Tampered Data Recovery in WSNs through Dynamic PCA and Variable Routing Strategies

Roberto Mag án-Carri ón, Fernando Pulido-Pulido, Jos é Camacho, and Pedro Garc á-Teodoro
Department of Signal Theory, Telematics and Communications, Faculty of Computer Science and Telecommunications – CITIC, University of Granada, Periodista Daniel Saucedo Aranda, s/n, E-18071 GRANADA (Spain)
Email: rmagan@ugr.es, fpulido@correo.ugr.es, josecamacho@ugr.es, pgteodor@ugr.es

***Abstract*—**Wireless sensor networks (WSNs) are highly sensible to data integrity attacks, which have an important impact on a number of relevant deployments and services. This paper introduces a tolerance approach to fight against data modification attacks in WSNs, which is based on a missing data imputation scheme. The proposal relies on two principal contributions: (1) a multivariate statistical technique where the dynamics of the sensor measurements for the monitored area are captured through the use of dynamic PCA (DPCA), and (2) a variable routing strategy that improves the recovering performance by spreading the effects of the data tampering attack. On the other hand, a complementary multivariate statistical anomaly detection module is implemented to determine the occurrence of data tampering attacks and trigger the subsequent reaction procedure to recover the affected data. As shown by the results obtained, the proposed tolerance approach improves the robustness of a WSN against data tampering attacks, and so its survivability and normal operation over time.

***Index Terms*—**missing data imputation, anomaly detection, Multivariate analysis, response/tolerance schemes, survivability, wireless sensor networks

I. Introduction

Sensor networks are composed of a number of sensing elements deployed to monitor some physical variables for a given area [1]. Although this kind of networks can be structured, a common trend is the use of wireless links in what it is called a wireless sensor network, or WSN. Whatever the case, there usually exists a central unit (CU) to gather and analyze the data generated by the sensors for the covered area. In a WSN, the data collected by the sensors can be transmitted directly to the CU (e.g., through GPS connections), but it is more usual to transmit such information by means of the so-called multi-hop routing strategies, where sensor devices operate not only as measuring nodes but also as relaying nodes to raise origin-destination connectivity. Multi-hop routing is more efficient in terms of energy consumption, which is desirable to reduce the WSN maintenance.

Given the wide adoption of WSNs for several relevant applications (military actions, crisis management and disaster recovery, medical, industry, etc.) [2], implementing security mechanisms to strengthen the services provided must be encouraged. For that, it is important to be aware of the existence of several inherent threats in WSNs [3]: packet dropping, route poisoning, identity spoofing, etc. In this context, the so-called data tampering, environmental tampering or tampering attack [3] [4], which affect data integrity, can lead to disastrous consequences. For instance, take the case of fire monitoring, which the application is considered in the present paper. Due to a malicious tampering attack, the fire may not be detected until it is too late to put it out. This way, a potential pyromaniac who wants to start a fire inside a monitored area only needs to physically alter the measurement provided by one or more sensors to deviate the attention of the fire brigade. The true fire, which is located at a different area, would progress without resistance.

In consequence, the deployment of efficient security schemes to reduce risks and threats are necessary. This can be tackled by providing proper mechanisms to obtain or estimate the compromised data, which will contribute to improve the survivability of the network, understood as *"the ability of a system to fulfill its mission, in a timely manner, in the presence of attacks, failures or accidents"* [5].

With this principal aim, the present work proposes the application of multivariate analysis techniques to recover WSN environments from data tampering attacks. For that, we first introduce the use of principal component analysis (PCA) [6] [7] to monitor and detect anomalies in the system behavior over time. After that, PCA-based trimmed scores regression (TSR-PCA) [8] [9] is introduced to subsequently recover tampered data. In order to improve the recovering performance of the tolerance approach, a dynamic version of PCA (dynamic PCA, or DPCA) is considered. DPCA incorporates the temporal information that is the auto-correlation and lagged cross-correlations of the sensor measurements, into the behavioral model of the WSN. Additionally, we show how the specific routing strategy chosen for multi-hop retransmissions has a relevant influence on the impact of data integrity attacks on the WSN performance. Some scenarios are analyzed to evidence the consequences on the number of sensors affected depending on both, the routing algorithm used and the location of the specific sensor attacked. A number of

Manuscript received June 15, 2013; revised October 24, 2013.
doi:10.12720/jcm.8.11.738-750

©2013 Engineering and Technology Publishing

variable routing strategies are studied and proposed to spread the effect of an attack event over time. This minimizes the consequences of the attack and thus improves the recovery performance.

The main contributions of the present paper are as follows:

- The development of a dynamic variant of a multivariate statistical-based reaction scheme to recover tampered data in WSNs. The recovery system is complemented with a multivariate statistical-based anomaly detection module that determines the occurrence of attack events against the target system and triggers the necessary alarm to execute the recovery procedure [10].
- The use and evaluation of variable routing schemes to spread the consequences of the attack over time and thus improve the recovery results.

The result of adopting these proposals, variable multi-hop routing strategies and dynamic PCA, is the enhancement of the recovery performance with respect to that obtained employing static multi-hop routing strategies and static PCA modeling [10].

The rest of the paper is organized as follows. Section II presents some relevant works related to the subject under study. Section III discusses the fundamentals of the multivariate analysis techniques used in the present work, both for monitoring and detection purposes and for missing data recovery. Section IV discusses the arrangement of the data collected in a WSN for standard (static) PCA and for dynamic PCA. Section V introduces variable routing strategies, as they will be subsequently studied to improve missing data recovery. After that, Section VI describes the general framework considered for testing our tolerance proposal, and the results obtained in terms of anomaly detection and missing data recovery are presented and discussed. Moreover, a discussion about the consequences of choosing different types of variable routing strategies is also presented in this section. Finally, Section VII summarizes the principal conclusions and remarks on this work as well as new future research directions.

II. Related Work

For network survivability [5] at least two principal security mechanisms must be deployed: attack detection (recognition) and response (recovery). Regarding the detection of non-legitimate network events, several solutions have been proposed in the literature. Among others, statistical-based, rule-based and data mining-related techniques can be found [11] to determine unauthorized events in monitored environments.

Those non-legitimate events which are detected need to be subsequently solved to guarantee the continuity of the network and the affected services. For that, response mechanisms must be devised and provided. In this context, missing data imputation techniques are used as response techniques. For example, the authors in [12] propose an anomaly detection scheme and missing data imputation algorithm based on neural networks to improve the performance of the classification process performed by the neural network. The network is first partitioned into clusters, and then the missing data imputation algorithm selects the nearest neighbor or the most repeated value of the neighborhood to estimate the missing value for the target sensor. If there are no neighbors, the predicted value is the last one obtained from the corresponding sensor. In reference [13], the authors provide a data mining-based technique addressing the missing data imputation problem in mobile sensor networks. They divide the monitored mobile sensor area into sub-areas. In each of them there exists a virtual static sensor which monitors the real mobile sensor readings by computing the mean of the corresponding real sensor measurements. Through the relationships among the virtual static sensors, their proposal exploits the spatial and temporal correlation to predict missing data from the real sensors. Another study in reference [14], addresses a robust method to recover missing data using two temporal predictors and one spatial. The algorithm selects the best predictor by assessing each one when there are missing data, thus showing how sampling rate and packet loss affect recovery accuracy.

A missing data recovery proposal using sparsity-spatial interpolation with a fixed discrete cosine transform (DCT) basis is addressed in [15]. An improvement of the previous work is proposed in [16] through the use of a sparsity-based missing data recovery method. By using and over-complete dictionary conformed by 2D DCT basis and past frames, their approach estimates a given missing data frame as a linear combination obtained from the dictionary by solving the l_1 norm optimization problem, thus accounting for the temporal data correlation. Using dynamic Bayesian networks, the authors in [17] develop a novel anomaly detection and missing data imputation technique by exploiting the spatial and temporal correlation existing between samples. They are capable of distinguishing if an anomaly is taking place by comparing a normality sensor model (data calibration-based model) and the actual sensor value. The imputation or recovery method is then addressed by inferring the most likely sensor value from both the current and immediate past values.

