2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Inconsistent Sensor Data Detection/Correction: Application to Environmental Systems

Miquel À. Cugueró, Joseba Quevedo, Vicenç Puig and Diego García

***Abstract*— In this paper, a data detection/correction approach is proposed for a real environmental monitoring system, in order to provide a reliable dataset when sensor faults occur. This is the case of communication faults that may prevent the acquisition of a complete dataset, which is of paramount importance in order to successfully apply further system tasks such as fault diagnosis. Sensor detection/correction method presented here is based on the combined used of spatial and time series models. Spatial models take advantage of the physical relation between different variables emplaced in the system (temperature sensors here) while time series models take advantage of the temporal redundancy of the measured variables, by means of Holt-Winters models here. The proposed approach is successfully applied to the rock collapse forecasting system in the *Torrioni di Rialba* located in Lombardy (Italy).**

I. Introduction

Sensor network monitoring real environments requires a real-time scheme ensuring high performance and ease of maintenance under unfavourable conditions, such as sensor malfunction due to faults or aging, which may jeopardise overall system performance. To deal with this problem, the use of an on-line Fault Diagnosis Systems (FDS) able to detect such faults and correct them by means of different techniques is highly desirable. Also, the FDS should identify which fault has occurred including both hardware and software faults.

Generally, two main strategies may be found in the literature when addressing the FDS problem, which are hardware redundancy, based on the use of extra sensors and preferred in critical systems, and analytical redundancy, based on the use of software sensors or models, combining information gathered by the sensor measurements or using physical description for the process. Nevertheless, the use of hardware redundancy in sensor network systems is very expensive and it is not a commonly viable solution. Hence, in a real environment sensor network as considered here, the FDS must take advantage of software sensors or models, based on the spatial and temporal relationships among sensors in this network.

In this paper, an innovative framework investigating the application of data validation/correction methodology in [1] for sensor networks monitoring real environments, is proposed. Sensor detection/correction method is based on the combined used of spatial (both static and dynamic) and time series models of the sensors set. Spatial models take advantage of the relation between different variables in the system (temperature sensors here) while time series models take advantage of the temporal redundancy of the measured variables by means of Holt-Winters time series models. The proposed approach is successfully applied to the rock collapse forecasting system located in the *Torrioni di Rialba* (Lombardy, Italy).

The structure of the paper is as follows: In Section II, the methodology to validate/correct the sensor data, in order to provide a reliable dataset when faulty situations occur within the sensor set, is proposed. In Section III, the application case study is described, based on the environmental sensor network located in Torrioni di Rialba (Lombardy, Italy), measuring several real magnitudes of interest such the environmental temperature, and considering several real-world scenarios. In Section IV, the results obtained applying the proposed methodology to the case study in Section III are detailed. Finally, conclusions of this work are outlined in Section V.

II. Data detection/correction approach

In real systems such the one considered here, there is usually a telemeasurement system acquiring, recording and validating data gathered from different kind of sensors at each sample time to accurately real-time monitor the whole system [1]. In this process, problems in the communication system e.g. between sensors and data loggers or in the telemeasurement system itself, often arise and produce data loss which may be of great concern in order to have valid historic records. When this is occurring, lost data should be replaced by a set of estimated data which should be representative and coherent. Also, another common problem in real system monitoring is caused by the unreliable sensors, which may be affected by some fault (e.g. offset, drift, freezing) in their measurements. These unreliable data should also be detected and replaced by estimated data, since they may be used for system management tasks such that maintenance, planning, investment plans and system fault detection and isolation (Figure 1).

Different types of data detection methods with distinct degrees of complexity may be considered according to the available system knowledge. Generally, two types of methods are considered, one for elementary 'low-level' signal based methods and another for 'high-level' model-based methods. The first type uses simple heuristics and limited statistical information from the sensors [2] [3] and is typically based on

Miquel À. Cugueró, Joseba Quevedo, Vicenç Puig and Diego García are with Intelligent/Advanced Control Systems research groups (SIC/SAC) in Polytechnic University of Catalonia (UPC), Terrassa (Barcelona), Catalonia (email: {miquel.angel.cuguero, joseba.quevedo, vicenc.puig, diego.garcia}@upc.edu).

