


Article

# Off-Line Data Validation for Water Network Modeling Studies

Marcos Quiñones-Grueiro <sup>1\*</sup>  0000-0001-5391-6774, Lizeth Torres <sup>2</sup> and Cristina Verde <sup>2</sup>

<sup>1</sup> Universidad Tecnológica de La Habana José Antonio Echeverría, Calle 114 No. 11901, CUJAE, Marianao, La Habana, Cuba, CP:19930; marcosqg88@gmail.com

<sup>2</sup> Universidad Nacional Autónoma de México, Instituto de Ingeniería, México

‡ These authors contributed equally to this work.

Version November 5, 2019 submitted to Water

**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 data-sets 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 requires the features of the different events to be identified. The proposal is applied to data-sets acquired from a Mexican water management utility located in the center part of Mexico. The data-sets were 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 data-sets correspond to dates between 25/06/2019 @ 00:00 and 25/09/2019 @ 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 were validated with the reports of the DMA managers. The abnormal events identified were 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 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 data sets of nominal operating conditions to identify a validation model. Therefore, from 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 to classify the data associated with

the nominal behavior of the network from the data which 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 which do not correspond to the behavior of the normal consumers.

## 2. Case Study: El Charro DMA

The proposed procedure is 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* comprises a middle-class neighborhood, a public hospital and a bus station. The EPANET layout of the DMA and their main characteristics are presented in Fig. 1.

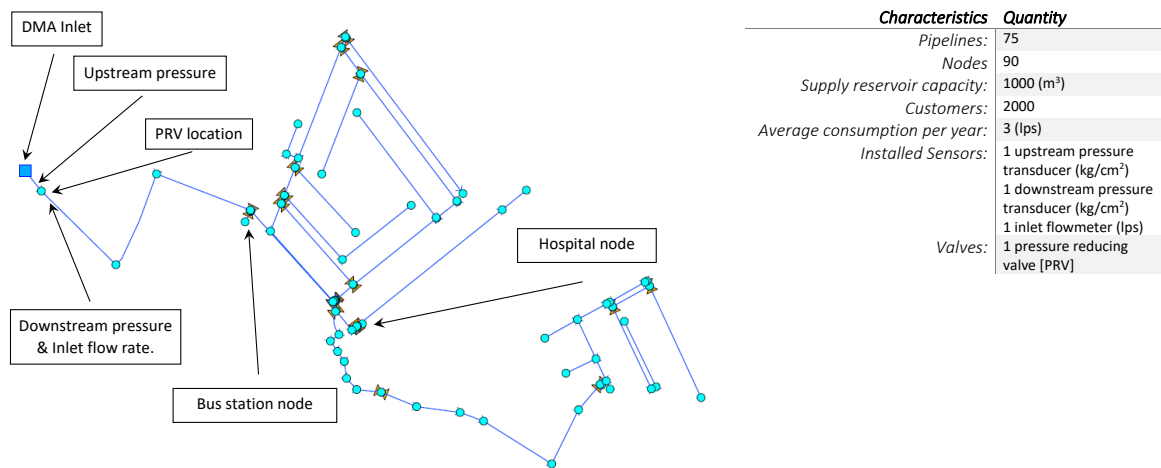


Figure 1. EPANET layout of the DMA *El Charro*

The raw database or sample set denoted  $R_D$  is 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 is located at the inlet of the DMA. The  $R_D$  corresponds to dates between 25/06/2019 @ 00:00 and 25/09/2019 @ 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 with a regular (uniform) time separation for the three variables of  $R_D$ .

The pre-processing is achieved in two steps. Firstly, the set of samples  $R_D$  for each month were linearly interpolated considering that the estimated values are separated by a regular (uniform) period of time  $\tau$  [5]. Secondly, a univariate test was performed to remove values of the data that lies far from the means. This step is 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  $q_{min}$  is 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 generate the new array  $P_s = [P_{s_1}, P_{s_2}, P_{s_3}]$  it is 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 can be considered as systematic computational processes used to handle huge of data which can be classified according to their similarities and differences without *a priori* knowledge of the classes of groups [6]. Thus, a clustering process can 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, involves data pre-processing methods to perform the following tasks: normalization, noise filtering, missing data

66 recovering and so on [5]. For the El Charro DMA the pre-processing task was explained in the  
 67 previous section. Feature selection, which is the second step, consists of determining the features of  
 68 the pre-processing data set  $P_s$  to be analyzed. In our case, we only considered the straightforward  
 69 values of the variables. Thus, the feature array is given by  $F = P_s$ .

