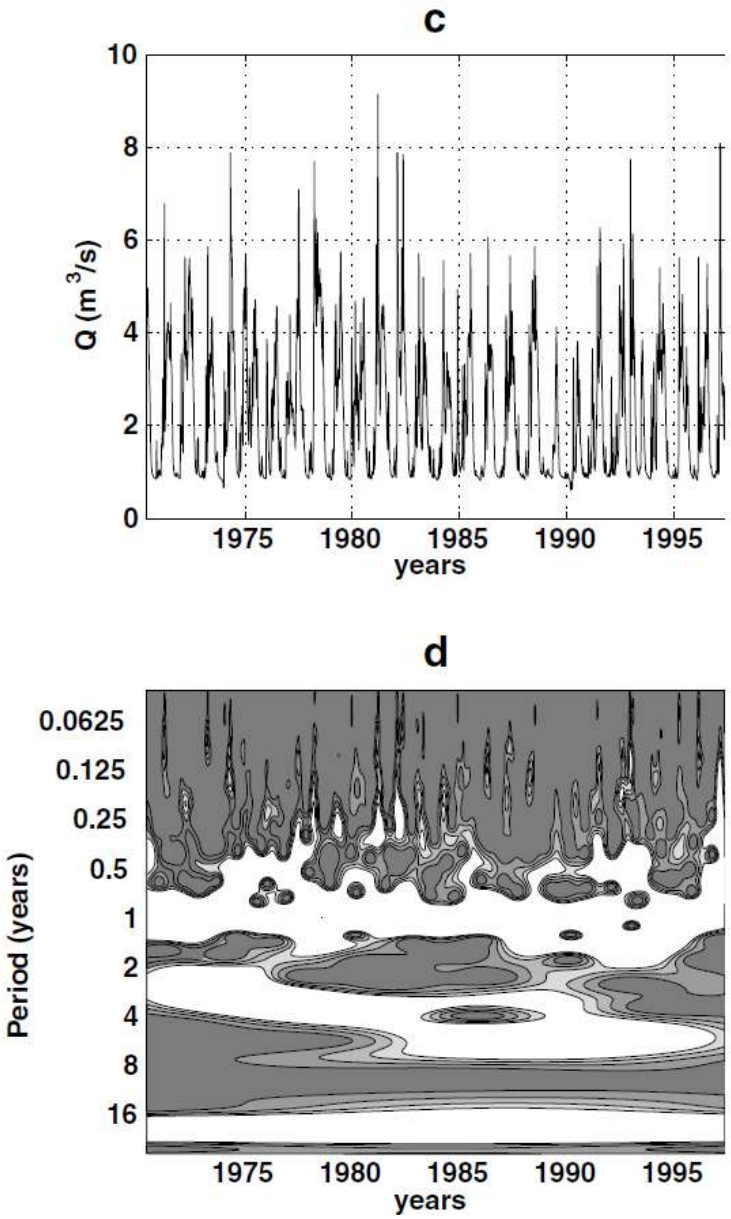


Illustration 16 : Débit hebdomadaire de la source de Fontestorbes (c) et spectre en ondelettes de Morlet du signal (d). Les coefficients élevés (périodicité significatives) correspondent à des couleurs claires.

Des éléments de définition mathématique des ondelettes sont donnés en Annexe 1

- **Cas des données non régulières**

Comme les analyses de Fourier, l'analyse par ondelettes ne peut pas être appliquée si les données ne sont pas régulières. Forster (1996) a développé une méthode qui permet une analyse équivalente sur des données avec un pas d'échantillonnage irrégulier.

A notre connaissance, très peu d'études ont utilisé cette méthode pour analyser des données de qualité des eaux souterraines. Notons tout de même l'exemple suivant (Kang et Lin, 2007), qui montre la mise en œuvre de cette méthode pour l'analyse que chroniques de qualité (nitrates, chlorures et sodium) des eaux souterraines dans un bassin versant agricole de Pennsylvanie.

L'Illustration 17 présente les résultats de l'analyse en ondelettes des chroniques de nitrates, chlorures et sodium dans un piézomètre du bassin versant. On peut identifier sur les graphiques présentés les périodicités les plus significatives avec la couleurs rouge. On peut ainsi remarquer qu'après 1991, les concentrations en nitrates montrent des périodicités intermittentes mais importantes, que l'on ne retrouve pas avec la même puissance pour les paramètres chlorures et sodium, probablement car les sources de ces 2 éléments sont plus diverses. Avant 1986, l'absence de périodicité pour les trois paramètres considérés est principalement due à la présence de nombreuses lacunes dans les chroniques. Bien que la méthode utilisée puisse traiter les données non régulièrement distribuées, des lacunes trop importantes peuvent empêcher la détection de périodicités.