This work is supported by the European Union Seventh Framework Programme, in the framework of the iSense project, under grant agreement No. INSFO-ICT-270428.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

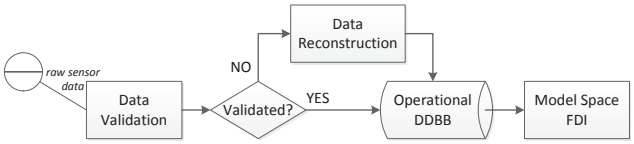

Fig. 1
RAW DATA DETECTION/CORRECTION AND SYSTEM FDI APPROACH

checking either signal values or variations, whilst the second type uses models for consistency-checking of the sensor data [4]. Here, the first type of data detection methods has been used to deal with sensor communication faults.

A. Data detection process

The data detection process is inspired by the Spanish AENOR-UNE norm 500540 [1]. The methodology presented here applies a set of consecutive detection tests to a given dataset (Figure 2), to finally assign a certain quality level depending on the tests passed.

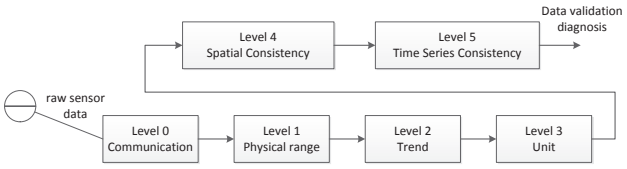

Fig. 2
DATA DETECTION TESTS

In a system like the one considered in this paper and in telemeasurement systems in general, one of the most common faults occurring are sensor communication faults. These type of faults are related with level zero of the sensor data detection methodology in [1]. This level checks whether the data is properly recorded, assuming that data acquisition systems sample data at a certain fixed rate. Hence, this level allows detecting problems in the data acquisition or communication system.

Here, communication faults are considered as the faults affecting the sensor of the telemeasurement system, and the data detection/correction procedure is used as a prefilter to estimate the missing data when this type of faults is occurring.

B. Data correction process

The output of the data detection process (Figure 2) is used to identify the invalidated data that should be reconstructed. Static and Dynamic Spatial Models (SSM and DSM, respectively), both related with Level 4 in Figure 2, and Time Series Models (TSM), related with Level 5 in Figure 2, are used for this purpose, depending on the performance of each model.

SSM and DSM take advantage of the relation between different variables physically related. SSM can be obtained using linear regression input-output models among higher correlated sensors as follows

$$\hat{x}_{s_n}(k) = a_n x_{f_n}(k) + b_n \tag{1}$$

where, given $f-s$ pairs of higher correlated sensors, $\hat{x}_{s_n}(k)$ is the estimation of s_n sensor data, given its high correlated paired sensor measurement $x_{f_n}(k)$ and the linear regression model parameters $[a_n, b_n]$ for model n.

Also, different wide used structures of DSM have been considered for the same purpose, including AutoRegressive with eXternal input (ARX) models, AutoRegressive-Moving-Average with eXternal input (ARMAX) models and Output Error (OE) models, for given $f-s$ pairs:

ARX_{sf}:

$$A(q)\hat{y}_s(t) = B(q)u_f(t) + e(t) \tag{2}$$

$ARMAX_{sf}$:

$$A(q)\hat{y}_s(t) = B(q)u_f(t) + C(q)e(t) \tag{3}$$

OE_{sf}:

$$\hat{y}_s(t) = \left[\frac{B(q)}{F(q)}\right] u_f(t) + e(t) \tag{4}$$

where $u_f(t)$ is the system input, $\hat{y}_s(t)$ is the forecasted system output, $e(t)$ is white noise of variance λ, and A, B, C, F are the parameter polynomials expressed in terms of the time-shift operator q^{-1}, so $q^{-1}u_f(t) = u_f(t-T)$, being T the sampling interval.

