LCSQA

Laboratoire Central de Surveillance de la Qualité de l'Air

Note technique

Travaux financés par le ministère chargé de l'environnement

ÉTUDE DE FAISABILITE POUR LA CONSTRUCTION D'UN OUTIL DE SYNCHRONISATION DES DONNEES DE SYSTEMES CAPTEURS

Salomon Morgane, Fievet Amandine et Spinelle Laurent (Ineris)

SYNTHESE

L'utilisation de systèmes capteurs pour la mesure de la qualité de l'air entraîne généralement la production d'une grande quantité d'informations que ce soit des données de mesures de polluants atmosphériques, météorologiques ou encore des informations de fonctionnement du système testé souvent appelées méta-données ou metadata. Ces technologies de mesures donnent accès à des mesures en temps réel qu'il est souvent nécessaire de retraiter (moyennes minute, quart-horaire ou horaire) mais également de synchroniser entre elles. Cependant, cette synchronisation des données sur un pas de temps commun devient rapidement compliquée lorsqu'elle fait intervenir plusieurs systèmes autonomes de par la grande quantité de données recueillies, la multiplicité des systèmes ayant chacun un pas de temps différents ou des horloges internes désynchronisées ne pouvant être synchronisées en amont des essais.

Ainsi, et pour répondre aux demandes des Associations agréées de surveillance de la qualité de l'air (AASQA) exprimées lors d'un atelier portant sur les capteurs durant les Journées Techniques des AASQA en 2018, le LCSQA-Ineris s'est proposé de conduire une étude de faisabilité pour construire un outil de synchronisation des données capteurs. À ce stade, une première version est disponible, nécessitant une mise en œuvre par les auteurs de la note[1].

[1] Contact : Spinelle Laurent - laurent.spinelle@ineris.fr

MAI 2020

ABSTRACT

The use of sensors systems for air quality monitoring usually results in the generation of a large amount of information, such as measurement of atmospheric pollutants data, meteorological data or working information regarding the tested device often referred to as metadata. These measurement technologies give access to real-time measurement that should often be reprocessed (minute average, 15 minutes or hourly averages) but also synchronised with each other. However, this data synchronisation on a common time base can becomes complicated when it involves several autonomous systems with a large amount of collected data, a multiplicity of systems having each one a different time base or desynchronised internal clocks that can't be synchronised before the experiments.

Thus, and to bring an answer to the questions from the local French air quality network (AASQA) raised during a workshop on sensors at the annual technical meeting of the AASQA (JTA) in 2018, the LCSQA-Ineris proposed to conduct a feasibility study to build a sensor data synchronisation tool.

1. INTRODUCTION

L'utilisation de systèmes capteurs pour la mesure de la qualité de l'air entraîne généralement la production d'une grande quantité d'informations que ce soit des données de mesures de polluants atmosphériques, météorologiques (température, humidité relative, pression atmosphérique, direction et vitesse de vent) ou encore des informations de fonctionnement du système testé souvent appelées méta données ou metadata (e.g. tension de la batterie, débit de pompe, données brutes de mesure). De plus, les technologies de mesures capteurs donnent accès à des mesures en temps réel ou à minima avec un pas de temps d'une seconde. Il est donc souvent nécessaire de retraiter ces données, notamment par l'utilisation de moyennes basées sur un pas de temps plus important (e.g. minute, quart-horaire ou horaire).

Ces systèmes capteurs sont en règle générale utilisés dans le cadre de campagnes de mesure sur le terrain en colocation avec des moyens de mesures dits « classiques » (e.g. analyseurs de gaz) et notamment des méthodes de référence. Une étape importante est donc la mise en commun des mesures effectuées avec des appareils différents, ayant chacun leur propre horloge et leur propre pas de temps d'acquisition lorsqu'ils ne peuvent être ajustés. Cependant, cette synchronisation des données sur un pas de temps commun devient rapidement compliquée lorsqu'elle fait intervenir plusieurs systèmes autonomes, de par :

- la quantité de données importante ;
- la multiplicité des sources ;
- des pas de temps différent ;
- des horloges désynchronisées.

