Flowmeter data validation and reconstruction methodology to provide the annual efficiency of a water transport network: the ATLL case study in Catalonia

ABSTRACT
The object of this paper is to provide a flowmeter data validation/reconstruction methodology that determines the annual economic efficiency of a water transport network. In this paper, the case of Aigües Ter Llobregat (ATLL) company, which manages 80% of the overall water transport network in Catalonia (Spain), will be used for illustrating purposes. Economic network efficiency is based on daily data set collected by the company using about 200 flowmeters of the network. Data collected using these sensors are used by remote control and information storage systems and they are stored in a relational database. All information provided by ATLL is analysed to detect inconsistent data using an automatic data validation method deployed in parallel with the network efficiency evaluation. As a result of the validation process, corrections of flow measurements and of billed water volume are introduced. Results from ATLL water transport network corresponding to year 2010 will be used to illustrate the approach proposed in this paper.

Key words | data validation, data reconstruction, efficiency analysis, water network

INTRODUCTION
Water network performance can be measured by the economic performance as a ratio between the annual water volume entering the network \( (V_{in}) \) and the billed water volume measured by flowmeters in the network \( (V_{out}) \).

In Aigües Ter Llobregat (ATLL) network, the telecontrol system acquires, stores and validates data from different flow and level sensors (collected at different sampling rates: 10 min, 1 hour, 1 day) to achieve accurate monitoring of the whole network. Frequent operating problems in the communication system between the set of the sensors and the data logger, or in the telecontrol itself, generate missing data during certain periods of time. Stored data are sometimes uncorrelated and of no use for historic records. Therefore, missing data must be replaced by estimated data. A second common problem is the lack of data sensor reliability (offset, drift, breakdowns, etc.) leading to false measurements.

Data sensors are used for several complex system management tasks such as planning, investment plans, operations, maintenance, security and operational control (Quevedo et al. 2010b). Therefore, wrong data must be detected and replaced by estimated data. Recorded data quality is a basic requirement to determine water network efficiency and further assess the non-revenue water of the network (Lambert 2003).

The study presented in this paper covers the performance analysis of the full network and of the 99 sectors belonging to the ATLL company network, as well as of the 10 zones containing those sectors. This study identifies sectors with the lowest economic performance. It also proposes where new flowmeters should be installed for a better assessment of the network performance by defining new zoning and sectorisation, and it helps locating which flowmeters need to be recalibrated.

doi: 10.2166/ws.2013.203
The main aim of this paper is to carefully analyse all raw data obtained through the telemetry system using a set of validation tests. Invalidated data are reconstructed with a set of models available for data validation.

PROPOSED METHODOLOGY

Commonly, data validation is mostly carried out manually by process experts, with the assistance of basic statistical data tools and graphic visualisation tools (Jorgensen et al. 1998). However, such procedure can only be applied for a limited amount of data (Mourad & Bertran-Krajewski 2002; Bertrand-Krajewski et al. 2005). The complete reliance on human judgments to cope with abnormal signals has become increasingly difficult due to a variety of malfunctions (Venkatasubramanian et al. 2003). Furthermore, in real time applications, it is difficult to manually execute data validation due to time constraints. In data collection platforms, automatic data validation techniques or software can increase the level of data reliability and decision support systems. For example, NIKLAS (Lempio et al. 2010) has been developed as a module for real-time and non-real-time evaluation of meteorological data. This module validates time series for the parameters of precipitation, global radiation, sunshine duration, air temperature, dew point temperature, relative humidity, wind speed, and air pressure. NDBC (2009) uses software that can automatically validate observations for measurement taken from a network of buoys, and coastal-marine automated network stations. Edthofer et al. (2010) developed a system for fault tolerant control in online drinking water quality monitoring, in which a data validation module is incorporated. In general, commercial supervisory control and data acquisition (SCADA) systems apply simple bounds and rates tests to single signals to detect outliers, and the reconstruction of these abnormal values are computed by a simple linear interpolation of raw pre- and post-validated data.

