A black-box model for generation of site-specific WWTP influent quality data based on plant routine data

Markus Ahnert, Conrad Marx, Peter Krebs and Volker Kuehn

ABSTRACT

This paper presents a simple method for the generation of continuous influent quality datasets for wastewater treatment plants (WWTPs) that is based on incomplete available routine data, only, without referring to any further measurement. In the approach, Weibull-distributed random data are fitted to the available routine data, such that the resulting distribution of influent quality data shows the identical statistical characteristics. Beside the description of the method, this paper contains a comprehensive analysis of robustness and universality of the approach. It is shown that incomplete datasets with only 10% remaining influent quality data can be filled with this method with nearly the same statistical parameters as the original data. In addition, the use with datasets of different WWTP plants sizes results always in a good agreement between original and filled datasets.

INTRODUCTION

In wastewater treatment plants (WWTPs) measurements are required for documentation of influent and effluent quality, process support and control, as well as process optimisation. In general, there are two types of data aggregation that can be distinguished:

Automatically measured data, which are mostly used for quantitative parameters (e.g. flow meter), are logged in a high frequency, convertible into different time scales and available at the Supervisory Control and Data Acquisition (SCADA) system.

Analytically determined data, which are mostly used for quality parameters (e.g. influent and effluent quality), are realised as grab or composite samples (over a period of 24 h) and are available in SCADA or operations diary after manual input.

Depending on regulations, as well as on the size of the WWTP, especially analytically determined data, are usually available with a low time resolution. In practice, quality measurements are carried out once a week, only at working days, or up to seven days per week on a daily basis. The advantage of using routine data for analysis purposes is their availability for long time periods with no additional costs (Spindler 2014).

Available databases can be used for further purposes, besides the ones depicted above. A practical application is presented by Ahnert et al. (2010). The SCADA data for temperature with a high time resolution is used as an alternative conservative measure instead of a real tracer. With this measure, the residence time distribution of tanks in a WWTP can be estimated.

The main application of influent data is certainly the use for design and layout of plant extension (reactors, aeration system and technical equipment). Another task becoming more important in the last decades is the use of plant data for setting up WWTP models. Models can serve as a tool for dynamic long-term simulation and to validate the quality of extensive datasets over one or more years. For this purpose, it is not necessary to have a dynamic intra-daily information of influent concentration variation.

Substantial effort has been invested in establishing a systematic approach for WWTP model identification and modelling (e.g. Langergraber et al. 2004; Rieger et al. 2010, 2012), while it was concluded that modelling requires a consistent influent load database. This is in contrast to many of the routine datasets from WWTP. In particular, at smaller plants, the operation is controlled on the basis of low data resolution, which complicates or hinders enhanced applications in terms of data processing and model identification. Therefore, adequate methods to improve data
resolution and to substitute missing values are needed. Martin & Vanrolleghem (2014) reviewed possible interpolation methods and tools differing in complexity and in resulting time-resolution (Devisser et al. 2006; Gevaert et al. 2009; De Keyser et al. 2010; Gernaey et al. 2011). All of them require more or less intense preliminary work (e.g. measurement campaign) as well as site-specific data and information, which results in a time-consuming procedure.

In this paper, we will present a straightforward method for generating a continuous influent dataset for a WWTP with rather low routine-data density including missing data of quality measures. The advantage of the presented approach is the integration of specific stochastic information describing the fluctuations in that respective catchment area combined with a very simple method. For our method, we exclusively used routine data such as daily mean values (e.g. for flows) and 24 h composite samples without any further measurement. In order to include a wide range of flow rates, a catchment area with a large proportion of combined sewers was chosen for this study.

**MATERIAL AND METHODS**

As noted in the section above, the required influent data for the proposed method is straightforward. Over a time range of one or several years, the following dataset is needed:

- daily mean values of inflow measurement (preferred from a combined sewer system; data from separate systems show a smaller range of inflow rate);
- available 24 h composite samples of chemical oxygen demand (COD);
- available 24 h-composite samples of total nitrogen (N_{tot}) and total phosphorus (P_{tot}), or otherwise;
- realistic ratios for N_{tot}/COD and P_{tot}/COD for this catchment area or from the literature.

As a first step, a commonly performed pre-processing is necessary for the elimination of outliers and the identification of fluctuations like weekly or seasonally oscillating trends. If no significant oscillations can be determined (e.g. seasonal variation due to holiday seasons or industrial influence) and only a minor part of storm water is directly discharged into the receiving water, it is accepted to assume a close-to-constant influent load for WWTBs predominantly treating municipal wastewater, if working with influent data covering longer time periods of one or more years.

