IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON AMBIENT INTELLIGENCE ENVIRONMENTS WITH WIRELESS SENSOR NETWORKS FROM THE POINT OF VIEW OF BIG DATA AND SMART & SUSTAINABLE CITIES

Received May 3, 2018, accepted June 12, 2018, date of publication June 15, 2018, date of current version July 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2848100

Stochastic and Information Theory Techniques to Reduce Large Datasets and Detect Cyberattacks in Ambient Intelligence Environments

BORJA BORDEL[1], RAMÓN ALCARRIA[2], TOMÁS ROBLES[1], AND ÁLVARO SÁNCHEZ-PICOT[1]

[1]Department of Telematic Systems Engineering, Universidad Politécnica de Madrid, 28040 Madrid, Spain
[2]Department of Geospatial Information, Universidad Politécnica de Madrid, 28040 Madrid, Spain

Corresponding author: Borja Bordel (bbordel@dit.upm.es)

This work was supported in part by the Ministry of Economy and Competitiveness through SEMOLA Project under Grant TEC2015-68284-R and in part by the Autonomous Region of Madrid through MOSI-AGIL-CM Project under Grant P2013/ICE-3019, co-funded by EU Structural Funds FSE and FEDER. The work of B. Bordel was supported by the Ministry of Education through the FPU Program under Grant FPU15/03977.

ABSTRACT Ambient intelligence refers a new technological paradigm, where everyday environments behave in a smart way and are sensitive to their inhabitants. In order to reach this objective, complex pervasive sensing platforms are deployed, together with artificial intelligence solutions. In these new, complex, and highly interdependent systems, traditional security policies and defense strategies are not effective, as thousands of heterogeneous cyber and physical elements are mixed and connected. New security solutions try to learn about the expected behavior from the system and its components, so if a strange event occurs; adequate preventive, corrective, and/or reactive security actions to detect and stop the potential cyber-physical attack being performed are triggered in an intelligent way. In order to learn about the system and select and apply the adequate security actions, very large datasets containing records of previous behaviors should be analyzed, sometimes in a very fast way. This fact enormously complicates the implementation of these new security solutions, as it is necessary a huge storage capacity, which many domestic systems do not have, and it is needed to work with huge data sets whose processing time prevents making decisions with the required speed. Therefore, in this paper, we investigate and propose a procedure to reduce large datasets, with the objective of enabling new security techniques to detect cyberattacks in a fast and efficient way. The proposed procedure is based on the calculation of small sets of samples, whose statistic configuration is as similar as desired to the original large dataset. Stochastic models and information theory techniques and theorems are composed and combined in order to define a mathematical framework which allows the obtention of these equivalent reduced datasets. We also describe and evaluate a first implementation of the proposed solution, using both, a simulation scenario and a real deployment.

INDEX TERMS Ambient intelligence, security, big data, cybersecurity, stochastic models, information theory.

I. INTRODUCTION

Future engineered systems are envisioned to be composed of ubiquitous deployments including thousands of hardware and software components, very heterogeneous and managed in an unmanned and non-centralized way [1].

Various solutions based on this new paradigm have been proposed during the last ten years: Cyber-Physical Systems (CPS) [2], Industry 4.0 [3], Smart environments [4], etc. Although the final applications that can support all these technologies are different, all of them are based on a pervasive sensing platform, which enables final applications to infer some relevant information that is implicit in the acquired data [2].

One of the most important and popular proposals in this area is Ambient Intelligence (AmI) [5]. AmI refers to a new technological paradigm where everyday environments behave in a smart way and are sensitive to their inhabitants. In AmI solutions, then, sensors and other processing devices are employed as implicit interfaces to interact with people and help them in their daily life.

In fact, the idea of using technology to enhance people experience and help them in their daily living activities is not

new (some first references to the concept of ''smart house'' can be found even in works of 1960). However, the current technological state enables, for the first time, considering AmI as a reality and as a discipline with a unique set of contributions. In particular, current information technologies can fulfill the requirements and essential characteristics of AmI systems: sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent.

Although all these characteristics are equally important, the aspect that affects the most the entire system configuration and architecture is transparence. In AmI system, hardware devices and other technologies tend to disappear [6]; so engineered deployments are indistinguishable from standard environments and daily living scenarios. This requirement practically forces the use of resource-constrained devices, as they are, in general, smaller in size, and allow the use of tiny batteries (two essential facts to create transparent technologies and unobtrusive embedded devices).

These resource-constrained devices are constantly acquiring information through specific sensors, so the AmI system may be considered sensitive. Besides, as AmI deployments must be ubiquitous, hundreds, thousand or even hundreds of thousands of tiny resource-constrained devices are continuously generating data, which are usually sent to the intelligent subsystems or to the equivalent module (see Figure 1).

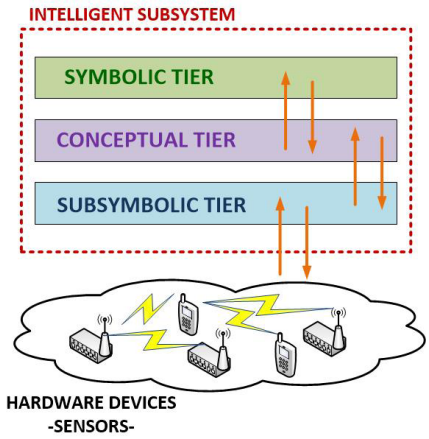

FIGURE 1. Basic architecture for an AmI environment.

Although the use of tiny heterogeneous resource-constrained sensor nodes has enabled the creation of viable AmI systems; it has introduced new challenges to be addressed. In particular, it is known that new engineered systems (such as Internet of Things deployments) are characterized for being unsecure nowadays [7]. This fact, even though it is transversal to all new technological systems, is especially critical in AmI solutions. Actually, as security solutions are supported by tiny sensor nodes, they cannot be implemented in hardware devices, or at network level. This weakness might be exploited to attack AmI systems using a new type of cyberattacks: the cyber-physical attacks [8].

In these attacks, changes (accidental or not) in hardware or software may appear, but due to the highly interdependency of components, the effects may influence in any other part of the system, and thousands of components could be the final objective of the attack. The genuine approach of cyber-physical attacks is acting on the weakest elements in the system (in this case the sensor nodes) as these elements can cause a fail in the critical components. Using this philosophy, for example, intelligent components could be forced to make a fake decision which might be fatal for the habitants of the environment where the AmI system is deployed.

Traditionally, security policies are preventive but due to their amplitude (there is an infinite amount of ways to perform a successful cyber-physical attack) and the inevitable weaknesses of AmI systems (devices are deployed in a public space, etc.), protection techniques against cyber-physical attacks are reactive.

As in industrial scenarios, where very complex and heterogeneous systems are controlled by supervisory control modules which detect anomalous behaviors, in future AmI systems intelligent components which supervise the global system evolution and provide security to the entire deployment should be included.

The described intelligent components, considering past information about the system behavior, will detect anomalous phenomena in the AmI deployment and will compare the obtained observations with known patterns. Thus, a decision about if a cyber-physical attack is running, what type of attack is being performed, and the most adequate actions to isolate the effects of the attack, will be made.

First works about how to develop this process have been reported [7]. However, a second challenge is still pending. In systems were hundreds of thousands of sensors and devices are continuously generating information flows, security components should study very large datasets to understand and learn about the real system behavior. Processing time, then, will increase exponentially.

Information, on the other hand, is seldom stored, as it is employed at real-time to make a decision and then it is immediately removed. If storage solutions to maintain very large datasets were included in AmI systems, important characteristics (as for example their transparence) would be affected.

Although standard Big Data techniques seem the perfect solution, as storing all past events and information is not always guaranteed, the application of this technique is not always possible. In conclusion, a new solution to reduce large data sets and detect cyberattacks in Ambient Intelligence Environments is required.

Therefore, in this paper we investigate and propose a procedure to reduce large datasets, with the objective of enabling new security techniques to detect cyberattacks in a fast and efficient way in AmI deployments. The proposed procedure is based on the calculation of small sets of samples, which will be easy to update, evaluate and maintain; and whose statistic configuration is as similar as desired to the original large

dataset. Stochastic models and information theory techniques and theorems are composed and combined in order to define a mathematical framework which allows the obtention of these equivalent reduced datasets.

The reminder of this paper is organized as follows. Section II describes the state of the art on security techniques and dataset reduction technologies for AmI applications. Section III presents the future security techniques for AmI systems, and the proposed dataset reduction technique. Section IV describes the experimental validation carried out. Finally, Section V presents the obtained results and Section VI concludes this work.

