Validation and Reconstruction of Flow Meter Data in the Barcelona Water Distribution Network

J. Quevedo♦, V. Puig♦, G. Cembrano♦♦, J. Blanch♦, J. Aguilar♣, D. Saporta♥, G. Benito♠ , M. Hedo♠, A. Molina♠

♦*Automatic Control Department*

Technical University of Catalonia, Rambla Sant Nebridi, 10, 08222 Terrassa, Spain

phone : +34 9373986327 fax : +34 9373986328

e-mail: vicenc.puig@upc.edu

♦♦*Industrial Robotics Institute, CSIC, C/ Llorens i Artigas 4-6, 08028 Barcelona, Spain*

♣*LAAS-CNRS, 7 Avenue du Colonel Roche, 31077 Toulouse, France*

♥*AGBAR Barcelona Water Company, C/Diputació, 351, 08009 Barcelona, Spain*

♠ *Adasa Sistemas, C/Pedrosa B, 30, 32, 08908 L'Hospitalet de Llobregat, Spain*

Abstract: This paper presents a signal analysis methodology to validate (detect) and reconstruct the missing and false data of a large set of flow meters in the telecontrol system of a water distribution network. The proposed methodology is based on a two time-scale forecasting models: a daily model is based on a ARIMA time series, while the 10-minute model is based on distributing the daily flow using a 10-minute demand pattern. The demand patterns has been determined using two methods: correlation analysis and an unsupervised fuzzy logic classification, named LAMDA algorithm. Finally, the proposed methodology has been applied to the Barcelona water distribution network providing very good results.

Keywords: Water Distribution Network, Tele-control System, Flow meter, Fault Detection, Sensor Failure, Unsupervised Classifier, Fuzzy Logic Classifier

1. INTRODUCTION

In a complex water distribution network, such as the case of Barcelona city, a telecontrol system must acquire,

store and validate data from many flow meters and other sensors every few minutes to achieve an accurate monitoring of the whole network in real time. Frequent operation problems in the communication system between the set of the sensors and the data-logger, or in the telecontrol itself, generate missing data during a certain periods of time. The stored data are sometimes uncorrelated and of no use for the historic records. Missing data must, therefore, be replaced by a set of estimated data. A second common problem is the lack of reliability of the flow meters (offset, drift, breakdowns) producing false flow data readings. These false data must also be detected and replaced by estimated data, since flow data is used for several network water management tasks, namely: planning, investment plans, operations, maintenance and billing/consumer services and operational control (Cembrano, 2000). The same type of problem can be found in gas or electricity networks (Matheson, 2004). Therefore, the methodology presented in this paper could also be applied to these networks.

According to the nature of the available knowledge, different kinds of data validation can be built, 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 1997; Mourad and Bertrand-Krajewski, 2002).

Elementary signal based methods use simple heuristics and limited statistical information of a given sensor (Burnell, 2003) (Jorgensen et al., 1998) (Maul-Kötter et al., 1998). Typically, these methods are based on validating either signal values or signal variations data validation. 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 (Tsang, 2003). 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 fault in the system. Otherwise, it is considered that the system is working properly. Models are usually derived using either multivariate procedures exploiting the correlation or the analytical relations between several quantities obtained using first principles, sometimes measured at different times ("*temporal redundancy*") and/or locations ("*spatial redundancy*"). The result of data validation may be either a binary variable indicating whether the data is considered valid or not, or a continuous validity index interpreted as a degree of confidence in the data. When the degree of confidence is too low, data can be either discarded or replaced by an estimate computed using a statistical or physical model (see, e.g., Petit-Renaud and Denoeux, 1998). Moreover, a subproduct 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. Some examples of these methods in literature applied to the water domain are:

- *Time-series analysis techniques* (Prescott and Ulanicki, 2001; Lobanova and Lobanova, 2003; Bennis et al.,

1997; Bennis and Kang, 2000; Crobeddu and Bennis, 2006).

- *Kalman filters* (Piatyszek et al. 2000; Pastres et al. 2004; Ciavatta et al. 2004).
- *Parity equations* (Ragot and Maquin, 2006; Hamioud et al. 2005a, 2005b; Boukhris et al. 2001).
- *Pattern recognition methods* (Valentin and Denoeux 2001).
- and *Principal Component Analysis* (Nelson et al., 1996; Arteaga, 2002; Harkat et al. 2006)

However, neither of the existing methods satisfies data validation and reconstruction specifications established by the Barcelona water company. In particular, the company specifies that every flow meter at water distribution level should only be validated and reconstructed using their own data by exploiting the "*temporal redundancy*" that exists since flow readings follow the consumer demand patterns. No mass balances between different sensors could be used since they involve relating the information systems of the Barcelona transport and distribution networks. This is not currently possible because those systems are not integrated. The transport network is in charge of the right choice of water sources that enter the system at each moment (quality and quantity of the water supply) and of the bulk water transferences between sources and reservoirs, while the distribution network is responsible of delivering water from the reservoirs to the consumers.

To address these problem specifications, this paper proposes a model based methodology for data validation (detection) and replacement of faulty/missing flow meter measurements in a water distribution network with district metering areas (DMA), based on the use of time series analysis in combination with short-term consumption patterns. The number of consumption patterns have been studied using correlation analysis and the fuzzy classification algorithm LAMDA. The proposed data validation and replacement algorithm exploits temporal redundancy existing in the demands because of the existence of weekly periodic behaviors associated to human habits (Brdys and Ulanicki, 1994). This algorithm is inspired on previously developed algorithms to predict water flow demands (Quevedo, 1988).

The structure of the paper is the following: in *Section 2,* the problem of flow meter data validation and reconstruction in Barcelona water distribution network information system is presented and the need for an algorithm to do it automatically is justified. In *Section 3*, the proposed methodology is introduced and motivated. *Section 4* presents the two-level model used for such methodology that uses at the first level a time-series and flow patterns at the second level to describe the flow time series recorded at each flow meter. *Section 5* describes the proposed methodology of data validation and reconstruction. Results of application of the proposed methodology in several real scenarios coming from the Barcelona water distribution network are presented in *Section 6,*. Finally, in *Section 7*, the conclusions and further developments are presented.

2. PROBLEM DESCRIPTION

2.1 Barcelona water distribution network

The Barcelona water distribution network is comprised of some 200 sectors (DMA's) with approximately 400 control points (Fig. 1). At present, the Barcelona information system receives, in real time, data from 200 control points, mainly flow meters and a few pressure sensors. Most of these flow meter control points correspond to the single supply point of a DMA, so that their evolution is closely related to that of the water demand in that DMA.

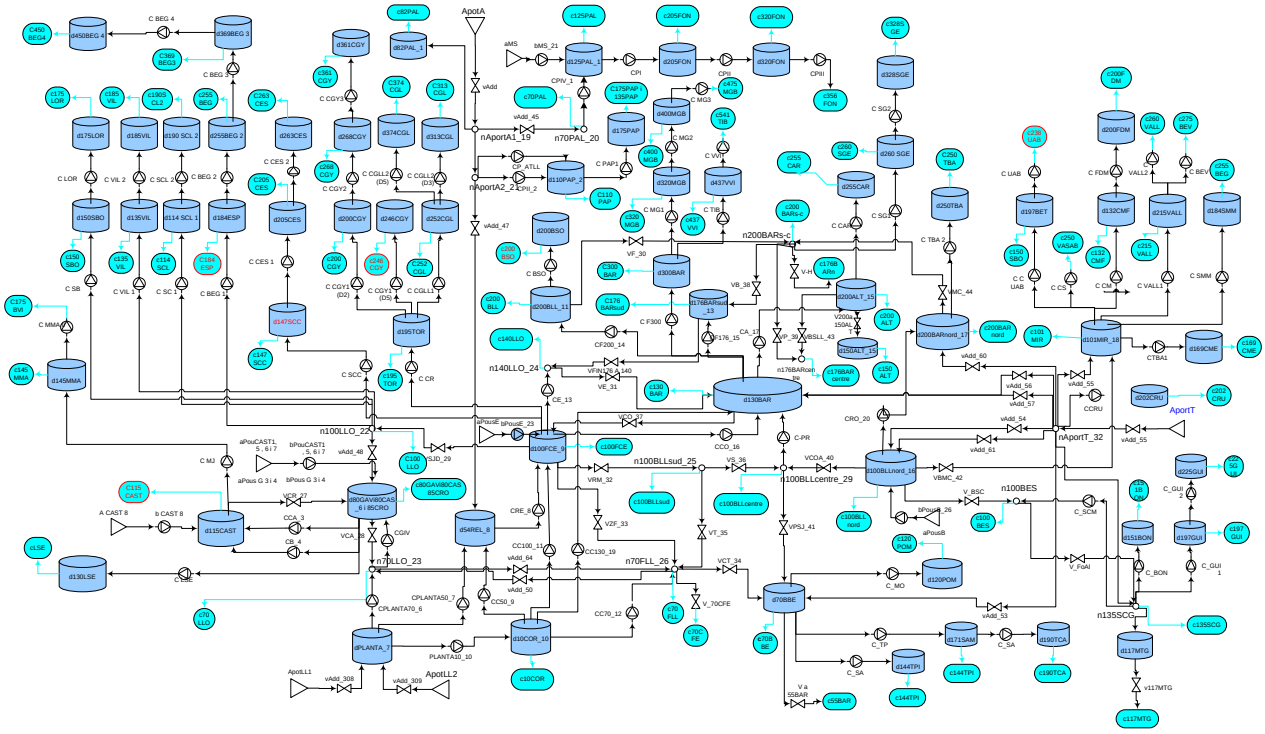
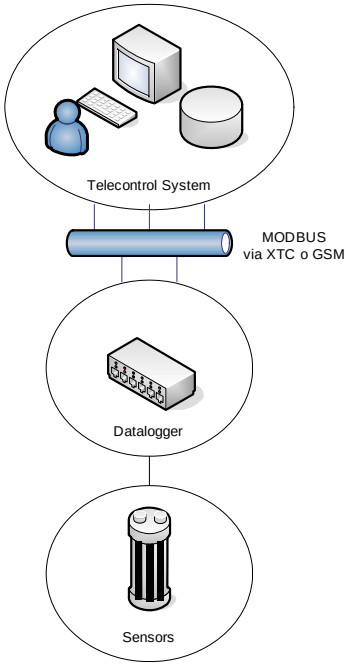