With the assumption of constant daily influent loads, the influent COD concentration of a single day can be calculated from the recorded inflow using the following equation:

\[ C_{COD} = \frac{L_{COD}}{F} = L_{COD} \cdot F^{-E} \]  

(1)

with \( C_{COD} \) being the COD inflow concentration \([\text{g} \cdot \text{m}^{-3}]\), \( L_{COD} \) being the COD input load \([\text{g} \cdot \text{d}^{-1}]\), \( F \) the inflow \([\text{m}^3 \cdot \text{d}^{-1}]\) and \( E = 1 \).

In theory, an inversely proportional relationship between flow and concentration is assumed, applying a constant of \( E = 1 \) in Equation (1). In case of field data, the correlation can be improved when fitting the exponent \( E \) to data (in that case, the coefficient \( L \) corresponds no longer to the COD load). An example is given in Figure 1. Real data of influent flow (daily mean values) and COD concentration (24 h-composite samples) from a WWTP with 490,000 p.e. are plotted in the left part of Figure 1. The general aspect of the relationship between flow and wastewater quality from Equation (1) is evident and was also shown by others (e.g. Rousseau et al. 2003). In general, there is a good match between the measured data and both fits. The slightly better approximation quality for the fit using \( E \neq 1 \), in terms of the coefficient of determination \( R^2 \), results from the characteristics of the sewer systems. For example, a combined sewer overflow event will result in lower input loads to the WWTP at wet-weather conditions (high flow rates) and affects the relationship described in Equation (1). Hence, the use of \( E \) as a fitting parameter accounts for a variety of characteristics of the catchment area and is more appropriate to approximate the influent routine data. In order to point out the advantages of the presented approach, the relationship between flow rate \( F \) and concentration \( C \) using \( L \) and \( E \) as a constant to be fitted to data, first was chosen as a main principle for estimating missing concentration values in the influent dataset.

For a more detailed analysis of the relationship between flow rate and resulting influent concentration, the dataset was divided into groups according to the flow rate (example at the right-hand side of Figure 1). The COD concentration distributions were plotted for two of the subsets, i.e. for low and high flow rates to the WWTP, respectively. As it would be expected, the mean COD concentration is lower for the subset of high flow rate and vice versa. Interestingly, the distributions are not symmetric in every case, where they show a skewness in the opposite direction. Based on this observation, a systematic approach of splitting the data into sub-groups was applied. In a first step of data aggregation, the dataset was divided in groups with a constant
number of data points, sorted by the flow rate. For every group, the distribution of COD concentration values was plotted. An example with six groups (A–F) is shown in Figure 2. Plot G shows the entire dataset similar to Figure 1, but with relative flow rates normalised by the maximum value of the complete dataset in order to be able to compare the data with those of other catchments. Some statistical parameters of the dataset from Figures 1 and 2 can be found in Table 1. Resulting values are typical for combined sewer systems in Germany and other countries too.

Distribution plots are depicted for every group A to F with increasing influent flow rate range, i.e. from (A) with the concentration values of the group with the lowest flow rates and (F) with the highest flow rates. It can easily be seen that the distributions exhibit different shapes with differing skewness. While the first distributions have their maximum on the right side it moves to the left in distributions with higher flow rates.

For mathematical description of such distributions with varying skewness, the Weibull function is appropriate. There is a wide use of the Weibull distribution for e.g. calculating wind speed distributions (Chellali et al. 2013) and in the field of environmental sciences (e.g. Gendig et al. 2003). It is based on two parameters for shape and scale, respectively.
Equation (2) shows the Weibull probability density function.

\[ f(x) = \lambda x^{k-1} e^{-(\lambda x)^k} \]  
(2)

with \( x \) as representative for relative flow rate \( F/F_{\text{max}} \), \( \lambda > 0 \) (scale parameter) and \( k > 0 \) (shape parameter).

Regarding groups A–F, putting both parameters of the Weibull function into relation to the normalised influent flow rate yields an exponential relationship as shown in plot H for shape parameter and plot I for scale parameter (see Figure 2). Plots H and I show a strong correlation of Weibull scale and shape parameters for an exponential fit. It can also be seen that the point on right-hand side in scale parameter plot (I) slightly differs from the course of the others. This point results from the group with the highest flow rate (F). This group includes the widest range of flow rates due to rareness of days with wet-weather flow compared to dry weather days. Using the described method especially the last group of data should be investigated carefully. Also it should be emphasized that a check of the goodness-of-fit of the shape and scale parameter is good enough because the resulting exponential functions are the base of generation of missing data values.

