proceedings

MDPI

Proceedings

Off-Line Data Validation for Water Network Modeling Studies †

Marcos Quiñones-Grueiro [1,‡*], Lizeth Torres [2,‡] and Cristina Verde [2,‡]

[1] Departamento de Automática y Computación, Universidad Tecnológica de La Habana José Antonio Echeverría, Calle 114 No. 11901, CUJAE, Marianao, La Habana 19930, Cuba

[2] Instituto de Ingeniería, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico; ftorreso@iingen.unam.mx (L.T.); verde@unam.mx (C.V.)

* Correspondence: marcosqg88@gmail.com

† Presented at the 4th International Electronic Conference on Water Sciences, 13–29 November 2019; Available online: https://ecws-4.sciforum.net/.

‡ These authors contributed equally to this work.

Published: 12 November 2019

Abstract: The success of the analysis and design of a Water Network (WN) is strongly dependent on the veracity of the data and a priori knowledge used in the model calibration of the network. This fact motivates this paper in which an off-line approach to verify datasets acquired from WN is proposed. This approach allows the data separation of abnormal and normal events without requiring high expertise for a large raw database. The core of the approach is an unsupervised classification tool that does not require the features of the different events to be identified. The proposal is applied to datasets acquired from a Mexican water management utility located in the center part of Mexico. The datasets are pre-processed to be synchronized since they were recorded and sent with different and irregular sampling times to a web platform. The pressures and flow-rate conforming the datasets correspond to the dates between 25 June 2019 @ 00:00 and 25 September 2019 @ 00:00. The District Metered Area (DMA) is formed by 90 nodes and 78 pipes, and it provides service to approximately 2000 consumers. The raw data identified as generated by abnormal events are validated with the reports of the DMA managers. The abnormal events identified are communication problems, sensor failures, and draining of the network reservoir.

Keywords: off-line data validation; water networks; abnormal data classification

1. Introduction

Data acquisition systems in Water Networks (WNs) collect measurements from the in situ sensors and transform them into mathematical values that represent a physical quantity. This value set R_D (known as raw data) must be validated before being used for network operation purposes or statistics studies to assure the reliability of the captured information. Some common problems caused by sensors malfunctions are offset, drift, and freezing of the measured variable [1]. Moreover, data from abnormal events that occur in the network must be identified to avoid incorrect studies and the construction of false models.

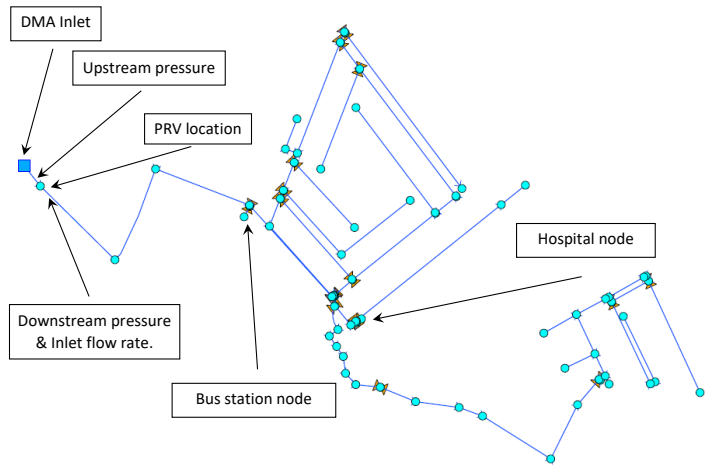
In general, WN operating data are required to build mathematical and data- driven models, which are significantly affected by the uncertain demand patterns and the quality of the data used in the model calibration [2]. Thus, if raw data R_D are not validated before they are used for diverse purposes, the resulting studies and models could not be representative of the real behavior of the network in normal operating conditions. Previous contributions have proposed data validation techniques for on-line applications [3,4]. These proposals, however, require large datasets of nominal operating

conditions to identify a validation model. Therefore, from a practical point of view, it makes more sense to validate, as a first step, the raw data in any study of WNs.

