Quality Zones Automatically Identified in Water Distribution Networks by applying Data Clustering Methods to Conductivity Measurements

Pierre Mandel¹, Yue Wang², Anatole Parre³, Cédric Féliers⁴, Véronique Heim⁵

¹Veolia Eau d'Ile-de-France, 92000 Nanterre, France (corresponding author). E-mail: pierre.mandel@veolia.com

²Veolia Eau d'Ile-de-France, 92000 Nanterre, France. Permanent e-mail: yue.wang@eleves.enpc.fr

³Veolia Eau d'Ile-de-France, 92000 Nanterre, France. Permanent e-mail: anatole.parre@gmail.com

⁴Veolia Eau d'Ile-de-France, 92000 Nanterre, France. E-mail: cedric.feliers@veolia.com

⁵Syndicat des Eaux d'Ile-de-France, 75006 Paris, France. E-mail: v.heim@sedif.com

Highlights:

- 4-year operational conductivity measurements from 215 probes were studied
- Main characteristics of the network studied: 8500 km pipes, 4.6M customers
- Conductivity can be used to characterize water origin and water residence time
- Mixing zones and tank-influenced zones can be isolated with the proposed method
- Conductivity-based clusters offer a prior tool for contamination warning systems

Abstract: This paper presents a clustering study showing how conductivity measured every five minutes by 215 probes over four years can be used to determine specific quality zones for a large Water Distribution Network (WDN): 8500 km of pipes, 4.6M customers. Conductivity time-series are compared using Dynamic Time Warping. Then, probes are ordered using a density-based method, and probe clusters are extracted automatically. The clusters are a sound representation of water quality in the WDN, both in terms of water origin and water residence time. More specifically, zones directly impacted by plants or by external water imports, mixing zones and zones influenced by tanks, can be isolated and analyzed. Globally, 82% of the probes were found to be clustered, consistent with expert knowledge on the WDN operation; 13% were unclassified; 3% were erroneously clustered; and 1% seemed to be reasonably clustered, without any physical understanding yet. Besides providing users with an increased understanding of water quality in WDNs, conductivity-based clusters offer an interesting prior tool for contamination warning systems.

Keywords: Water distribution system, Water origin detection, Conductivity time-series, OPTICS clustering, Dynamic Time Warping, Automatic cluster extraction method

1. Introduction

1.1. Summarizing water quality in Water Distribution Networks: challenges and current approaches

Installing probes and improving hydraulic models are the two common approaches for enhancing the operational monitoring of water quality in Water Distribution Networks (WDNs). Both approaches provide operators with data that must be processed, in order to summarize how networks operate globally, and how this reflects on water quality. On the one hand, using probes to monitor and possibly detect abnormal values is generally limited to a few critical points treated independently, without network-wide fusion or a correlation approach (Lambrou et al. (2014); Riss et al. (2018)). Some authors, however, have proposed methods that integrate multiple probes, but these studies still raise challenges in terms of representativeness and extrapolation, both spatially and temporally (e.g., Oliker and Ostfeld (2015); Housh and Ohar (2017)). On the other hand, the description given by hydraulic model outputs is too detailed to convey simple, yet complete understanding of network operation, and more especially water quality (Perelman and Ostfeld (2012)). In particular, compiling modeled data is no easy task and can be imprecise due to modeling errors (Martínez-Solano et al. (2017); Housh and Ohar (2017)).

Graph representations of WDNs can provide a reliable means of analyzing, understanding and summarizing overall network operation. By introducing the notion of *strongly connected components* into graphs representing typical flows in WDNs, Perelman and Ostfeld (2006) proposed an adaptive aggregation procedure to preserve the dynamic tracer concentrations observed on critical nodes during an extended period of simulation. This procedure generates a specific simplified network when coupled with a classical hydraulic aggregation algorithm (Ulanicki et al. (1996)) for a given set of traced/contaminated nodes and critical/observation nodes, and can reproduce tracer concentrations. However, the simplified network depends on the set of traced and critical nodes. Subsequently, a clustering method simplifying WDNs was developed (Perelman and Ostfeld (2012)), by extending the notions of strongly and weakly connected components defined on graphs representing WDNs. The method was not intended to preserve hydraulic or quality results obtained with the original WDN, but to provide a unique model that summarized the operation of the original model. However, the analysis of the propagation of tracers in strongly and weakly connected components showed the limitations of using static simplified representations of WDNs in the case of dynamic contaminations (Lifshitz and Ostfeld (2018)).

Modeled quality outputs can be clustered to define homogeneous zones, without taking into account the graph structure of WDNs. Residual chlorine, water age, THMs, TAAs, TOC and BDOC, all modeled for the pipes of a WDN, were processed using first Self-Organizing Maps (SOM) to rearrange and represent the data, and then clustering algorithms such as k-means or fuzzy c-means were implemented (Chang et al. (2011)). By clustering traced concentrations of the different inputs of a large WDN, zones with the same dynamic water origin were defined (Mandel et al. (2015)). The zones were then validated by statistical studies of conductivity measurements. Similar tracer concentrations were clustered by (Qin and Boccelli (2017)) and more recently by (Ennouri et al. (2019)).

With different objectives, researchers use clustering algorithms to simplify experimental time-series measured in WDNs. The OPTICS clustering algorithm was applied to a feature matrix representing spatial and hydraulic characteristics to determine representative locations for pressure monitoring probes (Xie et al. (2019)). Focusing on the understanding of microbial mechanisms, other authors used SOM to report on how various microbial parameters correlate with modeled water ages and are influenced by temperature (Blokkeer et al. (2016)). In the field of hydrology, studies on travel times of water passing through a stream bed (Schmidt et al. (2012)) or flowing through sewer systems (Dürrenmatt et al. (2013)) highlighted the interest of treating time-series of conductivity with elastic matching, such as Dynamic Time Warping (DTW), to identify “batches of water”. Once resynchronized using DTW, signals can be clustered using classical algorithms as k-means (Dupas et al. (2015)). DTW can be used as similarity measure among measured pressure transients to reveal patterns across multiple sensing locations in a WDN (Xing and Sela (2019)). It may also be used to detect outliers, as illustrated with water demand time-series in a small WDN (Huang et al. (2018)). Recently, DTW was applied as a goodness-of-fit metric to measure the similarity between modeled and measured conductivity time-series obtained from experimental tracer experiments in a WDN (Woo et al. (2019)). However, to the best of the authors’ knowledge, operational conductivity measurements have never been considered previously in the literature as a univariate marker to define water quality zones and detect water origins in WDNs. Moreover, large-scale conductivity datasets have not yet been studied in the literature on WDNs.