These flow-dependent relationships can be used for the generation of COD concentration values. Therefore, a random number for every missed concentration data (based on flow rate) is generated with a distribution function based on the shape and scale parameters (H and I) from the Weibull fit of the different probability distribution groups. After that, missing concentration values for \( N_{\text{tot}} \) and \( P_{\text{tot}} \) can be calculated using available \( N_{\text{tot}}/\text{COD} \) and \( P_{\text{tot}}/\text{COD} \) ratios or mean ratios.

In summary, the procedure consists of the following steps.

1. Calculation of shape and scale parameter for a given influent flow rate using available COD concentrations by:
   - classification of data in groups (see plots A–F);
   - calculation of Weibull parameters for every group (plots A–F);
   - generation of an exponential function for both Weibull parameters (plots H and I).

2. Filling the gaps (i.e. missing data) by generating COD concentrations with shape and scale parameter of the Weibull function depending on the inflow rate:
   - calculation of Weibull parameters for a given influent flow rate;
   - generation of a Weibull distributed random number (= COD concentration) with calculated Weibull parameters for the stochastic part of the generated data;
   - Calculation of missing \( N_{\text{tot}} \) and \( P_{\text{tot}} \) based on monitored \( N_{\text{tot}}/\text{COD} \) and \( P_{\text{tot}}/\text{COD} \) ratios or
   - Generation of a normally distributed random number for \( N_{\text{tot}} \) and \( P_{\text{tot}} \) regarding to existent distributions of \( N_{\text{tot}}/\text{COD} \) and \( P_{\text{tot}}/\text{COD} \) ratios, respectively.

The resulting concentration data for COD, \( N_{\text{tot}} \) and \( P_{\text{tot}} \) exhibit the same properties as the real data, while they have a stochastic part based on the Weibull-distributed random number generation that includes site-specific characteristics to create a realistic influent dataset.

### RESULTS AND DISCUSSION

This section focuses on step 2 (generation of COD concentrations) of the described methodology, while the calculation of \( N_{\text{tot}} \) and \( P_{\text{tot}} \) concentration using steps 3a or 3b is rather trivial and detailed results will not be presented in this paper but there is demand for some additional explanations. The generation of missed COD concentration data is based on a stochastic method resulting in a distribution that is fitted against real data.

For \( N_{\text{tot}} \) and \( P_{\text{tot}} \) also datasets exist that normally correspond to the days with COD measures. Now it is possible to calculate a fixed ratio of \( N_{\text{tot}} \) or \( P_{\text{tot}} \) to COD (3a) or to use a distribution of these ratios respectively (3b). In the first case, a comparatively small variance in the resulting \( N_{\text{tot}}/\text{COD} \) or \( P_{\text{tot}}/\text{COD} \) ratio is the final result. In case of use of

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Mean</th>
<th>SD</th>
<th>rel. SD</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow rate</td>
<td>m³ d⁻¹</td>
<td>118,487</td>
<td>44,305</td>
<td>37.4%</td>
<td>311,923</td>
<td>72,077</td>
</tr>
<tr>
<td>COD conc.</td>
<td>g m⁻³</td>
<td>552</td>
<td>150</td>
<td>27.2%</td>
<td>882</td>
<td>105</td>
</tr>
<tr>
<td>COD load</td>
<td>kg d⁻¹</td>
<td>60,132</td>
<td>11,680</td>
<td>19.4%</td>
<td>122,731</td>
<td>23,930</td>
</tr>
</tbody>
</table>

### Table 1

Statistical parameters of dataset from Figures 1 and 2 (n = 1,397 daily composite samples and daily mean flow rates)
distribution data of $N_{\text{tot}}/\text{COD}$ or $P_{\text{tot}}/\text{COD}$, a random value depending from the distribution has to be chosen. Using this second option (step 3b) in modelling studies with WWTP leading to fluctuating results in the effluent, especially in nitrogen fractions. A possible reason could be a relative strong difference within the $N_{\text{tot}}/\text{COD}$ ratio from one day to another. Therefore, a restriction should be integrated in the generation routine, but this is part of further investigation. From the current state it is recommend to use a fixed ration for calculation of $N_{\text{tot}}$ or $P_{\text{tot}}$ from generated COD concentration data.

The described method will be used to examine (i) the robustness of the approach using a complete dataset with various fractions of missing data and (ii) the universality of the approach on the datasets of three WWTPs of different sizes.

**Test of robustness**

In a first test, a complete dataset of a WWTP without gaps was used. Gaps were produced artificially by deleting every second or fifth concentration data point (remaining data fraction 50% and 80%). The remaining dataset was the starting point for data re-generation and gap filling using the proposed substitution method based on the Weibull distribution as described above.

