HYPE

Outil d'analyse statistique des séries temporelles d'évolution de la qualité des eaux souterraines

Manuel d'utilisation

Rapport final

N. Croiset, B. Lopez (BRGM)

Document élaboré dans le cadre de :
La Directive Cadre sur l'Eau

- **AUTEURS**

Nolwenn CROISET, hydrogéologue (BRGM), n.croiset@brgm.fr

Benjamin LOPEZ, hydrogéologue (BRGM), b.lopez@brgm.fr

- **CORRESPONDANTS**

Onema : Nolwenn BOUGON, Nolwenn.Bougon@onema.fr

Onema : Gaelle DERONZIER, Gaelle.Deronzier@onema.fr

Partenaire : Laurence GOURCY, correspondante Onema (BRGM), l.gourcy@brgm.fr

Droits d'usage : Accès libre

Niveau géographique : National

Couverture géographique : Métropole

Citations locales : Masses d'eau souterraines

Niveau de lecture : Professionnels, experts

- **RESUME**

Afin de répondre aux exigences de la Directive 2000/60/CE et de la Directive fille eaux souterraines 2006/118/CE qui demandent aux Etats Membre d'identifier, selon une approche statistique, les tendances d'évolution des concentrations en contaminants dans les eaux souterraines les agences et les offices de l'eau ont exprimé, dès 2009, un besoin d'accompagnement scientifique sur ces thématiques spécifiques.

Plusieurs actions ont ainsi été menées depuis 2010 dans le domaine de l'identification des tendances d'évolution de la qualité des eaux souterraines dont, à l'échelle nationale :

- Revue bibliographique et tests des méthodes statistiques existantes pour l'identification des tendances d'évolution des contaminants dans les eaux souterraines réalisés sous convention ONEMA-BRGM en 2010 (Lopez et Leynet, 2011).

- Projet national tendance (convention ONEMA-BRGM 2012) mené en partie dans le cadre de la révision de l'état des lieux 2013 avec l'élaboration de fiches « tendance » par masse d'eau souterraine (Lopez et al., 2013).

Ces actions nationales, complétées d'autres études menées aux échelles des grands bassins (Baran et al., 2009 ; Lopez et Baran, 2011 ; Lopez et al., 2012), ont permis d'acquérir l'expérience théorique et pratique nécessaire à l'élaboration d'une méthode robuste pour l'identification des tendances et des ruptures d'évolution temporelle de la qualité des eaux souterraines.

Après avoir pris connaissance de la méthode proposée, les agences, les offices de l'eau et les DEAL ont exprimé le besoin d'avoir un outil leur permettant de mettre en œuvre facilement ces approches statistiques. Cet outil dédié à l'analyse statistique des séries temporelles de la qualité des eaux souterraines répond à deux impératifs supplémentaires :

- fonctionner à partir du format SANDRE des données extraites d'ADES afin de prendre en compte les analyses réalisées au sein des réseaux RCS et RCO dont une des finalités est « d'identifier les tendances à la hausse significative et durable des polluants dans les eaux souterraines » (2006/118/CE) ;

- être compatible avec le format de l'outil SEEE (système d'évaluation de l'état des eaux) développé dans un langage informatique proche de R.

Un outil d'analyse statistique des séries temporelles d'évolution de la qualité des eaux souterraines, appelé HYPE, a ainsi été développé sous environnement R dans le cadre des conventions ONEMA-BRGM 2012 et 2013. Il permet à la fois de caractériser les séries temporelles d'évolution des contaminants dans les eaux souterraines en calculant les statistiques de base de manière automatique et d'identifier des tendances et des ruptures des séries chronologiques.

L'ensemble des méthodes compilées dans l'outil et des tests effectués sur des chroniques réelles tirées de la base de données ADES sont présentés dans le rapport relatif à l'étude (Lopez el al., 2013). Le présent document correspond au manuel d'utilisation détaillé de l'outil HYPE. Une plaquette synthétique de prise en main rapide de l'outil est fournie en guise de synthèse opérationnelle du manuel d'utilisation (partie 5).

- **MOTS-CLES : OUTIL INFORMATIQUE, LANGAGE R, ANALYSE STATISTIQUE, TENDANCE, RUPTURE, INVERSION, SERIES TEMPORELLES, QUALITE DES EAUX, EAUX SOUTERRAINES**

SYNTHESE POUR L'ACTION OPERATIONNELLE

Utilisation de HYPE - Commandes à saisir

Le répertoire de travail doit contenir les différents modules de l'outil et dans le cas où les unités sont définies par leur libellé complet, le fichier Unite_SANDRE.txt »

Définition du répertoire de travail

```
> setwd("chemin")
```

"*chemin*"est votre répertoire de travail.
Par exemple : "D:/Travail/tendance/outilR/"

Lecture des données

```
> source("lecture.r")
```

Sélection du fichier contenant les données d'entrée

Caractérisation des données

```
> source("caracterisation.r")
```

→ Tableau récapitulatif des statistiques de base

→ Représentation graphique des chroniques et de la distribution des données en boîte à moustache

Recherche de tendances et de ruptures

```
> source("tendances_ruptures.r")
```

→ Tableau récapitulatif des tests effectués et de leurs résultats

→ Représentation graphique des chroniques avec les tendances et les ruptures significatives identifiées

Kendall saisonnier

```
> source("mk_saisonnier.r")
```

→ Tableau récapitulatif des tests effectués et de leurs résultats

Kendall régional

```
> source("mk_regional.r")
```

→ Tableau récapitulatif des tests effectués et de leurs résultats

Echantillonnage

```
> source("reechantillon.r")
```

→ Fichier texte contenant les données échantillonnées

Document public

HYPE

Outil d'analyse statistique des séries temporelles d'évolution de la qualité des eaux souterraines

Manuel d'utilisation

BRGM/RP-63099-FR

Étude réalisée dans le cadre de la convention Onema - Brgm 2013-2015

N. Croiset, B. Lopez

Vérificateur :	**Approbateur :**
Nom : Laurence Gourcy	Nom : Nathalie Dorfliger
Date : 20/12/2013	Date :
Signature **:**	Signature **:**

En l'absence de signature, notamment pour les rapports diffusés en version numérique, l'original signé est disponible aux Archives du BRGM.

Le système de management de la qualité du BRGM est certifié AFAQ ISO 9001:2008.

Géosciences pour une Terre durable
brgm

Mots-clés : OUTIL INFORMATIQUE, LANGAGE R, ANALYSE STATISTIQUE, TENDANCE, RUPTURE, INVERSION, SERIES TEMPORELLES, QUALITE DES EAUX, EAUX SOUTERRAINES

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

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.

Sommaire

Liste des illustrations

1. Contexte du développement de l'outil HYPE

Afin de répondre aux exigences de la Directive 2000/60/CE et de la Directive fille eaux souterraines 2006/118/CE qui demandent aux Etats Membre d'identifier, selon une approche statistique, les tendances d'évolution des concentrations en contaminants dans les eaux souterraines les agences et les offices de l'eau ont exprimé, dès 2009, un besoin d'accompagnement scientifique sur ces thématiques spécifiques.

Plusieurs actions ont ainsi été menées depuis 2010 dans le domaine de l'identification des tendances d'évolution de la qualité des eaux souterraines dont, à l'échelle nationale :

- Revue bibliographique et tests des méthodes statistiques existantes pour l'identification des tendances d'évolution des contaminants dans les eaux souterraines réalisés sous convention ONEMA-BRGM en 2010 (Lopez et Leynet, 2011).

- Projet national tendance (convention ONEMA-BRGM 2012) mené en partie dans le cadre de la révision de l'état des lieux 2013 avec l'élaboration de fiches « tendance » par masse d'eau souterraine (Lopez et al., 2013).

Ces actions nationales, complétées d'autres études menées aux échelles des grands bassins (Baran et al., 2009 ; Lopez et Baran, 2011 ; Lopez et al., 2012), ont permis d'acquérir l'expérience théorique et pratique nécessaire à l'élaboration d'une méthode robuste pour l'identification des tendances et des ruptures d'évolution temporelle de la qualité des eaux souterraines.

Après avoir pris connaissance de la méthode proposée, les agences, les offices de l'eau et les DEAL ont exprimé le besoin d'avoir un outil leur permettant de mettre en œuvre facilement ces approches statistiques. Cet outil dédié à l'analyse statistique des séries temporelles de la qualité des eaux souterraines répond à deux impératifs supplémentaires :

- fonctionner à partir du format SANDRE des données extraites d'ADES afin de prendre en compte les analyses réalisées au sein des réseaux RCS et RCO dont une des finalités est « d'identifier les tendances à la hausse significative et durable des polluants dans les eaux souterraines » (2006/118/CE) ;

- être compatible avec le format de l'outil SEEE (système d'évaluation de l'état des eaux) développé dans un langage informatique proche de R.

Un outil d'analyse statistique des séries temporelles d'évolution de la qualité des eaux souterraines, appelé HYPE, a ainsi été développé sous environnement R dans le cadre des conventions ONEMA-BRGM 2012 et 2013. Il permet à la fois de caractériser les séries temporelles d'évolution des contaminants dans les eaux souterraines en calculant les statistiques de base de manière automatique et d'identifier des tendances et des ruptures des séries chronologiques.

L'ensemble des méthodes compilées dans l'outil et des tests effectués sur des chroniques réelles tirées de la base de données ADES sont présentés dans le rapport relatif à l'étude (Lopez el al., 2013). Le présent document correspond au manuel d'utilisation détaillé de l'outil HYPE. Une plaquette synthétique de prise en main rapide de l'outil est fournie en guise de synthèse opérationnelle du manuel d'utilisation (partie 5).

2. Descritption des tests statistiques implémentés dans l'outil

2.1. CALCUL DES STATISTIQUES DE BASES DE LA CHRONIQUE

Les statistiques de base sont calculées dans le module « Caractérisation » et en préalable à plusieurs traitement statistiques dans les autres modules de l'outil HYPE.

2.1.1. Moyenne

La moyenne des données de concentrations calculée par l'outil HYPE est une moyenne arithmétique :

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

où n est le nombre d'observations, x_i le vecteur des données et $\bar{x}$ leur moyenne.

Lorsque le résultat de l'analyse est inférieur à la limite de détection ou de quantification, la valeur de concentration prise en compte est la moitié de la limite de quantification ou de détection indiquée, ceci en accord avec les recommandations de la Directive 2000/60/CE reprises dans la Directive fille sur les eaux souterraines (2006/118/CE).

2.1.2. Ecart-type

L'écart-type calculé est **l'écart-type d'un échantillon, non biaisé**, défini par la formule suivante :

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où n est le nombre d'observations, x_i le vecteur des données et $\bar{x}$ leur moyenne.

2.1.3. Médiane/Quartiles

Lorsque le nombre de données dans la série est impair, la valeur de la médiane correspond à la valeur centrale de la série dont les données sont classées par rang. Lorsque le nombre de données est pair, la valeur de la médiane est la moyenne arithmétique des deux valeurs centrales.

Les quartiles sont les valeurs qui divisent le jeu de données en 4 parts égales. Le premier quartile est la valeur qui sépare le jeu de données entre les 25% les plus bas et le reste des analyses. Le troisième quartile sépare les 25% des analyses les plus élevées du reste. La procédure de calcul est détaillée sur l'Illustration 1.

Calcul des quartiles

De nombreuses définitions existent pour le calcul des quartiles. Hyndman and Fan (1996) listent 9 méthodes différentes utilisées dans les logiciels de calcul statistique. Nous avons choisi d'utiliser la définition notée « définition 7 » dans l'article de Hyndman et Fan qui est communément utilisée. La définition est donnée ci-dessous

Soit n est le nombre de valeurs dans la chronique et $x_1, x_2, \dots, x_n$ les valeurs ordonnées de la plus petite à la plus grande.

Le $i$ème quartile est défini par interpolation entre x_j et x_{j+1}.où j est la partie entière de $i/4 * (n-1) + 1$ et g est la partie fractionnaire de ce nombre. On a donc $g = i/4 * (n-1) + 1 - j$

Le $i$ème quartile est ainsi défini par :

$$Q_i = x_j + g(x_{j+1} - x_j)$$

Illustration 1 : Calcul des quartiles dans HYPE

Dans les colonnes « remarque sur la médiane/quartiles » un texte est ajouté si la médiane ou l'un des quartiles est supérieur à la limite de quantification maximale.