The use of multivariate techniques has been widely adopted in the scientific literature by the research community, but their application to WSNs is still limited. These methodologies fit well in this kind of environments, because multivariate techniques can model the high temporal and spatial correlation among sensors.

The limited number of multivariate analysis proposals for WSNs are mainly devised for intrusion or anomaly detection purposes. As an example, a PCA-based detection system for routing attacks is proposed in [18]. The network is partitioned into groups with a monitor per group with two PCA models: one for its own traffic and one for global traffic, which is obtained by exchanging its local PCA model with other monitors. Authors conclude that a PCA global distributed modeling achieves better detection performance than the centralized modeling for sinkhole attacks. Another PCA-based anomaly detection system is also proposed in [19]. Two phases are involved in this system: data modeling and anomaly detection. The first one is intended to improve the PCA modeling against outliers and inconsistent data. The anomaly detection procedure is carried out by comparing the calibration data and the new incoming data using the Mahalanobis distance.

On a different matter, several routing strategies are studied and proposed in the literature aimed at improving alternative aspects: load balance, confidentiality, etc. This way, a number of specific routing algorithms for ad hoc networks such as AODV, DSR, LSR coexist. However, to the best of our knowledge, no studies are available where the relationship between the underlying routing algorithm and the effect of attacks such data tampering attacks is analyzed.

In this context, the contribution of the present paper is twofold: (a) the use of multivariate techniques where the system dynamics are incorporated to the model, and (b) the use of variable routing strategies, where time-varying next hop nodes are considered to establish the origin-destination routes.

III. Multivariate Statistical Analysis and Missing Data Recovery

Most natural and man-made processes are multivariate systems, as their adequate characterization requires the joint use of several variables. For instance, weather forecasting depends on wind, atmosphere pressure and temperature, among many other factors.

Data description and modeling, discrimination and classification, or regression and prediction [20] are the usual fields for applying multivariate techniques. In the following sections, the fundamentals of multivariate statistical analysis in the context of this work are provided, both for monitoring and detection and for data recovery.

A. Principal Component Analysis (PCA)

The main goal of *principal component analysis*, or PCA, is data compression. PCA identifies a number of linear combinations of the original variables, the so-called *principal components* (PCs), which contain most of the relevant information (variability) in a data set **X**. This is a change of variables from the original variables in the **X** space to the PCs subspace. If **X** is a data matrix with *J* variables, PCA reduces its dimension from *J* variables to A PCs by finding the *A* - dimensional latent subspace of most variability captured.

PCA follows the next equation:

$$\boldsymbol{X} = \boldsymbol{T}_A \cdot \boldsymbol{P}_A^T + \boldsymbol{E}_A \qquad (1)$$

where $\mathbf{P}_A$ is the $J \times A$ loading matrix, $\mathbf{T}_A$ is the $I \times A$ score matrix and $\mathbf{E}_A$ is the $I \times J$ residual matrix. The maximum variance directions are obtained from the eigenvectors of $\mathbf{X}^T \cdot \mathbf{X}$, and they are ordered as the columns of $\mathbf{P}_A$ by explained variance. The rows of $\mathbf{T}_A$ are the projections of the original *I* observations in the new latent subspace. $\mathbf{E}_A$ is the matrix that contains the residual error, and it plays a crucial role in anomaly detection. The score or projection on the PCA subspace of a new observation is obtained as follows:

$$\boldsymbol{t}_{new} = \boldsymbol{x}_{new} \cdot \boldsymbol{P}_A \qquad (2)$$

The number of PCs in a model *A*, can be selected using several methods, including cross-validation [21]. The authors in [22] conclude that the *element-wise k-fold* (ekf) algorithm is a valid choice for PCA cross-validation when the model is used for missing data imputation, as in the present work.

B. Dynamic PCA

The loadings in PCA capture the relationships among the data variables. In traditional PCA, only static relationships are captured. To address this limitation, dynamic PCA (DPCA) was proposed by *Ku et al.* [23] in order to incorporate time inter-dependencies into the model. Modeling data dynamics is of relevance in different practical applications and engineering disciplines such as automatic control and system modeling [24].

DPCA performs PCA by extending the original **X** data matrix with the addition of the same variables lagged in time. The new matrix $\mathbf{X_d}$, augmented with *d* lags, contains in each row the observations from sampling time $k-d$ to sampling time *k*. This means that the number of variables in $\mathbf{X_d}$ grows with *d*, following:

$$\mathbf{X}_d = \begin{bmatrix} x(1) & x(2) & \cdots & x(d+1) \\ x(2) & x(3) & \cdots & x(d+2) \\ \vdots & \vdots & \ddots & \vdots \\ x(I-d) & x(I-d+1) & \cdots & x(I) \end{bmatrix} \qquad (3)$$

where $\mathbf{x}(k) = [x_1(k)x_2(k)\ldots x_J(k)]$ is the J -dimensional observation vector (one observation per sensor) at sampling time *k*, *d* the time lag and *I* the total number of observations per sensor or variable. The new matrix obtained $\mathrm{X_d}$ is $(I-d) \times J \cdot (d+1)$-dimensional. This way, applying PCA to the extended matrix, the dynamic relationships in the data are captured by the model.

The choice of the lag d can be performed using the multiple methods available in the literature to investigate the dynamics of time-series data. This includes the use of auto-correlation and partial auto-correlation plots [25].

C. PCA-based Monitoring and Anomaly Detection System

To recover the original data after a tampering attack, it is necessary to detect the attack and identify the tampered measurements. We carry out this by implementing an anomaly detection system based on multivariate analysis, which alerts a human supervisor when an anomaly occurs. This supervisor is in charge of discerning between "mere" measurement anomalies and actual tampering attacks. If an attack is determined, a subsequent recovery process is launched to restore the original sensor values affected by the attack.

In this line, one of the most extended applications of PCA is process monitoring and anomaly detection and diagnosis. During monitoring, Q and T^2 [26] statistics are commonly used. Q compresses the residuals in each observation and T^2 compresses the scores. With the statistics computed from the calibration data under normal conditions, control limits can be established with a certain confidence level. New data are subsequently monitored using these limits.

Q and T^2 statistics for a specific observation can be computed using the following equations:

$$T_i^2 = \sum_{a=1}^{A} \left(\frac{\tau_{ai} - \mu_a}{\sigma_a} \right)^2 \tag{4}$$

$$Q_i = \sum_{j=1}^{J} (e_{ij})^2 \tag{5}$$

where τ_{ai} represents the score of the i-th observation on the a-th PC, μ_a and σ_a stand for the mean and standard deviation of the scores of that PC in the calibration data, respectively, and e_{ij} represents the residual value corresponding to the i-th observation and the j-th variable.

The details of the usage and performance of a PCA-based anomaly detection system can be found in [10]. There, control limits for the T^2 and Q statistics are defined such that the occurrence of an anomaly is concluded when the limits are exceeded for three consecutive sampling times. Considering that the control limits are commonly chosen so that 95% of the observations gathered in the calibration stage fall below the limits, this means that the theoretical probability for false positives is $0,05^3 = 0,000125$. When an anomaly is detected, contribution plots [27] help the supervisor to elucidate the potential causes. If an attack is determined, a subsequent recovery process is launched. Although human intervention can be seen as a shortcoming of the proposed approach, the relevance of timely fire detection suggests such an intervention in a practical system. In a similar line, it is also usual to find this configuration in industrial process monitoring [27]. Although automatic supervision is also possible [10], this aspect does not constitute the core of the present work as we focus our attention in the subsequent stage of recovering tampered data.

