Knowledge extraction from raw data in water networks: Application to the Barcelona supramunicipal water transport network

Joseba Quevedo[*], Vicenç Puig[*], Diego Garcia[*], Josep Pascual[*], Jordi Saludes[*], Santiago Espin[**], Jaume Roquet[**], and Fernando Valero[**]

[*] Center of Supervision, Safety and Automatic Control (CS2AC), BarcelonaTech (UPC), Campus de Terrassa, Sant Nebridi, 10 08222 Terrassa, Barcelona, Spain, {joseba.quevedo, vicenc.puig, diego.garcia, josep.pascual, jordi.saludes}@upc.edu

[**] ATLL Concessionària de la Generalitat de Catalunya S.A. Sant Martí de l'Erm, 30. 08970 Sant Joan Despí, Barcelona, Spain

Abstract

Critical Infrastructure Systems (CIS) such as the case of potable water transport network are complex large-scale systems, geographically distributed and decentralized with a hierarchical structure, requiring highly sophisticated supervisory and real-time control (RTC) schemes to ensure high performance achievement and maintenance when conditions are non-favorable due to e.g. sensor malfunctions (drifts, offsets, problems of batteries, communications problems,...).

Once the data are reliable, a process to transform these validated data into useful information and knowledge is key for the operating plan in real time (RTC). And moreover, but no less important, it allows extracting useful knowledge about the assets and instrumentation (sectors of pipes and reservoirs, flowmeters, level sensors, ...) of the network for short, medium and large term management plans.

In this work, an overall analysis of the results of the application of a methodology for sensor data validation/reconstruction to the ATLL water network in the city of Barcelona and the surrounding metropolitan area since 2008 until 2013 is described. This methodology is very important for assessing the economic and hydraulic efficiency of the network.

Keywords: Fault diagnosis, Data validation, Water networks, Network efficiency.

1. INTRODUCTION

The competence to detect any malfunction of the information system, the capability to determine which is the origin and severity of the problem, which is the faulty device and which are the wrong data, and finally, the ability to estimate or reconstruct the wrong data by other instruments combined with models, are key functions to be included in CIS systems to keep their safe integrity.

To deal with this problem, the use of an on-line fault diagnosis system able to detect and to isolate faults and correct them by activating different kind of techniques e.g. data validation / reconstruction of sensor faults is desirable . Furthermore, the fault diagnosis process intends to identify which fault is causing the monitored events, such as the case of several contributions (Mourad et al., 2002), (Burnell, 2003) in potable water networks or (Jorgensen et al., 1998), (Maul-Kotter et al., 1998) and (Schultze et al., 2004) in urban waste water networks.

In Quevedo et al. (2009), a methodology to compute network efficiency taking into account raw flowmeter data and the network topology is presented. This methodology allows to take into account the estimated flowmeter uncertainty is when evaluating the network water balance, obtaining confidence intervals for key performance indices plus the economic efficiency corresponding to each zone. Moreover, the overall network efficiency is obtained and analysed helping to improve instrumentation (i.e. sensor location, recalibration) and to define new plans for the network maintenance to locate leaks in pipes. Furthermore, in Quevedo et al. (2010, 2012) and in S.Espin et al. (2012), a more general tool is developed to check raw flowmeter and level sensor data consistency taking into account not only spatial models but also temporal models (i.e. flowmeter time series) and internal models corresponding to several components in local units (e.g. pumps, valves, flows, levels, etc.). The latter approach allows the robust isolation of wrong data that must be replaced by valid estimated data.

2. PROPOSED METHODOLOGY

This methodology consists of the following steps:

2.1.- Flowmeter data validation tests

A methodology is developed for data validation and reconstruction of sensors installed in the water network. This methodology takes 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.) allowing the robust isolation of the wrong data that it is replaced by adequate estimated data.