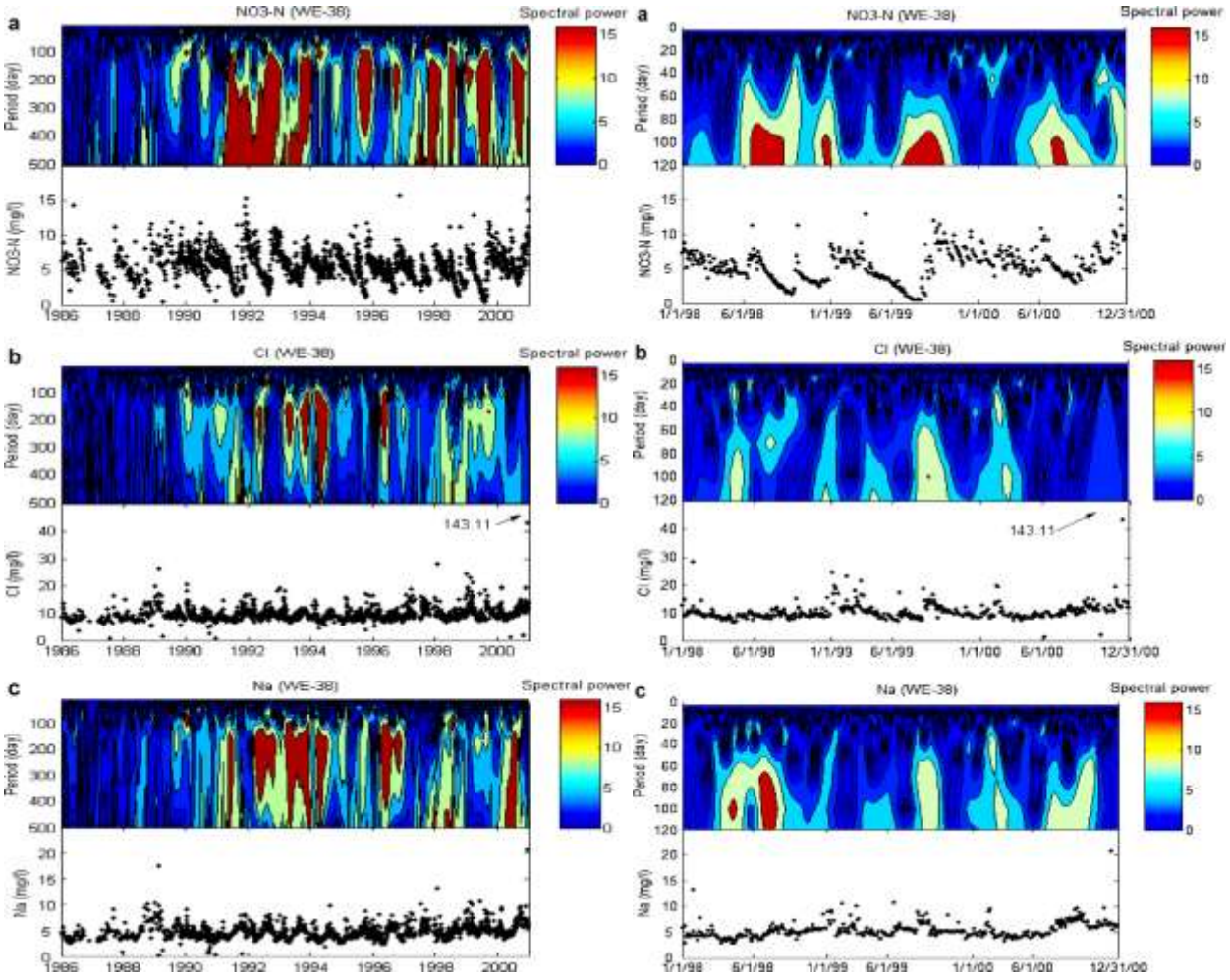


Illustration 17 : Spectre en ondelettes des concentrations en nitrates (a), chlorures (b) et sodium (c) et évolution temporelle des concentrations dans le piézomètre WE-38 sur la période 1986-2000 (à gauche) et 1998-2000 (à droite).

4.4. METHODE RECOMMANDEE POUR L'IDENTIFICATION DES VARIATIONS CYCLIQUES

Les tests d'identification des variations cycliques ont été réalisés sur des données réelles extraites d'ADES (http://www.ades.eaufrance.fr/). Comme pour les tests des méthodes de régression et de stationnarisation, seules les chroniques disposant d'au moins 25 données ont été utilisées, et ce quelle que soit la substance considérée. Tous les tests réalisés ne sont pas présentés dans le rapport afin de ne pas le surcharger.

La méthode basée sur le périodogramme de Lom-Scargle est recommandée pour une application sur des chroniques de qualité des eaux souterraines en raison de sa relative simplicité de mise en œuvre et parce qu'elle accepte des données irrégulièrement espacées. Les paragraphes suivant présentent donc quelques exemples d'application de la méthode basée sur le périodogramme de Lomb-Scarge sur des chroniques de qualité des eaux souterraines aux caractéristiques variables.

4.4.1. Application de la méthode de Lomb-Scargle sur des données stationnaires

- **Cas simple**

L'Illustration 18 montre la recherche des variations cycliques sur une chronique d'évolution des concentrations en nitrate au point 00955X0050/F1 (nappe du Bajocien). Cette chronique est stationnaire (test de Mann-Kendall, p-value = 0,57) c'est-à-dire qu'elle ne montre pas de tendance.

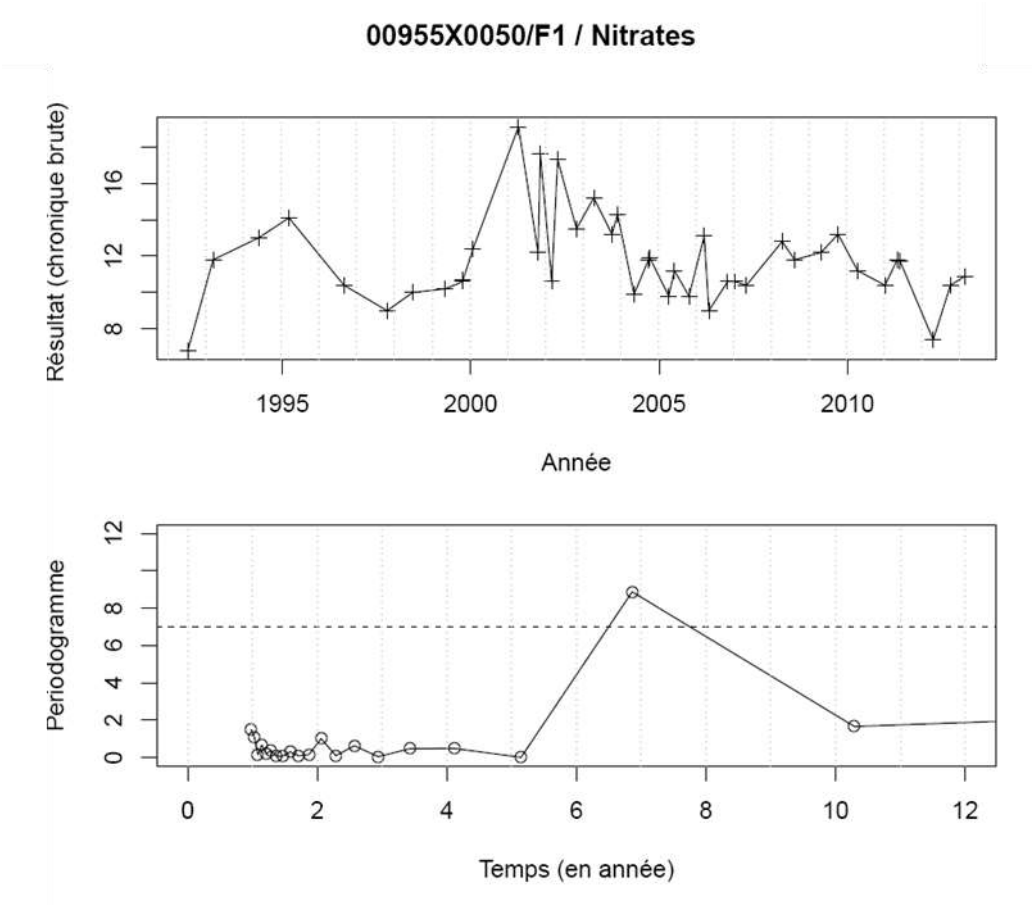


Illustration 18 : Chronique d'évolution des concentrations en nitrate au point 00955X0050/F1 et périodogramme de Lomb-Scargle associé.