70 Step 3, which is the main contribution of this paper, consists in the use of unsupervised machine  
 71 learning techniques to do anomaly detection. The goal of the anomaly detection task is to isolate the  
 72 events in the data-set which do not correspond to the normal consumption of the users. This is a  
 73 fundamental problem because if the data-set is not validated then it cannot be used for water modeling  
 74 tasks, i.e. demand modeling, WN model calibration, etc. Finally, in step 4, the resulting clusters that  
 75 represent the normal consumption patterns are integrated into a single data set.

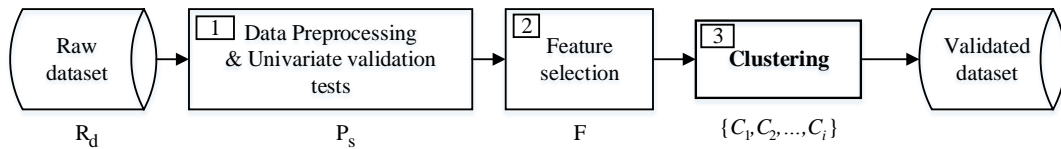


Figure 2. Off-line raw data validation procedure

### 76 3.1. Clustering patterns with DBSCAN

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

82 To define a cluster by considering the  $n$  observation set  $F = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$  with  $\mathbf{f}_i \in \mathbb{R}^m$ , the  
 83 concept of *Neighborhood* and *Density reachable objects* are required [7]

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

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

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

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

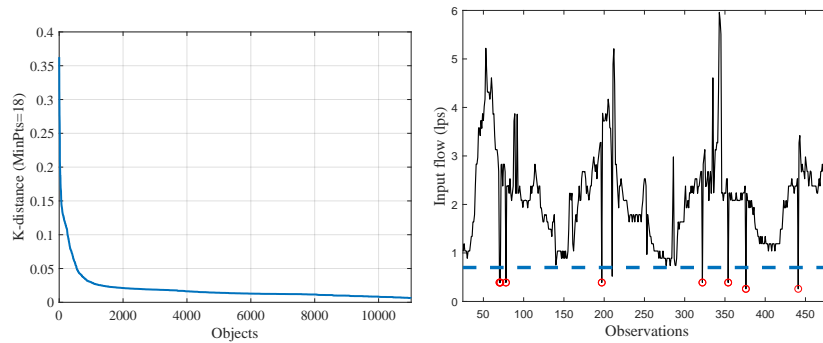
95 In general, if after processing all objects in  $F$  an object is not density reachable it is considered as  
 96 an outlier or unstructured data. From the above definitions one can see that two parameters define a  
 97 cluster:  $MinPts$  and  $d_{th}$ . The former one defines the minimum number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

100 For the application of DBSCAN to data from a DMA we considered that the minimum number of  
 101 observations that form a pattern are defined by the duration of the minimum night flow (MNF) regime  
 102 corresponding to the time period from 3 am to 6 am. Given that the sampling time of our system

**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**  $f_j \in F$  **do**  
   **if**  $f_j$  is not classified **then**  
      $DR_j = DensReach(f_j)$   
     **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**  $f_j$  is not a border-object **then**  
        $No = No \cup f_j$   
     Label  $f_j$  as classified  
**end**

**Algorithm 1:** DBSCAN Algorithm

103 is approximately 10 min a minimum of  $MinPts = 18$  observations is selected such that any cluster satisfies the condition  $|C| \geq 18$ . The applied steps for the search of the threshold  $d_{th}$  are summarized



**Figure 3.** Partial results of parameters and data of the DMA: (a) Sorted objects vs K proximity metric with a minimum cluster of 18 objects; (b) Preprocessing flow rate considering the MNF for a time window of 450

104

105 as follows and the specific graphic for the El Charro is shown in the left plot of Figure 3.