In view of the forgoing arguments, this paper presents a semi-automatic procedure for off-line validation of raw data acquired from WNs. The procedure, based on artificial intelligence tools, consists of four steps that require minimal setup, and it allows classifying the data associated with the nominal behavior of the network and the data that are generated from abnormal events. The procedure is applied to validate data acquired from a real DMA called El Charro, which is located in a small city in Mexico. It is demonstrated hereafter that it is possible to identify different anomalous events that do not correspond to the behavior of the normal consumers.

2. Case Study: El Charro DMA

The proposed procedure was applied to a raw database coming from a District Metered Area (DMA) located in a small city in the center part of Mexico. El Charro is comprised of a middle-class neighborhood, a public hospital, and a bus station. The EPANETlayout of the DMA and their main characteristics are presented in Figure 1.



Characteristics	*Quantity*
Pipelines:	75
Nodes	90
Supply reservoir capacity:	1000 (m^3)
Customers:	2000
Average consumption per year:	3 (lps)
Installed Sensors:	1 upstream pressure transducer (kg/cm^2) 1 downstream pressure transducer (kg/cm^2) 1 inlet flowmeter (lps)
Valves:	1 pressure reducing valve [PRV]

Figure 1. EPANETlayout of the District Metered Area (DMA) El Charro. PRV, Pressure Reducing Valve.

The raw database or sample set denoted as R_D was composed of upstream and downstream pressure, as well as flow-rate data, which were recorded and sent to a website platform from an IoT (Internet of Things) station that was located at the inlet of the DMA. The R_D corresponded to the dates between 25 June 2019 @ 00:00 and 25 September 2019 @ 00:00. The pressures and flow-rate records were sent to the website platform by different non-synchronized telemetry devices with irregular intervals between 10 and 11 minutes. Thus, the database was pre-processed to have the same dimension as a regular (uniform) time separation for the three variables of R_D.

The pre-processing was achieved in two steps. Firstly, the set of samples R_D for each month was linearly interpolated considering that the estimated values were separated by a regular (uniform) period of time τ [5]. Secondly, a univariate test was performed to remove values of the data that lied far from the means. This step was designed according to the expert knowledge about the physical variables from the Mexican DMA. Here, the univariate estimation for the flow rate q_k lower than the minimum night flow q_{min} was applied. Thus, q_k were replaced by the interpolated value from the previous and after values q_{k-1} and q_{k+1}, respectively. Thus, these preprocessing steps generated the new array $P_s = [P_{s_1}, P_{s_2}, P_{s_3}]$, which was the input array of the validation process with three rows for the three months of register data of the DMA.

3. Clustering Procedure

The unsupervised clustering algorithms could be considered as systematic computational processes used to handle a huge mount of data, which could be classified according to their similarities and differences without a priori knowledge of the classes of groups [6]. Thus, a clustering process could be used in a WN to reveal the organization of patterns into groups and to separate normal data from abnormal data.

The proposed procedure is described in Figure 2. The first step, as usual, involved data pre-processing methods to perform the following tasks: normalization, noise filtering, missing data recovering, and so on [5]. For the El Charro DMA, the pre-processing task was explained in the previous section. Feature selection, which is the second step, consisted of determining the features of the pre-processing dataset P_s to be analyzed. In our case, we only considered the straightforward values of the variables. Thus, the feature array is given by $F = P_s$.

Step 3, which is the main contribution of this paper, consisted of the use of unsupervised machine learning techniques to perform anomaly detection. The goal of the anomaly detection task was to isolate the events in the dataset that did not correspond to the normal consumption of the users. This was a fundamental problem because if the dataset was not validated, then it could not be used for water modeling tasks, i.e., demand modeling, WN model calibration, etc. Finally, in Step 4, the resulting clusters that represented the normal consumption patterns were integrated into a single dataset.