Alternatively, TSM take advantage of the temporal redundancy of the measured variables. A wide used method for time series modelling is the Holt-Winters (HW) approach [6]. This method is widely used because of its simplicity. There are different versions of this method e.g. additive or damped trend, additive or multiplicative seasonality, single or multiple seasonality [5]. Here, good performance has been attained with the additive single seasonality version, which estimated value is obtained for a forecasting horizon ℓ

$$\hat{x}_{TS}(k) = \bar{R}(k-\ell) + \ell\bar{G}(k-\ell) + \bar{S}(k-L) \tag{5}$$

where $\bar{R}$ is the level estimation removing seasonality,

$$\begin{aligned}\bar{R}(k-\ell) =& \alpha\left(x(k-\ell) - \bar{S}(k-L-\ell)\right) \\ &+ (1-\alpha)\left(\bar{R}(k-\ell-1)\right. \\ &\left.+\bar{G}(k-\ell-1)\right) \quad 0<\alpha<1\end{aligned} \tag{6}$$

$\bar{G}$ is the trend estimation,

$$\begin{aligned}\bar{G}(k-\ell) =& \beta\left(\bar{R}(k-\ell) - \bar{R}(k-\ell-1)\right) \\ &+ (1-\beta)\,\bar{G}(k-\ell-1) \quad 0<\beta<1\end{aligned} \tag{7}$$

$\bar{S}$ is the seasonal component estimation,

$$\begin{aligned}\bar{S}(k-\ell) =& \gamma\left(x(k-\ell) - \bar{R}(k-\ell)\right) \\ &+ (1-\gamma)\,\bar{S}(k-\ell-L) \quad 0<\gamma<1\end{aligned} \tag{8}$$

and L is the season periodicity, α, β and γ are the HW parameters (level, trend and season smoothing factors, respectively), x is the measured value and $\hat{x}_{TS}$ is the TSM estimated value.

Hence, analysing the historic records of a certain sensor, a HW TSM model can be obtained and used to estimate missing data of this element when a fault is affecting its readings.

The models accuracy is measured by the Mean Squared Error (MSE) of each model, evaluated in the m previous values to k

$$MSE(k) = \frac{1}{m}\sum_{j=k-m}^{k} e(j)^2 \tag{9}$$

where m is the number of data, $e(j) = x(j) - \hat{x}(j)$ is the error at instant j, $x(j)$ is the measured value at instant j, $\hat{x}(j)$ is the estimated value by the model (SSM, DSM or TSM, respectively) at instant j and k is the actual time instant. The model having best MSE index when the communication fault is produced (i.e. when a fault is detected by the detection process) is used to produce the reconstructed sensor signal.

III. Case Study: Environmental Application

The case study considered here involves the rock collapse forecasting system installed in the *Torrioni di Rialba*, which is located in the Alps of Lombardy in northern Italy (Figure 3). This is a real-time rock-fall monitoring system designed by Politecnico di Milano, including several types of sensors such Micro Electro-Mechanical Systems (MEMS) accelerometers, geophones, inclinometers, crackmeters and temperature sensors, in order to non-invasively check for micro-acoustic bursts, which are related with the formation of cracks in the rocks conforming the environment. The information gathered by these sensors is monitored from a Control Room located in Lecco city (Figure 4). The system has been installed in some critical areas of the Italian-Swiss Alps, such Saint Martino mountain (April, 2010), Val Canaria (Ticino, Switzerland, August 2011), Gallivaggio (July 2012) or the case considered here (Torrioni di Rialba, July 2010). The implemented monitoring system is conformed by a network of systems, which are meant to detect and localize micro-acoustic emissions from the rock surface while keeping a low energy consumption, which is of great importance in environmental monitoring setups. The benchmark used here is related to the measurements gathered by a new generation of intelligent clinometer sensors, which incorporate an internal thermal sensor used to compensate and correct the online measurements provided by these units. Here, an study on the data loss of these temperature sensors is performed. Concretely, there are three temperature sensors installed in this emplacement, which may suffer communication faults, a widely spread cause of missing data problems in real telecontrolled systems [1].