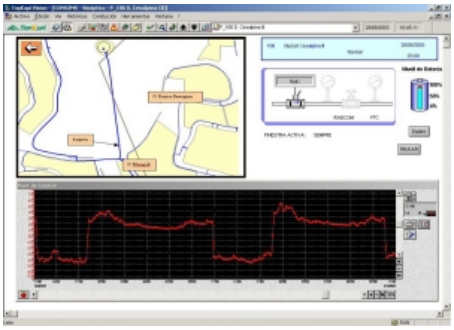
Fig. 1. Barcelona water network simplified scheme

Every 10 minutes, a data-logger stores the data from the sector sensors and, once a day, these data are sent to an Operational Data Base of the Telecontrol Information System implemented using TOPKAPI SCADA system (http://www.areal.fr/) via telephone XTC network or GSM radio using the ModBus communication protocol (Fig. 2). Communication with each of the dataloggers is ensured daily through 20 XTC lines managed simultaneously by the two servers. The data is recovered by using the TOPKAPI SCADA facilities and transferred to an ORACLE database.

The sensor data used by the telecontrol system must follow two functional procedures, previous to their integration for use in the Management System: raw data insertion process and data validation and replacement.

Telecontrol centre





Telecontrol information system implemented using TOPKAPI SCADA software

Fig. 2. Telecontrol of Barcelona water distribution system

2.2 Raw data insertion process

This process consists of data acquisition from the data logger towards the operational database in the TOPKAPI telecontrol system. In this process, missing sensor or data-logger data, as well as communication failures occur, which must be recovered using artificial data, for further use in statistic and hydraulic balance studies (Fig. 3). In this case, the failure detection is trivial: There is a gap in the data and an attached error message. A more interesting problem is the replacement of the missing data with a virtual value, which approximates the real value of the missed reading.

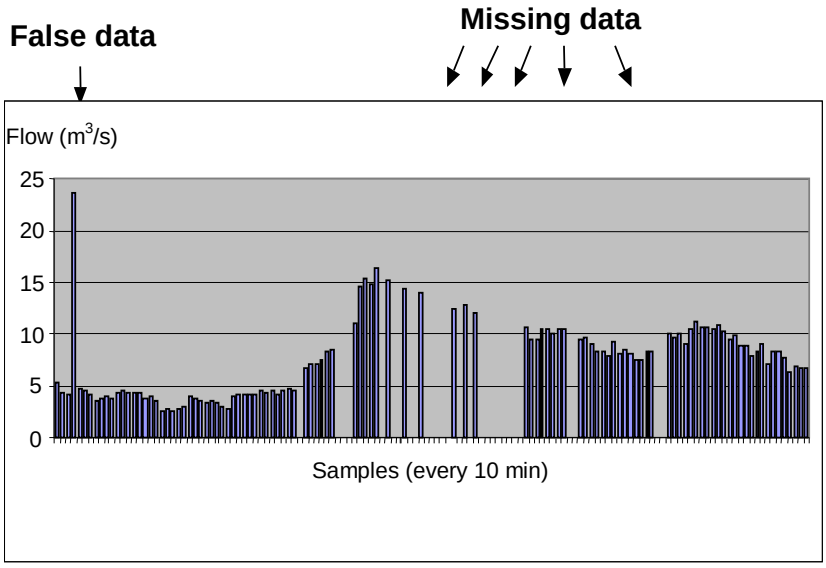


Fig. 3. Raw data with false and missing data

2.3 Data validation and reconstruction

At present, there is an automatic data validation procedure implemented at each control point. This validation checks each reading value against a validity range (daytime and nighttime minimum and maximum values). Whenever a reading value lies outside this range, the reading is invalidated. Additionally, the previous reading and the next one are also invalidated. However, these invalid data have been, so far, left intact and propagated for further use in hydraulic balances and statistics.

Invalid data should be replaced by virtual data (software or virtual sensor) before further computations are performed. A procedure for estimating missing or invalid data is necessary for the complete water network, being this the aim of this paper. This procedure, described in the next sections, can also be used to improve the actual implemented data validation method as it will also be described.

3. MOTIVATION AND OVERVIEW OF THE PROPOSED METHODOLOGY

3.1 Motivation

The proposed methodology for flow data validation and reconstruction is based on an analytical redundancy approach. A model is developed for each flow meters in the Barcelona water distribution distribution network based only on the actual and past readings of the same flow meter. This model will be able to exploit the "*temporal redundancy*" that exists in such readings since they follow the patterns of consumer demands, which, in most DMA's present clear a daily cycles (with changes in this pattern during the weekends and holydays).

It should be noticed that the methodology proposed can only be applied with reference to a particular type of measuring station location. Since, the water demand forecasting model is mainly based on the representation of the typical behaviour of water demands, the methodology proposed should be used only when the data can be considered really representative of water demands. This is the case, for example, of a flow meter located in the single feeding point of a DMA or a sector. In such a case the flow measured are generally quite representative of the demands of the area considered. On the other hand, the methodology could fail when the data analysed are relative to flows which can be highly influenced by the status of the control devices of the system and not only by the water demands. For example, considering a flow meter located on a generic pipe in a highly looped water distribution system, the flows in that pipe could increase or decrease not only according to the water demand pattern but also according to different valve settings.

The previous limitation could be overcome taking into account the current instrumentation in the network to build models that relate several sensors since in every sector only there is a sensor that measures the supplied flow. So, there are no additional sensors inside of the sector that can be used to establish balances between sensors (flow meters or pressure sensors) exploiting the redundancy between sensors. It is left as a second phase of the project presented in this paper to upgrade the existing instrumentation system including new sensors that allow to exploit such redundancy.

3.2 Overview

Once the model for each flow meter has been obtained, this model is be used for two purposes:

- for data invalidation through establishing a threshold for the difference between the real and predicted flow meter data reading. This threshold is determined using time series statistical methods. This invalidation methodology complements and enhances the heuristic existing validation methodology existing in the Barcelona water distribution network.
- for data reconstruction through the substitution of the real flow meter data reading for the estimated one using the model, once the data has been invalidated.

The general algorithm for flow meter data validation/reconstruction is presented in Figure 4.

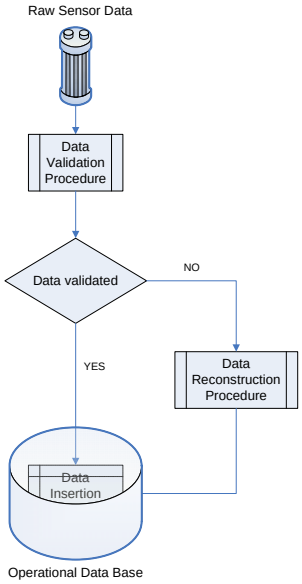


Fig. 4. Overview of the data validation/reconstruction procedure

4. DESCRIPTION OF THE FLOW TIME SERIES MODEL

4.1 Proposed time series modelling approach

A normal row flow meter series with readings every ten minutes in any of the control points of Barcelona water distribution network can be viewed as a time series, closely related to the DMA demand. Therefore, this study takes

advantage of demand forecasting techniques to model flow meter readings. In urban areas, such as the Barcelona network, the flow meter series presents: a daily periodicity (the demands presents a repetitive pattern every day that vary only in the week-ends and holydays) a weekly periodicity (the demand decreases during the week-ends) and seasonal changes (the demand changes according to the season because of weather, holydays, etc.).