D. Missing data Recovery Through Multivariate Statistical Methodologies

There are several methods to estimate missing data with PCA. These methods can be classified into two groups: regression and non-regression-based methods, the former ones exhibiting better performance [8]. Among the regression-based techniques, the trimmed scores regression (TSR) presents a good trade-off between simplicity and estimation performance [9].

The TSR method estimates the value of the scores from the trimmed scores, i.e., the scores obtained by filling the missing values with zeros. For data centered before PCA, this is equivalent to using the average value of a variable to give an initial estimation of its missing values.

Without loss of generality, let us assume an incomplete observation x_{inc} with available measurements on the first k variables and where the values of the remaining variables are missing. The trimmed scores of x_{inc} are calculated in PCA as follows:

$$\tau_A^* = (\mathbf{P}_{A,k}^*)^T \cdot x_{inc}^* \tag{6}$$

where

$$\mathbf{P}_{A,k}^* = \begin{bmatrix} p_{1,1} & \cdots & p_{A,1} \\ \vdots & \ddots & \vdots \\ p_{1,k} & \cdots & p_{A,k} \end{bmatrix} \tag{7}$$

$$x_{inc}^* = [x_1, \ldots, x_k]^T \tag{8}$$

and where $p_{a,j}$ is the loading corresponding to the j-th variable in the a-th PC. Only the available variables in x_{inc} and their corresponding loadings are thus used to compute the trimmed scores.

TSR makes use of the complete calibration data $\mathbf{X}$ and the trimmed scores to improve the estimation of the scores from incomplete observations. Let us call $\mathbf{X}^*$ the sub-matrix of $\mathbf{X}$ with the available variables in x_{inc}. The matrix of trimmed scores corresponding to the calibration data can be computed as follows:

$$\mathbf{T}_A^* = \mathbf{X}^* \cdot \mathbf{P}_A^* \tag{9}$$

The complete score matrix $\mathbf{T}_A$ can be regressed on the trimmed scores $\mathbf{T}_A^*$, such that

$$\mathbf{T}_A = \mathbf{T}_A^* \cdot \mathbf{B} + \mathbf{F} \tag{10}$$

where the matrix of regression coefficients $\mathbf{B}$ may be computed from least squares, as the inversion of $(\mathbf{T}_A^*)^T \cdot \mathbf{T}_A^*$ is typically nicely conditioned. If it is not, biased methods such as partial least squares (PLS) [28] [29] can be used to estimate $\mathbf{B}$. Afterwards, $v\mathbf{B}$ is used to improve the score estimation as follows:

$$\tau_A^{TSR} = (\mathbf{P}_A^* \cdot \mathbf{B})^T \cdot x_{inc}^* \tag{11}$$

Finally, the score τ_A^{TSR} can be used to estimate the incomplete observation, including its missing elements:

$$\hat{x} = \mathbf{P}_A \cdot \tau_A^{TSR} \tag{12}$$

TSR is more efficient as the inter-variable correlation in the original data set increases, since the variables with missing data for a given observation are computed from available values in the others.

IV. Data Arrangement for System Modeling

Once the fundamentals of multivariate statistical analysis are presented, this fourth section is mainly devoted to discuss how the incoming sensor data in WSNs are arranged to derive the intended PCA-based model for data recovery. This is a relevant matter since the data arrangement procedure can have a significant impact in the performance of a multivariate model [30].

In the first sub-section, we introduce the arrangement in traditional (static) PCA. Afterwards, the model is extended to the so-called DPCA through the addition of temporal lags in the sensor measurements, incorporating the system dynamics to the model.

E. *Static PCA Modeling*

The static PCA model is calibrated from the data gathered from the WSN and arranged in matrix form as shown in Fig. 1. The data corresponding to each single sensor are arranged as a column, and the data corresponding to each single sampling time are arranged as a row. Thus, the matrix of data X from which PCA is calibrated contains *J* variables, with *J* the number of sensors in the WSN, and *I* observations, with *I* the number of sampling times for each sensor. Thus, in this case, the corresponding model refers to a matrix $\mathbf{X}$ of dimension $I \times J$.

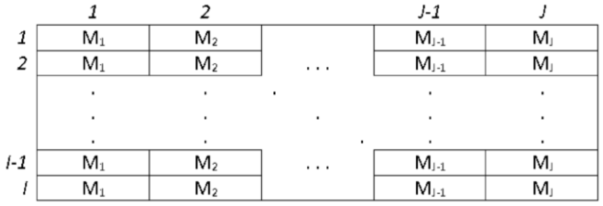


Figure 1. Traditional static PCA-based arrangement for the calibration data, conformed by I observations of J variables. M_j, with $j = 1, 2, \cdots, J$ stands for measurement of sensor j.

It is important to note that the actual location of a sensor in the data arrangement does not have any influence in terms of model calibration or data recovery performance.

F. *Dynamic PCA Modeling*

The traditional static PCA modeling does not consider the temporal inter-dependence among the system variables. However, the dynamic nature is inherent in WSNs, as the sensors are gathering information continuously. Therefore, dynamics should be considered by the model to improve the system representation.

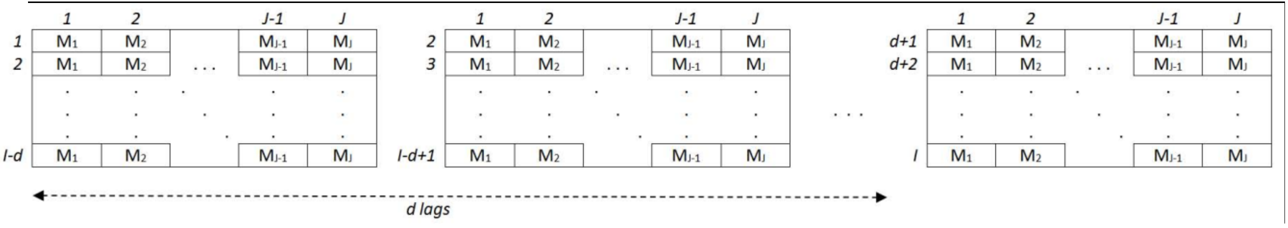


Figure 2. Dynamic PCA-based arrangement ($\mathrm{X_d}$) from the original static PCA data matrix (X) conformed by $J \times (d+1)$ variables and $I - d$ observations, J being the number of original variables, I the number of original observations and d the time lag considered. As in Fig. 1, M_j, with $j = 1, 2, \cdots, J$ stands for measurement of sensor j.

Fig. 2 depicts the data arrangement for DPCA from the original data matrix X used in the static version, according to the description provided in Section III-B. In this case, the corresponding model is fitted from matrix $\mathrm{X_d}$ whose dimension depends on the temporal lags used. Generalizing, the dimension of is $(I-d) \times J \cdot (d+1)$, with I the number of observations, J the number of the system variables and d the time lag.

V. Variable Routing Strategies for Multi-hop Transmissions

A main concern in WSN deployments is energy consumption. Low energy consumption is desired in order to reduce sensors maintenance. For this, a multi-hop routing scheme can be used so that the range of wireless communications is reduced as much as possible while the overall connectivity is maintained. Most of the existing routing algorithms are static, that is, the same routes from origin to destination are used throughout the WSN operation time, provided that the network conditions are sustained over time [31].

As previously discussed, DPCA incorporates system dynamics into the model. The auto-correlation and lagged cross-correlation information included is used to improve the recovery performance of the response scheme. However, if static routing strategies are used for re-transmissions, the measurements affected by a given tampered attack are always those corresponding to the same sensors, that is, those that forward their measurements through the tampered sensor. In this case, the auto-correlation and lagged cross-correlation information included in the model becomes useless for data imputation when the tampering affects a high number of consecutive time measurements.