1.2. Aims and general approach

The objective of this paper is to show how conductivity time-series measured by 215 probes over four years can be clustered to summarize water quality (both in terms of water origin and water residence time) in a large and complex WDN (8500 km pipes, 4.6M customers). To this end, a novel methodology relying on a density-based clustering algorithm (OPTICS) is proposed, and the resulting water quality zones are presented and discussed. This paper presents an approach complementary to that of (Mandel et al. (2015)), where modeled tracer concentrations of the different inputs of a WDN were clustered to define quality zones, and then validated using conductivity measurements. This study proceeds inversely: clusters are based solely on experimental conductivity measurements, and then explained and validated using hydraulic models and expert knowledge on the WDN’s operation.

The outline of the paper is as follows. The dataset and methodology are presented in section 2. DTW, used as similarity measure among conductivity time-series, is presented briefly. DTW post-treatments and cluster extraction methods are presented in detail, as they represent new developments. Probe clusters are presented and discussed in section 3. In particular, the statistical analysis of the strongest connections between couples of probes is discussed. Finally, the conclusion sums up the findings, discusses the potential application of this study to other use-cases, and provides potential prospects for the future.

2. Material and methods

2.1. WDN and dataset studied

The WDN studied belongs to the Syndicat des Eaux d'Ile-de-France (SEDIF). It is the largest drinking water service in France, providing water to 4.6 million inhabitants of the Paris metropolitan area. More than $7.5 \cdot 10^5$ m³ of water are produced daily and distributed in 37 District Metered Areas (DMAs) through 8500 km of pipes. The WDN is fed by surface water and groundwater, treated respectively by three drinking water plants (P_A , P_B and P_C) and four drilling wells (W_1 , W_2 , W_3 and W_4). P_A , P_B and P_C together account for around 95% of the water volume produced. 215 KaptaTM (Neraxis, Neuchâtel, Switzerland) probes, measuring active chlorine, conductivity, pressure and temperature every five minutes (McDougle et al. (2012)), are deployed in the WDN. The locations of the probes are represented in Fig. 1 by symbols and colors corresponding to the theoretical water origin (including mixing zones), as defined by the operators and hydraulic engineers. More specifically, the main zones fed by P_A , P_B , P_C and W_1 were defined by considering the extension of the DMAs ; the mixing zones involving W_2 , W_3 and W_4 were defined by modeled hydraulic tracing tests, as well as the $M_{P_A-P_B}$ zone. The two remaining zones $M_{P_B-P_C}$ and $M'_{P_B-P_C}$ correspond to specific DMAs, which can be fed either by P_B or P_C .

Conductivity, due to its conservative character, is regarded as a good indicator of water origin in WDNs (Woo et al. (2019)) and sewer systems (Dürrenmatt et al. (2013)), and so was selected as a "natural tracer" of water origin. Temperature data were not considered since 95% of the water produced originates from nearby rivers having approximately the same temperature (for example, the temperature differences of P_C with P_A and P_B exceed 0.5°C only 15 days per year in average). Moreover, water temperature can be affected by the surrounding soil temperature (up to +1.5°C in fall and -1.5°C in spring in our dataset). Hence, temperature could not be considered a reliable tracer. Temperature may nevertheless be used to refine certain mixing zones (especially for groundwater, which generally has high and constant temperature values). The dataset in this study gathered the conductivity measured by the probes and sensors located at the three plants from January 2015 to April 2019 (included), except for the five probes located in the mixing zone $M'_{P_B-P_C}$, which were installed in September 2018. The measurement acquisition period is 5 minutes. Drilling well W_3 was exploited until July 2017 while drilling well W_4 has been exploited since November 2018. Conductivity measured at the drilling wells was not taken into account as it is measured every month.

Fig. 1. Location of the probes installed in the WDN and sensors at drinking water plants. Paris is situated in the center of the network (see the Eiffel tower logo). P_A , P_B , P_C , W_1 zones: zones influenced by a unique water origin; M_{X-Y} zone: mixing zone influenced by water originating from X and Y. A simplified description of the average yearly conductivity profiles can be found in Appendix A.

2.2. Data pre-treatments

Few pre-treatments were applied to the data. Basically, missing values were replaced by NaN (Not a Number) values and "small" NaN gaps were filled using interpolation. All the gaps could not be filled using interpolation, as some of them were very long (several days); interpolating our data on such intervals would have yielded poor quality results. So, the maximal length of a gap to be filled by interpolated values had to be determined. The following procedure was developed: (i) two time-series with a complete data set were chosen, the first exhibiting frequent jumps due to alternating water origin, the second having a smoother profile; (ii) artificial gaps of various lengths, replacing original values with NaN, were created; (iii) artificial gaps were filled by different interpolation methods; (iv) methods were evaluated considering the error between original (i.e. true) and interpolated values. Consistent results were obtained using a piecewise cubic hermite interpolating polynomial on intervals smaller than 2h (95% of the interpolated points had a distance to the original value of less than $10 \mu S \cdot cm^{-1}$, which corresponds to the median of the minimal absolute difference between probes). Missing data represented 26.9% of the dataset before interpolation and 23.7% afterwards. Non-NaN data represented $71.9 \cdot 10^6$ values before interpolation and $75.1 \cdot 10^6$ afterwards.