Fig. 3
Torrioni di Rialba emplacement

Fig. 4
Torrioni di Rialba, area map

IV. Results

The scenario dataset is depicted in Figure 5. The time range of the measurements is from 2012-07-25 17h00 to 2012-10-17 12h30, with a sampling rate of 10 min. The sampling rate for all the scenario is depicted in Figure 6 where it can be observed how a regular sample period of 600 s (10 min) is respected for almost all the samples. However, there are some samples which present longer sampling rate (i.e. when a communication problem is occurring), having its maximum at 4200 s. This is a relatively low gap between samples (around 7 missing measurements).

Hence, a simulated communication fault enduring a whole day (i.e. 144 samples) has been used as a test scenario here. The available dataset has been divided into different parts: identification dataset (first 1440 samples), validation dataset (next 576 samples) and test dataset (remaining data), where the described simulated communication fault has been applied at sample 3250 (enduring 1 day). Regarding the linear regression SSM and the TSM used for this benchmark, the sensor pairs and parameters for each model are shown in Table I.

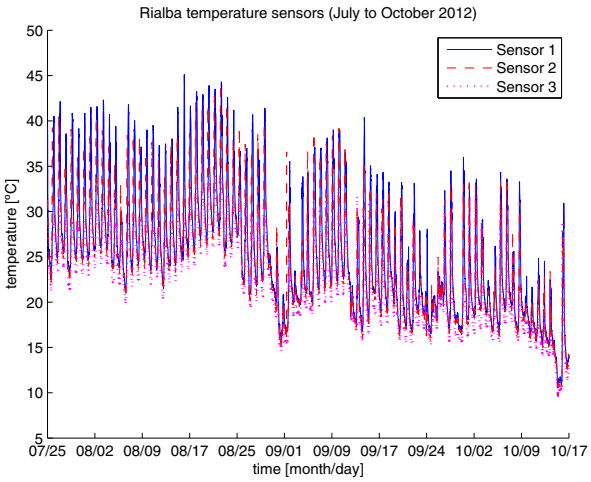

Fig. 5
RIALBA SCENARIO

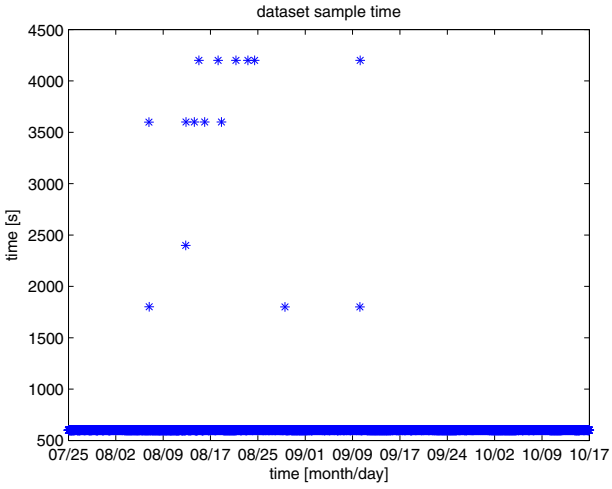

Fig. 6
RIALBA SCENARIO SAMPLE TIMES

Regarding DSM, different type of model structures have been used: ARX, ARMAX and OE. In order to obtain the models, mean and trend of the data have been removed for identification purposes. This process helps estimating linear models in a more accurate way, since they cannot cope arbitrarily differences between input and output signal levels. For steady state data (the case considered) this process should be applied in both input and output model data. The best models have been obtained according to Rissanen Minimum Description Length (MDL) and Akaike Information Criterion (AIC). The quality of the identified models is measured in terms of Loss Function, Akaike's Final Prediction Error (FPE) and Data Fit [7].