If a variable routing is used instead, (e.g., random next-hop selection) the measurements of the sensors will be sent to the CU through different routes. Thus, the sensors affected by a tampering attack vary with time. As a consequence, the effect of the attack is spread and the auto-correlation information is preserved.

Three variable routing schemes will be considered in this work. They differ in the rule to decide which of the n

©2013 Engineering and Technology Publishing

closest nodes to the current one in the route towards the CU is chosen as the next hop:

- *Random routing (RR)*: random selection of the next node, so that each one of the next available n nodes has the same probability to be selected $(1/n)$.
- *Differential random routing (DRR)*: random selection among the next $n-1$ nodes which were not selected in the previous sampling time.
- *Switching‐based routing (SR)*: select the next node following a deterministic pattern intended to vary the routes in time as much as possible.

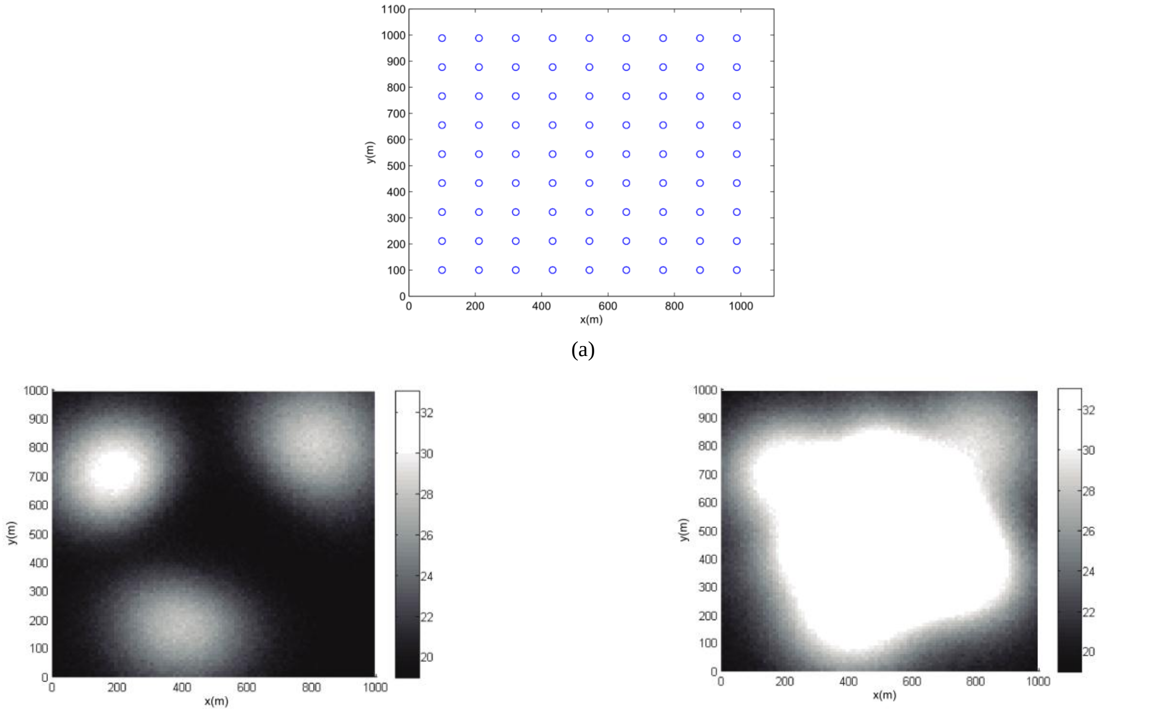
In all the discussed routing strategies the negative effect of a failure or an attack is spread over different sensors over time, which will preserve correlation information and thus the recovery results are expected to be improved.

VI. Experimental Framework

As a proof of concept, our study is focused on a WSN for firefighting in a forestry area. The main reason for this choice is the social and economical relevance of this kind of environments. In what follows, both the specific simulation scenario considered to validate our proposals and the experimental results obtained are discussed. Moreover, some further discussions about the consequences of using different routing strategies are established.

A. Simulation Scenario

A simulator based on Matlab 2009b to obtain the temperature evolution of a forestry area has been developed for experimental purposes. It is based on reference [32], where temperatures at specific locations are computed as a function of the distance to a number of temperature focuses. Both normal temperature focuses and fire focuses are modeled using a 2D gaussian distribution. Fig. 3 shows the simulation scenario, where Fig. 3(a) corresponds to the distribution of the sensors in the area. Two types of temperature maps are also presented. Derived from normal conditions, Fig. 3(b) shows three normal temperature (in $^\circ C$) sources representing the hottest areas, which may be valleys among cooler zones representing mountains. Fig. 3(c) illustrates a fire situation where the fire has a central focus and has burnt out over more than half of the total area.



(a)

(b)

(c)

Figure 3. Simulation scenario: (a) Sensor locations, (b) Temperature map under normal conditions, (c) Temperature map with a fire focus.

We assume a 1000 m×1000 m square area of forestry where 81 (9×9) sensors are regularly distributed, i.e., each sensor is located ~100 m away from its neighbors (Fig. 3(a)). The dimension of the scenario and the number of sensors are inspired by a real system provided by the Libelium company (http://www.libelium.com/wireless_sensor_networks_to_detec_forest_fires/). A regular topology is assumed for the sake of simplicity. Nevertheless, the extension to a non-regular topology is straightforward by using distance metrics among sensors [10] except for the implementation of some dynamic routing schemes, as it will be discussed afterwards.

Every sensor acquires the ambient temperature for predefined sampling times and sends the measurements towards the CU. A multi-hop routing scheme is assumed in the simulation. Nodes in the leftmost column of the topology measure the temperature and send the measurements towards the right. Nodes in the intermediate seven columns of the topology also send their measurements towards the right. Additionally, they forward the measurements collected by other sensors at their left. Finally, the nodes in the rightmost column of the topology forward all the collected information to the CU. The simplest routing scheme following the previous description is a linear (left-to-right) routing scheme, as in the MFCA routing protocol [1]. Fig. 4 shows an example of how this routing algorithm works and the effect of a tampering attack compromising a node at the rightmost column and affecting all the information routed through it.

The simulation tool is first employed to generate a data set (hereafter, CAL data set) used to calibrate the PCA models for both anomaly detection and data recovery. The CAL data matrix $\mathbf{X}$ contains $I = 100$ observations of $J = 81$ variables (the temperatures obtained by each sensor) under normal temperature conditions, i.e., without fire influence. A situation in which a fire focus evolves over time is then simulated (hereafter, FIR data set). The FIR data set is used to simulate the tampering attacks and to compute the estimation error of the recovery system. To simulate the attacks, the worst case is always considered.

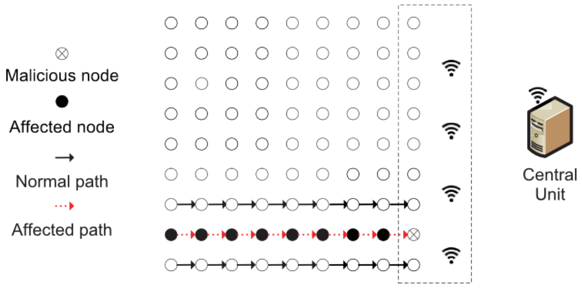


Figure 4. A malicious relay node compromises the sensing values gathered by the other sensors that are routed through the former one employing linear left-to-right static routing.

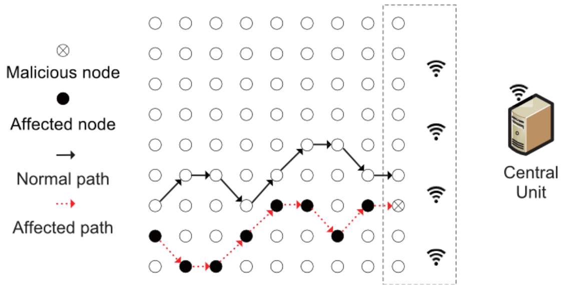


Figure 5. A malicious relay node compromises the sensing values gathered by the other sensors that are routed through the former one employing variable and probabilistic (e.g., with a random next node selection pattern) left-to-right routing.