Ainsi, et pour répondre aux demandes des Associations agréées de surveillance de la qualité de l'air (AASQA) exprimées lors d'un atelier portant sur les capteurs durant les Journées Techniques des AASQA en 2018, le LCSQA-Ineris s'est proposé de conduire une étude de faisabilité pour construire un outil de synchronisation des données issues de systèmes capteurs.

Un premier outil a pu être mis en place et est disponible auprès des auteurs de la note[2]. Cette note technique propose de détailler ses fonctionnalités, à son stade de développement actuel, et les différentes étapes de son utilisation.

2. REFERENCE ET REMERCIEMENT

Cet outil a été développé en utilisant le langage de programmation statistique R[3] avec des bibliothèques telles que *Shiny*[4] et *openair*[5]. La bibliothèque *Shiny* permet de créer des outils de traitement et visualisation de données intégrés dans une interface interactive directement à partir de R. La bibliothèque *openair* propose des fonctions dédiées à l'analyse de données de qualité de l'air.

3. PRESENTATION DE L'OUTIL

L'objectif premier de cet outil est de faciliter le traitement des données générées par les systèmes capteurs pour la mesure de la qualité de l'air. Il permet de synchroniser plusieurs fichiers sur une même base de temps, de les fusionner en un seul et même fichier csv, mais également de générer des moyennes sur une base de temps variable. Ainsi :

- la **synchronisation/fusion** permet d'unifier des fichiers CSV ou TXT en se basant sur la date et l'heure de l'échantillon ;
- la **moyenne** peut quant à elle être réglée en fonction des besoins, de moyennes secondes à moyennes pluri-journalières ;
- enfin, il est possible de **fusionner plusieurs fichiers et de moyenner** les données présentes en une seule et même étape.

À cette étape du développement, une phase de test préliminaire auprès d'utilisateurs internes au LCSQA a permis de soulever les verrous techniques d'une future mise en ligne :

- limitation des capacités de calcul à des fichiers inférieurs à 100 Mo ;
- l'absence de stockage ne permet pas le chargement de fichiers provenant de différents dossiers.

Pour le moment, l'utilisation de l'outil nécessite ainsi une mise en œuvre par le LCSQA. L'objectif final étant de mettre à disposition l'outil auprès des AASQA qui le souhaiteraient, ces dernières sont invitées à contacter les auteurs de la note pour qu'ils puissent le mettre en œuvre sur leurs jeux de données capteurs.

[2] Contact : Spinelle Laurent - laurent.spinelle@ineris.fr

[3] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application, Framework for R. R package version 1.4.0.2. https://CRAN.R-project.org/package=shiny

[5] Carslaw, D. C. and K. Ropkins, (2012) openair --- an R package for air quality data analysis. Environmental Modelling & Software. Volume 27-28, 52-61.

À terme, il est envisagé de rendre l'outil accessible directement aux AASQA, par exemple par l'utilisation d'un identifiant unique et d'un mot de passe qui seront gérés par le LCSQA.

La **Erreur ! Source du renvoi introuvable.** présente la fenêtre d'introduction de l'outil qui rappelle ses fonctionnalités.

Figure 1: Fenêtre d'introduction de l'outil

Un menu permet d'accéder aux différentes étapes d'utilisation de l'outil (**Erreur ! Source du renvoi introuvable.**).

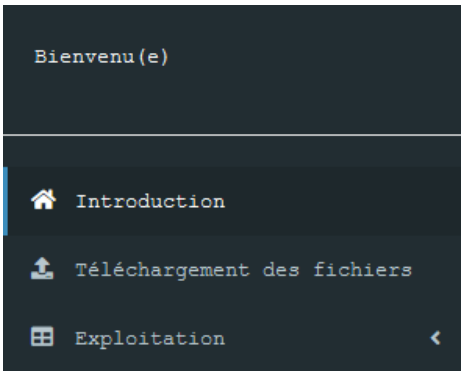

Figure 2: Table de navigation dans l'outil

La seconde étape "**Téléchargement des fichiers**" permet de charger les fichiers de travail (**Erreur ! Source du renvoi introuvable.**).