II. STATE OF THE ART

In the last ten years, various original security solutions for AmI and other similar systems (such as CPS) have been reported.

In general, four different fields may be distinguished when talking about security in AmI systems [17]: physical security, intrusion tolerance, active protection and IT (information technologies) security (see Figure 2).

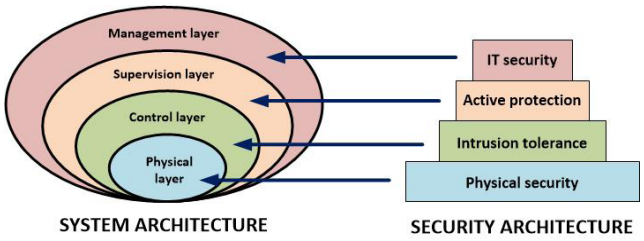

FIGURE 2. Research areas in security for AmI environments.

Physical security is typical from industrial systems, where critical components are physically isolated and protected from shocks, chemical damage, etc. These techniques, as said, are not usually employed in AmI scenarios as sensors are deployed in public spaces, and system architectures in AmI deployments are not as hierarchical as in industrial solutions. However, some sparse proposals on this topic, based on the definition of safety instrumented systems -SIS- (modules which sense the hardware components to detect ''physical aggressions'' to them) have been reported [9].

Intrusion tolerance is the most important research area in security for AmI solutions. In this context, intrusions are accepted as inevitable events, so technologies to guarantee the system continues its normal operation even if a cyberattack is running are investigated. Most of these works are focused on the design of enhanced control loops [10], [11], but cyber-attack taxonomies have been also described [12], [13]. Proposals about working schemes for CPS [14] and other systems under cyber-physical attacks [15] have been also described. The problem of all these proposals is that they are focused on attacks which introduce perturbations or known malicious signals in the system, so solutions are fixed and rigid. In consequence, these proposals reduce their usability if new or slightly different attacks are performed.

Finally, the concept of cyber-physical attack has been investigated. Abstract taxonomies and description languages have been proposed [16]. These instruments are very useful to classify and infer the use of certain types of cyber-physical attacks.

Active protection policies are those technologies that modify the basic behavior of systems to inject controls and evaluation points that protect and avoid cyberattacks to the system. For example, typical authentication solutions based on a user ID and a password are active protection techniques. In general, works about active protection solutions for AmI environments are sparse, and focused on access control: i.e. on how regulating what a user/device can do and what the programs are allowed to execute on behalf of the user/device [18], [19].

IT security is focused on traditional indicators: integrity, availability, confidentiality and fingerprinting among other parameters [25]. In fact, different works about how traditional security solutions (i.e. firewalls, computer shields, etc.) could be applied to AmI scenarios have been recently described [20]. However, this approach only partially covers the problems and vulnerabilities associated with cyber-physical attacks, so more general solutions are required. Actually, new ideas associated to security in AmI solutions such as veracity, plausibility, witnesses and physics [25] cannot be covered using these traditional instruments. Furthermore, general reviews about the problems associated with security in AmI scenarios have been also reported. Works about critical problems (such as new social or mathematical attacks [24]) in several relevant scenarios [21]–[23] may be found.

Enhanced intrusion tolerance techniques and the inclusion of new ideas, such as plausibility, will necessarily require the use of intelligent solutions. These technologies will learn about the system behavior using information about past events, and using the acquired knowledge will ensure the behavior of the system within plausible limits, even when an attack or intrusion has occurred.

In order to perform these learning processes very large datasets including information about past events should be stored, maintained and processed. As this is a very inefficient approach, techniques to reduce large datasets and calculate some equivalent smaller and more efficient sets have been investigated.

In general, techniques to reduce datasets are based on pattern recognition solutions. Traditional proposals were based on the exploration of the complete dataset (after organizing it as a tree, for example), so redundant or useless segments could be pruned [26], [27]. Although this approach was valid twenty years ago, nowadays, the size of datasets makes difficult their exploration in a limited amount of time.

More recent proposals, on the other hand, employ pattern recognition technologies to, as in previous works, remove redundant information [28]. These solutions are still pretty inefficient, so most recent contributions try to calculate statistical models to synthesize the entire dataset in much

more manageable structures. Gaussian models are, probably, the most popular [29].

Our proposal follows the most novel approach employing stochastics models to reduce large datasets in AmI environments. In order to enhance the efficiency of these basic techniques, information theory concepts are integrated in the calculation process.

III. REDUCED DATASETS FOR CYBERATTACK DETECTION

In this Section the future security policies for AmI systems are slightly introduced and described (first subsection). After the need of reducing large datasets in this future scenario is explained, the proposed new technology (including the stochastics models and the employed theorems of information theory) is described with details.

A. INTELLIGENT SECURITY SOLUTIONS FOR AmI DEPLOYMENTS

A cyber-physical attack may be described using only six different fields [16]. Figure 3 presents the basic structure of a cyber-physical attack.

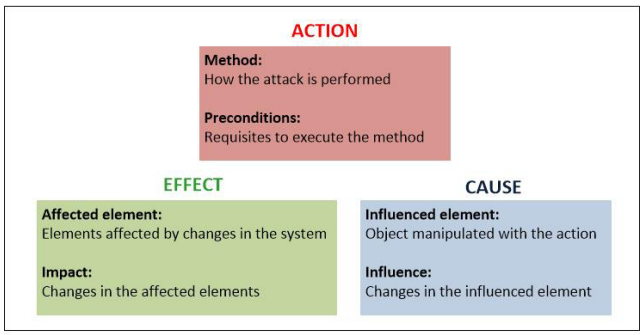

FIGURE 3. Structure of a cyber-physical attack.

Below we provide a short description of each one of the six named fields:

- Method: It represents the procedure employed to affect the system.
- Preconditions: They list the requirements of the system so that the method can be effective. Together with the ''method'' this list made up the ''action'' of the cyber-physical attack.
- Influenced element: It refers the elements which have been manipulated through the described action.
- Influence: The produced changes in the influenced element. Together with the ''influenced element'', it describes the ''cause'' of the cyber-physical attack.
- Affected element: A list of the elements being affected by the changes in the system. Usually they are the objective of the attack.
- Impact: A description of the changes in the system. Together with the ''affected element'', it describes the ''effect'' of the cyber-physical attack.

```
<Action>
   <Method>
      <Category>
           //Physical, cyber or hybrid
      </Category>
      <Description>
         //Free text
      </Description>
    </Method>
    <Precondition>
      <Category> … </Category>
      <Description>
           …
      </Description>
    </Precondition >
        …
</Action>
<Cause>
     <InfluencedElement>
        <Category> … </Category>
        <Name> … </Name>
     </InfluencedElement>
     <Influence>
        …
     </Influence>
</Cause
<Effect>
     <AffectedElement>
        …
     </AffectedElement>
     <Impact>
        …
     </Impact>
</Effect>
```

FIGURE 4. Example of a generic XML description of a cyber-physical attack using the CP-ADL description language.

This systematic manner of describing cyber-physical attacks enables the use of security techniques based on artificial intelligence technologies. Using the presented description method, cyber-physical attacks that may suffer AmI deployments can be modeled with XML documents, employing (for example) the CP-ADL (Cyber-Physical Attack Description Language) description language (see Figure 4) [16]. Thus, these documents could be used as patterns to be discovered in the system behavior. If one of these behavior patterns is recognized, then it may be deducted that a cyber-physical attack is being performed.

Both, the pattern construction process and the pattern recognition process at real-time during the system operation should be based on intelligent solutions which learn from the information about the past events in the AmI deployment.

These learning technologies, in combination with some techniques to support the defense and protection strategy (such as the game theory), have been proved to be successful as security solutions for AmI deployments.

Contrary to traditional security solutions (such as ciphers, certificates, etc.) which show important scalability, synchronization and other similar problems when implemented in systems as complex, ubiquitous and heterogeneous as AmI environments [30]; the described learning and pattern recognition techniques match perfectly the requirements and characteristics of AmI deployments.

In this work we are not describing in detail any intelligent security solution for AmI environments, as it is

not the objective of this paper. However, it is possible to find several examples of these systems in the most recent state-of-the-art [6].