The problem of modelling time series coming associated to the water demand has been addressed in the literature by many researchers. Depending on the timescales, there are methods that are more suitable for long, mid or short term prediction. In the particular application of data validation and reconstruction the model that is required should allow short term prediction (every 10 minutes). Some references in short-term water demand forecast are (Shvarser et al., 1993)(Zhou et al, 2002) (Maidment & Parzen, 1984a; 1984b)(Maidment et al., 1986)(Smith, 1988)(Miaou, 1990). Without exception, all these papers refer to the recurring patterns and periodicities that exist in water demand data at different levels of temporal aggregation. For this reason, a convenient way to model this type of time series, with clear daily and weekly seasonality, is to use a two-level model: a daily flow model, in combination with a daily 10-minute distribution pattern inspired in a previous work by the authors used for operational real-time control of water networks (Quevedo, 1988). A similar approach has recently been proposed in (Alvisi, 2007). The data aggregation in daily values provides more insight into the statistical properties of the series, which may be otherwise obscured by the variance of short-term readings. Additionally, the auto-correlation analysis in the 10-minute-reading series would require very long delays to cater for daily and weekly seasonality.

The proposed flow prediction procedure consists of two levels (Figure 5):

- a time-series modelling to represent the daily aggregated flow values and
- a set of daily flow distribution patterns that takes into account the variation during the week-ends and holidays periods. Every pattern consist of 144 10-minute values for each daily pattern.

The daily series of 10-minute values of the flow estimate is computed as a product of the daily aggregated flow value and the appropriate 10-minute distribution pattern.

4.2 Aggregate daily flow model

The aggregate daily flow model is built on the basis of a time series modeling approach using ARIMA modeling (Box and Jenkins, 1970). A time series analysis was carried out on several daily aggregate series, which consistently showed a weekly seasonality, as well as the presence of deterministic periodic components (Abraham and Box, 1975). A general expression for the aggregate daily flow model, to be used for a number of sensors in different locations was derived using three main components:

a) one-week-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 $s_{1-2}=\pm 2\pi/7\, j$ or equivalently, in z-plane: $z_{1-2}= e^{s_{1-2}} = \cos(2\pi/7) \pm \sin(2\pi/7)j$). Then, the oscillating polynomial is:

$$y(k) = 2\cos(2\pi/7)y(k-1) - y(k-2) \quad (1)$$

b) an integrator take into account possible trends and the non-zero mean value of the flow data:

$$y(k) = y(k-1) \quad (2)$$

c) an autoregressive component to consider the influence of previous flow values within a week. For the general case, the influence of 4 previous days is considered. However, after parameter estimation and significance analysis, the models are usually reduced to a smaller number of parameters.

$$y(k) = -a_1 y(k-1) - a_2 y(k-2) - a_3 y(k-3) - a_4 y(k-4) \quad (3)$$

Combining the three components (1)-(3) in the following way:

$$\Delta y_{\text{int}}(k) = y(k) - y(k-1)$$
$$\Delta y_{osc}(k) = \Delta y_{\text{int}}(k) - 2\cos(2\pi/7)\Delta y_{\text{int}}(k-1) + \Delta y_{\text{int}}(k-2)$$
$$y_p(k) = -a_1\Delta y_{osc}(k-1) - a_2\Delta y_{osc}(k-2) - a_3\Delta y_{osc}(k-3) - a_4\Delta y_{osc}(k-4)$$

the structure of aggregate daily flow model for each sensor is then:

$$y_p(k) = -b_1 y(k-1) - b_2 y(k-2) - b_3 y(k-3) - b_4 y(k-4) - b_5 y(k-5) - b_6 y(k-6) - b_7 y(k-7) \quad (4)$$

where:

$b_1 = a_1 - (2\cos(2\pi/7)+1)$
$b_2 = a2 - (2\cos(2\pi/7)+1)a1 + 2\cos(2\pi/7)+1$
$b_3 = a_3 - (2\cos(2\pi/7)+1)a_2 + (2\cos(2\pi/7)+1)a_1 - 1$
$b_4 = a_4 - (2\cos(2\pi/7)+1)a_3 + (2\cos(2\pi/7)+1)a_2 - a_1$
$b_5 = -(2\cos(2\pi/7)+1)a_4 + (2\cos(2\pi/7)+1)a_3 - a_2$
$b_6 = (2\cos(2\pi/7)+1)a_4 - a_3$
$b_7 = -a_4$

The parameters of this model should be adjusted using parameter estimation methods (as for example, the least-square methods) and historical data free of faults.

4.3 10-minutes flow model

The 10-minute flow model is based on distributing every 10-minutes the daily flow prediction provided by the time-series model (4) using a 10 min-flow pattern that takes into account the daily/month variation in the following way

$$y_{p10}(k+i) = \frac{y_{pat}(k,i)}{\sum_{j=1}^{144} y_{pat}(k,j)} y_p(k) \quad i = 1,\cdots,144 \quad (5)$$

where: $y_p(k)$ is the predicted flow for the day t using model (4) and $y_{pat}(k,i)$ is the prediction provided by the 10 min-flow pattern considering the flow pattern class day/month of the actual day k.

In order to determine the number 10-minute flow patterns to be considered, two different approaches have been used for comparison as will be discussed below. One uses a correlation study between different groupings of days, such as,

workdays and weekends in a month or a trimester. The other is an unsupervised classifier based of fuzzy logic called LAMDA (Aguilar-Martin, 1982).

Once, the number of patterns have been established, their composition is given. Every pattern consist of 144 10-minute values that will be stored in the Operational Data Base of the Telecontrol Information System. Each pattern 10 minute value is determined by computing the mean 10-minute values of the days that has been associated to this flow pattern. The algorithm that computes the 10-minutes flow prediction using (5) will use the adequate pattern taking into account the day of the week and the month of year.

4.3.1 10-Minute flow pattern determination using correlation analysis

For each sensor, the 10-minute daily records were first aggregated into averages for each weekday (Monday through Sunday) and for each month (84 patterns) for a whole year. Data for several years is also analysed to check if patterns change from year to year. In order to obtain a more reduced, but representative number of daily pattern classes, the correlation between the pattern curves was analyzed for different groupings (working days, Saturdays, Sundays) for each month and semester. The correlation factor r is computed as follows for two signals x_i and x_j with means given by $\overline{x}_i = E(x_i)$ and $\overline{x}_j = E(x_j)$, respectively:

$$r(x_i, x_j) = \frac{cov(x_i, x_j)}{\sqrt{cov(x_i, x_i)\, cov(x_j, x_j)}} \tag{6}$$

$$cov(x_i, x_j) = E\left[(x_i - \overline{x}_i)(x_j - \overline{x}_j)\right] \tag{7}$$

and it is used as a measure of similarity between two signals or curves. The correlation study showed that, for the majority of the sensors, workdays (Monday through Friday) could consistently be grouped in one pattern, while Saturdays and Sundays required one or two different daily flow distribution patterns, depending on the residential or industrial use of water in the area (See *Section 6* for details). Moreover, the effect of the seasonality can be easily handled considering one pattern per month for each class. In some sensors, depending on the water use in the sector, it is even possible to have a pattern for a whole trimester.

4.3.2 10-Minute flow pattern determination using fuzzy classification

LAMDA is a classification method based on the evaluation of the *adequacy* ("degree of membership") of an element to each class that has been developed by (Aguilar-Martin, 1982). LAMDA allows non-supervised and supervised

learning. In case of non-supervised learning classes are not known a priori, while in the supervised case need to be known. LAMDA will be applied to the 10-minute flow meter data to automatically discover the different classes of patterns that there exist by using the non-supervised classification mode in LAMDA.

An essential difference with other clustering methods (e.g. Linear Discriminate, Fuzzy C-Means, GK-Means), is that the LAMDA classification analysis is not based on minimization criteria (e.g. minimum distance between points, minimum square error) but on the evaluation of the contribution of each component x_j to its adequacy to a given class (*Marginal Adequacy Degree*, *MAD*). The global adequacy (*Global Adequacy Degree*, *GAD*) of an element to each class, is obtained by means of a fuzzy aggregation function applied to of the *MAD*s (Piera and Aguilar-Martin, 2002). The actual implementation of this algorithm used in this paper is the one included in the tool SALSA developed by (Kempowsky, 2003; 2006).

In the LAMDA classification method, each element to be classified is represented by a vector *X* with *d* components named *descriptors*: $\{x_1,\cdots,x_d\}$. The vector *X* can be seen as a point $X \in \Theta \subset \mathbb{R}^d$ where Θ is the space of possible values for each descriptor. If the element is quantitative, the descriptor values must be normalized to fit the unit interval taking into account the maximum and minimum descriptor values:

$$x_j^{norm} = \frac{x_j - x_j^{\min}}{x_j^{\max} - x_j^{\min}} \tag{8}$$

When applying LAMDA to classify the water demands patterns, every day (element to be classified) is characterized with a vector *X* of *d* =144 descriptors corresponding each one to the 10-min flow.