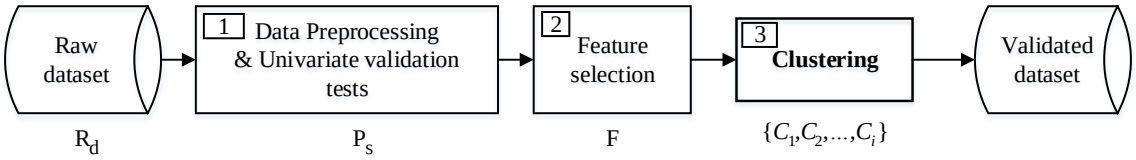


Figure 2. Off-line raw data validation procedure.

Clustering Patterns with DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a non-parametric, density-based clustering technique [7]. Namely, the goal of the algorithm is to partition the dataset formed by the feature array F into sub-sets. In this work, an object is understood as a feature observation $\mathbf{f}_i \subset F$ for all $i = 1, 2, ..., n$. This method in particular identifies regions in the data space with a high density of objects.

To define a cluster by considering the n observation set $F = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ with $\mathbf{f}_i \in \Re^m$, the concept of neighborhood and density reachable objects are required [7]

Definition 1 (Neighborhood of $\mathbf{f}_i$)**.** *The neighborhood of object $\mathbf{f}_i$ denoted as $D_i = \{\mathbf{f}_j \in F\}$ is defined by the set of objects $\mathbf{f}_j$ such that a proximity measure between $\mathbf{f}_i$ and $\mathbf{f}_j$ is satisfied. This means that $D_i = \{\mathbf{f}_j \in F \mid ||\mathbf{f}_j - \mathbf{f}_i|| < d_{th}\}$ where d_{th} is a user-defined threshold that characterizes the size of the neighborhood. In this context, all $\mathbf{f}_j$ are called neighbors of $\mathbf{f}_i$.*

An object $\mathbf{f}_i^*$ is called a core-object if the number of objects in its neighborhood D_i is larger than a user-defined number $MinPts \in Z$. The rest of the objects inside the neighborhood of a core-object are called border objects.

Definition 2 (Density reachable objects)**.** *If there exist a set of core-objects $\{\mathbf{f}_1^*, \mathbf{f}_2^*, ..., \mathbf{f}_i^*\}$ that are neighbors, then any object of their respective neighborhoods $D_1, D_2, ..., D_i$ is density reachable by any of the core-objects.*

Definition 3 (Cluster)**.** *In the framework of DBSCAN, a cluster is defined by the set of density reachable objects $\mathcal{C} = D_1 \bigcup D_2 \bigcup ... \bigcup D_i$.*

In general, if after processing all objects in F, an object is not density reachable it is considered as an outlier or unstructured data. From the above definitions, one can see that two parameters define a cluster: $MinPts$ and d_{th}. The former one defines the minimum number of objects required to consider the existence of a cluster, and the latter characterizes how close these objects must be in the data space. The DBSCAN algorithm is shown in Algorithm 1.

Data: F, $MinPts$, d_{th}, C: set of clusters, No: set of noise objects, i: number of clusters
Label all objects as not classified, $C = \emptyset$, $No = \emptyset$, $i = 0$;
for $\mathbf{f}_j \in F$ **do**
  if $\mathbf{f}_j$ *is not classified* **then**
    $DR_j = DensReach(\mathbf{f}_j)$
    if $|DR_j|$>1 **then**
      Form a new cluster with all density reachable objects
      Label cluster' objects as classified
      $C_i = DR_j$, $C = \{C, C_i\}$, $i = i + 1$
    if $\mathbf{f}_j$ *is not a border-object* **then**
      $No = No \cup \mathbf{f}_j$
  Label $\mathbf{f}_j$ as classified
end

Algorithm 1: DBSCAN algorithm.