Any case, as intelligent systems, these new security solutions must learn about AmI deployments using datasets containing information about past events. These datasets use to be very large, as they must contain information about all devices in the AmI environment. Typically, datasets contain all data generated in the system for a long time. Thus, intelligent security systems should manage and store large amounts of information, which is very costly and prevents to operate at real-time (a basic requirement for these security solutions).

Therefore, techniques to reduce these datasets without errors or information loss; or to calculate smaller equivalent datasets with a controlled level of "deformation" are required to improve the efficiency of security solutions for AmI environments.

B. STOCHASTICS MODELS TO REDUCE DATASETS

To train an intelligent security solution for AmI environments, in general, it is employed a non-structured dataset $\mathbb{D}$. The size (cardinality) of this dataset grows up as time passes (1), as it is formed by accumulating the historical data generated by all devices and components in the AmI environment.

$$card\left\{\mathbb{D}\right\}_{t\to\infty} \to \infty \tag{1}$$

Among the different types of non-structured datasets we can find, $\mathbb{D}$ might be considered half-structured. In fact, within $\mathbb{D}$ it is possible to identify subsets Λ_i with a uniform structure (2). The index i takes values from a set of indexes I, employed as identifiers (they may be numbers, n-tuples, etc.)

$$P = \{\Lambda_i, : i \in I\} \tag{2}$$

When dividing $\mathbb{D}$ in a set P of sub-datasets Λ_i (disjoint and non-empty) a partition of $\mathbb{D}$ is constructed, which (thus) verifies a series of mathematical properties (3).

$$\begin{aligned} &\forall i \in I, \Lambda_i \subseteq \mathbb{D} \quad and \quad \Lambda_i \neq \emptyset \\ &\forall i,j \in I, \quad i \neq j, \ \Lambda_i \cap \Lambda_j \neq \emptyset \\ &\bigcup_{i\in I} \Lambda_i = \mathbb{D} \end{aligned} \tag{3}$$

To correctly identify and construct the sub-datasets Λ_i, it is possible (and advisable) to follow a systematic procedure.

In this procedure, and depending on the application, a set of classification (and relevant) variables C must be created as first step. Each one of these classifiers must refer available meta-information about data. For example, in the most typical case, the device type that generated the datum and its location identify subsets with a homogenous structure.

It can be seen that, as more classifiers (or classification variables) are defined (i.e. as the cardinality of C goes up) a higher number of structured sub-datasets Λ_i with a smaller sizer will made up the partition P. Thus, P depends on the set of classifiers C (4).

$$P = P\left(C\right) = P_C \tag{4}$$

In general, we will say that a partition P_α is a refinement of a partition P_β of $\mathbb{D}$, if each element Λ_i^α of P_α is a subset of some element Λ_i^β of P_β. It is also said that P_α is finer than P_β, or that P_β is coarser than P_α. In practice, this concept implies that P_α is constructed fragmenting even more the partition P_β. In our context, that means the set of classifiers α has more elements (it includes more classification variables) than β (5).

$$card\left\{\alpha\right\} \geq card\left\{\beta\right\} \Longrightarrow P_\alpha \leq P_\beta \tag{5}$$

The selection of the set of classification variables is very important, because as the more variables are considered, the fragmentation level of the original dataset $\mathbb{D}$ increases (the generated partition P_C is finer); and subsets Λ_i are more strongly structured. Sub-datasets with a stronger structure are easier to process, but (on the other hand) finer partitions are composed by more elements (sets), so learning algorithms must analyze more datasets. For every system, then, there is an equilibrium point between the structuration level of sub-datasets and the number of subsets generated; which enables processing the historical dataset in the minimum time.

In general, the number of possible partitions to be defined for a certain set is calculated using the Bell number B. In our particular case, this number may be computed using a recursive expression (6).

$$B_{\mathbb{D}} = \sum_{k=0}^{card\{\mathbb{D}\}-1} \binom{card\left\{\mathbb{D}\right\}-1}{k} B_k \quad \text{being } B_0 = B_1 = 1 \tag{6}$$

Once the set of classification variables C has been defined in the first step (taking into account all previous considerations), in the second step it must be decided the range of values V_i that each of the variables c_i can take (7).

$$\forall c_i \in C \exists V_i = \left\{v_i^j, j = 1, \ldots, N_i\right\} \vdots c_i = v_i^j \tag{7}$$

Most common classification variables are physical parameters (e.g. the geographical location of sensors), and, therefore, they are continuous variables. This is not acceptable to create efficient processing application, so these variables must be quantified and discretized in order to define the ranges V_i where classification variables take values. Each variable c_i may take N_i different values.

At this point, it is possible to define a mathematical vector relation G, called "chopping function" that generates all subsets Λ_i in a systematic manner (8).

$$\begin{aligned} &G\left(\mathbb{D}, C, V_1, \ldots, V_i, \ldots, V_N\right) \\ &\quad = \begin{bmatrix} g_1\left(\mathbb{D}, C, V_1, \ldots, V_i, \ldots, V_N\right) \\ \cdots \\ g_S\left(\mathbb{D}, C, V_1, \ldots, V_i, \ldots, V_N\right) \end{bmatrix} = P_C \\ &\quad g_i\left(\mathbb{D}, C, V_1, \ldots, V_i, \ldots, V_N\right) = \Lambda_i \end{aligned} \tag{8}$$

The classification problem is said to be well-posed if for the selected classification variables and their ranges of values, the chopping function generates a completed partition of $\mathbb{D}$ (9).

$$\begin{aligned}&\forall d \in \mathbb{D} \exists \Lambda_i \in P_C : d \in \Lambda_i \\ &\quad being\ (v_1, \ldots, v_i, \ldots, v_N)\ metainformation\ of\ d \\ &\quad and\ metainformation\ of \quad \forall \lambda_i \in \Lambda_i\end{aligned} \tag{9}$$

Basically, for each N-tuple of values it must be obtained a sub-set Λ_i; so, there is a relation between the values N_i and the number of elements S in the partition P_C (10).

$$S = \prod_{k=1}^{N} N_k \tag{10}$$

Once the second step is completed, in the third step, if the number of structured sub-datasets that we have constructed is too large, it may be reduced by merging elements to obtain a coarser partition (11).

$$\Lambda_i^* = \bigcup_{k=i,j,\ldots} \Lambda_k \tag{11}$$

When obtained the partition P_C of $\mathbb{D}$ with all the desired structured sub-datasets Λ_i, we can work with each subset individually, taking advantage of its structure. Hereinafter we continue the analysis focusing on a unique sub-dataset: to process the entire dataset it is only necessary to repeat the operation for every element in the partition P_C.

Reducing a dataset Λ_i is finding a second dataset Λ_i^R with a smaller number of elements (i.e. with a smaller cardinality) whose ''distance'' to the original set is the minimum (12).

$$\Lambda_i^R = \min_{\tilde{\Lambda}} \mathcal{J}\left(\Lambda_i, \tilde{\Lambda}\right) \tag{12}$$

The distance function $\mathcal{J}(\cdot,\cdot)$ represents any metric to be used for the calculation of the reduced dataset. The distance function may be defined in time, geometrically, or considering statistical elements among other possibilities. The distance definition is very important as it allows maintaining the most relevant characteristics of the original dataset for a certain given application. In that way, if the distance function is not correctly selected for our application, the reduced dataset will not adequately represent the original one even if calculations are correct.

On the other hand, the optimization problem (12) defined by the metric $\mathcal{J}$ is solved in a different manner depending on the type of considered distance. Therefore, the first step is to select an adequate distance function.

As can be seen, besides, the optimization problem presents a degree of freedom: the cardinality (size) M of the reduced dataset (13). In general, as M goes up, better approximations (reduced sets) may be obtained; however, sets are larger and more time is required to process them. In this context, M is also called ''order of the reduced set''.

$$M = card\left\{\Lambda_i^R\right\} \tag{13}$$

In our application, as datasets are created to train intelligent security systems, the most important aspect to be preserved is the statistical distribution of values (as the final objective is to detect anomalous behaviors).

