Identification of industrial wastewater by clustering wastewater treatment plant influent ultraviolet visible spectra

David J. Dürrenmatt and Willi Gujer

ABSTRACT

A procedure is proposed which allows the detection of industrial discharge events at the inlet of a wastewater treatment plant without the need for measurements performed at the industry, for special equipment and for exact knowledge of the industrial sewage. By performing UV/Vis measurements at the inlet of a plant and analyzing them with a two-staged clustering method consisting of the self-organizing map algorithm and the Ward clustering method, typical sewage clusters can be found. In an experiment performed at a mid-sized Swiss plant, one cluster of a cluster model with five clusters could be attributed to an industrial laundry. Out of 95 laundry discharging events measured in a validation period, 93 were correctly detected by the proposed algorithm, two were false positives and five were false negatives.

Key words | data mining, industrial wastewater, self-organizing maps, UV/Vis photospectrometry, Ward clustering

INTRODUCTION

Industrial wastewater can have a significant impact on the performance of a wastewater treatment plant (WWTP). Discharged at times when the plant is running at full capacity, e.g. during peak hours, industrial sewage can cause overloading and thus exceeding effluent concentration constraints.

WWTP operators are generally not provided relevant information by the industrial sites in the catchment area, neither on the type of sewage, the temporal pattern of discharging, nor on the exact load which is released to the sewer system, although this information is important for optimal plant control. Analyzing WWTP inflow with focus on the detection of industrial dischargers and the attribution of a discharge event to its producer, on the other hand, is difficult.

Methods for the detection of unusual changes in the wastewater composition and abnormal influent characteristics can be found in the literature. Langergraber et al. (2004), for example, present a method to generate alarm parameters from measured influent UV/Vis absorption spectra. Although their method can be used for early warning, it does not directly provide information on the discharger. A method to search the sources of wastewater which inhibits nitrification is given by Kroiss et al. (1992); the method however requires sampling of the industrial sewages. This paper, in contrast, proposes a two-staged clustering approach which uses influent UV/Vis absorption spectra only to reveal information on the wastewater producers in a catchment area and on the pattern of their discharging, which is novel. Clustering is a fundamental task in data mining and aims at grouping data instances into subsets such that similar instances are grouped together (Rokach & Maimon 2005). Using a different approach to cluster UV/Vis spectra measured at a fuel park WWTP, Lourenço et al. (2006) show that information can be extracted from the spectra and can be used for qualitative monitoring. However, their sample size was small and the spectra were sampled at several locations within the plant, thus exhibiting clear differences.

The proposed approach consists of a self-organizing map (SOM) which is used to generate a smaller but still representative data set using preprocessed UV/Vis absorption spectra. In the second step, the reduced data set is clustered
by Ward’s hierarchical agglomerative clustering algorithm and the clusters are manually labeled. The installation of an UV/Vis sensor at the WWTP inlet is sufficient (and may serve other purposes) and there is no need for the industry to install special measuring equipment, data storage and transmission infrastructure. The deployed clustering model detects and distinguishes different sewages and attributes them to producers.

MATERIALS AND METHODS

In-situ UV/Vis photospectrometry

UV/Vis photospectrometry measures the absorption of light from the ultraviolet to the visible range. Although there are methods which allow the extraction of quantitative information on concentrations of chemical compounds which absorb in the given wavelength range, they cannot directly be applied for wastewater analysis because of physical or chemical interference (Thomas & Cerda 2007). However, taking into account that a UV/Vis absorption spectrum is unique for a certain sewage composition, it can be considered as a fingerprint and be used for further analysis.

The spectra for this study were recorded with a submersible spectrometer probe (spectro::lyzer by s::can Messtechnik GmbH, Vienna, Austria) with 5 mm optical path length which measures the turbidity compensated absorbance between the wavelengths of 200 and 742.5 nm in 2.5 nm steps (for more information on the sensor, see e.g. Langergraber et al. 2003).

Site description

This experiment is performed in a Swiss community with about 20'000 population equivalents. The average discharge originating from the catchment at the WWTP inlet is 80 L/s. An industrial laundry is situated in the catchment approximately 1 km upstream of the plant, which corresponds to an estimated sewer flow time of 26 minutes. During the irregularly occurring discharge events, 10 L/s of laundry sewage are pumped into the sewer system for a variable amount of time (18 minutes on average). The composition of laundry wastewater is compared to domestic wastewater in Table 1.

