gapIT : Un outil visuel pour l'imputation de valeurs manquantes en hydrologie

Olivier Parisot, Laura Giustarini, Olivier Faber, Renaud Hostache, Ivonne Trebs, and Mohammad Ghoniem

Centre de Recherche Public – Gabriel Lippmann, Belvaux, Luxembourg
parisot@lippmann.lu

Résumé. Les données manquantes sont problématiques en hydrologie, car elles gênent le calcul de statistiques interannuelles et sur de longues périodes, ainsi que l'analyse et l'interprétation de la variabilité des données. Dans cet article, nous présentons gapIT, une plateforme d'analyse de données permettant d'inspecter visuellement les données manquantes et ensuite de choisir la méthode de correction adéquate. Nous avons utilisé l'outil pour estimer les données manquantes dans des séries temporelles correspondant aux débits mesurés par des stations hydrométriques du Luxembourg.

1 Introduction

Traditionnellement, les données hydrométriques se présentent sous la forme de séries temporelles, représentant des mesures effectuées régulièrement par des stations : ces mesures peuvent concerner différents aspects comme les hauteurs et les débits de l'eau dans les cours d'eau, les quantités de précipitations, etc. Comme ces mesures sont souvent prélevées par un réseau distribué de capteurs, le problème des données manquantes est inévitable. Allant d'une simple valeur manquante à une longue plage de valeurs manquantes, les lacunes peuvent avoir des causes multiples : dysfonctionnement des capteurs, maintenance des stations de mesure, erreurs humaines, etc. (Harvey et al. (2010)).

Le réseau hydrométrique au Luxembourg fournit un bon cas d'utilisation. Il est constitué de différentes stations hydrométriques permettant de mesurer notamment les débits des cours d'eau. Les mesures sont ensuite fréquemment utilisées dans les modèles numériques de prévision hydrologique ou pour calculer des statistiques sur les écoulements (e.g. temps de retour des crues ou des sécheresses).

En conséquence, lorsque certaines séries de mesures présentent beaucoup de lacunes (par exemple : les données de la station de HallerBach au Luxembourg, Figure 1), cela pose de nombreux problèmes et il est nécessaire d'apporter un soin très particulier à combler ces lacunes avec une bonne précision.

Afin de combler ces lacunes, les méthodes classiques d'analyse de données ont été appliquées dans le domaine hydrologique (Salas (1980)), et des travaux récents tentent de fournir des solutions toujours plus efficaces (Harvey et al. (2010); Mwale et al. (2012)). Or, il est souvent difficile de choisir la bonne méthode de calcul parmi toutes celles qui existent car

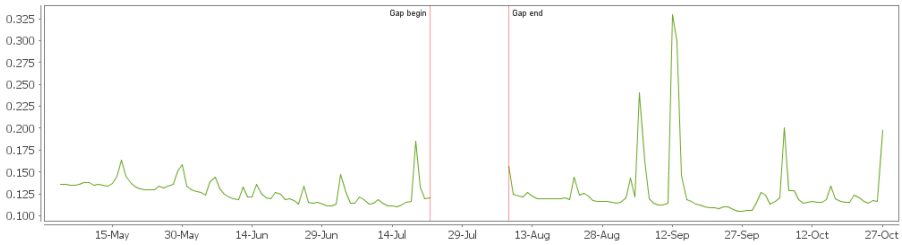

FIG. 1 – *Un exemple de trou dans une série temporelle montrant l'évolution du débit dans la station de Hallerbach (Luxembourg). L'abscisse représente le débit en* m^3/s.

cette tâche dépend grandement du contexte et de l'expert en charge de corriger les données (Gyau-Boakye et Schultz (1994)) :