In this context, we can imagine the data generated by sensors in the AmI environment follows a certain and unknown model $\mathcal{M}$ (14) [33]. This model, for example, may represent the physical laws that control the evolution of the sensed variables [34]. This model considers two parameters: the state variables $\vec{s}$ representing the current state of the system; and the control variables $\vec{\eta}$ representing the initial and/or spatial configuration of the environment. Sets U, V and Z are appropriate vector spaces. As model $\mathcal{M}$ represents a physical system, then, sets U, V and Z are defined on the field of real numbers.

$$\begin{aligned}&\mathcal{M}: U \times W \to Z \\ &\mathcal{M}(\vec{s}, \vec{\eta}) = 0 \vec{s} \in U \subseteq \mathbb{R}^{p_u} \vec{\eta} \in W \subseteq \mathbb{R}^{p_w}\end{aligned} \tag{14}$$

However, this deterministic model does not correctly represent reality, as there are some uncertainties $\vec{\theta}$ in all real systems that must be also considered (15). $\vec{\theta}$ represents the total addition of uncertainties caused by all randomness sources: geometry, initial conditions, approximations in the model, etc [33].

$$\mathcal{M}\left(\vec{s}, \vec{\eta}; \vec{\theta}\right) = 0 \quad \vec{\theta} \in U \subseteq \mathbb{R}^{p_u} \tag{15}$$

Then, we could encapsulate all parameters in relation to the system state $\vec{x} = \left[\vec{s}, \vec{\theta}\right]$. As uncertainties are not deterministic (they are random components), the new solution of model $\mathcal{M}$ is not a vector, but a random variable.

At this point we consider a probability space $(\Omega, \mathcal{F}, \wp)$ that represents the sensor output, and the metric space (T, Σ) with the distance Σ. Being Ω the sample space, $F \subseteq 2^{\Omega}$ the event space and $P : \mathcal{F} \to [0, 1]$ the probability measure. Then, a measurable function $X : \Omega \to T$ is a random function with domain Ω and range T if it verifies that $X^{-1}(B(T)) \subset \mathcal{F}$, being $\mathcal{B}(T)$ the Borel σ-field generated by the open sets in T under the metric Σ.

For our application, where sensors of AmI environments produce measures represented by real numbers, we must consider that $T = \mathbb{R}$ and $\mathcal{B}(T) = \mathcal{B}(\mathbb{R})$, the Borel σ-algebra.

With these hypotheses, set Λ_i can be understood as a historical result record of the random experiment represented by the probability space $(\Omega, \mathcal{F}, \wp)$. Then, random function X may be reconstructed calculating its statistical parameters (moments μ, marginal distributions $F(\cdot)$ and correlation matrix r) from values in Λ_i. Although Λ_i only contains particular realizations of the random experiment, the probability theory guarantees that for an enough great number of results, the statistical parameters of the set converge to the real ones (16).

$$\begin{aligned}\mu(q) &= E\left[X^q\right] \\ r &= E[X \cdot X] \\ F(x_i) &= P(X \leq x_i)\end{aligned} \tag{16}$$

In our case, as the number of realizations in Λ_i increases with time, this procedure must be performed after a certain minimum initialization time. This conclusion, furthermore, is coherent with the theory of learning and intelligent systems.

Then, we could construct a second random variable $\tilde{X}$ (17) that is called ''reduced order model for X'', built with only M different elements; but whose statistical behavior is equivalent to the behavior of the original function X [31].

$$\tilde{X}(\Omega) = (\widetilde{x_1}, \ldots, \widetilde{x_M}) \subset X(\Omega) \tag{17}$$

Together with each element $\tilde{x_k}$ it is necessary to define a probably value $\tilde{p_k}$, so the pairs $(\tilde{x_k}, \tilde{p_k})$ provide a satisfactory approximation of the target function, and describe a random function with similar probability laws to X. In other words, it must be found two optimal vectors $\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}$ such that, for a given value of M parameter, it is obtained the closest possible random function $\tilde{X}$ to the target function X.

Considering function X is defined by sub-dataset Λ_i, and comparing the above description to the expression (12), it is easy to understand that the reduced order model for X, $\tilde{X}$, is in fact the reduced sub-dataset Λ_i^R we are looking for. It is important to note that, the reduced sub-dataset will have only $2M$ elements (or two M-dimensional vectors), in contrast to the original sub-data set that may have hundreds of thousands of elements.

For this reduced order model it is also possible to define the statistical parameters (18), as in expression (16). It must be noted that function $\mathbf{1}(\cdot)$ returns the unit if the logical expression in the argument is true, and zero in any other case.

$$\begin{aligned}
\tilde{\mu}(q) &= \sum_{k=1}^{M} \tilde{p_k} \cdot \widetilde{x_k}^q \\
\tilde{r}(j) &= \sum_{k=1}^{M} \tilde{p_k} \cdot \widetilde{x_k} \cdot \widetilde{x_{k+j}} \\
\tilde{F}(x_i) &= \sum_{k=1}^{M} \tilde{p_k} \mathbf{1}(\widetilde{x_k} \leq x_i)
\end{aligned} \tag{18}$$

At this point, in order to finally obtain the reduced sub-data set Λ_i^R, it is necessary to solve two last problems: the selection of the metric J, and the selection of vectors $\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}$.

For our purpose we employ a special type of reduced order models called Stochastic Reduced Order Models (SROM) [32], where distance between random functions are defined considering their most important statistical characteristics. Errors in the estimation of these statistical elements are weighted and aggregated in order to evaluate the distance (19).

$$\mathcal{J}\left(X, \tilde{X}\right) = \sum_{k=1}^{3} \rho_k e_k \left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) \tag{19}$$

Weights ρ_k may be defined as desired, depending on the application and the importance of each statistical element. The first error element is due to differences in the moments of the two random variables (20). In particular, it is obtained the aggregated relative quadratic deviation for moments from first order to q_{max} order.

$$e_1\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2} \sum_{q=1}^{q_{max}} \left(\frac{\tilde{\mu}(q) - \mu(q)}{\mu(q)}\right)^2 \tag{20}$$

The second component is calculated from errors in the correlation matrix (21); and the third element is obtained from differences in marginal distributions (22).

$$e_2\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2} \sum_{j=1}^{M} \left(\frac{\tilde{r}(j) - r(j)}{r(j)}\right)^2 \tag{21}$$

$$e_3\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2} \int \left(\tilde{F}(x_i) - F(x_i)\right)^2 dx_i \tag{22}$$

In order to calculate vector $\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}$ first the value of M must be selected [31]. This approach permits choosing the value of M depending on the computing resources in the AmI environment (as more powerful instruments are available, higher datasets may be considered). Besides, it is much easier to solve the optimization problem only with vectors $\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}$, than considering the value of M variables as well.

Furthermore, to simplify calculations in the optimization problem, vectors $\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}$ are not computed together but sequentially. First $\overrightarrow{\tilde{x}^{opt}}$ is obtained by means of any method guaranteeing the possibility to reach the optimum solution. Later, the optimization problem is finally addressed considering only $\overrightarrow{\tilde{p}^{opt}}$ as variable.

In order to calculate the optimum vector $\overrightarrow{\tilde{x}^{opt}}$, it is not possible to select M random samples from the original data set Λ_i, as it is not guaranteed they represent values in Λ_i in the best manner. On the other hand, processing the entire sub-dataset Λ_i to obtain values that represent the entire variation range is not possible in terms of computation time and software resources. Therefore, we select the following strategy.

First it is obtained a vector $\vec{\xi}$ composed of R samples, independent and randomly taken from Λ_i. It must be guaranteed that $M \ll R$. Then, using an appropriate technique, vector $\overrightarrow{\tilde{x}^{opt}}$ is extracted from vector $\vec{\xi}$. There are different valid approaches to do that [31], such as pattern classification of integer optimization; however, as in this work our objective is to create a lightweight and efficient security mechanism, we choose the technique named as ''dependent thinning''.

Dependent thinning, basically, employs any of the available algorithms to obtain samples of a hard-core Poisson process to thin vector $\vec{\xi}$ and calculate $\overrightarrow{\tilde{x}^{opt}}$. In these algorithms, elements in $\overrightarrow{\tilde{x}^{opt}}$ must fulfill a geometric requirement (23), that guarantees values from all the variation ranges are taken.

$$\mathfrak{d}\left(\tilde{x_i}, \tilde{x_j}\right) > d_0 \quad i, j = 1, \ldots, M; i \neq j \tag{23}$$

Function $\mathfrak{d}(\cdot, \cdot)$ represents a Euclidean distance such as $\mathfrak{d}_1$, $\mathfrak{d}_\infty$ or the traditional $\mathfrak{d}_2$. Besides, d_0 is a constant value

employed as control variable to construct the set $\overrightarrow{\tilde{x}^{opt}}$. It is important to note that, in these algorithms, there is only one degree of freedom, i.e. only one parameter (d_0 or M) can be selected freely, the other depends on the selected value. For example, for $d_0 = 0$, necessarily $M = R$.

