Automatic validation of flowmeter data in transport water networks: Application to the ATLLc water network

Diego Garcia[1], Joseba Quevedo[1], Vicenç Puig[1], Jordi Saludes[1], Santiago Espin[2], Jaume Roquet[2], and Fernando Valero[2]

[1] Advanced Control Systems (SAC), Universitat Politècnica de Catalunya (UPC), Campus de Terrassa, Rambla Sant Nebridi, 10 08222 Terrassa, Barcelona, Spain
{diego.garcia,joseba.quevedo, vicenc.puig,jordi.saludes}@upc.edu
[2] ATLL Concessionària de la Generalitat de Catalunya S.A. Sant Martí de l'Erm, 30. 08970 Sant Joan Despí, Barcelona, Spain

Abstract. In this paper, a methodology for data validation and reconstruction of flow meter sensor data in water networks is presented. The raw data validation is inspired on the Spanish norm (AENOR-UNE norm 500540). The methodology consists in assigning a quality level to data. These quality levels are assigned according to the number of tests that data have passed. The methodology takes into account not only spatial models but also temporal models relating the different sensors. The methodology is applied to real-data acquired from the ATLLc Water Network. The results demonstrate the performance of the proposed methodology in detecting errors in measurements and in reconstructing them.

1 Introduction

In any water network, a telecontrol system must acquire, store and validate data obtained in real time from sensors periodically (e.g. every few minutes) to achieve an accurate monitoring of the whole network.

The sensor measures a physical quantity and converts it into a signal that can be read by an instrument. The measuring system then converts the sensor signals to values aiming to represent certain "real" physical quantities. These values, known as "raw data", need to be validated before they can be used in a reliable way for several network water management tasks, namely: planning, investment plans, operations, maintenance and billing/consumer services and operational control (Quevedo et al., 2010a).

Frequent operation problems in the communication system between the set of the sensors and the data loggers, or in the telecontrol itself, generate missing data during some periods of time. Therefore, missing data should be replaced by a set of estimated data obtained from other spatially related sensors.

A second common problem is the lack of reliability of the water system meters (e.g. due to offset, drift and breakdowns) producing false flow data readings. These false data must also be detected and replaced by estimated data.

According to the nature of the available knowledge, different types of data validation can be implemented, with varying degrees of sophistication. In general, one may distinguish between elementary signal-based ("*low-level*") methods and model-based ("*higher level*") methods (see, e.g. Denoeux et al., 1997; Mourad & Bertrand-Krajeswski, 2002). Elementary signal based methods use simple heuristics and limited statistical information of a given sensor (Burnell, 2003; Jorgensen et al., 1998; Maul-Kotter & Einfalt, 1998).

Typically, these methods are based on validating either signal values or signal variations. In the signal value-based approach, data are assessed as valid or invalid according to two thresholds (a high one and a low one); outside these thresholds data are assumed invalid. On the other hand, methods based on signal variations look for strong variations (peaks in the curve) as well as lacks of variation (flat curve).

Model-based methods rely on the use of models to check the consistency of sensor data. This consistency check is based on computing the difference between the predicted value from the model and the real value measured by the sensors. Then, this difference, known as residual, will be compared with a threshold value (zero in the ideal case). When the residual is bigger than the threshold, it is determined that there is a problem in the sensor or in the system. Otherwise, it is considered that everything is working properly.

The result of data validation may be either a binary variable indicating whether the data are considered valid or not, or a continuous validity index interpreted as a degree of confidence in the data. Moreover, a sub-product of using model-based approaches for sensor data validation is that the prediction provided by the model can be used to reconstruct the faulty sensor.

2 Proposed Methodology

This section presents a methodology for data validation/correction of sensor data taking into account not only spatial models but also temporal models (time-series of each flow meter) and internal models of the several components in the local units (pumps, valves, flows, levels, etc.). This proposal allows for robust isolation of wrong sensor data which should be replaced by adequate estimated data. The methodology is applied to flow and level meters, since it exploits their temporal redundancy of data.

2.1 Data Validation Methodology

Raw data validation is inspired on the Spanish norm (AENOR-UNE norm 500540). The methodology is based on assigning a quality level to the considered sensor dataset. Quality levels are assigned according to the number of tests that have been passed, as represented in Figure 1. An explanation of each level is as follows:

- *Level 0*: The ***communications*** level simply monitors whether the data are recorded at the fixed sampling time taking used for the supervisory system to collect data (e.g. this could not be the case due to problems in the communication system).

- *Level* 1: The ***bounds*** level checks whether the data are inside their physical range. For example, the maximum values expected by the flow meters are obtained by pipes' maximum flow parameters.

- *Level 2:* The ***trend*** level monitors the data rate. For example, level sensor data cannot change more than several centimetres per minute in a real tank.

