Notice

Clément LAROCHE : clement-laroche@hotmail.fr

Dernière mise à jour : 29/03/2022

On présente dans cet onglet la notice d'utilisation de l'application. On rappelle que cette application a pour objectif d'aider l'utilisateur à localiser des signaux anormaux de concentration dans le temps et dans le territoire. Cet outil *n'a pas vocation ou prétention* d'expliquer la présence de phénomènes anormaux. Cela relève de l'expertise de l'utilisateur.

Onglet présentation

Le premier onglet donne des informations générales sur les données qui ont été chargées dans l'application.

Encadré 1 : Informations générales

On introduit la définition suivante :

Définition 1 (maximum journalier) : En se fixant une date d, on définit un maximum journalier comme le maximum des relevés de concentrations ayant eu lieu dans la région d'étude le jour d. L'information de quantification du maximum est également conservée :

- si le maximum journalier correspond à une valeur de LOQ alors il est dénoté comme non quantifié.
- si le maximum journalier ne correspond pas à une valeur de LOQ alors il est dénoté comme quantifié.

Un fois cette définition posée, on présente les informations générales sur l'étude que l'on s'apprête à mener en utilisant l'application :

- on donne le nom de la susbtance sur laquelle on travaille (on ne peut charger qu'une substance à la fois), correspond à l'entrée **Paramètre** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne la période temporelle d'étude, correspond aux entrées **Date de début** et **Date de fin** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne le nom de la région d'étude, correspond à l'entrée **Zone administrative** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne le nombre de prélèvements effectués dans cette zone et durant cette période, cela correspond au nombre de lignes du tableau **Analyses.csv** issues de la requête sur http://www.naiades.eaufrance.fr/
- on donne le pourcentage de prélèvements des relevés, on utilise la colonne **MnemoRqAna** du tableau **Analyses.csv** dont on compte le nombre d'occurences telle que la description indique que le résultat est au dessus du seuil de quantification
- on donne le nombre de jours au cours desquels un ou plusieurs prélèvements ont été effectués durant la période d'observation et dans la zone d'étude. Cela donne accessoirement le nombre de maximum journaliers que l'on peut calculer dans les données chargées dans l'application.

- on donne le pourcentage de quantification des maximums journaliers (cf. **Définition 1**) obtenus à partir des données.
- on affiche le nombre de stations actives (qui ont effectuées au moins un prélèvement) durant la période de temps définie et dans la zone d'étude choisie.

Encadré 2

Tous les maximums journaliers de la période d'étude et de la zone administrative choisie sont tracés en fonction du temps. Les points sont colorés selon leur statut de quantification (quantifié ou non quantifié). Les points rouge sont non quantifiés, les points bleu sont les maximums de concentrations quantifiés.

Encadré 3

Définition 2 (composante hydrographique) : on appelle ici composante hydrographique un réseau d'eaux de surface tel que pour n'importe quel couple de point (x, y) dans ce réseau, il existe un chemin d'eau reliant x et y. Le sens d'écoulement du courant n'est pas considéré.

On introduit la géographie de cette étude. Sont tracés sur une carte :

- toutes les stations ayant effectué au moins un prélèvement durant la période d'étude.
- en noir, les contours de la région administrative choisie pour cette étude.
- en bleu, les principaux cours d'eaux de la région (ceux qui disposent d'un code cours d'eau dans la base BD TOPO IGN)
- les stations sont coloriées selon la composante hydrographique auxquelles elles appartiennent (cf **Défintion 2**).

Toutes les coordonnées géographiques utilisées sont dans le Système de Coordonnées de Références **WGS84**.

Onglet Détection

Cet onglet constitue le principal outil qui permet de mettre en lumière des signaux anormaux présents dans les données chargées. Cette page comporte 4 encadrés. Les encadrés 1 et 2 concernent la partie détection temporelle. Les encadrés 3 et 4 sont portés sur le clustering spatial et la détection de clusters anormaux. Les choix effectués dans l'encadré 1 déterminent les informations qui seront présentes dans les encadrés suivants. L'encadré 2 est complémentaire du premier et présente des informations supplémentaires qui varient selon les choix effectués en encadré 1. Les choix de l'encadré 3 détermineront également les information dans l'encadré 4.

Encadré 1 : Détection temporelle

Cet encadré comporte 4 éléments, on commence par les deux plus intuitifs :

- un **graphique des maximums journaliers** portant le titre **Résultat de la segmentation**, les résultats de détection de ruptures temporelles sont appliqués sur la série des maximums journaliers. Les segments temporels sont représentés en blanc. Il est possible de sélectionnner un segment temporel, si c'est le cas il est surligné en noir. Choisir un segment change l'intégralité des informations de la section **Encadré 2 : Informations complémentaires sur le segment sélectionné**, de la section **Encadré 3 : Détection spatiale** et l'intégralité des informations de la section **Encadré 4 : Informations complémentaires sur la détection spatiale**
- une **tick-box échelle-log**, la représentation des maximums journaliers étant très écrasée vers les seuils de LOQ mais comportant tout de même des valeurs élevées, l'échelle logarithmique permet d'obtenir une représentation mieux répartie sur l'axe des concentrations. Ce bouton change le **graphique**

des maximums journaliers ainsi que les **violins plots saisonniers** de la section **Encadré 2 : Informations complémentaires sur le segment sélectionné**

