An improved tool of water data analytics for flowmeters data

Application to the Barcelona supra-municipal distribution water network

Joseba Quevedo
CS^{2}AC Research Center
Universitat Politecnica de Catalunya (UPC)
Terrassa (Barcelona), Spain
Email: joseba.quevedo@upc.edu

Jose Luis Pesantez
CS^{2}AC Research Center
Universitat Politecnica de Catalunya (UPC)
Terrassa (Barcelona), Spain
Email: jose.pesantez.corral@estudiant.upc.edu

Santiago Espin
ATLL Concessionaria de la
Generalitat de Catalunya, S.A
Barcelona, Spain

Jaume Roquet
ATLL Concessionaria de la
Generalitat de Catalunya, S.A
Barcelona, Spain

Fernando Valero
ATLL Concessionaria de la
Generalitat de Catalunya, S.A
Barcelona, Spain

***Abstract*—This paper presents an improved tool for data validation and reconstruction of flowmeters. These sensors are installed in the Catalonia regional water network from Barcelona (Spain). Here a new time series model with exogenous variable is proposed with excellent results for data validation. It is postulated that the integration of the electronics alarms, along with other tests about the daily data accumulated and a later analysis of the data reconstruction allow to improve the results of the existing tools. This is accomplished by decreasing the false alarms and missing alarms of more than 6000 hourly data retrieved from more than 200 flowmeters each day. This new tool provides reliable information daily reliable information of the state of the water network. This information could potentially contribute to optimally control and manage this large and complex water network.**

***Index Terms*—data analytics, data validation, time series, flowmeters, water network.**

I. Introduction

Critical Infrastructure Systems (CIS) such as the case of potable water transport network are large-scale systems, geographically distributed and decentralized with a hierarchical structure. These water networks require sophisticated supervisory and real-time control (RTC) schemes to ensure high performance achievement and maintenance when conditions are non-favorable [1], [2] due to e.g. sensor malfunctions (drifts, offsets, problems of batteries, communications problems, etc.). Reliable information is the basis for decision making processes. In the operation mode, reliable information aid to optimize energy costs and reduce water losses while ensuring supply to consumers in quantity and quality, regardless of fluctuating demands. The main task is to validate the raw data of the sensors and, if that the data is non validated, the tool replace this wrong data with an estimation data to reconstruct a reliable and complete database of the system. This procedure allows to process, filter, debug and complete all the received raw data and to transform them into useful information.

The case study here presented is the Catalonia regional water network. This facility is managed by ATLL Company which supplies water to the Barcelona metropolitan area (Fig. 1) where most of the Catalonian population is concentrated. This network transports the drinking water from the main water treatment plants (ETAPs), which take the water from two different rivers (Llobregat and Ter), towards the main storing and buffer tanks of 116 municipalities in the Barcelona metropolitan area, using about 1045 km of pipes of up to 3m diameter.

Fig. 1. ATLL water network of Barcelona metropolitan area.

The network is composed by 170 storage tanks, 67 pumps and 212 demand sectors. Every 10 min the data of more than 200 flowmeters and 115 tank level sensors are recorded in a SCADA system. ATLL supplies 4.5 million inhabitants with an approximate yearly demand of 210 cubic hectometres and its responsibility ends at municipal head tanks.

It is complex to efficiently operate this large network in real time. Operators might face difficulties in activities such as the management of supply and demand, changes of direction, controlled retention times to minimize the formation of sub-products, water mixtures (contributions from different origins, involving valve movements) and water path through pipes from 200 to 2400 mm with hundreds of sensors.

The reliability of quantitative measurements (basically the flows in the pipes and the volume of water produced and delivered to the municipalities) is important to control the revenue and non-revenue water of the network, to generate monthly the bills for the delivered water to the municipalities, to supervise in real-time the efficiency of around 100 sectors, 11 zones and overall network and also for early event detection (leakages, sensor or actuators malfunctions, etc.).

In order, to deal with this problem, the use of an on-line fault diagnosis system is advised. Such a system is capable to detect and to isolate such faults and correct them by activating different kind of techniques. Furthermore, the fault diagnosis process intends to identify which fault is causing the monitored events. Once the data are reliable a process to transform these validated data in useful information and knowledge is key for the operating plan in real time (RTC) and also but no least important to extract useful knowledge about the assets and instrumentation (sectors of pipes and reservoirs, flowmeters, level sensors, etc.) of the network for short, medium and large term management plans.

