International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software - Ft. Collins, Colorado, USA - June 2018

Jun 27th, 10:40 AM - 12:00 PM

Linear prediction techniques for performance enhancement and maintenance of water networks using SCADA data

Arnau Martí
University of Vic - Central University of, arnau.marti@uvic.cat

Pere Marti-Puig
University of Catalonia, pere.marti@uvic.cat

Moisès Serra-Serra
University of Vic - Central University of Catalonia, moises.serra@uvic.cat

Manuel Pardo
AVSA Aigües Vic, mpardo@aiguesvic.com

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

Linear prediction techniques for performance enhancement and maintenance of water networks using SCADA data

Arnau Martí[a,b], **Pere Marti-Puig**[a], **Moises Serra-Serra**[a], **Manuel Pardo Méndez**[b]

[a] *Data and Signal Processing Group, U Science Tech, University of Vic - Central University of Catalonia, c/ de la Laura 13, 08500 Vic, Catalonia, Spain (arnau.mati,pere.marti,moises.serra@uvic.cat)*

[b] *AVSA Aigües Vic, Carrer de la Riera, 6, 08500 Vic, Catalonia, Spain(amarti,mpardo@aiguesvic.com)*

Abstract: From the analysis of the data captured in real time through a SCADA system, a contribution to improving the management of drinking water distribution and the early detection of anomalies is presented. In a real water network, the SCADA system must periodically acquire, store and validate the data collected by sensor measurements to achieve accurate network monitoring. For each sensor measurement, the raw data is usually represented by one-dimensional time series which must be validated before further use to ensure the reliability of the results obtained. In the present approach, we use linear predictors to verify data, detect outliers and restore missing values as well as to forecast different variables at different time intervals. The comparison of the predictions with measurements also serves to generate an error which is reported to an expert through warnings when it is unusually high. This human operator tries to associate significant prediction errors with pump configuration changes or system failures. The work mainly focuses on the predictor's configuration at different temporal levels.

***Keywords**:* water network, SCADA system, Linear prediction, condition-monitoring.

1 INTRODUCTION

In condition monitoring systems working on data collected by SCADA systems, it is critical to verify the integrity of the data received. Otherwise, errors are introduced in the very first step of the processing chain to propagate all along. The presented work is done in the Potable Water Treatment Station (PWTS) of Aigües de Vic S.A. where there are two basic problems with the sensors received information. Sometimes samples of the data series measurements are lost or suddenly false data are received. The first case usually may be because of a sensor failure. The second is caused by errors in the communications system that transmits and stores the information in the central database. To solve or minimize the effect of these troubles, the data of several PWTS sensors, stored by the SCADA system, have been analysed. As indicated by some investigations (Lamrini et al. 2011) it is possible to verify the integrity of the data and validate the obtained samples using predictive techniques. With these tools, even, we can reconstruct discarded samples.

The steps to make the analysis of the data will be the following: collect the historical signal data, make a validation of each sample according to basic criteria, perform a prediction of each sample to discard the incoherent samples and finally reconstruct the discarded samples (Quevedo et al. 2016; Blanch et al. 2009). The method presented and tested to predict the signal is based on a Wiener FIR (Finit Impulse Response) filter, which uses the previous data samples to make the estimation of the next sample. The difficult will be to get a filter configuration that allows estimating the lost or discarded data with reliability. Some modifications will be made to improve the initial algorithms and refine the predictions.

2. MATERIALS AND METHODS

2.1 Data Acquisition

The data history provided by SCADA d'Aigües de Vic S.A. will be used. All signal sensors contained in the ETAB remains in a SQL database in the Aigües de Vic S.A. server, where information has been stored since the beginning of 2014. There are mainly three types of signals. Those provided by flow meters installed in the pipes of the PWTS, which can be data of instant flow or volume of accumulated water. And those supplied by the level sensors of the water tanks. The study initially focuses on the level of water containers. The SCADA that stores the information is centralized on a computer which causes loss of data when this computer fails or is not available. The SCADA combines different communication systems to transmit the data from the sensors to the central computer. In some cases, it is necessary to use repeaters via radio that also causes data loss often. The data is collected every 5 minutes, except for some remote sensors that may have a minor frequency.