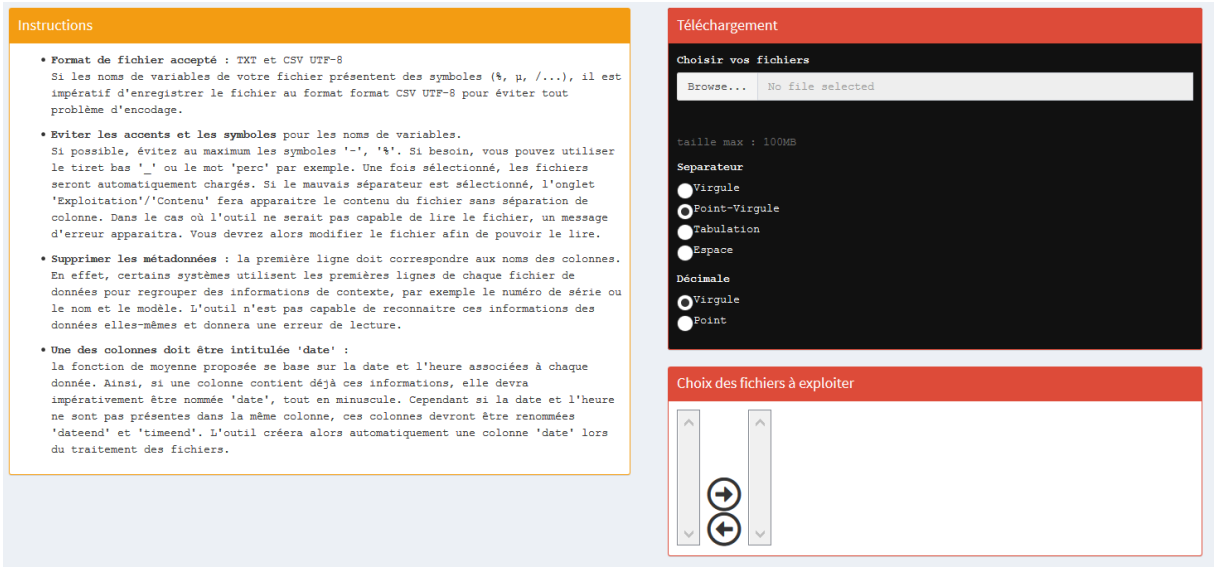

Figure 3: Téléchargement des fichiers et instructions spécifiques

L'utilisation de l'outil est soumise à certaines règles détaillées dans les instructions de la plateforme (**Erreur ! Source du renvoi introuvable.**) et reprisent ci-dessous :

- **Formats de fichier supportés :** TXT et CSV UTF-8. Si les noms de variables de votre fichier présentent des symboles (%, µ, /...) il est impératif d'enregistrer le fichier-au format CSV UTF-8 pour éviter tout problème d'encodage ;
- **Éviter les accents et les symboles pour les noms de variables.** Si possible, évitez au maximum les symboles "-", "%". Si besoin, vous pouvez utiliser le tiret bas "_" ou le mot "perc" par exemple. Une fois sélectionné, les fichiers seront automatiquement chargés. Si le mauvais séparateur est sélectionné, l'onglet "Exploitation"/"Contenu" fera apparaitre le contenu du fichier sans séparation de colonne. Dans le cas où l'outil ne serait pas capable de lire le fichier, un message d'erreur apparaitra. Vous devrez alors modifier le fichier afin de pouvoir le lire ;
- **Supprimer les métadonnées :** la première ligne du fichier doit correspondre aux noms des colonnes. En effet, certains systèmes utilisent les premières lignes de chaque fichier de données pour regrouper des informations de contexte, par exemple le numéro de série ou le nom et le modèle. L'outil n'est pas capable de reconnaitre ces informations des données elles-mêmes et donnera une erreur de lecture ;
- **Une des colonnes doit être intitulée 'date' :** la fonction de moyenne proposée se base sur la date et l'heure associées à chaque donnée. Ainsi, si une colonne contient déjà ces informations, elle devra impérativement être nommée "date", tout en minuscule. Cependant si la date et l'heure ne sont pas présentes dans la même colonne, ces colonnes devront être renommées 'dateend' et 'timeend'. L'outil créera alors automatiquement une colonne 'date' lors du traitement des fichiers.