For the application of DBSCAN to the data from a DMA, we considered that the minimum number of observations that formed a pattern was defined by the duration of the Minimum Night Flow (MNF) regime corresponding to the time period from 3 am to 6 am. Given that the sampling time of our system was approximately 10 min, a minimum of $MinPts = 18$ observations was selected such that any cluster satisfied the condition $|\mathcal{C}| \geq 18$.

The applied steps for the search for the threshold d_{th} are summarized as follows, and the specific graphic for El Charro is shown in the left plot of Figure 3.

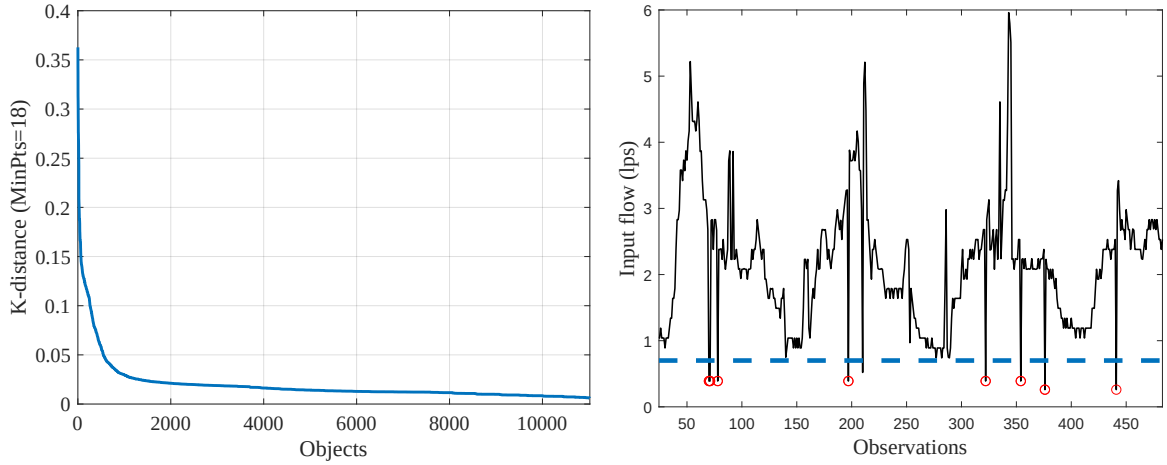


Figure 3. Partial results of the parameters and data of the DMA: (**a**) sorted objects vs. the Kproximity metric with a minimum cluster of 18 objects; (**b**) preprocessing flow rate considering the Minimum Night Flow (MNF) for a time window of 450.

- Compute the distances of each object f_i with respect to its nearest neighbors, and sort them in ascending order, for all objects.
- Define the distance d_i that corresponds to the 18$^{\text{th}}$ position of the classification, for all objects.
- Sort all the measures $d_s = \{d_1, d_2, ..., d_n\}$ according to the magnitudes in descending order and plot them according to their respective magnitude.
- Choose the $d_{th} = K$-metric where the sorted object and the K-metric are given by the first valley.

4. Results and Discussion

This section describes the main results of the validation process for the data R_D and discusses the performance of the proposition by considering the study case. To visualize the effects of data management clearly, only short time windows are shown in the figures.

A time series of 450 interpolated and synchronized values is shown in the right side of Figure 3. The blue line corresponds to the value of the MNF regime, and data below this value were replaced by interpolated data and marked with the symbol ∘ in the graphic. One can see that the circles are isolated points and without any dynamics. Thus, these do not correspond to abnormal events.

By applying the DBSCAN algorithm to all the array F, two clusters were obtained in the data space. To clarify the interpretation of the results, the projection of the two identified clusters in each plane of F is shown in Figure 4. The ∘ symbol in red color denotes an object in the normal cluster, and the × symbol in blue color means an abnormal event. Thus, the cluster that represents the normal consumption is identified, and the other cluster is separated with unstructured data that represent anomalies. The classified data shown in the three projections indicate the relationship between three features of the measured variables: the upstream pressure, which was measured before a Pressure Reducing Valve (PRV) installed at the DMA inlet, the downstream pressure, which was measured after the PRV, and the flow rate, which was measured after the PRV and whose behavior depended on the demand for water by the DMA users.