Considering e.g. the models obtained for the sensor 1 using the other available sensors (i.e. sensor 2 and sensor 3, respectively), the following models are obtained when sensor 2 is used as input:

ARX_{12}:

$$\begin{aligned} A_{12}(q) &= 1 - 1.476q^{-1} + 0.4151q^{-2} - 0.001438q^{-3} \\ &\quad + 0.06736q^{-4} + 0.01919q^{-5} \\ B_{12}(q) &= 0.02426q^{-1} - 0.01991q^{-2} - 0.005864q^{-3} \\ &\quad - 0.007134q^{-4} + 0.03287q^{-5} \\ fit &= 50.74\% \end{aligned}$$

$ARMAX_{12}$:

$$\begin{aligned} A_{12}(q) &= 1 - 1.925q^{-1} + 0.932q^{-2} \\ B_{12}(q) &= 0.007096q^{-2} \\ C_{12}(q) &= 1 - 0.5005q^{-1} \\ fit &= 51.2\% \end{aligned}$$

OE_{12}:

$$\begin{aligned} B_{12}(q) &= 0.019q^{-2} + 0.07228q^{-3} \\ F_{12}(q) &= 1 - 0.9193q^{-1} \\ fit &= 62.33\% \end{aligned}$$

The outputs provided by these models are depicted in Figure 7. Alternatively, when sensor 3 is considered as the model input, the following models are obtained:

ARX_{13}:

$$\begin{aligned} A_{13}(q) &= 1 - 1.965q^{-1} + 0.9707q^{-2} \\ B_{13}(q) &= 0.5012q^{-1} - 1.174q^{-2} + 0.8603q^{-3} \\ &\quad - 0.1734q^{-4} - 0.02324q^{-5} + 0.0572q^{-6} \\ &\quad - 0.2332q^{-7} + 0.409q^{-8} - 0.2168q^{-9} \\ fit &= 41.9\% \end{aligned}$$

$ARMAX_{13}$:

$$\begin{aligned} A_{13}(q) &= 1 - 1.906q^{-1} + 0.9175q^{-2} \\ B_{13}(q) &= 0.03744q^{-1} - 0.03458q^{-2} - 0.01869q^{-3} \\ &\quad + 0.005067q^{-4} + 0.03134q^{-5} - 0.02729q^{-6} \\ &\quad - 0.0084q^{-7} + 0.1485q^{-8} - 0.1195q^{-9} \\ C_{13}(q) &= 1 - 0.4621q^{-1} - 0.2066q^{-2} \\ fit &= 53.81\% \end{aligned}$$

TABLE I

SSM/TSM PARAMETERS, RIALBA BENCHMARK

	SSM					TSM		
n	sensor s_n	sensor f_n	correlation	a_n	b_n	α_n	β_n	γ_n
1	1	3	0.847024	0.442	13.105	1	1×10^{-5}	0.237
2	2	3	0.847488	0.519	11.213	1	0.762	1×10^{-5}
3	3	2	0.847488	1.383	-7.405	0.502	3.89×10^{-4}	0.344

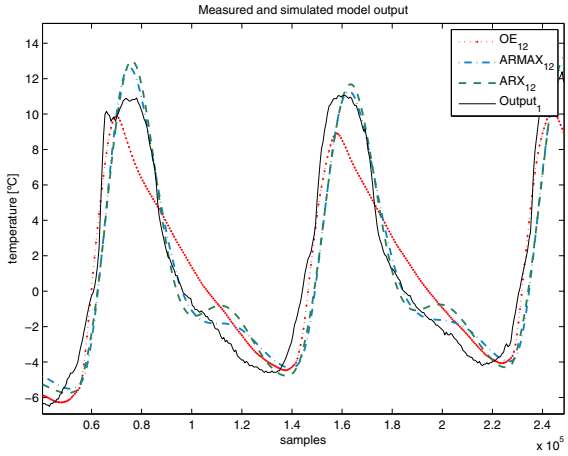

Fig. 7

SENSOR 1 MODELS PERFORMANCE USING SENSOR 2 AS MODEL INPUT

OE_{13}:

$$B_{13}(q) = 0.3978q^{-1} - 0.1375q^{-2}$$
$$F_{13}(q) = 1 - 0.8095q^{-1}$$
$$fit = 51.23\%$$

The outputs provided by these models are depicted in Figure 8.