2.1.4. Fréquence de quantification

Le résultat d'une analyse est considéré comme une quantification dès lors que la valeur de concentration est reportée supérieure à la limite analytique de quantification. Ceci correspond à la définition du code remarque 1 associé à l'analyse dans le référentiel SANDRE. Ainsi, pour le traitement de données directement extraites de la base ADES, toutes les analyses accompagnées d'un code remarque égal à 1 sont considérées comme quantifiées. Aucune vérification n'est effectuée sur la valeur du résultat de l'analyse.

La fréquence de quantification est égale au rapport du nombre d'analyses supérieures à la limite de quantification (c'est-à-dire ayant un code remarque égal à 1 pour les données extraites de la base ADES directement) par le nombre total d'analyses.

2.1.5. Limites de quantification

Deux couples de limites de quantification minimum et maximum sont indiquées. Le premier correspond aux analyses indiquant une concentration inférieure à la limite de quantification (code remarque égal à 10 selon le référentiel SANDRE). Le second correspond à toutes analyses qui ne sont pas des quantifications (code remarque différent de 1). Cette dernière limite peut alors correspondre à une limite de détection.

2.1.6. Fréquence d'échantillonnage

Pour caractériser la fréquence d'échantillonnage, deux valeurs caractéristiques sont calculées : la moyenne et l'écart-type des écarts entre deux analyses.

L'écart-type calculé est l'écart-type non biaisé (pour la définition, voir § 2.1.2).

2.1.7. Normalité de la distribution des données

La normalité de la distribution des données est testée, notamment afin de savoir s'il faut appliquer des tests paramétriques ou non paramétriques à ces données.

Elle est testée pour les chroniques disposant d'au moins 3 analyses en appliquant le test de Shapiro-Wilk (Illustration 2).

Si la p-value du test de Shapiro est inférieure à 0,05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle les données sont normalement distribuées.

Le test de normalité de Shapiro-Wilk (Shapiro & Wilk, 1965)

Le test de Shapiro-Wilk teste l'hypothèse nulle selon laquelle un échantillon provient d'une population normalement distribuée.

La statistique calculée est :

$$W = \frac{(\sum_{i=1}^{n} a_i x_i')^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où $\bar{x}$ est la moyenne de l'échantillon, x' est le vecteur contenant les données triées (x_i' est donc la $i$ème valeur la plus petite), n est le nombre d'observations0, et :

$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$, où $m = (m_1, \ldots, m_n)^T$ et les m_i sont les valeurs attendues d'une distribution normale pour un échantillon de taille n et V est la matrice de variance-covariance correspondante.

La p-value du test est une valeur exacte pour $n = 3$, sinon des approximations différentes sont utilisées pour $4 \leq n \leq 11$ d'une part et $n \geq 12$ d'autre part. L'algorithme utilisé pour le calcul de la p-value est celui proposé par Royston (1995).

Illustration 2: Test de normalité de la distribution des données dans HYPE

Pour les chroniques qui ne suivent pas une distribution normale, il faudra utiliser de préférence des tests non paramétriques pour la recherche de tendances et de ruptures.

2.1.8. Autocorrélation des données

Estimer l'autocorrélation des données revient à se demander si une valeur observée à un temps t dépend de ce qui a été observé dans le passé.

Il est important d'estimer cette autocorrélation des données car elle peut biaiser le niveau de significativité des tests statistiques (Renard, 2006). Dans des conditions environnementales, l'autocorrélation totale est toujours positive, ce qui tend à diminuer la significativité des tests statistiques.

L'intervalle de confiance à 95% est calculé pour être comparé à l'autocorrélation au rang 1.

Le test d'autocorrélation

L'autocorrélation au rang k se calcule par la formule suivante :

$r_k = C_k/C_0$, 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 suivants.

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

$r_{lim} = \frac{1}{\sqrt{n}} qnorm\left(\frac{1+\propto}{2}\right)$, où $qnorm$ est la fonction quantile d'une loi normale centrée réduite. $\propto$ est le seuil de significativité. Nous l'avons choisi égal à 0,95.

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

Illustration 3 : Test d'autocorrélation

2.2. TESTS DE TENDANCE

2.2.1. Test de tendance non paramétrique (Mann-Kendall)

Le test de Mann-Kendall (décrit dans l'Illustration 4) est associé au calcul de la pente de Sen (décrit sur l'Illustration 5). Il est appliqué sur les toutes les chroniques non stationnaires disposant d'au moins 10 analyses.

La tendance est dite significative d'un point de vue statistique lorsque la p-value du test est inférieure à 5%.

Le test de Mann-Kendall (Kendall, 1938, repris par Renard, 2006)

L'hypothèse H0 testée est l'absence de tendance.

La statistique calculée est définie comme suit :

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn\left[(y_j - y_i)(x_j - x_i)\right]$$

où la fonction sgn est la définie par : $sgn(X) = 1$ pour $X > 0$; $sgn(X) = 0$ pour $X = 0$ et $sgn(X) = -1$ pour $X < 0$.

Mann (1945) et Kendall (1975) ont démontré que

$$E(S) = 0$$

$$Var(S) = n(n-1)(2n+5)/18$$

Dès que l'échantillon contient une dizaine de données, la loi de la statistique de test Z ci-dessous peut-être approché par une gaussienne centrée-réduite.

$$Z = \frac{S-1}{(Var(S))^{1/2}} \text{ si } S > 0$$

$$Z = 0 \text{ si } S = 0$$

$$Z = \frac{S+1}{(Var(S))^{1/2}} \text{ si } S < 0$$

S'il y a des ex-aequo dans la série, la variance de S est corrigée de la façon suivante :

$$Var(S) = 1/18\left[n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(p-1)(2p+5)\right]$$

où t_p est le nombre d'égalités impliquant p valeurs.

Illustration 4 : Test de tendance de Mann-Kendall (Kendall, 1938)

Calcul de la pente de Sen et de l'ordonnée à l'origine

La pente de la droite de régression (appelée pente de Kendall-Theil ou pente de Sen) est estimée par la méthode de Sen (Sen, 1968).

La pente est la médiane de toutes les pentes calculées entre chaque paire de point. L'estimation de l'ordonnée à l'origine peut être calculée de différentes manières. La méthode retenue est celle recommandée par Helsel et Hirsch (2002) utilisant la pente médiane et la médiane des variables (Conover, 1980).

$$pente_{Sen} = mediane_{i<j}\left\{\frac{y_j - y_i}{x_j - x_i}\right\}$$

$$ordonnée_{orig} = mediane(y) - pente_{Sen} * mediane(x)$$

Illustration 5 : Calcul de la pente de Sen et de l'ordonnée à l'origine pour le test de Mann-Kendall

2.2.2. Test de tendance paramétrique : la régression linéaire

La régression linéaire (détaillée sur l'Illustration 6) est calculée pour les chroniques disposant d'au moins 5 analyses et ayant des données normalement distribuées. Les hypothèses sous-jacentes à l'application d'une régression linéaire sont : la normalité de la distribution des données, l'homogénéité de la variance et une relation linéaire entre la variable expliquée et la variable explicative

Le r^2 est calculé ainsi que la pente de régression.

La p-value de la régression est donnée. La tendance est dite significative d'un point de vue statistique lorsque la p-value du test <0,05 (5%).

La régression linéaire (selon Renard, 2006)

L'hypothèse H0 testée est que les données ne sont pas linéairement dépendantes du temps.

Ce test est basé sur le modèle paramétrique suivant :

$$X = \propto + \beta t + \varepsilon$$

où les erreurs ε suivent une loi normale centrée.

La valeur de la pente et l'ordonnée à l'origine sont définies de façon à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs de la droite de régression.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x}$$

La variance de l'estimateur de tendance peut être estimée par :

$$Var(\hat{\beta}) = \frac{12\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\alpha} - \hat{\beta}x_i)}{(n-2)n(n^2+1)}$$

Le test de la régression linéaire consiste alors à vérifier que l'estimateur du coefficient β est proche de 0. Pour cela on compare la statistique Z définie ci-dessous aux quantiles d'une loi de Student à n-2 degré de liberté :

$$Z = \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}$$

Illustration 6 : Régression linéaire

2.2.3. Test de Mann-Kendall modifié (Hamed, 1998)

Ce test ne peut être appliqué que si la chronique dispose 40 analyses ou plus. Ce test permet de prendre en compte l'autocorrélation des données dans la série chronologique.

Le principe repose sur une modification du test S de Mann-Kendal plutôt que de modifier les données elles-mêmes :

$$Var_\rho(S) = \gamma Var_{\rho=0}(S)$$

où γ est un facteur correctif appliqué à la variance.

Ainsi, seule la p-value du test de Mann-Kendall est modifiée ; la pente est la même que celle calculée pour le test de Mann-Kendall non modifié.

Test de Mann-Kendall modifié (Hamed, 1998)

La modification du test correspond au fait qu'un échantillon autocorrélé positivement de taille n se comporte comme un échantillon indépendant de taille $n^* < n$ (et inversement pour un échantillon autocorrélé négativement).

Plusieurs méthodes de calcul de γ sont relevées dans la littérature. Nous retiendrons la méthode proposée par Hamed et Rao (1998), légèrement plus puissante que la formule proposée par Yue et Wang (2004), d'après les tests effectués par Renard (2006). La formule proposée par Hamed et Rao se base sur une formule empirique spécifiquement calculée pour corriger la statistique de Mann-Kendall. Elle prend en compte les autocorrélations des résidus de régression calculées aux différents rangs si celles-ci sont significatives :

$$\gamma = 1 + \frac{2}{n(n-1)(n-2)} \sum_{k=1}^{n-1} (n-k)(n-k-1)(n-k-2)\rho_k$$

où *n* est le nombre de données et ρ_k est l'autocorrélation à l'ordre *k*, si elle est significative,

$\rho_k = 0$ sinon

Le seuil de significativité choisi pour l'autocorrélation est 5%.

Illustration 7 : Test de Mann-Kendall modifié pour la prise en compte de l'autocorrélation.

2.3. ANALYSE DE LA VARIABILITE ENTRE PERIODES

Deux tests sont utilisés par l'outil afin d'étudier la variabilité entre périodes : un test paramétrique, l'analyse de la variance (ANOVA), si les données sont normalement distribuées et un test non paramétrique, le test de Kruskal-Wallis, si les données ne sont pas normalement distribuées.

Ce test permet d'évaluer si une variable explicative a une influence sur les données. Deux variables explicatives qui correspondent à deux périodes sont testées dans l'outil : le trimestre (saison) et le mois.

2.3.1. ANOVA

Ce test est appliqué pour les chroniques présentant au moins 10 valeurs et ayant une distribution normale.

Si la p-value du test est inférieure à 0,05, on peut considérer qu'au moins deux périodes sont significativement différentes l'une de l'autre.

Analyse de la variance (ANOVA) à un facteur

L'hypothèse nulle de ce test est que les moyennes des différentes périodes sont identiques.

La statistique calculée est la statistique de Fisher, définie comme suit :

$F = \frac{s_\alpha^2}{s_r^2}$, où s_α^2 et s_r^2 sont respectivement la variance due au facteur (ou variance inter-classe)

et la variance résiduelle (ou variance intra-classe).

$$s_{\alpha}^2 = \frac{\sum_{i=1}^{g} n_i (\overline{x_i} - \bar{x})^2}{k-1}$$

$$s_r^2 = \frac{\sum_{i=1}^{g} (n_i - 1)\, {\sigma_i}^2}{n-k}$$

g est le nombre de périodes disposant de n_i échantillons, $\overline{x_i}$ est la moyenne et ${\sigma_i}^2$ la variance des valeurs dans le groupe i. $\bar{x}$ est la moyenne globale de l'échantillon.

La statistique est ensuite comparée aux quantiles d'une loi de Fisher à $(g-1)$ degrés de liberté au numérateur, et $(n-g)$ degrés de liberté au dénominateur.

Illustration 8 : Analyse de la variance (ANOVA) à un facteur.

2.3.2. Test de Kruskal-Wallis

De même que le test d'analyse de la variance, le test de Kruskal-Wallis permet de déterminer si les données d'une période sont significativement différentes d'une autre. Il est appliqué pour les chroniques dont les données ont une distribution non normale et disposant d'au moins 10 valeurs.

Si la p-value du test est inférieure à 0,05, il y a une différence entre les données d'au moins deux périodes.

Test de Kruskal-Wallis

L'hypothèse nulle est que les périodes ne sont pas différentes les unes des autres.

La statistique calculée est définie comme suit :

$$K = (n-1)\frac{\sum_{i=1}^{g} n_i(\bar{r_i} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

où g est le nombre de période, n_i est le nombre d'observation dans la période i, r_{ij} est le rang de l'observation j dans le groupe i. n est le nombre total de données, $\bar{r_i} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ et $\bar{r} = \frac{1}{2(n+1)}$

La statistique est ensuite comparée aux quantiles d'une loi du chi 2 à (g-1) degrés de liberté.