Each element is assigned to the class with maximum *GAD*, once the *MAD* for each class has been determined. Elements with very small adequacies are assigned to a Non Informative Class (*NIC*). The *MAD* for an element *X* to the Class C_k considering the j^{th} descriptor x_j, that is, the *j* direction of the space $\Theta \subset \mathbb{R}^d$, is given by the *MAD* function. LAMDA can work with several *MAD* functions (Binomial, Gaussian), each one characterized by its parameters. In the water demand pattern classification application, the Fuzzy Binomial function has been chosen with only one parameter ρ_{kj} (Aguilar-Martin, 1982). Then, the *MAD* for an element *X* to the Class C_k considering the j^{th} descriptor x_j is given by

$$MAD(x_j \,|\, \rho_{kj}) = \rho_{kj}^{\,x_j} \left(1-\rho_{kj}\right)^{(1-x_j)} \tag{9}$$

In the self-learning procedure, the parameters of the *MAD* function are estimated using the arithmetic mean recursive equation given by (Kempowsky, 2003)(Kempowsky, 2006):

$$\hat{\rho}_{kj} = \rho_{kj} + \frac{x_j - \rho_{kj}}{N+1} \tag{10}$$

where N is the total number of elements in class C_k, and x_j is the normalized value of j^{th} descriptor of the last element X assigned to the class C_k.

The *GAD* of an element X to the class C_k is computed through an aggregation function based on combining the marginal adequacies (*MAD*s) through a lineal interpolation between *t-norm* and *t-conorm fuzzy operators* (Piera and Aguilar-Martin, 1991). In the water demand pattern classification problem, the *t-norm* and *t-conorm* respectively used are *min* and *max*, what leads to computed the *GAD* as

$$GAD\left(x|C_k\right) = \alpha\, max\left(MAD\left(x_1|C_k\right),...,MAD\left(x_d|C_k\right)\right) + (1-\alpha)\, min\left(MAD\left(x_1|C_k\right),...,MAD\left(x_d|C_k\right)\right) \tag{11}$$

where the parameter $\alpha \in [0,1]$ is called *exigency index*. If α=0 the classification of an element in a class is not strict while in the case α =1 very strict.

The results obtained fuzzy classification study using LAMDA to the discovery of water demands patterns confirmed the results obtained with the correlation analysis study (See *Section 6* for details).

4.4 Model validation and accuracy

The two-level models (daily and 10-min) have being validated using data that has not been using for calibrating model parameters. Model accuracy was measured by the explained variance (EV),

$$EV = 1 - \frac{\sum_{i=1}^{n}(e_i - \mu_e)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2} \tag{12}$$

the root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{13}$$

and the mean absolute percentage error (MAE%)

$$MAE\% = 100\frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{\mu_y}\right| \tag{14}$$

where n is the number of observed data, y and $\hat{y}$ the measured and predicted values, $e = y - \hat{y}$ the errors, μ_e the mean error and μ_y the mean of the measured values.

5. DESCRIPTION OF THE PROPOSED DATA VALIDATION/RECONSTRUCTION METHODOLOGY

The proposed data validation and reconstruction procedure works as follows: The daily time-series models represent the dynamic behavior of the daily flow (or demand) aggregations based on historic records. These models allows to validate the aggregate daily flow obtained from raw 10-minute flow data. The daily flow corresponding to day *k*, *y(k)* is considered validated when

$$y(k) \in \left[\hat{y}(k) - c_\alpha \sigma_{\hat{y}}, \hat{y}(k) + c_\alpha \sigma_{\hat{y}} \right] \tag{15}$$

where $\hat{y}(k)$ is the prediction provided by the daily flow model, $\sigma_{\hat{y}}$ is the standard deviation of this prediction and c_α is a coefficient that depends on the degree α of confidence of the interval. $\sigma_{\hat{y}}$ and c_α. The validation of the daily corresponding to day *k* automatically validates the 144 10-minutes samples associated to this day. In case that the daily data is invalidated then, the 10-minute flow prediction model (5) described in *Section 4.3* is used to reconstruct the 10-minutes data corresponding to this day. Figure 5 summarizes the proposed data/validation and reconstruction methodology.

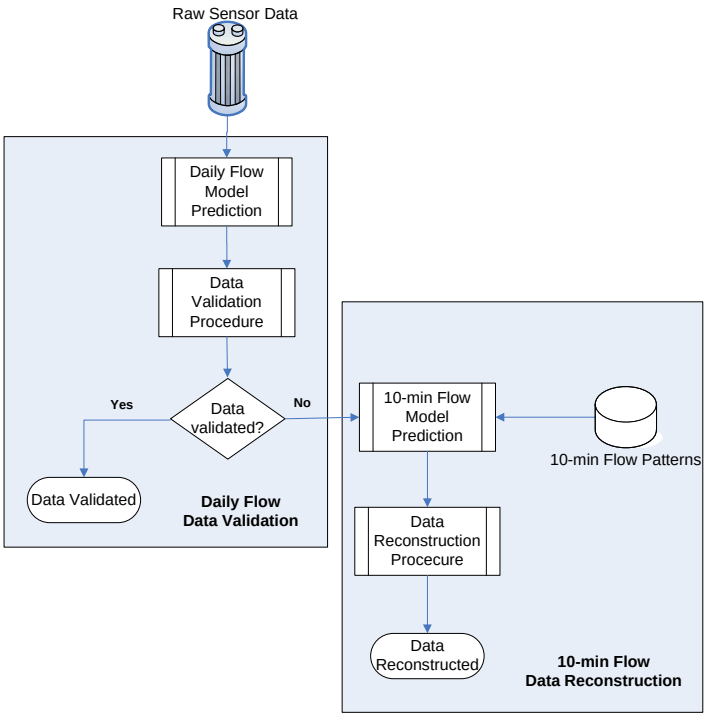


Fig. 5. Flow data validation and reconstruction procedure

6. RESULTS

The proposed methodology for validating and reconstructing missing and false data has been applied to the whole set of flow meters of the Barcelona network corresponding to single feeding points of 200 DMA's. Here, some representative examples of the results in some DMA's are presented.

6.1 Results of the daily flow model

The results presented in this section are based on the 10-minute daily records of one year (from June 1, 2003 to May 31, 2004). These records include several occurrences of missing and incorrect data that have been removed to calibrate the models and patterns. 10-minute data were processed to obtain hourly and daily flow values.

The results of applying a aggregate daily flow model, described in *Section 4,* are shown below. Using a selected range of daily flow data that does not contain faulty data, the parameters of a time series model in Eq. (4) were identified using the least-squares method. Figure 6 shows the real flow values and those predicted one day ahead by this model at sensor "*Avinguda Sarrià"*. Figure 7 and 8 show, respectively, how the quality prediction indicators presented in *Section 4.4* and confidence intervals (15) vary when the prediction horizon increases. The behaviour of the estimation is considered satisfactory, since the prediction errors (the difference between the real data and the estimated value) with a year mean absolute percentage error (MAE%) of less than 5%. Similar results were obtained for the rest of sensors of the Barcelona water distribution network.

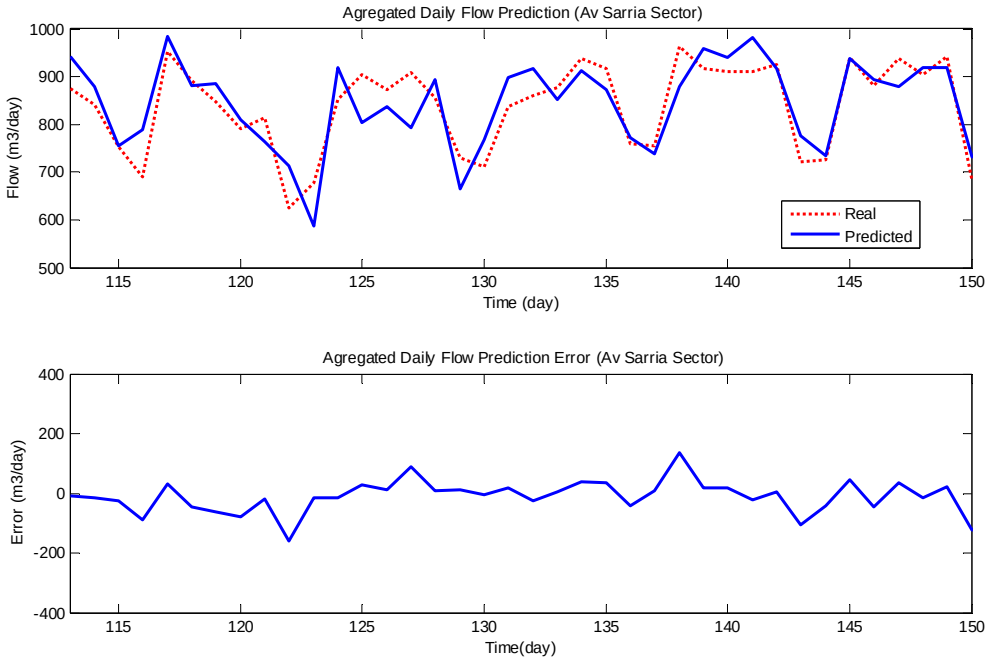


Fig. 6. Daily real/predicted data and prediction error corresponding to *Avinguda Sarrià"* sector