In Quevedo et al. (2009), a methodology to compute network efficiency taking into account raw flowmeter data and the network topology is presented. Basically, raw flowmeter data consistency is analysed using spatial network models (i.e. mass balance equations for each sector). Wrong or missing data are removed and replaced by estimated data using models, and filtered data are analysed to compute the performance of each sector. Estimated flowmeter uncertainty is taken into account in the network water balance evaluation obtaining confidence intervals for key performance indices following the approach of Taylor (2010). Finally, economic efficiency corresponding to each zone and to the overall network are obtained and analysed to improve instrumentation (i.e. sensor location, recalibration) and to define new plans for the network maintenance to locate leaks in pipes. Further, in Quevedo et al. (2009), 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. In this work, a methodology that combines both previous approaches is presented. This methodology consists of the following steps:

1. Flowmeter data validation tests.
2. Wrong data reconstruction based on model estimations.
3. Model generation based on filtered data.
4. Flowmeter data inaccuracy computation.
5. Sector and network efficiency computation.

Raw data validation and wrong data reconstruction (steps 1 and 2)

Raw flowmeter data validation is inspired in the Spanish AENOR-UNE norm 500540 (UNE 2004). The methodology consists of 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.

An explanation of each level is as follows.

- **Level 0**: The **communications** level simply monitors whether data are properly recorded, taking into account that the supervisory system is expected to collect data at a fixed sampling rate. Problems in sensors or in the communication system can be detected at this stage.
- **Level 1**: The **bounds** level checks whether data lie inside their physical range. For example, the maximum values expected for flowmeters can be determined by a simple analysis of the pipe flow capacity.
Level 2: The *trend* level monitors the data rate. For example, in a real tank, level sensor data cannot change more than several cm per minute.

Level 3: The *models* level uses three parallel models:
- **Valve/Pump**: the valve model supervises the possible correlation that exists between the flow measurement and the opening valve command in a given pipe or pump element.
- **Time series**: 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 can be used to compare and to validate recorded data.
- **Up-Downstream**: the up-downstream model checks correlation models between historical data corresponding to sensors located at different local stations in the same pipe (Quevedo et al. 2009). For example, sensor set reliability can be checked based on data corresponding to flowmeters located at different points of the same pipe in a transport water network.

A decision tree method has been developed to invalidate data at level 3, combining the results obtained with the three models. The *Up-Downstream* model is very useful not only to detect problems in sensor data but also to detect leaks in pipes and to compute the water balance in transport network sectors.

Once all test levels have been run, if data inconsistency is detected, the next step involves fault isolation by combining the previous tests. For instance, if an inconsistency involving a set of two flowmeters is detected, historical data and other features of both flowmeters are analysed to diagnose the cause of the problem and to identify the faulty sensor. Then, all data corresponding to this faulty sensor are replaced by the data corresponding to the healthy sensor on the same pipe. At this stage, the outputs of the models derived at level three are very useful to generate such reconstructed data.

### Network models and performances based on filtered data (steps 3, 4 and 5)

A water transport network can be divided into a set of interconnected sectors (see Figure 2). 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 study, pipes are considered pressurised. Hence, no delays are assumed 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 around measured values (‘normal’).
- Random errors between measurement instruments are uncorrelated (‘independence’).