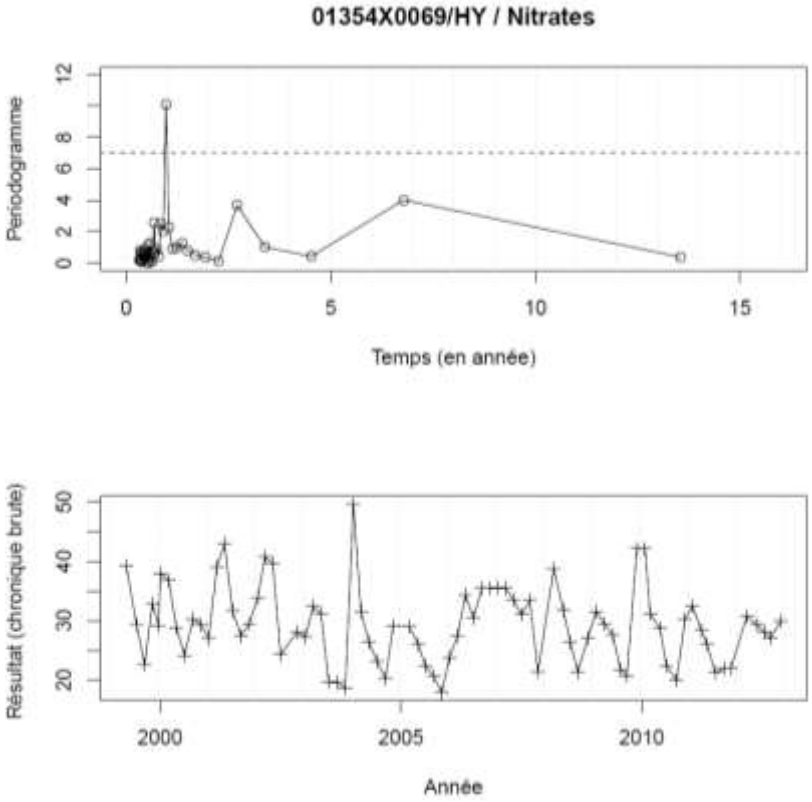
S'il n'est pas évident d'identifier un ou des cycles à l'œil nu sur la chronique brute, la méthode de Lomb-Scargle révèle en revanche une cyclicité du signal selon une période de 7 ans environ (cette périodicité est très fréquente sur des nappes à double cycles annuels et pluriannuels et ainsi que sur des hydrogrammes de cours d'eau importants). Le résultat est représenté sur le périodogramme par un point situé au-dessus de la ligne en pointillé qui représente le seuil de significativité du test avec un intervalle de confiance de 95%. Cette périodicité serait à comparer

aux grands cycles hydroclimatiques qui impactent la région d'où cette chronique est extraite. Une corrélation positive entre les cycles pluriannuels d'évolution des concentrations en nitrate et les évolutions climatiques pourrait alors expliquer les augmentations des concentrations observées durant les périodes 1993-1996, 2001-2004 et 2008-2011. Ces informations constituent des éléments de gestion importants car elles permettent de prédire la réponse de l'aquifère en terme de concentration en nitrate en fonction des futures fluctuations hydroclimatiques.

- **Effet du nombre de données dans la chronique**

L'Illustration 19 compare la recherche des variations cyclique selon les méthodes du corrélogramme simple et avec le périodogramme de Lomb-Scargle. La série temporelle est une chronique d'évolution des concentrations en nitrate, stationnaire, à fréquence de prélèvements élevée (6 analyses par an en moyenne) et régulièrement échantillonnée.

a)



b)

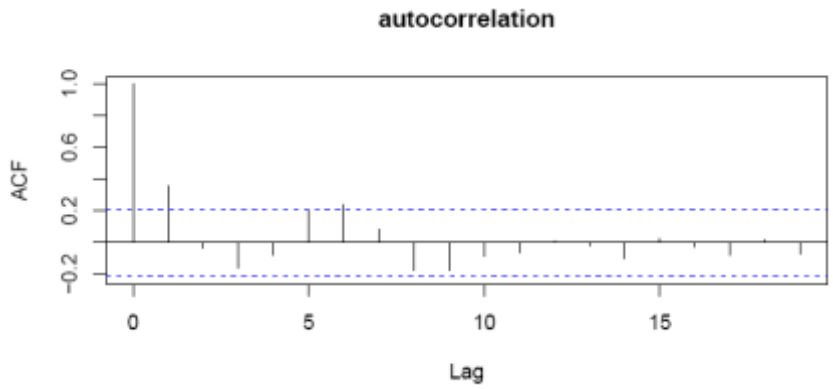
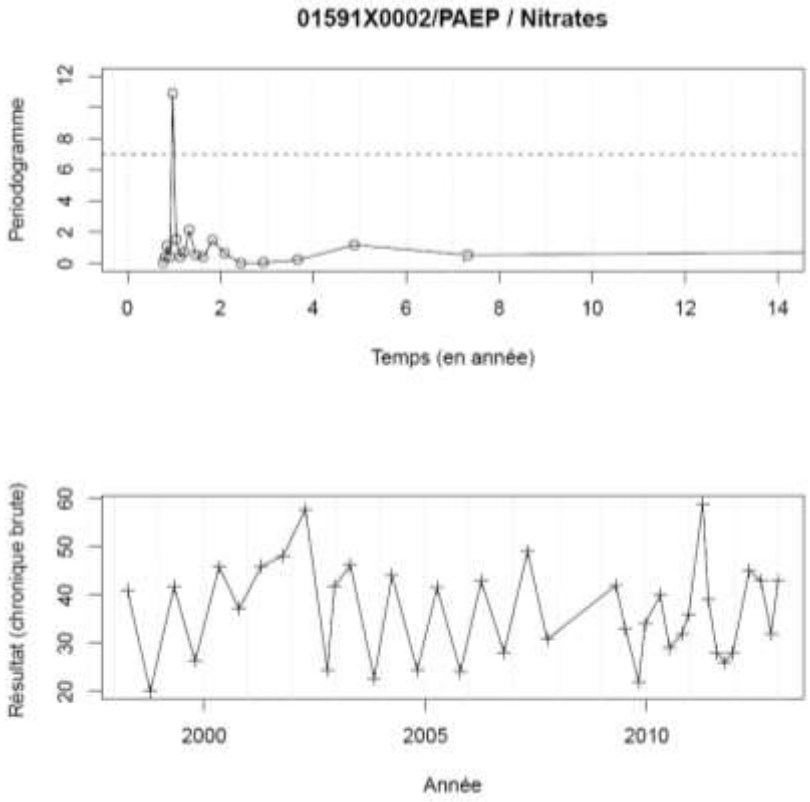


Illustration 19 : Comparaison du périodogramme de Lamb-Scarge (a) et du corrélogramme (b) calculé à partir d'une chronique d'évolution des concentrations en nitrate avec beaucoup d'analyses et régulièrement échantillonnée.