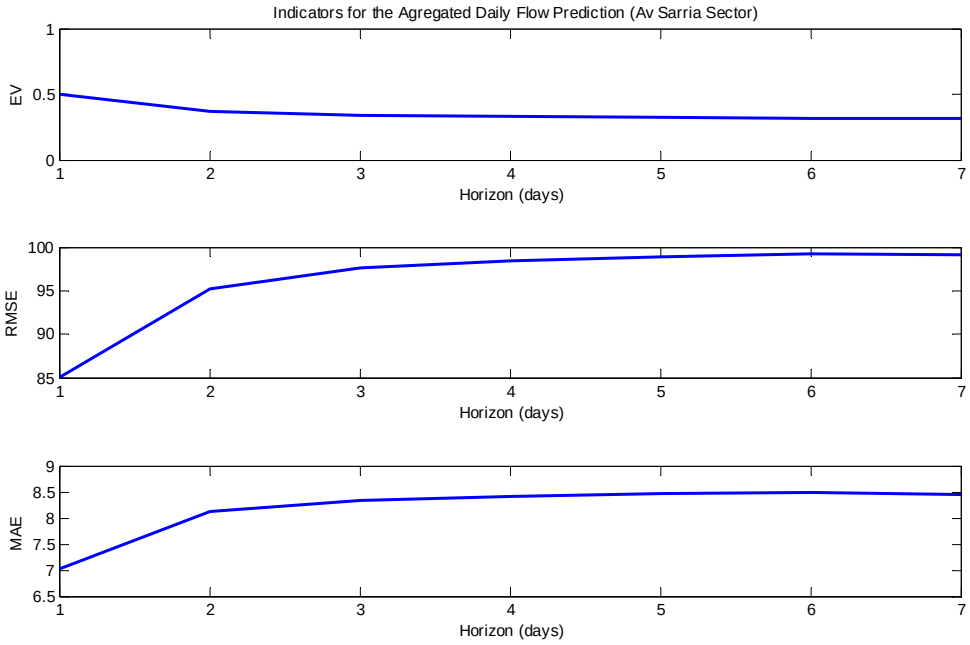


Fig. 7. Evolution of the daily flow prediction quality changing the horizon corresponding to "*Avinguda Sarrià*" sector

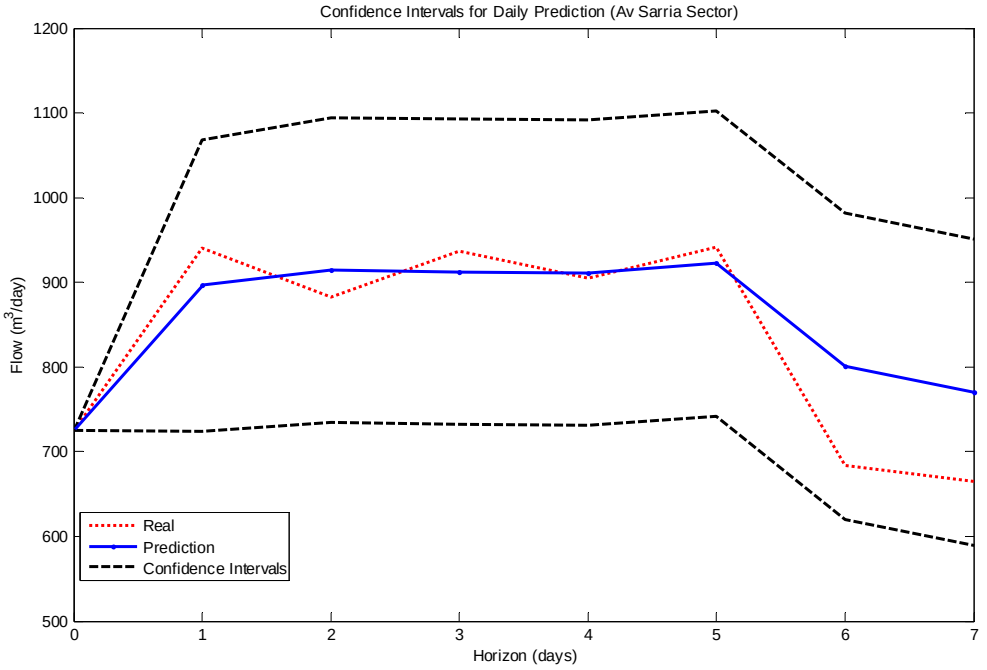


Fig. 8. Evolution of the daily flow prediction confidence intervals changing the horizon corresponding to "*Avinguda Sarrià*" sector

6.2 Results of 10-minute flow pattern study

6.2.1 Results based on the correlation analysis

Table 1 shows some results of the correlation study of the "*El Papiol"* flow meter corresponding to the month of July. In particular, it shows high correlation factors for different pairs of weekdays, as well as for each of these weekdays with a monthly workday average. Lower correlation factors of each of these with Saturday/Sunday patterns are also apparent. On the hand Saturdays and Sundays present also a high correlation compared to the workdays. This

is an example of results indicating that two different behaviors exist and daily distribution patterns may be classified into two patterns: workdays and Saturdays/Sundays.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Workday Moth	Workday Trim	Workday Year
Mon	1,00	0,91	0,92	0,96	0,94	0,65	0,44	0,97	0,97	0,96
Tue	0,91	1,00	0,92	0,93	0,93	0,68	0,50	0,96	0,95	0,94
Wed	0,92	0,92	1,00	0,93	0,94	0,64	0,45	0,97	0,96	0,96
Thu	0,96	0,93	0,93	1,00	0,95	0,70	0,51	0,98	0,97	0,96
Fri	0,94	0,93	0,94	0,95	1,00	0,66	0,49	0,98	0,97	0,97
Sat	0,65	0,68	0,64	0,70	0,66	1,00	0,71	0,68	0,69	0,68
Sun	0,44	0,50	0,45	0,51	0,49	0,71	1,00	0,49	0,46	0,45
Workday Month	0,97	0,96	0,97	0,98	0,98	0,68	0,49	1,00	0,99	0,99
Workday Trim	0,97	0,95	0,96	0,97	0,97	0,69	0,46	0,99	1,00	0,99
Workday Year	0,96	0,94	0,96	0,96	0,97	0,68	0,45	0,99	0,99	1,00

Table 1. Correlation factors between each pair of daily average flow patterns in one month at the "*El Papiol*" flow meter

Figure 9 shows the daily flow distribution patterns for workdays corresponding to July 2003 (red lines) and the aggregated (average) pattern.

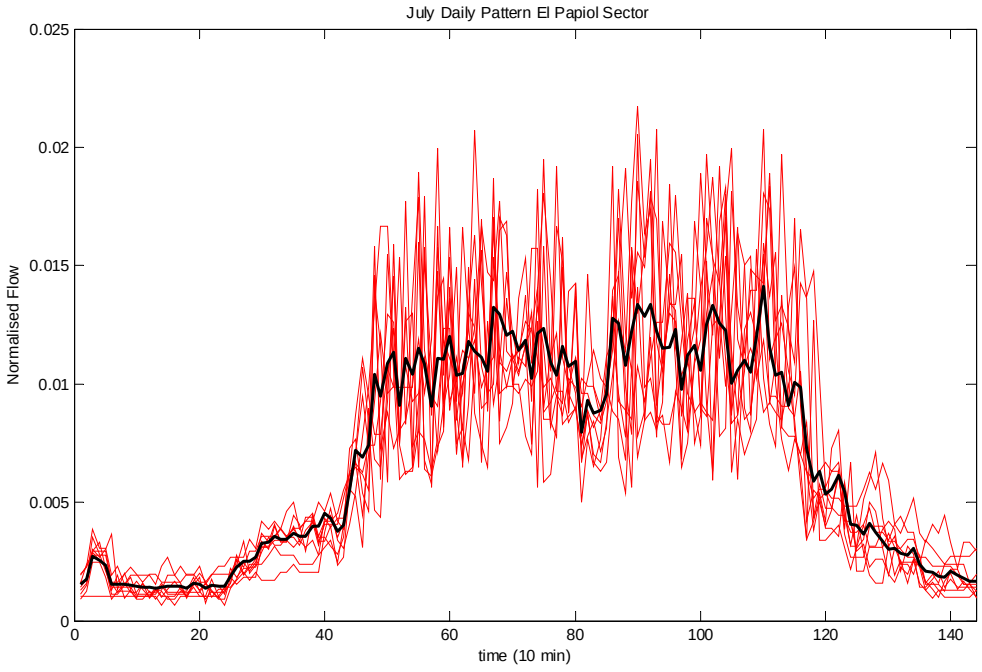


Fig. 9. Daily flow pattern of "*El Papiol*" sector corresponding to the month of July

6.2.2 Results based on the LAMDA classification method

Here some of the results that have been obtained with the LAMDA method using the software SALSA (Kempowski, 2003; 2004b) applied to one year data set of "*El Papiol*" sensor are presented. Two types of days were identified: weekends and workdays, represented respectively by classes 1 and 2.