Given a sector with several flowmeters at both input and output (see, for example, sectors 1, 3 and 4 in Figure 2), the model is

\[
\sum_{j=1}^{n_{\text{in}}} F_{\text{in}_j}(t) = K \sum_{l=1}^{n_{\text{out}}} F_{\text{out}_l}(t) + M
\]

where \(\sum_{j=1}^{n_{\text{in}}} F_{\text{in}_j}(t)\) and \(\sum_{l=1}^{n_{\text{out}}} F_{\text{out}_l}(t)\) 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.
Considering that input and output flowmeters have errors, named respectively $e_{in}$ and $e_{out}$, Equation (1) is rewritten as follows:

$$
\sum_{j=1}^{n} (F_{in}(j) + e_{in}(j)) = K \sum_{j=1}^{n} (F_{out}(j) + e_{out}(j)) + M
$$

Thus model residuals are given by

$$
e(t) = \sum_{j=1}^{n_{in}} F_{in}(t) - K \sum_{l=1}^{n_{out}} F_{out}(l) - M = \sum_{j=1}^{n_{out}} e_{out}(j)
$$

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

$$
\sigma = \frac{\sigma_{fit}}{\sqrt{n_{in} + n_{out}}}
$$

Given a confidence interval $\alpha$ with a standard deviation radius $\lambda(\alpha)$, the relative error is

$$
\text{Flowmeter error (\%)} = 100 \frac{\lambda(\alpha)\sigma}{\text{mean (flowmeter)}}
$$

Network efficiency is computed as the ratio between the network output flow $V_{out}$ and the network input flow $V_{in}$,

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

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

$$
[R_{\text{min}}(n), R_{\text{max}}(n)] = \left[ \frac{V_{out} - \lambda(\alpha)\sqrt{n_{out}} \cdot \sigma_{out}}{V_{in} + \lambda(\alpha)\sqrt{n_{in}} \cdot \sigma_{in}}, \frac{V_{out} + \lambda(\alpha)\sqrt{n_{out}} \cdot \sigma_{out}}{V_{in} - \lambda(\alpha)\sqrt{n_{in}} \cdot \sigma_{in}} \right]
$$

where $n$ is the number of days taken into account in the efficiency calculation horizon (e.g. $n = 365$ for a year).

This analysis is very useful to detect problems in sensors and leaks in network sectors. The combined use of

![Figure 2](image-url)
efficiency interval $[R_{\text{min}}, R_{\text{max}}]$, and flowmeter error (%) provide the following logic rules:

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector includes the accepted efficiency interval provided by the ATLL company: $R = [0.98, 1.02]$ (98–102%) and if the flowmeter error is coherent with the sensor imprecision provided by the manufacturer (1–2% for the electromagnetic flowmeter used in ATLL), then this sector and their sensors are in normal operation.

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector includes the accepted efficiency interval provided by the ATLL company and if the flowmeter error is higher (three times) than the sensor imprecision provided by the manufacturer (1–2% for the electromagnetic flowmeter used in ATLL), then this sector is in normal operation but a sensor problem has been detected and a maintenance operation could be required.

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector is always lower than the accepted efficiency interval provided by the ATLL company whereas the flowmeter error is coherent with the sensor imprecision provided by the manufacturer (1–2% for the electromagnetic flowmeter used in ATLL), then the sector could have a leakage problem while their sensors are in normal operation.

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector is always lower than the accepted interval efficiency provided by the ATLL company and if the flowmeter imprecision is higher (three times) than the sensor imprecision provided by the manufacturer (1–2% for the electromagnetic flowmeter used in ATLL), then the sector could have a leakage problem and/or a maintenance operation could be required by sensors.

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector is always bigger than the accepted interval efficiency provided by the ATLL company and if the flowmeter error is coherent with the sensor imprecision provided by the manufacturer (1–2% for the electromagnetic flowmeter used in ATLL), then the sector could have a sensor calibration problem but sensors are in normal operation.

- If the efficiency interval $[R_{\text{min}}, R_{\text{max}}]$ for a given sector is always bigger than the ideal value $R = 1$ (100%) and if the flowmeter error is higher (three times) than the sensor imprecision provided by the manufacturer (1–2% for the used electromagnetic flowmeter used in ATLL), then the sector could have a sensor calibration and operation problem, so a maintenance operation could be required by sensors.