The resulting cumulative COD loads are shown in Figure 3 (left) depicting the measured data (orange), two datasets based on the power function regarding the constant COD input load (black and grey), and two datasets applying the Weibull distribution (light and dark blue), each for remaining data ratios of 50% and 80%. In addition, an illustrative example of the relationship between COD concentration and inflow rate is plotted on the right-hand side of Figure 3 to show the visual effect of the Weibull method compared to the use of a power function.

Regarding the cumulative load distribution of the simple power function (Equation (1)), a significant deviation from the original dataset can be noticed at the lower and upper range of the influent COD load (Figure 3, left). In contrast, the cumulative function of the COD loads resulting from the Weibull distribution based substitution approach shows a significantly better agreement between measured and generated data which is due to the stochastic part in the Weibull distribution-based data generation.

In order to further analyse the deviating behaviour of both models, artificial datasets were generated with 10–90% remaining data. For comparison purposes, the mean, standard deviation and relative standard deviation of the original complete dataset is calculated to 80,905 kg COD/d, 16,095 kg COD/d and 19.9%. Some statistical parameters of generated datasets can be found in Table 2. While the mean COD load is nearly identical for both methods, the other measures show a better fit for Weibull-based method.

The comparison between the relative deviations of the various generated datasets from the measured data, where the generated datasets were produced applying the power function and the Weibull distribution function, is plotted in Figure 4, upper graph. It can be seen that the generated data based on the Weibull-based approach has a similar relative deviation as the measured values, which is approximately 20%, irrespective of the degree of missing data. The use of power function instead leads to a much smaller relative deviation of 10.2–18.7% depending on the...
extent of missing data, which does not sufficiently reflect the characteristics of the original dataset. The indirect dependence between an increasing data gap and the decreasing relative standard deviation can be explained by the convergence to the datasets' mean with the constant mean input load.

As another indicator for the quality of the generated data, the Coefficient of Efficiency CoE (Nash & Sutcliffe 1970) was adapted as a measure for the agreement of the shape between the cumulative distribution of measured and generated datasets referring to the data from Figure 3. The fit is better for values of CoE near 1 and thus, again, the Weibull-based approach yields clearly better substitution of missing values (Figure 4, lower graph).

It can be seen that for both quality measures only minor differences between the two approaches exist up to a ratio of 30% missing data. Beyond that value, the Weibull-based method shows an increasingly better fit due to a minor decrease in CoE and a nearly constant relative deviation close to the one of the measured data (19.9%).

Because of the stochastic nature of the Weibull method, the fluctuation range of different generated datasets is of great interest. Therefore, a set of 50 datasets was filled with the Weibull-based method and sorted according to their cumulative distribution. For every generated data point, the relative deviation between the 50 datasets was calculated. Overall, the relative deviation increases from 0.5% to 1.14% (Table 3), depending on the substituted fraction of the dataset (from 10% to 90%, as considered in Figure 4). This fluctuation range is very small.

It is shown that the Weibull-based methodology is better suited than the classical power function approach for a wide range of extent of missing data. The resulting influent load characteristics exhibit the same stochastic characteristics as the real measured data and proves to be closer to reality than the data where gaps were substituted using mean load data.

**Universality of the approach**

The second test for the Weibull-based method was the application to three WWTPs of different size and a significant degree of missing data points to provide information on whether the proposed approach is universally applicable. Table 4 provides some plant details like size and some information on the available datasets.
The fraction of missing data amounts to more than 50% in all plants and a large amount of data has to be generated for a continuous influent dataset. Results of the application of the Weibull-based method are presented in Figure 5.

At the left-hand side of Figure 5, the cumulative distribution functions show that for all three WWTPs a good agreement between the measured data and the Weibull distribution based method was achieved. The simpler power function-method only returns an approximate fit to the measured data, which is supported by the corresponding statistical parameters on the right-hand side of Figure 5. In some cases, the generated mean COD load of power function method is slightly better than for Weibull-based method (e.g. WWTP 2). Main interest of the method is not the mean value but the distribution and the variance. The variance in influent COD load cannot be described well by the power function method. Due to the substitution of missing data applying the constant mean COD load, the calculated variance after the gap filling procedure is about half of the real variance in measured data. The coefficient of efficiency also shows a better agreement between the distribution of the measured data and the Weibull-generated data.

The application of the power-function based data substitution, e.g. for dimensioning of aeration equipment, will lead to serious errors in the low and high range of the required air input due to underestimation of the real variance. Hence, the use of the Weibull distribution is a suitable and universal method to generate missing data for specific catchment characteristics and use them, e.g. for activated sludge modelling.