A l'heure actuelle, l'outil est capable de lire deux formats de fichiers texte et reconnaitre quatre types de séparateurs de données et deux types de séparateurs décimaux. Cependant, tous les fichiers chargés simultanément doivent être formatés de la même manière **: même format de fichier, même séparateur de données, même séparateur décimal.**

Pour le moment, pour profiter de cet outil et avec leur accord, les données peuvent être fournies aux auteurs qui se chargeront de l'utilisation direct de l'outil.

Une fois les fichiers de travail téléchargés, la troisième et dernière fonction peut être activée : "**Exploitation**". Cette fenêtre se divise en deux onglets : "**Données**" et "**Fusion des fichiers**".

L'onglet " **Données**" (**Erreur ! Source du renvoi introuvable.**) permet de visualiser les différents fichiers chargés de manière individuelle. Il permet ainsi de contrôler si l'outil a correctement importé les différentes variables (colonnes) et de revenir à l'étape précédente dans le cas contraire pour modifier les fichiers ou les paramètres d'importation.

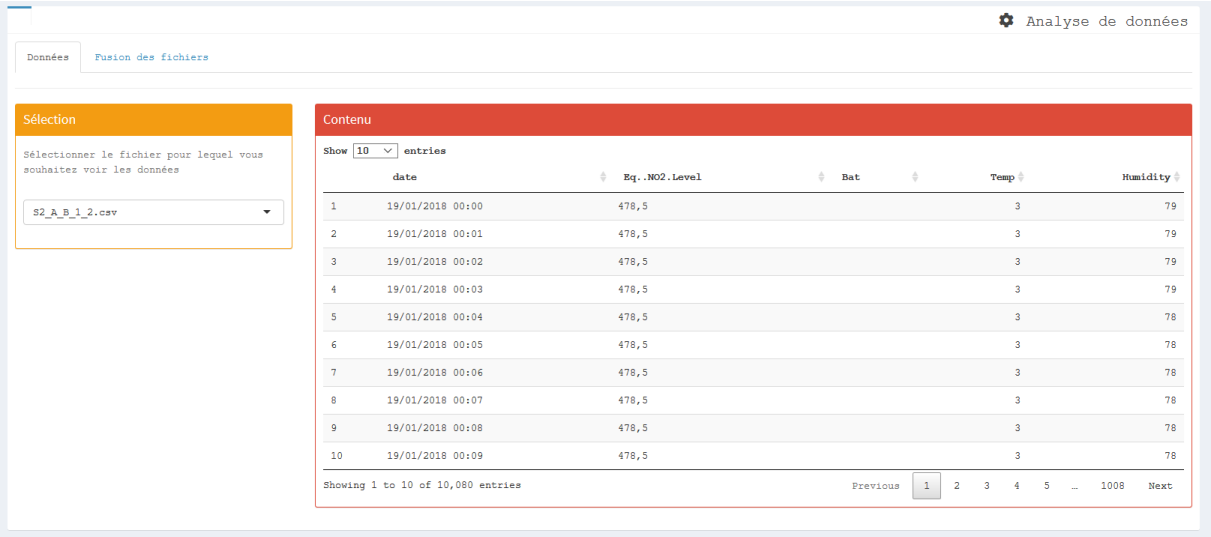

Figure 4: Visualisation des données possible avant fusion et calcul

Une fois les fichiers correctement importés, l'onglet "**Fusion des fichiers**" (**Erreur ! Source du renvoi introuvable.** donne accès aux fonctionnalités principales de l'outil. Ainsi la partie gauche de l'écran permet de sélectionner le type de traitement de donnés souhaité, ainsi que le pas de temps en cas de calcul de la moyenne.

