Extracting Temporal Patterns for Contamination Event Detection in a Large Water Distribution System

Nicolas Cheifetz[1], Selma Kraiem[1], Pierre Mandel[1], Cédric Féliers[1], Véronique Heim[2]
[1] Veolia Eau d'Ile de France, Le Vermont, 28, Boulevard de Pesaro, Nanterre F-92751, France
[2] Syndicat des Eaux d'Ile de France (SEDIF), 120 Boulevard Saint-Germain, Paris F-75006, France
[1]*nicolas.cheifetz@veolia.com*

ABSTRACT

Monitoring water quality in a drinking Water Distribution Network (WDN) is typically based on several sensors that are deployed on different locations of the WDN. Each sensor measures one or multiple signals, and the real challenge resides in detecting significant contaminations by analyzing these water quality signals. In practice, some detection methods may fail due to a specific design to certain forms of contaminants as well as a generic formulation with model-free algorithms. Both cases might trigger false alarms or produce no detection when a contamination occurs. Moreover the problem is hardly addressed when the underlying hydraulic regime changes over time causing fluctuations in the detection statistics. Such variation can dramatically decrease the detection performance or result to misleading analysis. Events like source shifting or tank filling are normal situations and can occur frequently depending on the operating conditions. This paper aims to deal with the variability of the contamination events under various operational conditions in water networks. The procedure is fully data-driven and leads to the extraction of meaningful temporal patterns using various data analysis techniques. This methodology can be used as a preprocessing stage to improve the performance of any algorithm to detect contaminations. The proposed approach is illustrated on a large real-world network in France and a qualitative interpretation is given to highlight a better understanding of the hydraulic regimes over time.

Keywords: Water distribution system; Early warning detection system; Sequential pattern mining.

1 INTRODUCTION

In order to ensure a healthy water distribution in the network, it's primordial to detect any intentional and accidental contamination of public water supply [1]. The design of sensor-based contaminant warning systems (CWS) is a promising approach for the mitigation of contamination risks in drinking water distribution systems [2]. Traditional detectors are based on data-driven techniques to analyze the collected signals at each monitoring station independently [3] or after synchronization [4] using statistical, heuristics or machine learning methods.

Besides, the hydraulics conditions are hardly the same in operation (varying water sources, tank levels, etc.) which implies the emergence of changes in the water quality [5]. A quality monitoring system should not trigger alarms for such normal operating changes, by using hydraulic modeling for example [6]. It is classically assumed that the detector can discriminate the presence of a specific pollutant using some drinking quality parameters (e.g. free chlorine residual, conductivity, pH, turbidity, etc.) [7]. Here, we tackle a more general event detection problem in WDNs and we believe that extracting some prior knowledge would enhance the performance of any detector, like the event detection software CANARY [8] for instance.

This paper is organized as follows: the overall methodology is decomposed in two consecutive stages and the case study is described in section 2. First, the extraction of elementary motifs from water flow time series is introduced in section 3. Then, a Levenshtein distance is formulated to measure the difference between two sequences of elementary motifs in section 4. A density-based clustering algorithm is used to identify different groups of sequence using the dedicated distance and a procedure of parameter estimation based on a greedy algorithm is drawn. Two resulting temporal patterns are finally illustrated and a realistic interpretation is given in section 5. The article ends with a discussion about the achieved results and draws some future applications and prospects.

2 GENERAL METHODOLOGY AND DATA DESCRIPTION

The aim of this paper is to identify automatically the most representative operating periods in terms of water flow incoming and outgoing a Water Distribution System (WDS). A multi-stage methodology is designed to address this problematic, as illustrated in Figure 1. The first step consists in extracting elementary motifs from water flow time series. Each pattern characterizes a simplified hydraulic state defined by constant flow values, using a classical K-means algorithm. This multivariate discretization is used to compute a dedicated Levenshtein distance that compares pairs of pattern sequence. The DBSCAN algorithm [9] is finally used to regroup similar sequences and their medoids, called temporal patterns, are determined.