Illustration 9 : Test de Kruskal-Wallis.

2.4. TESTS DE RUPTURE

Deux types de rupture sont recherchés : une rupture dans la moyenne et une inversion dans la tendance.

2.4.1. Changement de moyenne

Pour rechercher un changement de moyenne dans la chronique, deux tests d'homogénéité peuvent être appliqués. Si les données sont normalement distribuées, le test appliqué est le test paramétrique de Buishand. Dans le cas contraire, on applique le test non paramétrique de Pettitt.

Si une rupture significative est détectée, les moyennes arithmétiques sur les tronçons pré- et post-rupture sont calculées.

a) Test non paramétrique de Pettitt

Le test de Pettitt est un test non paramétrique qui dérive du test de Mann-Whitney. Ce test est appliqué sur les chroniques non stationnaires disposant d'au moins 3 données et dont la distribution est non normale.

La rupture est dite significative d'un point de vue statistique lorsque la p-value du test est inférieure à 5%.

Le test de Pettitt (Pettitt, 1979)

Le test de Pettitt est non paramétrique. Il dérive du test de Mann-Whitney. L'hypothèse nulle est l'absence de rupture dans la chronique. Elle est testée par la statistique $U_{t,n}$ considérée pour l'ensemble des valeurs de t telles que $1 \leq t \leq n$:

$U_{t,n} = \sum_{i=1}^{t} \sum_{j=t+1}^{n} D_{ij}$ où : $D_{ij} = sgn(X_i - X_j)$ où X_i est le vecteur des données trié par date et la fonction sgn est définie par :

$sgn(X) = 1$ pour $X > 0$; $sgn(X) = 0$ pour $X = 0$ et $sgn(X) = -1$ pour $X < 0$

On utilise alors la variable K_n pour tester H_0 telle que $K_N = max|U_{t,n}|$.

Si k correspond à la valeur de K_n, la probabilité de dépassement de la valeur k est donnée par :

$$\Pr(K_n > k) \sim 2 \exp\left[-\frac{6k^2}{(n^3 + n^2)}\right]$$

Si $\propto$ est supérieur à cette probabilité, H_0 est rejetée. La série présente alors une rupture au temps t définissant K_n.

Illustration 10 : Test d'homogénéité de Pettitt

b) Test paramétrique de Buishand

Le test de Buishand est un test paramétrique. Il est appliqué pour les chroniques disposant d'au moins 10 valeurs et dont la distribution est normale. Ce test suppose un non changement de la variance de la série.

Le test de Buishand (Buishand, 1982, 1984)

L'hypothèse H_0 est l'absence de rupture dans la chronique.

Ce test est construit à partir des écarts cumulés à la moyenne jusqu'à un rang k :

$$S_k = \sum_{i=1}^{t} (x_i - \bar{x})$$

La statistique de test est obtenue par la division des valeurs S_k par la déviation standard :

$$Z = \max\left(\frac{|S_k|}{\sqrt{n}\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}} \right)$$

Une valeur de Z élevée est un signe d'une rupture dans la chronique. La significativité du test est calculé en comparant la valeur de Z à des valeurs critiques.

Les valeurs critiques prises en compte sont celles évaluées par Buishand (1982) par la génération de séquences aléatoires. Ces valeurs sont données dans l'Illustration 11.

N	Z au niveau de confiance		
	α = 0,10	α = 0,05	α = 0,01
10	1,05	1,14	1,29
20	1,10	1,22	1,42
30	1,12	1,24	1,46
40	1,13	1,26	1,50
50	1,14	1,27	1,52
100	1,17	1,29	1,55
∞	1,22	1,36	1,63

Illustration 11: Valeurs critiques de Z du test de Buishand

Illustration 12 : Test paramétrique d'homogénéité (Buishand, 1982)

2.4.2. Inversion de la tendance

Darken propose dans sa thèse (1999) une méthode pour détecter un changement de signe de la pente (et non un changement dans la magnitude de la tendance sans changement de signe) qui est valable pour des données non normalement distribuées. Le test est décrit en Illustration 13.

Si la p-value du test est inférieure à 5%, la chronique présente une inversion de tendance significative. La date de rupture est déterminée ainsi que les tendances avant et après cette date de rupture.

Changement de pente de Darken (1999)

Darken (1999) propose dans sa thèse deux méthodes basées sur le tau de Kendall.

Pour un changement de tendance (= changement de signe), il propose la statistique suivante :

$$Z = \frac{\tau_1 - \tau_2}{\sqrt{Var(\tau_1) + Var(\tau_2)}}$$

Les variances sont calculées comme décrit par Kendall (1976).

La date de rupture la plus probable est la date pour laquelle Z est maximum.

La p-value du test est calculée en comparant la statistique Z pour la date de rupture identifiée aux quantiles d'une loi normale centrée réduite (Darken, 1999).

Illustration 13 : Test d'inversion de tendance (Darken,1999)

2.5. TEST DE KENDALL SAISONNIER

Ce module permet d'effectuer un test de Kendall saisonnier. Ce test, proposé par Hirsch et al. (1982) permet d'estimer des tendances de séries cycliques saisonnières.

Le principe est identique à celui du test de Mann-Kendall mais le caractère saisonnier de la série est pris en compte. Autrement dit, pour des données mensuelles ayant une saisonnalité de 12 mois, on ne va pas chercher à savoir s'il y une croissance au global sur la série, mais simplement si, d'un mois de janvier à l'autre, d'un mois de février à l'autre et ainsi de suite, il y a une tendance.

L'outil peut être utilisé sur deux périodes différentes : le trimestre (de janvier à mars, avril à juin, juillet à septembre et octobre à décembre) et le mois calendaire.

La statistique S_k de Kendall se calcule à partir de la somme des statistiques pour chaque saison (Hirsch, 1982).

$S_k = \sum_{i=1}^{s} S_i$ où s est le nombre de saisons et S_i sont les statistiques S de Mann-Kendall

$$S_i = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn\ [(y_{j_i} - y_{k_i})(x_{j_i} - x_{k_i})]$$

Et $\sigma_{S_k} = \sqrt{\sum_{i=1}^{S} n_i(n_i - 1)(2n_i + 5)/18}$ où n_i est le nombre de données pour la saison i.

La statistique calculée est :

$$Z_{S_k} = \frac{S_k}{\sigma_{S_k}}$$

Si le produit du nombre de saisons par le nombre d'années est supérieur à 25, la distribution des S_k peut être approximé par une distribution normale avec et la variance égale à la somme des variances. S_k est standardisé par soustraction de sa moyenne (égale à 0) et division par son écart-type. Le résultat est évalué en comparant à une table de distribution normale standard.

L'hypothèse nulle est à rejeter à un niveau de significativité α si $|Z_{Sk}| > Z_{crit}$ où Z_{crit} est la valeur de la distribution normale avec une probabilité de dépassement de $\propto/2$.

Illustration 14 : Test de Kendall saisonnier

Le calcul de la pente de Sen et l'ordonnée à l'origine sont légèrement modifiés par rapport au test de Mann-Kendall (cf Illustration 5) : la pente est la médiane des pentes calculées entre deux analyses **au sein d'une même période**, les pentes calculées entre des analyses contenues dans des périodes différentes ne sont pas prises en compte.

2.6. TEST DE KENDALL REGIONAL

Ce test permet d'étudier la présence de tendance à l'échelle d'une région d'étude, appelée ici unité spatiale, comprenant plusieurs sites.

Pour utiliser ce module, il faut que le tableau des données contienne une colonne «UNITE_SPATIALE » précisant à quelle unité spatiale est rattaché le point.

Le principe du test est de déterminer si une tendance cohérente peut être mise en évidence à partir des différentes chroniques. Il suit exactement le même principe que le test de Kendall saisonnier. On pourra donc se reporter au paragraphe précédent pour la mise en application de ce test ; la « saison » est à remplacer par l'« unité spatiale ».

3. Avant d'utiliser l'outil…

3.1. INSTALLATION DU LOGICIEL R

Le logiciel R a été créé en 1993 par Robert Gentleman et Ross Ihaka. Ce logiciel est libre et gratuit. Il peut fonctionner sur différents systèmes d'exploitation (Linux, Windows, MacOS).

R dispose d'une version de base comprenant la plupart des fonctionnalités utiles pour la statistique de base et de nombreux « packages » (ou « extensions »), mis librement à disposition. Quelques-uns de ces « packages » sont nécessaires au fonctionnement de l'outil HYPE.

Une version postérieure à la version 2.1.0 est nécessaire au fonctionnement de l'outil.

Il est donc nécessaire de télécharger d'abord la version de base de R comme présenté ci-dessous, puis les packages nécessaires (voir paragraphe 3.3.1).

Pour installer le logiciel, il faut cliquer sur le lien suivant : http://cran.r-project.org

Il faut ensuite choisir votre système d'exploitation (Linux, Windows ou MacOS), puis cliquer sur « base » puis télécharger la dernière version.

D'abord, enregistrez le fichier nommé « R-2.XX.X-win32.exe » sur votre disque de façon à le retrouver (par défaut, le fichier s'enregistre le plus souvent dans le dossier « téléchargement » du répertoire « mes documents »).

Ensuite, double cliquez sur ce fichier et suivez les consignes d'installation en laissant les valeurs par défaut. Une fois l'installation terminée, vous aurez la possibilité de lancer le programme depuis le menu « démarrer » de Windows ou en cliquant sur l'icône  sur le bureau.

3.2. LE MINIMUM VITAL A SAVOIR SOUS R

L'Illustration 15 présente un aperçu de la console R, visible au démarrage du logiciel.

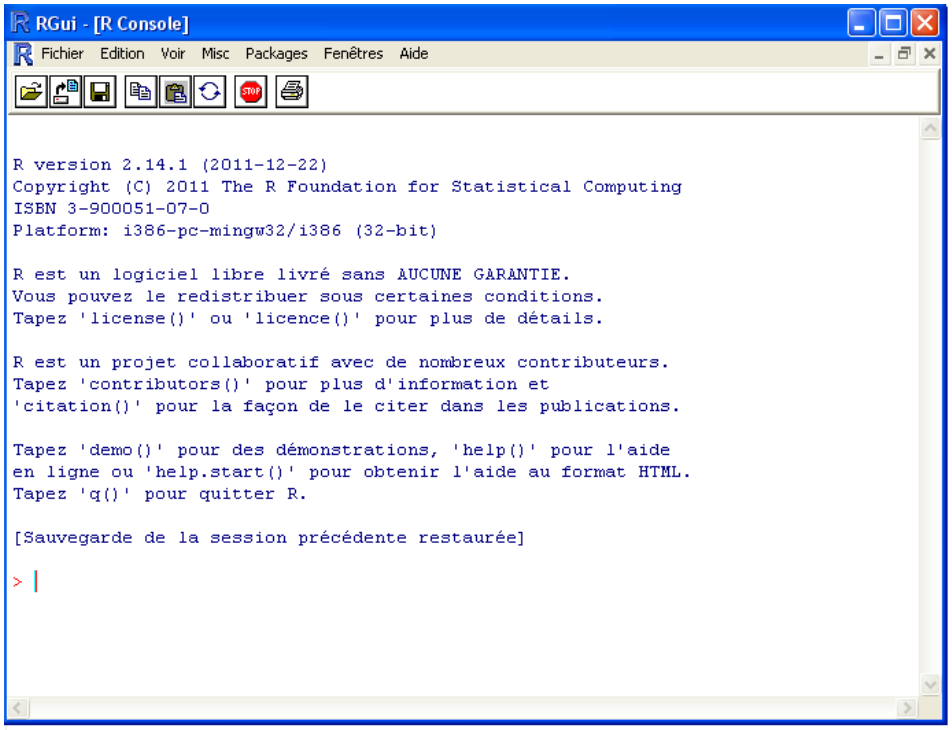

Illustration 15: Console de R au démarrage du programme

`>` invite l'utilisateur à entrer une commande.

`+` indique que la commande précédente n'est pas terminée. On peut stopper un processus avec la combinaison Ctrl+C.

`q()` est la commande pour quitter R. On peut ensuite sauver ou non son travail en tapant y/n/c (oui, non, annuler)

On peut naviguer dans les anciennes commandes avec les flèches ↑ et ↓ du clavier.

Les commandes tapées par l'utilisateur s'affichent en rouge sur l'interface graphique.