**ATLL NETWORK RESULTS**

The methodology described in the previous section has been applied to ATLL network (Figure 3) from 2007 to determine
Figure 4 | Results corresponding to the sector 1 zone 1.
the annual economic efficiency using the installed flowmeters. Most flowmeters in the ATLL network are electromagnetic, and can be input flowmeters, billing flowmeters, output flowmeters and control flowmeters depending on where they are installed. Input flowmeters are installed at water transport network inputs which correspond to production plant outputs. There are 10 input flowmeters distributed in three production plants (Abrera, Cardedeu and ITAM Llobregat), and all of them are electromagnetic. Billing flowmeters are installed at all demand points, generally, in the input of tanks, or in the output when tanks are the property of ATLL. The number of billing flowmeters is 185. These are electromagnetic except for nine, which are mechanical flowmeters. Output flowmeters are used to measure water that is returned to production plants if it is needed. There are five output electromagnetic flowmeters, all of them in Abrera production plant. Finally, control flowmeters are distributed along the entire network. They are mostly electromagnetic flowmeters although some of them are ultrasonic flowmeters. They provide additional measurements to check volumes measured by other flowmeters and to validate mass balances in the network. All the data obtained for these flowmeters is collected in real time by the SCADA system.

The application of the proposed methodology has allowed analysing the network efficiency evolution and quantifying the effects of different actions (new instrumentation, maintenance plans, etc.) applied during these years in the overall network. This methodology has been firstly applied in a sector by sector basis in order to discriminate the real efficiency of all network components, and further, to each zone and the whole network. In order to show the results of this methodology, this section presents the analysis for the year 2010 in two sample sectors. The proposed methodology is compared with the conventional method used in commercial SCADA systems. In such systems the detection of faulty data is only based on the analysis of simple tests on single signals (level 0: communication faults, level 1: bounds limits and level 2: trend limits) and the use of a simple linear interpolation method to filter wrong values. The first sample sector is composed of one input flowmeter and three output flowmeters.

Figure 4(a) displays daily upstream and downstream raw flowmeter data. Figure 4(b) is a scatter plot relating input raw flowmeter data versus output raw flowmeter data and its linear approximation. The Pearson coefficient of the linear approximation provides a quite good indicator of the quality of the up-downstream model (Pearson coefficient equal to 1 corresponds to the ideal case).

Figure 4(c) displays daily upstream and downstream filtered flowmeter data using the conventional SCADA approach, considering the following limits: \( F(t_i) > 0 \) & \( \text{Abs}[F(t_i)-F(t_{i-1})] < 4,000 \text{ m}^3/\text{day} \). Figure 4(d) is a scatter plot relating filtered input flowmeter data versus filtered output flowmeter data and its linear approximation. The outliers have been replaced by linear interpolation.

Figures 4(e) and 4(f) show the daily upstream and downstream filtered flowmeter data following the proposed methodology. This includes at level 3 Holt Winters time series models (Taylor 2005) of input and output daily flows and up-downstream models (see Equation (2)). All data that have not passed a test are marked as invalid data (or outliers).

In all figures, outliers which have been replaced by estimated data obtained from other validated data are displayed and marked.