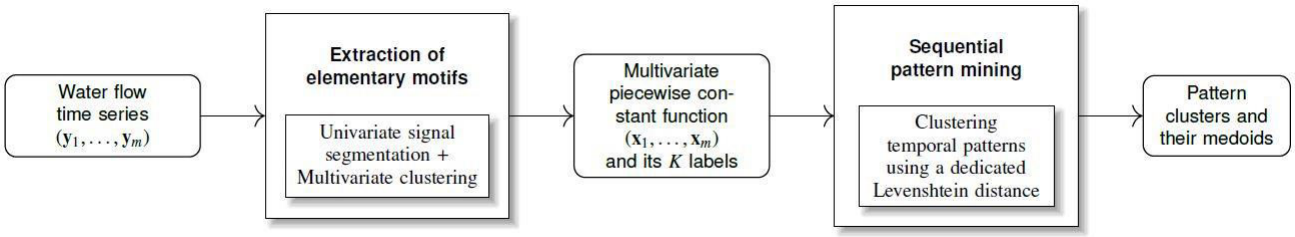


Figure 1. Block diagram describing the proposed methodology.

The proposed approach is illustrated on a large real-world network in France. The Syndicat des Eaux d'Ile-de-France (SEDIF) is an association including 150 municipalities that ensures the production and the distribution of drinking water to more than 4.5 million inhabitants of suburban Paris. The network of the SEDIF is the largest drinking WDN in France with about 8,600 km of pipes, almost 600,000 active connections and more than 750,000 m^3 of water produced each day. The water is produced in three large Drinking Water Treatment Plants (DWTP) located on the three main rivers of the Seine river basin, as shown in Figure 2. This paper is focused on a major part of the SEDIF network, mainly supplied by the Neuilly-sur-Marne DWTP and located on the Marne river. This subnetwork is depicted as the green area in Figure 2 and can be represented by a single hydraulic model including multiple sectors with different elevations. This hydraulic model is simplified as a system only characterized by water flows collected in 2015. As the SEDIF network is fully interconnected (e.g., large interconnections between the production plants, illustrated in Figure 2), the various operational conditions are strongly impacting the water propagation into the entire WDN. Indeed, any point into the network can be under the influence of multiple sources depending on its location and time.