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

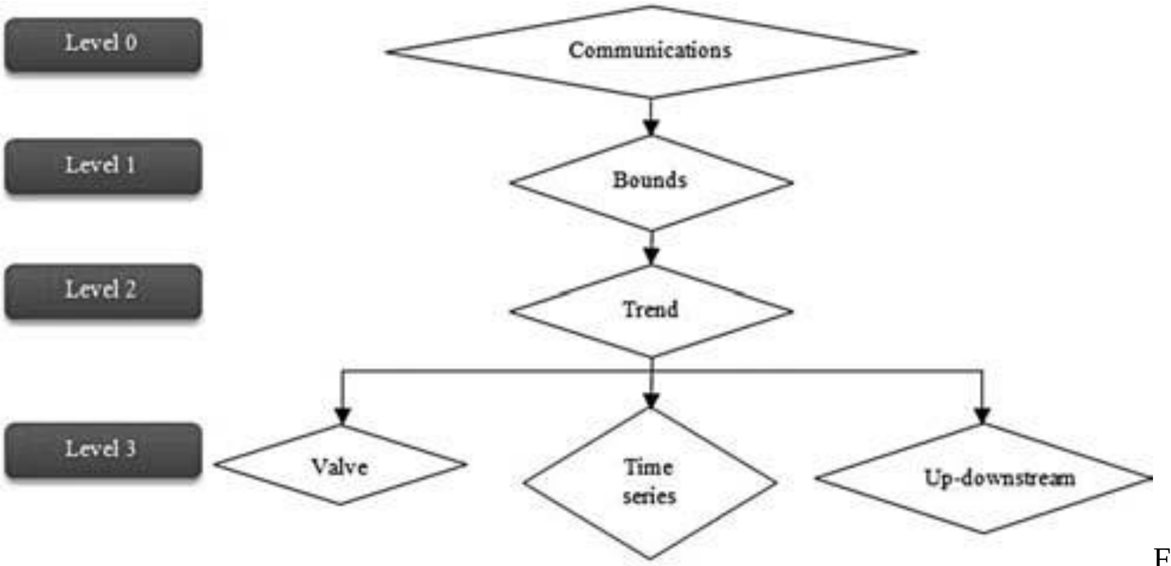


Figure 1. Data validation diagram

An explanation of each level is as follows:

Level 0: The communications level simply monitors whether the data are recorded or not taking into account that the supervisory system is expected to collect data at a fixed sampling time (e.g. due to problems in the sensor or 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 flowmeters 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 centimeters 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 opening valve command in the same pipe or pump element).

- 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 the model is used to compare and to validate the recorded data.

- Spatial model: The spatial model checks the correlation between historical data of sensors located in different but near local stations in the same pipe (Quevedo et al., 2010, 2012), as e.g. data from flowmeters located at upstream/downstream points of the same pipe in a transport water network.

A decision tree method has been developed to invalidate data in level 3. This method detects invalid data from the result of the three models. In particular, the spatial model is very useful not only to detect problems in sensor data but also to detect leakages in pipes and to compute the balance in transport network sectors.

Once the data have passed all test levels, if any data inconsistency is detected, next step is to isolate the fault by combining the previous tests. For instance, if the three tests detect an inconsistency in a set of two flowmeters, the system analyses the historical data and other features of both flowmeters to diagnose the cause of the problem and to identify the sensor in faulty operation. And then, all the data of this faulty sensor are replaced by the data of a healthy sensor located in the same pipe.

2.2.- Wrong data reconstruction based on model estimations

The levels 0, 1, 2, 3a, 3b and 3c in Figure 1 are used to validate the raw data coming 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. The structure of these models is further explained in Section 3.
The best of these three models considered is used to reconstruct by the non-validated data at time *k*, according to their Mean Square Error (MSE):

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

where *y* is the non-validated data, *ye* is the reconstructed data and *L* is the number of previous data samples used to compute the MSE.

2.3.- Sector model generation based on filtered data

A water transport network can be divided into a set of interconnected sectors. Usually, a sector is composed of demand nodes, tanks, pipes and flowmeters. Flowmeters measure sector inputs and outputs. External demand is considered as an output. In this paper, pipes are considered pressurized. Hence, no delays are considered in pipes. The sector model is based on mass balance equations and the following hypotheses are assumed:

- Flowmeters are maintained and calibrated by the water management company following a maintenance program (which is the case for ATLL Company network in Catalonia).
- Flowmeters have been installed and operated fulfilling the manufacturer recommendations, thus avoiding systematic measurement errors (‘unbiased’).
- Random errors are normally distributed with zero mean (‘normal’).
- Random errors between measurement instruments are uncorrelated (‘independence’).