Once obtained the vector $\overrightarrow{\tilde{x}^{opt}}$, it is possible to solve the optimization problem to calculate $\overrightarrow{\tilde{p}^{opt}}$, and the reduced sub-dataset Λ_i^R is constructed, as the solution of a SROM model.

C. REDUCING DATASETS WITH NO ERROR USING THE INFORMATION THEORY

The previously described procedure presents two practical problems. First, it is necessary to calculate the statistical parameters of the complete and original sub-dataset Λ_i, which is very costly in terms of processing time, software resources, memory, etc., as this set may contain hundreds of thousands of measures. And, second, the proposed method assumes the original model $\mathcal{M}$ is time-invariant, which is not true in general, as physical conditions, system configurations, etc., tend to change with time.

In order to address these problems, we introduce some information theory techniques into the previously described method based on SROM models. As SROM model enable the construction of reduced datasets as close as desired (with an error as small as desired) to the original dataset; in terms of information theory we say is a lossless procedure (i.e. compression). In this section we try to simplify calculations but also using lossless techniques. Sub-section D will describe a lossy technique.

As model $\mathcal{M}$ may change in time, its solution is no longer a random variable, but a stochastic process $\phi(t; x)$. This stochastic process, as variations in physical laws are soft, is locally stationary. In a locally stationary process, statistical distributions change in time, but we can consider they remain constant within time windows with a size of T_{win} time units (or N_{win} samples) (24).

$$X = \phi\left(t_0; x\right) = \phi\left(t_0 + \varepsilon; x\right) \quad \forall \varepsilon \leq T_{win} \tag{24}$$

On the other hand, as usual in information theory solutions, we are considering the stochastic process is ergodic (thus the process statistic parameters match temporary).

Only for the objective of this paper, we consider each sub-dataset Λ_i has a matrix structure: rows represent time instants, columns represent different devices (see Figure 5).

Then, we are processing this matrix column by column. First, we take a time window $w[n]$ so the stochastic process is stationary within it. Using this window as a sliding window we perform a spectral analysis of each column. This analysis, based on the Short-Time Fourier Transform (STFT), enables us to process data using digital processing techniques and reduce the size of the original dataset Λ_i. Figure 6 shows the block diagram of the proposed algorithm.

As can be seen, we consider a square window as it allows representing variations in the frequency spectrum in more resolution. In general, the STFT of the samples within

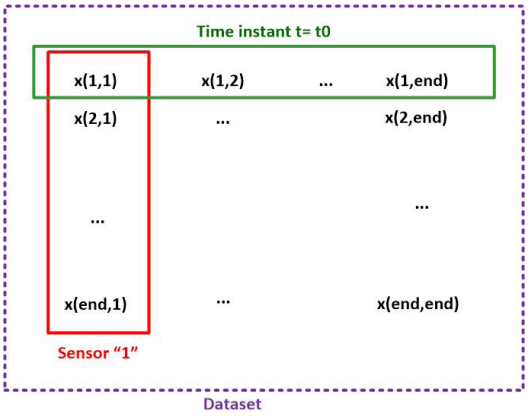

FIGURE 5. Matrix structure of the sub-datasets.

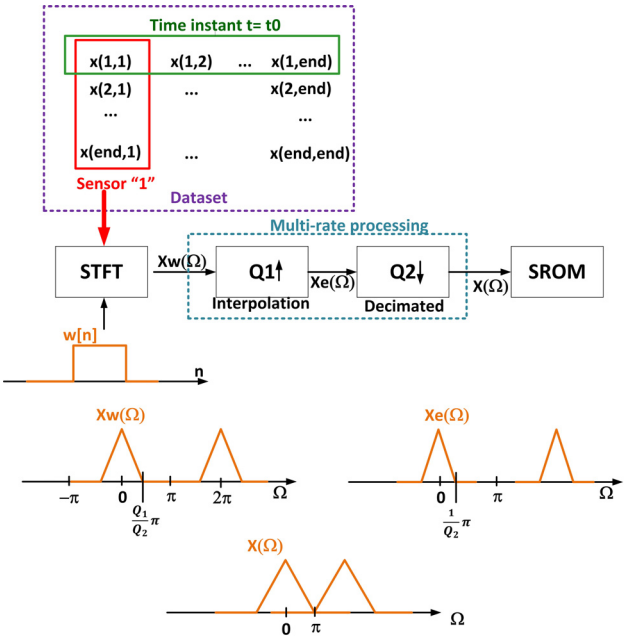

FIGURE 6. Block diagram of the proposed processing algorithm.

the window will only have relevant values up to a certain limit or bandwidth f_b (25).

$$f_b = \frac{Q_1}{Q_2}\pi \leq f_{Nyq} = \pi \tag{25}$$

In terms of the number of data in the dataset, evolution of the sensed physical variable has been sampled using a rate above the minimum required value (the Nyquist frequency f_{Nyq}). In particular, the sampling rate is $\frac{Q_2}{Q_1}$ times above the limit. Therefore, it is possible (using interpolation and decimated techniques) to remove some data with loss of information. At the end, it is possible to remove Q_2 samples per group of Q_1 data.

Details about how to implement interpolation and decimated devices are not provided in this paper. They are very well-known components (basic elements for the digital processing field), whose effects in the frequency spectrum can be

seen in Figure 6. However, a complete mathematical analysis of these elements is not the objective of this work and it can be found in the state-of-the-art.

Performing this processing using time analysis is not always feasible, as time reference is not often provided together with data in AmI environments. Nevertheless, frequency analysis is possible even if no time information is available.

Once the number of samples has been reduced as much as possible, the statistical parameters for this i-th time window in the l-th column are calculated (26)

$$\begin{aligned} \mu_{i,l}(q) &= E\left[X_{i,l}^{q}\right] \\ r_{i,l} &= E\left[X_{i,l} \cdot X_{i,l}\right] \\ F_{i,l}(x_j) &= P\left(X_{i,l} \leq x_j\right) \end{aligned} \tag{26}$$

Once obtained all these parameters, the window is sliced to the next position. This process should be repeated for each time windows in each column. However, if sub-datasets are very large, that may still require too much time. Therefore, in order to make processing time acceptable, statistical parameters will only be obtained for some time windows. Results for the other windows will be obtained by interpolation [35].

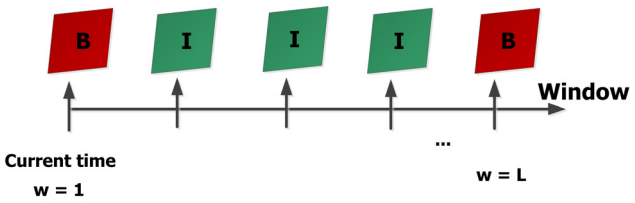

FIGURE 7. Sequence of intra-results and interpolated results.

In general, statistical results will be calculated only for one window per group of L time windows (27). Figure 7 represents the sequence of calculations. Two different types of results will be then obtained:

- Intra-result: It represents the real statistical parameters for a certain time window $w = w_0$. These states will be noted as S^B or B. At least, two intra-results must be calculated at the beginning of the processing algorithm.
- Interpolated result: It refers to the results obtained by interpolating two intra-results. These results are really fast to obtain but present a bigger uncertainty. The error goes up when more temporal distance exists between the intra-results employed to interpolate. These states will be notes as S^I or I.