La définition du répertoire de travail (dans lequel se trouvent les scripts à lancer) s'effectue par la commande `setwd("chemin")`

Pour lancer un programme sous R, on utilise la commande `source("nom du fichier")`

3.3. INSTALLATION DES PACKAGES NECESSAIRES AU FONCTIONNEMENT DE L'OUTIL

R est composé d'un socle commun et d'une bibliothèque de fonctions implémentées par les utilisateurs, appelées *packages*, mises à disposition de tous. Plusieurs de ces *packages* sont nécessaires au fonctionnement de l'outil

Il faut les **installer** (c'est-à-dire les télécharger sur Internet ou à partir d'un fichier zip) puis les **charger** (c'est-à-dire les rendre accessibles).

3.3.1. Installation des *packages*

La première chose à faire est de télécharger les *packages* qui seront nécessaires à l'utilisation de l'outil. L'installation est pérenne, il n'y a pas besoin de réinstaller les *packages* à chaque démarrage de session.

Les *packages* à installer sont les 4 *packages* suivants : `fume`, `plotrix`, `Kendall` et `chron`.

a) *Depuis Internet*

Pour télécharger les *packages* depuis Internet, il y a deux solutions :

- Par la commande `install.packages(nomdupackage)`
- En cliquant sur « Installer le package » dans le menu « Package »

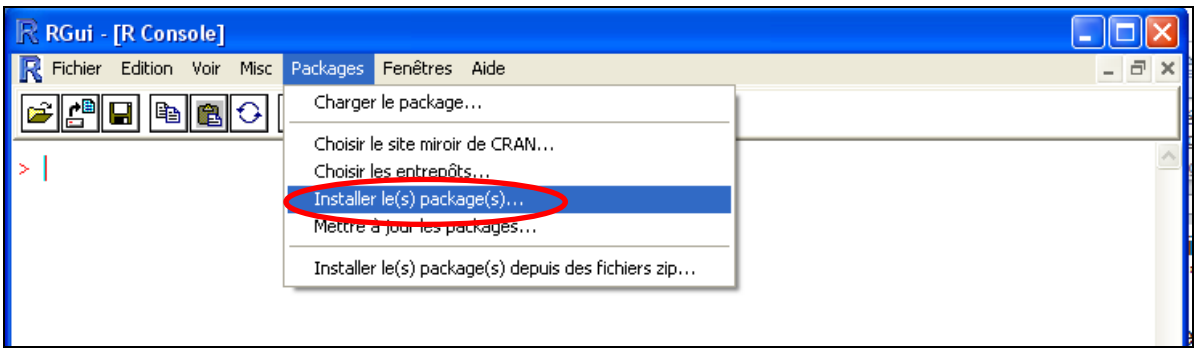

Illustration 16 : Installation d'un package depuis Internet – Aperçu d'écran

Il faut alors sélectionner un site miroir sur lequel le package sera téléchargé (voir Illustration 17). Choisissez le site le plus près de chez vous.

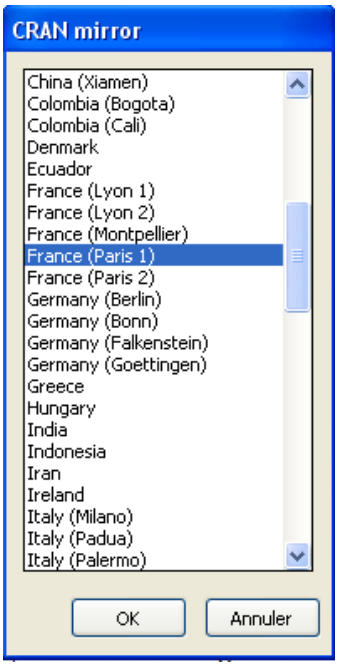

Illustration 17 : Copie d'écran du logiciel R lors du choix du site miroir depuis lequel télécharger un package

Choisissez ensuite, dans la liste qui s'affiche, le package que vous voulez installer.

Remarque : Lorsque vous installez les packages depuis Internet, le package `chron` s'installe automatiquement lors de l'installation du package `fume`, il vous suffit donc d'installer les 3 packages suivants : `fume`, `plotrix` et `Kendall`.

b) *Depuis un fichier .zip*

- Cliquez sur la fonction "Installer(s) le(s) package(s) depuis un fichier zip..." dans le dans le menu « Package »

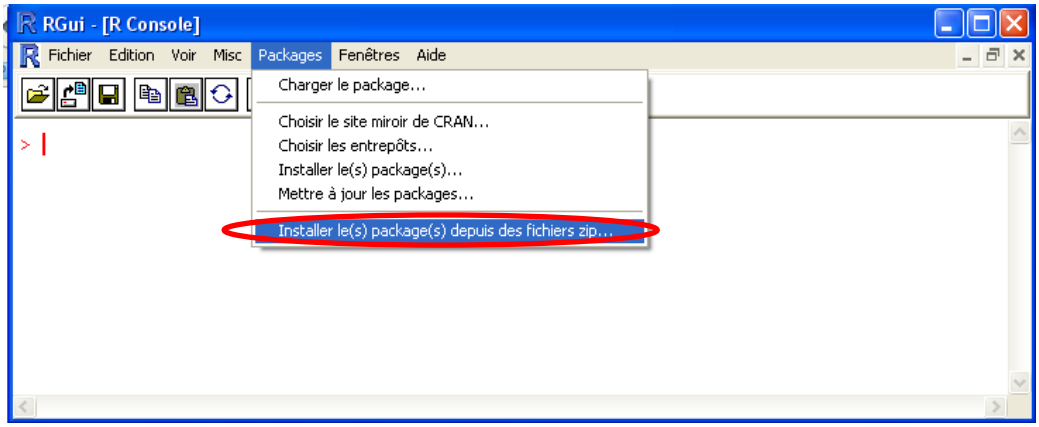

Illustration 18 : Installation d'un package depuis un fichier zip - Aperçu d'écran

Choisissez ensuite les packages à installer dans l'arborescence qui vous est proposée.

3.3.2. Chargement des *packages*

Une fois les *packages* installés, il faut les charger. Contrairement à l'installation, le chargement doit normalement se faire à chaque démarrage de session.

Chargement à chaque démarrage de R

Là encore deux possibilités :

- Soit utiliser la commande `library("nom du package")`
- En cliquant sur « Charger le package » dans le Menu « Package » comme présenté sur l'illustration ci-dessous.

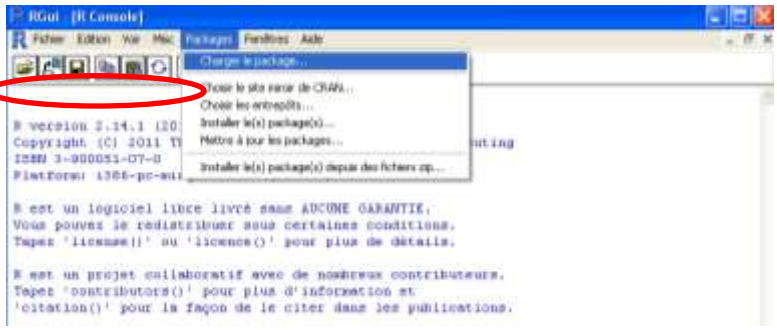

Illustration 19 : Chargement d'un package - Aperçu d'écran.

Chargement automatique à chaque démarrage de R

Pour éviter de charger les mêmes *packages* à chaque session, il est possible de les indiquer dans un fichier de configuration, *Rprofile.site*, lu au démarrage de R.

Dans les environnements Windows, le fichier *Rprofile.site* est situé par défaut dans le répertoire « C:\Program Files\R\R-2.14.1\etc ».

Pour que les *packages* nécessaire au fonctionnement de l'outil soit chargés à chaque utilisation, il faut ouvrir le fichier dans un éditeur de texte et ajouter les trois lignes suivantes :

```
library("fume")
library("Kendall")
library("plotrix")
```

Le fichier ainsi modifié doit être enregistré.

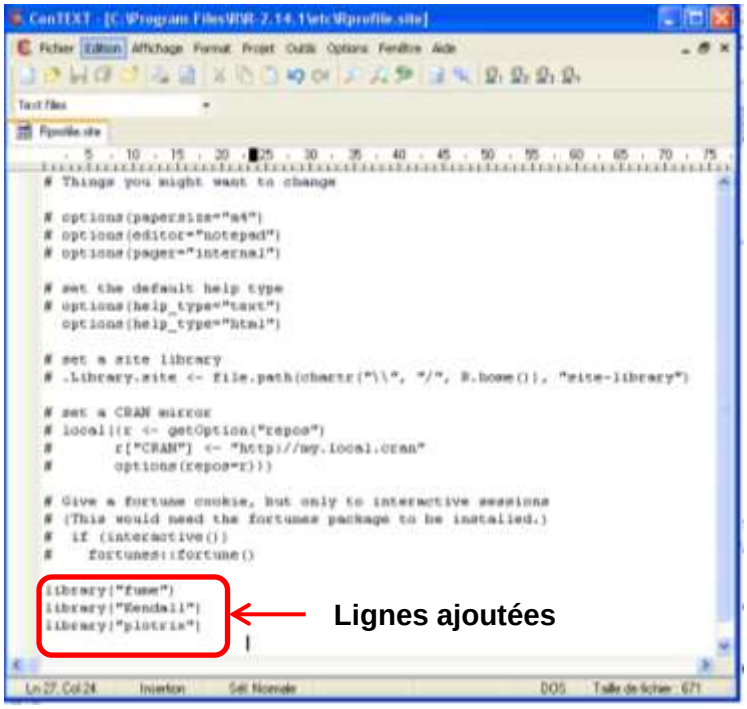

Illustration 20 : Aperçu du fichier Rpofile.site après modification pour que les packages nécessaires soient chargés à chauqe démarrage de R

4. Les différents modules de HYPE

4.1. FORMAT DES DONNEES D'ENTREE

Les données à traiter doivent être sous forme d'un **fichier texte**. Plusieurs mises en forme sont acceptées afin qu'il soit possible de traiter des données provenant d'un export de la base de données ADES réalisé depuis le site public ou producteur (sous réserve que les formats d'export ne soient pas modifiés) et des données provenant d'un fichier Excel personnel.

Les données traitées peuvent contenir des analyses sur plusieurs points de prélèvement et/ou de plusieurs paramètres. Chaque chronique, identifiée par un point BSS et un nom de paramètre sera traitée séparément.

Le caractère séparateur de colonne peut être :

- soit le caractère « | », comme c'est le cas dans les fichiers provenant d'exports ADES ;

- soit une tabulation, ce que l'on obtient facilement lorsque l'on enregistre un fichier d'un tableur (Excel par exemple) en un fichier au format texte.

Le fichier peut comporter un nombre illimité de colonnes mais certaines sont obligatoires.

Le séparateur décimal doit obligatoirement être un point.

Les titres des colonnes doivent se situer sur la première ligne et doivent correspondre **exactement** aux titres décrits ci-dessous.

Les colonnes obligatoires, récapitulés sur Illustration 21, sont :

- dans le cas où votre fichier comprend plusieurs chroniques, un identifiant de la chronique défini par un identifiant du point de prélèvement et/ou un paramètre :

    - l'identifiant du point de prélèvement est appelé **CODE_BSS** ou **Code national BSS**. Cet identifiant peut être un nombre ou une chaîne de caractère. Il apparaîtra dans le titre des graphiques

    - la substance analysée est appelée **Paramètre** ou **LIBELLE_PARAMETRE**. Ce paramètre peut être un nombre ou une chaîne de caractère. Il apparaîtra dans le titre des graphiques.
      S'il n'y a aucune de ces deux colonnes dans le fichier d'entrée, l'ensemble des analyses est considéré former une seule et même chronique.

- la date de prélèvement, appelée **DATE_DEBUT_PRELEVEMENT** ou **Date prélèvement**. Les dates doivent **impérativement** être **sous la forme *jj/mm/aaaa*** avec possibilité de rajouter l'heure (*jj/mm/aaaa hh :mm*)

- un code correspondant à la qualification du résultat comme renseigné dans ADES, appelé **CODE_SIGNE** ou **Code remarque analyse**. Ce paramètre permet notamment le calcul des taux de quantification. Le tableau ci-dessous décrit les valeurs prises par ce paramètre selon le résultat de l'analyse.

CODE_SIGNE	
1	Résultat supérieur au seuil de quantification
2	Résultat inférieur au seuil de détection
10	Résultat inférieur au seuil de quantification
7	Traces (<seuil de quantification et > seuil de détection

- la valeur du résultat, appelé **RESULTAT** ou **Résultat de l'analyse** ou **Résultat analyse.** Cette colonne doit obligatoirement contenir des nombres et comme précisé plus haut, le séparateur décimal doit obligatoirement être un point. Dans le cas où le résultat n'est pas disponible (analyse non faite par exemple), la cellule correspondante doit être vide.