2.3. Similarity measure using DTW

2.3.1. The method

In order to cluster the conductivity measurement of the probes, it is necessary to compute a similarity measure between time-series. Due to the propagation of water in pipes, comparing two time-series at the same dates

and times would serve no purpose. Instead, time-series have to be “resynchronized” beforehand. DTW is a method that aligns two time-series optimally by non-linearly warping the time-axis to maximize their similarity. Originally proposed for spoken word recognition (Sakoe and Chiba (1978)), this method has been used in a variety of fields, in particular those in which signals are susceptible to propagate such as music (Müller et al. (2006)), seismology (Li et al. (2010)), social network messages (Jendoubi et al. (2015)), hydrology (Dupas et al. (2015)), etc. The alignment of two time-series must comply with certain constraints:

- Continuity condition: each point of the first time-series must be aligned with one or more points of the second time-series, and vice versa;
- Boundary conditions: the first and last points of the first time-series must be aligned with the first and last points, respectively, of the second time-series;
- Monotonicity condition: the temporal ordering in time-series must be respected.

In addition, a Sakoe-Chiba band constraint specifying a maximal delay between two time-series (or sliding window) of $\pm$ one day was added. This constraint has two advantages: it decreases computational time by reducing alignment possibilities, and it focuses on neighborhood relations, by excluding the alignment of time-series over long periods of time. It is particularly adapted to a density-based clustering algorithm like OPTICS, which explores the dataset using neighborhood notions (see section 2.4.).

2.3.2. Use of DTW and post-treatments

Applying DTW to measured time-series must be done carefully as parts of some conductivity signals may be missing, exhibit noise, or offset differences up to $\pm 25 \mu\text{S}\cdot\text{cm}^{-1}$ due to the probe calibration procedure. Firstly, time-series having more than 20% of NaNs were considered as non-associable with others and given infinite DTW measures. To overcome the problem of noise, once resynchronized by DTW, time-series were compared using absolute differences instead of squared differences (Mandel (2010)). Offset issues were tackled by normalizing data beforehand; this practice is common when using DTW as the method seeks shape similarities (Keogh and Kasetty (2003); Shokoohi-Yekta et al. (2017)). Finally, the alignments proposed by DTW were post-treated (see Fig. 2 as illustration) to ensure that the results remained consistent with the operation of a WDN:

- Excessive time dilatations: alignments where one point of the first time-series is aligned with more than seven points of the second time-series were discarded as they do not have a physical background (this corresponds to alignments colored in black in Fig. 2). The same problem was observed for flat portions of conductivity time-series by (Woo et al. (2019)) and tackled by adding a Gaussian white noise to the original data;
- Extreme and isolated delays: once excessive time-dilatations have been removed, associations that have extreme (maximal or minimal) delay values representing more than 20% of the 5 nearest delays are recursively eliminated (this corresponds to alignments colored in green in Fig. 2). In particular, when at least 20% of delays between two time-series i and j are equal to the width of the Sakoe-Chiba band ($\pm$ one day), $\text{DTW}(i,j)$ is given an infinite value.
- Finally, if more than 70% of the points of a time-series have been eliminated by DTW post-treatments, the time-series is considered as non-associable with other time-series.

Fig. 2. Alignments proposed by DTW for two conductivity time-series (the blue one was shifted by $+300 \mu\text{S.cm}^{-1}$ and only 1/10 alignment is represented for the purpose of illustration). Insert: cumulative bar chart displaying the distribution plot of the delays calculated by DTW between the two time-series.

2.4. Clustering algorithm: OPTICS

2.4.1. The OPTICS algorithm

Ordering Points To Identify the Clustering Structure (OPTICS) is based on Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. (1996)), a density-based algorithm which clusters points that are close to each other and marks as outliers points that lie in low-density regions. Neighboring solutions are characterized by two parameters: a radius of neighborhood ϵ and a number of points required to form a cluster *MinPts*. OPTICS extends DBSCAN by considering a dynamic ϵ which adapts to clusters with different densities (Ankerst et al. (1999)), so that only *MinPts* remains as parameter. OPTICS proceeds by exploring the dataset in a linear order, beginning with an arbitrary point and always choosing as next point that closest to any of the already explored points. The Reachability Distance (RD) is computed for every point x ; it expresses the required density to form a cluster with x and one of the points already explored. The results are represented in an RP (Reachability Plot), i.e. a bar graph representing the RDs with the exploration ordering. Since points belonging to a cluster have low RDs, “valleys” in the RP are typical for clusters; the deeper the valley, the denser the cluster. The first point to be explored is given an infinite RD value; points that are not associable with other points are also given infinite RD values. *MinPts* was set to 1 to permit detecting small clusters.

2.4.2. Automatic cluster extraction methods

OPTICS does not produce an explicit clustering for the dataset, making it necessary to implement a method to extract clusters. Two methods were tested: the first, proposed by the research group that developed OPTICS, is commonly used in the literature; the second, proposed in this study, is an extension of the first.

1. Gradient-Clustering (GC) proposed by (Brecheisen et al. (2004)) computes the gradient of the RDs by considering three consecutive RDs in the RP. By introducing geometric calculations, the method determines all the inflexion points that are candidates as starting and ending points for the clusters. Two parameters have to be specified: w , the interval between two bars of the RP; and t (between 0° and 180°), the threshold angle for detecting inflexion points. In addition, we introduced *MaxPts*, a parameter specifying the maximal size of the extracted clusters. Optimal values for parameters ($w = 10^{-2}$, $t = 170$, *MaxPts* = 30) were determined by considering the silhouette index, using cluster

representatives as defined in (Brecheisen et al. (2004)). The detailed results on the determination of the optimal parameter values can be found in Appendix B.