After six years the research group CS2AC of the UPC and the technical staff of ATLL Company have developed a tool for validation and reconstruction of the flowmeters data in the network. At this time, the aforementioned tool is operative in all the network [4]–[6]. A detailed daily report is generated with all the non-validated data, the tag of the flowmeter and occurred time, the test not exceeded, the reconstruction data and other key useful information.

In this paper an improvement of the tool and the results of its real application are presented. The paper is structured as follows. In Section 2, the methodology of the actual tool for on-line data validation and reconstruction is described. Then, the improvements and the recent results of this tool are presented in Section 3. Finally, the main conclusions and recommendations are drawn in Section 4.

II. METHODOLOGY FOR ON-LINE DATA VALIDATION AND RECONSTRUCTION

The basic methodology for on-line data validation and reconstruction has been already presented and compared with other techniques in [7], [8]. In summary, it applies a set of consecutive validation tests to a given dataset (Fig. 2) to finally assign a certain quality level depending on the tests passed. The six different quality levels are the following:

- Test 1: This test allows to easily detect missing data due to any data acquisition or communication errors.
- Test 2: Based on a sensors operational measurement interval, i.e. maximum and minimum limits, values above or below this range are non-validated.
- Test 3: The trend level takes into account the data changes over time. This allows to detect unexpected changes.
- Test 4: This level allows to check the variables in a given unit, e.g. a flowmeter cannot measure a non-zero value if the valve located at the same pipe is totally closed.
- Test 5: This level evaluates the sensor's measurements against estimated data given by a time series (TS) model based on historical data.
- Test 6: This level checks the correlation between different neighboring sensors. It is defined as Spatial (SP) model, e.g. two flowmeters located in the inflow and outflow of the same pressured pipe without any element in the middle must measure the same quantity.

The raw data must be overcome all aforementioned tests to become validated data. If this is not the case, the raw data are labelled as non-validated and removed data and they must be reconstructed for any estimation technique.

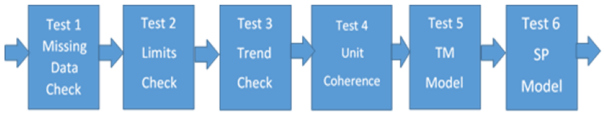

Fig. 2. The validation set of tests for the raw data.

Time Series (TS) models and Spatial (SP) models are used for two purposes, first to validate the raw data checking the difference (residuals) of raw data with the estimations of these models in tests 5 and 6 and second, to reconstruct the missing and non-validated data.

The selected TS model is the well-known and useful autoregressive (AR) model for time series [2]. The model is described for hourly-sampled data by the following difference equation, which describes a periodicity of 24 hours:

$$\hat{y}_{TS} = -a_1 y(k-1) - a_2 y(k-2) - a_3 y(k-3) \ldots - a_{24} y(k-24) \quad (1)$$

While the SP models are linear regression models relating different spatially correlated measurements in the system, stated as follows:

$$\hat{y}_{SPout}(k) = a y_{SPin}(k) + b \quad (2)$$

In equation 2, a and b are the parameters of the model to be calibrated. For instance, two flowmeters (in and out flow) installed in the same pipe and separated by a certain distance should measure a similar value if there is no element e.g. reservoir or node located between them i.e. a=1; b=0.

The model with the lowest Mean-Square-Error (MSE) over the m-estimations (usually m=24 is enough) previous to k is the selected candidate to estimate the invalid/missing k-sample:

$$MSE(k) = \frac{1}{m} \sum_{j=k-m}^{k} (y(j) - \hat{y}(j))^2 \quad (3)$$

III. IMPROVEMENTS OF THE METHODOLOGY

Any data fault detection technique has to minimize two opposing key commitments. Firstly, to reduce the number of false alarms or false positives (false non-validated raw data). Secondly, to decrease the missing alarms or false negatives (validating wrong raw data) and the quality of any data fault detection is measuring the distance to ideal case, zero of these false positives and negatives. The experience in the last years have shown that the fusion of all the possible homogeneous or heterogeneous information of the system (e.g. actuators or components states, electronic alarms, etc.) and the improvement of the models quality, helps to improve the performances of a given data fault detection technique. In this sense, we propose in this work the following improvements:

A. ARMA models with exogenous inputs

24 hours auto-regressive models represent the behavior of flows in the pipes connected directly to the consumption nodes without any reservoirs between them. However, when the pipes are controlled with valves and pumps to supply water to reservoirs (Fig. 3 and 4) the AR models are not good estimators (Fig. 5). Main reason is because the behavior is discontinuous and not periodic.