2.2 Software used

The Matlab program is used to manipulate and process all the information. This program allows data processing using scripts and includes a toolbox to connect to SQL databases so that all the database information can be easily captured and transferred to vectors with which the Matlab performs any mathematical treatment.

2.3 Data Preprocessing

Before processing the data, the system must ensure that certain integrity conditions are met (Cugueró-Escofet et al. 2016; Quevedo Casín et al. 2017). For this reason, the data vectors and associated dates are checked by validating each of the samples according to their physical and temporal characteristics. The following items are checked. The data read must satisfy the periodicity condition set by the SCADA system that stores the information. This means that the samples must have a temporal separation of 5 minutes and synchronized. For instance, that the minutes must always be multiples of 5 and the seconds of 0. When an inconsistent time is found we decided to eliminate the sample as, it has been observed that, in most cases, the value is also inconsistent. NaNs (Not a Number) are set in temporary positions where no data is available. The units of each magnitude are checked and correct if proceed. The tank level signals collected in the PWTS are done as a percentage and therefore can only take values from 0 to 100. It is checked that the increments of the magnitudes do not exceed the physical restrictions (the level in the tanks cannot vary more than 1-5% in 5 minutes).

2.4 The Wiener predictors

To predict lost samples, we propose a linear estimator that is implement with a finite impulse response filter (FIR). Linear estimators correspond to linear filter structures that are designed following a statistical criterion of minimizing the estimation error, in our case the Wiener filtering technique, which obtain the filter coefficients by minimizing the cost function $E[|e(n)|^2]$, where e(n) is the estimation error, E[·] denotes the expectation operator, and and |·| the Euclidian norm. Figure x, shows a diagram and its development for the case that applies. So, we have:

$$e_n = x_n - \hat{x}_n = x_n - \mathbf{a}^T \boldsymbol{x} \tag{1}$$

Taking into account that vector **a** contains the filter coefficients and vector $\boldsymbol{x}$ the samples used to predict $\hat{x}_n$, as follows:

$$\mathbf{a}^T = [a_0 \quad \cdots \quad a_{L-1}] \; ; \quad \boldsymbol{x}^T = [x_{n-1} \quad \cdots \quad x_{n-L}] \tag{2}$$

The Wiener filter is designed so as to minimize the mean square error (MMSE) criteria stated as:

$$a_l = \arg min\, E[|e_n|^2] \tag{3}$$

Using vector notation the FIR derivation to predict $\hat{x}_n$is quite straightforward. Let's see:

$$E[|e_n|^2]= E[e_n {e_n}^T] = E[(x_n - \mathbf{a}^T \boldsymbol{x})(x_n - \boldsymbol{x}^T \mathbf{a})] = E[x_n x_n - x_n \boldsymbol{x}^T \mathbf{a} \; - \mathbf{a}^T \boldsymbol{x} x_n + \mathbf{a}^T \boldsymbol{x}\boldsymbol{x}^T \mathbf{a}] \tag{4}$$

$$E[|e_n|^2] = E[x_n x_n] - 2E[x_n \boldsymbol{x}^T]\mathbf{a} \ + \mathbf{a}^T E[\boldsymbol{x}\boldsymbol{x}^T]\mathbf{a} \tag{5}$$

Being a quadratic function, the resolution of the equation will provide us the minimum of the function

$$\frac{\partial E[|e_n|^2]}{\partial \mathbf{a}} = -E[x_n \boldsymbol{x}^T] + E[\boldsymbol{x}\boldsymbol{x}^T]\mathbf{a} = 0 \tag{6}$$

which is:

$$\mathbf{a}{=}E[\boldsymbol{x}\boldsymbol{x}^T]^{-1}E[x_n \boldsymbol{x}^T]{=}\boldsymbol{R_{xx}}^{-1}\boldsymbol{r}_{xx} \tag{7}$$