Table 1 Characterization of domestic and laundry wastewater (sampled on 25th May 2010 12:00)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Laundry</th>
<th>Domestic</th>
</tr>
</thead>
<tbody>
<tr>
<td>COD [mg/L]</td>
<td>1970</td>
<td>800</td>
</tr>
<tr>
<td>Soluble COD [mg/L]</td>
<td>1556</td>
<td>290</td>
</tr>
<tr>
<td>NH4-N [mg/L]</td>
<td>1.1</td>
<td>25</td>
</tr>
<tr>
<td>NO3-N [mg/L]</td>
<td>2.8</td>
<td>0.3</td>
</tr>
<tr>
<td>NO2-N [mg/L]</td>
<td>0.45</td>
<td>0.04</td>
</tr>
<tr>
<td>PO4-P [mg/L]</td>
<td>0.5</td>
<td>2.8</td>
</tr>
<tr>
<td>pH [-]</td>
<td>8.1</td>
<td>7.2</td>
</tr>
<tr>
<td>Conductivity [mS/cm]</td>
<td>1.5</td>
<td>1.3</td>
</tr>
</tbody>
</table>

In the process information system of the plant, data of common operating parameters are readily available. In a measuring campaign from 1st June 2009 00:00 to 30th June 2009 24:00, a total of 51,614 UV/Vis absorption spectra were recorded after the fine screen (6 mm) with a sampling interval of 1 minute.

At the laundry, the discharge events were recorded in a measuring campaign from 2nd June 2009 00:00 to 12th June 2009 12:00. To be used for model validation, the events measured at the laundry must be synchronized with the events detected at the WWTP by taking into account the flow in the sewer. This is done by adding the estimated mean sewer flow time to the recorded events.

Two-stage clustering approach: self-organizing maps and ward clustering

UV/Vis data has high dimensionality, is subject to significant noise and suffers from outliers, this makes clustering difficult. Thus we have chosen a two-staged approach which has proven powerful especially when dealing with “noisy and messy” data (Vesanto & Alhoniemi 2000; Canetta et al. 2005). The approach is illustrated in Figure 1.

Data preparation

Because the concentrations of the domestic sewage and the laundry wastewater vary over time, a preprocessing method has been developed which reduces dilution phenomena of a UV/Vis fingerprint while emphasizing its characteristic shape. This is achieved by normalizing the absorption spectra $a(\lambda)$ (absorbance $a$ at wavelength $\lambda$) to have zero mean and unit variance and shifting it so that all measured absorption rates of wavelengths greater than a wavelength $\lambda_c$ are aligned.
Mathematically speaking, spectra are transformed by the non-intuitive equation

\[
\tilde{a}(\lambda) = \frac{1}{\sqrt{\text{var}(a)}} \left( a(\lambda) - \frac{\int a(\lambda) d\lambda}{\int_\lambda d\lambda} \right).
\]

After transformation, all \( \tilde{a}(\lambda) \) with \( \lambda \geq \lambda_c \) are cropped.

The preprocessed spectra are split into a calibration set, which contains the UV/Vis data from 13th to 30th June 2009 (no laundry data available) and a validation set, which contains the UV/Vis data and laundry discharge data from 2nd to 12th June 2009: The former is used to build the model and to detect and label the clusters (see below), the latter to assess the model performance.

**Self-organizing maps**

The self-organizing maps are a variant of artificial neural networks based on unsupervised learning, originally proposed by Kohonen (2001). They learn to cluster groups of similar input data in a non-linear projection from a high-dimensional data-space onto a lower-dimensional discrete lattice of neurons on an output layer (feature map, cf. Figure 1) in an orderly fashion (Kalteh et al. 2008). This is done in a topology preserving way which means that neurons physically located close the each other have similar input patterns. Additionally, the SOM is tolerant to noise which is especially helpful when dealing with experimental measurements.

Each neuron has assigned a prototype vector having the same dimensionality as the input data. During training, these vectors are optimized in order to represent the whole input data; the set of prototype vectors is therefore representative for the data set. The optimization of the prototype vectors is proportional to a learning rate and a neighborhood function, both monotonically decreasing during the ordering process. The former is a scalar; the latter forms a smoothing kernel around the prototype vector and makes sure that only input vectors within a certain neighborhood affect the prototype vector. Both are important to allow a regular smoothing.