In this case, a significant improvement is to use autoregressive moving average (ARMA) models with exogenous variables such as the hourly mean opening of the valves, or the hourly pulse width modulation of the pumps or the deviation of the real level in the reservoirs respect to their set-points. In our application, we use hourly mean opening of the valves as exogenous variable (u) of the ARMA models in actual and previous hourly instants (k and k-1) of time:

$$\hat{y}_{TS} = -a_1 y(k-1) - a_2 y(k-2) - a_3 y(k-3) - \ldots - a_{24} y(k-24) + b_0 u(k) + b_1 u(k-1) \quad (4)$$

The proposed model, ARMA time series with valve exogenous variable used in the overall ATLL water network, shows a 10 fold improvement in respect to the fit of the prediction of previous AR models (Fig. 6).

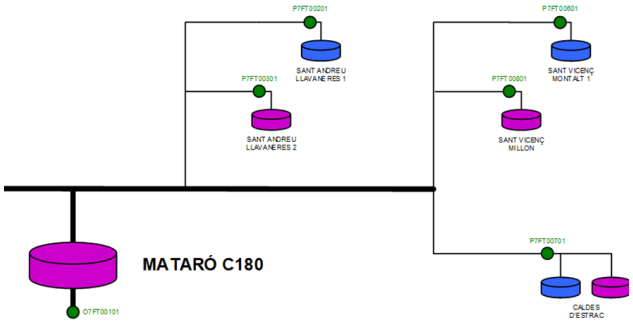

Fig. 3. A part of the ATLL network with several measured inflows to demand tanks.

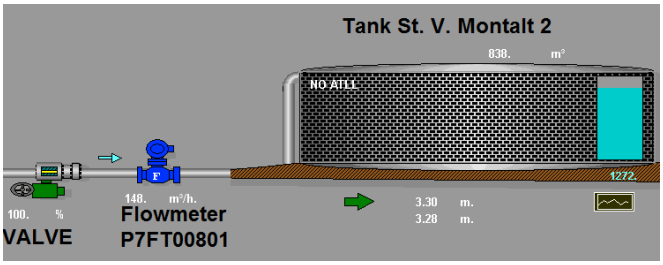

Fig. 4. Synoptic of the flowmeter P7FT00708 and the associated valve and tank.

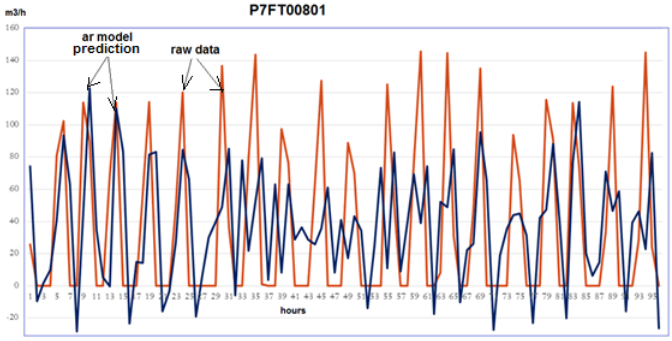

Fig. 5. An example of a 24 hours AR model one-hour step prediction (blue) compared with the raw data (red).

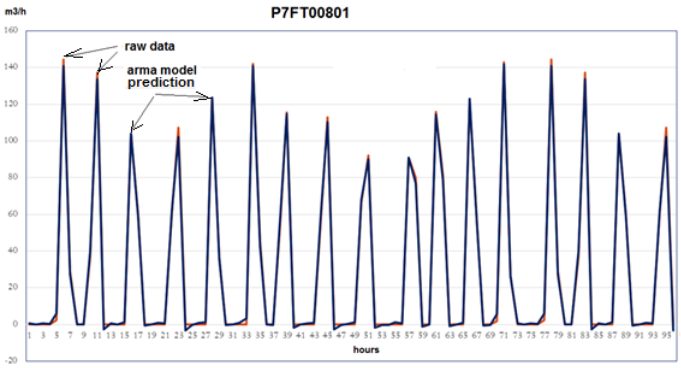

Fig. 6. An example of a 24 hours ARMA with valve exogenous input variable model one-hour step prediction (blue) compared with the same raw data (red).

B. Electronics alarms

The new tool takes into account the alarms received from remote flow-metering stations. A single station might contain one or more flow-measuring devices. In the case of alarm detection, the new tool verifies if it is within the group of alarms susceptible to compromise the measurement made by the device (for example voltage failure, flowmeter anomaly or PLC failure, among others). If so, this test invalidates the raw data and performs the value estimation process to present it as a proposal on its reconstruction.

