From online data to model input:

a flexible open source data analysis tool

**C. De Mulder**1\*, T. Flameling2, J. Langeveld3,

S. Weijers2, I. Nopens1 and Y. Amerlinck1

1BIOMATH, Department of Mathematical Modeling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium

2Waterschap De Dommel, PO Box 10.001, 5280 DA Boxtel, The Netherlands

3Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands

\*Corresponding author, chaim.demulder@ugent.be

**Keywords:** online data, data analysis, gap filling

**Introduction**

Models have since a long time been valuable tools in the operational and performance optimization of Water Resource Recovery Facilities (WRRF). A WRRF model with good predictive power requires good data, as an input and for calibration and validation purposes. Especially for models aimed at capturing the dynamics of an installation, a high-frequency input is necessary (Cierkens *et al.*, 2012, Martin & Vanrolleghem, 2014), which is often generated by means of different influent generator options (Langergraber *et al.*, 2008, Gernaey *et al.*, 2011). In case there is indeed high frequency data available, still these influent generators are used, because the on-line measurements are subject to noise, sensor failure or other classic problems hindering on-line data acquisition (Rieger *et al.*, 2010). This way, a lot of information contained in the online data is lost. This contribution presents a Python™ package that allows the user to both separate valuable data from faulty data and replace the faulty data with, for example, data based on an influent model. The data train from online measurements to useable model input is hereby made flexible and reproducible.

**Materials and methods**

The data used to illustrate the developed code was provided by Waterboard De Dommel and acquired at the WRRF of Eindhoven in the year 2013. Data for the whole year was made available, containing on-line sensor data for the three sewer flows entering the plant (COD, soluble COD and TSS) and for the combined flows (ammonia and phosphate). A data point was available every 5 minutes.

The data analysis and gap filling code was written in Python™ (Python Software Foundation, Oregon, USA). The code includes (1) simple detection of noise and sensor failure, (2) calculation of data characteristics (e.g. an average daily profile, correlations between data series…) (3) the possibility to use these characteristics or external influent modeling results to fill the gaps in the on-line data on a scientifically sound basis and (4) plotting functionalities to visualize the whole data analysis train.

In the current case, small gaps in the data are filled based on linear interpolation, large gaps are filled with influent model output. This influent model was described by Langeveld *et al.* (2014) and implemented in WEST (MIKEbyDHI, Denmark), specifically for the WRRF of Eindhoven. It showed good agreement with on-line data after calibration and was therefore deemed a good option for gap filling.

**Results**

The possible use and results of the described package are illustrated with data analysis and filling of ammonia data for a period of three weeks in October 2013 (**Figure 1**). At the end of the proposed work flow (the lower of the four data plots), a relatively smooth dataset is obtained, containing reliable, dynamic and still highly frequent information on the actual influent. **Figure 2** illustrates the flexibility of the code functionalities, in the sense that the user can also decide within what time range (s)he wants data points to be replaced. Using this flexibility, an increasingly smooth dataset is obtained, possibly avoiding numerical issues when running a WRRF model fed with this input.

Data availability based on the used filter functions varied from as low as 4% to 84% for the whole year 2013. Especially for measurements where higher percentages of reliable data are available, the proposed methodology, thanks to its time-saving potential, allows for increased use of the available data. Datasets filled with this procedure have been used as an input to the model of the Eindhoven WRRF, yielding very good validation results. An example of such a result is shown in **Figure 3**.

**Conclusions**

The results show that the developed Python™ code not only assists in vastly automating the data filtering, filling and the visual tracking of the process, but also in making the workflow transparent and easily reproducible. Use of the package and workflow also implies that the valuable information contained in online measurements can still be recovered in a modeling context, while this data is otherwise often discarded due to the time constraints of data reconciliation. In case an influent model is not available, although it is the preferred choice, other filling techniques are already implemented (filling based on daily averages, correlations, previous data points…) and will gladly be illustrated in case of acceptance.

**References**

K. Cierkens, S. Plano, L. Benedetti, S. Weijers, J. de Jonge, and I. Nopens. Impact of influent data frequency and model structure on the quality of WWTP model calibration and uncertainty. Water Sci Technol, 65(2):233– 42, 2012.

K. V. Gernaey, X. Flores-Alsina, C. Rosen, L. Benedetti, and U. Jeppsson. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environmental Modelling & Software, 26:1255–1267, 2011.

G. Langergraber, J. Alex, D. Weissenbacher, Norbert ad Woerner, M. Ahnert, T. Frehmann, N. Halft, I. Hobus, M. Plattes, V. Spering, and S. Winkler. Generation of diunral variation for influent data for dynamic simulation. Water Science and Technology, 57(9):1483–1486, 2008.

J. Langeveld, R. Schilperoort, P. Van Daal, L. Benedetti, Y. Amerlinck, J. de Jonge, T. Flameling, I. Nopens, and S. Weijers. A new empirical sewer water quality model for the prediction of WWTP influent quality. In Proceedings of the 13th Conference on Urban Drainage, 2014.

C. Martin and P. A. Vanrolleghem. Analysing, completing, and generating influent data for WWTP modelling: A critical review. Environmental Modelling & Software, 60:188 – 201, 2014.

L. Rieger, I. Takacs, K. Villez, H. Siegrist, P. Lessard, P. A. Vanrolleghem, and Y. Comeau. Data reconciliation for wastewater treatment plant simulation studies - planning for high-quality data and typical sources of errors. Water Environment Research, 82(5):426–438, 2010.



**Figure 1** Plots of the complete filtering and filling procedure, here shown for 3 weeks of ammonia measurements. Top to bottom: original data; original data points and points filtered out by noise filtering and sensor failure detection; original and filtered data points, along with points replaced by interpolation; original data points and points filled by interpolation or by filling with modeled values; the influent flow rate at the WWTP.



**Figure 2** The filled dataset for ammonia measurements, making use of an extended range within which to use modeled values for replacement.



**Figure 3** Example validation of the ammonia concentration in the Activated Sludge Tanks at the Eindhoven WRRF. Simulations were run with influent datasets obtained from on-line data and subsequent data validation and gap filling as described throughout this contribution.