- une colonne précisant l'unité du résultat. Pour cette colonne, il y a deux possibilités :

    - renseigner directement l'unité par son abréviation qui apparaîtra dans les légendes des graphiques. Le titre de la colonne doit être alors **UNITE_GRAPH** ou **Unité du graphique**

    - renseigner l'unité par son libellé complet – ce qui est le cas dans un export ADES – dans une colonne appelée **UNITE** ou **Unité**. Le programme fera alors appel au fichier texte « Unite_SANDRE.txt », livré avec l'outil pour convertir cette unité en son abréviation. **Ce fichier texte doit impérativement se trouver dans le répertoire de travail défini en début de session**.

- de manière optionnelle, dans le cas où l'on veut effectuer un test de Kendall régional, il est possible de renseigner une colonne précisant à quelle unité spatiale appartient le point. Le titre de la colonne doit être **UNITE_SPATIALE** ou **Unité spatiale.** En l'absence d'une telle colonne, le test régional est effectué sur la totalité des chroniques du fichier d'entrée.

- de manière optionnelle, une colonne précisant le code SANDRE du paramètre étudié, appelé **CODE_PARAMETRE**. Cette colonne sera uniquement utilisée pour l'élaboration de la légende des graphiques : si le code paramètre correspond à un élément chimique, l'axe des ordonnées des graphiques sera appelé « Concentration », si le code correspond à la température, le pH, la conductivité, l'oxygène dissous ou le potentiel d'oxydo-réduction, l'axe des ordonnées prendra le nom du paramètre. Par défaut, le libellé de l'axe est « Concentration ».

Titre des colonnes du fichier à traiter Option 1	Titre des colonnes du fichier à traiter Option 2	Contenu des colonnes	Obligatoire ou facultatif	Descriptif	Format
CODE_BSS	Code national BSS	**Identifiant de la chronique**	**Obligatoire si plusieurs chroniques**, facultatif sinon (l'une ou l'autre ou les deux colonnes peuvent être renseignées)	Identifiant du point de prélèvement	Numérique ou chaîne de caractère
LIBELLE_PARAMETRE	Paramètre			Identifiant du paramètre	Numérique ou chaîne de caractère
DATE_DEBUT_PRELEVEMENT	Date prélèvement	**Date de l'analyse**	**Obligatoire**		jj/mm/aaaa au moins (l'heure peut être ajoutée)
CODE_SIGNE	Code remarque analyse	**Code de qualification du résultat**	**Obligatoire**		Numérique
RESULTAT	Résultat de l'analyse *ou* Résultat analyse	**Valeur du résultat**	**Obligatoire**		Numérique (séparateur décimal : point)
UNITE_GRAPH	Unité du graphique	**Unité**	**Obligatoire** (l'une ou l'autre ou les deux colonnes peuvent être renseignées)	Abréviation	Numérique ou chaîne de caractère
UNITE	Unité			Libellé complet	Doit correspondre exactement à la nomenclature SANDRE
UNITE_SPATIALE	Unité spatiale	**Unité spatiale**	**Obligatoire pour un test de Kendall régional si plusieurs unités spatiales**		Numérique ou chaîne de caractère
CODE_ PARAMETRE		**Code paramètre**	Facultatif		Doit correspondre exactement à la nomenclature SANDRE

Illustration 21 : Récapitulatif des colonnes obligatoires dans le fichier d'entrée.

4.2. MODULE « LECTURE DES DONNEES »

Ce module permet de charger en mémoire les données sur lesquelles vous voulez travailler. Il est obligatoire d'exécuter ce module avant tout traitement de vos chroniques.

4.2.1. Définition du répertoire de travail

Avant l'exécution d'un module, l'utilisateur doit définir le répertoire de travail. Le répertoire de travail doit contenir les différents modules (avec l'extension .r) et dans le cas où l'unité est définie par son libellé complet et non par l'abréviation renseignée sur le graphique, le fichier « Unite_SANDRE.txt ».

Il y a deux possibilités pour définir le répertoire de travail :

- A l'aide de la commande `setwd("chemin")`. Le chemin doit être renseigné entre guillemets, la séparation des répertoires s'indique par une barre oblique (slash) comme montré dans l'exemple ci-dessous. La barre oblique en fin de chemin est optionnelle.

Remarque : pour éviter les erreurs de frappe, vous pouvez ne taper que les premières lettres du nom des répertoires et utiliser la touche de tabulation, qui complète automatiquement le nom des répertoires.

- En sélectionnant « Changer le répertoire courant… » dans le Menu « Fichier » comme présenté sur l'Illustration ci-dessous.

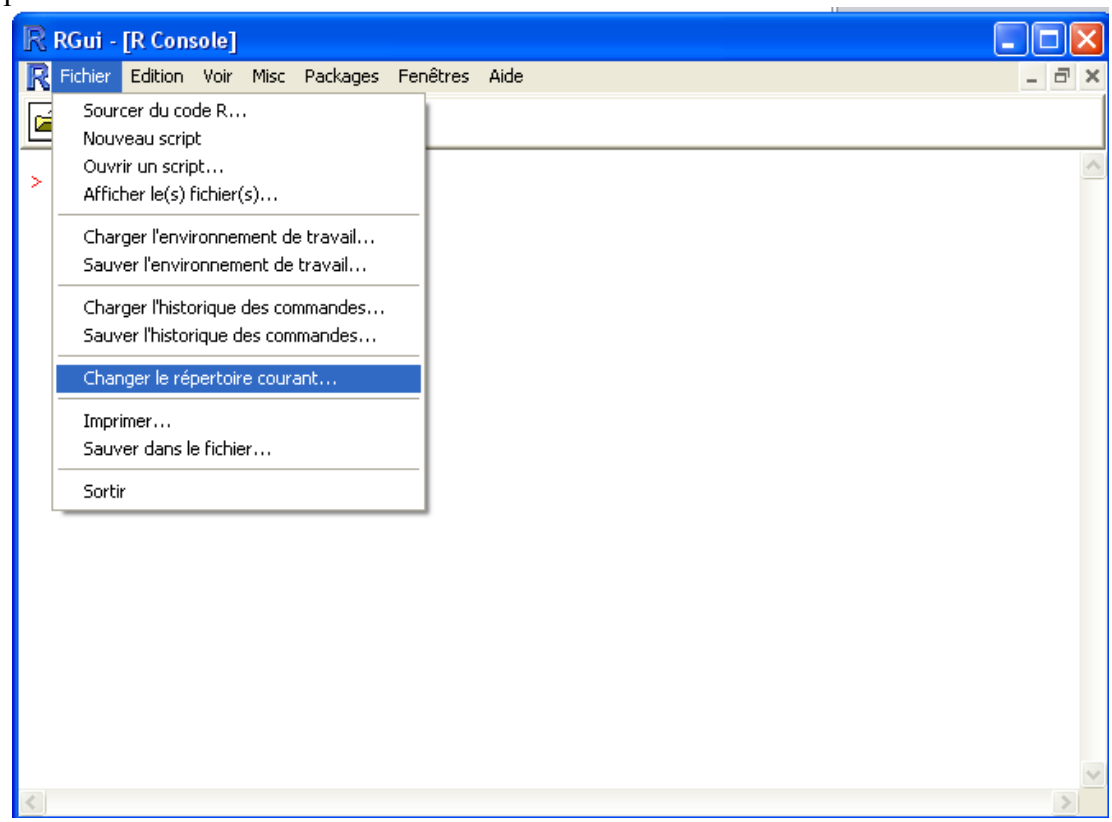

Illustration 22 : Choix du répertoire de travail- Aperçu d'écran

4.2.2. Exécution du module de lecture des données

Le lancement de ce module s'effectue ensuite en tapant la commande `source("lecture.r")`

Exemple de commandes tapées pour le lancement du script de lecture des données :

```
> setwd("D:/Travail/Tendance/outilR/")

> source("lecture.r")
```

Au lancement du script, une fenêtre contenant l'arborescence des fichiers s'ouvre, vous permettant de sélectionner le fichier texte contenant vos données.

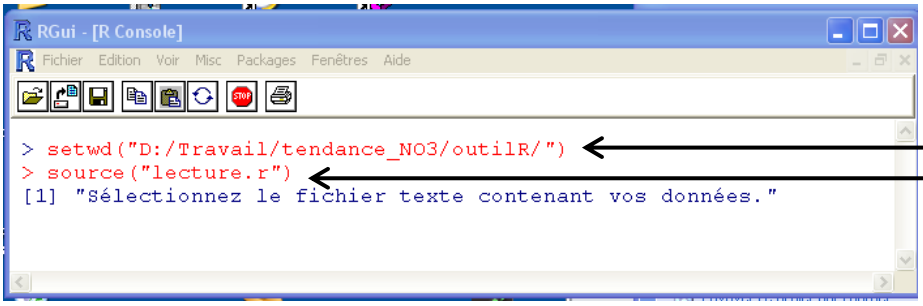

Illustration 23 : Copie d'écran de l'interface graphique de R à l'exécution du script « lecture.r »

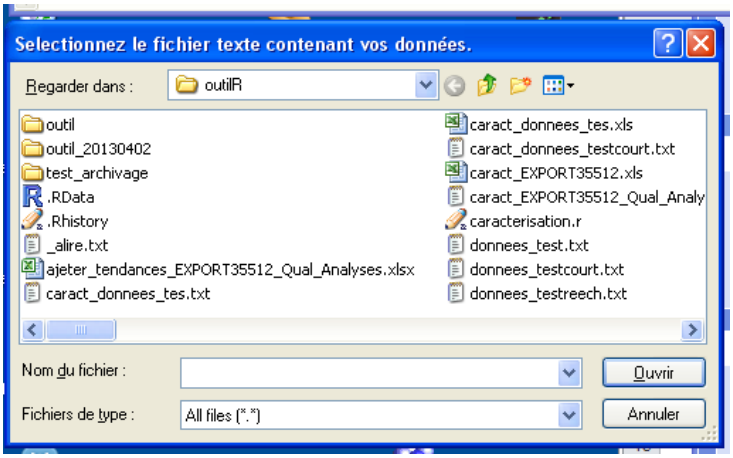

Illustration 24 : Fenêtre contenant l'arborescence des fichiers permettant de sélectionner le fichier contenant vos données.

Il vous est demandé ensuite quel est le caractère séparateur de colonnes utilisé dans le fichier contenant les données. Il vous faut alors simplement taper 1 ou 2, suivant le caractère utilisé puis appuyer sur la touche entrée.

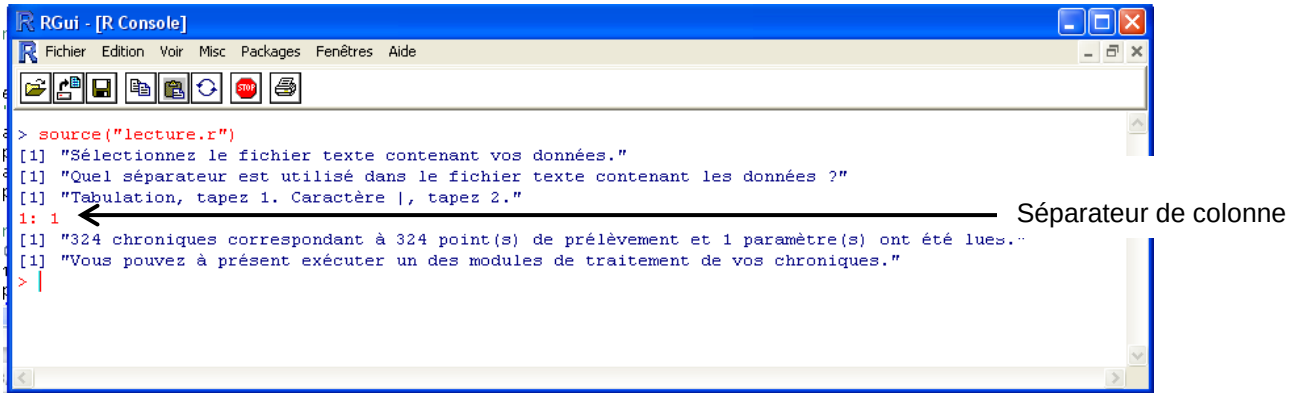

Illustration 25 : Aperçu d'écran à l'exécution du script lecture.r

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"X chroniques correspondant à Y point(s) de prélèvement et Z paramètre(s) ont été lues.