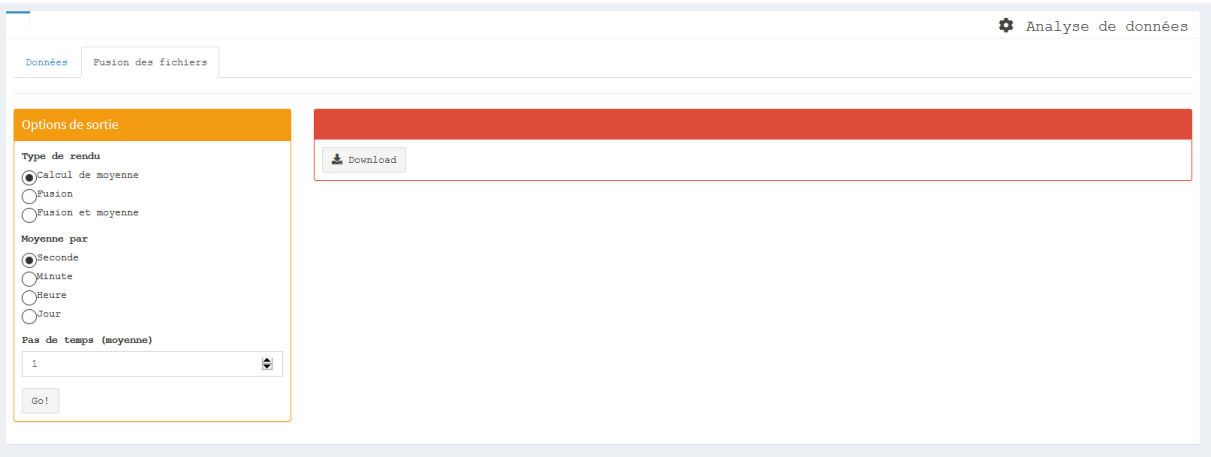

Figure 5: Fenêtre de fusion de données

Les trois traitements possibles sont :

- **le calcul de moyenne** sur une base de temps déterminée en fonction des options décrites ci-dessous ;
- **la fusion** qui permet d'unifier plusieurs fichiers de données en se basant sur la date et l'heure de l'échantillon. Dans cette option, l'intégralité des données de chaque fichier sera conservée. En cas de données manquantes, l'outil n'est à l'heure actuelle pas capable de compléter ou d'extrapoler ces données. Il se contentera d'indiquer NA dans les cellules vides.
- enfin, il est possible de **fusionner** plusieurs fichiers et de **moyenner** les données présentes en une seule et même étape.

Concernant le calcul de moyenne, la durée peut- être fixée au moyen de deux paramètres :

- l'**unité de temps** : "**Seconde**", "**Minute**", "**Heure**" et "**Jour**" ;
- le **pas de temps**, soit le nombre associé à l'unité de temps définit précédemment. Cette option est un champ libre, par défaut cette valeur est fixée à 1.

Par exemple, dans le cas de moyennes quart-horaire, il suffit de sélectionner "Minute" en unité de temps et d'inscrire "15" dans le champ « pas de temps ». Une fois les paramétrages réalisés, le bouton "**Go!**" permet de lancer le traitement. Le résultat obtenu s'affiche dans la partie droite de l'écran (**Erreur ! Source du renvoi introuvable.**).