The classification parameters were the Fuzzy Binomial function for *MAD*, and the *min-max* aggregation operators for *GAD* with exigency index of 0.47.

Since self-learning classification is used, days with some missing measurement values were discarded before hand, assuming that they correspond to sensor malfunction; therefore only 316 data, over 365 in the year, have been analyzed. Figure 10 and 11 shows, respectively, the flow pattern for the two identified classes: class 1 (weekends) and class 2 (workdays). The *x*-axis corresponds to 144 10-minute samples that are used as a descriptors for the LAMDA classification algorithm while in the *y-axis* corresponds the mean value of 10-minute flow of the days included in the class.

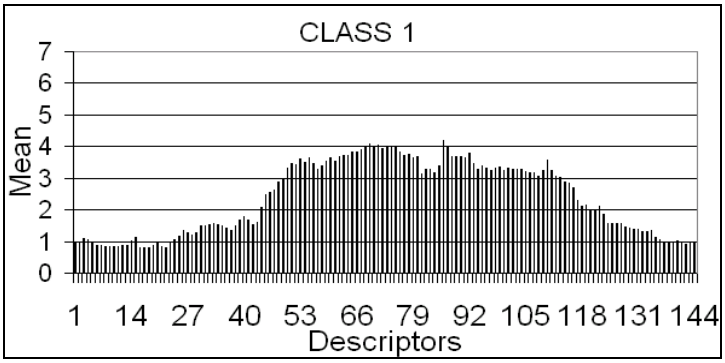


Figure 10. Flow pattern corresponding to Class 1 (weekends)

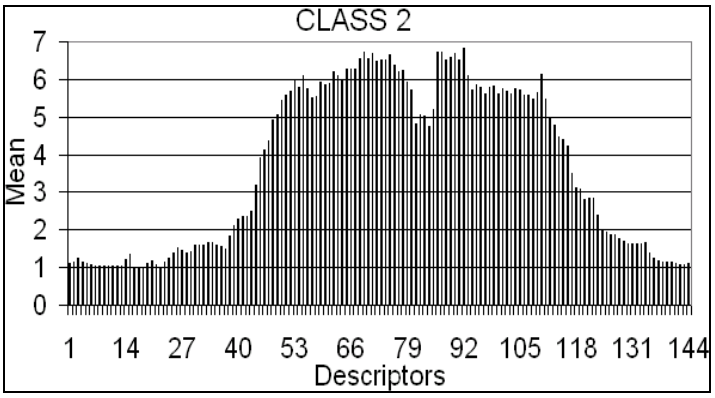


Figure 11. Flow pattern corresponding to Class 2 (workdays)

The 96 elements assigned to class 1 (weekend days), among them 7 are holidays and 7 are misclassified working days. In class 2, there are 224 days, where 11 Saturdays and 1 Sunday are misclassified elements. The results of the LAMDA classification method for other sensors confirm that the workday pattern is different from Saturdays and Sundays. In some sensors, depending on the water use (domestic or industrial) in the sector, the method discriminates three classes: workdays, Saturdays and Sundays, just like the method based on correlation does.

6.3 Results of 10-minute flow model

The 10-minute flow model prediction is determined using the procedure described in *Section 4.3* that is based on distributing every 10-minutes the aggregate daily flow prediction provided by the time-series model (described *Section 4.2*) using a 10 min-flow pattern determined either using correlation analysis or by the LAMDA classification method. Figure 12 presents the result of the 10-minute flow model prediction compared against the raw data coming from the data logger in the case of the "*Avinguda Sarrià*" sector. Figure 13 shows how the quality prediction indicators presented in *Section 4.4* vary when the prediction horizon increases. This prediction method presents a year mean predicted error of less than 5% . Similar results were obtained for the rest of the Barcelona network sensors.

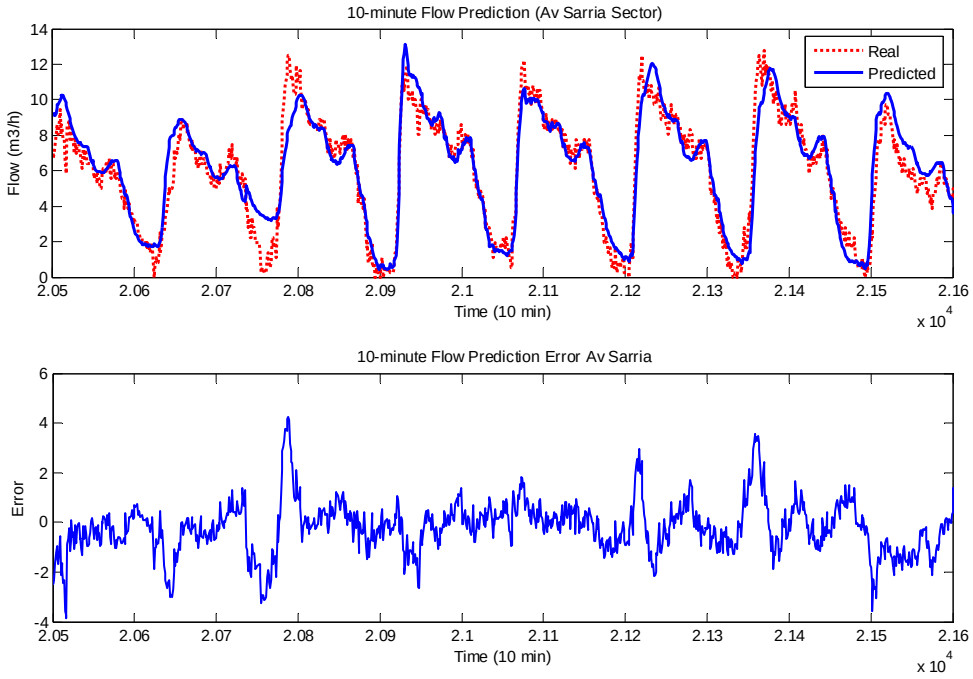


Fig. 12. "10-minute real/predicted data and prediction error corresponding to *Avinguda Sarrià"* sector

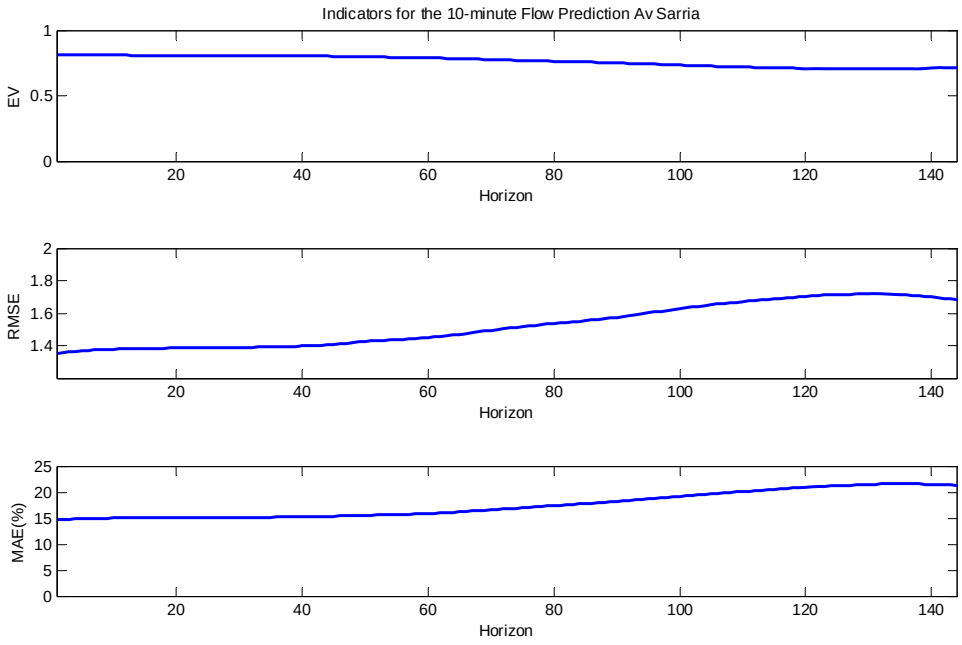


Fig. 13. Evolution of the 10-minute flow prediction quality changing the horizon corresponding to "*Avinguda Sarrià*" sector

6.4 Results of data validation/reconstruction

This section presents results of the data validation and reconstruction of data registered at the sensor "*Avinguda Sarrià*" . Figure 14 shows the daily flow corresponding to this sensor where in the period from 12/07/2004 to 14/07/2004 missing and faulty data were present in the raw 10-minute data (see Figure 15). The daily flow corresponding to these days is invalidated using the confidence intervals corresponding the aggregate daily time-series model as it can be seen in Figure 14. Figure 15 shows how the missing and faulty data corresponding to these days is replaced by the prediction provided by the 10-minute prediction.

