15/06/2020

Note méthodologique

Tendances sur les séries chronologiques

pour le

SERVICE PUBLIC DE WALLONIE
AGRICULTURE, RESSOURCES NATURELLES, ENVIRONNEMENT

DEPARTEMENT DE L'ENVIRONNEMENT ET DE L'EAU (DEE)

Wallonie environnement SPW

Ephesia

Historique du document

Version	Date	Auteurs	Commentaires
1.0	24/02/2020	Dimitri D'Or et Denis Allard	Version initiale

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

1. INTRODUCTION

Une première étude en 2014

En 2014, la Commission européenne a adressé au Service Public de Wallonie une mise en demeure concernant la protection des eaux contre la pollution par les nitrates à partir de sources agricoles.

Dans ce contexte, la Région wallonne a fait réexaminer les données de concentrations en nitrates dans les eaux souterraines et dans les eaux de surfaces afin de mettre en évidence d'éventuelles tendances ainsi que des ruptures de tendance sur les séries chronologiques de mesures de concentration en nitrates.

Un workflow complet avait été élaboré et appliqué (D'Or et Allard, 2014) à l'analyse des séries chronologiques disponibles pour 986 sites pour les eaux souterraines, 55 sites pour les eaux de surfaces et 158 séries issues de modélisation du transfert des nitrates vers les nappes (65 relatives aux eaux de surface, 65 à la zone racinaire et 28 au niveau de la nappe de base).

Les 986 séries correspondant aux eaux souterraines étaient réparties en 33 masses d'eau. Une analyse globale par masse d'eau a également été réalisée en agrégeant les données de toutes les séries chronologiques qui en font partie. Cette étude était purement statistique et ne cherchait pas à proposer des interprétations hydrogéologiques. L'objectif était de porter un regard purement quantitatif, le plus objectif possible, sur les séries mesurées et d'en extraire des conclusions de type statistique sur les éventuelles tendances.

Un complément d'étude en 2020

Sur cette base, le Service Public de Wallonie souhaite aujourd'hui revoir l'outil et le pérenniser de façon que l'analyse puisse être mise à jour avec de nouvelles données ou reproduites sur d'autres variables.

Dans ce cadre, **ce document détaille l'ensemble des méthodes statistiques** mises en œuvre, avec les nitrates comme exemple. La même méthodologie peut s'appliquer à n'importe quel autre polluant que les nitrates.

Afin de porter un jugement statistique sur la significativité d'une tendance et de sélectionner le modèle adéquat pour celle-ci, nous croisons deux approches statistiques, à savoir la théorie statistique des tests d'hypothèses et la sélection de modèles par vraisemblance pénalisée, et deux cadres mathématiques faisant des d'hypothèses plus ou moins restrictives : les cadres paramétrique et non paramétrique. Nous proposons également une solution permettant de filtrer l'effet de la hauteur d'eau de la nappe lorsque des données piézométriques sont disponibles. En amont de toute analyse, nous proposons d'éliminer de l'analyse les éventuels outliers par une procédure automatisée.

Le plan de ce document est le suivant. Au chapitre 2, nous détaillons les méthodes statistiques mobilisées. En fin de chapitre nous présentons le workflow qui est appliqué sur chaque série chronologique. Au chapitre 3, nous présentons la méthode pour réaliser une analyse agrégée à l'échelle d'une masse d'eau entière. Au chapitre 4, nous détaillons la méthodologie adoptée pour la classification des eaux souterraines.

2. LES METHODES STATISTIQUES MOBILISEES

2.1 Rappel sur les approches paramétriques et non-paramétriques

On parle **d'approches paramétriques** lorsque l'on fait des hypothèses sur la ou les distributions statistiques des données. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide des paramètres estimés sur l'échantillon ; les procédures d'estimation et de test subséquente ne portent alors que sur ces paramètres, et on utilise des résultats mathématiques sur la distribution des estimateurs de ces paramètres. Le plus souvent, on fait une hypothèse de normalité sur les données, ou sur les résidus à un modèle de régression (cf ci-dessous pour les différents modèles de régression possibles) ; sous cette hypothèse la moyenne et la variance suffisent pour caractériser complètement la distribution.

Les approches **non paramétriques** ne font aucune hypothèse sur la distribution sous-jacente des données ou des résidus à une modèle de régression. Pour réaliser des tests dans ce cadre, on transforme les valeurs en rang, allant de 1 (pour la plus petite) à n (pour la plus grande), et les tests se font sur des grandeurs calculées à partir de ces rangs. En ne faisant aucune hypothèse sur les distributions des données, les tests non paramétriques élargissent le champ d'application des procédures statistiques. En contrepartie, ils sont moins puissants que les tests paramétriques lorsque leurs hypothèses sont compatibles avec les données.

Que l'approche soit paramétrique ou non-paramétrique, dans la suite nous aurons à considérer plusieurs modèles pouvant décrire l'évolution au cours du temps de la teneur en nitrate mesurée dans un ouvrage. Le premier modèle, noté M_0, qui est le plus simple correspond à une absence de changement. Le second modèle, noté M_1, correspond à une variation linéaire avec le temps ; si la pente de la variation est nulle, cela correspond au modèle M_0. Le dernier modèle, noté M_2, est le plus riche. On considère qu'il existe une année de changement (appelée rupture) et que l'on observe deux évolutions linéaires avec des pentes différentes avant et après cette date. A nouveau, le modèle M_1 peut être vu comme un cas particulier du modèle M_2 si la date de rupture correspond à la première ou à la dernière data, ou si les deux pentes sont égales. Mathématiquement, nous écrirons les choses de la façon suivante.

Dans un ouvrage donné, il y a n mesures de nitrate (ou tout autre polluant) Y_i qui ont été effectuées à des dates t_i. On a donc un tableau de n lignes et des valeurs (t_i, Y_i) à chaque ligne $i = 1, \dots, n$. On fera l'hypothèse qu'il existe une fonction, notée f, qui peut décrire la teneur moyenne en nitrate au cours du temps. Ainsi, on aura

$$E[Y(t)] = f(t),$$

où $E[Y]$ désigne l'espérance mathématique de Y. Nous considérerons trois modèles d'évolution par la suite :

- Le modèle de base, M_0, considère que la moyenne est constante au cours du temps, c'est-à-dire que
$$E[Y(t)] = f(t) = a.$$
- Le premier modèle M_1, consiste à faire l'hypothèse que l'espérance mathématique des teneurs en nitrates suit une évolution linéaire en fonction du temps
$$E[Y(t)] = f(t) = a + bt.$$
- Enfin, pour le modèle M_2, on pose
$$E[Y(t)] = f(t) = a + bt, t < T;$$
et
$$E[Y(t)] = f(t) = c + dt, t \geq T\ .$$
On impose en outre la continuité de la courbe à la date de rupture, notée T. A cette date, on a donc :
$$f(T) = a + bT = c + dT.$$

2.2 Méthodes d'estimation pour le modèle linéaire M_1

Cas paramétrique

Dans le cas paramétrique, on réalise l'estimation d'un modèle linéaire classique

$$Y_i = a + bt + \varepsilon\_i, \qquad i = 1, \dots, n$$

Où n est le nombre de données, et pour lequel on obtient une estimation de la pente, de l'ordonnée à l'origine et des résidus. Dans le cadre habituel du modèle linéaire sur résidus Gaussien, on obtient également une p-valeur de la pente, et donc une statistique de test pour rejeter H_0 contre H_1.