2. *Extended-Gradient-Clustering (e-GC)* reduces the number of unclassified points by using a tree representation, based on the notions of predecessor and successor introduced by (Brecheisen et al. (2004)). This representation is richer than an RP since a point has only one predecessor, but may have several successors. Clusters determined by GC are kept and successors that are unclassified are colored using the same color as that of its predecessor. Constraints of depth and breadth of coloring, parameters D and B , were added to avoid excessive extensions of the clusters. An example is presented in Fig. 3. Optimal values of D and B ($D=1$ and $B=3$) were determined by considering the silhouette index, using cluster representatives as defined in (Brecheisen et al. (2004)). The detailed results on the determination of the optimal parameter values can be found in Appendix B.

Fig. 3. Tree-based cluster extractions, where colors represent clusters (dark blue circles represent unclassified probes). a) GC extraction; b) e-GC extraction with $D = \infty$ and $B = \infty$; c) e-GC extraction with $D = 1$ and $B = 3$.

2.5. Time segmentation and type of analysis

The 4.5 year time-series were segmented into 226 weeks before being processed by DTW (see second step in the flowchart presented in Fig. 4). On the one hand, DTW needs time-series with noticeable shape changes (for example, peaks) in order to work properly. Such changes may not be visible on conductivity time-series shorter than a day. On the other hand, DTW computations are time-consuming: even with a Sakoe-Chiba band constraint, the algorithm complexity is $O(n^2)$ when comparing two time-series of length n . Preliminary tests with five segmentations (year, month, week, day and hour) showed that hourly and daily segmentations yielded on average 15% and 6%, respectively, of null RDs, because of insufficient shape changes in conductivity. Weeks were found to be the smallest segments yielding consistent DTW values.

Weekly and overall analyses were conducted (see Fig. 4), applying OPTICS and GC cluster extraction methods respectively to the 226 weekly DTW similarity matrices, and to the average matrix of the 226 weekly matrices (i.e. the (i,j) component is the mean of all the 226 (i,j) components discarding infinite values). In order to summarize the weekly analysis, the 226 clustering results obtained on the weekly similarity matrices were processed statistically, by analyzing the frequency of the pairs of probes clustered together.

Fig. 4. Methodology flowchart

3. Results and discussion

3.1. Overall clustering results

Compared to the water quality zones identified by operators presented in Fig. 1, overall clusters offer a consistent yet richer analysis of water quality in the WDN. A spatial visualization of the GC and e-GC cluster extractions is shown in Fig. 5, where 21 probe clusters are defined for the WDN, based uniquely on conductivity measurements. Furthermore, a comparison of the distribution of probes in the zones presented in Fig. 1 and in the overall clusters is given in Table 1. Clusters obtained by e-GC are sound extensions of those obtained with GC; they are presented in the following:

- Clusters #1 to #5 are grouped in the P_A zone; clusters #1 to #4 differ from each other in terms of water residence time in the network, as revealed by the manual analysis of the noticeable changes in the conductivity signals from the different probes. Cluster #1 groups P_A and the probes directly influenced by the plant (this was confirmed by a modeled tracing test not presented herein). Cluster #2 groups more distant probes, including two entire DMAs of the P_A zone; however, one probe belonging to the M'_{PB-PC} zone is erroneously classified in cluster #2 (see Table 1). Cluster #3 corresponds exactly to a DMA of the P_A zone. Cluster #4 comprises three probes belonging to two neighboring DMAs. Cluster #5 groups all the probes located in a specific city. This grouping can be explained by the presence of a valve with the neighboring M_{PA-PB} zone which may partly feed cluster #5 with a particular blend of P_A and P_B .
- Clusters #6 to #9 are grouped in the P_B zone; clusters #6 and #7 differ from each other in terms of water residence time in the network (manual analysis of the conductivity signals). Cluster #6 groups P_B (defined as representative probe) and probes directly influenced by the plant; one probe of the M_{W4-PB} zone also belongs, erroneously, to cluster #6 (see Table 1). Cluster #7 corresponds exactly to a specific DMA. Cluster #8 groups four neighboring probes which are close to an exchange point with the neighboring M_{PA-PB} zone. The conductivity signals of the probes of cluster #8 exhibit the same baseline variations as the probes belonging to cluster #6 with regular drops ranging from $80 \mu S.cm^{-1}$ to $200 \mu S.cm^{-1}$ that can be observed from December to May, a period during which the conductivity levels of P_A and P_B are different (see Fig. A1 in Appendix A). Moreover, the chronological order of the drops observed on the four probes suggests that water is moving eastwards. This is confirmed by the fact that the drops correlate with the flowrate of the water coming from the neighboring M_{PA-PB} zone. Cluster #8 can therefore be considered as another mixing zone between P_A and P_B . The probes of cluster #9 are close to an exchange point with a neighboring WDS, where water is imported, with volumes ranging from $200 m^3.d^{-1}$ to $3000 m^3.d^{-1}$.
- Clusters #10 to #15 are grouped in the P_C zone; clusters #10, #12 and #13 differ from each other in terms of water residence time in the network (manual analysis of the conductivity signals). They are more difficult to characterize, as zone P_C is complex in terms of hydraulics (16 DMAs, compared to 9 and 5, for zones P_A and P_B respectively); however, the clusters are spatially consistent. Cluster #10 groups P_C and the probes directly influenced by the plant. Cluster #11 groups the probes from the

eastern part of the P_C zone; the probes grouped in this cluster may represent a mixing zone between P_B and P_C as $14,000 \text{ m}^3 \cdot \text{d}^{-1}$ are on average transferred from P_B to P_C . Cluster #12 groups, among others, 4/5 of the probes of a specific DMA. Cluster #13 groups three of the four probes of a specific DMA. Clusters #14 represents a small geographical zone for which no physical explanation can be given yet; however, its extension by e-GC seems questionable as shown by the statistical analysis of weekly clusters (section 3.2). The probes of cluster #15 are partly fed by W_2 as suggested by the presence of synchronous peaks in conductivity and temperature (W_2 water has a constant temperature of 25-26°C). Therefore, cluster #15 can be seen as a new mixing zone M'_{W_2-PC} .