Given a sector with several flowmeters at both input and output, the model is:

$$\sum_{j=1}^{n\ input} (F_j(t)) = K \sum_{l=1}^{n\ output} (F_l(t)) \; + \quad M \tag{2}$$

where F_j and F_l are the daily flows measured by input and output sensors, respectively. Parameters K and M are determined using the least squares parameter estimation approach and real data. In the ideal case, they should be equal to $K= 1$ and $M= 0$, respectively.

2.4.- Flowmeter data inaccuracy computation

Considering that input and output flowmeters have errors, named respectively e_j and e_l, Equation (2) is rewritten as follows:

$$\sum_{j=1}^{n\ input} (F_j(t) + \, e_j(t)) = K \sum_{l=1}^{n\ output} \big(F_l(t) + e_l(t)\big) + \, M \tag{3}$$

Thus, model residuals are given by:

$$e(t) = \sum_{j=1}^{n\ \ input} (F_j(t)) - K \sum_{l=1}^{n\ output} \big(F_l(t)\big) - \, M \; = \sum_{j=1}^{n\ input} e_j(t) - K \sum_{l=1}^{n\ output} e_l(t) \tag{4}$$

$$e(t) \sim \; N\big(0, K^2 n_{ouput}\sigma_{output}^2 + n_{input}\sigma_{input}^2\big) \tag{5}$$

Consider that input and output sensors have the same characteristics, i.e. it is assumed that $\sigma_{input} = \sigma_{output} = \sigma$. If main sectors are close to the ideal case K=1. Then, the residual error e(t) is normally distributed $N\big(0, \sigma_{fit}^2\big)$ with $\sigma_{fit}{}^2 = (n_{input} + n_{output})\sigma^2$ and the variance of the error of each flowmeter can be estimated by:

$$\sigma = \frac{\sigma_{fit}}{\sqrt{(n_{input} + n_{output})}} \tag{6}$$

Given a confidence interval α with a standard deviation radius λ(α), the relative error is:

$$Flow\ meter\ error\ \% = \frac{\lambda(\alpha)\sigma}{mean(flowmeter)} \tag{7}$$

2.5. Sectors, zones and the whole network efficiency computation

A sector, a zone or whole network efficiency can be computed as the ratio between the network output flow V_{out} and the network input flow V_{in}:

$$R = \frac{V_{in}}{V_{out}} \tag{8}$$

As these two quantities are affected by flowmeter errors, the network efficiency calculation has an uncertainty that can be quantified by means of the following interval:

$$[R_{min}, R_{max}] = \left[\frac{V_{out} - \lambda(\alpha)\sigma_{output}\sqrt{365n_{output}}}{V_{in} + \lambda(\alpha)\sigma_{input}\sqrt{365n_{input}}}, \frac{V_{out} + \lambda(\alpha)\sigma_{output}\sqrt{365n_{output}}}{V_{in} - \lambda(\alpha)\sigma_{input}\sqrt{365n_{input}}}\right] \tag{9}$$

3. RESULTS

The proposed methodology has been applied to ATLL network in the last 6 years, from 2008 to 2013. ATLL network supplies drink water to 4.5 million inhabitants in Catalonia (Spain) with an approximate yearly demand of 240 cubic hectometers through 829 km of piping with diameters up to 3000 mm and its responsibility ends at municipal head tanks (Figure 2).