- *Level 3:* The models level uses three parallel models:

    - ***Local station related variables model***: the local station model supervises the possible correlation existing between the different variables in the same local station (i.e. flow and the command in the same valve).

    - ***Time series model***: This model takes into account a data time series for each variable (Blanch et al., 2009). For example, analysing historical flow data in a pipe, a time series model can be derived and the output of this time series model is used to compare and validate the recorded data.

    - ***Spatial model****:* The up-downstream model checks the correlation models between historical data of sensors located in different but near local stations in the same pipe (Quevedo et al., 2010b, 2012). For example, data of flow meters located at different points of the same pipe of the water network allows checking the sensor set reliability.

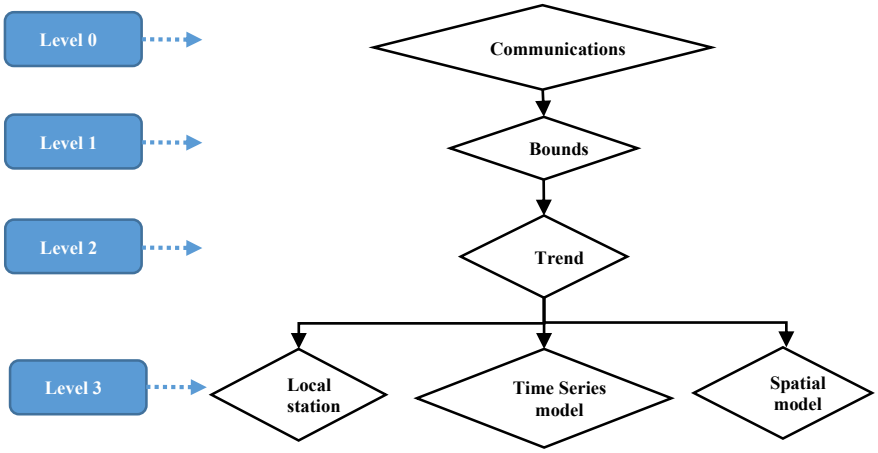

Fig. 1. Raw flowmeter data validation tests

2.2 Data Reconstruction Methodology

The levels 0, 1, 2, 3a, 3b and 3c in Figure 1 are used to validate the raw data from the sensors. If any of these levels does not validate the raw data, reconstructed data is provided by the best of the three models considered in level 3 (see Section 3). The best of these three models considered is used to reconstruct the non-validated data at time *k*, according to their Mean Square Error (MSE)

$$MSE = \frac{1}{L}\sum_{i=k-L}^{k-1}(y(i)-\hat{y}(i))^2 \tag{1}$$

where y is the non-validated data, $\hat{y}$ is the reconstructed data and L is the number of previous data samples used to compute the MSE.

3 Models for data validation and reconstruction

In this section, the different models used for data validation and reconstruction will be described.

3.1 Spatial Model

The water network model constitutive elements and their basic relationships are introduced in this section. The mass balance expression for the *i*-th tank is stated as a discrete-time difference equation

$$y_i(k+1) = y_i(k) + \frac{\Delta t}{A_i}\left(q_{in_i}(k) - q_{out_i}(k)\right) \tag{2}$$

where $y_i(k)$ is the tank level, A_i is the tank section, $q_{in_i}(k)$ is the manipulated inflow and $q_{out_i}(k)$ is the outflow, which may include manipulated tank outflow and consumer demands, both given in m^3/s.

Moreover, in a water network system nodes are represented as intersections of mains, which mass balance may be expressed as the static equation

$$\sum_i q_{in_i}(k) = \sum_i q_{out_i}(k) \tag{3}$$

where, similarly to Equation (2), $q_{in_i}(k)$ and $q_{out_i}(k)$ correspond to the inflow and outflow of the *i*-th subnet node, also given in m^3/s.

3.2 Time-series Model

Usually the flow in the pipes have a daily repetitive behaviour that can modelled using a Time Series (TS) model. TS models take advantage of the temporal redundancy of the measured variables. Thus, for each sensor with periodic behaviour, a TS model can be derived:

$$\hat{y}_{ts}(k) = g(y_m(k-1),...,y_m(k-L)) \tag{4}$$

where g is the TS model, for data exhibiting a periodicity of L samples.