- Cluster #16 corresponds exactly to the W_1 zone. The cluster extension proposed by e-GC is meaningful as the two missing probes of the W_1 zone are added at this stage.
- Clusters #17 to #19 correspond to the mixing zones identified in Fig. 1.
 - Cluster #17 gathers all the probes of the M_{PA-PB} zone, one probe of the M_{W_3-PB} zone, the two probes of the M_{W_2-PC} zone and an isolated probe of the P_C zone. As W_3 has stopped operation since July 2017, the M_{W_3-PB} zone cannot be identified easily during the period (January 2015-April 2019). The inclusion of the probes of the M_{W_2-PC} zone, most probably because their conductivity signals also exhibit abrupt changes due to the discontinuous operation of W_2 , and the isolated probe of the P_C zone, which is located directly on an exchange point with a neighboring WDS, are clear errors of the method (see Table 1).
 - Cluster #18 corresponds exactly to the M_{PB-PC} zone.
 - Cluster #19 gathers four of the five probes located in the M'_{PB-PC} zone (the latter being classified in cluster #2).
- Cluster #20 corresponds to a mixing zone between P_B and P_C . Both probes grouped in this cluster are close to the mains transporting water. As a result, they exhibit abrupt changes of conductivity as can be seen in Fig. 6.
- Cluster #21 gathers 29 unclassified probes, mostly located in zones P_C (20 probes), P_A (5 probes) and P_B (2 probes). Unsurprisingly, due to the unsteady operation of W_3 and W_4 , two probes belonging to M_{W_3-PB} and M_{W_4-PB} are unclassified (see Table 1).

Table 1. Matching matrix between the quality zones presented in Fig. 1 and overall clusters. The figures represent the number of probes (in red: probes clustered erroneously).

Zone Cluster#	P_A	P_B	P_C	W_1	M_{PA-PB}	M_{PB-PC}	M'_{PB-PC}	M_{W_2-PC}	M_{W_3-PB}	M_{W_4-PB}
1	20									
2	18						1			
3	6									
4	3									
5	3									
6		34								1
7		5								
8		4								
9		2								
10			24							
11			12							
12			10							
13			3							
14			3							
15			3							
16				5						
17			1		13			2	1	
18						6				
19							4			
20		2								
21	5	2	20						1	1

Globally, 177 (82 %) probes are found to be clustered consistently with expert knowledge of the operation of the WDN. 3 (1 %) probes seem to be reasonably clustered, yet without providing any physical understanding (cluster #14). 29 (13 %) probes are unclassified and 6 (3 %) probes are clustered erroneously (one in cluster #2, one in cluster #6 and four in cluster #17). Typical conductivity patterns for the clustered probes can be found in Fig. 6, where the conductivity signals of the representative probes are shown for a three-week period.

Fig. 5. Spatial extension of the 21 overall clusters found after GC and e-GC extractions.

Fig. 6. Example of conductivity measurements for the representative probes of the 20 overall clusters (April 2019); RP#X: Representative Probe of cluster#X

3.2. Results obtained using a non-warping measure between time-series

The benefit of using DTW as similarity measure between conductivity time-series has been evaluated by comparing the results presented in 3.1. to those obtained with a simpler non-warping measure, called *CL* in the following (*Constant Lag*). In order to compute the measure $CL(u,v)$ between two conductivity time-series u and v , the optimal time lag $T_{u,v}$ that minimizes the absolute value between the normalized time-series $\bar{u}$ and $\bar{v}$ is calculated (the formal definition of $T_{u,v}$ is given in Equation 1). Then, $CL(u,v)$ becomes simply Equation 2. As previously, similarity measures are computed weekly and then the average CL matrix is computed (see Fig. 4).

$$T_{u,v} = \underset{-1 \text{ day} \leq j \leq 1 \text{ day}}{\operatorname{argmin}} |\bar{u}(t) - \bar{v}(t + j)|$$

$$CL(u,v) = |u(t) - v(t + T_{u,v})|$$

The clusters based on the CL measure (see Appendix C for a detailed presentation of the results) give a broader description of the WDN operation, with larger main clusters and smaller secondary ones. The number of probes clustered erroneously is the same as when using DTW (6, see Table C1 in Appendix C) and several clusters are comparable (e.g., clusters #1 and 2) or equal (cluster # 9) to those presented in Fig. 5. Despite a larger number of unclassified probes (32 compared to 29), some zones presented in Fig. 1 are more precisely detected with the CL measure: clusters #17-2 and #19 correspond respectively to M_{W3-PB} and M'_{PB-PC} . However, the definition of several small clusters is questionable (clusters #14, 15-1, 15-2 and 15-3), as they do not reflect any known operational reality. Oppositely, DTW, by focusing on specific shape changes in the time-series (peaks, drops), yields better results when considering smaller mixing zones (clusters #5, 8, 11 and 15 of Fig. 5). In most cases, these small mixing zones are included in the main clusters (clusters #5, 11 and 15 of Fig. 5) or unclassified (cluster #8 of Fig. 5) when using the CL measure.

3.3. Statistical analysis of weekly clustering results

In order to summarize the clustering results obtained for the 226 weeks of the dataset, the statistical frequency of the pairs of probes clustered together was studied. The results are presented graphically in Fig. 7, where pairs of probes belonging to the same cluster more frequently than a threshold value are linked by a blue line. In Fig. 7, different values of the threshold were selected manually to permit the visual assessment of the intensity of the probes' associations, and define the cores of the clusters.

The three main zones associated with the plants can be seen with a relatively small threshold value. For a threshold value of 0.11, zones P_A , W_1 and M_{PB-PC} are totally separated from the rest of the WDN. There is no direct link between zones P_B and P_C . However, zone M_{PA-PB} remains connected to certain probes of the P_B zone and to the two probes of cluster #20 (see Fig. 5). This is due to the rectangular shape of the conductivity signals in both zones (see Fig. 6); their similarity is enhanced when the signals are normalized.