Sous ces conditions, on constate que les deux méthodes donnent les mêmes résultats avec l'identification des cycles annuels qui structurent significativement la chronique.

L'Illustration 20 présente la même comparaison dans le cas d'un suivi des concentrations en nitrate en eau souterraine régulier mais moins fréquent (2 analyses par an en moyenne).

a)



b)

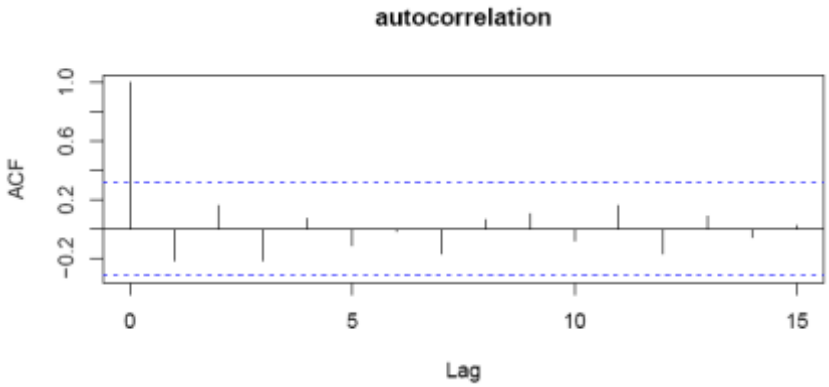


Illustration 20 : Comparaison du périodogramme de Lamb-Scarge (a) et ducorrélogramme (b) calculé à partir d'une chronique d'évolution des concentrations en nitrate régulièrement échantillonnée mais à fréquence d'échantillonnage faible

Dans ce cas, il est intéressant de constater que la méthode du calcul ducorrélogramme simple ne permet pas d'identifier de variation cyclique alors que le périodogramme de Lomb-Scarge montre clairement une période significative d'un an. Cette méthode apparait donc plus puissante pour identifier les cycles potentiellement dans les chroniques de qualité des eaux souterraines, notamment dans le cas de faibles fréquences d'échantillonnage.

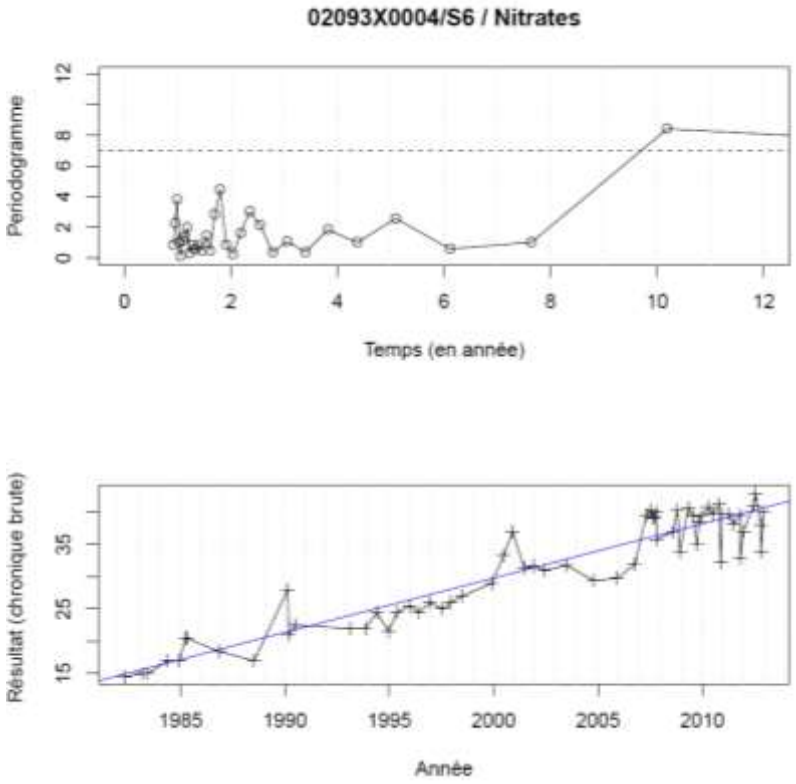
4.4.2. Application de la méthode de Lomb-Scargle sur des données non-stationnaires

Il peut être parfois nécessaire de recourir à un traitement préalable des données brutes avant de rechercher les variations cycliques. C'est par exemple le cas lorsque la chronique n'est pas stationnaire comme présenté en exemple dans l'Illustration 21a. La chronique d'évolution des concentrations en nitrate montre clairement une tendance à la hausse bien détectée par le test de Mann-Kendall (p-value = $2{,}9 \times 10^{-31}$). Le périodogramme de Lomb-Scargle calculé sur cette chronique ne permet pas d'identifier de cycles périodiques. De plus, il dérive à partir de 10 ans ce qui indique un problème d'application de la méthode.