- L'interpolation est une solution simple et efficace si les séries sont plutôt de type *continu* (sans variations importantes, par exemple) et si le trou n'est pas trop grand.
- Les régressions linéaires sont également utilisées (Bennis et al. (1997)), notamment lorsqu'il existe des liens évidents entre les réseaux de capteurs : elles peuvent permettre de capturer les relations entre les mesures effectuées en amont et en aval d'un cours d'eau.
- Les arbres de régression sont également de bons candidats : ils sont plus expressifs que des formules issues de régressions linéaires, et ils sont simples à visualiser et à interpréter ((Witten et al., 2011, section 3.3), Kotsiantis (2013)).
- Les réseaux de neurones artificiels ont récemment été utilisés, notamment via des *perceptrons* (Tfwala et al. (2013)) ou des cartes auto-adaptatives (Mwale et al. (2012)).
- L'algorithme espérance-maximisation (EM) est souvent utilisé pour reconstruire des données manquantes (Van Hulse et Khoshgoftaar (2008)).
- Enfin, différentes techniques de prédiction de séries temporelles peuvent être utilisées suivant les caractéristiques des séries (ARMA, ARIMA, etc.).

Habituellement, les experts en hydrologie se servent de divers scripts pour corriger les données (*R*, *MATLAB*). Ainsi est-il important pour eux de pouvoir disposer d'un outil interactif pour à la fois intégrer les diverses sources de données, visualiser les séries temporelles, et enfin choisir le mode d'estimation de valeurs manquantes le plus adapté.

2 gapIT : un outil pour estimer les données manquantes

2.1 Technologie

gapIT est une application développée en JAVA et basée sur le logiciel d'analyse et de traitement de données Cadral (Pinheiro et al. (2014)). Les données en entrée sont de différents types : *a*) les séries temporelles correspondant aux mesures prises aux différentes stations ; *b*) les coordonnées géographiques des stations ; *c*) les dépendances amont/aval entre stations (suivant les cours d'eau).

L'interface graphique permet une navigation interactive dans les données : les séries temporelles sont visualisées via la librairie graphique JFreeChart [1]. De plus, une carte interactive des stations de mesure et les indispensables fonctionnalités de filtrage sont fournies. Il est également possible de visualiser le graphe de dépendance entre stations grâce à la librarie JUNG [2].

Pour la partie calculatoire (régressions, réseaux de neurones, etc.), le système utilise les fonctionnalités de la librairie de fouille de données WEKA (Witten et al. (2011)), ainsi que des composants de la librairie Apache Commons Math [3].

2.2 Inspection et caractérisation des valeurs manquantes

Les valeurs manquantes ne se corrigent pas de la même manière selon le contexte (taille des trous, saison durant laquelle les valeurs sont manquantes, probabilité qu'une crue soit en cours, etc.) (Gyau-Boakye et Schultz (1994)). Par conséquent, ces informations sont calculées et affichées dans l'interface graphique (Figure 2).