Cas non paramétrique

Dans le cas non paramétrique, on ne fait pas d'hypothèse sur la distribution de ε. Nous suivrons la démarche préconisée dans le document du BRGM (Lopez *et al.*, 2013). Nous utiliserons le **test robuste** de Mann-Kendall pour tester un lien monotone entre les variables Y_i et t (Kendall, 1938, repris par Renard, 2006), en lien avec la **méthode de la pente robuste**. Ce test est lié au coefficient de corrélation de Kendall. Il est basé sur les rangs des variables Y_i. L'idée générale est que si Y_i est une variable croissante, les rangs de la variable sont également croissants, indépendamment de la loi de croissance de Y (et inversement dans le cas où Y_i est décroissant). La pente du modèle M_1 est estimée selon la technique de Kendall-Theil, ou méthode de Sen (Sen, 1968 ; Helsel and Hirsch, 1992). L'estimation de la pente robuste est la médiane de toutes les pentes calculées sur les $n(n-1)/2$ couples de points.

2.3 Méthodes d'estimation pour le modèle avec rupture, M2

Nous imposons au modèle M2 la continuité à la date de la rupture, à la différence de l'approche du BRGM qui, n'imposant pas cette continuité, aboutit à des modèles présentant un saut de discontinuité (potentiellement non négligeable) à la date de rupture.

Cas paramétrique

Pour toutes les dates de rupture possibles, on calcule la variance des résidus

$$\varepsilon_i = Y_i - \hat{f}(t_i)$$

où $\hat{f}(t_i)$ est l'estimation de Y_i par le modèle M_2. La date de rupture retenue est celle ayant la variance des résidus la plus faible.

Cas non paramétrique

Pour toutes les dates de rupture possibles, on réalise une estimation robuste des pentes de la régression de part et d'autre de la date de rupture, à l'aide de la méthode de la ligne robuste décrite ci-dessus. L'analyse robuste ne fournissant que des pentes, nous recherchons l'ordonnée à l'origine qui minimise la somme des carrés des résidus. Nous recherchons ensuite la date de rupture qui minimise le carré des résidus entre la variable mesurée et celle prédite par le modèle M_2 ainsi ajusté.

Ces méthodes ne permettent pas de calculer une valeur BIC ou une p-valeur de M_2 contre M_1, puisque nous n'avons pas de modèle pour calculer une vraisemblance. Le choix du modèle M_2 doit se faire sur la base de la significativité des pentes et d'un score SSR plus favorable.

2.4 Les tests d'hypothèses

Un test d'hypothèse est une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (ici la série temporelle des teneurs en nitrates) en confrontant cette hypothèse nulle à une hypothèse dite alternative. A partir de calculs réalisés sur la série observée (les échantillons), nous émettons des conclusions sur l'évolution de la qualité de l'eau mesurée (la population), en leur rattachant des risques de se tromper. Les calculs sont menés en fonction du modèle statistique choisit. Un modèle statistique est un ensemble d'hypothèses mathématiques décrivant la distribution statistique de la population.

Pour mener un test statistique il faut :

1. Formuler une hypothèse neutre, parfois appelée hypothèse nulle, notée H_0. En règle générale il s'agit d'une hypothèse d'absence de changement. Nous poserons tout au long de ce travail que **l'hypothèse neutre correspond à une teneur en Nitrate constante tout au long du temps**. L'hypothèse neutre correspond donc au modèle M_0.

2. Formuler une hypothèse alternative, notée H_a. Nous avons fait le choix de poser que l'hypothèse **alternative est la présence d'un changement**, celui-ci étant à la diminution ou à l'augmentation. Lorsque l'hypothèse neutre sera rejetée en faveur de l'hypothèse alternative, nous associerons le signe de la variation au résultat du test. Ainsi par exemple,

    a. on peut tester l'hypothèse H_0 contre l'hypothèse H_1 qui considère que le modèle alternatif est le modèle M_1 : $f(t) = a + b(t)$. Dans ce cas, tester H_0 contre H_1 revient à tester H_0 : $b = 0$ contre H_1 : $b \neq 0$.

    b. Si l'hypothèse H_2 considère que le modèle alternatif est le modèle M_2, tester H_0 contre H_a revient à tester H_0 : $c = d = 0$ contre H_2 : $c \neq 0$ ou $d \neq 0$.

    c. Nous pourrions également construire un test statistique visant à tester H_1 contre H_2. La statistique de test utilisée dans ce cas, est un rapport de somme de carrés. Nous ne suivrons pas cette approche, car nous privilégions l'approche décrite à la section suivante. [1]

3. Choisir un **niveau de confiance,** noté $1\text{-}\alpha$. La valeur α est la probabilité de rejeter à tort H_0, alors que celle-ci est vraie. Le plus souvent on fait le choix α=0.05.

4. Se donner un modèle statistique pour la distribution des données, permettant ainsi de calculer sous H_0 la probabilité de rejet (la p-valeur).

    a. Dans le cas d'un test paramétrique, nous choisissons de faire une hypothèse Gaussienne **pour la distribution des résidus**, c'est-à-dire que nous faisons l'hypothèse classique que les ε_i sont des variables aléatoires Gaussiennes indépendantes et identiquement distribuées, d'espérance nulle et de variance

    $$\sigma^2 : \varepsilon_i \sim N(0, \sigma^2), \quad \text{indépendantes.}$$

    b. Dans le cas des tests non paramétriques, nous ne faisons aucune hypothèse sur les ε_i. Les calculs se font sur les rangs des valeurs.

Remarques

- Rejeter H_0 revient à considérer qu'il existe un changement jugé significatif, compte tenu des données observées ;
- Ne pas rejeter H_0 peut se produire soit parce qu'il n'y a effectivement pas de changement, soit parce que les données observées ne sont pas assez nombreuses pour établir que le changement est significatif.

[1] Pour le lecteur intéressé, la procédure pour réaliser un tel test est détaillée dans le rapport "The EU Water Framework Directive: Statistical aspects of the identification of groundwater pollution trends, and aggregation of monitoring results" (Grath et al., 2001, p. 58-60).