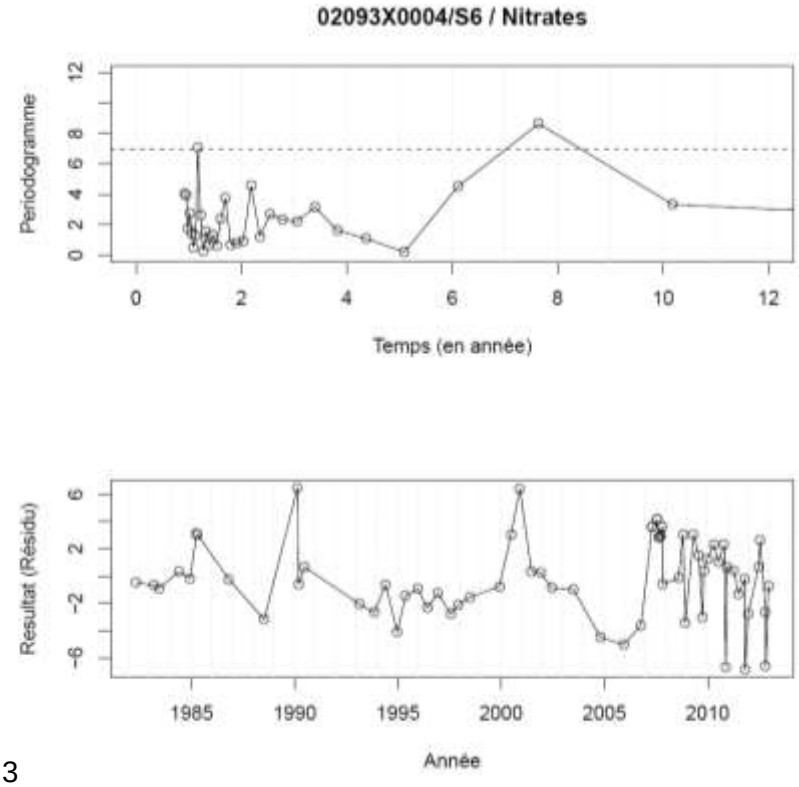
La procédure recommandée est de stationnariser la chronique avant de rechercher les variations cycliques éventuelles. Cette opération consiste à modéliser la tendance (monotone ou plus complexe) afin de la soustraire aux données brutes. Dans l'exemple en Illustration 21b, la tendance monotone identifiée par le test de Mann-Kendall et modélisée avec le test de Sen a été retranchée à la chronique brute d'évolution des concentrations en nitrate. La chronique résultant de cette opération, le résidu, est stationnaire. Le périodogramme de Lomb-Scargle calculé à partir du résidu révèle maintenant que l'évolution des concentrations en nitrate au point 02093X0004/S6 est structurée selon 2 cycles : le premier de période annuelle et le second d'une période de 8 ans environ.

Ce dernier exemple montre bien l'intérêt de stationnariser les chroniques avant de rechercher les variations cycliques. Il confirme aussi la nécessité de procéder à l'analyse « cas par cas » des chroniques d'évolution de la qualité des eaux souterraines lorsque l'on souhaite en faire la décomposition. Ce constat suggère l'impossibilité d'automatiser complétement la procédure de décomposition des chroniques. Le développement d'un outil informatique de décomposition des séries temporelles ou son intégration dans un outil existant comme HYPE devra prendre en compte cette contrainte en laissant la possibilité à l'opérateur d'orienter la succession des tests à réaliser.

a)



b)



3

Illustration 21 : Calcul du périodogramme de Lomb-Scargle sur une chronique brute d'évolution des concentrations en nitrate non stationnaire (a) et sur les résidus après soustraction de la tendance monotone (b).

5. Conclusions et perspectives

- **Bilan de l'action**

De nombreuses méthodes aux caractéristiques différentes permettent de décomposer des signaux temporels. Toutefois, parmi le panel des possibles, peu de méthodes apparaissent adaptées aux caractéristiques spécifiques des chroniques d'évolution de la qualité des eaux souterraines. L'étude menée dans le cadre de la convention ONEMA-BRGM 2013-2015 oriente le choix des méthodes à appliquer en proposant un arbre de décision qui établit une procédure pas à pas pour décomposer des chroniques d'évolution des concentrations en éléments dans les eaux souterraines (Illustration 22).

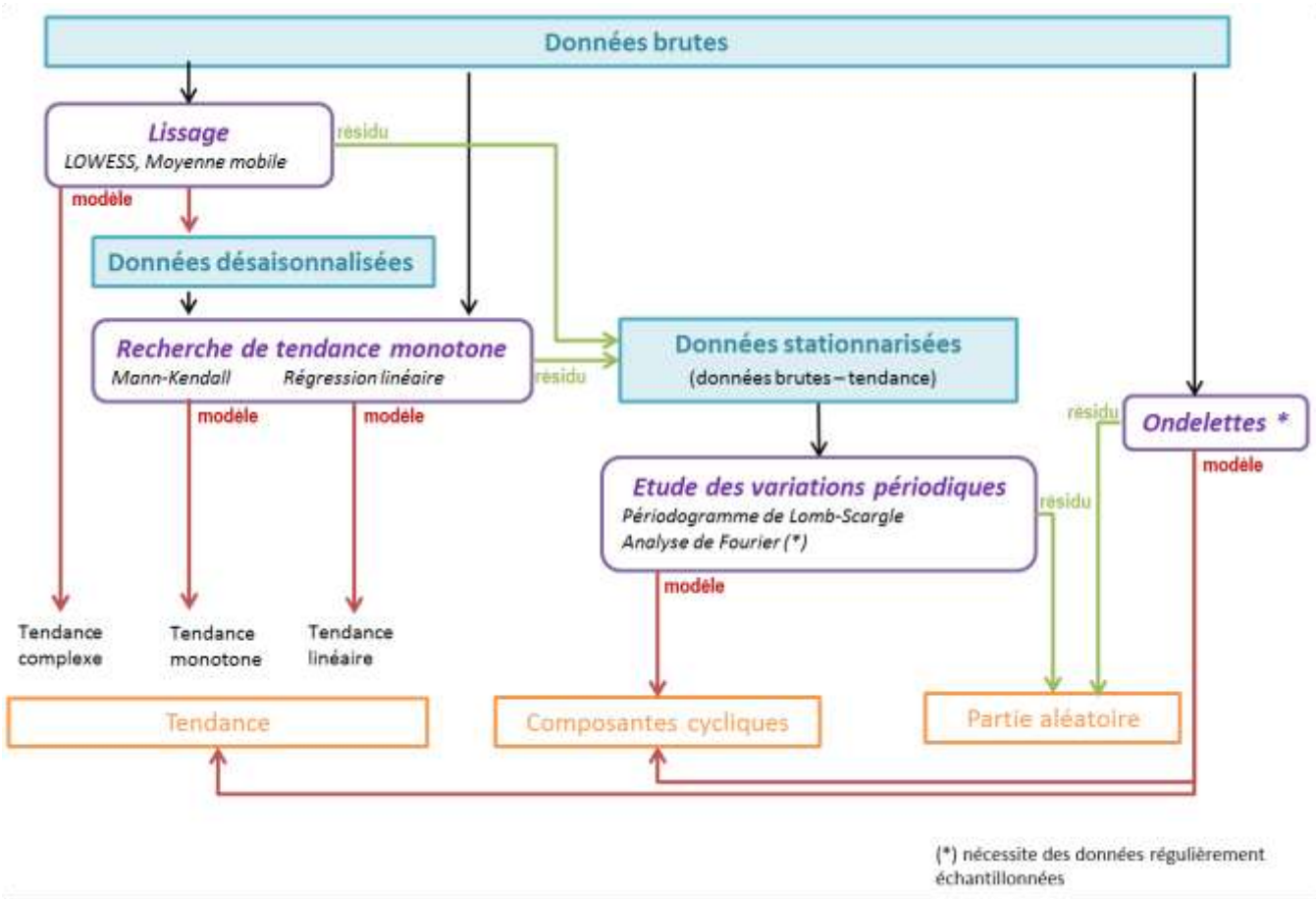


Illustration 22 : Arbre décisionnel pour la décomposition des chroniques de qualité des eaux souterraines selon les méthodes statistiques recommandées par Croiset et al. 2016.