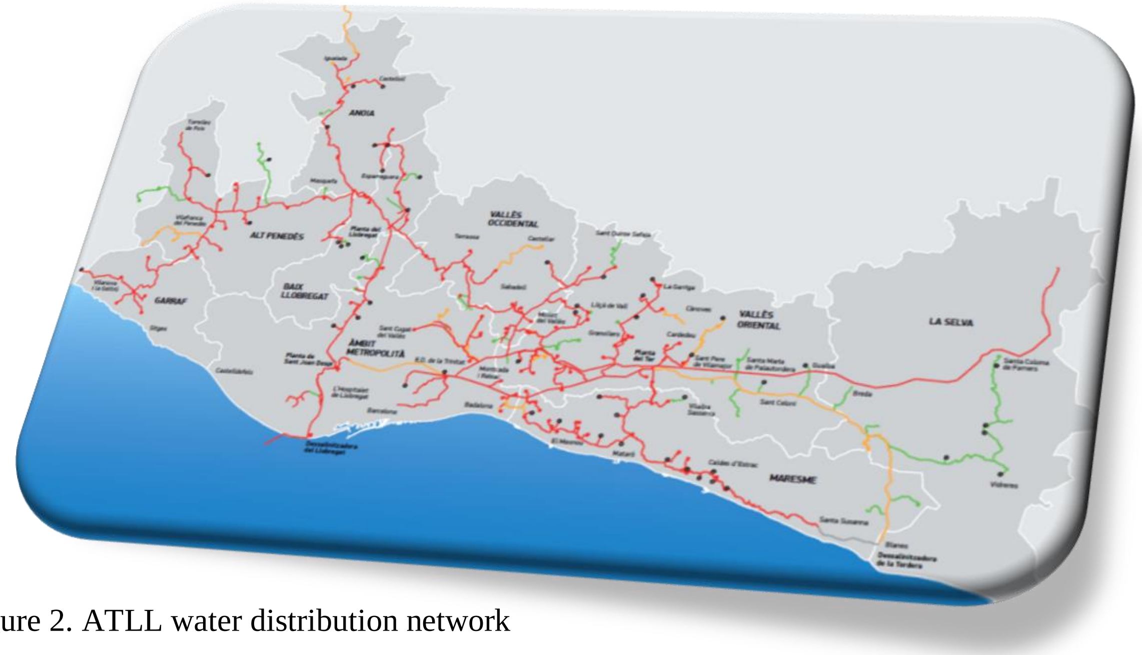


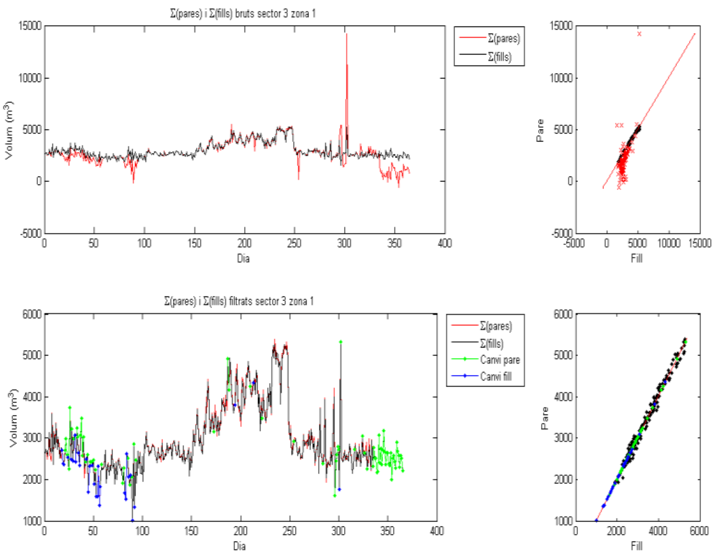
Figure 2. ATLL water distribution network

During the considered period, 6 annual reports have been developed (analysing all the daily data per year of more than 200 flowmeters and 115 level sensors in the tanks) to provide the hydraulic and economic efficiency of more than 90 sectors, 10 zones and the whole ATLL network. The concept of network hydraulic efficiency analyzed in this study is calculated as the ratio between the volume of authorized consumption (CA) and the volume of water entering the network (VED). The CA includes the sum of consumption measured or not, but which have been authorized. On the other hand, the economic efficiency is calculated as the ratio between the volume of water billed division (VAF) and the volume of water entering the network (VED).

For this reason, the methodology requires a preliminary analyzes of incidents due to unmetered consumption that have been authorized (like emptying tanks or pipes, etc ..) and other events (failures, power failures, problems communications, ..) that have been identified, documented and corrected by staff ATLL throughout 2013. This pre-analysis is manually developed by the university research group taking into account all the information recorded by ATLL operators validating, or not, these unmetered consumptions. After that, all the raw data of each sector, zone or whole network are validated and reconstructed allowing finally obtaining several index of performances: interval hydraulic efficiency, imprecision of the sensors, quality of the raw data regarding the number of non-validated raw data. Figure 3 shows an example of a sector information.