Taking into account all the possible models for sensor 1 data reconstruction (Figure 7 and Figure 8), an ARMAX model using sensor 2 is chosen since it obtains a good fit (i.e. 51.2 %) with a reduced order (i.e. 2) when compared with the rest of the models obtained. This model also provides a qualitative good prediction (see Figure 9) when compared with the others. Similar procedure has been applied to the rest of the sensors in order to select the best models to reconstruct their data when a communication fault is affecting them. Similarly as explained for sensor 1, the trade-off between model fit and order is considered. As a result of this selection procedure, a different DSM has been obtained for each sensor, as detailed in Table II. The model expressions for each sensor are as follows

Sensor $n = 1$:

$$A_{12}(q) = 1 - 1.925q^{-1} + 0.932q^{-2}$$
$$B_{12}(q) = 0.007096q^{-2}$$
$$C_{12}(q) = 1 - 0.5005q^{-1}$$
$$fit = 51.2\%$$

Sensor $n = 2$:

$$A_{21}(q) = 1 - 0.9446q^{-1} - 0.9779q^{-2} + 0.9238q^{-3}$$
$$B_{21}(q) = 2.054q^{-1} - 2.99q^{-2} - 1.167q^{-3} + 2.934q^{-4} - 0.6256q^{-5} - 0.1925q^{-6} - 0.08352q^{-7} + 0.1556q^{-8} - 0.08346q^{-9}$$
$$C_{21}(q) = 1 + 0.06096q^{-1}$$
$$fit = 51.84\%$$

Sensor $n = 3$:

$$B_{32}(q) = 0.5547q^{-7} - 0.5428q^{-8}$$
$$F_{32}(q) = 1 - 0.9871q^{-1}$$
$$fit = 42.98\%$$

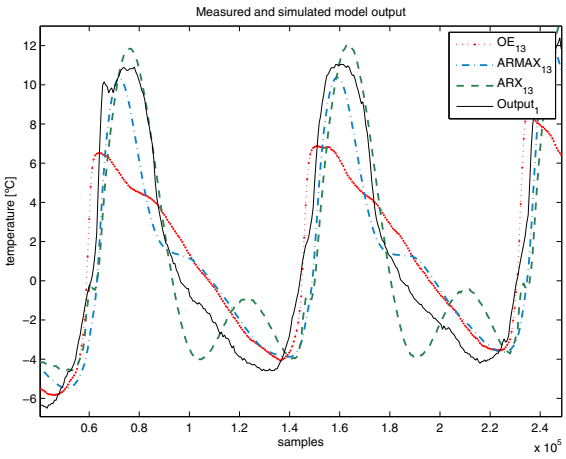

Fig. 8

SENSOR 1 SELECTED MODEL PERFORMANCE USING SENSOR 3 AS MODEL INPUT

The outputs provided by these DSM models for sensors 2 and 3 are presented in Figure 10 and Figure 11, respectively.

Finally, the performance of the best DSM models obtained is compared with the SSM and the TSM models in order

TABLE II

DSM PARAMETERS, RIALBA BENCHMARK

n	sensor s_n	sensor f_n	model structure	Loss Func.	FPE	Fit
1	1	2	ARMAX	0.109607	0.109753	51.2 %
2	2	1	ARMAX	1.14146	1.14641	51.84 %
3	3	2	OE	0.840637	0.841478	42.98 %