The prototype vector \( w_i \) of neuron \( i \) which has minimum distance to an input vector \( \tilde{a}_t \) is the winning neuron for this...
input vector and is called the best-matching unit (BMU). The distance between the input vector and its BMU is called quantification error and given by the following Euclidian distance:

\[ q.e. = \| \tilde{a}_t - w_t \| \]

In this study, the software SOM toolbox for Matlab (Vesanto et al. 2000) was used to train and evaluate the SOMs.

**Ward clustering**

When applying the hierarchical agglomerative Ward clustering method (Ward 1963) on the SOM prototype vectors, each vector first forms its own cluster. Then, subsequently, the two clusters with minimum Ward distance are merged (the Ward distance between two clusters is defined as amount of variance added when two clusters are merged). The aim is to have small variance within a cluster and high variance between the clusters.

The optimum number of clusters is estimated by taking into account the Davies-Bouldin index (db-index; Davies & Bouldin 1979), which is the averaged similarity between each cluster and its most similar one (Halkidi et al. 2002). Because clusters with minimum similarity are aimed, the db-index is minimized.

The task of the expert is it now to identify the cluster in which the UV/Vis spectra of laundry/domestic wastewater mixtures lie (hereafter named “laundry cluster”).

**RESULTS AND DISCUSSION**

**Clustering model**

The clustering model given in Figure 2 was trained following the approach illustrated in Figure 1 using the preprocessed spectra of the calibration set (17,215 measurements, 54% of the data set, cropped at \( \lambda_c = 324 \) nm). A SOM with a two dimensional feature map (the neurons are arranged in a hexagonal grid) with map size \( 50 \times 13 \) was trained using a Gaussian neighborhood function with \( \sigma_t \) linearly decreasing with time from 7 to 1 and an inverse learning rate function with \( \alpha_0 = 0.5 \) (see Figure 1). Please note that for all distance measures, the Euclidian distance was used.

The topographic error (the proportion of all data vectors whose first and second BMUs are not adjacent vectors, cf. Kohonen 2001) of the SOM is 0.064 and the average quantization error 0.153. The Ward algorithm was applied for a cluster size of five, which had the lowest db-index of 0.67.

**Laundry cluster detection**

In order to find the cluster which contains BMUs for the UV/Vis spectra measured when laundry wastewater is being discharged, the use of three visualizations was advantageous:

1. The plot of the cluster centroids (cf. Figure 2c) is helpful to detect clusters significantly deviating from the others and having a shape consistent with available prior knowledge. Cluster 2 deviates from the others.
2. Considering a time series plot showing the particular cluster overlaid with the measured inflow parameters

![Figure 2](https://iwaponline.com/wst/article-pdf/63/6/1153/445541/1153.pdf)
(Figure 3a), one would again select Cluster 2 (higher temperature, slightly elevated discharge).

3) A ring map revealing the periodicity of the discharging (Figure 3b) exhibits an obvious, albeit irregular, operational schedule of Monday-Friday 7:00 to 23:00 and frequently of Saturday 7:00 to 12:00 for Cluster 2, which corresponds to the production schedule of the laundry.

As a result, one can say that laundry discharging events are contained in Cluster 2 with high probability. It is important to notice that the quantification error plotted in Figure 3a remains small during most of the detected events which shows that the UV/Vis spectra measured during an event are appropriately represented on the SOM.

To set up the model and select the associated cluster no quantitative information on the time and duration of the discharging events is needed. In the next section, the clustering model will be validated using quantitative information recorded at the industrial site.

Given the case that the cluster cannot be identified, heavy dilution effects, insufficient absorption of characteristic compounds or interference due to mixtures of several similar sewages could be hindering reasons. In bigger catchment areas where there are many industries with interfering spectra the integration of other operating data or mounting the measuring devices upstream of the plant might help.

Model performance

The model performance is assessed by evaluating the clustering model with the validation data set (14,399 spectra, 46% of the data set) and comparing the time periods when Cluster 2 is detected with the measured discharging events at the laundry shifted by the mean flow time in the sewer.

The clustering model predicts a total of 95 events. Comparing these to the 98 events measured, 93 were assigned correctly, two were false positives and five were false negatives (cf. Figure 4a). This corresponds to a failure rate of 7%.