2.5 La sélection de modèle : une approche alternative aux tests usuels

Nous proposons d'utiliser également une seconde approche. La théorie statistique qui permet de confronter des modèles concurrents, non nécessairement emboités est la sélection de modèle. Il s'agit de pénaliser la log-vraisemblance par un terme qui dépend du nombre de paramètres et du nombre de données. La justification de cette approche résulte de l'observation suivante : il est toujours possible d'augmenter la vraisemblance en augmentant les paramètres d'un modèle. Cela se fait au prix d'une augmentation de la complexité du modèle, qui risque alors le sur-ajustement. On constate empiriquement qu'un modèle sur-ajusté sera moins performant en prédiction sur un jeu de donnés différent du jeu de données ayant servi à l'ajustement du modèle. Un modèle sur-ajusté sera également moins performant en situation d'extrapolation, par exemple pour prédire l'évolution dans les années à venir. On pénalise alors la complexité d'un modèle par un terme qui dépend à la fois du nombre de paramètres et du nombre de données. Le critère BIC (Schwarz, 1978) est défini par :

$$BIC = -2 \ln L + p \ln n,$$

où L est la vraisemblance calculée à son maximum, p est le nombre de paramètres et n est le nombre de données. Entre deux modèles dont les paramètres sont estimés par maximum de vraisemblance, on doit choisir le modèle avec le BIC le plus faible. Toutes autres choses étant égales par ailleurs, BIC augmente avec la variance du résidu et avec le nombre de paramètres. Ainsi, un BIC plus faible est le signe d'un meilleur ajustement ou d'un nombre de paramètres plus faible, ou les deux. Dans les deux cas, le BIC le plus faible doit être privilégié. Dans l'approche par sélection de modèle, on ne parle pas de significativité, mais de « force de la preuve » (« strength of evidence » en anglais). D'après Kass and Raftery (1995), le poids de la preuve en faveur du modèle avec le BIC le plus faible peut s'interpréter de la façon suivante

Δ BIC	Force de la preuve
0 à 2	Simple mention
2 à 6	Positif
6 à 10	Fort
> **10**	Très fort

Il est évidemment possible de comparer plus de deux modèles entre eux. Afin de pouvoir aisément distinguer la sélection de modèle des tests d'hypothèses, nous noterons M_0, M_1 et M_2 les trois modèles statistiques correspondant aux hypothèses H_0, H_1 et H_2.

1. **Il est important de souligner que la sélection de modèle par critère BIC nécessitant le calcul d'une vraisemblance, se fait nécessairement dans un cadre paramétrique.**

2. **La sélection de modèle nécessite que l'hypothèse de normalité des résidus soit vérifiée. Aussi, après ajustement d'un modèle, on teste que les résidus sont bien gaussiens. Si c'est le cas, le test est valide. Si ce n'est pas le cas, on réalise un test non-paramétrique, qui demande des calculs plus longs, et qui sont un peu moins puissants (ie, capables de détecter une tendance lorsqu'elle existe).**

2.6 Corrélation avec les hauteurs piézométriques

Sur certains ouvrages, la hauteur piézométrique est mesurée en même temps que la teneur en nitrate. Dans les formations crayeuses, notamment en Hesbaye, en pays de Herve et dans le pays de Mons, on observe des corrélations importantes entre teneur en nitrates et hauteur piézométrique, en particulier lorsque l'épaisseur de la zone non saturée (ZNS) est importante. Ce phénomène s'explique par un stockage de nitrate dans la ZNS, qui est remobilisé puis lessivé lorsque le niveau de la nappe remonte. A l'inverse, lorsque la nappe est basse, les nitrates restent stockés dans la ZNS. Lorsque c'est le cas, il peut s'avérer intéressant de filtrer l'effet de la hauteur d'eau pour analyser l'évolution tendancielle en nitrates. Nous suggérons de pratiquer les analyses par tests statistiques et par sélection de modèles après avoir retiré l'effet de la hauteur d'eau.

Estimation paramétrique

Dans le cadre paramétrique, nous construisons un modèle linéaire multivarié dans lequel intervient également la cote piézométrique. Ainsi, nous écrirons

$$Y_i = f(t_i) + \beta Z_i + \varepsilon\_i$$

où Z_i désigne la cote piézométrique et β est le coefficient de la régression de Y_i sur Z_i. Dans le cadre des modèles linéaires, nous pouvons tester la significativité du coefficient βou, ce qui est équivalent, de la significativité du coefficient de corrélation linéaire entre Y_i et Z_i. En utilisant la régression multivariée[2] de Y sur (t, Z), nous pouvons également calculer la valeur des pentes des modèles M_1 et M_2, ainsi que leur significativité en tenant compte de la présence de la variable Z.

Il faut souligner un point important : sur une série donnée, la cote piézométrique n'est pas connue pour chaque valeur mesurée de la teneur en nitrate : il existe des données manquantes. Le nombre de données permettant l'analyse statistique multivariée est inférieur au nombre de données utilisées pour une analyse (t_i, Y_i). De ce fait, les modèles estimés avec ou sans la cote piézométrique ne sont pas directement comparables par leur BIC ou par la somme des carrés des résidus, SSR.

Estimation non paramétrique

Les méthodes non-paramétriques utilisées ici ne permettent pas l'utilisation d'une covariable. Aussi, l'analyse par les méthodes non paramétriques impose de travailler sans l'utilisation des cotes piézométriques.

2.7 Détection de données isolées

Avant toute analyse, une étape de pre-processing est lancée afin de retirer de l'analyse certaines données isolées dans le temps. En effet, sur de nombreuses séries, on observe quelques mesures isolées, séparées des mesures plus récentes et plus régulières par une période de plusieurs années. Inclure ces données dans l'analyse aurait pu mener à des résultats statistiquement peu robustes. Il a donc été décidé de retirer de l'analyse les groupes ayant un nombre de mesures inférieur ou égale à 5 données, séparées d'au moins 4 années des autres mesures (ces valeurs peuvent être paramétrées et modifiées par l'utilisateur).

2.8 Détection des outliers

Après avoir retiré de l'analyse les données isolées temporellement, et avant de chercher à identifier des tendances et des ruptures sur les séries chronologiques, cette étape vise à identifier les outliers potentiels. En effet, ces points éloignés du reste des données peuvent avoir une influence importante sur les pentes estimées. La présence d'outliers entraîne également des statistiques de test en général très élevée, menant au rejet systématique de l'hypothèse nulle. Nous éliminerons donc des données qui s'écartent trop d'une évolution « normale ».

[2] Cette analyse se fait en appelant la fonction **lm** dans **R**.

Sans cotes piézométriques

La procédure utilisée consistera à réaliser une régression non paramétrique robuste de Y_i contre t_i à l'aide la procédure LOESS dans R. Celle-ci, expliquée en détail dans Grath *et al*. (2001, p. 55-56) produit une estimation de la variance du bruit autour de la fonction de régression. On retirera des données servant à l'estimation et aux tests de tendance toutes les données se situant localement en dehors d'un intervalle de confiance à 99.5%, c'est-à-dire toutes les données Y_i telles que

$$Y_i \notin [\hat{Y}_i(t_i) - q\sigma_i; \hat{Y}_i(t_i) + q\sigma_i]$$

où q est le quantile associé à l'intervalle de confiance à 99.5%. L'intervalle à 99.5% a été choisi car il correspond au seuil conventionnel pour l'identification des outliers dans la représentation des boxplots. Il constitue un bon compromis entre taux de variations admissibles et taux d'élimination.

Avec cotes piézométriques

Lorsque les cotes piézométriques sont mesurées, la détection des outliers est modifiée de la façon suivante. La régression non paramétrique du filtre LOESS est bivariée, portant à la fois sur t_i et Z_i. On retirera les données se situant localement en dehors de l'intervalle de confiance issu de la prédiction bivariée.