The three projections shown in Figure 4, respectively, have the following relations: upstream pressure-flow rate, downstream pressure-flow rate, and downstream pressure-upstream pressure. In the left graph, it can be noted that there are many data that indicate that the behavior of the flow rate is not related to the upstream-pressure normal behavior since a low pressure is not feasible with a relative high flow. This situation is not perceived in the center graph, since the number of data showing a disassociation between the flow-rate behavior and the downstream-pressure behavior is smaller. This was an indicator that an abnormal event was out of the network. Finally, in the right graph, a large amount of data can be seen that highlight an anomalous condition between the upstream pressure and the downstream pressure. Thus, it was concluded that the abnormal event was associated with a low upstream pressure.

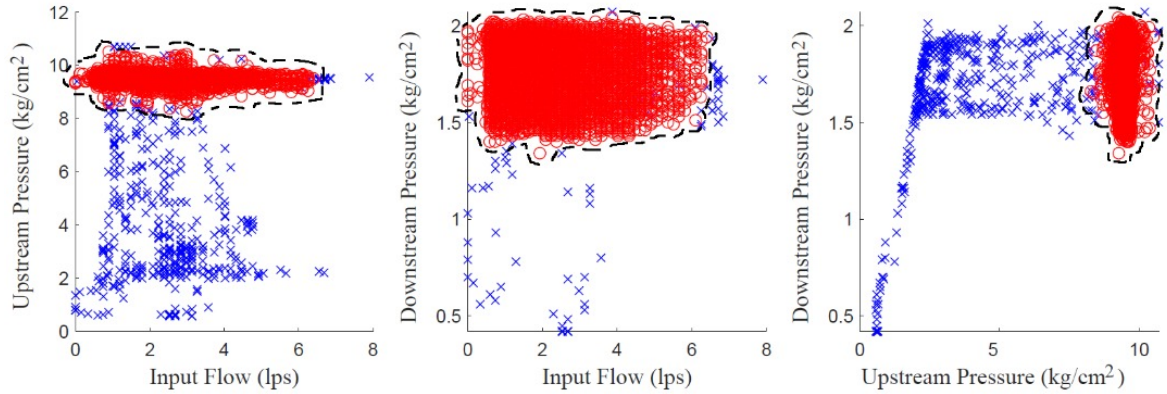


Figure 4. Data space projections of the features: normal conditions in red ∘, anomalies conditions in blue ×.

To analyze the data results in the time domain, windows from the samples 1700 to 2400 identified as abnormal data are shown in Figure 5. The abnormal event produced a sharp upstream pressure drop and deviations in both directions of the flow rate. On the contrary, the downstream pressure was only reduced drastically in a small sample interval. These behaviors of the variables could be diagnosed as a reservoir draining. This analysis was coherent with the cluster remarks made by analyzing Figure 4. This conclusion was verified with the operator register. Therefore, the tank draining behavior was isolated from the normal events. This subset of data could not be used to model the nominal behavior

of the network. Thus, the data corresponding to these time periods should not be used for any study of the DMA, except for fault diagnosis purposes.

Since downstream pressure and flow were measured after a PRV and since the relationship between both variables seemed to have only one pattern, one could infer that an anomaly existed before the PRV. An explanation for this inference can be found in Figure 5b, which shows the behavior of the upstream pressure. In particular, it was observed that the upstream pressure dropped three times. According to the DMA managers, these drops were due to problems in supplying the reservoir. More precisely, the pumps used to feed the reservoir failed. Figure 5c shows that only one of these three drops affected the downstream pressure, which was thanks to the PRV, which worked as long as the upstream pressure was greater than the downstream pressure. As can be seen in Figure 5b,c, the upstream pressure was lower than the downstream pressure only once around the 1800th observation.