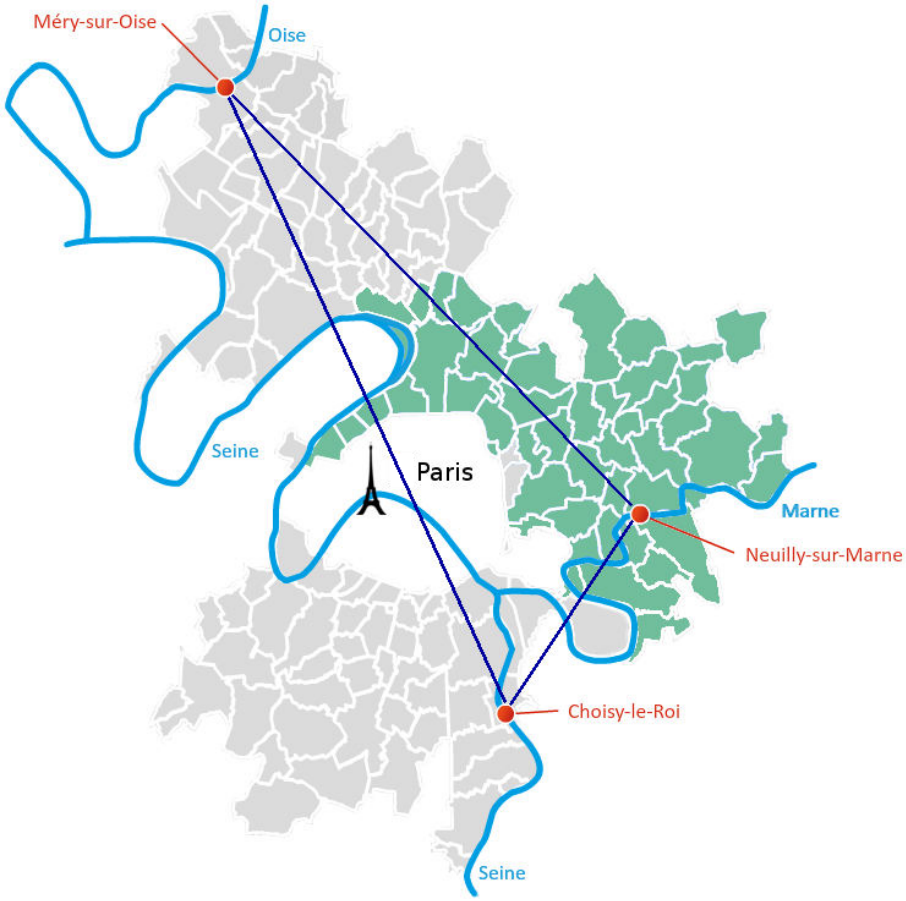


Figure 2. The SEDIF perimeter around Paris, the three main drinking water treatment plants in red and their interconnections in dark blue. The WDS studied in this paper is highlighted in green.

The paper presents a procedure to give an insight on the recurrent operating periods over a year from a single water distribution system. Both the pattern extraction and pattern mining methods are described in the following sections.

3 EXTRACTION OF ELEMENTARY MOTIFS

Let $\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_m\}$ denote a set of m time series, where each one of them $\boldsymbol{y}_j = \left(y_{1j}, \dots, y_{T_j j}\right)$ corresponds to water flows recorded at the border of the WDS. That is to say y_j is a univariate time series and $y_{tj} \in R$ is an incoming or outgoing flow. Note that no assumption is made about synchronization between time series and the production plant flow is omitted due to its value predominance and relative stability.

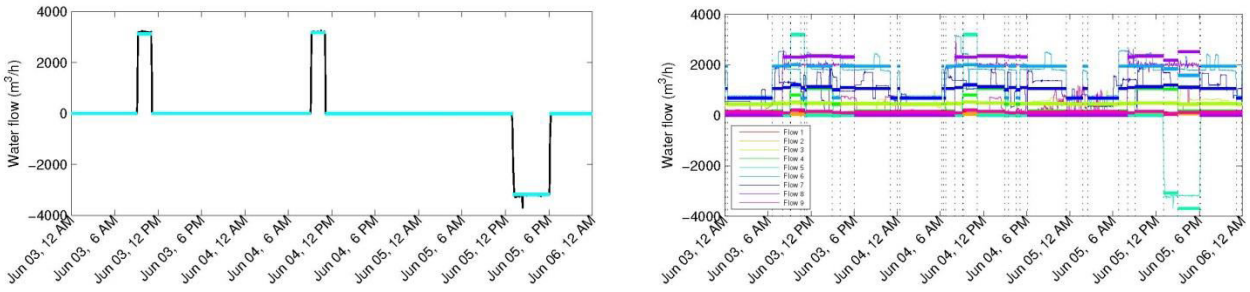
3.1 Univariate signal segmentation

The original time series are recorded with a fine granularity where the time step is 2'30 and present classical issues like noisy data and missing values. The dataset representing more than 200,000 points in a year per water flow, needs to be simplified using a piecewise approximation for instance. The TVD-MM algorithm [10] is used to denoise each time series independently while preserving the signal changes and aims to minimize the objective function

$$\sum_{t=1}^{T} |y_t - x_t|^2 + \lambda \sum_{t=2}^{T} |x_t - x_{t-1}|\,,$$

where $\lambda > 0$ is the regularization parameter, $(y_1, \dots, y_T)$ represents the original signal and $(x_1, \dots, x_T)$ is the smoothed signal. The higherλ, the smoothest the resulting signal. Note that the number of segments is not required and the segment values are modeled as constants. The method is notably suitable when the water flow signal can be approximated by piecewise constant functions as illustrated in Figure 3a.

Obviously, segmenting independently each flow signal is simpler than tackling a multidimensional water flow. As the WDN is considered as a system, the segmented time series are then aggregated into a single matrix $\boldsymbol{x}$ sharing all the change-points which can be seen as multivariate time series $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)$ over the same time grid $\{1, \dots, n\}$. In other words, this matrix is composed by n multivariate segments where each segment is an m-dimensional vector revealing m constant flows for a specific period, as shown in Figure 3b. In our case, this segmentation step allows to reduce significantly the size of the overall dataset by a factor 6.



(a) Univariate segmentation. *(b) Multidimensional time series.*

Figure 3. Univariate segmentation (3a) using the TVD-MM algorithm over three days in June 2015 – the raw data of water flow are painted in black and the piecewise constant functions are in cyan. The segmented multivariate time series (3b) with change points as vertical dotted lines.