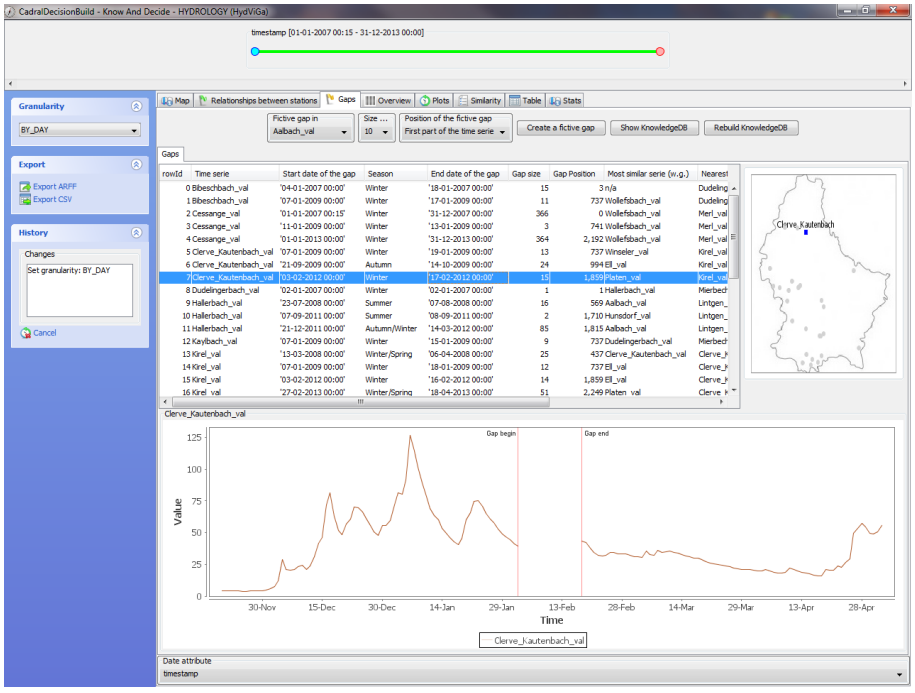

FIG. 2 – *gapIT permet d'inspecter visuellement les lacunes, de les caractériser en fonction du contexte afin de déterminer la technique d'approximation la plus adaptée.*

1. http ://www.jfree.org/jfreechart/
2. http ://jung.sourceforge.net/
3. http ://commons.apache.org/proper/commons-math/

2.3 Estimation des valeurs manquantes

Pour un trou donné, l'outil propose un module pour estimer les valeurs manquantes (Figure 3) en suivant les trois phases suivantes : sélection des stations de référence, sélection de l'algorithme d'estimation, et enfin visualisation et évaluation des résultats.

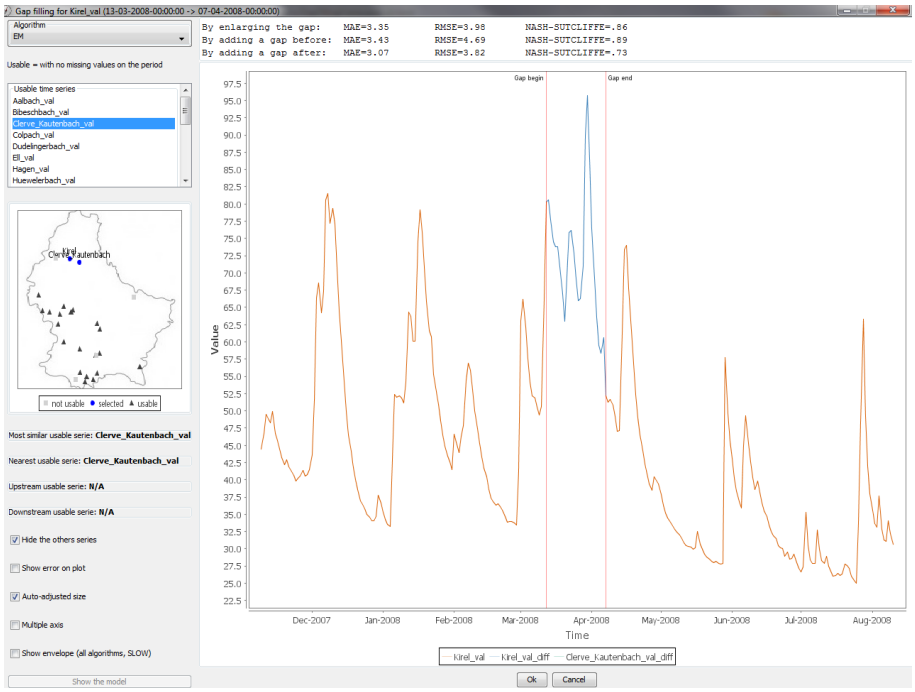

FIG. 3 – *Pour remplir un trou donné, l'utilisateur peut choisir les stations de référence et l'algorithme d'estimation. En résultat, il obtient une estimation des valeurs manquantes, complétées par des indications quant à la précision des résultats (MAE, RMSE, Nash Sutcliffe).*