Water quality sub-clusters can be identified for a threshold value of 0.19. In particular, M_{PB-PC} (cluster #18) is split into two sub-clusters comprising four probes in the northern part and two probes in the southern part. This subdivision is confirmed by hydraulic tracing of the nearby tank, showing that the two southern probes are fed by water coming mostly from the tank (ranging from 50 % to 70 %). The probes of the M'_{PB-PC} zone are completely separated from the rest of the WDN. The associations kept at lower threshold values show that M'_{PB-PC} is mostly influenced by water coming from P_C , which is consistent with the hydraulics. An alternate cluster #14 is also proposed that appears to be more consistent with the hydraulics. The M_{W2-PC} zone is identified and separated from the M_{PA-PB} zone, which is perfectly identified (originally both zones belong to cluster #17 in Fig. 5).

The sub-clusters of the P_A zone appear for a threshold value of 0.28. In particular, cluster #2 is split into two (eastern and western) sub-clusters which are consistent with DMA frontiers. For values higher than 0.28, the associations of the probes can be described as core zones (see zones M_{PA-PB} and P_A , for which the plant is still associated with other probes). In particular, the two probes corresponding to the M_{W2-PC} zone (belonging to cluster #17), cluster #9 and cluster #20 are clustered for more than 50% of the weeks. The most frequent association of probes involves two neighboring probes of zone M_{PA-PB} that are associated in more than 64% of the weeks.

Fig 7. Graphical illustration of the popularity of the pairs of probes clustered together for weekly clusters.

The overall and the weekly clustering analyses are complementary to each other, the second giving more insight to the results of the first, by allowing:

- the strength of probes' associations to be assessed and visualized (e.g., clusters #9, #19 and #20 appear to be very strongly defined)
- the cores of the clusters to be identified (see e.g., clusters #1 and #17)
- specific sub-clusters to be specified (e.g., northern and southern parts of cluster #18, eastern and western parts of cluster #2, M_{W2-PC} as a subcluster of cluster #17)

4. Conclusion

This article presented a novel methodology for defining specific water quality clusters by analyzing conductivity signals measured by probes in a WDN over a pluriennial period. Conductivity time-series were compared using a measure that automatically resynchronizes signals to take into account variable propagation delays between probes. Then, the probes were ordered using a density-based method. Finally, clusters of probes were extracted with a geometric method.

Without any prior knowledge of the WDN's operation or of the probes' locations, the methodology provides operators and hydraulic engineers with a sound and practical representation of water quality in the WDN, both in terms of water origin and water residence time in the network. In particular, the results showed that the probe clusters enhance understanding of how water is distributed within the network, whether in zones directly impacted by plants or by external water imports, mixing zones, or zones influenced by tanks. Globally, 82% of the probes were found to be clustered consistently with expert knowledge on the WDN's operation and 13% of the probes were unclassified.

The methodology proposed will be particularly suitable to WDNs that operate with waters having very different conductivity signals, for instance a 50-50 split between groundwater and surface water. Furthermore, conductivity-based clusters will offer an interesting prior tool for contamination warning systems, as they will reduce the search space, for both the detection and propagation steps. Future work will investigate the extension of the method to a dataset comprising temperature measurements to refine the clusters proposed. Moreover, the method will be coupled to a hydraulic regime detection approach (Cheifetz et al. (2017)) to produce probe clusters representing quasi-dynamic water origin zones.

Data availability

Original conductivity data were provided by SEDIF. Requests for data should be made directly to SEDIF. The codes belong to Veolia Eau d'Ile-de-France, and cannot be shared.

Acknowledgements

The authors thank Gabriel Damay for his work on the DTW post-treatments. They also thank Sylvia Figueira, Catherine Henriet, Catherine Elain and Salma Monjid for validating the results of the algorithm in terms of water-quality and hydraulics.

References

- Alam, I. A. and M. Sadiq (1989). "Metal contamination of drinking water from corrosion of distribution pipes." *Environmental Pollution* **57**(2), 167-178.
- Ankerst, M., M. M. Breunig and H.-P. Kriegel (1999). "OPTICS: ordering points to identify the clustering structure." *SIGMOD Rec.* **28**(2), 49-60.
- Blokker, M. E. J., R. W. Furnass, J. Machell, R. S. Mounce, G. P. Schaap and B. J. Boxall (2016). "Relating Water Quality and Age in Drinking Water Distribution Systems Using Self-Organising Maps." *Environments* **3**(2).
- Brecheisen, S., H.-P. Kriegel, P. Kröger and M. Pfeifle (2004). "Visually Mining Through Cluster Hierarchies". *SIAM International Conference on Data Mining*
- Chang, K., J. L. Gao, W. Y. Wu and Y. X. Yuan (2011). "Water quality comprehensive evaluation method for large water distribution network based on clustering analysis." *Journal of Hydroinformatics* **13**(3), 390-400.
- Cheifetz, N., S. Kraiem, P. Mandel, C. Féliers and V. Heim (2017). "Extracting Temporal Patterns for Contamination Event Detection in a Large Water Distribution System". *CCWI 2017: Computing and Control for the Water Industry Conference*, Sheffield, UK.
- Cyna, B., G. Chagneau, G. Bablon and N. Tanghe (2002). "Two years of nanofiltration at the Méry-sur-Oise plant, France." *Desalination* **147**(1), 69-75.
- Dupas, R., R. Tavenard, O. Fovet, N. Gilliet, C. Grimaldi and C. Gascuel-Odoux (2015). "Identifying seasonal patterns of phosphorus storm dynamics with dynamic time warping." *Water Resources Research* **51**(11), 8868-8882.
- Dürrenmatt, D. J., D. Del Giudice and J. Rieckermann (2013). "Dynamic time warping improves sewer flow monitoring." *Water Research* **47**(11), 3803-3816.
- Ennouri, O., N. Cheifetz, P. Mandel, C. Féliers and V. Heim (2019). "Graph Clustering for Quality Event Detectors in a Large Water Distribution Network". *MARAMI 2019: The 10th Conference on Network Modeling and Analysis*, Dijon, France.
- Ester, M., H.-P. Kriegel, J. Sander and X. Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". *KDD'96 (Second International Conference on Knowledge Discovery and Data Mining)*.
- Housh, M. and Z. Ohar (2017). "Integrating physically based simulators with Event Detection Systems: Multi-site detection approach." *Water Research* **110**, 180-191.
- Huang, P., N. Zhu, D. Hou, J. Chen, Y. Xiao, J. Yu, G. Zhang, and H. Zhang (2018). "Real-Time Burst Detection in District Metering Areas in Water Distribution System Based on Patterns of Water Demand with Supervised Learning" *Water* **10**(12), 1765.
- Jendoubi, S., A. Martin, L. Liétard, B. Ben Yaghlane and H. Ben Hadji (2015). "Dynamic Time Warping Distance for Message Propagation Classification in Twitter". *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Cham, Springer International Publishing.
- Keogh, E. and S. Kasetty (2003). "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration." *Data Mining and Knowledge Discovery* **7**(4), 349-371.
- Lambrou, T. P., C. C. Anastasiou, C. G. Panayiotou and M. M. Polycarpou (2014). "A low-cost sensor network for real-time monitoring and contamination detection in drinking water distribution systems." *IEEE sensors journal* **14**(8), 2765-2772.
- Li, M., Z. Liu, Y. Chen and Y. Hai (2016). "Characteristics of iron corrosion scales and water quality variations in drinking water distribution systems of different pipe materials." *Water Research* **106**, 593-603.
- Li, Y., H. Chen and Z. Wu (2010). "Dynamic Time Warping Distance Method for Similarity Test of Multipoint Ground Motion Field." *Mathematical Problems in Engineering* **2010**, 12.
- Lifshitz, R. and A. Ostfeld (2018). "Clustering for Analysis of Water Distribution Systems." *Journal of Water Resources Planning and Management* **144**(5).
- Mandel, P. (2010). "Modelling Ozonation Processes for Disinfection By-Product Control in Potable Water Treatment: From Laboratory to Industrial Units". Ph.D. thesis, Université Rennes 1.