C. Accumulated daily flows

Another interesting improvement here postulated is to consider the accumulated raw hourly data, where a daily value is compared with the accumulated one-hour estimations of the time series model. This is undertaken in a daily estimation for the same pipe.

Then, if both accumulated daily values are close, then all 24-hour raw data are validated. If not, the non-validation hourly data are maintained as alarms.

D. Comparison of the reconstruction and raw hourly

The non-validated raw data must be reconstructed by the best estimator. However, in some cases it has been evidenced that the appear that the new proposed values are significantly close to the raw data. In other words, in these cases both validated and non-validated data produce practically the same results. In these situations, it is not necessary to replace the raw data for the estimated data.

E. Improved performances

The improved validation and reconstruction tool has been checked with more than 200 flowmeters of ATLLs water network with more than 6 thousand hourly raw data of the ATLL during the last 6 months. Thereafter, the results of this study showed that less than 2% of the raw data are non-validated hourly data, when in the previous tool was in mean of 5% and the false alarms and missing alarms are less than 1% in mean, 4 times less than the initial tool.

IV. CONCLUSIONS

In this paper, a data analytics improved tool is presented. Its aim is to overcome sensor issues and eventually provide reliable data of a critical infrastructure system state, such as water networks. To accomplish this aim, a validation methodology based on a set of data quality tests allows to detect suspicious erroneous data. Then, a reconstruction scheme is defined using Temporal and Spatial Models to provide an estimation based on the model having the best fit. In addition, a set of new tests and models are proposed to minimize the false and missing alarms.

The continuous and intense collaboration of this Company with the CS^2AC Research Center in the last years have produced an interesting data analytics tool, useful for operation and maintenance plans in order to achieve an efficient management of this complex water transport network.

ACKNOWLEDGEMENTS

This work is supported by the project CICYT HARCRICS DPI2014-58104-R and by the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (2017-SGR-482). Our research is also continuously and strongly supported by the Company ATLL Concessionaria de la Generalitat de Catalunya, S.A.

REFERENCES

[1] M. Schutze, A. Campisano, H. Colas, W. Schilling, PA. Vanrolleghem (2004) Real time control of urban wastewater systemswhere do we stand today? J Hydrol 299(34), pages 335−348

[2] M. Mourad, J. Bertran-Krajewski (2002) A method for automatic validation of long time series of data in urban hydrology. Water Sci Technol 45(45), pages 263−270

[3] J. Quevedo, D. Garcia, V Puig, J Saludes, MA. Cuguer, S Espin, J Roquet and F. Valero (2017), Chapter 10 Sensor Data Validation and Reconstruction, Real-Time Monitoring and Operational Control of Drinking-Water Systems Book, Ed.Springer, 2017.

[4] D. Garcia, J. Quevedo, V. Puig, J. Saludes, S. Espin, J. Roquet, F. Valero (2015), Knowledge extraction from raw data in water networks. Application to the Barcelona supramunicipal distribution water network, A: New Developments in IT Water Conference. "2nd New Developments in IT Water Conference, 8-10 February 2015, Rotterdam (Holland)". Rotterdam: 2015, pages 1-9.

[5] D. Garcia, J. Quevedo, V. Puig, J. Saludes, S. Espin, J. Roquet, F. Valero (2014) Automatic Validation of Flowmeter Data in Transport Water Networks: Application to the ATLLc Water Network, IDEAL 2014: 15th International Conference, Salamanca, Spain, September 10-12, 2014. Proceedings, pages 118-125.

[6] J. Quevedo, V. Puig, J. Saludes, J. Pascual, S. Espin, J. Roquet, F. Valero (2014). Flowmeter data validation and reconstruction methodology to provide the annual efficiency of a water transport network: the ATLL case study in Catalonia Water science and technology: water supply. International Water Association (IWA). 14-2, pages 337-346. ISSN 1606-9749.

[7] J. Quevedo, J. Blanch, V. Puig, J. Saludes, S. Espin, J. Roquet (2010) Methodology of a data validation and reconstructions tool to improve the reliability of the water network supervision. In: International conference of IWA water loss, Sao Paolo, Brazil.

[8] D. Garca, J. Quevedo, V. Puig, MA. Cuguero(2014) Sensor data validation and reconstruction in water networks: a methodology and software implementation. In: 9th International conference on critical information infrastructure security, Limassol, Cyprus.