The processing time goes down as a higher value is selected for L parameter. However, due to the introduced uncertainties during the interpolation process, the difference between the best reduced sub-dataset and the original one is bigger than if only intra-results are employed.

$$Result = \begin{cases} B & w = kL \quad k \in \mathbb{N} \\ I & others \end{cases} \tag{27}$$

One finished the described procedure three different matrix will be obtained: the matrix $\mu(q)$ of moments, the matrices r of correlations and the matrix F of marginal distributions. In all these three matrices, the cell (i, j) represents the value of the corresponding statistical parameter for the i-th time window in the l-th column (device). The statistical distance is, then, defined according to this new situation (28-30).

$$e_1^{i,j}\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2}\sum_{q=1}^{q_{max}}\left(\frac{\widetilde{\mu_{i,j}}(q) - \mu_{i,j}(q)}{\mu_{i,j}(q)}\right)^2 \tag{28}$$

$$e_2^{i,j}\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2}\sum_{j=1}^{M}\left(\frac{\widetilde{r_{i,j}}(j) - r_{i,j}(j)}{r_{i,j}(j)}\right)^2 \tag{29}$$

$$e_3^{i,j}\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) = \frac{1}{2}\int\left(\widetilde{F_{i,j}}(x_k) - F_{i,j}(x_k)\right)^2 dx_k \tag{30}$$

At this point three different decisions can be made to obtain the reduced sub-dataset Λ_i^R:

- Calculating a reduced dataset of $2M$ elements per time window and device (column). In this case, expressions proposed in the previous section are directly applied to reduce the dataset using SROM models. With this option, the total obtained reduction is the smallest, but intelligent system can only be trained to detect very small distortions in the system behavior.
- Calculating a reduced dataset of $2M$ elements per device (column) or time window. In this case, we employ the same reduced sub-dataset to represent an entire device of time window. Expressions related to SROM models must be slightly modified to aggregate errors due to time windows or devices (depending on the calculation being performed) (31) [34]. This option provides equilibrium between the calculation of a very detailed dataset (first option) and a very reduced sub-dataset (the third option).
- Calculating a reduced dataset of $2M$ elements for all devices (columns) or time windows. In this case only one sub-dataset represents the entire stochastic process. With this option the most reduced dataset is obtained. Expressions related to SROM models must be modified to aggregate errors due to time windows and devices (32).

$$\mathcal{J}\left(X, \tilde{X}\right) = \sum_{i\,or\,j}\sum_{k=1}^{3}\rho_k e_k^{i,j}\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) \tag{31}$$

$$\mathcal{J}\left(X, \tilde{X}\right) = \sum_{i,j}\sum_{k=1}^{3}\rho_k e_k^{i,j}\left(\overrightarrow{\tilde{x}^{opt}}, \overrightarrow{\tilde{p}^{opt}}\right) \tag{32}$$

Considering any of the described decisions, a reduced sub-dataset will be obtained by following a lossless compression procedure.

D. REDUCING DATASETS ARBITRARILY USING THE INFORMATION THEORY

In the previous section we have improved the efficiency of calculations about SROM models, using information theory techniques to reduce the original sub-datasets.

However, we have proposed a design to avoid the loss of information, what has imposed a limit to the data reduction we can obtain. If certain loss of information was tolerated, then, it would be possible to go beyond these limits. This section is focused on this hypothesis.

Thus, in this case, before performing the spectral analysis proposed in the previous section, we are quantifying the values stored in the original dataset Λ_i (33). This procedure is irreversible and generates a loss of information that degrades the data in the dataset.

$$\forall \lambda \in \Lambda_i : \lambda \in [\lambda_{min}, \lambda_{max}] \xrightarrow{Quant} \widehat{\lambda} \in \widehat{\Lambda}_i : \widehat{\lambda} \in \{\lambda_1, \ldots, \lambda_Q\} \tag{33}$$

In fact, as no information about the quality of the stored data is available, in all previous sections we are considering they have an infinite quality. In this last proposal the quality is reduced to a finite value.

In order to quantify the data in the dataset we are employing Q different values (33). The value of Q may be chosen freely, but it determines the quality of the obtained quantified sub-dataset $\widehat{\Lambda}_i$.

If we consider the second theorem of the information theory (Shannon's source coding theorem) together with the Hartley's theorem, we obtain a theorem that can be used to estimate the quality of the data after their quantification (34-35).

$$Q = \sqrt{1 + (S/N)} \tag{34}$$

$$(S/N) = Q^2 - 1 \tag{35}$$

In this context the Signal-Noise relation (SNR) represents the quality of the data after the quantification. In this case, the added noise must be understood as numerical, stochastic or quantification noise (it does not have an electrical origin) As Q goes up and more values are employed to quantify, the final quality is higher. In general, thus, as the (S/N) ration is closer to zero the quality of the quantified dataset $\widehat{\Lambda}_i$ decreases.

In fact, this process does not reduce the size of the original sub-dataset Λ_i, but it allows removing much more samples during the spectral analysis. Of course, this improvement is much more important as Q goes down (although the obtained quality is worse).

This effect is possible as, in general, quantified signals present several times in a row the same value, which means they evolve slower. Then, in the frequency spectrum, relevant values will be around zero, which means that the sampling frequency is well above its minimum value (the Nyquist's limit). Thus, using interpolation and decimation techniques a lot of samples will be removed (without decreasing the input quality or SNR). Figure 8 represents this phenomenon.

The second and last technique which will allow us to arbitrarily reduce the size of a dataset Λ_i after quantifying it, is the reduction of the sampling rate below its minimum value

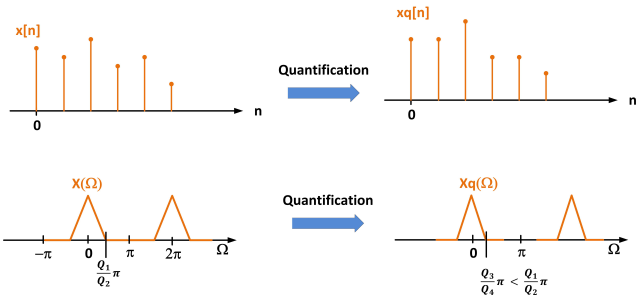

FIGURE 8. Frequency spectrum of a signal and its quantified version.

(the Nyquist's limit). This procedure is also irreversible, and produces in general the "smearing" of data. The reduction in the quality of data may be estimated, so it is possible to select (for each application) the compression level in terms of the minimum admissible quality of the input data.

We assume we want to reduce the sampling rate below the Nyquist limit (36). Then, as discrete signals are 2π-periodic in the frequency spectrum, collisions and interferences between different replicas will appear.

$$f_b = \frac{Q_1}{Q_2}\pi > f_{Nyq} = \pi \tag{36}$$

The signal power of the spectrum replicas that collide with the main area in the frequency spectrum (i.e. $\Omega \in [-\pi, \pi]$), represent a new noise (this time with an electrical origin) that causes the reduction in the signal (data) quality. This new noise N_c may be easily estimated (37).

$$x[n] \xrightarrow{STFT(FFT)} X[k] \quad N_c = 2 \cdot \sum_{k \geq \pi} |X[k]|^2 \tag{37}$$

Then, if we consider the quantified sub-dataset has a quality of $SNR_0 = (S/N_0)$, it is possible to estimate the deterioration in the data quality because of the use of sub-sampling techniques. From the calculated frequency spectrum we can obtain a ratio that includes the signal power, the stochastic noise and the electric noise (38).

$$SNR_c^* = \frac{S + N_0}{N_c} \tag{38}$$

In order to obtain the final quality of the data after being quantified and sub-sampled, we only have to operate both previous results (39). With this procedure we can also obtain the deterioration percentage $D(\%)$ due to sub-sampling (40).

$$SNR_c = (S/N_c) = \frac{SNR_c^* \cdot SNR_0}{SNR_0 + 1} \tag{39}$$

$$D(\%) = 100 \cdot \left(1 - \frac{SNR_c^*}{SNR_0 + 1}\right) \tag{40}$$

The calculation of the final reduced sub-dataset is equal to the described solution in the previous section once performed the sub-sampling.

IV. EXPERIMENTAL VALIDATION

In order to evaluate the performance of the proposed solutions, six different experiments were carried out. The first five experiments were based on numerical simulations. The last experiment was performed in a real scenario deployment.

The first experiment was designed to evaluate the impact of parameter M in the precision and statistical closeness between the original dataset and the reduced one. A dataset is constructed with fifteen million (15M) entries, produced by one hundred and fifty (150) different sensors. Using SROM models (as explained in section III.B, without considering any information theory technique) the proposed dataset is reduced, considering different values for M parameter. Distance (error) between the original statistical parameters and the parameter of the reduced set is evaluated.

The second experiment is proposed to evaluate the improvement in the processing time that is obtained when employing information theory techniques in a lossless scenario. Using the same original dataset proposed for the first experiment, the mean processing time required to reduce the dataset is evaluated, considering only the SROM models and considering, besides, information theory techniques. The experiment is repeated for different values of the M parameter, the L parameter (number of interpolated results) and the N_{win} parameter (window size where the stochastic process is considered stationary).

The third experiment was very similar to the second one. Using the proposed dataset, it was reduced using SROM models and information theory techniques in a lossless scenario, as explained in Section III.C. The experiment was repeated for different values of the M parameter, the L parameter (number of interpolated results) and the N_{win} parameter (window size where the stochastic process is considered stationary). Distance (error) between the original statistical parameters and the parameter of the reduced set is evaluated.