The aggregate hourly flow model may be built on the basis of a time series modelling approach using ARIMA modelling (Box &Jenkins, 1970) or using Holt-Winters Time Series Model. A TS analysis is carried out on several daily aggregate series, which consistently showed a daily seasonality, as well as the presence of deterministic periodic components. A general expression for the hourly time series model can be derived using three main components (Quevedo, 2010a):

- One-day-period oscillating signal with zero average value to cater for cyclic deterministic behaviour, implemented using a second-order (two-parameter) model with two oscillating modes, in s-plane $s_{1-2}=+/-2\pi/24\ j$ or equivalently, in z-plane: $z_{1-2} = cos(2\ \pi/24)+/- sin(2\ \pi/24)j$. The oscillating polynomial is

$$y(k)= 2cos(2\pi/24)\ y(k-1) - y(k-2) \tag{5}$$

- An integrator that taking into account possible trends and non-zero mean values of the flow data is described by

$$y(k) = y(k-1) \tag{6}$$

- An autoregressive component of order 21 to consider the influence of previous values within the series is considered

$$y(k) = -a_1y(k-1) - a_2y(k-2) - a_3y(k-3)- \ ... \ - a_{21}y(k-21) \tag{7}$$

Component (6) plus the orders of the two components presented in (4) and (5) leads to a final order of 24 (i.e. number of samples within a day for sampling period of 1 h) for the obtained model with the following structure

$$\begin{aligned} y_p(k) = &-b_1y(k-1) - b_2y(k-2) - b_3y(k-3) - b_4y(k-4) - b_5y(k-5) \\ &- b_6y(k-6) - \ ...– b_{24}y(k-24) \end{aligned} \tag{8}$$

Thus, this TS model of order 24 is consistent with the daily pattern (see Figure 3).

4 Application to the ATLLc Water Network

The methodology presented in previous section has been applied to ATTLc Water Network. The methodology presented in previous section exploits the "*spatial redundancy*" existing in the networks by means of spatial models relating upstream and downstream flow meters. The methodology is applied through the following steps to search outliers and reconstruct data when they are found. First, in case of two flow meters in the same pipe, a linear model given by

$$\sum_{j=1}^{n_{in}} q_{in_j}(t) = K \sum_{l=1}^{n_{out}} q_{out_l}(t) + M \tag{9}$$

is found, where $\sum_{j=1}^{n_{in}} q_{in_j}(t)$ and $\sum_{l=1}^{n_{out}} q_{out_l}(t)$ are the hourly flows measured by the input and output sensors, respectively (see Fig. 2). If there is a tank between the input and output sensors, data from the sensor level is included in the input sensor data.

Fig. 2. Two flowmeters in the same pipe

Parameters K and M are estimated by using real data and using the least-squares method. In the ideal case, those parameter should be $K = 1$ and $M = 0$, respectively. Then, with the residuals obtained by this model and using a threshold of 3σ (three times the standard deviation), outliers can be found and removed.

Additionally, a 24 hours ARIMA time series models are found for both input and output sensors. They are used to determine if the outlier values belongs to the input or to the output sensor. Finally, the invalidated data are been reconstructed by the model that provides better prediction according to MSE in (1).

Figure 3 shows the results of a flow meter with a spatial model including two level sensors in case of a tank with two bodies and five flow meters (Figure 4) and two time series models for the reconstruction phase. Most of the data invalidated by limits test and valves flow meter incoherence test have been reconstructed by the spatial model presented in Figure 5.

5 Conclusions

In this paper, a methodology for automatic data validation and reconstruction of sensor data of the water network has been developed taking into account not only spatial models but also temporal models (time series of each flowmeter) and internal models of the several components in the local units (pumps, valves, flows, levels, etc.). The methodology consists in assigning a quality level to data and quality levels are assigned according to the number of tests that have been passed.