Note than for real magnitudes$\boldsymbol{R_{xx}}$ is a symmetric Toeplitz matrix:

$$\boldsymbol{R_{xx}} = E[\boldsymbol{x}\boldsymbol{x}^T] = \begin{bmatrix} r_0 & \cdots & r_{L-1} \\ \vdots & \ddots & \vdots \\ r_{L-1} & \cdots & r_0 \end{bmatrix} \qquad \boldsymbol{r}_{xx}{=}E[x_n \boldsymbol{x}^T]{=}\begin{bmatrix} r_1 \\ \cdots \\ r_L \end{bmatrix} \tag{8}$$

A sliding time window is used to estimate the autocorrelation values (r_0,⋯, r_L). The window and the filter sizes are selected after a set of experiments.

2.5 Proposed method for signal reconstruction

It is clear that predictions are getting worse and worse as they try to anticipate more time in the future. This is a problem when bursts of values are lost. To minimize errors and reconstruct the databases, we propose to make a prediction called forward $\hat{f}_i$ to fill the lost values from the start of the burst and a backward prediction, $\hat{b}_i$ that supplies the missing values from the end of the burst to the past. As the values are more reliable at the extremes, to fill a burst of *N* missing samples (i goes from 1 to N) we weigh the estimate as follows:

$$\hat{x}_i{=}\frac{\hat{f}_i(N-i) + \hat{b}_i i}{N} \tag{9}$$

3. RESULTS

Two experiments have been designed to find the filter length and the optimal number of samples necessary for calculating the correlation. The first one computes the MSE for various values of the order of the filter, to be able to see the effect of each one of the parameters. It was determined that from an order of 245 coefficients there is practically no improvement in the MSE. Regarding the delay, as it was to foresee, more delay more error, in an entirely straightway.To be able to choose the number of samples required to obtain good estimations of the correlation, since they are not stationary systems, a test is performed in which the MSE is calculated for several values of the number of samples used. In this experiment, it is determined that around 250 samples are the most indicated. Some samples were removed from the original data to simulate intervals of lost data to verify de performance of the method. The graphic shows the result of the three proposed method, Normal FIR, Reverse FIR and mix FIR.

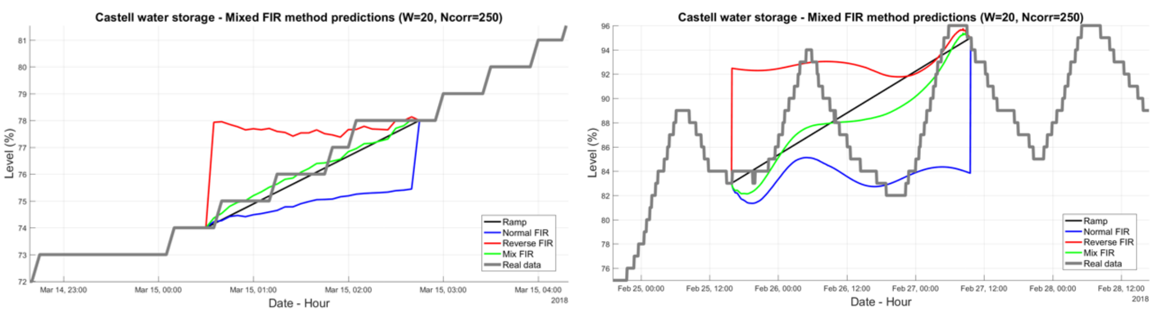

Figure 1 Two examples of simulating a burst of sample loss and reconstruction. The results are compared with the true data