Trois composantes ont été identifiées comme pertinentes à extraire des chroniques : l'évolution non-stationnaire ou tendancielle, l'évolution cyclique périodique et l'évolution aléatoire de la chimie des eaux souterraines. Chaque évolution étant reliée à un forçage spécifique du comportement des éléments dans les eaux souterraines (impact des activités anthropiques, impact des cycles naturels hydroclimatiques et bruit environnemental couplé à l'incertitude analytique respectivement) leur individualisation permet d'orienter les plans de gestion en estimant les facteurs explicatifs. Les méthodes proposées servent à la fois à extraire une

composante spécifique des séries temporelles mais aussi à prétraiter les données brutes pour l'application d'autres tests statistiques proposés dans l'arbre décisionnel.

Pour extraire la composante tendancielle des chroniques d'évolution de la qualité des eaux souterraines, les auteurs recommandent le test de Mann-Kendall (et ses dérivés) pour l'identification et l'extraction des tendances monotones. Pour les tendances plus complexes, l'algorithme de LOWESS apparaît tout à fait adapté. Cette méthode est d'ailleurs adoptée par certains pays européens comme l'Autriche et la Roumanie dans le cadre de l'identification des tendances demandée par la DCE. Un coupage des 2 méthodes, lissage de LOWESS suivi d'un test de Mann-Kendall sur les résidus, permet d'identifier certaines tendances non détectables sans prétraitements. Toutefois, cette méthode, si robuste et puissante soit-elle, nécessite de faire des choix sur les paramètres de lissage, notamment sur la taille de la fenêtre de lissage. Or le choix de ce paramètre, qui impacte fortement la qualité de l'identification des tendances complexes, ne peut être automatisé.

Les variations cycliques périodiques qui structurent potentiellement les chroniques d'évolution de la qualité des eaux souterraines peuvent être mises en évidence et extraites grâce au périodogramme de Lomb-Scargle. Cette méthode présente l'avantage de pouvoir être appliquée sur des données irrégulièrement espacées ce qui est très souvent le cas dans le domaine des suivis environnementaux. Les tests et les comparaisons de méthodes menés sur des données réelles de la qualité des eaux souterraines ont révélé la puissance du périodogramme de Lomb-Scargle pour identifier les cycles périodiques, notamment dans les conditions de faibles fréquences d'échantillonnage. Cette méthode nécessite néanmoins de stationnariser les chroniques avant traitement, les tendances « masquant » les potentielles évolutions périodiques des concentrations en éléments dans les eaux souterraines.

- **Perspectives**

Après avoir extrait les composantes des chroniques d'évolution de la qualité des eaux souterraines, il est intéressant de les comparer à des variables potentiellement explicatives, comme par exemple des chroniques piézométriques ou pluviométriques qui traduisent les évolution hydroclimatiques naturelles, les historiques d'utilisation et d'application d'une substance donnée pour les évolutions anthropiques ou bien la valeur de l'incertitude analytique qui affecte les analyses des concentrations.

Dans l'exemple suivant, Helsel et Hirsch (1992) cherchent à estimer l'évolution des particules en suspension dans une rivière. Ce paramètre est visiblement corrélé avec le débit de la rivière. Dans un premier temps les auteurs calculent la régression par l'algorithme LOWESS de la masse de particules transportée en fonction du débit (cf Illustration 23).

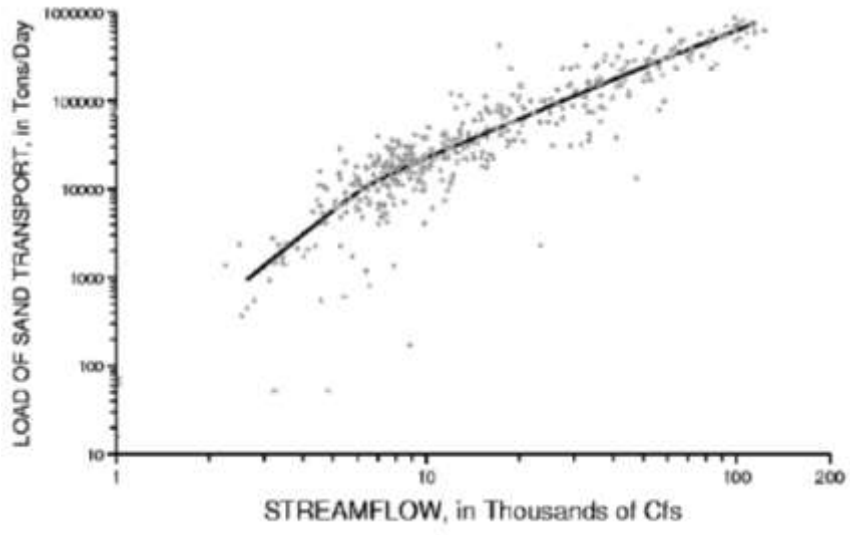


Illustration 23 : Masse de sable transporté en fonction du débit de la rivière Colorado à Lees Ferry (Colorado) entre 1949 et1964

Il est alors possible de retrancher au signal ce qui peut être expliqué par cette variable. La tendance d'évolution de la masse de particules qui n'est pas expliquée par les variations de débit en fonction du temps peut ainsi être calculée.

Un autre moyen d'estimer les corrélations entre les différents signaux est de calculer la corrélation croisée entre 2 variables. Ce calcul est sensiblement équivalent à celui de l'autocorrélation. Toutefois la corrélation croisée nécessite, comme le calcul de l'autocorrélation, que les chroniques soient régulièrement échantillonnées ce qui est très rarement le cas des données environnementales. Ainsi, de manière similaire à ce qu'il a proposé pour le calcul de l'autocorrélation, Scargle a proposé une méthode pour calculer la corrélation croisée de données non régulièrement échantillonnées.