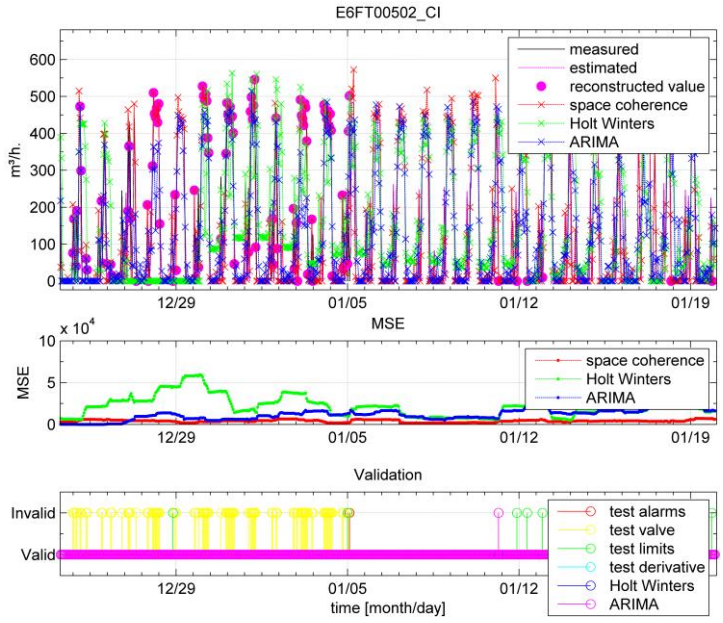

Fig. 3. Validation and reconstruction results of the flowmeter E6FT00502

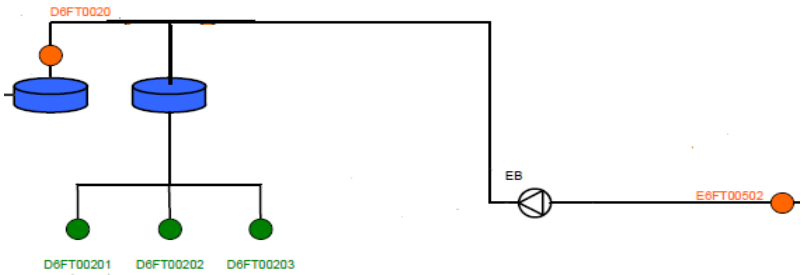

Fig. 4. Spatial relationship between 5 flowmeters and 2 level sensors of a tank

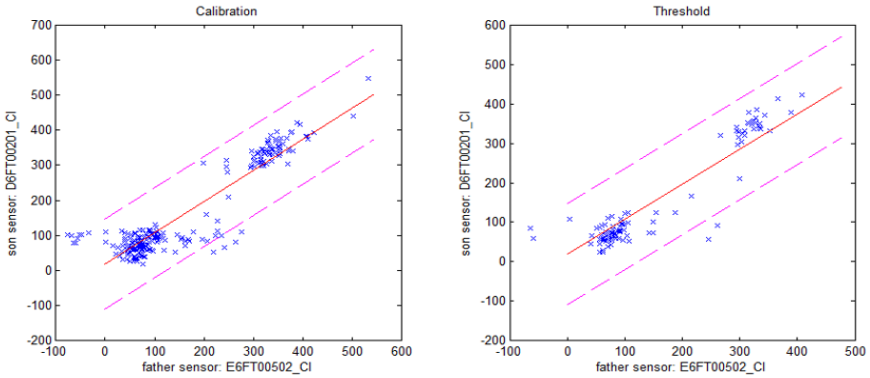

Fig. 5. Calibration and threshold of the spatial model

Acknowledgements

This work is partially supported by CICYT SHERECS DPI-2011-26243 of the Spanish Ministry of Education, by EFFINET grant FP7-ICT-2012-318556 of the European Commission and by AGAUR Doctorat Industrial 2013-DI-041.

References

1. Blanch, J.; Puig, V.; Saludes, J.; Quevedo, J. "ARIMA Models for Data Consistency of Flowmeters in Water Distribution Networks". 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes. (2009). pp. 480 – 485
2. Box, G.E.P., Jenkins,G. M. Time series analysis forecasting and control. Holden-Day (1970).
3. Burnell D. “Auto-validation of district meter data” Advances in Water Supply Management- Maksimovic, Butler, Memon eds., Swets & Zeitlinger Publish(2003).
4. Denoeux, T., Boudaoud, N., Canu, S., Dang, V.M., Govaert, G., Masson, M., Petitrenaud, S., Soltani, S. (1997). “High level data fusion methods”. Technical Report CNRS/EM2S/330/11-97v1.0, Université de Technologie de Compiègne (1997).
5. Jörgensen H.K, Rosenörn S., Madsen H., Mikkelsen P. “Quality control of rain data used for urban run-off systems”.Water Science and Technology, 37, 113-120 (1998)
6. Maul-Kötter, B., Einfalt T. (1998). “Correction and preparation of continuously meas-ured rain gauge data: a standard method in North Rhine-Westphalia”. Water Science and Technology, 37(11), pp 155-162. (1998)
7. Mourad, M., Bertrand-Krajeswski, J.L. “A method for automatic validation of long time series of data in urban hydrology”. Water Science and Technology Vol. 45, No 4-5, pages 263-270, (2002)
8. Quevedo, J.; Pascual, J.; Puig, V.; Saludes, J.; Espin, S.; Roquet, J. "Data validation and reconstruction of flowmeters to provide the annual efficiency of ATLL transport water network in Catalonia". New Developments in IT & Water. (2012)
9. Quevedo, J., Puig, V., Cembrano, G., Blanch. J. “Validation and reconstruction of flowmeter data in the Barcelona water distribution network”. Control Engineering Practice Journal, 18 (6), pp. 640-651. (2010a)
10. Quevedo, J.; Blanch, J.; Puig, V.; Saludes, J.; Espin, S.; Roquet, J. “Methodology of a data validation and reconstruction tool to improve the reliability of the water network supervision”, Water Loss Conference 2010, Sao Paulo, Brazil. (2010b)
11. UNE, “Redes de estaciones meteorológicas automáticas: directrices para la validación de registros meteorológicos procedentes de redes de estaciones automáticas: validación en tiempo real”. AENOR UNE 500540 (2004).