This corresponds to the case in which one relay node in the rightmost column is tampered. The tampering of such a node affects all the measurements forwarded by the node. For illustration purposes, the nodes affected by a worst-case tampering attack in a WSN using the left-to-right static and dynamic routing schemes are showed in Fig. 4 and Fig. 5, respectively.

The variable routing strategies introduced in Section V are analyzed here for a number of nodes $n = 3$. RR and DRR strategies as well as three variants of the SR strategy, referred to as SRI, SRII and SRIII, are evaluated.

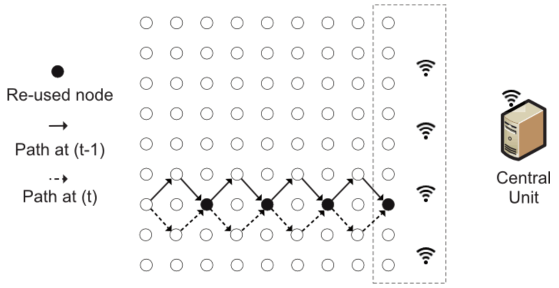


Figure 6. First switching-based routing strategy (SRI): from leftmost node to rightmost routing four nodes are re-used at consecutive sampling times $t - 1$ and t.

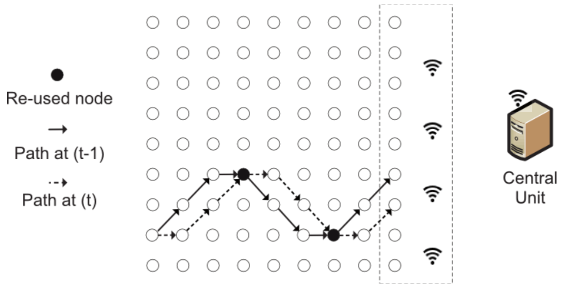


Figure 7. Second switching-based routing strategy (SRII): from leftmost node to rightmost routing only two nodes are re-used at consecutive sampling times $t - 1$ and t.

SRI defines a pattern where each row of sensors switch between forwarding towards the next upper right or the next lower right neighbor node. As shown in Fig. 6, this simple pattern does not achieve a high degree of variability in the routes, since half of the nodes in a route also coincide in the same route after switching. In SRII a more complex pattern is defined for each three rows of the network. This scheme reduces the number of node coincidences in alternative routings (see Fig. 7). The third variant of SR, SRIII, introduces an initial random route selection at the beginning for only the first sampling time. Then, each node switches continuously and independently among the remaining three alternative next nodes. It should be noted that the SRI and SRII schemes require some sort of synchronization among the nodes, while SRIII can be naturally implemented initializing the nodes in slightly different times.

B. Anomaly Detection Results

Through the comparison of the PCA model obtained from calibration to the new observations under monitoring (i.e., the test data set), anomalies in the environmental behavior can be detected. The Matlab

©2013 Engineering and Technology Publishing

PLS-toolbox [33] is used to illustrate how the monitoring process is carried out by employing the monitoring graphics such as those presented in Figs. 8 and 9. These figures depict the fire evolution (inverted triangles) in the FIR data set. Almost from the beginning of the fire, the observations surpass the established limits for normal conditions, calculated from the CAL data set (dark circles). Thus, the fire is timely detected with the monitoring system.

Fig. 8 corresponds to the case where only a fire situation is taking place. Instead, in Fig. 9 a tampering attack is performed during the fire evolution, which is intended to disrupt the operations of the fire brigade.

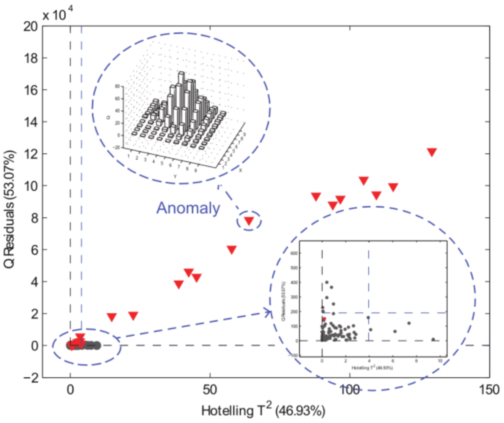


Figure 8. Monitoring graphic under fire influence: initial calibration data (dark circles) and control limits (vertical and horizontal dashed lines), from which anomalies are detected (inverted triangles). Q contribution plot detailing an anomalous observation (top left) corresponding to the fire evolution is also depicted.

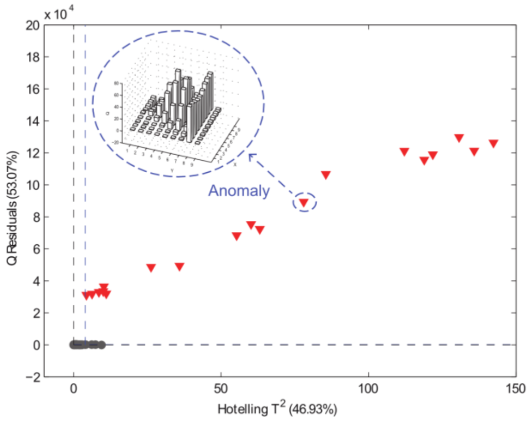
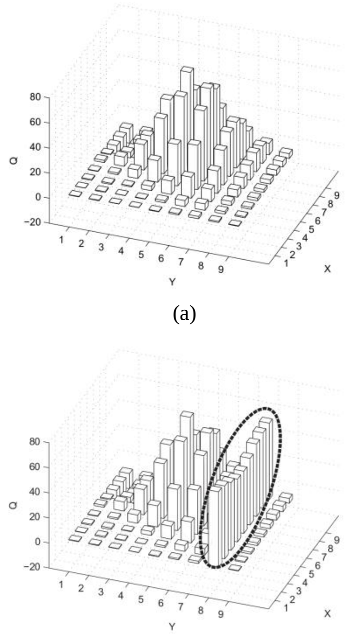


Figure 9. Monitoring graphic under fire influence and data tampering: initial calibration data (dark circles) and control limits (vertical and horizontal dashed lines), from which anomalies are detected (inverted triangles). Q contribution plot detailing an anomalous observation (top left) corresponding to the joint influence of fire and data tampering attack is also depicted.

At this point, it is important to note that the monitoring system can not distinguish between anomalies due to actual fire events and anomalies due to tampering attacks or sensor malfunctions. To discern between both situations, human intervention is recommended. Such an intervention is common for this kind of PCA-based monitoring systems [27], while it is not a limitation of our central recovery related proposal. This way, when an alarm is triggered the system computes the T^2 and/or Q contribution plot for the corresponding observation. From the visualization of the contribution plot, the human supervisor is in charge of diagnosing the potential causes for the alarm. Fig. 10(a) shows the detail of the Q contribution in one of the anomalous observations found in Fig. 8, while Fig. 10(b) shows the pattern obtained under attack for an anomalous observation in Fig. 9. In the monitoring system, tampering attacks are shown as sharp artifacts which shape depends on the routing scheme used and which are clearly different to the smooth contribution of a true fire.



(a)

(b)

Figure 10. Q contribution plots: (a) profile generated from the fire situation without any malicious presence; (b) profile generated from the fire situation and with the presence of the rightmost malicious node affecting to all the sensor measurements routed through it. The dashed circle highlights the artifact due to the attack occurred.