Les méthodes d'estimation des corrélations entre les différentes composantes des chroniques d'évolution de la qualité des eaux souterraines et les variables potentiellement explicatives pourraient intégrer une procédure globale d'analyse des signaux temporels en hydrogéologie. Cette procédure pourrait être implémentée dans un outil informatique facilitant la mise en œuvre des méthodes recommandées. Cet outil ne pourrait cependant pas traiter de manière totalement automatique un jeu complet de données comme le fait actuellement l'outil d'identification des tendances HYPE. L'opérateur devrait en effet intervenir à différentes étapes de la procédure de décomposition des chroniques, à la fois pour choisir le test ou le prétraitement à appliquer, et pour définir les paramètres de certaines méthodes comme le lissage par l'algorithme de LOWESS par exemple.

6. Bibliographie

Baran N., Gourcy L., Lopez B., Bourgine B., Mardhel V. (2009) – Transfert des nitrates à l'échelle du bassin Loire-Bretagne. Phase 1 : temps de transfert et typologie des aquifères. Rapport BRGM RP-54830-FR, 105 p.

Cleveland W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association **74**, 829-836.

Croiset N., Lopez B. (2013) – HYPE : Outil d'analyse statistique des séries temporelles d'évolution de la qualité des eaux souterraines – Manuel d'utilisation. BRGM/RP-63066-FR. 64 p., 33 fig.

Daliakopoulos I.N, Coulibaly P., Tsanis I.K. (2005). Groundwater level forecasting using artificial neural networks. Journal of Hydrology **309**, 229-240.

Eckner A. (2012). Algorithms for Unevenly-Spaced Time Series: Moving Averages and Other Rolling Operators. Working document available at http://www.eckner.com/research.html

Forster G. (1996). Wavelets for period analysis of unevenly sampled time series. Astronomical Journal **112**, 1709-1729.

Hamed K. H. et Rao A. R., (1998) - A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology **204,** 182-196.

Helsel D.R., Hirsch R.M., (1992) - Statistical method in water resources, Studies in Environmental Science 49, Elsevier, Amsterdam

Jones A.L. et Smart P.L. (2005). Spatial and temporal changes in the structure of groundwater nitrate concentration time series (1935-1999) as demonstrated by autoregressive modelling. Journal of Hydrology **310**, 201-215.

Kang S. et Lin H. (2007). Wavelet analysis of hydrological and water quality signal in an agricultural watershed. Journal of Hydrology **338**, 1-14.

Kwiatkowski D., Philips P., Schmidt P., Shin Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics **55**, 159-178.

Labat D., Abadou R. et Mangin A. (2000). Rainfall-runfoff relations for karstic springs. Part II: continuous wavelet and discrete orthogonal multiresolution analyses. Journal of Hydrology **238**, 149-178.

Lalot E. (2014). Analyse des signaux piézométriques et modélisation pour l'évaluation quantitative et la caractérisation des échanges hydrauliques entre aquifères alluviaux et rivières – Application au Rhône. Thèse de l'Ecole Nationale Supérieure des Mines de Saint-Etienne

Lopez B. et Leynet A. (2011). Evaluation des tendances d'évolution des concentrations en polluants dans les eaux souterraines. Revue des méthodes statistiques existantes et recommandations pour la mise en œuvre de la DCE. Rapport BRGM/RP-59515-FR

Lopez B., Baran N., Bourgine B., Brugeron A., Gourcy L. (2012) - Pollution diffuse des aquifères du bassin Seine-Normandie par les nitrates et les produits phytosanitaires : temps de transfert et tendances. Rapport final BRGM/RP-60402-FR ; 326p.

Lopez B., Croiset N., Surdyk N., Brugeron A. (2013) – Développement d'outils d'aide à l'évaluation des tendances dans les eaux souterraines au titre de la DCE. Rapport final. BRGM/RP-61855-FR, 98 p., 45 ill., 1 ann.

Lopez B., Baran N., Bourgine B., (2015) - An innovative procedure to assess multi-scale temporal trends in groundwater quality: Example of the nitrate in the Seine–Normandy basin, France. Journal of Hydrology 522, 1–10.

Rehfeld K., Marwan N., Heitzig J. et Kurths J. (2011). Comparison of correlation analysis techniques for irregularly sampled time series. Nonlinear Processes in Geophysics **18**, 389-404.

Scargle J.D. (1981). Studies in Astronomical Time Series Analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal **263**, 835-853.

Schulz M. et Statteger K. (1997). Spectrum : Spectral analysis of unevenly spaced paleoclimatic time series.

Welch A.H., Lico M.S. et Hughes J.L. (1988). Arsenic in Ground Water of the Western United States. Ground Water **26**, 333-347

Yue S. et Wang C. (2004). The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. Water Resources Management **18**, 201-218.

Annexe 1 : Description de méthodes d'identification des variations cycliques

- Analyse spectrale

Dans les analyses spectrales, les fréquences que l'on peut étudier sont bornées entre zéro (l'origine des axes) et la plus haute fréquence détectable $f_{Nyquist}$ (fréquence de Nyquist). Les puissances ont des valeurs arbitraires dépendantes du nombre total de données analysées, mais aussi de la fonction de densité spectrale employée et donc de la méthode d'analyse spectrale.

De manière simple, on cherche à décomposer le signal observé en une somme d'harmoniques. La j[ème] harmonique représente un phénomène d'une période égale à $(n-1)*\Delta t/j.\ j$ varie de 1 à $(n-1)/2$, donc, en théorie, les périodes de temps des harmoniques identifiables varient de $2\Delta t$ à $(n-1)\ \Delta t$. En pratique, les cycles ayant des périodes comprises entre $4\Delta t$ et un quart (ou même un sixième) du temps total, peuvent être identifiés.

Les méthodes d'analyse spectrale exigent la stationnarité et (théoriquement) l'ergodocité des données (cf définitions en 2.3) et également, pour la plupart des méthodes, linéarité et régularité.