- 106
- Compute the distances of each object  $f_i$  with respect to its nearest neighbors and sort them in ascending order, for all objects.
  - 107
  - 108 • Define the distance  $d_i$  that corresponds to the 18th position of the classification, for all objects.
  - 109 • Sort all the measures  $d_s = \{d_1, d_2, \dots, d_n\}$  according to the magnitudes in descending order and plot them according to its respective magnitude.
  - 110
  - 111 • Choose  $d_{th} = K$ -metric where the sorted object and the  $K$ -metric is given by the first valley.

## 112 4. Results and Discussion

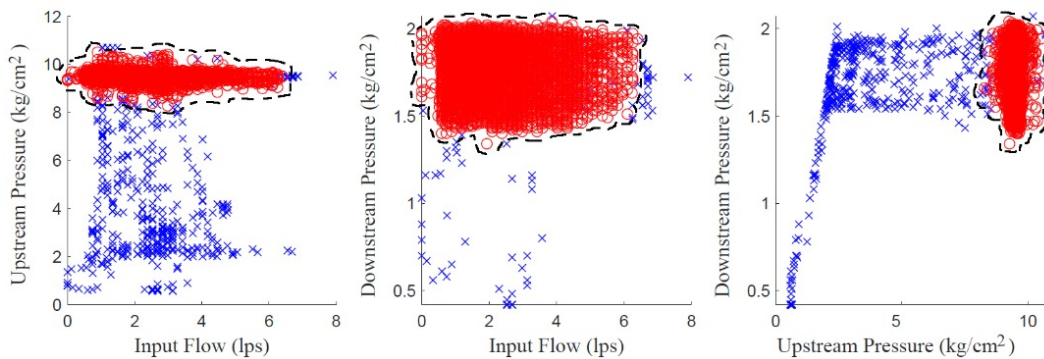
113 This section describes the main results of the validation process for the data  $R_D$  and discusses the performance of the proposition by considering the study case. To clearly visualize the effects of data management, only short time windows are shown in the figures.

116 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  $\circ$  in the graphic. One can see that the circles are isolated points and without any dynamic. Thus, these do not correspond to an abnormal event.

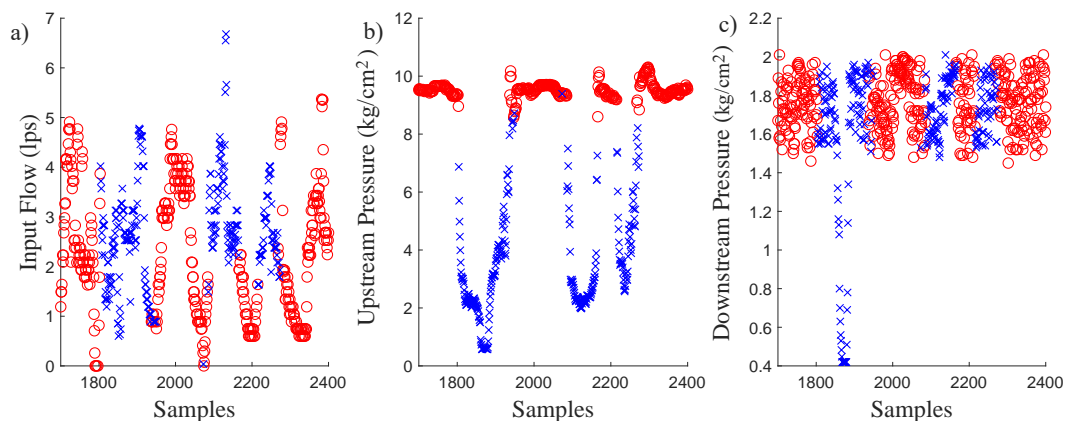
120 By applying the DBSCAN algorithm to all the array  $F$ , two clusters were obtained in the data space. To clarify the interpretation of the results the projection of the two identified clusters in each plane of  $F$  are shown in Figure 4. The  $\circ$  symbol in red color denotes an object in the normal cluster

123 and the  $\times$  symbol in blue color means abnormal event. Thus, the cluster that represents the normal consumption is identified and the other cluster is separated with unstructured data that represents  
 124 consumption is identified and the other cluster is separated with unstructured data that represents  
 125 anomalies. The classified data shown in the three projections indicates the relationship between  
 126 three features of the measured variables: the upstream pressure, which is measured before a pressure  
 127 reducing valve (PRV) installed at the DMA inlet, the downstream pressure, which is measured after  
 128 the PRV, and the flow rate, which is measured after the PRV and whose behavior depends on the  
 129 demand for water by the DMA users.