ZONE 1 - SECTOR 3

Upstream flowmeters and level sensors: N9FT00402, N9FT00201, N9LT00201, N9LT00401
Downstream flowmeters: N9FT00403, N9FT00401
Yearly volume of upstream [m3]: 1.080.272
Yearly volume total fills [m3]: 1.074.280
Model: P = -0.397 + 1.006 · F
Number of non-validated raw data: 105
Precision of up and down stream flowmeters [%]: 2.860 and 2.876
Raw hydraulic efficiency [%]: 106.910
Filtered hydraulic efficiency [%]: 99.445
Interval of filtered hydraulic eficiency [%]: [99.024 - 99.869]



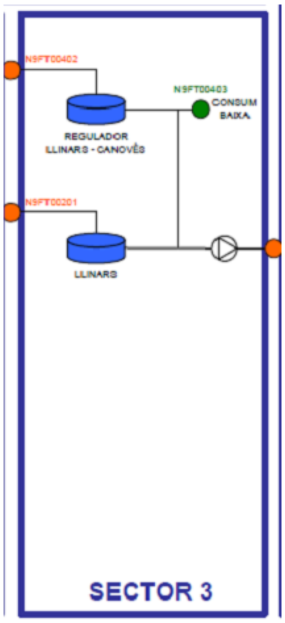
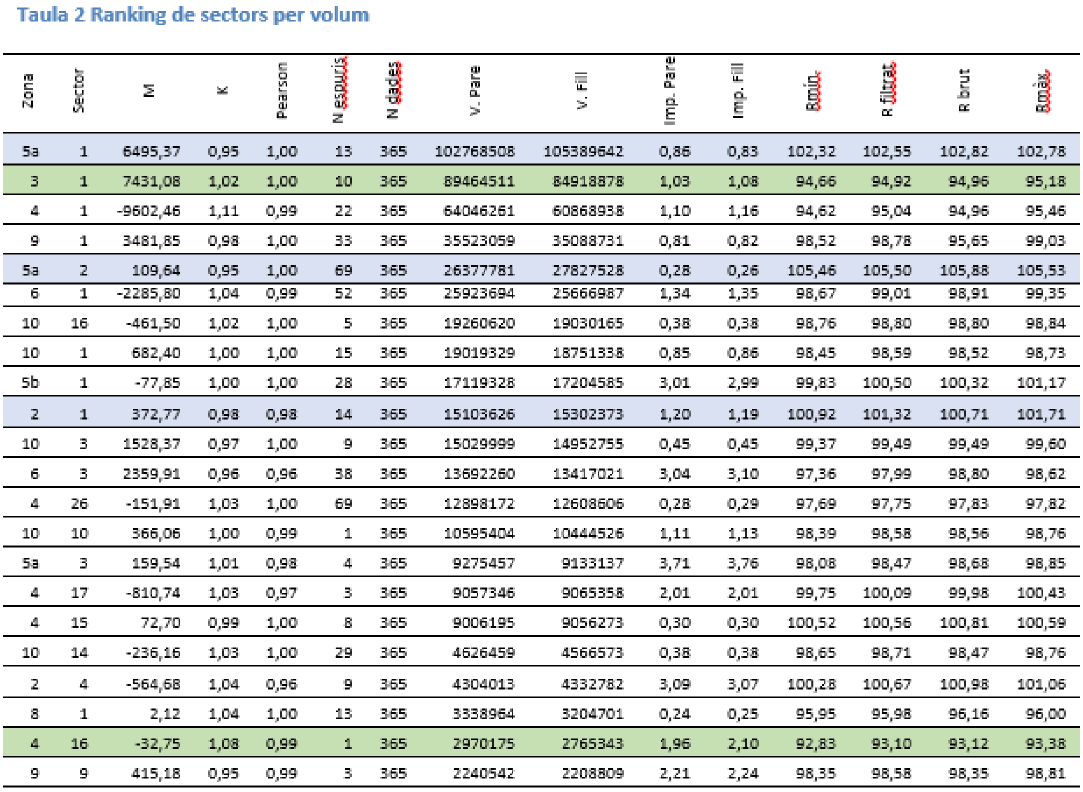


Figure 3. The results of a sector

The annual report also contains several ranking of all the 90 sectors ordered from the larger to the smaller volume, by efficiency, by sensor imprecision and by the quality of the data. The Figure 4 shows a piece of ranking table according to the volume per sector. These rankings are very useful to extract recommendations for ATLL Company and a list of recommendations is proposed every year as a report. Moreover, the actions developed by the ATLL Company from the previous recommendations are also studied in this report.