On considère un processus discret $x(n)$ où $n=0,\pm1,\pm2,\dots,\pm N)$ aléatoire et stationnaire du second ordre de moyenne nulle et dont la fonction d'autocorrélation est :

$$r_x(k)=\frac{E\{(x(n+k)-\bar{x})*(x(n)-\bar{x})\}}{\sigma^2}$$

L'autocorrélation est simplement calculée comme la moyenne sur l'intervalle fini $[0,N]$:

$$\hat{r}_{Per}(\propto)=\frac{1}{N}\sum_{n=0}^{N-\propto-1}x^*(n)x(n+\propto)$$

Pour $\propto<0$, on utilisera la propriété $\hat{r}_x(-\propto)=\hat{r}_x^*(\propto)$

La transformée de Fourier est le spectre de $x(n)$: $S_x(\omega)=\sum_{k=-\infty}^{\infty}r_x(k)\exp(-i\omega k)$. Elle ne s'applique que sur des séries infinies. Le périodogramme est le carré du module de la transformée de Fourier

Pour mettre en évidence les composantes périodiques d'un signal, il faut qu'elles soient à caractère infini. Pour satisfaire à cette condition, avant d'appliquer la Transformée de Fourier, toute série de données à caractère fini est multipliée par une fenêtre rectangulaire. Des discontinuités apparaissent cependant au niveau des extrémités de chaque fenêtre répétée, dues à l'effet des bords créé par les fenêtres rectangulaires. La Transformée de Fourier montre un pic principal entouré par des petits pics secondaires dus à l'effet de troncation. Il existe plusieurs sortes de fenêtres spectrales de pondération (Bartlett, Tukey, Parzen, Welch…) qui permettent d'éliminer les pics « parasites ».

Si l'on n'échantillonne pas un signal pour un nombre entier de points par période, il y aura une erreur dans les coefficients de la série de Fourier (appelée fuite spectrale). La série de Fourier discrète calculée donnera des fréquences autres que celles du signal. Dans le cas idéal, il y a

un nombre entier de points par période. En pratique, cela n'est pas possible car on ne connaît pas a priori les fréquences d'un signal.

- Périodogrammes

Dans le cas des périodogrammes, on cherche à estimer l'autocorrélation d'un signal de longueur finie. Ces méthodes font l'hypothèse d'une autocorrélation nulle en dehors d'une certaine fenêtre. Cette estimation s'appuie sur un signal modifié $x_N(n)$, qui est la troncature du signal initial $x(n)$:

$x_N(n) = x(n)w_N(n)$ où $w_N(n) = 1$ si $0 \leq n \leq N$ et $w_N(n) = 0$ sinon.

Dans la méthode du périodogramme modifié, la fenêtre de troncature est modifiée. La fenêtre est un compromis entre la résolution spectrale et la fuite de spectre. La fenêtre rectangulaire permet d'avoir la meilleure résolution spectrale mais pose de problèmes de fuite spectrale. Les autres fenêtres les plus couramment utilisées sont les fenêtres triangulaires, de Hamming et de Blackman.

- Périodogramme de Lomb-Scargle

Cette méthode permet d'estimer la densité spectrale pour les séries non régulières. Au lieu de considérer le produit de cosinus et sinus directement, Scargle a modifié le périodogramme standard pour trouver d'abord le temps T tel que la paire de sinusoïdes soient orthogonales aux temps d'échantillonnage t_j

$FT_X(\omega) = F_0 \sum[A \, \mathrm{X_N} \cos \omega t'_n + iB_N \sin \omega t'_n]$ où

$F_0(\omega) = \left(\frac{N}{2}\right)^{1/2} \exp(-i\omega t), \quad A(\omega) = (\sum \cos^2(\omega t'_n)\,)^{-1/2}, \quad B(\omega) = (\sum \sin^2(\omega t'_n)\,)^{-1/2}\,, \quad t'_n = t_n - \tau(\omega)$ et $\tau(\omega) = \frac{1}{2\omega} \tan^{-1}\left(\frac{\sum \sin(2\omega t_n)}{\sum \cos(2\omega t_n)}\right)$

Dans R sont codés le périodogramme classique, le périodogramme de Lomb-Scargle, celui de Welch, la méthode de Thomson Multitaper.

Scargle (1982) a proposé une méthode basée sur la transformée de Fourier. Il propose de calculer dans un premier temps la transformée de Fourier de la série du carré de la valeur absolue de la transformée de Fourier, qui est en fait le spectre de puissance.

$$P_X = |FT_X(\omega)|^2$$

$$\rho_X(t) = \mathcal{F}^{-1}[P_X(\omega)]$$

Cette méthode est décrite plus en détail dans le paragraphe 4.2.2.

Il peut être intéressant de prendre en compte l'autocorrélation liée à l'échantillonnage. Scott (1976) a montré que la fonction d'autocorrélation d'un processus observé et la fonction d'autocorrélation théorique $\rho_X^{true}(t)$sont reliées par :

$$\rho_X(t) = \rho_X^{true}(t) * \rho_S(t)$$

Où $\rho_S(t)$ est la fonction d'autocorrélation liée à l'échantillonnage. Scargle propose d'estimer cette fonction ainsi :

$FT_S(\omega) = F_0 \sum[A \cos \omega t'_n + i \sin \omega t'_n]$ et $\rho_S(t) = \mathcal{F}^{-1}[|FT_S(\omega)|^2]$

- Ondelettes :

Dans la transformée de Fourier, on décompose le signal dans une base de fonctions à un paramètre : $\{\exp(\mathrm{i}\omega_\mathrm{n}\mathrm{t}), \mathrm{n} \in \mathbb{R}\}$. Dans l'analyse en ondelettes, on utilise la base à deux paramètres : $\{\psi_{\mathrm{a},\tau}(\mathrm{t}), (\mathrm{a},\tau) \in \mathbb{R}_+^* \times \mathbb{R}\}$. Cette base est construite par translation et dilatation d'une unique fonction ψ :

$$\psi_{\mathrm{a},\tau}(\mathrm{t}) = \frac{1}{\sqrt{\mathrm{a}}}\psi\left(\frac{\mathrm{t}-\tau}{\mathrm{a}}\right)$$

Le paramètre a peut être interprété comme le facteur de dilatation ($\mathrm{a} > 1$) ou de contraction ($\mathrm{a} < 1$) de la fonction d'ondelette $\psi(\mathrm{t})$, correspondant à plusieurs échelles d'observations. Le paramètre τ peut être interprété comme la translation dans le temps de la fonction $\psi(\mathrm{t})$, qui permet l'étude du signal $\mathrm{x}(\mathrm{t})$ localement autour de τ.

Il existe de nombreuses formes d'ondelettes, le choix de l'ondelette à utiliser dépend de l'application envisagée. La plus couramment utilisée est l'ondelette de Morlet. La transformée en ondelettes s'écrit de manière générale :

$$C_f(a,b) = \sum_{\alpha=1}^{N} f(t_\alpha)\overline{\psi_{(a,b)}(t_\alpha)}$$

Centre scientifique et technique
Direction Eau, Environnement et Ecotechnologies
3, avenue Claude-Guillemin
BP 36009 – 45060 Orléans Cedex 2 – France – Tél. : 02 38 64 34 34
www.brgm.fr