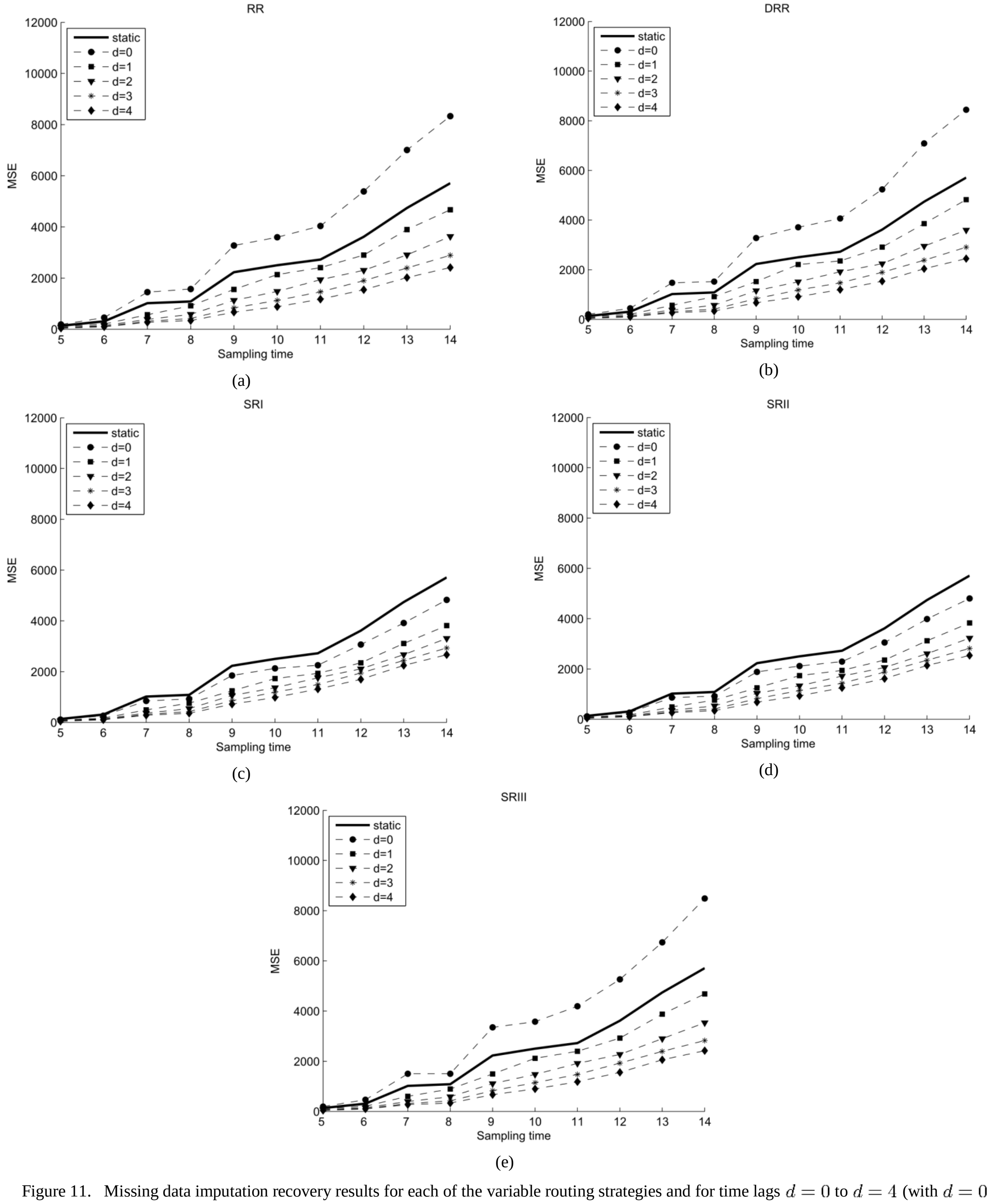
From the knowledge of the routing scheme followed in the WSN, the shape of common artifacts caused by tampering attacks can be predicted. Thus an automatic solution is possible. The authors in [10] propose a window-filter-based approach that is applied to the previous obtained contribution plots in order to perform automatic detection. This filter highlights specific artifacts. It should be noted that when unpredictable origin-destination routes are established, such as in RR and DRR strategies, the artifacts for attacks are also unknown and the detection procedure becomes a more challenging task.

Whatever the detection method used to determine the occurrence of attack or malfunction situations, either manual or automatic, a missing data recovery process is afterwards executed to solve the situation and recover the affected data. This process, which is the focus of the present paper, is evaluated below.

C. *Missing Data Recovery Results*

Once an attack alarm is generated, a response mechanism should be performed to mitigate the consequences of the threat and achieve the system survivability in terms of the continuity of the services provided. In the present work, the anomalous values detected as tampered data are treated as missing values and estimated using missing data recovery techniques.

Authors in [10] consider static and deterministic routing algorithms, including the linear routing of Fig. 4, together with PCA modeling. These schemes are used by the recovery system described in Section III-D. We try to improve these results through two means: incorporating time correlations into the model for the target system, and employing variable routing strategies as explained in Section V and VI-A.



(a)

(b)

(c)

(d)

(e)

Figure 11. Missing data imputation recovery results for each of the variable routing strategies and for time lags $d = 0$ to $d = 4$ (with $d = 0$ meaning that no time lag is considered): (a) RR, (b) DRR, (c) SRI, (d) SRII and (e) SRIII.

©2013 Engineering and Technology Publishing

To evaluate the performance of the new data recovery approach, the mean square error (MSE) of estimation of tampered data recovery is computed for 10 consecutive observations (from 5th to 14th) in the FIR data set, where the evolution of a fire is measured. Worst-case tampering attacks, to those nodes in the rightmost column of the WSN, are considered. The results are depicted in Fig. 11 for each of the routing approaches considered and from $d=0$ to $d=4$ ($d=0$ means that no time lag is used). The MSE evolution for the static routing strategy (Fig. 4) is also shown as a baseline for comparison.

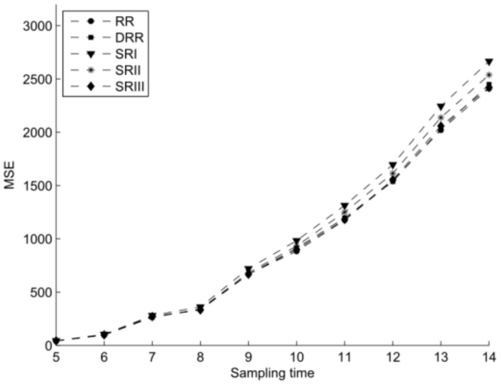


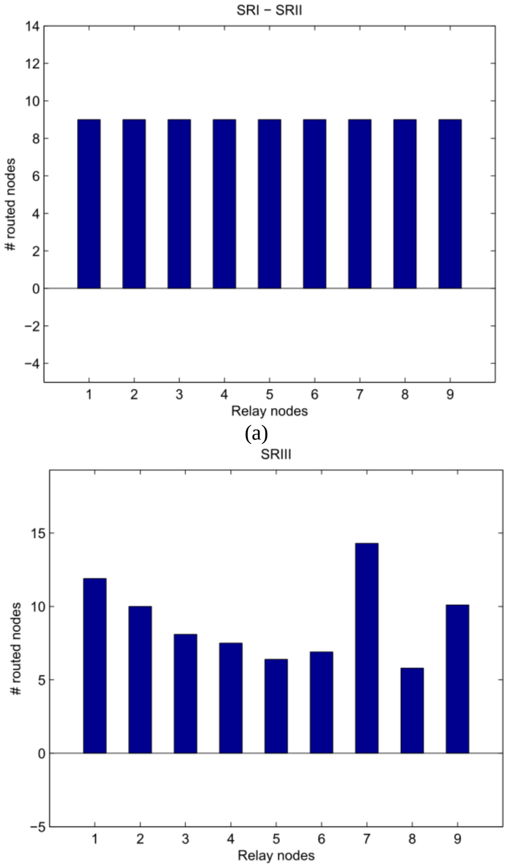
Figure 12. MSE evolution for all considered variable routing strategies and for $d=4$