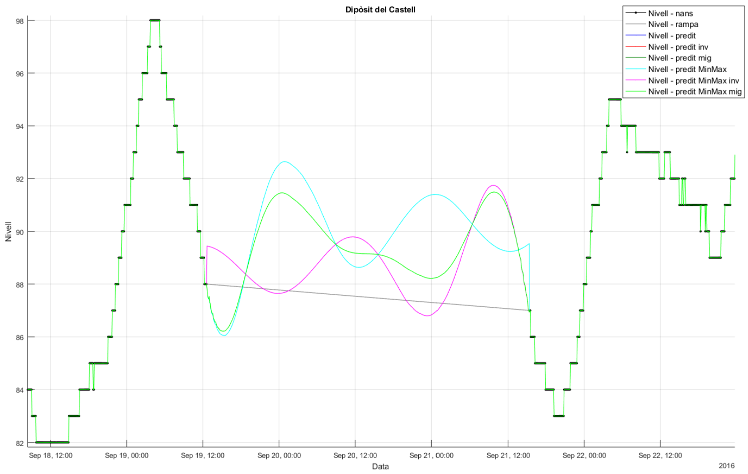

Figure 2 Reconstruction of a lost data stream.

1.5 CONCLUSIONS

The normal method with the FIR filter works correctly in the case of few consecutive lost samples. But in this case, a simple ramp or line joining the last data, before the first lost data, and the next data received after the last lost data, is a good approximation too. When the lost data burst is large (more than 25 samples, which represents some hours) the FIR method tends to get away from the reality. The filter configuration is not simple, because of the order and the number of samples for the autocorrelation calculation is critical and depends on the number of consecutive sample reconstructions that it's necessary to estimate. If the configuration is not sufficiently accurate, the FIR filter prediction error increases the longer is the burst of lost data. In this case, the proposed mix method between normal FIR and reverse FIR minimizes the prediction error. In addition, this method maintains the signal coherence, even if it does not exactly estimate the data burst.

ACKNOWLEDGMENTS

This work is partially supported by the Catalan Government under grant 2015DI40. Just thank the company Aigües de Vic S.A. that allowed us access to their databases to perform this research.

REFERENCES

Blanch, Jordi, Vicenç Puig, Jordi Saludes, and Joseba Quevedo. 2009. ARIMA Models for Data Consistency of Flowmeters in Water Distribution Networks. IFAC Proceedings Volumes 42 (8). Elsevier: 480–85. https://doi.org/10.3182/20090630-4-ES-2003.00080.

Cugueró-Escofet, Miquel À., Diego García, Joseba Quevedo, Vicenç Puig, Santiago Espin, and Jaume Roquet. 2016. A Methodology and a Software Tool for Sensor Data Validation/reconstruction: Application to the Catalonia Regional Water Network. Control Engineering Practice 49 (April): 159–72. https://doi.org/10.1016/j.conengprac.2015.11.005.

Lamrini, B., El-K. Lakhal, M-V. Le Lann, and L Wehenkel. 2011. Data Validation and Missing Data Reconstruction Using Self-Organizing Map for Water Treatment. Neural Computing and Applications 20 (4): 575–88. https://doi.org/10.1007/s00521-011-0526-5.

Quevedo, J., J. Pascual, S. Espin, and J. Roquet. 2016. Data Validation and Reconstruction for Performance Enhancement and Maintenance of Water Networks. IFAC-PapersOnLine 49 (28). Elsevier: 203–7. https://doi.org/10.1016/J.IFACOL.2016.11.035.

Quevedo Casín, Joseba Jokin, Diego García Valverde, Vicenç Puig Cayuela, Jordi Saludes Closa, Miquel Àngel Cugueró, Santiago Espin Basany, Jaume Roquet, and Fernando Valero Cervera. 2017. Real-Time Monitoring and Operational Control of Drinking-Water Systems. Edited by Vicenç Puig, Carlos Ocampo-Martínez, Ramon Pérez, Gabriela Cembrano, Joseba Quevedo, and Teresa Escobet. Advances in Industrial Control. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50751-4.