Premièrement, la phase de sélection des stations *de référence* est critique, car elle permet à l'expert de choisir les séries temporelles qui serviront à compléter les séries temporelles lacunaires. Pour ce faire, l'outil propose plusieurs approches complémentaires :

- la sélection des stations les plus proches géographiquement ;
- la sélection des stations se trouvant sur le même cours d'eau (en amont et/ou en aval) ;
- la sélection des stations ayant les séries temporelles les plus similaires sur la période concernée ; la similarité est calculée en utilisant la déformation temporelle dynamique (*Dynamic Time Warping*) (Berndt et Clifford (1994)).

Deuxièmement, l'outil propose diverses méthodes d'estimation : interpolation, régressions multiples, arbres de régressions, réseaux de neurones (perceptron multi-couches), algorithme des plus proches voisins, etc.

Troisièmement, l'utilisateur peut ensuite appliquer via le logiciel la méthode d'estimation sélectionnée en se basant sur les séries temporelles choisies pour évaluer les valeurs man-

quantes. Or, il est important de déterminer la précision des estimations produites. Pour ce faire, des trous fictifs sont créés dans la fenêtre de temps proche du trou réel à remplir (par exemple : un trou avant, un trou après, élargissement du trou en cours d'examen ou à un endroit choisi par l'utilisateur). Ensuite, les mesures suivantes sont calculées : l'erreur absolue moyenne (MAE), la racine carrée de l'erreur quadratique moyenne (RMSE) et le coefficient *Nash-Sutcliffe* car c'est un indicateur très commun en hydrologie (Nash et Sutcliffe (1970)).

Pour finir, l'outil est capable de calculer automatiquement la *configuration optimale* (stations de référence et algorithme). Dans ce cas, l'utilisateur garde la possibilité de modifier la configuration à son gré de manière à obtenir un nouveau résultat plus proche de ses attentes.

3 Exemple : les débits des cours d'eau au Luxembourg

gapIT a été utilisé pour estimer les débits d'écoulement manquants pour des stations sélectionnées au Luxembourg, sur la période allant du 1er janvier 2007 au 31 décembre 2013. Le jeu de données utilisé correspond à des mesures effectuées toutes les quinze minutes dans 24 stations. Afin de tester l'efficacité de l'outil, un ensemble de trous fictifs a été créé pour ces stations. Pour obtenir un ensemble représentatif, les trous générés sont de différentes tailles, se situent durant différentes saisons, etc. Ensuite, pour chacun de ces trous fictifs, toutes les techniques d'estimation proposées par l'outil ont été testées, et pour chacun des cas, les taux d'erreur ont été mesurés (MAE, RMSE, Nash-Sutcliffe).

Ainsi, nous avons constaté que les réseaux de neurones et les arbres de régression permettent d'obtenir les taux d'erreur les plus faibles. De plus, si l'on considère le meilleur résultat concernant chaque trou, alors on voit que les taux d'erreur sont globalement très faibles. Cela signifie que pour ces cas, une estimation correcte est possible en utilisant les données présentes (Figure 4). En revanche, dans un certain nombre de cas, les meilleurs taux d'erreurs sont élevés. Après analyse, il s'avère que pour les stations concernées, il n'existe pas suffisamment de stations assez proches, similaires ou dépendantes afin de créer une estimation assez précise.