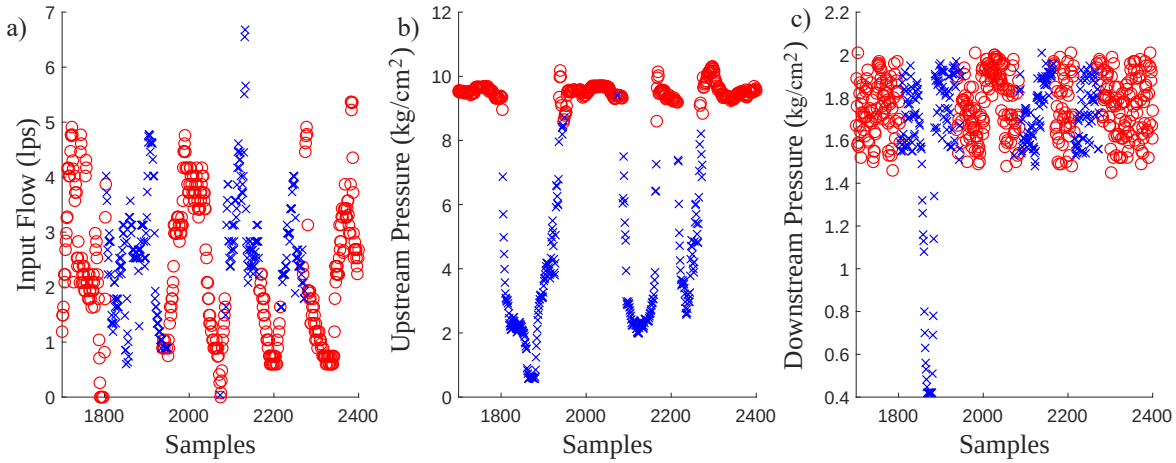


Figure 5. Classified raw data produced by reservoir draining.

5. Conclusions

This paper presented an off-line approach to data validation in WN for modeling studies. The core of the proposal was the application of an unsupervised classification tool that did not require the features of the different events to be identified. The advantages of the proposal were illustrated with a datasets acquired from a Mexican water management utility. The abnormal events identified in the data were validated with the reports of the DMA managers. In particular, the unsupervised method allowed the identification of a systematic anomaly: the draining of the reservoir. On the basis of these results, the network operators concluded the convenience of the pressure reducing valve.

Funding: This research was funded by IT100519-DGAPA-UNAMand CONACYT Convocatoria de Proyectos de desarrollo científico para atender problemas nacionales 2017, Proyecto 4730 Estaciones de diagnóstico y monitoreo para redes de distribución de agua con conexión a Internet.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kanakoudis, V.; Tsitsifli, S. Water pipe network reliability assessment using the DAC method. *Desalin. Water Treat.* **2011**, *33*, 97–106, doi:10.5004/dwt.2011.2631.
2. Bartkiewicz, E.; Zimoch, I. Impact of Water Demand Pattern on Calibration Process. *Proceedings* **2017**, *2*, 191, doi:10.3390/ecws-2-04961.
3. Quevedo, J.; Puig, V.; Cembrano, G.; Blanch, J.; Aguilar, J.; Saporta, D.; Benito, G.; Hedo, M.; Molina, A. Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract.* **2010**, *18*, 640–651, doi:10.1016/j.conengprac.2010.03.003.

4. Cugueró-escofet, M.À.; García, D.; Quevedo, J.; Puig, V.; Espin, S.; Roquet, J. A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network. *Control Eng. Pract.* **2016**, *49*, 159–172, doi:10.1016/j.conengprac.2015.11.005.
5. Han, J.; Kamber, M.; Pei, J. *Data Mining Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2012; p. 703.
6. Theodoridis, S. Koutroumbas, K. *Pattern Recognition*; Elsevier: Amsterdam, The Netherlands, 2009.
7. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996; pp. 226–231.