The fourth experiment was designed to evaluate the performance of the proposed solution in scenarios where certain information losses are tolerated. This experiment was very similar to the previous one. The proposed dataset was reduced using SROM models and information theory techniques, as explained in Section III.D. The experiment was repeated for different values of the Q parameter (i.e. different data qualities). Distance (error) between the original statistical parameters and the parameter of the reduced set is evaluated.

Finally, the fifth experiment was proposed to evaluate the improvement in the processing time that is obtained when employing information theory techniques in a lossy scenario. Using the same original dataset proposed for the first experiment, the mean processing time required to reduce the dataset is evaluated, considering only the SROM models and (on the other hand) lossy algorithms. As in the fourth experiment, this last validation was repeated for different values of the Q parameter (i.e. different data qualities).

In all these first five experiments, we employ the MATLAB software as simulation platform, where very efficient algorithms to calculate statistical parameters and other similar elements are available. In particular, MATLAB 2017b was employed. This tool was deployed in a Linux (Ubuntu 16.04) machine with 8GB of RAM memory and an Intel i7 processor.

All previously described experiments are focused on evaluating the performance of the proposed dataset processing techniques. However, it is necessary to prove that the obtained reduced dataset is valid to train future intelligent security systems. Therefore, in the sixth and final experiment this validation was performed.

In this experiment, a real deployment of twenty-five sensors based on the Samsung Artik 020 platform was developed (see Figure 9). These sensors were producing data, each one, at a rate of 1 $datum/s$. Sensors were deployed in the first floor of the B-building in the Telecommunication school (at Technical University of Madrid).

FIGURE 9. Sensors based on the Samsung Artik 020 architecture.

The produced dataset was employed to train an intelligent security system based on the game theory [7]. Results about the success rate when detecting cyber-physical attacks were registered.

The dataset constructed by sensors was also reduced using the three proposed techniques: SROM models, lossless information theory techniques, and lossy algorithms. SROM models were configured to produce datasets of $M = 20$ data. Other important parameters such as L or Q were configured as indicated: $L = 50$, $N_{win} = 3600$, $Q = 32$.

A security solution was trained with each one of these reduced datasets. Results about the success rate when detecting cyber-physical attacks were registered. Results are compared to the success rate obtained during the first part of the experiment.

V. RESULTS

This section presents and discusses some results of the experiments described in the previous section.

In order to remove from the results of the experiments (as much as possible) variations in the simulations due to exogenous variables (e.g. delays in the operations performed by the operating systems), for each case twelve different simulations were performed. The average of all these measures was obtained to calculate the final results.

Figure 10 shows the results of the first experiment. As can be seen, the total error decreases as M goes up, being practically zero for values above $M = 600$. Any case, for very small values of M parameter, such as $M = 2$, the total error

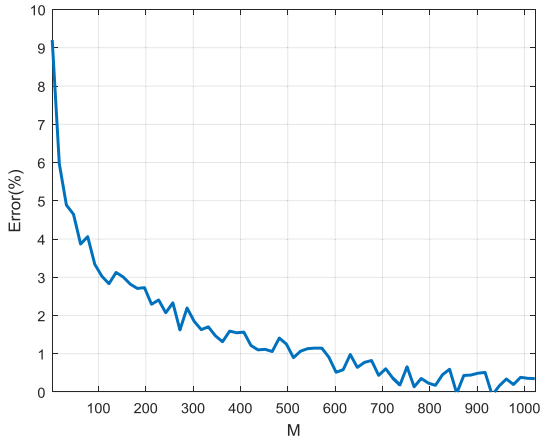

FIGURE 10. Results of the first experiment. Statistical distance between the original dataset and the reduced one.

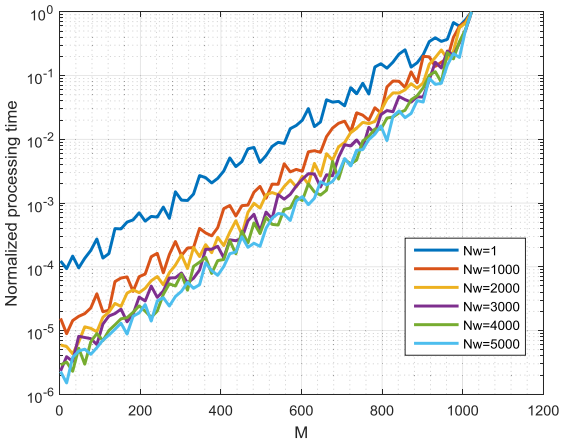

FIGURE 12. Results of the second experiment. Improvement in the processing time depending on N_{win}: lossless scenario.

is also assumable, as it is around 10% which is the typical error for standard measurement systems.

In fact, obtained results for the first experiments are coherent with previously reported experiences in the state-of-the-art about SROM models [32].

For the second experiment, to be able to compare results when using information theory technologies to results when no information theory technique is considered, we are obtaining only one reduced dataset for all time windows and devices in the original sub-dataset.

With this consideration, it is possible to perform this second experiment. Results are shown on Figure 11 and Figure 12.

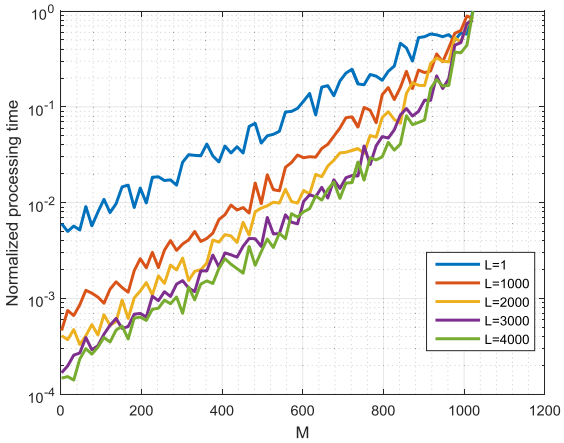

FIGURE 11. Results of the second experiment. Improvement in the processing time depending on L: lossless scenario.

As it can be seen, the behavior in both figures is similar, as (at the end) both increasing the value of L parameter or increasing the value of N_{win} parameter, causes the number of time windows to be evaluated to decrease.

Results show that, using information theory technologies, it is possible to reduce the processing time in one magnitude order (for small values of M parameter). However, as M goes up, the required time to solve the underlying optimization problem in SROM models increases exponentially and dominates in the aggregated total time, so all curves converge. This point is reached around $M = 1000$.

The second experiment proves the proposed information theory techniques are useful to reduce the processing time, however, it is necessary to evaluate the error associated with this new approach. Figure 13 and Figure 14 show the results of the third experiment.

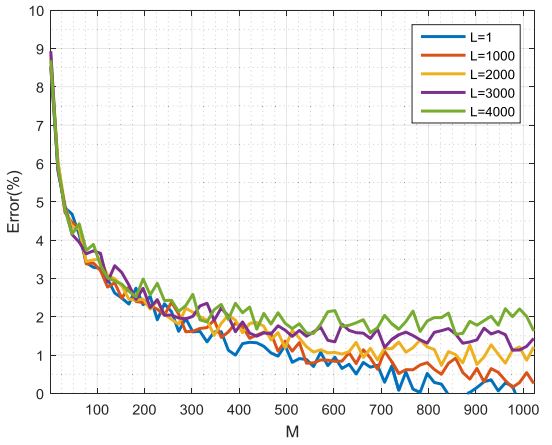

FIGURE 13. Results of the third experiment. Statistical distance between the original dataset and the reduced one depending on L: lossless scenario.

As it can be seen, as in the previous experiment, the behavior is similar in both figures. When M presents a small value, error associated to the SROM model dominates the total error, and there is no dependency on L or N_{win}. However, as M increases and the error due to the statistical calculations goes down, errors caused by interpolation and the assumption that the stochastic process is locally stationary are relevant.

There is, nevertheless, an important difference. As errors due to interpolation techniques are higher, as L parameter goes up (being M a high value) the total error can reach the

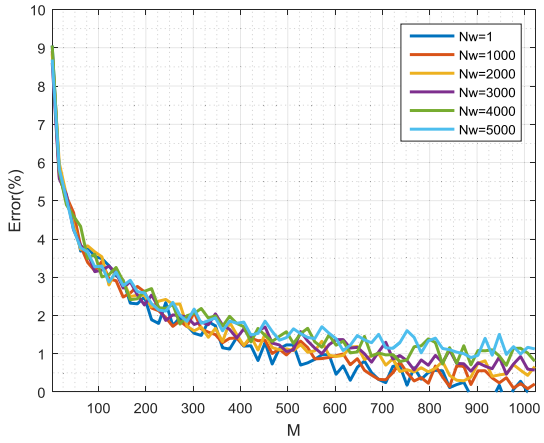

FIGURE 14. Results of the third experiment. Statistical distance between the original dataset and the reduced one depending on N_{win}: lossless scenario.