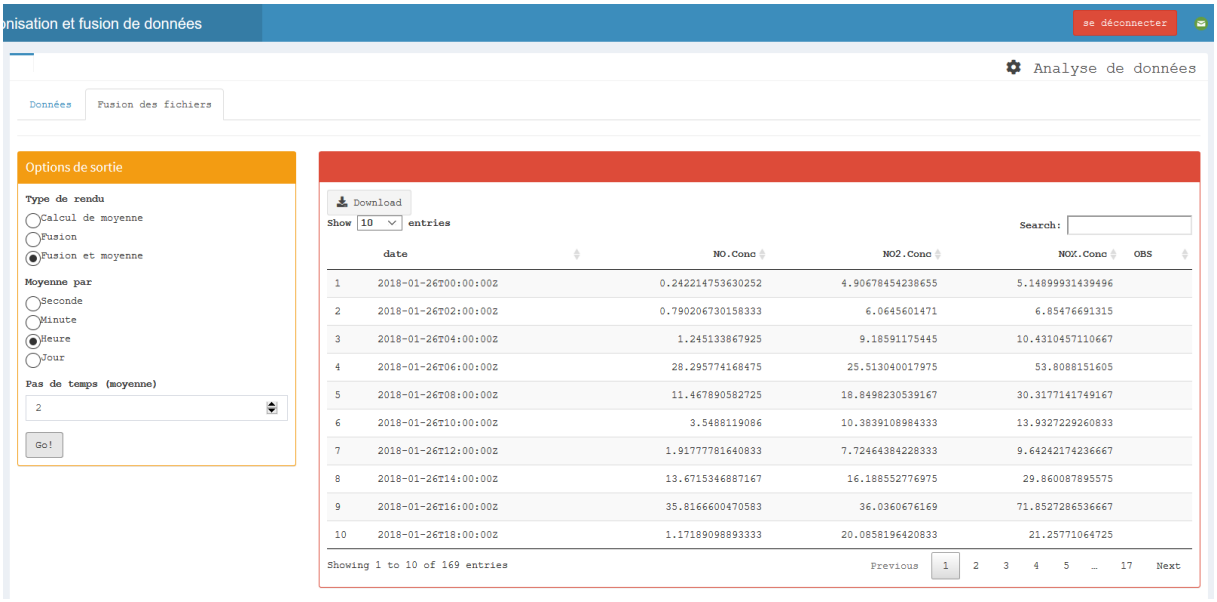

Figure 6: Résultat de la fusion de données paramétrée avec un pas de temps de 2 heures

Ces résultats sont alors téléchargeables sous forme d'un unique fichier "data_exporte.csv" à l'aide du bouton "Download" situé au-dessus du tableau de résultats. Pour le moment, l'export est disponible dans **un seul format :**

- **Fichier ".csv"** ;
- **Séparateur de données : "," aussi appelé "virgule" ou "comma"** ;
- **Séparateur numérique : "."**.

L'utilisateur final devra donc ensuite ouvrir et/ou convertir ces données avec ses propres logiciels de traitement pour continuer son étude.

4. CONCLUSION DE L'ETUDE DE FAISABILITE

L'objectif de cette étude de faisabilité était de construire un outil de synchronisation des données de systèmes capteurs en vue de leur exploitation. Cette note technique reprend ainsi les principales étapes de son utilisation. Son développement a pu être concrétisée par la mise à disposition auprès d'utilisateurs internes au LCSQA pour une phase de test préliminaire qui a permis d'ajuster certaines erreurs et de soulever les principaux verrous techniques pour une future mise en ligne de cet outil.

En particulier, nous avons pu observer la limitation des capacités de calcul dans le cas de fichiers lourds, dépassant les 100 Mo. Nous avons également mis en exergue la difficulté de charger l'ensemble des fichiers en une seule fois. En effet, l'absence de stockage ne permet pas le chargement de fichiers provenant de différents dossiers. Ainsi, si l'outil est pleinement fonctionnel, pour profiter de ses fonctionnalités, il nécessite pour le moment de s'adresser aux auteurs pour pouvoir l'utiliser.

En termes de perspective, l'Ineris souhaiterait pouvoir mettre cet outil à disposition, au moyen d'un système d'authentification individuelle ou par catégorie d'utilisateur (Programme de travail 2020 du LCSQA). Il serait également intéressant d'ajouter des fonctionnalités graphiques simples, comme par exemple des séries temporelles.