The major advantages of the developed method compared to simpler substitution approaches at one side and complex influent generation methods requiring additional

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Increase of mean relative deviation in random-generated datasets with different ratio of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of missing data</td>
<td>10%</td>
</tr>
<tr>
<td>Mean relative deviation within 50 generated datasets</td>
<td>0.49%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Size and COD-dataset characteristics of considered WWTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (p.e.)</td>
<td>Measured data</td>
</tr>
<tr>
<td>WWTP 1</td>
<td>20,000</td>
</tr>
<tr>
<td>WWTP 2</td>
<td>285,000</td>
</tr>
<tr>
<td>WWTP 3</td>
<td>670,000</td>
</tr>
</tbody>
</table>

Figure 5 | Cumulative distribution functions of influent COD load of the considered WWTP (left) and statistical parameters (right).
measurements at the other side can be summarized as follows.

**Based on available routine data**

Depending on the size and corresponding regional legal regulations, only a limited number of daily influent samples are taken and analysed, especially regarding WWTP with less than 50,000 p.e. Mostly, a continuous influent quality dataset is not available. In Table 5, information on the number of WWTP of different dimensions are provided by the European Environment Agency (EEA 2014). It is evident that only 12% of WWTP have a capacity for more than 50,000 p.e., where detailed routine data can be expected. For process optimisation, design of technical equipment as well as process modelling, a continuous and reliable dataset on the influent quality is necessary. The application of the developed method for influent quality generation allows overcoming information gaps due to the shortage of analytical data. Then it is already possible to use modelling tools for WWTP starting with a plant size of 10,000 p.e., representing a large part of WWTP in EU.

**Simple and easy-to-use method**

The application of the described method does not require additional data beside the available routine influent quality data and inflow measurements. This is an advantage to more powerful but also more complex tools as presented, e.g. by Gernaey et al. (2011), Devisscher et al. (2006) or De Keyser et al. (2010). The calculation is based on known statistical measures and can be performed with standard office software or special mathematical tools.

**Realistic site-specific characteristics of influent load**

Instead of a simple substitution of missing data with mean values, the described Weibull distribution based method implements a stochastic part that is fitted to the available measured data. Although no additional site-specific information is needed for the procedure, the generated data exhibit a similar variance as the real data. Special attention should be paid to the weekly distribution of available data. For a large number of WWTP no quality data were available from weekend days. Hence, it must be verified that no significant difference exists between the load pattern at weekends and working days. Otherwise, the method has to be processed twice – once for working days and once for weekends (only possible in case of available measuring data), respectively.

In contrast to the cited grey-box models for influent generation, the described method is only suited for reconstruction of existing datasets but not for scenario generation with varying conditions (e.g. population changes, climate change influences). This handicap is not only valid for the described Weibull-based method but also for other black-box models too. Within a complex modelling study with long-term predictions, it may be suited to use a grey-box influent generator but in case of work with an existing WWTP and the available routine data for modelling of existing time series the Weibull-based method should have an advantage.

**CONCLUSIONS AND OUTLOOK**

The method described in this paper is based on commonly available routine data and can be used for generation of a continuous influent quality dataset of an existing WWTP. For calculations, only the daily influent flow rates and some quality measurements (e.g. COD) are needed. All other required data can be generated even from relatively scarce data by analysing the data distribution and approximate them with Weibull distribution functions.

The resulting dataset has the same variance as the measured data and is, therefore, well suited for any process modelling or plant design approach that will need continuous influent data.

There are some relevant options for application of this method. For all kinds of modelling of the activated sludge process, the influent COD load is needed. A continuous dataset can be created with the available measured data mixed with generated ones to substitute missing daily values. Depending on the time-scale of the modelling task, the daily-based data can be overlaid by a diurnal variation to generate a realistic influent variation at a time-scale of hours (e.g. Langergraber et al. 2008).

With such detailed influent quality data, modelling can be used as a tool for process optimisation, analysis of energy consumption of technical equipment (when coupled
with energy models), improvement of control strategies or analysis of general capability.

A general application in other regions worldwide is possible if minimum routine quality data recording exists that allows to calibrate the Weibull distribution in the described method. Then this method could be part of a preliminary study for consultants to get an overview of existing WWTP and their datasets, e.g. in other countries without comprehensive field work.

The presented method was also used for a systematic analysis of more than 30 WWTP routine datasets of a size between 10,000–800,000 p.e.

From the results, some empirical relationships can be derived to develop a method that could serve for a general substitution of measured influent quality data. This work is still in progress.

REFERENCES


First received 6 April 2016; accepted in revised form 20 September 2016. Available online 11 October 2016