Vous pouvez à présent lancer un des modules de traitement de vos chroniques."
```

S'il manque une des colonnes obligatoires, un message d'erreur s'affiche vous indiquant la colonne manquante.

Les données sont gardées en mémoire tant que l'interface graphique de R n'est pas fermée ou que le script `lecture.r` n'a pas été relancé.

4.3. MODULE « CARACTERISATION »

4.3.1. Objectifs

Ce module permet d'appréhender la structure des données en proposant d'une part leur représentation graphique et d'autre part le calcul des statistiques de base. L'exécution de ce script entraîne la création d'un fichier texte qui récapitule les principales caractéristiques de chaque chronique et, si l'utilisateur le demande, un fichier pdf de représentation graphique des données.

Les fichiers de sortie créés sont enregistrés dans le répertoire dans lequel se situent vos données d'entrée.

Le détail des calculs effectués est présenté dans la partie 2 du document.

4.3.2. Exécution du module

Avant d'exécuter ce script, il faut avoir au préalable exécuté le script `lecture.r` afin que les données soient chargées (§ 4.2).

Le lancement de ce module s'effectue en tapant la commande suivante :

```
> source("caracterisation.r")
```

Si des fichiers portant le même nom que des fichiers de sortie à créer sont ouverts au moment du lancement du module, un message d'erreur apparaît vous demandant de fermer les fichiers. Vous devrez ensuite relancer l'exécution du module.

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"Le module caractérisation a bien été exécuté. Les fichiers résultats ont été créés dans le répertoire contenant votre fichier de données"
```

4.3.3. Fichiers de sortie

Les fichiers de sortie créés sont enregistrés dans le répertoire dans lequel se situent vos données d'entrée.

a) Tableau récapitulatif

Le tableau récapitulatif est nommé « caract_*nom du fichier d'entrée*.txt».

Il comprend, pour chaque chronique, les informations présentées dans le tableau ci-dessous.

Titre de la colonne	Explication
CODE_BSS	Identifiants de la chronique
LIBELLE_PARAMETRE	
Date min	Date de la première analyse de la série temporelle étudiée
Date max	Date de la dernière analyse de la série temporelle étudiée
Nbre analyses	Nombre de données qui composent la série temporelle étudiée
Longueur de la chronique (jours)	Durée de la chronique, correspond à la différence temporelle entre Date max et Date min.
Moyenne des résultats	La moyenne est calculée en considérant les valeurs inférieures à la LQ égale à LQ/2
Médiane des résultats	Pour le calcul de la médiane tous les résultats sont pris en compte tels quels, qu'ils correspondent à une quantification ou non.
Remarque médiane	Remarque dans le cas où la médiane est inférieure à la limite de quantification/détection la plus élevée
Ecart-type des résultats	Pour le calcul de l'écart-type les résultats de. L'écart-type calculé est l'écart-type non biaisé.
Premier quartile des résultats	Pour le calcul du premier décile tous les résultats sont pris en compte tels quels, qu'ils correspondent à une quantification ou non.
Remarque premier quartile	Remarque dans le cas où le premier décile est inférieure à la limite de quantification/détection la plus élevée
Dernier quartile résultats	Pour le calcul du troisième décile tous les résultats sont pris en compte tels quels, qu'ils correspondent à une quantification ou non.
Remarque dernier quartile	Remarque dans le cas où le troisième décile est inférieur à la limite de quantification/détection la plus élevée
Taux de quantification	Le taux de quantification est calculé à partir de la colonne de qualification du résultat. Il est égal au rapport du nombre d'analyses ayant un code remarque égal à 1 par le nombre total d'analyses.
LQ min tt codes	Limite de quantification minimum sur la chronique en considérant tous les codes signes différents de 1.
LQ max tt codes	Limite de quantification maximum sur la chronique en considérant tous les codes signes différents de 1.
LQ min code 10	Limite de quantification minimum sur la chronique en considérant uniquement les codes signes égaux à 10.
LQ max code 10	Limite de quantification maximum sur la chronique en considérant uniquement les codes signes égaux à 10.
Moyenne du nombre de jours d'écarts entre deux analyses consécutives	Somme des durées entre deux analyses consécutives divisée par le nombre d'intervalles entre deux analyses consécutives
Ecart-type du nombre jours d'écarts entre deux analyses consécutives	L'écart-type calculé est l'écart-type non biaisé.
Moyenne du nombre de jours d'écarts entre deux analyses sans outliers	Mêmes statistiques que les deux précédentes en retirant les outliers de la série des nombres de jours d'écart en deux analyses (valeurs situées au-delà de deux écart-type autour de la moyenne)
Ecart-type du nombre jours d'écarts entre deux analyses sans outliers	Ecart type non biaisé sans outliers
p-value - test de Shapiro	P-value du test de Shapiro. Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle les données sont normalement distribuées.

Normalité de la distribution des données	Interprétation du résultat du test de Shapiro ou, si le test n'a pas été effectué, raisons de la non-application
Valeur de l'autocorrélation au rang 1	Valeur calculée après analyse de l'autocorrélogramme de la série chronologique
Significativité de l'autocorrélation	Interprétation de l'autocorrélation ou, si le test n'a pas été effectué, raisons de la non-application

Illustration 26 : Récapitulatif des paramètres de sortie du module « caractérisation » (script : caracterisation.r).

Tous les détails relatifs aux calculs de ces différentes valeurs sont exposés dans le rapport Lopez et al. (2013).

b) Graphiques

Si vous avez saisi « oui » à la question de savoir si vous vouliez une représentation graphique, un fichier pdf nommé « graphes_caract_*nom du fichier d'entrée*.pdf» va être créé. Il contient une page par chronique.

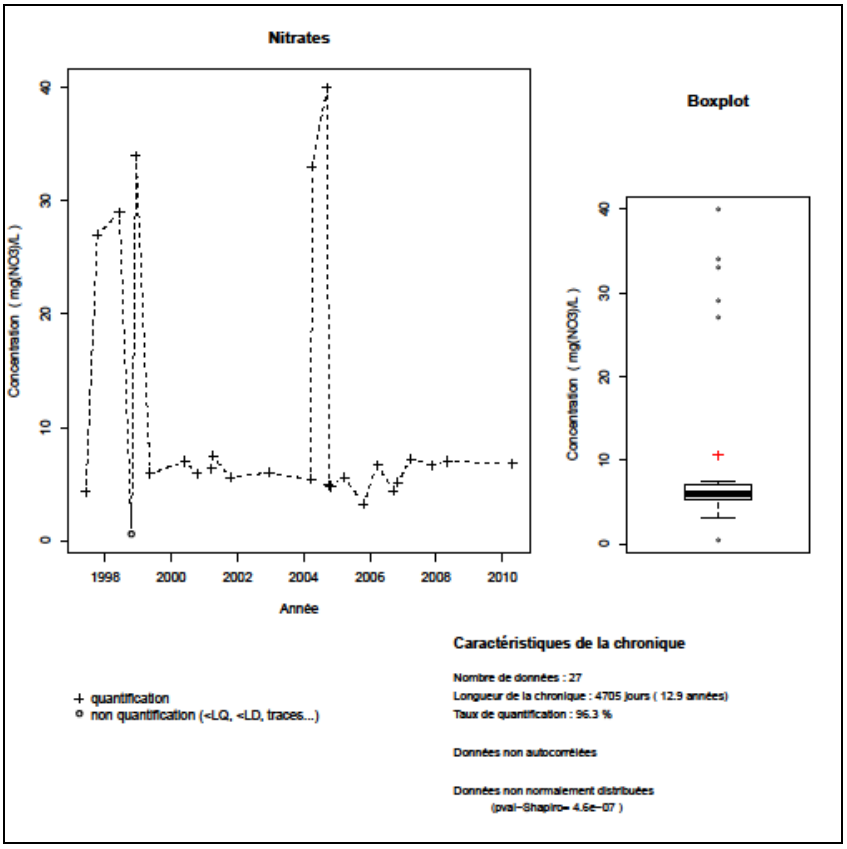

Illustration 27 : Exemple de sortie graphique du module « caractérisation » de HYPE appliqué sur une chronique d'évolution des concentrations en nitrate dans les eaux souterraines.

Comme présenté sur l'Illustration 27, sur chaque page sont représentées :

- Dans le grand quart en haut à gauche, la chronique, en différentiant les valeurs quantifiées des valeurs inférieures aux limites de quantification ou de détection. Dans ce dernier cas, la valeur reportée sur le graphique est la limite analytique de quantification (ou de détection).

- En haut à droite, un diagramme en boîte à moustache, qui représente la répartition statistique des données de la série. Les informations fournies par ce type de diagramme sont présentées sur l'Illustration 28. Des détails sur la construction des boîtes à moustaches sont donnés dans le rapport correspondant à l'étude (Lopez et al., 2013).

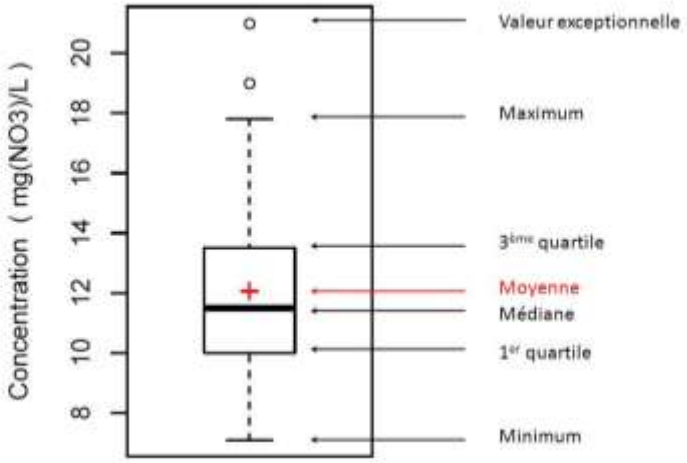

Illustration 28 : Aide à la lecture d'un diagramme en « boîte à moustache ».

- En bas à gauche la légende du graphique de la chronique.

- En bas à droite, sont données quelques caractéristiques de la chronique : nombre d'analyse, longueur de la chronique, résultats des tests de normalité et d'autocorrélation

4.4. MODULE « TENDANCES & RUPTURES »

4.4.1. Objectifs

Ce module permet d'appliquer des tests statistiques de détection de tendance et de rupture sur les chroniques.

En fonction des caractéristiques des chroniques (nombres d'analyses, normalité de la distribution, autocorrélation…), l'outil détermine automatiquement les tests les plus robustes à appliquer. L'Illustration 29 présente l'arbre décisionnel appliqué dans l'outil HYPE.

Dans ce module, les résultats d'analyses reportés inférieurs à une limite (de quantification ou de détection) sont substitués par la valeur correspondante de la limite de quantification ou de détection. Il convient alors **d'interpréter avec grande prudence les chroniques présentant des taux de quantification faibles**.

La description des tests statistiques est donnée détaillée en partie 2.

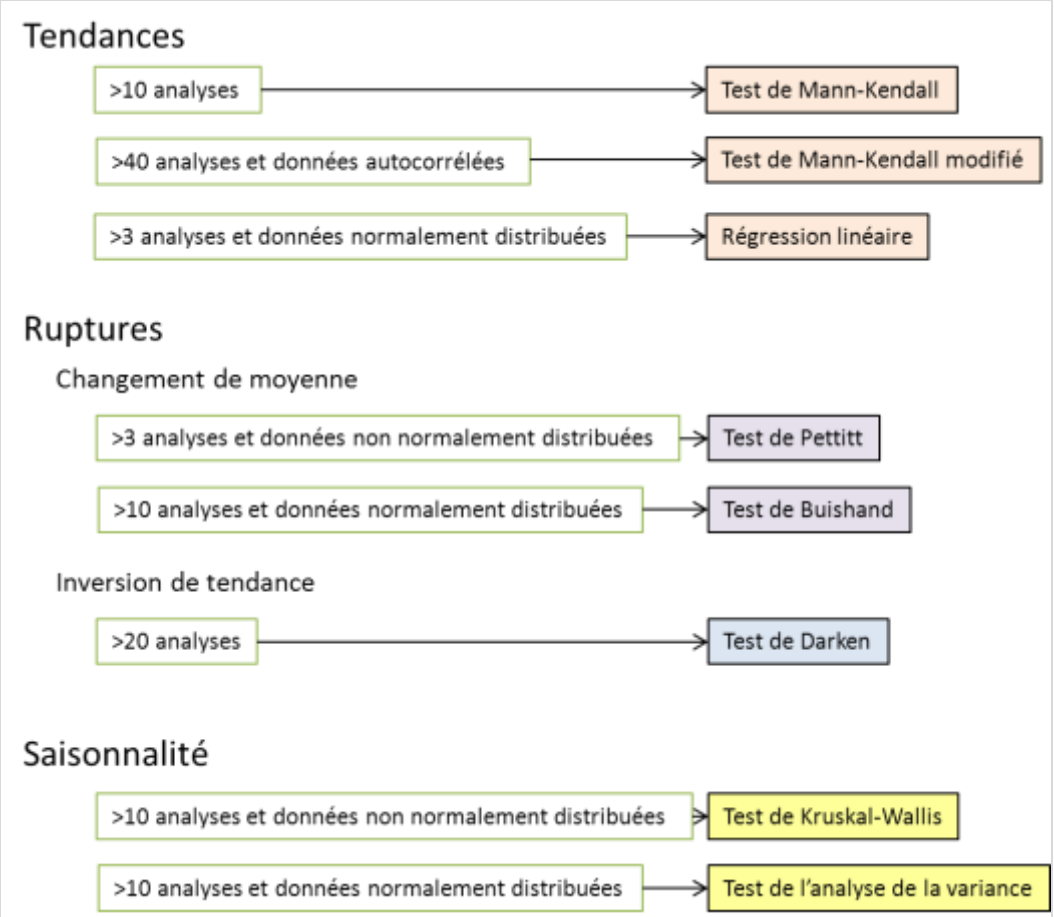

Illustration 29 : Schéma récapitulatif des critères de sélection automatique des tests appliqués dans dans le module « tendances et ruptures » de HYPE en fonction des conditions intiales des données compilées.

c) *Recherche de tendance*

Deux tests de tendance peuvent être appliqués. Dans le cas où les données sont normalement distribuées, une régression linéaire est appliquée si la chronique comporte au moins 3 données et un test de Mann-Kendall est également appliqué si la chronique comporte au moins 10 données. Dans le cas où les données ne sont pas normalement distribuées, seul le test de Mann-Kendall est appliqué si la chronique compote au moins 10 données.