The average duration is 18.0 minutes for the measured and 20.8 minutes for the predicted events. The error of the predicted event duration is thus $2.8 \pm 5.9$ minutes (mean ± standard deviation); the error of the predicted event start is $0.9 \pm 2.0$ minutes and $3.7 \pm 3.6$ minutes of the event end (the distribution is given in Figure 4b). Some of this error may be caused by neglecting the variable flow velocity and dispersion in the sewer when shifting by the mean sewer flow time.

Detection limit

The detection limit of the clustering model for the detection of laundry discharge events depends on the dilution of laundry and domestic sewage at the WWTP inlet.

The maximum dilution of laundry sewage which can still be detected by the clustering model was experimentally estimated by measuring and clustering the UV/Vis spectra of different dilutions of domestic and laundry sewage (both grab-sampled on 25th May 2010 12:00). Dilutions which contained more than 7% laundry sewage were detectable. This corresponds to a minimum fraction of the COD load of 12%. Comparing the discharge of 10 L/s of the laundry with the discharge at the inlet of the WWTP which is 124 L/s (85% percentile), one can conclude that the majority of the discharging events are detectable.
It was observed that dilution caused by rainfall events only has minor effects on the model performance. This can be explained: Although pollutant concentrations are lower during rainfall events, the ratio between domestic and laundry sewage approximately remains the same. Five laundry discharging events out of six which occurred during three rainfall events (WWTP inflow greater than 200 L/s) in the validation period were detected correctly.

Long term validity

To ensure long term validity, it must be ascertained that the conditions under which the clustering model was calibrated remain constant. That is, the quality of the laundry sewage does not change, e.g. due to changed processes, and that there is no other interfering sewage originating from the catchment area which exhibits a similar UV/Vis fingerprint and activates a BMU in the same cluster on the feature map of the SOM.

To detect possible changes, it is advisable to i) track the quantification error during clustering (high quantification errors indicate novel/abnormal patterns) and to ii) recompute the cluster model regularly and check whether the number of clusters remains the same and the clustering results are similar (disappearing and appearing clusters indicate changes in the catchment area). Recomputing can be highly automated and does not need additional measurements to be performed. However, when changes are detected and validation is advisable, additional measurements are required.

POTENTIAL APPLICATIONS

The authors see four practical tasks for which the proposed method is helpful: (1) The localization of an unknown sewage producer (by inferring rules from the clustered time series and by estimating sewer flow distance by analyzing the trajectories on the feature map), (2) the verification of legal compliance (i.e. if a producer violates agreed discharging loads), (3) the implementation of source related cost-allocation relying on detected events and (4) the use of the SOM to detect abnormal sewage compositions (by analyzing the quantification error).

CONCLUSIONS

Using a robust and reliable UV/Vis sensor mounted at the inlet of a WWTP in combination with the proposed two-staged clustering approach (SOM in combination with Ward’s clustering algorithm), it is possible to detect and distinguish different sewage compositions, thus reveal information about the catchment area. If the sewage types can further be linked to their producers the temporal pattern of discharging can be visualized and used for further investigations. In the given example, the approach generated a SOM with five characteristic clusters, of which one could be assigned to an industrial laundry. Model validation revealed that out of 95 events of variable duration which occurred in 12 days, 93 could be detected with two false positives and five false negatives only.

It has to be stressed that some of the simplifications and assumptions which were justified in this example do not hold in more complex catchments. For instance, having two industrial sites A and B in the catchment which produce sewages whose UV/Vis spectra do not interfere, thus whose discharging can be differentiated by the SOM, one already has to identify four clusters: “A”, “B”, “A and B” and “none”.

Figure 4

(a) Measured and predicted events for the validation period (the width of the bar corresponds to the duration of the event; bold circles indicate false positives, triangles false negatives). (b) Distribution of the error when comparing the predicted with the measured events.
ACKNOWLEDGEMENTS

The WWTP staff is acknowledged for providing open access to their SCADA system and Michael Burckhardt for the provision of the laundry discharge data.

REFERENCES

Canetta, L., Cheikhrouhou, N. & Glardon, R. 2005 Applying two-stage SOM-based clustering approaches to industrial data analysis. *Production Planning & Control* 16(8), 774–784.