Trois points importants sont à signaler :

1. La procédure ci-dessus ne peut pas fonctionner si la cote piézométrique est manquante. Dans ce cas, on réalise d'abord une prédiction de la cote piézométrique en fonction du temps, puis on utilise cette cote prédite pour calculer les enveloppes supérieures et inférieures de la valeur Y_i. Une donnée située en dehors de l'intervalle défini par ces enveloppes est considérée comme un outlier.
2. La représentation graphique (t_i, Y_i) ignore la valeur de la cote. De ce fait, les valeurs prédites de Y_i présentent des fluctuations importantes, liées aux variations de cote. Afin de produire un rendu visuel comparable à celui obtenu en absence de données piézométriques, on superpose une courbe dans laquelle les cotes piézométriques sont toutes remplacées par la moyenne des cotes mesurées. Ces valeurs ne sont utilisées que pour la représentation graphique et nullement pour le calcul des statistiques. Un point non considéré comme outlier (en bleu sur les figures) pourrait donc se trouver en dehors de l'enveloppe.
3. Les outliers statistiques seront intégrés dans les analyses non paramétriques. En effet, celles-ci sont par construction robustes à la présence d'outliers. Leur exclusion ne se justifie donc plus.

3. WORKFLOW DE L'ANALYSE PAR SERIE

Nous appliquons le workflow suivant à chaque série temporelle (Figure 1) :

1. **Identification des données isolées temporellement**

   Les données isolées temporellement sont écartées de l'analyse.

2. **Test sur le nombre de données disponibles**

   Si (Nombre de données disponibles <= 10), on arrête le workflow.

3. **Détection des outliers statistiques**

   Les données considérées comme outliers *statistiques* sont identifiées pour les représentations graphiques. Elles ne participent pas à l'analyse statistique paramétrique ultérieure. Cette détection peut se faire avec ou sans l'utilisation de la cote piézométrique.

4. **Analyse paramétrique**

   L'analyse paramétrique peut se faire avec ou sans l'utilisation de la cote piézométrique.

   a. On calcule les différentes régressions pour les modèles M_0, M_1 et M_2
   b. On calcule les p-valeurs de H_0 contre H_1 et H_2 et la p-valeur de H_1 contre H_2.
   c. On calcule les valeurs BIC pour les modèles M_0, M_1 et M_2
   d. On sélectionne un modèle selon les critères exposés plus bas.
   e. On représente le modèle sélectionné par l'analyse

5. **Test de normalité des résidus**

   Si (résidus Gaussiens), on arrête le workflow.

6. **Si (résidus non Gaussiens)**

   Régression et tests non paramétriques

   a. On réintègre les outliers statistiques (mais pas les données isolées temporellement)
   b. On calcule les différentes régressions non paramétriques pour les modèles M_1 et M_2
   c. On calcule les p-valeurs de H_0 contre H_1 ; pour tester le modèle M_2, on calcule les p-valeurs des deux pentes à l'aide du test de Mann-Kendall.
   d. On sélectionne un modèle sur base des critères exposés ci-dessous et on le représente graphiquement.

Pour la sélection de modèles, les règles suivantes sont utilisées :

1. **Analyse paramétrique :**
   a. Le modèle sélectionné est celui pour lequel le critère BIC est minimum.
   b. Si un modèle M_2 est sélectionné et que la différence sur le critère BIC entre les modèles M_1 et M_2 est inférieure à 2, ou que les deux pentes sont non significatives, le modèle le plus simple, M_1, est conservé.
   c. Si un modèle M_1 est sélectionné et que la pente est non significative, on retombe sur un modèle M_0.
2. **Analyse non paramétrique :**
   a. Le modèle sélectionné est celui pour lequel la Somme des Carrés des Résidus (SSR) est minimum.
   b. Si un modèle M_2 est sélectionné et que les deux pentes sont non significatives, le modèle le plus simple, M_1, est conservé.
   c. Si un modèle M_1 est sélectionné et que la pente est non significative, on retombe sur un modèle M_0.

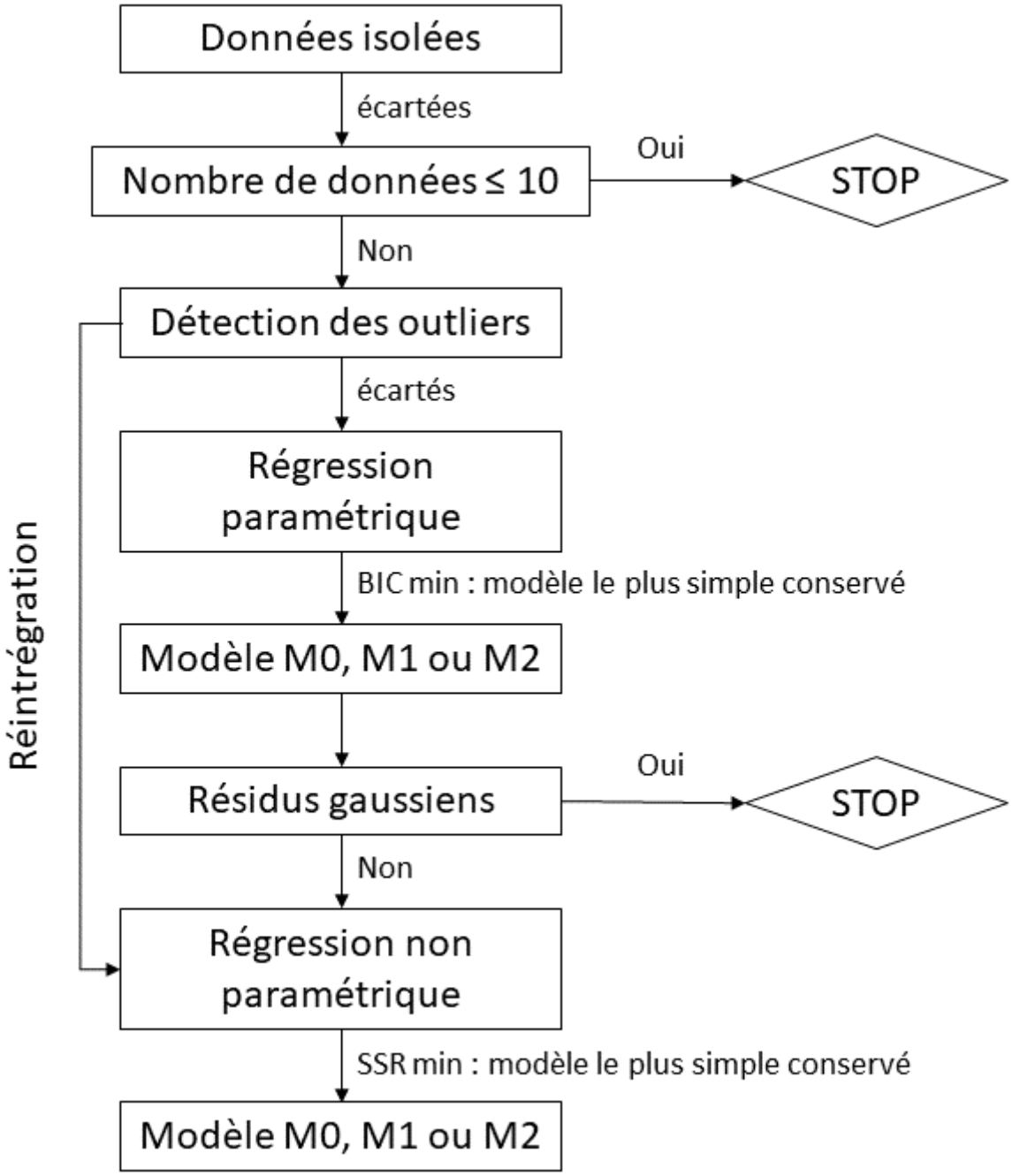


Figure 1 : Logigramme du workflow d'analyse des séries chronologiques.