De plus, si les données présentent une autocorrélation significative et si la chronique dispose d'au moins 40 données, un test de Mann-Kendall modifié est appliqué. La p-value de ce test est différente de celle du test non modifié ; elle tient compte de l'autocorrélation. En ce qui concerne la pente de Sen, ce test est équivalent au test de Mann-Kendall non modifié, elle n'est donc pas recalculée.

d) *Recherche de ruptures*

Deux types de ruptures sont recherchés dans les chroniques :

La présence d'un **changement significatif de moyenne** est recherchée à l'aide d'un test d'homogénéité :

- test de Buishand si les données sont normalement distribuées (Illustration 12),

- test de Pettitt si les données ne sont pas normalement distribuées (Illustration 10).

La présence d'une **inversion de tendance** est également recherchée. Pour cela une méthode tirée des travaux de Darken est appliquée si la chronique dispose d'au moins 20 données (Illustration 13).

e) *Etude du caractère saisonnier de la chronique*

Le caractère saisonnier des chroniques est étudié par deux tests :

- un test d'analyse de variance dans les cas où les données sont normalement distribuées (Illustration 8),

- un test de Kruskal-Wallis dans le cas où les données ne sont pas normalement distribuées (Illustration 9).

Ces tests permettent de déterminer si les données de chaque période sont significativement différentes les unes des autres. Les périodes considérées par HYPE sont les trimestres calendaires (1er janvier au 31 mars, 1er avril au 30 juin, 1er juillet au 30 septembre, 1er octobre au 31 décembre).

4.4.2. Exécution du module

Avant d'exécuter ce script, il faut avoir au préalable exécuté le script `lecture.r` afin que les données à traiter soient chargées. Si le script `lecture.r` a déjà été exécuté, pour le module « caractérisation » par exemple, il n'est pas nécessaire de refaire la manipulation.

Le lancement de ce module s'effectue en tapant la commande suivante :

```
> source("tendances_ruptures.r")
```

Si des fichiers portant le même nom que des fichiers de sortie à créer sont ouverts au moment du lancement du module, un message d'erreur apparaît vous demandant de fermer les fichiers. Vous devrez ensuite relancer l'exécution du module.

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"Le module tendances et ruptures a bien été exécuté. Les fichiers résultats
ont été créés dans le répertoire contenant votre fichier de données"
```

4.4.3. Fichiers de sortie

Les fichiers de sortie créés sont enregistrés dans le répertoire dans lequel se situent vos données d'entrée.

a) *Tableau récapitulatif*

Le tableau de sortie, appelé « tendances_*nom du fichier d'entrée*.txt», présente les différents tests effectués, leur significativité et leurs résultats.

L'ensemble des informations que l'on peut trouver dans le tableau est récapitulé dans le tableau ci-dessous.

Titre de la colonne	Remarque
CODE_BSS	Identifiants de la chronique
LIBELLE_PARAMETRE	
Date min	Date de la première analyse de la série temporelle étudiée
Date max	Date de la dernière analyse de la série temporelle étudiée
Nbre analyses	Nombre de données qui composent la série temporelle étudiée
Longueur de la chronique (jours)	Durée de la chronique, correspond à la différence temporelle entre Date max et Date min.
p-value - test de Shapiro	P-value du test de Shapiro. Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle les données sont normalement distribuées.
Normalité de la distribution des données	Interprétation du résultat du test de Shapiro ou, si le test n'a pas été effectué, raisons de la non-application
p-value - test de Mann-Kendall	P-value du test de Mann-Kendall . Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas de tendance.
Tau - test de Mann-Kendall	Statistique tau du test de Mann-Kendall
Pente de Sen - test de Mann-Kendall (unité/an)	Paramètres de la droite de tendance si elle est significative
Ordonnée à l'origine - test de Mann-Kendall	
Tendance - test de Mann-Kendall	Interprétation du résultat du test de Mann-Kendall ou raison pour laquelle le test n'a pas été effectué.
p-value - régression linéaire	P-value de la régression linéaire . Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas de tendance.
r carré - régression linéaire	R carré de la régression linéaire
Pente - régression linéaire (unité/an)	Paramètres de la droite de régression si elle est significative
Ordonnée à l'origine - régression linéaire	
Tendance - régression linéaire	Interprétation du résultat de la régression linéaire ou raison pour laquelle le test n'a pas été effectué.
Valeur de l'autocorrélation au rang 1	Valeur de l'autocorrélation au rang 1
Significativité de l'autocorrélation	Interprétation de l'autocorrélation ou raison pour laquelle le test n'a pas été effectué
p-value - test de Mann-Kendall modifié	P-value du test de Mann-Kendall modifié.
Tendance - test de Mann-Kendall modifié	Interprétation du résultat du test de Mann-Kendall modifié ou raison pour laquelle le test n'a pas été effectué.
p-value - test de changement de moyenne	P-value du test de changement de moyenne. Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas de rupture significative.
Date rupture - test de changement de moyenne	Date de changement de moyenne si elle est significative.
Test de changement de moyenne	Description du test de changement de moyenne appliqué et interprétation du résultat ou raison pour laquelle le test n'a pas été effectué.
Moyenne - tronçon pré-rupture	Moyenne pré- et post- rupture si le test de changement de moyenne est significatif

Moyenne - tronçon post-rupture	
p-value - test d'inversion de pente de Darken	P-value du test d'inversion de tendance. Si la valeur est inférieure à 0.05, on considèrera qu'on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas d'inversion de tendance.
Date rupture - test d'inversion de pente de Darken	Date de l'inversion de tendance si elle est significative.
Test d'inversion de pente de Darken	Interprétation du résultat du test d'inversion de tendance ou raison pour laquelle le test n'a pas été effectué.
Pente de Sen - tronçon pré-rupture (unité/an)	Paramètres des droites de tendance pré- et post- inversion de tendance si elles sont significatives
Ordonnée à l'origine - tronçon pré-rupture	
Pente de Sen - tronçon post-rupture (unité/an)	
Ordonnée à l'origine - tronçon post-rupture	
p-value du test de variabilité	P-value du test de variabilité entre saisons. Si la valeur est inférieure à 0.05, on considèrera qu'au moins deux saisons sont significativement différentes l'une de l'autre.
Analyse de la variabilité entre saisons	Description du test de variabilité entre saison appliqué et interprétation du test ou raison pour laquelle le test n'a pas été effectué.

Illustration 30 : Récapitulatif des paramètres de sortie du module « tendances et ruptures » (script : tendances_ruptures.r).

b) Sortie graphique

Si vous avez saisi « oui » à la question de savoir si vous vouliez une représentation graphique, un fichier pdf, appelé « graphes_tendances_*nom du fichier d'entrée*.txt», va être créé, contenant une page par point BSS.

L'Illustration 31 présente les différentes informations visibles sur les graphiques de sortie obtenus après application du module « Tendance et ruptures » de HYPE. Les caractéristiques principales des données temporelles sont rappelées dans le quart en bas à gauche tandis que les résultats numériques des différents tests statistiques réalisés sont reportés dans le quart en bas à droite de la sortie graphique. Le dessin de la chronique occupe la moitié supérieure de la sortie graphique. Les droites des tendances significatives sur la longueur totale de la série ainsi que sur les tronçons avant et après inversion de pente sont superposées à la représentation de la série temporelle. Il en est de même pour les moyennes dans le cas où la chronique n'est pas homogène.

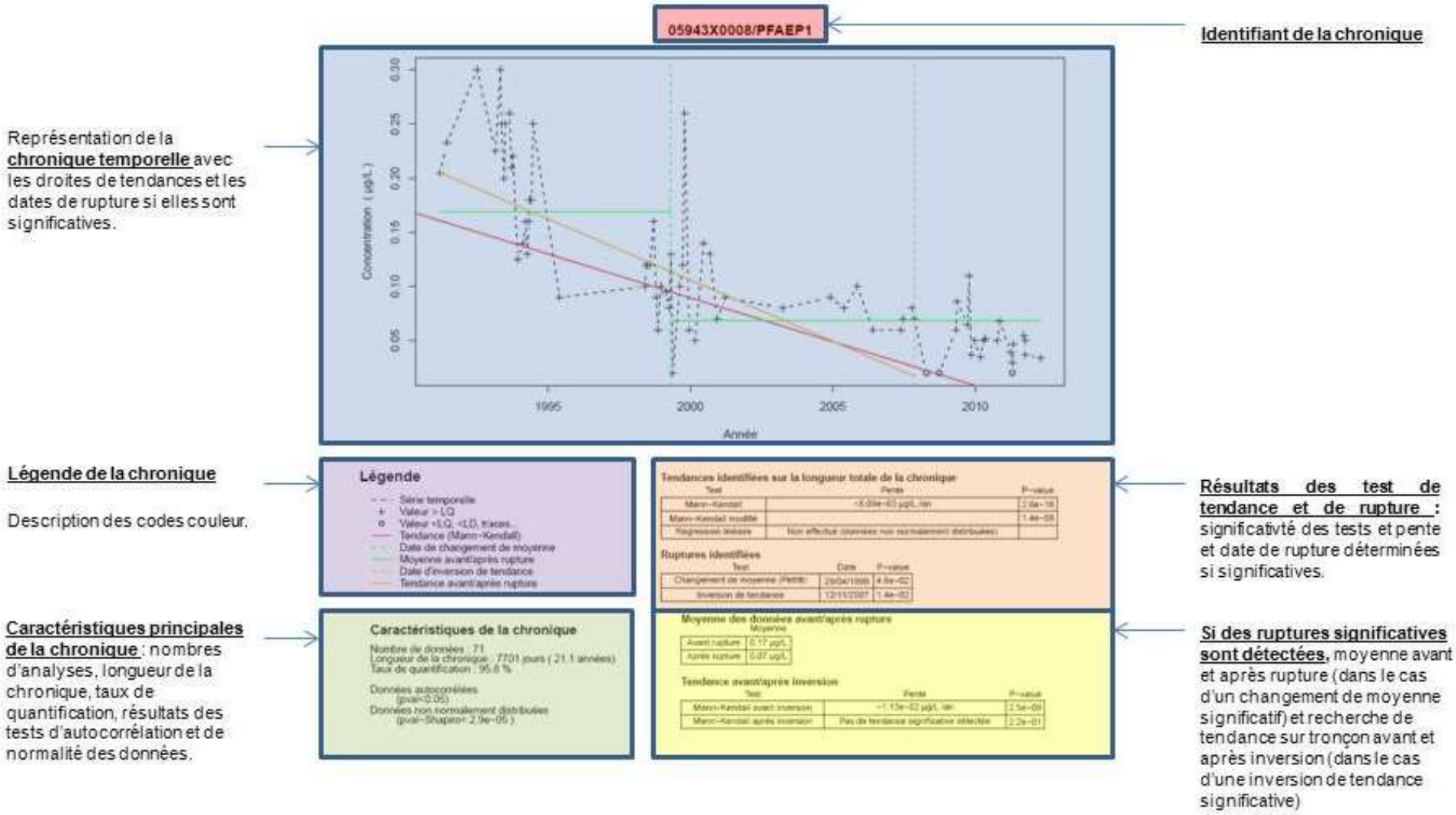

Illustration 31 : Exemple de sortie graphique obtenue avec le module « Tendances et ruptures » pour une chronique de nitrates au point BSS 05943X0008/PFAEP1.

4.5. MODULE « REGIONAL »

4.5.1. Objectifs

Ce module permet d'effectuer un test de Kendall régional. Ce test permet d'étudier la présence de tendance à l'échelle d'une région d'étude, appelée ici unité spatiale, comprenant plusieurs points de prélèvement ou plusieurs paramètres.

Le principe du test est de déterminer si une tendance cohérente peut être mise en évidence à partir des différentes chroniques appartenant à une même unité spatiale.

Pour utiliser ce module, le tableau des données peut contenir une colonne **UNITE_SPATIALE** ou **Unité spatiale** précisant à quelle unité spatiale est rattaché le point (Illustration 21). En l'absence d'une telle colonne, le test régional est effectué sur la totalité des chroniques du fichier d'entrée.

Pour les détails de la mise en œuvre du test, on se reportera au paragraphe 2.6.

4.5.2. Exécution du module

Avant d'exécuter ce script, il faut avoir au préalable exécuté le script `lecture.r` afin que les données soient chargées.

Le lancement de ce module s'effectue en tapant la commande suivante :