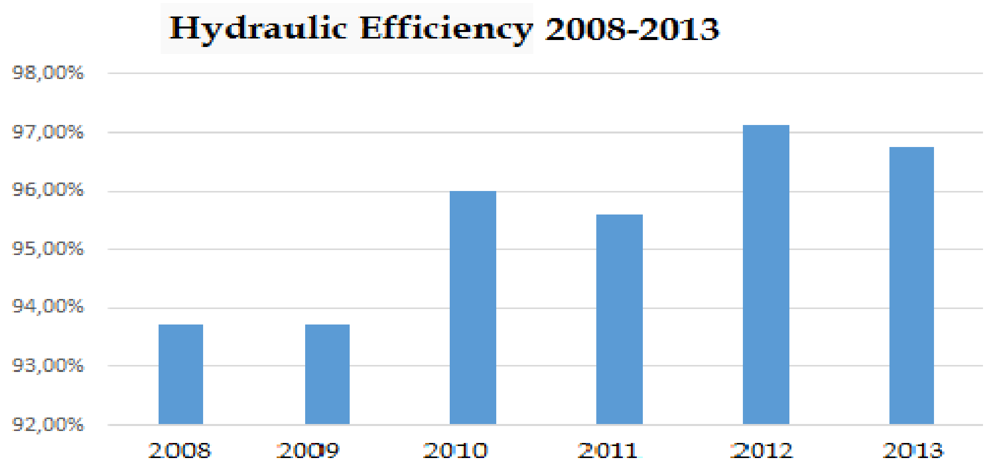
Taula 2 Ranking de sectors per volum

Zona	Sector	M	K	Pearson	N espuris	N dades	V. Pare	V. Fill	Imp. Pare	Imp. Fill	Rmín.	R filtrat	R brut	Rmàx.
5a	1	6495,37	0,95	1,00	13	365	102768508	105389642	0,86	0,83	102,32	102,55	102,82	102,78
3	1	7431,08	1,02	1,00	10	365	89464511	84918878	1,03	1,08	94,66	94,92	94,96	95,18
4	1	-9602,46	1,11	0,99	22	365	64046261	60868938	1,10	1,16	94,62	95,04	94,96	95,46
9	1	3481,85	0,98	1,00	33	365	35523059	35088731	0,81	0,82	98,52	98,78	95,65	99,03
5a	2	109,64	0,95	1,00	69	365	26377781	27827528	0,28	0,26	105,46	105,50	105,88	105,53
6	1	-2285,80	1,04	0,99	52	365	25923694	25666987	1,34	1,35	98,67	99,01	98,91	99,35
10	16	-461,50	1,02	1,00	5	365	19260620	19030165	0,38	0,38	98,76	98,80	98,80	98,84
10	1	682,40	1,00	1,00	15	365	19019329	18751338	0,85	0,86	98,45	98,59	98,52	98,73
5b	1	-77,85	1,00	1,00	28	365	17119328	17204585	3,01	2,99	99,83	100,50	100,32	101,17
2	1	372,77	0,98	0,98	14	365	15103626	15302373	1,20	1,19	100,92	101,32	100,71	101,71
10	3	1528,37	0,97	1,00	9	365	15029999	14952755	0,45	0,45	99,37	99,49	99,49	99,60
6	3	2359,91	0,96	0,96	38	365	13692260	13417021	3,04	3,10	97,36	97,99	98,80	98,62
4	26	-151,91	1,03	1,00	69	365	12898172	12608606	0,28	0,29	97,69	97,75	97,83	97,82
10	10	366,06	1,00	0,99	1	365	10595404	10444526	1,11	1,13	98,39	98,58	98,56	98,76
5a	3	159,54	1,01	0,98	4	365	9275457	9133137	3,71	3,76	98,08	98,47	98,68	98,85
4	17	-810,74	1,03	0,97	3	365	9057346	9065358	2,01	2,01	99,75	100,09	99,98	100,43
4	15	72,70	0,99	1,00	8	365	9006195	9056273	0,30	0,30	100,52	100,56	100,81	100,59
10	14	-236,16	1,03	1,00	29	365	4626459	4566573	0,38	0,38	98,65	98,71	98,47	98,76
2	4	-564,68	1,04	0,96	9	365	4304013	4332782	3,09	3,07	100,28	100,67	100,98	101,06
8	1	2,12	1,04	1,00	13	365	3338964	3204701	0,24	0,25	95,95	95,98	96,16	96,00
4	16	-32,75	1,08	0,99	1	365	2970175	2765343	1,96	2,10	92,83	93,10	93,12	93,38
9	9	415,18	0,95	0,99	3	365	2240542	2208809	2,21	2,24	98,35	98,58	98,35	98,81