3.2 Multivariate clustering and model selection

A classical clustering method is performed on the multivariate time series $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)$. The well-known K-means algorithm [11] is applied using various random initializations and the partition with the lowest intra-cluster inertia is selected. The number of clusters K is usually assigned by minimizing some information criterion (e.g., AIC or BIC) but here no clear minimum could be found due to the large size of the data. Then, the K-value is selected by minimizing a penalized and weighted version of the intra-cluster inertia defined by $C = D + \gamma \nu_K log(n(m+1))$, where D is a distance defined by the following equation, $\gamma > 0$ is the penalization parameter and $\nu_K = mK$ is the number of free parameters.

$$D = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \delta_i^2 \varphi\left(\boldsymbol{x}_i, \boldsymbol{\mu}_{z_i | z_i = k}\right)^2} \quad \text{with } (\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \sqrt{\tfrac{1}{m}\textstyle\sum_{j=1}^{m} (x_{ij} - \mu_{kj})^2}\,,$$

and δ_i is the duration (in hour) of the segment i, z_i is the label of segment i and $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{km})$ is the mean centroid of cluster k. Note that the distance D and C are expressed in m^3 (facilitating the interpretation), and D is linearly correlated to the inertia optimized by K-means.

The next section describes a strategy to extract meaningful "patterns" or subsequences in the time series $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n)$ where each temporal centroid $\boldsymbol{\mu}_i \in \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ is called an "elementary motif". The sequential pattern mining method is based on a dedicated Levenshtein distance.

4 SEQUENTIAL PATTERN MINING

4.1 Computing a dedicated Levenshtein distance

Let us introduce a distance in order to quantify the difference between sequences of elementary motifs that represent the pattern instances. Moreover, a reformulation of the Levenshtein distance [12] is adopted because this distance takes into account each sequence order and the three single-character operations (insertion, deletion and substitution). Some other distances (e.g., Hamming distance) do not share these features. The distance noted L is based on the function φ defined in subsection 3.2; let us note $\varphi(\mu_k, \mu_l) = \varphi_{k,l}$ and $\varphi(\mu_k, 0) = \varphi_k, \forall (k,l) \in \{1, \dots, K\}^2$. Considering two patterns u and v, the Levenshtein distance is defined as $(\forall i = 1, \dots, |u|, \forall j = 1, \dots, |v|)$

$$\begin{cases} L(0,0) = 0 \\ L(i,0) = L(i-1,0) + \delta_{u_i} \varphi_{z_{u_{i-1}}, z_{u_i}} \\ L(0,j) = L(0,j-1) + \delta_{v_j} \varphi_{z_{v_{j-1}}, z_{v_j}} \\ L(i,j) = min\left[L(i-1,j) + \delta_{u_i} \varphi_{z_{u_{i-1}}, z_{u_i}}, L(i,j-1) + \delta_{v_j} \varphi_{z_{v_{j-1}}, z_{v_j}}, L(i-1,j-1) + Sub(z_{u_i}, z_{v_j})\right] \end{cases}$$

and the substitution cost is defined by

$$Sub\left(z_{u_i}, z_{v_j}\right) = \begin{cases} \delta_{u_i} \varphi_{z_{u_{i-1}}, z_{u_i}} + \delta_{v_j} \min\left(\varphi_{z_{v_{j-1}}, z_{v_j}}, \varphi_{z_{v_j}, z_{v_{j+1}}}\right) & \text{if } i = |u| \\ \delta_{u_i} \min\left(\varphi_{z_{u_{i-1}}, z_{u_i}}, \varphi_{z_{u_i}, z_{u_{i+1}}}\right) + \delta_{v_j} \varphi_{z_{v_{j-1}}, z_{v_j}} & \text{if } j = |v| \\ \delta_{u_i} \min\left(\varphi_{z_{u_{i-1}}, z_{u_i}}, \varphi_{z_{u_i}, z_{u_{i+1}}}\right) + \delta_{v_j} \min\left(\varphi_{z_{v_{j-1}}, z_{v_j}}, \varphi_{z_{v_j}, z_{v_{j+1}}}\right) & \text{otherwise} \end{cases}.$$