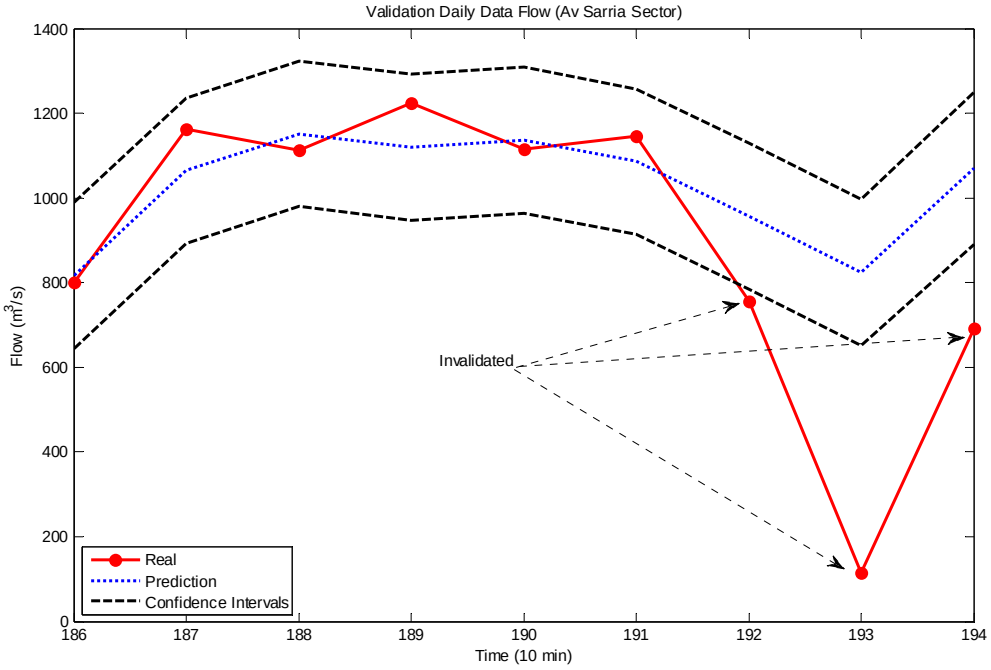


Fig. 14. Results of daily flow data validation

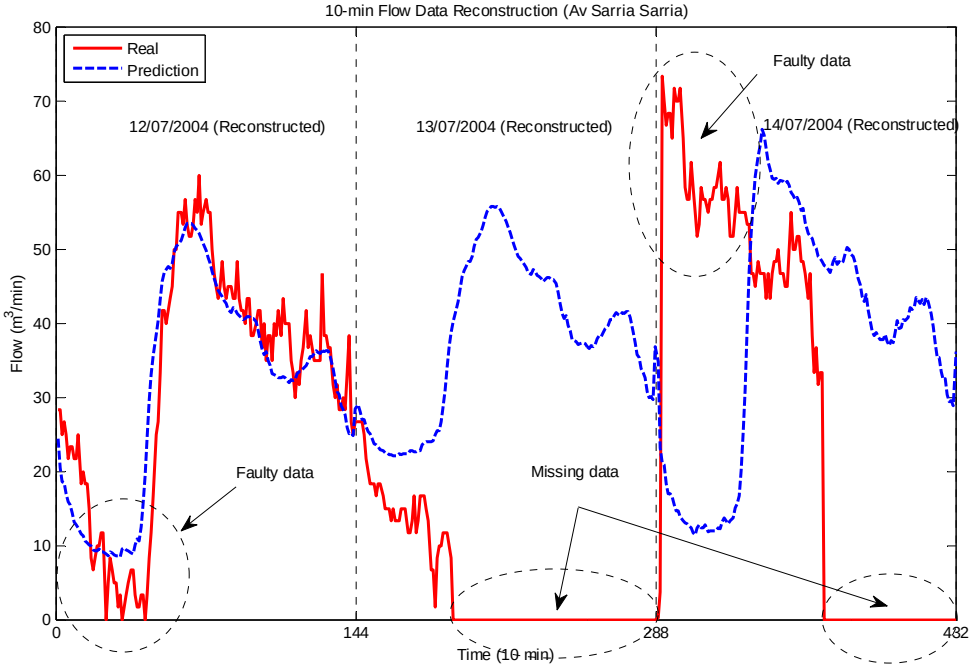


Fig. 15. Results of 10-minute flow data reconstruction

6.5 Software implementation

The proposed methodology has been implemented in Visual Basic (VB.NET) and PL/SQL DEVELOPER using an ORACLE Data Base and integrated to the Telecontrol Information System including new tools for Barcelona water distribution network management. The main functionalities are the pattern generation of the daily flow meters behaviour, the daily aggregated estimation based on time series, the reposition of invalidated or missing data and the monitoring of the real and estimated data. Figure 16 presents one the screens of the software implementation where

the 10-min flow model prediction compared against the raw data coming from the data logger in the case of the "*El Raval*" sector.

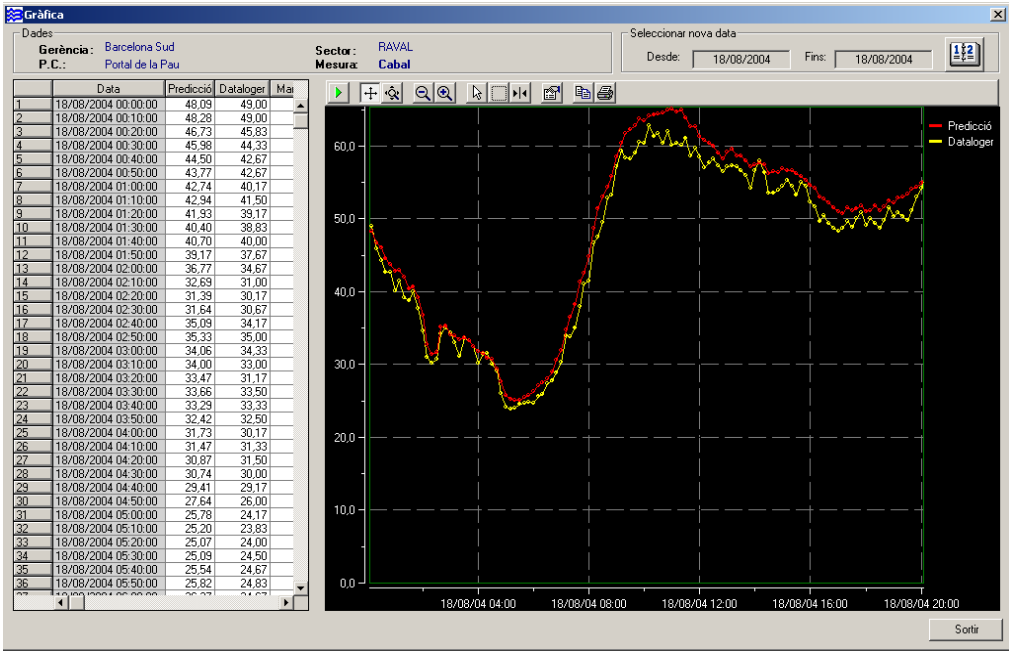


Fig. 16. Screen of the validation/reconstruction tool implementation

7. CONCLUSIONS

A two-phase methodology has been developed to estimate the replacement values for invalidated and missing data of the flow meters in a water distribution network with district metering. The project refers to flow meters at single feeding points of DMA's, so that the flow meter series is a measure of the demand in each metered area. An extension of this project will deal with DMA's containing a tank and a valve, so that the dynamic evolution of the reservoir and the valve control actions must also be taken into account when modelling the flow meter series. A time-series model based on daily aggregated flows provides the estimation one day ahead (or several days ahead) and a set of daily pattern for each 10 minute-interval make it possible to replace the values of invalidated or missing data. The daily pattern classification has been obtained by statistical and qualitative (fuzzy logic) methods with similar results. The satisfactory results with the historic data of flow meters in Barcelona have stimulated the authors to complete the implementation of this methodology as a new tool to improve the management of Barcelona water distribution network. Flow models in water network sectors are also expected to contribute to water leak detection in future developments.