Mandel, P., M. Maurel and D. Chenu (2015). "Better understanding of water quality evolution in water distribution networks using data clustering." *Water Research* **87**, 69-78.

Martínez-Solano, F. J., P. L. Iglesias-Rey, D. Mora-Meliá and V. S. Fuertes-Miquel (2017). "Exact Skeletonization Method in Water Distribution Systems for Hydraulic and Quality Models." *Procedia Engineering* **186**, 286-293.

McDougle, T., A. Monsorez, M. Maurel and C. Lemoine (2012). "Smart Sensor network case study for drinking water quality monitoring". *IWA World Water Congress & Exhibition*, Busan, South Korea.

Müller, M., H. Mattes and F. Kurth (2006). "An Efficient Multiscale Approach to Audio Synchronization". *6th International Conference on Music Information*.

Oliker, N. and A. Ostfeld (2015). "Network hydraulics inclusion in water quality event detection using multiple sensor stations data." *Water Research* **80**, 47-58.

Perelman, L. and A. Ostfeld (2006). "Aggregation of Water Distribution Systems for Contamination Detection." *Water Distribution Systems Analysis Symposium 2006*: 1-13.

Perelman, L. and A. Ostfeld (2012). "Water-Distribution Systems Simplifications through Clustering." *Journal of water resources planning and management* **138**(3), 218-229.

Qin, T. and D. Boccelli, L. (2017). "Grouping Water-Demand Nodes by Similarity among Flow Paths in Water-Distribution Systems." *Journal of Water Resources Planning and Management* **143**(8), 04017033.

Riss, G., M. Romano, K. Woodward, F. Memon and Z. S. Kapelan (2018). "Improving Detection of Events at Water Treatment Works: A UK Case Study". *HIC 2018, 13th International Conference on Hydroinformatics*. Palermo, Italy.

Sakoe, H. and S. Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43-49.

Schmidt, C., A. Musolff, N. Trauth, M. Vieweg and J. H. Fleckenstein (2012). "Transient analysis of fluctuations of electrical conductivity as tracer in the stream bed." *Hydrol. Earth Syst. Sci.* **16**(10), 3689-3697.

Shokoohi-Yekta, M., B. Hu, H. Jin, J. Wang and E. Keogh (2017). "Generalizing DTW to the multi-dimensional case requires an adaptive approach." *Data mining and knowledge discovery* **31**(1), 1-31.

Ulanicki, B., A. Zehnpfund and F. Martinez Alzamora (1996). "Simplification of Water Distribution Network Models". *Second International Conference on Hydroinformatics*. Zürich, Suisse, 493-500.

Woo, H., D. L. Boccelli, J. G. Uber, R. Janke, and Y. Su (2019). "Dynamic Time Warping for Quantitative Analysis of Tracer Study Time-Series Water Quality Data" *Journal of Water Resources Planning and Management* **145**(12), 04019052.

Xie, C. L., Q. S. Fang, H. Y. Zhang, and J. X. Zhang (2019). "Optimal layout of pressure monitoring points in water supply network based on Optics", *E3S Web of Conferences* **79**, 01015.

Xing, L., and L. Sela (2019). "Unsteady pressure patterns discovery from high-frequency sensing in water distribution systems." *Water Research* **158**, 291-300.

Appendix A: average yearly conductivity profiles

Whereas the zones influenced by a unique water origin exhibit different yearly conductivity profiles (especially zones P_A and W_1), mixing zones are more similar to each other, except for zone M_{W2-PC} . This can be seen in Fig. A1, where yearly conductivity profiles, as measured by the probes belonging to the zones defined in Fig. 1, are depicted. For each zone, conductivity data were grouped by month, and deciles were computed to assess the seasonal trends and dispersions of the values within the zones. The median values of plants P_A , P_B and P_C are also presented (Fig. A1 a). Differences between plants can be explained by: (i) the Intake Water (IW) characteristics (the IWs of P_B and P_C have a median absolute difference of $30 \mu\text{S}\cdot\text{cm}^{-1}$ while the IW of P_A has median absolute differences of $95 \mu\text{S}\cdot\text{cm}^{-1}$ and $134 \mu\text{S}\cdot\text{cm}^{-1}$, with IWs of P_B and P_C , respectively); and by (ii) plant P_A itself, which is the only one to use nanofiltration to treat water (Cyna et al. (2002)). Conductivities measured at P_B and P_C , and to a smaller extent at P_A , are lower than in their corresponding zones, most probably due to limited iron-release in the network (Alam and Sadiq (1989); Li et al. (2016)).