4.2 Clustering and extracting operating periods

A clustering algorithm exploiting the previous distance is used to aggregate similar patterns among the overall sequence of elementary motifs. It is worth noting that successive motifs with identical labels are merged. First, a sequence of p candidate patterns are enumerated according to some prior knowledge relative to the addressed problem. Then, the DBSCAN algorithm [9] groups candidate patterns in high density regions; that is to say, a similar pattern has a distance less than a given threshold $\varepsilon > 0$. This algorithm has a worst case complexity of $O(p^2)$ and does not require setting the number of clusters (unlike K-means). The estimation of the ε-value is needed and a greedy-like procedure is performed to identify few potential clusters, where each iteration is defined such as

1. Selection of the best pattern (depending on the addressed problem: most frequent, etc.);
2. Aggregation of candidate patterns similar to the best pattern instances (with a distance $< \varepsilon$).

Then, the DBSCAN algorithm is used on all the patterns identified by the greedy clusters. The final threshold ε is chosen such as the DBSCAN rate of good classification is maximized while its ε-value is the lowest. Note that pattern overlaping can occur between patterns of different clusters but not inside each cluster. Finally, the most meaningful operational periods are identified as the medoid pattern per cluster. The next part presents the results obtained using the proposed methodology in order to extract temporal periods for contamination event detection.

5 EXPERIMENTAL RESULTS AND DISCUSSION

Following the description of the case study in section 2, nine time series of water flow collected in 2015 are used for characterizing successive hydraulic states of a WDS with respect to water exchange with its adjacent hydraulic systems (two plants and six sectors). The number of elementary motifs is selected by minimizing a specific criterion based on the distance D defined by the equation in subsection 3.2. Figure 4 shows the evolution of this criterion according to several numbers of clusters and the minimum is reached when $K = 20$ (with $\gamma = 10^{-3}$) meaning that the average error of the piecewise approximation is about 100 m^3 at each time step per water flow.

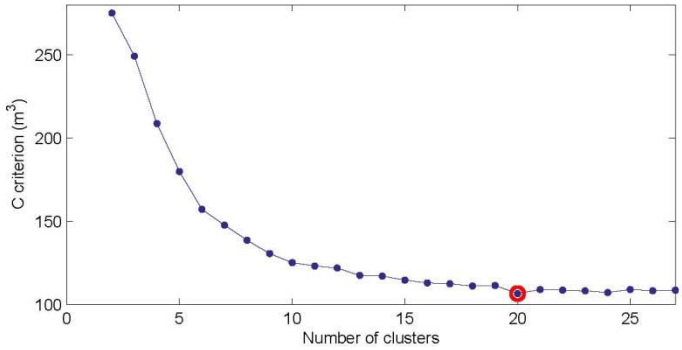


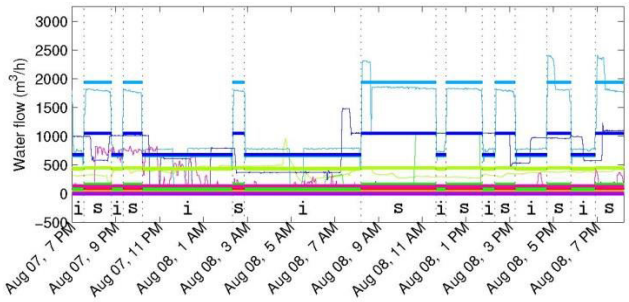
Figure 4. Evolution of a specific criterion selecting the number of elementary motifs.

The K-means result implies a partition of 20 different labels over the year, that is to say the definition of 20 elementary motifs occurring at different periods and explaining each time series. After merging successive identical labels, a sequence of 5,000 labels is built. Then, about 60,000 candidate patterns are enumerated using patterns composed of 7 to 17 consecutive labels where each pattern has one item with duration of about 24h. Indeed, the resulting operating periods are expected to be at least 1-day long in order to process extensive simulations of contamination using hydraulic models (calibrated over 24h). By selecting the most frequent pattern (with at least an instance of 24h long), the DBSCAN algorithm is set with $\varepsilon = 70$ using the greedy procedure defined in subsection 4.2. All the most representative patterns are finally obtained performing this algorithm on the symmetrical matrix of Levenshtein distance (of size 60,000 × 60,000).