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



**Figure 4.** Data space projections of the features: Normal condition red  $\circ$ , Anomalies conditions blue  $\times$



**Figure 5.** Evolution of the classified raw data produced by reservoir draining

140 To analyze the data results in the time domain, windows from the samples 1700 to 2400 identified  
 141 as abnormal data are shown in Figure 5. The abnormal event produced a sharp upstream pressure  
 142 drop and deviations in both directions of the flow rate. On the contrary, the downstream pressure is  
 143 only reduced drastically in a small sample interval. These behaviors of the variables can be diagnosed  
 144 as a reservoir draining. This analysis is coherent with the cluster remarks made by analyzing Figure 4.  
 145 This conclusion was verified with the operator register. Therefore, the tank draining behavior has been



146 isolated from the normal events. This subset of data cannot be used to model the nominal behavior of  
147 the network. Thus the data corresponding to these time periods should not be used for any study of  
148 the DMA, except for fault diagnosis purpose.

149 Since downstream pressure and flow are measured after a PRV, and since the relationship between  
150 both variables seems to have only one pattern, one can infer that an anomaly exists before the PRV.  
151 An explanation for this inference can be found in Fig. 5 b) that shows the behavior of the upstream  
152 pressure. In particular, it is observed that the upstream pressure drops three times. According to  
153 the DMA managers, these drops were due to problems to supply the reservoir. More precisely, the  
154 pumps used to feed the reservoir failed. Fig. 5 c) shows that only one of these three drops affected the  
155 downstream pressure, what it was thanks to the PRV, which works as long as the upstream pressure is  
156 greater than the downstream pressure. As can be seen in Figures 5 b) and 5 c), the upstream pressure  
157 was lower than the downstream pressure only once around the 1800th observation.

## 158 5. Conclusions

159 This paper presented an off-line approach to data validation in WN for modeling studies. The  
160 core of the proposal is the application of an unsupervised classification tool which does not requires  
161 the features of the different events to be identified. The advantages of the proposal were illustrated  
162 with a data-sets acquired from a Mexican water management utility. The abnormal events identified  
163 in the data were validated with the reports of the DMA managers. In particular, the unsupervised  
164 method allowed the identification of a systematic anomaly: the draining of the reservoir. On the base  
165 of these results, the network operators concluded the convenience of the pressure reducing valve.

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

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

## 170 References

- 171 1. Kanakoudis, V.; Tsitsifli, S. Water pipe network reliability assessment using the DAC method. *Desalination*  
172 *and Water Treatment* **2011**, *33*, 97–106. doi:10.5004/dwt.2011.2631.
- 173 2. Bartkiewicz, E.; Zimoch, I. Impact of Water Demand Pattern on Calibration Process. *Proceedings* **2017**,  
174 *2*, 191. doi:10.3390/ecws-2-04961.
- 175 3. Quevedo, J.; Puig, V.; Cembrano, G.; Blanch, J.; Aguilar, J.; Saporta, D.; Benito, G.; Hedro, M.; Molina, A.  
176 Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control*  
177 *Engineering Practice* **2010**, *18*, 640–651. doi:10.1016/j.conengprac.2010.03.003.
- 178 4. Cugueró-escofet, M.À.; García, D.; Quevedo, J.; Puig, V.; Espin, S.; Roquet, J. A methodology and a software  
179 tool for sensor data validation / reconstruction : Application to the Catalonia regional water network.  
180 *Control Engineering Practice* **2015**, pp. 1–14. doi:10.1016/j.conengprac.2015.11.005.
- 181 5. Han, J.; Kamber, M.; Pei, J. *Data Mining Concepts and Techniques*; Elsevier, 2012; p. 703.
- 182 6. Theodoridis, S. Koutroumbas, K. *Pattern Recognition*; Elsevier, 2009.
- 183 7. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large  
184 Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data*  
185 *Mining* **1996**, pp. 226–231.

186 © 2019 by the authors. Submitted to *Water* for possible open access publication under the terms and conditions  
187 of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).