We can observe that in all the cases the recovery performance increases with the time lag, since more dynamic information is captured by the model. Therefore, the combination of DPCA with variable routing strategies is effective in terms of recovery performance. The routing strategies which present some degree of randomization (RR, DRR and SRIII) have a similar behavior when the number of lags in DPCA is changed. This also happens for purely deterministic routing (SRI and SRII). Deterministic routing outperforms probabilistic routing for a low number of lags, but as the number of lags grows, the opposite is found. Since a better performance is obtained for a high number of lags, we can conclude that probabilistic routing is a better solution in terms of recovery performance. This can be explained by the fact that probabilistic methods perform a more variable distribution of the sensors in the routes. In consequence, the recovery data procedure gets more valid (non-tampered) temporal data. Fig. 12 presents the comparison of the MSE for $d=4$ and the five routing strategies. In this figure, the superiority of probabilistic methods is clear. Also, SRII outperforms SRI as a consequence of the lower amount of re-used nodes in the path for consecutive sampling times in the former: if the same node is tampered at $t-1$ and t, the temporal correlation is lost for this sensor. For this reason, the higher the variability in the routing, the better recovery performance. Table I shows numerical results of the recovery procedure for the 10th sampling time under fire influence. The same conclusions aforementioned are obtained from the table. We can see how the recovery performance is enhanced in more than 60% in comparison with the static approach.

TABLE I. MSE RESULTS FROM EACH VARIABLE ROUTING STRATEGY AT SAMPLING TIME $t=10$

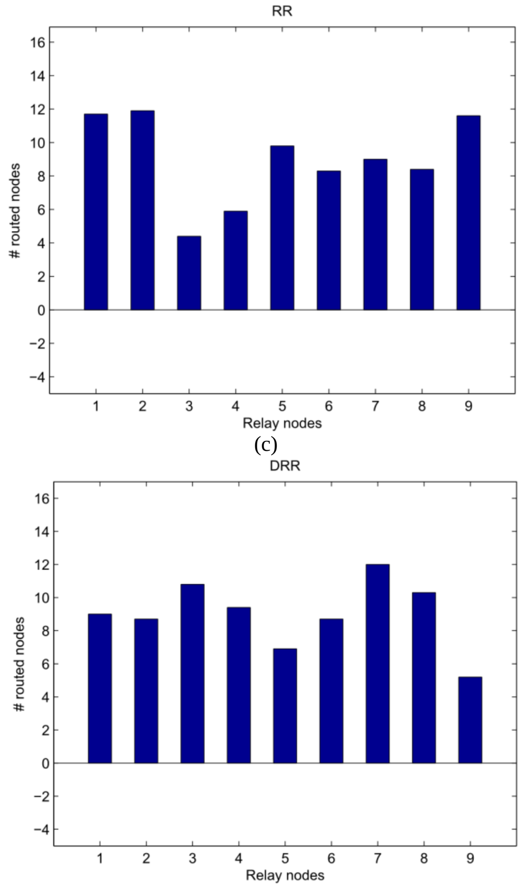
	MSE				
Algorithms	*d=0*	*d=1*	*d=2*	*d=3*	*d=4*
Static	2500	-	-	-	-
RR	3594.9	2135.7	1476.2	1127.9	883.2
DRR	3708.2	2207.7	1514.3	1180.4	912.1
SRI	2128.7	1727	1364.8	1191.7	983.2
SRII	2116.3	1735	1330.8	1143.7	933.2
SRIII	3576.8	2118	1472	1145.3	899.3

We can conclude that the joint use of DPCA modeling and variable routing strategies provides better recovery performance than static PCA modeling and static routing. Also, probability based routing outperforms the deterministic switching schemes considered. However randomness introduces a certain degree of traffic unbalancing in the relay nodes, as it is shown in Fig. 13, which is a negative effect in terms of energy consumption. It should be noted that this is corrected with time. In summary, the traffic distribution and the number of re-used nodes are highly relevant design requirements to devise variable routing algorithms for data recovery and energy consumption, which are critical aspects in WSNs.



(a)

(b)



(c)

(d)

Figure 13. Average number of sensors routed through each relay node at 10th sampling time: (a) SRI and SRII; (b) SRIII; (c) RR; (d) DRR. The traffic load balance observed in random algorithms tends to be similar to the switching-based case when averaging through longer time periods.

D. Discussion on the Consequences of using Variable Routing Strategies

Random and deterministic strategies have been assessed to yield variable routing patterns. The randomization of the route selection coupled with dynamic modeling improves the system recovery performance thanks to a better traffic distribution. This means that the number of sensors repeated in a route at different sampling times is lower. This improves the recovery performance, since one main source of information for the recovery of a lost measurement of a sensor is its previous measurement. A drawback of routing randomization is that some degree of traffic unbalance is introduced. Traffic balance has a relevant impact over the entire energy consumption, which is a critical aspect. On the other hand, an advantage of random routing is that it cannot be predicted by a malicious node/person. However, this is also a drawback since the control unit, which retrieves the WSN measurements, cannot infer the origin of a tampered or malformed packet. This complicates the application of the proposed recovery approach.

Deterministic strategies yield worse outcomes in data recovery in combination with dynamic models. They have also additional drawbacks. Non-regular network topologies complicate the design of the varying routing pattern and some degree of time synchronization of the sensors may be needed to limit the unbalanced traffic.

The most practical method considered was a deterministic switching scheme randomly initiated. This method can be easily extended to a non-regular network and deployed in practice. It shares the recovery performance of purely random approaches but with the advantage that a random generator is not needed in the nodes: random initialization is implicitly introduced for a non synchronized start up of the sensors.

VII. Conclusions and Future Work

This work presents a tolerance approach to recover missing data due to tampering attacks or failures in sensors or communications. The work relies on multivariate statistical techniques to model the target environment where the system dynamics are incorporated into the model by using dynamic principal component analysis (DPCA). Additionally, several routing strategies are studied to analyze the effect of the tampering attacks on the overall network performance. From this study we conclude the convenience of implementing variable routing strategies in which the traffic distribution has a relevant impact over the recovery performance and energy consumption, critical aspects in WSNs.

Despite the good results achieved when the proposal is deployed and evaluated in a simulation WSN scenario, some further research actions can be carried out to extend this work. More general non-regular topologies for the scenarios should be analyzed to generalize the application of the tolerance scheme introduced. Alternative routing strategies to those studied here might also be deployed and the results conveniently analyzed. Finally, and although it is out of the scope of our work, more specific and contrasted detection schemes can be incorporated to the global detection plus reaction system to improve the overall security and reliability of the proposed framework.

Acknowledgment

This work has been partially supported by Spanish MICINN through project TEC2011-22579 and by the FPU P6A grants program of the University of Granada.

References

[1] J. N. Al-Karaki and A. E. Kamal, “Routing techniques in wireless sensor networks: a survey,” *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–28, Dec 2004.

[2] J. Yick, B. Mukherjee, and D. Ghosal, “Wireless sensor network survey,” *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[3] X. Chen, K. Makki, K. Yen, and N. Pissinou, “Sensor network security: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 11, no. 2, pp. 52–73, 2009.

[4] A. D. Wood and J. A. Stankovic, “Denial of service in sensor networks,” *IEEE Computer*, vol. 35, no. 10, pp. 54–62, Oct 2002.

[5] M. Lima, A. dos Santos, and G. Pujolle, “A survey of survivability in mobile ad hoc networks,” *IEEE Communications Surveys & Tutorials*, vol. 11, pp. 66–77, 2009.