4. AGREGATION DES RESULTATS PAR MASSE D'EAU SOUTERRAINE

Une analyse agrégée est réalisée sur chaque masse d'eau après exclusion des données isolées et des outliers de chaque série.

Cette analyse fait une hypothèse implicite d'indépendance entre les différentes séries chronologiques au sein d'une masse d'eau. La DCE se place dans le cadre de cette hypothèse ainsi que les différents rapports du BRGM établissant la méthodologie utilisée en France. Il en va de même dans de nombreux autres pays Européens. Nous nous plaçons également dans ce cadre. En statistiques spatiales, il est bien connu que les valeurs estimées sont relativement peu affectées par la prise en compte ou non des dépendances spatiales. En revanche, les variances d'estimation et toutes les grandeurs dérivées de ces variances (p-valeurs, tests de significativité, etc.) sont très fortement affectées. C'est pourquoi, conscients que les dépendances spatiales ne sont pas prises en compte, nous ne calculons pas les p-valeurs associées au calcul des pentes agrégées. Tenir compte de ces dépendances nécessiterait des développements nouveaux en statistiques pour les données spatio-temporelles qui dépassent de très loin le cadre du présent travail.

4.1 Analyse paramétrique

Les modèles M_0, M_1 et M_2 sont ajustés sur l'ensemble des données d'une masse d'eau donnée. Comparée avec l'analyse réalisée pour une série particulière, cette analyse présente les différences suivantes :

- Les informations piézométriques sur les différents ouvrages (avec des données manquantes pour des dates différentes d'un ouvrage à l'autre), trop compliquées à prendre en compte dans ce cadre, ne sont pas prises en compte ici.
- Chaque série étant centrée autour de courbes différentes, les résidus mélangés ne peuvent pas suivre une distribution Gaussienne, mais au mieux une distribution qui serait un mélange de distributions Gaussiennes. Pour cette raison, on ne peut pas réaliser de test d'ajustement Gaussien sur les résidus de l'ajustement global.
- Pour cette raison, mais aussi à cause des possibles corrélations spatiales, on ne pourra utiliser ni les p-valeurs calculées sur les pentes ni le critère BIC pour sélectionner les modèles.
- La sélection du modèle se fait donc uniquement par le critère de écarts quadratiques, également appelé SSR (Sums of Squared Residuals). En cas d'égalité, le modèle le plus simple est retenu.

Un graphique représentant l'ensemble des séries et le modèle agrégé sélectionné est réalisé.

4.2 Analyse non paramétrique

Un ajustement non paramétrique est réalisé en appelant la fonction **loess.** Cette fonction calcule une régression polynomiale locale. Bien que localement paramétrique, le fait que la courbe finale soit le résultat d'une multitude de régressions locales transforme cette courbe finale en une fonction non paramétrique. Elle est par ailleurs réputée assez robuste. On obtient une pente (non paramétrique) actuelle à partir des positions de la courbe ajustée durant les quatre dernières années à la date de l'analyse, c'est-à-dire pour le présent rapport durant les années 2010-2013. Un graphique représentant l'ensemble des séries, puis l'ajustement **loess,** est réalisé.

4.3 Représentation agrégée des pentes

Le cross-plot entre les pentes paramétriques et non paramétriques calculé sur chacune des séries est représenté. Dans le cas où l'analyse non paramétrique n'est pas réalisée (car les résidus sont considérés Gaussiens), on la pose égale à l'analyse paramétrique. Les non-valeurs (NA dues à des séries trop courtes) sont exclues du graphique. Sur ce graphique, les cercles sont fonctions de la longueur de la série. Un code couleur indique si la teneur en Nitrate prédite par le modèle est inférieure ou supérieure au seuil réglementaire de 50 mg/l en 2014. Les pentes agrégées sont également représentées.

Ces graphiques permettent de vérifier la concordance entre les ajustements paramétriques et non paramétriques, de visualiser la dispersion des pentes et de voir où se place le modèle agrégé par rapport aux séries individuelles. Grâce à l'utilisation d'un code couleur, il permet également de repérer immédiatement si les ouvrages ayant des mesures supérieures au seuil réglementaire ont une pente en 2014 négative ou positive.

5. RESULTATS PRODUITS

La Section 5.1 présente les résultats produits pour chaque série analysée. Les résultats obtenus après l'agrégation p sur l'ensemble des séries, par masse d'eau ou par zone vulnérable sont décrits à la Section **Erreur ! Source du renvoi introuvable.**. La Section 5.3 détaille les différentes classifications u tilisées pour interpréter les résultats.

5.1 Résultats par série

Les figures suivantes sont produites et enregistrées dans le répertoire *\Nitrates\Programmes\R\figures\Masses\IDmasse* où *IDmasse* est l'identifiant de la masse d'eau. La nomenclature est la suivante :

- *Hist_Residus_Serie_IDserie.png* montre l'histogramme des résidus.
- *Modele_NP_Selectionne_Serie_IDserie.png* montre le modèle non paramétrique ajusté. Si un tel fichier existe pour une série, c'est lui qui contient le meilleur modèle. Dans le cas contraire, le modèle sélectionné est paramétrique.
- *Modele_Selectionne_Serie_IDserie.png* montre le modèle paramétrique sélectionné. Il est le meilleur modèle ajusté dans le cas où il n'existe pas de fichier *Modele_NP_Selectionne_Serie_IDserie.png* pour la même série.
- *Modeles_NP_Serie_IDserie.png* montre les 3 modèles non paramétriques (M_0, M_1 et M_2). Ce fichier n'existe que si une analyse non paramétrique a été réalisée.
- *Modeles_Serie_IDserie.png* montre les 3 modèles paramétriques (M_0, M_1 et M_2).
- *Outliers_q_5_Serie_IDserie.png* montre la série avec identification des outliers.

Les résultats de la procédure de sélection de modèle sont résumés dans des tableaux donnés dans les fichiers *SPW_NO3_Resultats_Masse_IDmasse.xlsx* où *IDmasse* est l'identifiant de la masse d'eau. La première feuille reprend tous les résultats bruts. Lorsqu'un modèle non paramétrique est sélectionné, le modèle paramétrique est également donné. Dans ce cas, seul le modèle non paramétrique est utilisé pour le calcul des statistiques et les cartographies, tel que disponible dans la feuille "Resume", qui contient les résultats remis en forme.

La colonne "intercept date min" (resp. "intercept date max") donne la valeur de la concentration en nitrates estimée par le modèle au début (resp. à la fin) de la période de mesures. Les colonnes "estim 2013", "estim 2015", "estim 2021" et "estim 2027" donnent respectivement les estimations au 31 décembre 2013, 2015, 2021 et 2027. Les colonnes "proba depassement 2015", "proba depassement 2021" et "proba depassement 2027" donnent respectivement la probabilité de dépasser la norme de 50 mg/l au 31 décembre 2015, 2021 et 2027. Les dates de prédiction peuvent être définies par l'utilisateur dans le fichier de paramètres.