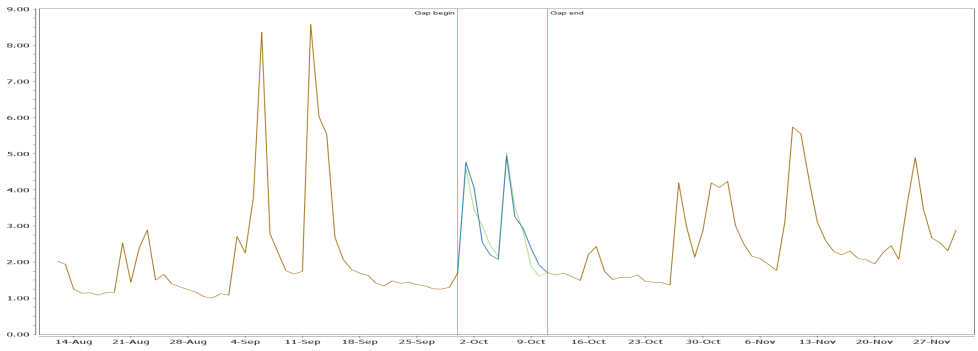

FIG. 4 – *Un trou fictif créé pour le débit mesuré par la station Hunsdorf en octobre 2008. En brun, les valeurs existantes* (m^3/s)*; en vert les valeurs effacées ; en bleu, les valeurs estimées via un réseau de neurones. Les taux d'erreur : MAE=0.28, RMSE=0.02, Nash-Sutcliffe=0.90.*

4 Conclusion et perspectives

Dans cet article, nous avons présenté gapIT, une plateforme permettant de visualiser, d'analyser et d'estimer les valeurs manquantes dans les séries temporelles hydrologiques. L'outil a été utilisé pour traiter les lacunes dans les débits mesurés par des stations hydrométriques au Luxembourg. Les futurs travaux concerneront l'extension de la solution pour intégrer simultanément différents types de données hydrologiques. De plus, une étude sera menée quant à l'utilisation de techniques d'analyse de flux afin d'analyser les données en temps réel.

Références

Bennis, S., F. Berrada, et N. Kang (1997). Improving single-variable and multivariable techniques for estimating missing hydrological data. *Journal of Hydrology 191*, 87 – 105.

Berndt, D. J. et J. Clifford (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, Volume 10, pp. 359–370. Seattle, WA.

Gyau-Boakye, P. et G. Schultz (1994). Filling gaps in runoff time series in West Africa. *Hydrological sciences journal 39*(6), 621–636.

Harvey, C. L., H. Dixon, et J. Hannaford (2010). Developing best practice for infilling daily river flow data. *Role of Hydrol. in Managing Consequences of a Changing Glob. Env.*.

Kotsiantis, S. (2013). Decision trees : a recent overview. *A.I. Review 39*(4), 261–283.

Mwale, F., A. Adeloye, et R. Rustum (2012). Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi–a SOM approach. *Phys. and Chem. of the Earth 50*, 34–43.

Nash, J. et J. Sutcliffe (1970). River flow forecasting through conceptual models part i : A discussion of principles. *Journal of hydrology 10*(3), 282–290.

Pinheiro, P., Y. Didry, O. Parisot, et T. Tamisier (2014). Traitement visuel et interactif dans le logiciel Cadral. In *Atelier GT-VIF, EGC 2014, Rennes, France*.

Salas, J. D. (1980). *Applied modeling of hydrologic time series*. Water Resources Publication.

Tfwala, S. S., Y.-M. Wang, et Y.-C. Lin (2013). Prediction of missing flow records using multilayer perceptron and coactive neurofuzzy inference system. *The Sc. World Journal 2013*.

Van Hulse, J. et T. M. Khoshgoftaar (2008). A comprehensive empirical eval. of missing value imputation in noisy software measurement data. *Journ. of Syst. and Soft. 81*(5), 691–708.

Witten, I. H., E. Frank, et M. A. Hall (2011). *DM : Pract. ML Tools and Techniques*. Elsevier.

Summary

Missing values in hydrological time series stand in the way of the creation of hydrological models, the computation of important summary statistics and the analysis and interpretation of flow variability. In this paper, we propose gapIT, a Visual Analytics system that aims to help experts to find the appropriate infilling technique according to the context. The tool has been used to infill gaps in water discharge series obtained from selected hydrometric gauging stations in Luxembourg.