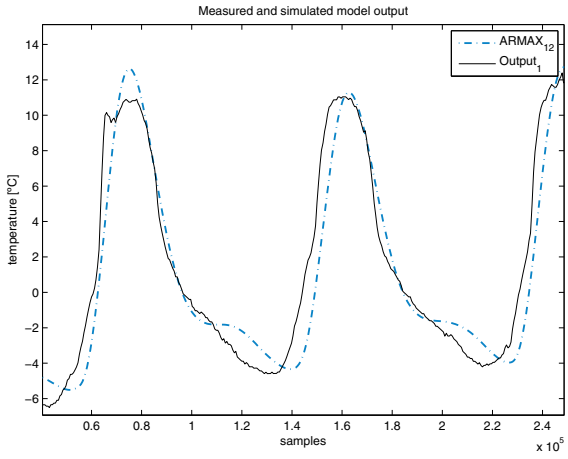

Fig. 9

SENSOR 1 SELECTED MODEL PERFORMANCE USING SENSOR 2 AS MODEL INPUT

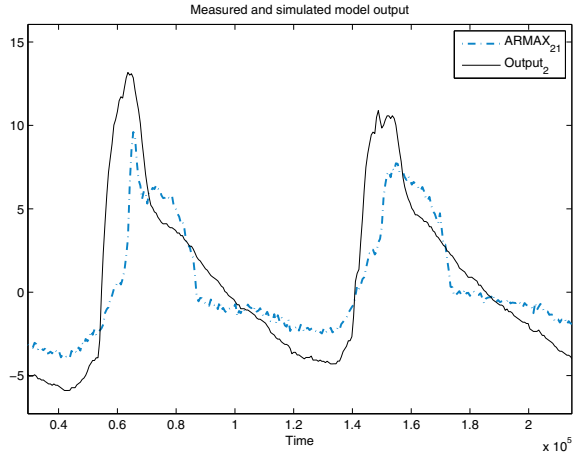

Fig. 10

SENSOR 2 SELECTED MODEL PERFORMANCE USING SENSOR 1 AS MODEL INPUT

to use the best among them to correct the sensor signal. The results attained with these models for the sensor communication faults are shown from Figures 12 to 14. In the latter figures, upper sub-plots depict the outputs obtained by SSM (dashed), DSM (dash-dotted) and TSM (thick dashed) models, gathering the estimated data best representing the measured data (solid) according to their MSE index, depicted in sub-plots below. For sensor 1 (see Figure 12), it may be observed how the model having the best performance (i.e. lowest MSE when the fault is produced) is the DSM, hence it is used for data reconstruction of this sensor. In the case of sensor 2, the model of choice according to the same criterion is the SSM (see Figure 13). In the latter figure, TSM plots are out of range due to bad performance of this model for this particular sensor and hence do not appear in the plot. In the case of sensor 3, the model considered to reconstruct the sensor signal is the DSM (see Figure 14).

V. CONCLUSION

In this paper, an effective sensor data detection/correction method has been proposed to provide reliable datasets in a real environmental data telemeasurement system. The method proposed detects and corrects sensor data communication faults using a data pre-filter, in order to provide reliable datasets to be used in further system operations such fault diagnosis or system monitoring. Regarding the sensor data pre-filter, sensor data validation and reconstruction using combined spatial and time series models have been implemented with successful results when communication faults occur in the telemeasurement system, which is one of the most common faults affecting this kind of systems. The proposed method has been successfully tested with the temperature sensors implemented in the rock collapse forecasting system in the Torrioni di Rialba, located in Lombardy (Italy).

ACKNOWLEDGMENT

This work is supported by the European Union Seventh Framework Programme, in the framework of the iSense project, under grant agreement No. INSFO-ICT-270428, and by the Spanish research project SHERECS, under grant agreement No. DPI2011-26243. The authors of this work would also like to thank the contribution of Prof. Cesare Alippi and Dr. Manuel Roveri from the Politecnico di Milano (PoliMi).

REFERENCES

[1] J. Quevedo, V. Puig, G. Cembrano, J. Blanch, J. Aguilar, D. Saporta, G. Benito, M. Hedo, A. Molina, Validation and reconstruction of flow meter data in the Barcelona water distribution network, Control Engineering Practice 18 (6) (2010) 640–651. doi:10.1016/j.conengprac.2010.03.003.

[2] D. Burnell, Auto-validation of district meter data, in: CCWI '03 Advances in Water Supply Management, London, 2003.

[3] H. Jörgensen, S. Rosenörn, H. Madsen, P. Mikkelsen, Quality control of rain data used for urban run-off systems, Water Science and Technology 37 (11) (1998) 113–120.
[4] K. Tsang, Sensor data validation using gray models, ISA Transactions 42 (2003) 9–17.
[5] S. Makridakis, S. Wheelwright, R. Hyndman, Forecasting methods and applications, John Wiley & Sons, 1998.
[6] P. R. Winters, Forecasting sales by exponentially weighted moving averages, Management Science 6 (52) (1960) 324–342.
[7] L. Ljung, System Identification: Theory for the User, Prentice Hall (2nd Edition), 1999.