ACKNOWLEDGMENTS

This works belongs to a research applied project granted by ADASA and AGBAR companies. The authors wish also to thank the support received by the Research Comission of the Generalitat of Catalunya (Grup SAC ref. 2005SGR00537) and by CICYT (ref. DPI-2005-05415) of Spanish Ministry of Education.

REFERENCES

Alvisi, S., Franchini, M., Marinelli, A. “A short-term, pattern-based model for water-demand forecasting”. Journal of Hydroinformatics, 9 (1), 2007.

Arteaga F, Ferrer A. (2002), “Dealing with missing data in MSCP: several methods, different interpretations, some examples”, Journal of Chemometrics, vol. 16, pp 408-418.

Abraham B., Box G. E. P. (1975) “Linear Models, Time Series and Outliers”. University of Wisconsin-Madison, Dept of Statistics. Tech. Report No. 438

Aguilar-Martin J., Lopez de Mantaras R. (1982), “The process of classification and learning the meaning of linguistic descriptors of concepts”, Approximate Reasoning in Decision Analysis, North Holland Publishing Company, pp 165-175.

Box G.E.P., Jenkins G. M. (1970) “Time Series Analysis Forecasting and Control” Holden-Day

Bennis S., Berrada F., Kang N. (1997). “Improving single variable and multivariable techniques for estimating missing hydrological data”. Journal of Hydrology, 191(1-4), pp. 87-105.

Bennis, S., Kang, N. (2000). A new methodology for validating historical hydrometric data with redundant measurements. In W. R. Blain and C. A. Brebbia (Eds), Hydraulic Engineering Software VIII, WIT Press, 2000.

Boukhris A., Giuliani S., Mourot G. (2001). “Rainfall-runoff multi-modelling for sensor fault diagnosis”. Control Engineering Practice, Vol. 9 (6), June 2001, pp. 659-671.

Brdys, M., Ulanicki, B. “Operational Control of Water Systems: Structures, Algorithms and Applications”. Prentice Hall International, 1994.

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

Cembrano, G., Wells G., Quevedo J., Pérez R., Argelaguet R. (2000). “Optimal Control of a Water Distribution Network in a Supervisory Control System”. *Control Engineering Practice,* 8(10), 1177-1188. Elsevier. Great Britain.

Ciavatta, S., Pastres, R., Lin, Z., Beck, M.B., Badetti, C., Ferrari, G. (2004). “Fault detection in a real-time monitoring network for water quality in the lagoon of Venice (Italy)”. Water Science and Technology, Vol. 50, No 11, pages 51-58, 2004.

Crobeddu. E., Bennis, S. (2006). “Data acquisition, validation and forecasting for a combined sewer network”. In V. Popov, A.G. Kungolos, C.A. Brebbia and H. Itoh (Eds), Waste Management and the Environment III, WIT Press.

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, Compiègne, France, November 1997.

Franklin, S.L., Maidment, D.R. (1986). "An evaluation of weekly and monthly time series forecast of municipal water". Water Resources Bulletin, 22(4), 611-621.

Hamioud, F., Joannis, C., Ragot, J. (2005a). Fault diagnosis for validation of hydrometric data collected from sewer networks. 10th International Conference on Urban Drainage, 10ICUD, Copenhagen, Denmark, August 21-26, 2005.

Hamioud, F., Joannis, C., Ragot, J. (2005b). Localisation de défauts de capteur pour la validation des mesures hydrométriques issues de réseaux d'assainissement. 20ème colloque sur le traitement du signal et de l'image GRETSI 2005 -- Louvain la Neuve Belgique, 6-9 septembre 2005.

Harkat, M.F., Mourot, G., Ragot, J. "An improved PCA scheme for sensor FDI: Application to an air quality monitoring network". Journal of Process Control, Vol. 16, Issue 6, July 2006, Pages 625-634.

Jörgensen H.K, Rosenörn S., Madsen H., Mikkelsen P.S. (1998) "Quality control of rain data used for urban run-off systems". Water Science and Technology, 37(11), pp 113-120.

Kempowsky, T., Aguilar-Martin, J., Subias, A., Le Lann, M.V. (2003). "Classification tool based on interactivity between expertise and self-learning techniques. IFAC Safeprocess, 2003, Washington DC, USA, 2003

Kempowsky T. (2004a), "Surveillance de procédés à base de methods de classification: Conception d'un outil d'aide pour la detection et le diagnostic des défaillances", Doctoral Thesis, Institut National des Sciences Appliquées de Toulouse (France).

Kempowsky T. (2004b), "SALSA Situation Assessment using LAMDA Classification Algorithm", User's Manual, LAAS-CNRS, 2004.

Kempowsky, T., Subias, A., Aguilar-Martin, J., (2006). "Process situation assessment: From a fuzzy partition to a finite state machine". Engineering Applications of Artificial Intelligence, Vol. 19 (5), August 2006, Pages 461-477.

Lobanova H.V. , Lobanova G. V. (2003). " Approach for Restoration of Missing Data, Long-term Time Series and Generalization of Results" in " Advances in Water Supply Management- Maksimovic, Butler, Memon eds., Swets & Zeitlinger Publishers, The Netherlands.

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, 37(11), pp 155-162.

Maidment, D. R., E. Parzen (1984a). "Time patterns of water uses in six Texas cities". Journal of Water Resources Planning and Management, ASCE, 110(1), 90-106.

Maidment, D.R., E. Parzen (1984b). “Cascade model of monthly municipal water use”. Water Resources Research, 20(1), 15-23.

Maidment, D.R., Miaou, S.P., M.M. Crawford (1985). “Transfer function models of daily urban water use” Water Resources Research, 21 (4), 425-432.

Matheson, D., Jing, C., Monforte, F. (2004). “Meter Data Management for the Electricity Market”. 8th International Conference on Probabilistic Methods Applied to Power Systems, Iowa State University, Ames, Iowa, September.

Miaou, S.P. (1990) “A class of time series urban water demand models with non-linear climatic effects”. Water Resources Research, 26(2), 169-178.

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, pages 263-270, 2002.

Nelson P., Taylor P., MacGregor J. (1996), “Missing data methods in PCA and PLS: Score calculations with incomplete observations”, Journal of Chemometrics and Intelligent Laboratory Systems, vol. 35, pp 45-65.

Pastres, R., Ciavatta, S., Solidoro, C. (2003). “The Extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data”. Ecological Modelling, Vol. 170, Issues 2-3, 15, Pages 227-235, 2003.

Petit-Renaud, S., Denoeux, T. “A neuro-fuzzy system for missing data reconstruction”. 1998 IEEE Workshop on Emerging Technologies, Intelligent Measurement and Virtual Systems for Instrumentation and Measurement, Saint-Paul, USA, May 1998.

Piatyszek, E., Voignier, P., Graillot, D. (2000). “Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test“. Journal of Hydrology, Volume 230, Issues 3-4, Pages 258-268, 2000.

Piera N., Aguilar-Martin J. (1991), “Controlling Selectivity in Nonstandard Pattern Recognition Algorithms”, IEEE Transactions on Systems, Man, and Cybernetics, nº January-February 1991, pp.71-81.

Prescott S.L., Ulanicki B. (2001) “Time Series Analysis of Leakage in Water Distribution Networks” in Water Software Systems Theory and Applications. Research Studies Press, England.

Puig, V., Quevedo, J., Figueras, J., Riera, S., Cembrano, G., Salamero, M., Wilhelmi, G. (2003). “Fault detection and isolation of rain gauges and limnimeters of Barcelona’s sewer system using interval models IFAC SAFEPROCESS, Washington, USA,.

Quevedo J., Cembrano G., Valls A., Serra J. (1988) “Time-series Modelling of Water Demand – A Study on Short-term and Long-term Predictions” in Computer Applications in Water Supply, B. Coulbeck and C. H. Orr eds. Research Studies Press, England.

Ragot, J., Maquin, D. (2006). “Fault measurement detection in an urban water supply network”. Journal of Process Control, Volume 16, Issue 9, Pages 887-902, 2006.

Shvartser, L., Shamir, U., Fedman, M. “Forecasting hourly water demands by pattern recognition approach”. Journal of Water Resources Planning and Management, ASCE, 119 (6), 51-64.

Smith, J.A. “A model of daily municipal water use for short-term forecasting”. Water Research Resources, 24(2), 201-206.

Tsang, K.M. (2003). “Sensor data validation using gray models”. ISA Transactions 42, 9–17.

Valentin, N., Denoeux, T. (2001) “A neural network-based software sensor for coagulation control in a water treatment plant”. Intelligent Data Analysis, 5:23-39.

Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J. (2002). “Forecasting operational demand for an urban water supply zone”. Journal of Hydrology, 259, 189-202.