Les colonnes avec les p-valeurs donnent les résultats des tests statistiques.

> La ***p-valeur*** est définie comme la probabilité d'obtenir la même valeur pour la statistique de test ou une valeur encore plus extrême si l'hypothèse nulle était vraie. Plus cette valeur est faible, moins vraisemblable est l'hypothèse nulle.

5.2 Agrégation des résultats

Le Tableau 1 présente la fiche de synthèse établie pour l'ensemble des séries disponibles ainsi que pour chaque masse d'eau ou chaque zone vulnérable. Il résume le nombre de séries chronologiques affectées à chaque catégorie. On peut y relever les observations suivantes :

- Le nombre de séries chronologiques n'ayant pu être analysées faute d'un nombre suffisant de données.
- Le nombre de séries chronologiques pour lesquelles aucune tendance n'a été mise en évidence.
- La répartition des séries chronologiques analysées selon le type d'évolution
- La répartition des séries chronologiques par type de modèle (M_0, M_1 ou M_2 ; paramétrique ou non paramétrique)
- Le nombre de séries chronologiques pour lesquelles la concentration en nitrates estimée par le modèle à la fin de la période de mesure est inférieure à la norme des 50 mg/l/an.

Tableau 1 : Synthèse des modèles ajustés sur les séries chronologiques d'eau souterraine du réseau Nitrates sur l'ensemble de la Région Wallonne.

	Nombre
Séries disponibles	838
Séries ayant fait l'objet de l'analyse	802
Tendance non-significative	382
Tendance significative	420

Tendance	Nombre
Situation et évolution favorables	603
Situation favorable et évolution légèrement défavorable	58
Situation favorable mais évolution défavorable	53
Situation favorable mais évolution très défavorable	24
Situation défavorable mais évolution favorable	10
Situation et évolution défavorables	54

Type de modèle	Nombre		Nombre	Concentration	Nombre
M0	247	P	442	< 50 mg/l/an	738
M1	84	NP	360	> 50 mg/l/an	64
M2	471				

Pour chaque masse d'eau ou zone vulnérables résultats suivants sont également produits :

- Une carte selon la classification en codes couleur (**Figure *2***, voir Section 5.3 pour la construction de la légende).
- Un graphique avec l'ensemble des séries de la masse d'eau ainsi que l'ajustement loess et le modèle agrégé sélectionné (**Figure *3***).
- Un cross-plot entre les pentes paramétriques et non paramétriques (exemple à la Figure 4). Le coefficient de corrélation entre les deux types de pentes indique que l'intensité de la relation entre modèles paramétriques et non paramétriques. Comme attendu, pour les masses d'eau comptant peu de séries chronologiques, il se peut que cette relation soit moins forte.

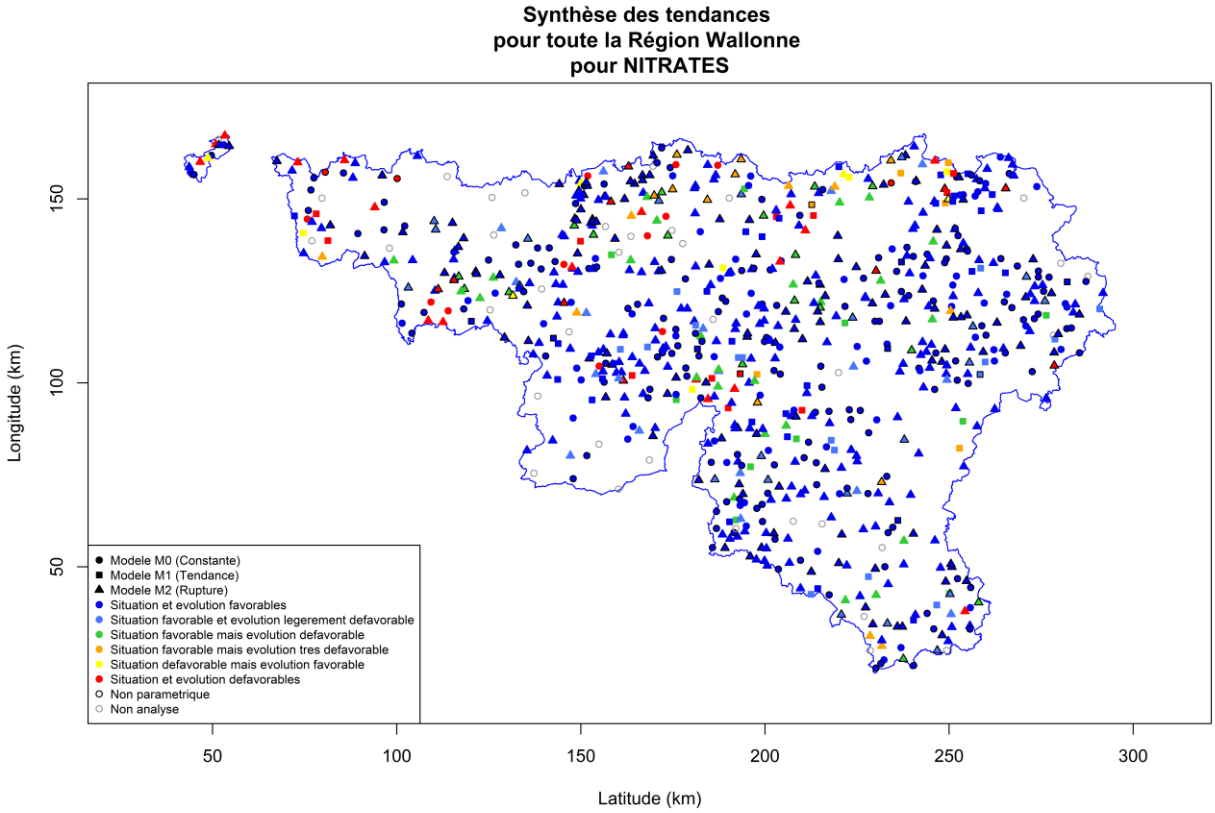


Figure 2 : Synthèse des tendances à l'échelle de la Région Wallonne pour le réseau Nitrates.