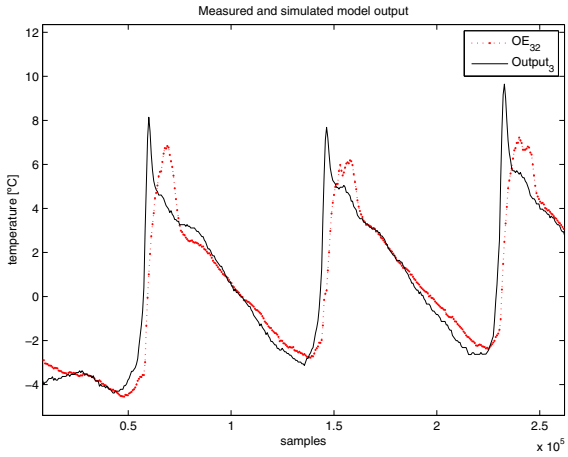

Fig. 11
SENSOR 3 SELECTED MODEL PERFORMANCE WITH SENSOR 2 INPUT

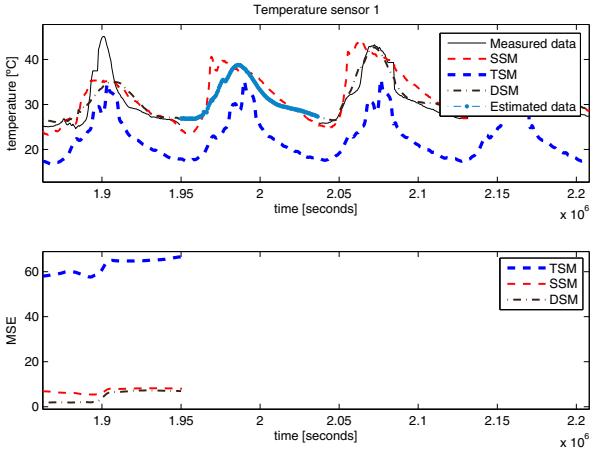

Fig. 12
SENSOR 1 COMMUNICATION FAULT

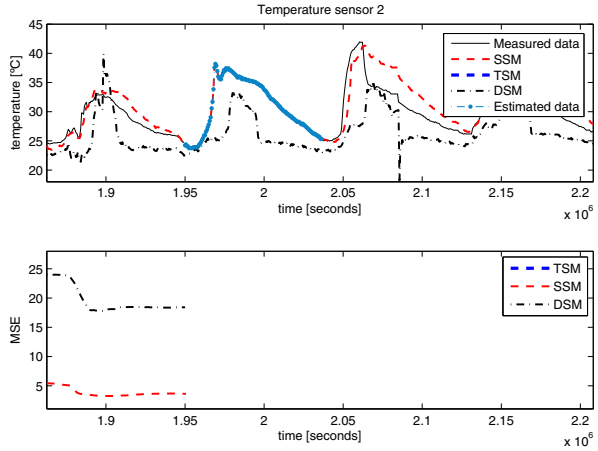

Fig. 13
SENSOR 2 COMMUNICATION FAULT

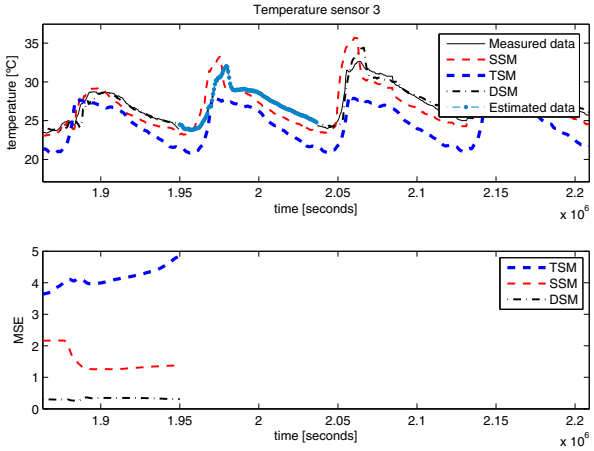

Fig. 14
SENSOR 3 COMMUNICATION FAULT