Figure 4. A piece of the volume per sector ranking table

Finally, the annual report provides the economic and hydraulic interval efficiencies of the whole network as well as the comparison with the results in previous years. As, it was clearly see in the Figure 5 for hydraulic efficiency, this indicator has been improved from 2008 to 2013 more than 2%.

Table 1. Historic evolution of the hydraulic efficiency of ATLL network (2008-2013)



4. CONCLUSIONS

In this work, an overall analysis of the results of the application of a methodology for sensor data validation/reconstruction to ATLL water network in the city of Barcelona and the surrounding metropolitan area since 2007 until 2013 is described. This methodology is very important for assessing the economic and hydraulic efficiency of the network. The proposed methodology has been applied to ATLL water network in the last 6 years, from 2008 to 2013 with satisfactory results. In particular, the hydraulic efficiency has been improved from 2008 to 2013 more than 2% as a result of the application of the proposed methodology and derived actions.

References

Burnell D. (2003) “Auto-validation of district meter data” Advances in Water Supply Management-Maksimovic, Butler, Memon eds., Swets & Zeitlinger Publishers.

Espin S. and Roquet J. (2012) “ Systematic control of efficiency and water losses reduction for Barcelona supramunicipal distribution network of 4.5 millions of inhabitants”, New Developments in IT & Water Conference, Amsterdam.

Jörgensen H.K, Rosenörn S., Madsen H., Mikkelsen P. (1998) “Quality control of rain data used for urban run-off systems”.Water Science and Technology, 37, 113-120.

Maul-Kötter, B., Einfalt T. (1998) “Correction and preparation of continuously measured rain gauge data: a standard method in North Rhine-Westphalia”. Water Science and Technology, Vol. 37(11), pp. 155-162.

Mourad, M., Bertrand-Krajeswski, J.L. (2002) “A method for automatic validation of long time series of data in urban hydrology”. Water Science and Technology Vol. 45, No 4-5, pp. 263-270.

Quevedo, J., Blanch, J., Saludes, J., Puig, V. & Espin, S. (2009). “Methodology to determine the drinking water transport network efficiency based on interval computation of annual performance. In Water Loss Conference 2009, Cape Town, May 2009.

Quevedo, J., Blanch, J., Puig, V., Saludes, J., Espin, S. & Roquet (2010). “Methodology of a data validation and reconstruction tool to improve the reliability of the water network supervision”. Water Loss Conference 2010, Sao Paulo, Brazil.

Quevedo, J.; Pascual Pañach, Josep; Puig Cayuela, Vicenç; Saludes Closa, Jordi; Espin, S; Roquet, J. (2012) “Data validation and reconstruction of flowmeters to provide the annual efficiency of ATLL transport water network”, New Developments in IT & Water Conference, Amsterdam

Quevedo J., J. Pascual, V. Puig, J. Saludes, R. Sarrate, A. Escobet, S. Espin and J. Roquet (2014). “Flowmeter data validation and reconstruction methodology to provide the annual efficiency of a water transport network”. Water Science & Technology: Water Supply. Vol. 14.2.

Schütze, M.,Campisano,A.,Colas,H.,Schilling,W.,Vanrolleghem,P.A.,(2004). “Real time control of urban waste water systems—where do we stand today?”. Journal of Hydrology 299,335–348.

UNE 500540 (2004) “Redes de estaciones meteorológicas automáticas: directrices para la validación de registros meteorológicos procedentes de redes de estaciones automáticas: validación en tiempo real”. AENOR.