Fig.A1. Yearly profiles of conductivity measurements for each zone defined: a) Zones influenced by a unique water origin (2nd to 8th deciles); b) mixing zones (3rd to 7th deciles).

Appendix B: Determination of the optimal values of the parameters of GC and e-GC methods

Optimal values for the parameters of the extraction method were determined by testing different values and assessing the quality of the results on the whole dataset (226 weeks of conductivity data for 215 probes). Weekly clusters were formed by OPTICS and then extracted using GC and e-GC methods.

The silhouette index is a measure used to assess the relevancy of clustering results. It relies on the comparison of intra-cluster distance (which shall be minimized to guarantee the homogeneity of the clusters formed) with inter-cluster distance (which shall be maximized to guarantee the separation of distant points). Equation B1 gives the formal definitions used in the tests. Once the silhouette index S_C has been calculated for each cluster C , the average silhouette index S is computed. Silhouette values should be maximized (high inter-cluster distance compared to intra-cluster distance).

$$\left\{ \begin{array}{l} S_C = \frac{b_C - a_C}{\max(b_C, a_C)} \\ S = \frac{1}{N} \sum_{i=1}^N S_i \\ b_C = \min_{L \neq C} (M_{i \in C}(D_{DTW}(i, RP_L))) \\ a_C = M_{i \in C}(D_{DTW}(i, RP_C)) \end{array} \right. \quad (\text{Equation B1})$$

- b_C is the inter-cluster distance of cluster C
- a_C is the intra-cluster distance of cluster C
- RP_L is the representative probe of cluster L
- $D_{DTW}(i, RP_L)$ is the DTW distance between probe i and the representative probe of cluster L
- $M_{i \in C}$ is the median value for all probes i belonging to cluster C

Different values were tested with GC in order to set the parameters w and t . The following values were tested:

- $w = 0.001 ; t = 150$
- $w = 0.001 ; t = 160$
- $w = 0.001 ; t = 170$
- $w = 0.01 ; t = 150$
- $w = 0.01 ; t = 160$
- $w = 0.01 ; t = 170$
- $w = 0.1 ; t = 160$
- $w = 0.1 ; t = 170$
- $w = 0.2 ; t = 170$

Once the optimal parameter values for GC were set, the different values tested for e-GC were:

- $D = 1 ; B = 1$
- $D = 1 ; B = 2$
- $D = 1 ; B = 3$
- $D = 1 ; B = 4$
- $D = 1 ; B = 5$
- $D = 2 ; B = 2$
- $D = 2 ; B = 3$
- $D = 2 ; B = 4$
- $D = 2 ; B = 5$
- $D = \infty ; B = \infty$
- $D = 0 ; B = 0$ (simply called $w = 0.01 ; t = 170$ in the following)

Results obtained with the GC method are presented in Fig. B1. Besides the silhouette indicator, the proportion of probes being clustered was considered to select the best couple of values for w and t . The distributions of

the values of the indicators over the 226 weeks and the 6 parameter settings are presented as boxplots. The results obtained are consistent with the method presented in (Brecheisen et al. (2004)): (i) the higher w , the smaller the clusters; (ii) the higher t , the larger the clusters. As it can be seen from Fig. B1, indicators are opposing (if the clusters are larger, the silhouette generally decreases). Therefore, selecting optimal values for w and t consists in finding the most acceptable compromise between the indicators. Values equal or greater than 0.1 for w , despite yielding high median silhouette values, were discarded due to the low proportion of classified probes and the high variability of the silhouette values. A good trade-off is obtained with $w = 0.01$ and $t = 170$: this setting leads to a significant increase of the proportion of classified probes (median value of 71%) without compromising quality (median silhouette index of 0.52).

Fig. B1. Results of the tests performed to determine the optimal parameter values of the GC extraction method (the central red line indicates the median, the bottom and top edges of the blue box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol)

Results obtained with the e -GC method are presented in Fig. B2. Besides the two previous indicators, the numbers of probes clustered erroneously (i.e. having a negative silhouette index) were also considered. The choice of the “best” values for the parameters was made considering the stabilization of the proportion of clustered probes, which, after a rapid increase, tends to stabilize with new successors being added to the clusters. The parameter values corresponding to the inflexion point were selected: $D = 1$ and $B = 3$. With a limited loss in quality (the median silhouette value decreases from 0.52 to 0.43), this setting increases the probes of probes clustered from 0.71 to 0.90.

Fig. B2. Results of the tests performed to determine the optimal parameter values of the e-GC extraction method (the central red line indicates the median, the bottom and top edges of the blue box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol)

Appendix C: Clustering results obtained with the *CL* measure

Fig. C1. Spatial extension of the 22 overall clusters found after GC and *e*-GC extractions. OPTICS results obtained with *CL* measure

Table C1. Matching matrix between the quality zones presented in Fig. 1 and overall clusters. The figures represent the number of probes (in red: probes clustered erroneously). OPTICS results obtained with *CL* measure

Zone Cluster#	P _A	P _B	P _C	W ₁	M _{PA-PB}	M _{PB-PC}	M' _{PB-PC}	M _{W2-PC}	M _{W3-PB}	M _{W4-PB}
1	18									
2	19									
3	10									
4	2									
5	2									
6		33							1	2
9		2								
10			25							
11			11							
12			11							
13			5							
14			3							
15-1			2							
15-2			2							
15-3			2							
16				3						
17-1		1			12					
17-2								2		
18				1		6				
19							5			
20		2			1					
21	4	11	15	1					1	