### Table 1 | Comparative results of the linear models obtained in the two sample sectors

<table>
<thead>
<tr>
<th></th>
<th>Raw data</th>
<th>Conventional SCADA filtered data (levels 0 to 2)</th>
<th>Proposed method filtered data (levels 0 to 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>( F_{in} = 0.978 ) ( F_{out} ) + 178.142</td>
<td>( F_{in} = 0.978 ) ( F_{out} ) + 156.788</td>
<td>( F_{in} = 1.007 ) ( F_{out} ) + 21.675</td>
</tr>
<tr>
<td>Pearson coefficient</td>
<td>0.970</td>
<td>0.945</td>
<td>0.999</td>
</tr>
<tr>
<td><strong>Example 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>( F_{in} = -0.904 ) ( F_{out} ) + 1071.889</td>
<td>( F_{in} = 0.413 ) ( F_{out} ) + 644.001</td>
<td>( F_{in} = 0.653 ) ( F_{out} ) + 382.820</td>
</tr>
<tr>
<td>Pearson coefficient</td>
<td>0.016</td>
<td>0.048</td>
<td>0.642</td>
</tr>
</tbody>
</table>
Figure 5 | Results corresponding to sector 2 zone 4.
Table 1 shows that the Pearson coefficient corresponding to the proposed methodology is closer to the ideal case than those corresponding to the results obtained directly using raw data or the conventional linear interpolation method used in SCADA systems. Moreover, the proposed methodology allows obtaining a flowmeter error (Equation (5)) and efficiency interval (Equation (7)) for this sector. In particular, the upstream flowmeter error is close to 0.5% whereas the downstream flowmeter error is close to 1.0%, and the efficiency interval is [98.7, 98.9%]. Thus, according to the rules presented in section network models and performances based on filtered data (steps 3, 4 and 5), this sector is working properly and the flowmeter error is within the imprecision of the ATLL flowmeters.

The second sample sector is only composed of one upstream flowmeter and one downstream flowmeter, but the quality of the time series corresponding to raw data is worse than in the first sample sector (see Figures 5(a) and 5(b)). In this case, the time series corresponding to the upstream flowmeter had an operating problem for almost half a year.

The validation method detects isolates and properly reconstructs wrong flowmeter data using downstream flowmeter data. Filtered data using the conventional linear interpolation approach in SCADA systems are shown in Figures 5(c) and 5(d), and filtered data using the proposed methodology are presented in Figures 5(e) and 5(f). Table 1 compares the results of different linear approximations. Note that the proposed methodology is the one that provides the best Pearson coefficient (0.642), although in this case is far from the ideal value because of the flowmeter problems. The coefficients of the linear approximation are $K = 0.655$ and $M = 582$ m$^3$. In this case, upstream and downstream flowmeter errors are close to 17% and the sector efficiency interval is [104.7, 108.5%]. Thus, following the rules described in section network models and performances based on filtered data (steps 3, 4 and 5), in this sector flowmeters could have a calibration problem and a maintenance action should be applied.

The same procedure has been applied to all ATLL water network sectors allowing ranking them from the best to the worst one, taking into account several performance indices: economic efficiency, sensor error, data quality (% of estimated data), etc. The 10 network zones and the overall network have also been addressed in order to obtain global performances of the ATLL network in this work.

**CONCLUSIONS**

In this work, a combined methodology to evaluate the economic efficiency of a water transport network, including the detailed analysis of all sectors and zones, is proposed. Before evaluating network efficiency, the methodology checks raw flowmeter data consistency using several tests and models, and replaces wrong data by model estimations. After the data validation/reconstruction process, the proposed methodology evaluates the efficiency of all sectors, zones and complete network taking into account sensor inaccuracies and providing a confidence interval. This confidence interval summarises network misbehaviours either due to leakage or sensor bad calibration. A tight confidence interval is indicative that the network is behaving properly. Otherwise, a wide confidence interval indicates the existence of some leaks or bad calibrated sensors.

**ACKNOWLEDGEMENTS**

This work belongs to a research applied project granted by ATLL Company. The authors wish also to thank the support received by the Spanish Ministry through SHERECS project (ref. DPI-2011–26243), and by iSense grant FP7-ICT-2009-6-270428 and by EFFINET grant FP7-ICT-2012-318556 of the European Commission.

**REFERENCES**


Jorgensen, H. K., Rosenorn, S., Madsen, H. & Mikkelsen, S. 1998 Quality control of rain data used for urban runoff systems. Water Science and Technology 37 (11), 113–120.


First received 28 June 2013; accepted in revised form 7 October 2013. Available online 12 November 2013