```
> source("mk_regional.r")
```

Si un fichier portant le même nom que le fichier de sortie à créer est ouvert au moment du lancement du module, un message d'erreur apparaît vous demandant de fermer le fichier. Vous devrez ensuite relancer l'exécution du module.

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"Le module régional a bien été exécuté. Le fichier résultat a été créé
dans le répertoire contenant votre fichier de données"
```

4.5.3. Fichier de sortie

Ce module crée en fichier de sortie un tableau comprenant, pour chaque unité spatiale, les résultats du test de Kendall régional : tau, p-value et pente de Sen.

Titre de la colonne	Remarque
Unité spatiale	Code commun aux différentes chroniques prises en compte pour le calcul de la tendance régionale
Nombre total de points	Nombre total de points dans l'unité spatiale
Nombre de points pour lesquels un test de Mann-Kendall a été effectué	Nombre de points pour lesquels un tes de Mann-Kendall a été réalisé (disposant d'au moins 10 analyses et chronique non stationnaire)
p-value - Kendall régional	p-value du test de Kendall regional
Tau – Kendall régional	Statistique tau du test de Kendall régional
Pente de Sen	Pente de Sen déterminée si le test est significatif
Tendance – Kendall régional	Interprétation du test de Kendall régional ou raison pour laquelle le test n'a pas été effectué

Illustration 32 : Récapitulatif des paramètres de sortie du module régional

4.6. MODULE « SAISONNIER »

4.6.1. Objectifs

Ce module permet d'effectuer un test de Kendall saisonnier. Ce test permet d'estimer des tendances de séries cycliques saisonnières.

Le principe est identique à celui du test de Mann-Kendall mais le caractère saisonnier de la série est pris en compte. Autrement dit pour des données mensuelles ayant une saisonnalité de 12 mois, on ne va pas chercher à savoir s'il y une tendance au global sur la série, mais simplement si, d'un mois de janvier à l'autre, d'un mois de février à l'autre, et ainsi de suite, il y a une tendance.

L'outil effectue le test saisonnier sur deux périodes différentes : le trimestre (de janvier à mars, avril à juin, juillet à septembre et octobre à décembre) et le mois calendaire.

Pour les détails de la mise en œuvre du test, on se reportera au paragraphe 2.5.

4.6.2. Exécution du module

Avant d'exécuter ce script, il faut avoir au préalable exécuté le script `lecture.r` afin que les données soient chargées.

Le lancement de ce module s'effectue en tapant la commande suivante :

```
> source("mk_saisonnier.r")
```

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"Le module saisonnier a bien été exécuté. Le fichier résultat a été
créé dans le répertoire contenant votre fichier de données"
```

4.6.3. Fichier de sortie

Ce module crée en fichier de sortie un tableau comprenant, pour chaque série temporelle, les résultats du test de Kendall saisonnier : tau, p-value et pente de Sen.

Titre de la colonne	Remarque
CODE_BSS	Identifiants
LIBELLE_PARAMETRE	
p-value - Kendall saisonnier trimestre	p-value du test de Kendall saisonnier en considérant comme période le trimestre
Tau - Kendall saisonnier trimestre	Statistique tau du test de Kendall saisonnier
Pente de Sen - Kendall saisonnier trimestre	Pente de Sen déterminée si le test est significatif
Tendance - Kendall saisonnier trimestre	Interprétation du test de Kendall saisonnier ou raison pour laquelle le test n'a pas été effectué
p-value - Kendall saisonnier mois	Mêmes paramètres mais en considérant comme période le mois calendaire.
Tau - Kendall saisonnier mois	
Pente de Sen - Kendall saisonnier mois	
Tendance - Kendall saisonnier mois	

Illustration 33 : : Récapitulatif des paramètres de sortie du module saisonnier.

4.7. MODULE « RE-ECHANTILLONNAGE »

4.7.1. Objectifs

Ce module permet d'échantillonner une chronique pour en extraire les analyses effectuées avec une périodicité particulière. Deux périodicités sont proposées par défaut par l'outil HYPE : annuelle et mensuelle. Les données ainsi extraites sont écrites dans un fichier résultat sous une forme lisible par les modules précédemment cités. L'ensemble des tests de caractérisation de la chronique et de recherche de tendances et de ruptures peuvent donc être effectués non plus sur les chroniques totales mais sur des chroniques partielles ré-échantillonnées.

L'échantillonnage avec périodicité annuelle permet d'extraire des analyses effectuées au même moment de l'année. L'outil n'extrait que des chroniques disposant d'analyses sur 10 années différentes au moins. Si pour une année, la chronique dispose de plusieurs analyses, la valeur prise en compte est la moyenne des résultats des analyses et si parmi ces analyses, il y a au moins une quantification, le résultat pour la période est considéré comme une quantification.

L'échantillonnage avec périodicité mensuelle permet d'extraire des analyses effectuées avec une régularité mensuelle. L'outil n'extrait que des chroniques disposant d'analyses sur 10 mois différents au moins et pour lesquelles au moins 70% des mois possèdent une analyse. A l'instar de l'échantillonnage annuel, si pour un mois, la

chronique dispose de plusieurs analyses, la valeur prise en compte est la moyenne des résultats des analyses et si parmi ces analyses, il y a au moins une quantification, le résultat pour la période est considéré comme une quantification

L'utilisateur doit préciser la taille de la fenêtre dans laquelle l'outil peut échantillonner la chronique, c'est–à-dire à plus ou moins combien de jours autour d'une date donnée, une analyse sera considérée dans la période fixée.

Au choix de l'utilisateur, l'outil peut extraire la chronique la plus longue ayant la périodicité choisie ou extraire une chronique autour d'une date donnée.

4.7.2. Exécution du module

Avant d'exécuter ce script, il faut avoir au préalable exécuté le script `lecture.r` afin que les données soient chargées.

Le lancement de ce module s'effectue en tapant la commande suivante :

```
> source("reechantillonage.r")
```

Trois paramètres sont ensuite à renseigner :

- la périodicité : mensuelle ou annuelle ;
- le mode d'échantillonnage : automatique (l'outil extrait la chronique la plus longue) ou autour d'une date fixée par l'utilisateur ;
- la taille de la fenêtre d'échantillonnage.

Lorsque l'exécution du script est terminée, le message suivant s'affiche

```
"Le module rééchantillonnage a bien été exécuté. Le fichier contenant
les chroniques rééchantillonnées a été créé dans le répertoire
contenant votre fichier de données "
```

4.7.3. Fichiers de sortie

Le module créé les fichiers suivants :

- Un tableau récapitulatif qui comprend :
    - L'identifiant de la chronique (identifiant du point de prélèvement et/ou paramètre analysé)
    - Un texte précisant un sous-échantillon de la chronique qui a pu être trouvé suivant les paramètres indiqués ;
    - La date centrale de la période de recherche
    - Le nombre d'années ou de mois disposant d'une analyse

    - La longueur totale en année ou en mois du sous-ensemble identifié

- Dans tous les cas, un fichier texte reprenant les analyses effectuées avec une périodicité annuelle pour les points BSS disposant d'une chronique annuelle de plus de 10 années. Seuls les champs obligatoires (code BSS, date, code signe, résultat, unité et code paramètre si présent initialement) sont repris dans ce fichier.

5. Synthèse opérationnelle

La plaquette suivante résume en une feuille l'ensemble des commandes qu'il est possible de saisir sous R afin de faire fonctionner l'outil HYPE.

Après avoir définit le répertoire de travail et lu les données à analyser, 2 modules principaux (« caractérisation » et « tendance et ruptures ») et 3 modules complémentaires (« saisonnier », « régional » et « ré-échantillonnage ») peuvent être envisagés. Les modules principaux proposent à la fois une sortie graphique en format pdf et une sortie sous forme de tableau en format txt. Les modules complémentaires ne proposent qu'une sortie sous forme de tableau.

Une fois bien assimilé les différentes opérations compilées dans l'outil HYPE, il devient possible de ne travailler qu'avec la plaquette d'utilisation qui sert alors de « pense bête » pour lancer les scripts appropriés. Il conviendra alors à l'opérateur d'interpréter les résultats fournis par l'outil afin notamment de déterminer si la significativité statistique révélée par HYPE trouve un écho d'un point de vue environnemental.

Utilisation de HYPE - Commandes à saisir

Le répertoire de travail doit contenir les différents modules de l'outil et dans le cas où les unités sont définies par leur libellé complet, le fichier Unite_SANDRE.txt »

Définition du répertoire de travail

```
> setwd("chemin")
```

"*chemin*"est votre répertoire de travail.
Par exemple : "D:/Travail/tendance/outilR/"

Lecture des données

```
> source("lecture.r")
```

Sélection du fichier contenant les données d'entrée

Caractérisation des données

```
> source("caracterisation.r")
```

- Tableau récapitulatif des statistiques de base
- Représentation graphique des chroniques et de la distribution des données en boîte à moustache

Recherche de tendances et de ruptures

```
> source("tendances_ruptures.r")
```

- Tableau récapitulatif des tests effectués et de leurs résultats
- Représentation graphique des chroniques avec les tendances et les ruptures significatives identifiées

Kendall saisonnier

```
> source("mk_saisonnier.r")
```

- Tableau récapitulatif des tests effectués et de leurs résultats

Kendall régional

```
> source("mk_regional.r")
```

- Tableau récapitulatif des tests effectués et de leurs résultats

Echantillonnage

```
> source("reechantillon.r")
```

- Fichier texte contenant les données échantillonnées

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-54884-FR, 105 p.

Buishand T.A. (1982). Some methods for testing the homogeneity of rainfall records. . Journal of Hydrology 58, 11-27.

Buishand T.A. (1984). Tests for detecting a shift in the mean of hydrogeological time series. Journal of Hydrology 73, 51-69.

Conover W.L. (1980). Practical nonparametric statistics, 2d ed.: New York, John Wiley and Sons, 493p.

Darken P.F. (1999). Testing for changes in trend in water quality data. PhD Faculty of Virginia Polytehcnic Institute and State University.

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

Hirsch R.M., Slack J.R., Smith R.A. (1982). Techniques of trend analysis for monthly water quality data. Water Resources Research 18, 107-121.

Hyndman R., Fan Y. (1996). Sample quantiles in Statistical Packages, The American Statistician, 50 (4), 361-365.

Lopez B., Baran N., (2011) - Etude de faisabilité pour l'estimation des tendances d'évolution de la qualité des eaux souterraines du bassin Rhin-Meuse. Rapport BRGM/RP-60649-FR, 138 p.

Lopez B., Leynet A. et al. (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 final. BRGM/RP-59515-FR, 166 p., 48 ill.

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, 93 p., 45 ill., 1 ann.

Pettitt A.N. (1979). A non-parametric approach to the change-point problem. Applied Statistics 28, 126-135.

Renard B. (2006). Détection et prise en compte d’éventuels impacts du changement climatique sur les extrêmes hydrologiques en France. Thèse de l’Institut National Polytechnique de Grenoble. Unité de Recherche Hydrologie-Hydraulique, Cemagref (Lyon).

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

Shapiro S.S. et Wilk M.B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52,2 and 3, 591-597.

Géosciences pour une Terre durable
brgm

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