Deux autres éléments permettent le réglage de la modélisation :

- **un curseur de valeur** permettant de régler la valeur de la pénalité. Ce curseur a une influence direct sur le **graphique des maximums journaliers**. Intuitivement, il correspond au poids que l'on associe au fait de poser une rupture dans ce graphique. Donc plus le poids est léger (plus la valeur de pénalité est faible), plus on va poser de ruptures dans le graphique. A l'inverse, plus le poids est lourd (plus la valeur de pénalité est élevée), moins on va poser de ruptures dans le graphique.

- **un graphique de coude** en dessous du **curseur de valeur**. C'est un nuage de point représentant le coût total d'une segmentation (l'ensemble des ruptures sur le graphique des maximums journaliers) en fonction du nombre de rupture. Le point rouge de ce graphique vous indique le coût de la segmentation que vous êtes en train d'explorer sur le graphique des maximums journaliers. En bougeant la valeur du **curseur**, d'autres segmentations seront explorées. Dans ce cas, le point rouge vous indiquera le coût de la segmentation que vous explorerez. On considère que la segmentation optimale (donc le nombre optimal de ruptures dans le signal) correspond au "coude" du graphique. Ce coude est indiqué ici par la droite verticale noire. Le **curseur de valeurs** commence directement à cette position (donc la segmentation optimale est affichée par défaut).

Détail : Pourquoi avoir laissé d'autres segmentations que la segmentation optimale (correspondant donc au coude) ? La démarche de cette application étant exploratoire, des segmentations autres que celle définie comme optimale permettront d'explorer des segments qui n'étaient pas présents dans la segmentation optimale et qui peuvent porter de l'information que les experts seront en mesure d'analyser.

Encadré 2 : Informations complémentaires sur le segment sélectionné

Cet encadré est entièrement dépendant de la section **Encadré 1** et notamment du segment sélectionné en noir dans le graphique des résultats de segmentation. Il est composé de trois élements que l'on peut décrire comme suit :

- le premier élement est un **texte informatif sur le segment sélectionné**. Sont affichées les informations suivantes :
  - les dates délimitant le segment sélectionné (surligné en noir) en **Encadré 1**
  - le nombre de maximum journaliers dans ce segment
  - le pourcentage de quantification de maximum journalier
  - le nombre de stations qui ont été actives entre les deux dates délimitant le segment
  - la valeur minimum des maximums journaliers dans ce segment
  - la valeur moyenne des maximums journaliers dans ce segment
  - la valeur médiane des maximums journaliers dans ce segment
  - la valeur maximum des maximums journaliers dans ce segment

- **un graphique de fonctions de répartition**, ce graphique permet de juger la qualité de l'estimation du modèle sous-jacent. En bleu, on trouve la fonction de répartition empirique (donc celles des données observées) que l'on compare avec la courbe rouge qui correspond à la courbe théorique obtenue par la modélisation. Plus ces deux courbes sont proches, plus cela signifique l'ajustement du modèle choisi est précis. Les droites noires verticales correspondent aux valeurs de LOQ présentes dans le segment sélectionné.

- **un boxplot de saisonnalité**, on trace ici le violin plot (objet dont la lecture est similaire à un box plot) des maximums journaliers de concentrations du segment selectionné dans l'**Encadré 1** en rouge. On prend la période temporelle sur laquelle s'étend le segment et on trace les violins plots des mêmes périodes pour les années précédentes et suivantes. Par exemple, si le segment sélectionné s'étend du "01-03-2017" au "24-04-2017", on trace les violin plots des "01-03-20XX" au "24-04-20XX" avec "20XX" étant les autres années disponibles dans le jeu de donnée. Si l'étendue du segment sélectionné dépasse une année, ce tracé n'est pas possible. On trace alors uniquement le violin plot du segment sélectionné.

Encadré 3 : Détection spatiale

Cet encadré est entièrement dépendant de la section **Encadré 1** et notamment du segment sélectionné en noir dans le graphique des résultats de segmentation. Il est composé de 3 éléments :