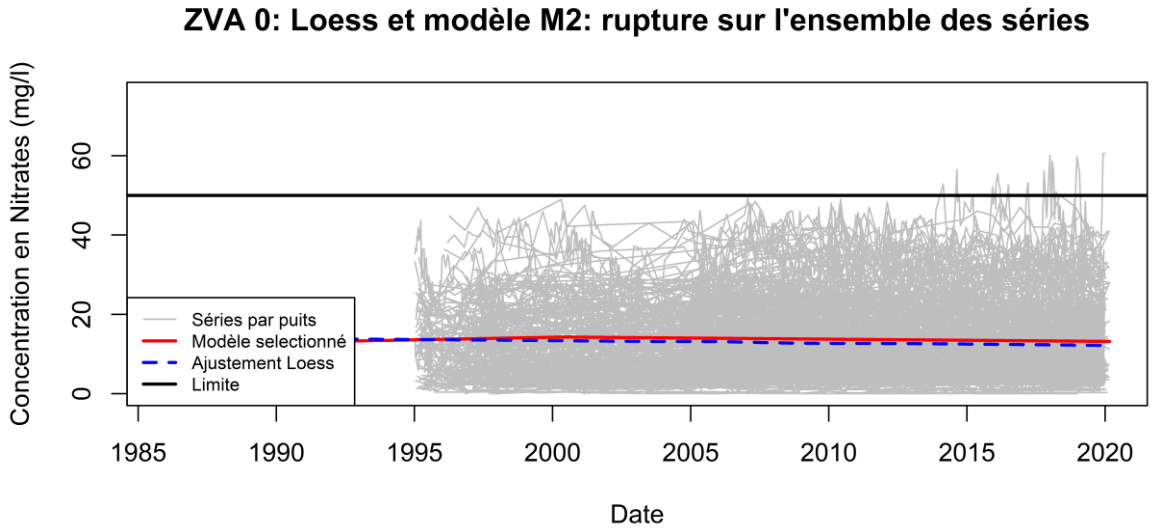


Figure 3 : Pentes agrégées pour la Zone Vulnérable 0 pour le réseau Nitrates.

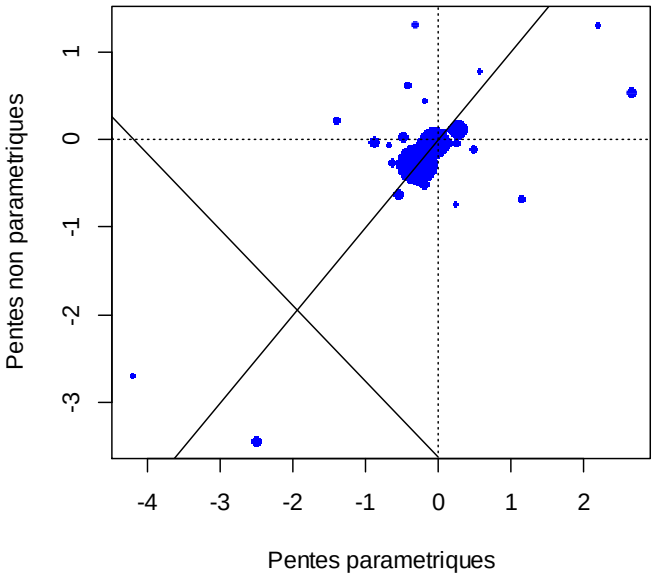


Figure 4 : Diagramme de dispersion entre les pentes paramétriques et non paramétriques agrégées. Chaque point représente une masse d'eau. La taille des points est proportionnelle au nombre de séries chronologiques analysées pour la masse d'eau en question.

5.3 Classifications des résultats

Des classifications différentes sont utilisées pour les eaux souterraines et les eaux de surface d'un côté, et pour les séries issues de la modélisation de l'autre côté.

Classification pour les eaux souterraines et les eaux de surface

Pour représenter les différents cas de figures, une classification synthétique a été mise au point. Elle vise à informer simultanément sur la tendance (à la diminution ou à l'augmentation), sur le dépassement éventuel de la norme des 50 mg/l et sur le type de modèle (M_0, M_1 ou M_2). La classification proposée comprend 35 classes. Les valeurs des limites sur les concentrations varient selon que l'on s'intéresse aux eaux souterraines (Figure 5) ou aux eaux de surface (Figure 6). Le nombre de classes et le choix des couleurs peut être modifié par l'utilisateur via le fichier de paramètres. Les figures montrées ici ne sont donc que des exemples.

Les données fournies pour les eaux de surface sont exprimées en mg d'azote par litre (mg N/l) et pas en mg de nitrate par litre (mg NO3/l). Pour éviter d'avoir des tableaux de classification pour les eaux de surface et les eaux souterraines qui n'emploient pas les mêmes unités, les limites de classe pour le tableau « eaux de surface » ont été fixées en mg NO3/l. En conséquence, pour pouvoir utiliser la classification proposée, il faut multiplier les valeurs des données exprimées en mg N/l par un facteur de conversion égal à 4.425.

Pour la cartographie, afin d'augmenter la lisibilité, ces 35 classes sont regroupées en 6 couleurs. Le Tableau 2 et le Tableau 3 donnent les légendes des symboles utilisés pour les séries relatives aux eaux souterraines et aux eaux de surface respectivement.

Dans la légende, le terme "Situation" fait référence à la position de la concentration en nitrates par rapport à la concentration en nitrates. La situation est dite favorable si la concentration est inférieure à la norme et défavorable si la concentration est supérieure à la norme. Le terme "Evolution" décrit, lui, la tendance. L'évolution est dite défavorable si la dernière pente du modèle est positive. Elle est dite favorable si cette pente est négative.

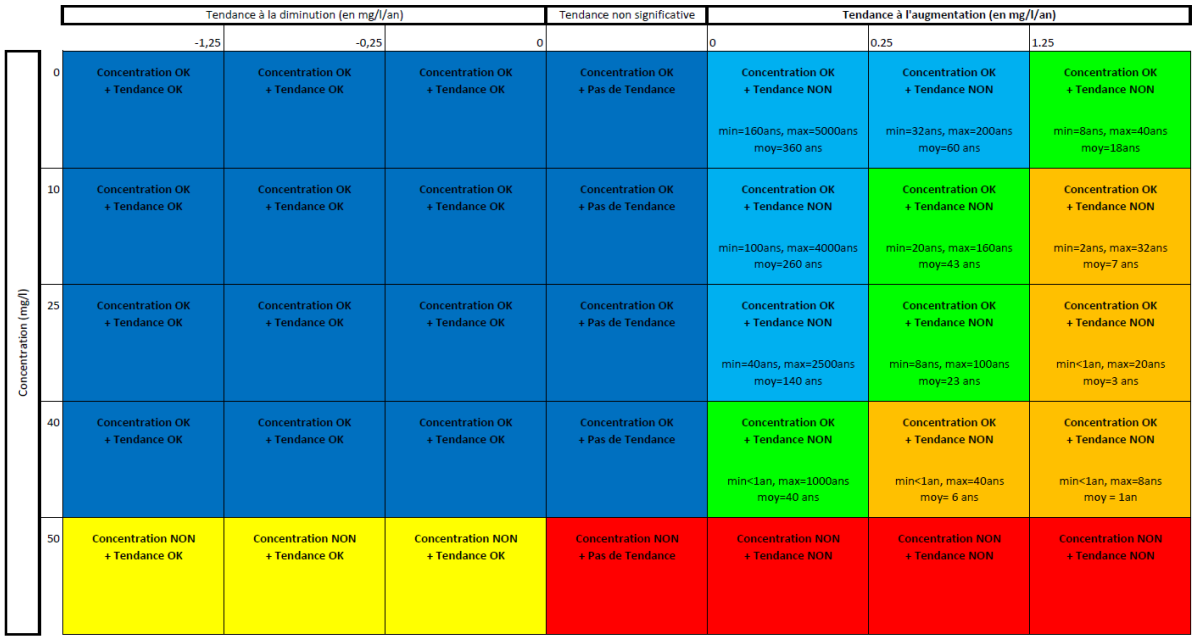


Figure 5 : Classification pour les séries chronologiques d'eau souterraines. Les couleurs sont utilisées pour faciliter la lecture des cartes.