[6] I. Jolliffe. Principal Component Analysis. (2002). [Online]. Available:http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4

[7] J. E. Jackson, *A User’s Guide to Principal Components*, Wiley Series in Probability and Statistics, 2004.

[8] F. Arteaga and A. Ferrer, “Dealing with missing data in MSPC: Several methods, different interpretations, some examples,” *Journal of Chemometrics*, vol. 16, no. 8-10, pp. 408–418, Aug 2002.

[9] F. Arteaga and A. Ferrer, “Framework for regression-based missing data imputation methods in on-line MSPC,” *Journal of Chemometrics*, vol. 19, no. 8, pp. 439–447, Aug 2005.

[10] R. Mag án-Carri ón, J. Camacho, and P. Garc á-Teodoro, “Survivability in wireless sensor networks: A multivariate statistical approach for tampered data recovery,” *Submitted to ACM Transactions of Sensor Networks*, 2013.

[11] M. Xie, S. Han, B. Tian, and S. Parvin, “Anomaly detection in wireless sensor networks: A survey,” *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1302–1325, Jul. 2011.

[12] Y. Li and L. E. Parker, “A spatial-temporal imputation technique for classification with missing data in a wireless sensor network,” in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Sep 2008, pp. 3272–3279.

[13] L. Gruenwald, M. S. Sadik, R. Shukla, and H. Yang, “DEMS: A data mining based technique to handle missing data in mobile sensor network applications,” in *Proc. Seventh International Workshop on Data Management for Sensor Networks*, New York, NY, USA: CM, 2010, pp. 26–32.

[14] J. C. Lim and C. J. Bleakley, “Robust data collection and lifetime improvement in wireless sensor networks through data imputation,” in *Proc. Fifth International Conference on Systems and Networks Communications*, Aug 2010, pp. 64–69.

[15] D. Guo, X. Qu, L. Huang, and Y. Yao, “Sparsity-based spatial interpolation in wireless sensor networks,” *Sensors*, vol. 11, no. 3, pp. 2385–2407, Feb 2011.

[16] D. Guo, Z. Liu, X. Qu, L. Huang, Y. Yao, and M. T. Sun, “Sparsity-based online missing data recovery using overcomplete dictionary,” *Sensors Journal*, *IEEE*, vol. 12, no. 7, pp. 2485–2495, July 2012.

[17] E. W. Dereszynski and T. G. Dietterich, “Spatiotemporal models for data-Anomaly detection in dynamic environmental monitoring campaigns,” *ACM Trans. Sen. Netw.*, vol. 8, no. 1, pp. 3:1–3:36, 2011.

[18] M. A. Livani and M. Abadi, “A PCA-based distributed approach for intrusion detection in wireless sensor networks,” in *Proc. International Symposium on Computer Networks and Distributed Systems IEEE*, Feb 2011, pp. 55–60.

[19] N. Chitradevi, K. Baskaran, V. Palanisamy, and D. Aswini, “Designing an efficient PCA based data model for wireless sensor networks,” in *Proc. 1st International Conference on Wireless Technologies for Humanitarian Relief*, New York, NY, USA: ACM, 2011, pp. 147–154.

[20] K. H. Esbensen, *Multivariate Data Analysis - in Practice*, Esberg, Aalborg University, Ed. CAMO, 2009.

[21] S. Wold, “Cross-validatory estimation of the number of components in factor and principal components models,” *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978.

[22] J. Camacho and A. Ferrer, “Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Theoretical aspects,” *Journal of Chemometrics*, vol. 26, no. 7, pp. 361–373, 2012.

[23] W. Ku, R. H. Storer, and C. Georgakis, “Analysis,” *Chemometrics and Intelligent Laboratory Systems*, *Disturbance Detection and Isolation by Dynamic Principal Component*, vol. 30, no. 1, pp. 179–196, Nov 1995.

[24] J. Chen and K.-C. Liu, “On-line batch process monitoring using dynamic PCA and dynamic PLS models,” *Chemical Engineering Science*, vol. 57, no. 1, pp. 63–75, 2002.

[25] A. Hatemi-J, “Multivariate tests for autocorrelation in the stable and unstable VAR models,” *Economic Modeling*, vol. 21, no. 4, pp. 661–683, July 2004.

[26] H. Hotelling, *Multivariate Quality Control. Techniques of Statistical Analysis*, New York, Ed., MacGraw-Hill, 1947.

[27] T. Kourti and J. MacGregor, “Multivariate spc methods for process and product monitoring,” *Journal of Quality Technology*, vol. 28, 1996.

[28] P. Geladi and B. R. Kowalski, “Partial least-squares regression: A tutorial,” *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[29] S. Wold, M. Sjstrm, and L. Eriksson, “PLS-regression: A basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct 2001.

[30] J. Camacho, J. Picó, and A. Ferrer, “Bilinear modelling of batch processes. Part I: theoretical discussion,” *Journal of Chemometrics*, vol. 22, no. 5, pp. 299–308, May 2008.

[31] P. Padilla, J. Camacho, G. Maci á-Fern ández, J. D ázVerdejo, P. Garc á-Teodoro, and C. G ómez-Calero, “On the influence of the propagation channel in the performance of energy-efficient geographic routing algorithms for wireless sensor networks (wsn),” *Wireless Personal Communications*, pp. 1–24, 2012.

[32] G. X. E. S. Manolakos and D. V. Manatakis, “Temperature field modeling and simulation of wireless sensor network behavior during a spreading wildfire,” in *Proc. 16th European Signal Processing Conference*, Lausanne, Switzerland: EURASIP, 2008.

[33] B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, and R. S. Koch. (2005). PLSToolbox 3.5 for use with Matlab. [Online]. Available: http://www.eigenvector.com/software/pls_toolbox.htm

Magán-Carrión, Roberto is a Ph.D. student at the Department of Signal Theory, Telematics and Communications of the University of Granada (Spain) and member of the research group “Network Engineering and Security Group (NESG)”. He received his MSc degree in Telecommunications from the University of Málaga in 2008. His research interests are focused on Mobile Ad hoc NETworks (MANETs) security and more specifically on response (IR, Intrusion Response) and tolerant (IT, Intrusion Tolerant) solutions as part of global cross-layer security challenge with the aim of achieving survivable systems.

Pulido-Pulido, Fernando received his MSc degree in Telecommunications from the University of Granada in 2013. Part of the present work contributed to develop his Master Thesis project.

Camacho, Jos é is Associate Professor at the Department of Signal Theory, Telematics and Communications of the University of Granada (Spain) and member of the research group “Network Engineering and Security Group (NESG)”. He holds a degree in Computer Science from the University of Granada (2003) and a Ph.D. in Control Systems and Industrial Computing from the Technical University of Valencia (2007). His Ph.D. was awarded with the second Rosina Ribalta Prize to the best Ph.D. projects in the field of

©2013 Engineering and Technology Publishing

Information and Communication Technologies (ICT) from the EPSON Foundation, and with the D.L. Massart Award in Chemometrics from the Belgian Chemometrics Society. His research interests include exploratory data analysis, monitoring, control and optimization with multivariate data analysis techniques.

Garc ı́a-Teodoro, Pedro received the B.Sc. degree in physics (electronics specialty) from the University of Granada, Spain, in 1989. In 1989, he received a grant from “Fujitsu Spain”, and during 1990, a grant from “IBM Spain”. From 1989 to 2011, he was Associate Professor and, since 2011, Full Professor and Director of the Department of Signal Theory, Telematics and Communications, University of Granada, and head of the research group “Network Engineering and Security Group (NESG)” of this University. His initial research interest was concerned with speech technologies, in which he developed his Ph.D. thesis in 1996. Since then, his professional interests have been in the field of computer and network security, especially focused on intrusion detection and denial of service attacks.

©2013 Engineering and Technology Publishing