- **un menu déroulant** contenant 4 options. Ces quatre options définissent les informations disposées sur la **carte interactive leaflet**. Les quatres propositions du menu sont les suivantes :
  - **Stations actives :** indiquent les stations qui ont effectuées au moins un prélèvement pendant la période du segment sélectionné dans la section **Encadré 1** (surligné en noir). C'est une variable binaire qui indique en *TRUE* que la station a bien effectué au moins un relevé, en *FALSE* que la station n'a pas effectué de relevé.
  - **Composantes hydrographiques :** cette carte reprend la carte affiché dans l'onglet **Présentation**. Cela permet de retrouver les composantes hydrographiques connectées (cf **Définition 2**) dans la région sans avoir à changer d'onglet. Elle est tout de même dépendante du choix de segment en section **Encadré 1** car seule les stations actives durant la période de temps sont colorées selon leur appartenance à une composante hydrographique. Les stations inactives sont affichées par des points noirs plus petits.
  - **Clusters spatiaux** : cette option affiche les résultats de clustering spatial sur les stations. On regroupe les stations selon leur distance dans le réseau hydrographique. Chaque cluster est associé à une couleur. Seules les stations actives sont colorées selon leur appartenance à un cluster. Plusieurs choix de clustering sont disponibles en utilisant le **curseur de nombre de clusters** au dessus du **menu déroulant**. Les stations inactives sont affichées par des points noirs plus petits.
  - **Valeurs du front de Pareto :** cette option correspond à la représentation spatial des résultats de détection d'anomalies. Plus une valeur de front de Pareto est faible plus le cluster portant cette valeur est anormal (au sens de deux critères explicités en section **Encadré 4**). Une valeur de front de Pareto est affectée à chaque cluster, les résultats sont donc dépendants de la valeur prise par **le curseur de nombre de clusters**.
- **Le curseur de nombre de cluster :** permet de choisir entre plusieurs clusterings (qui diffèrent donc selon leur nombre de clusters). Il est important de noter que plus le nombre de clusters est élevé plus leur résolution géographique est fine (plus leur étendue géographique est faible). Cependant, lorsqu'un cluster est petit, le nombre de station le composant est faible, il y a donc moins de données disponibles pour ce cluster. Ce paramètre de nombre de cluster peut être vu comme un compromis entre la résolution géographique et le nombre de données disponibles pour l'analyse de chaque cluster.
- **La carte interactive leaflet :** présente les informations choisie dans le **menu déroulant**. Elle est dépendant de la valeur prise par **le curseur du nombre de cluster**. L'utilisateur peut cliquer sur les stations affichées sur la carte. Cela change les informations présentées dans la section **Encadré 4**. Lorsque l'on clique sur une station, toutes les stations appartenant au même cluster sont surlignées en rouge. Cela permet de sélectionner un cluster que l'on pourra étudier plus en détail dans la section **Encadré 4**.

Encadré 4 : Informations complémentaires sur la détection spatiale

Cet encadré regroupe des informations des informations sur une résolution géographique plus fine que dans la section **Encadré 3** ainsi que des informations supplémentaires sur la détection d'anomalie. Il comporte trois éléments :

- **le graphique des concentrations de la stations selectionnée :** lorsque l'on clique sur une station dans la **carte interactive leaflet**, ce graphique donne dans son titre le code identifiant de la station et trace les valeurs de concentrations relevées par la station durant le segment temporel selectionné dans la section **Encadré 1**. Les informations de quantifications de ces relevées présentes sont indiquées.

- **le graphique de front de Pareto des clusters :** dans ce graphique chaque cluster composant le clustering spatial choisi dans la section **Encadré 3** est représenté par un point. Les deux axes correspondent au deux critères qui permettent de juger de l'anormalité d'un cluster. L'axe des abscisses se définit comme l'hétérogénéité des distributions des concentrations des stations composant chaque cluster. Intuitivement, plus l'abscisse d'un cluster est faible, plus les profils de concentrations relevés par les stations de ce cluster se ressemblent. On retrouvera dans les valeurs d'abscisses basses tous les clusters dont les stations n'ont relevé que des données non quantifiées (durant segment temporel selectionné en section **Encadré 1**). Inversement, les clusters comportant des abscisses élevées présenteront beaucoup d'hétérogénéité dans les distributions de concentrations de ses stations. Les concentrations des stations ne se ressembleront pas forcément d'une station à une autre. Pour l'axe des ordonnées, plus un cluster comportera une valeur faible, plus la présence de valeurs élevées et/ou quantifiées dans ce cluster sera faible. Inversement, une valeur élevée sur l'axe des ordonnées indiquera une présence de valeurs élevées et/ou quantifiées. Lorsque l'on clique sur une station dans **la carte interactive leaflet**, le cluster contenant cette station sera surligné en rouge sur ce graphique.

- **les informations complémentaires sur le cluster selectionné :** permettent de résumer des informations sur les données de concentrations comprises dans le segment temporel sélectionné en section **Encadré 1** ET dans le cluster spatial sélectionné dans la section **Encadré 3**. On y trouve les informations suivantes :
  - le nombre de relevés effectuées par les stations du cluster spatial durant la période de temps définie par le segment temporel sélectionné.
  - le pourcentage de quantifications de ces données
  - le nombre de stations composant le cluster
  - le minimum de concentration de ces données
  - la moyenne de concentration de ces données
  - le médiane de concentration de ces données
  - le maximum de concentration de ces données
  - toutes les valeurs de LOQ présentes dans ces données, la valeur de LOQ la plus présente dans les données comporte une astérisque
  - des informations sur la station du cluster ayant le plus grand pourcentage de quantification (son code identifiant, son pourcentage de quantification et le nombre de relevés effectués pendant la période de temps sélectionnée)