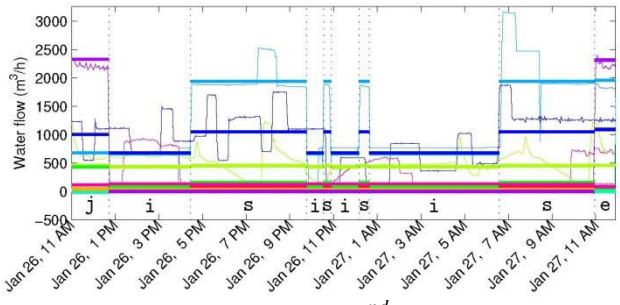
The medoids of the two most representative patterns are illustrated by Figure 5. The Subfigure 5a draws the first medoid which is defined by the sequence of labels ‘isisisisisisisis’, a succession of two elementary motifs ‘i’ and ‘s’. It lasts 25h12min early from the 7th to the 8th of August and its belonging cluster represent a cumulated duration of 42% in 2015. The motif s marks the intermittent significant flows that come from sectors 4 and 5. In addition, the medoid of the second cluster is drawn in 5b and its sequence of labels ‘jisisisise’ occurs from the 26th to the 27th of January for 1 day and 55 minutes. Its cluster represents a cumulated duration of 23% in 2015. The motifs j and e show the behavior of the WDS when it is fed by the plant 1.

For brevity, a short description of the four elementary motifs occurring in the two medoids is given in Figure 6. The motif ’i’ (Subfigure 6a) shows an incoming flow at about 700 m^3/h from sector 4 and 5, while the period ’s’ (Subfigure 6b) displays higher flows of about 2;000 m^3/h and 1;000 m^3/h respectively. Furthermore, the motifs ’j’ (Subfigure 6c) and ’e’ (Subfigure 6d) indicate flow

at about 2;300 m^3/h from plant 1 and 1;000 m^3/h from sector 5 - the motif 'ȷ' has a higher flow of 500 m^3/h from sector 4 while the motif 'e' shows a higher flow of 2;000 m^3/h from sector 6.

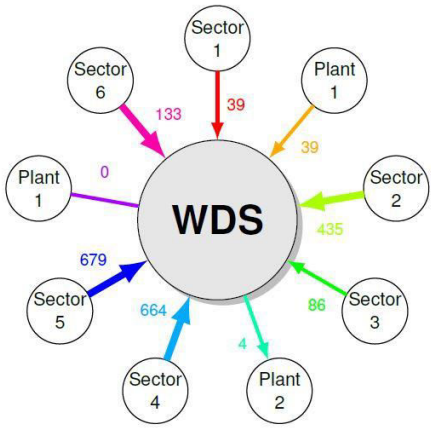


(a) Medoid of the 1st pattern.

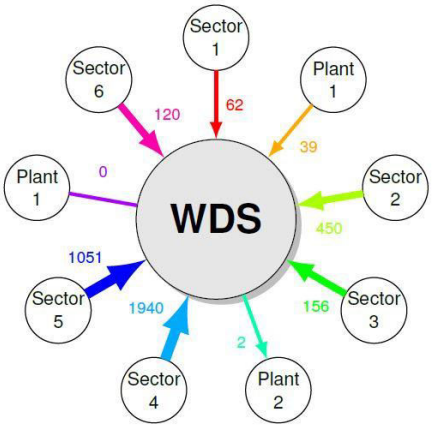


(b) Medoid of the 2nd pattern.

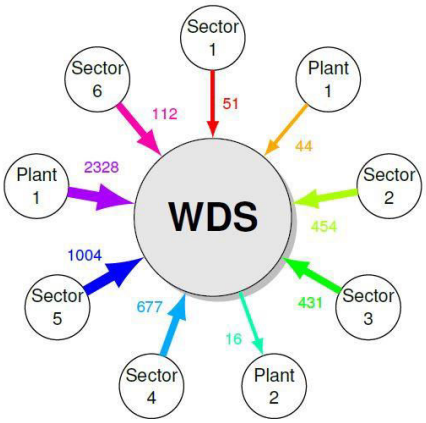
Figure 5. Medoids of the two most representative patterns.