Tableau 2 : Classification des résultats de la sélection de modèles pour les eaux souterraines.

Niveau (code graphique)	Symbole	Qualificatif pour l'évolution
1	●	Situation et évolution favorables
2	●	Situation favorable et évolution légèrement défavorable
3	●	Situation favorable mais évolution défavorable
4	●	Situation favorable mais évolution très défavorable
5	●	Situation défavorable mais évolution favorable
6	●	Situation et évolution défavorables
Type d'analyse	○	Analyse non paramétrique ○ Série non analysée par manque de données
Type de modèle	●	M0 ■ M1 ▲ M2

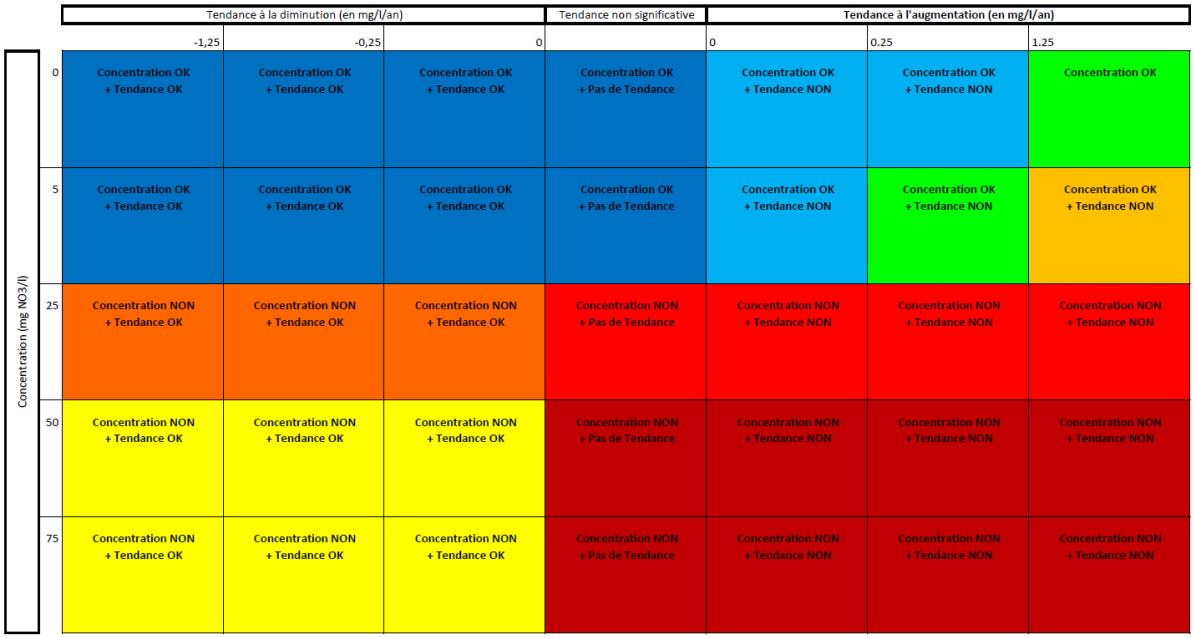


Figure 6 : Classification pour les séries chronologiques d'eau de surface. Les couleurs sont utilisées pour faciliter la lecture des cartes.

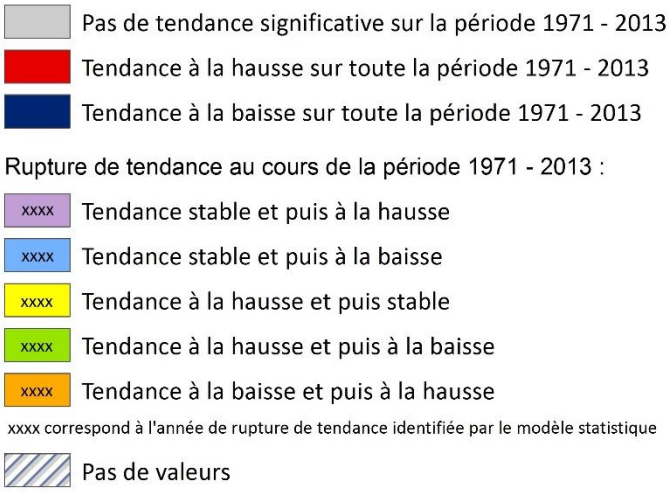
Tableau 3 : Classification des résultats de la sélection de modèles pour les eaux de surface.

Niveau (code graphique)	Symbole	Qualificatif pour l'évolution
1	●	Situation et évolution favorables
2	●	Situation favorable et évolution légèrement défavorable
3	●	Situation favorable mais évolution défavorable
4	●	Situation favorable mais évolution très défavorable
5	●	Situation défavorable mais évolution favorable (par rapport à la directive 91/676/CE)
6	●	Situation défavorable mais évolution favorable (par rapport à la directive 2000/60/CE)
7	●	Situation et évolution défavorables (par rapport à la directive 2000/60/CE)
8	●	Situation et évolution défavorables (par rapport à la directive 91/676/CE)
Type d'analyse	○	Analyse non paramétrique ○ Série non analysée par manque de données
Type de modèle	●	M0 ■ M1 ▲ M2

Classification pour les séries issues de la modélisation

Pour les séries issues de la modélisation, la classification est basée uniquement sur l'évolution. La légende est donnée au Tableau 4.

Tableau 4 : Classification des résultats de la sélection de modèles pour les séries chronologiques issues de la modélisation.



6. REFERENCES BIBLIOGRAPHIQUES

D'Or D. and Allard. D. (2014). Mise en évidence de tendances éventuelles sur les séries chronologiques présentées par les points du réseau de mesure wallon des eaux de surface et souterraines en ce qui concerne les nitrates. Rapport final. Rapport Ephesia Consult RP DGO3 2014002 - Novembre 2014.

Grath J., Scheidleder A., Uhlig S., Weber K., Kralik M., Keimel T. and Gruber D. (2001). "The EU Water Framework Directive: Statistical aspects of the identification of groundwater pollution trends, and aggregation of monitoring results". Final Report. Austrian Federal Ministry of Agriculture and Forestry, Environment and Water Management (Ref.: 41.046/01-IV1/00 and GZ 16 2500/2-I/6/00), European Commission (Grant Agreement Ref.: Subv 99/130794), in kind contributions by project partners. Vienna.

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

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90 (430): 773–795. doi:10.2307/2291091

Lopez B., Croiset N., Surdyk N. and 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.

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).

Schwarz, G. E. (1978). "Estimating the dimension of a model". *Annals of Statistics* 6 (2): 461–464. doi:10.1214/aos/1176344136.

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

Sohier C. (2011). Développement d'un modèle hydrologique sol et zone vadose afin d'évaluer l'impact des pollutions diffuses et des mesures d'atténuation sur la qualité des eaux en Région wallonne (thèse de doctorat). Université de Liège – Gembloux Agro-Bio Tech, 338 p., 30 tabl. 146 fig.

Williams J.R., Jones C.A., Dyke P.T. (1984). A modelling approach to determining the relationship between erosion and soil productivity. Transactions of the ASAE. 27, 129-144.