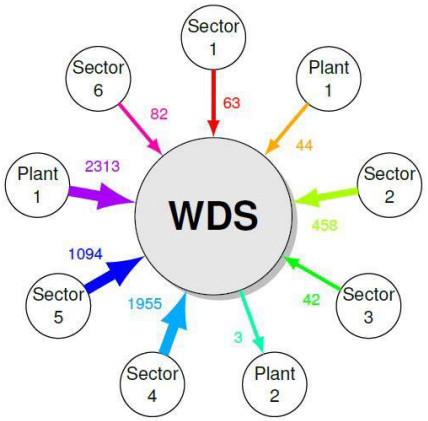
(a) Elementary motif 'i'.



(b) Elementary motif 's'.



(c) Elementary motif 'j'.



(d) Elementary motif 'e'.

Figure 6. Description of the elementary motifs appearing in the first two pattern medoids: arrows represent water flows (m^3/h) between the WDS and other adjacent hydraulic systems (two plants and six sectors). Each color is associated to a specific water flow: flow 1 in red, flow 2 in orange… up to flow 9 in mauve. Note that most of the differences between the four elementary motifs can be seen through water coming from the sector 3, 4, 5 and plant 1 (via the flow 8).

6 CONCLUSIONS AND FUTURE WORK

This article presents a general methodology to identify temporal patterns in a multidimensional time series. The proposed approach is applied to extract the most representative operational periods, defined as a sequence of constant multivariate flows incoming and outgoing a water distribution system. The two most representative temporal patterns are illustrated and a description of their motifs is shown according to the WDS. Based on real flow time series, the experimental results show a certain practicability in extracting operational periods. Future work will investigate a spatial segmentation on each relevant operational period using hydraulic simulations. Then, various EDS would be compared in terms of detection performance based on the proposed methodology.

Acknowledgements

The work presented in the paper is part of the French-German research project ResiWater that is funded by the French National Research Agency (ANR; project: ANR-14-PICS-0003) and the German Federal Ministry of Education and Research (BMBF; project: BMBF-13N13690).

References

[1] N. Sankary and A. Ostfeld, “Inline mobile sensors for contaminant early warning enhancement in water distribution systems,” *Journal of Water Resources Planning and Management*, vol. 143, no. 2, 2016.

[2] W. E. Hart and R. Murray, “Review of sensor placement strategies for contamination warning systems in drinking water distribution systems,” *Journal of Water Resources Planning and Management*, vol. 136, no. 6, 2010.

[3] X. Yang and D. L. Boccelli, “Bayesian approach for real-time probabilistic contamination source identification,” *Journal of Water Resources Planning and Management*, v. 142, 2013.

[4] C. Kühnert, M. Baruthio, T. Bernard, C. Steinmetz, and J.-M. Weber, “Cloud-based event detection platform for water distribution networks using machine-learning algorithms,” *Procedia Engineering*, vol. 119, pp. 901–907, 2015.

[5] M. Housh and Z. Ohar, “Integrating physically based simulators with event detection systems : Multi-site detection approach,” *Water Research*, vol. 110, pp. 180–191, 2017.

[6] N. Oliker, Z. Ohar, and A. Ostfeld, “Spatial event classification using simulated water quality data,” *Environmental Modelling & Software*, vol. 77, pp. 71–80, 2016.

[7] D. G. Eliades, D. Stavrou, S. G. Vrachimis, C. G. Panayiotou *et al.*, “Contamination event detection using multi-level thresholds,” *Procedia Engineering*, vol. 119, 2015.

[8] A. Leow, J. Burkhardt, W. E. Platten III, B. Zimmerman *et al.*, “Application of the CANARY event detection software for real-time performance monitoring of decentralized water reuse systems,” *Environmental Science: Water Research & Technology*, 2017.

[9] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *KDD*, vol. 96, no. 34, 1996, pp. 226–231.

[10] M. A. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak, “On total variation denoising: A new majorization-minimization algorithm and an experimental comparison with wavelet denoising,” in *IEEE International Conference on Image Processing*, 2006, pp. 2633–2636.

[11] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Symposium on Mathematical Statistics and Probability*. Univ. of California Press, 1967.

[12] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.