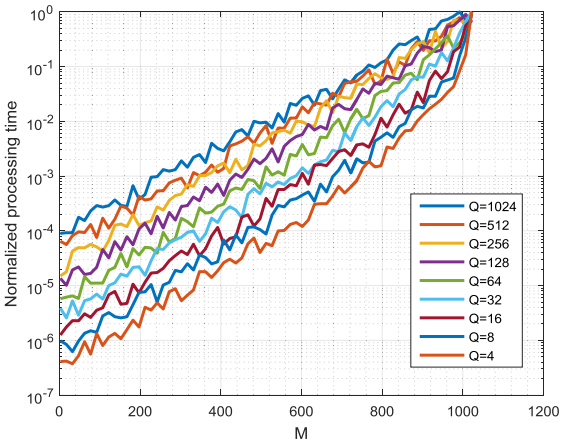

FIGURE 16. Results of the fifth experiment. Improvement in the processing time depending on Q: lossy scenario.

value of 2%. However, errors caused by considering stochastic process as locally stationary are much smaller, so the total error only reaches a maximum value around 1%.

The fourth and fifth experiments are focused on evaluating the proposed solutions for lossy scenarios. In the fourth experiment the statistical distance (error) associated to these techniques is evaluated. In the fifth experiment the improvement in the processing time is estimated. Figure 15 shows the results of the fourth experiment. Figure 16 shows the results of the fifth experiment.

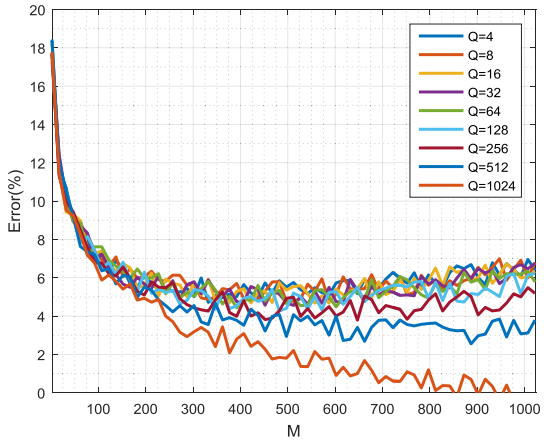

FIGURE 15. Results of the fourth experiment. Statistical distance between the original dataset and the reduced one depending on Q: lossy scenario.

As can be seen on Figure 15, in this case, error increases around 50% in respect to the use of common SROM models (see Figure 10), so the maximum error (for $M = 2$) is around 18%. That is because errors are accumulative, and error caused by the loss of information is never negligible; furthermore, it is similar to the maximum error generated by standard SROM models. It can be also seen how, when errors due to SROM models are practically zero ($M = 1000$) the total error remains around 8% (because of the reduction in the data quality).

This error drastically decreases when the number of quantification level increases. Thus, if we consider that $Q = 1024$, the behavior is similar to the one obtained for standard SROM models.

This degradation in the precision (quality) of the reduced dataset allows, on the other hand, an important reduction in the processing time (see Figure 16). A reduction of almost three orders of magnitude may be obtained if the number of quantification levels is reduced from $Q = 1024$ to $Q = 4$. In this case, however, as M parameter goes up, the total processing time converges to a unique value due to the important resource consumption when solving the optimization problem associated to SROM models.

Therefore, the use of information theory techniques (both lossless solution and lossy algorithms) is only advisable when M is configured with small and medium values (up to $M = 900$ approximately).

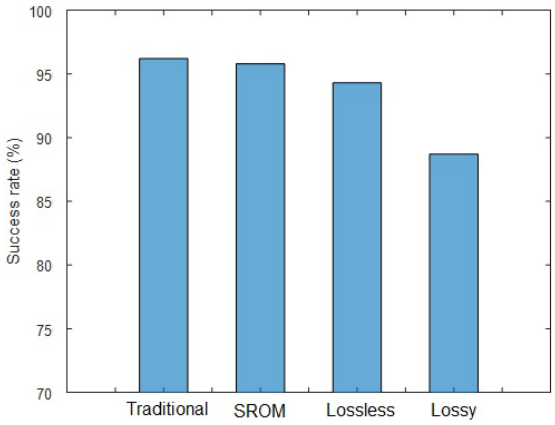

FIGURE 17. Results of the sixth experiment. Success rate.

Finally, during the sixth experiment, once the good performance of the proposed technologies was proved, we evaluate the utility of our proposal. Figure 17 shows the results of the sixth experiment where the success rate for the same security

system, but trained using different reduced and non-reduced datasets, is compared.

As can be seen, the success rate is very similar when employing the non-reduced dataset, a dataset reduced using only SROM models and a dataset reduced using SROM models and lossless information theory techniques. In all cases it is around 96%. However, this rate decreases around 10% when lossy information theory techniques are employed (the rate is near 88%). On the other hand, the processing time is much lower when employing lossy techniques, so (as the success rate is still acceptable) they are a very good option for systems operating at real-time.

VI. CONCLUSIONS

In this paper we have investigated and proposed a procedure to reduce large datasets, with the objective of enabling new security techniques to detect cyberattacks in a fast and efficient way.

The proposed solution is based on the use of Stochastic Reduced Order Models (SROM) which are complemented with information theory techniques to improve the processing time and the compression rate. Information theory techniques are valid for both lossless and lossy scenarios.

Using the Nyquist's and Shannon's theorems, as well as spectral analysis technologies, it is possible to reduce datasets before using SROM models, to calculate the final reduced datasets preserving the statistical properties of the original set.

Results showed that reduced datasets are valid solutions to train intelligent security systems, maintaining the success rate in the same level as if standard datasets were employed. Besides, numerical simulations proved that the obtained processing time enables the use of the proposed solution in real-time applications, contrary to traditional approaches.

REFERENCES

[1] N. F. S. Zurita, M. K. Colby, I. Y. Tumer, C. Hoyle, and K. Tumer, "Design of complex engineered systems using multi-agent coordination," *J. Comput. Inf. Sci. Eng.*, vol. 18, no. 1, p. 011003, 2018.

[2] B. Bordel, R. Alcarria, T. Robles, and D. Martín, "Cyber–physical systems: Extending pervasive sensing from control theory to the Internet of Things," *Pervasive Mobile Comput.*, vol. 40, pp. 156–184, Sep. 2017.

[3] B. B. Sánchez, R. Alcarria, D. Sánchez-de-Rivera, and Á. S. Picot, "Enhancing process control in industry 4.0 scenarios using cyber-physical systems," *J. Wireless Mobile Netw.*, vol. 7, no. 4, pp. 41–64, 2016.

[4] B. Bordel, R. Alcarria, D. Sánchez-de-Rivera, D. Martín, and T. Robles, "Fast self-configuration in service-oriented Smart Environments for real-time applications," *J. Ambient Intell. Smart Environ.*, vol. 10, no. 2, pp. 143–167, 2018.

[5] D. J. Cook, J. C. Augusto, and V. R. Jakkula, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive Mobile Comput.*, vol. 5, no. 4, pp. 277–298, Aug. 2009.

[6] M. Weiser, "The computer for the 21st Century," *IEEE Pervasive Comput.*, vol. 1, no. 1, pp. 19–25, Jan./Mar. 2002.

[7] B. Bordel, R. Alcarria, D. Sánchez-de-Rivera, and T. Robles, "Protecting industry 4.0 systems against the malicious effects of cyber-physical attacks," in *Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.* Cham, Switzerland: Springer, Nov. 2017, pp. 161–171.

[8] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *Proc. 50th IEEE Conf. Decis. Control Eur. Control Conf. (CDC-ECC)*, Dec. 2011, pp. 2195–2201.

[9] A. Gabriel, "Design and evaluation of safety instrumented systems: A simplified and enhanced approach," *IEEE Access*, vol. 5, pp. 3813–3823, 2017.

[10] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber-physical system security for the electric power grid," *Proc. IEEE*, vol. 100, no. 1, pp. 210–224, Jan. 2012.

[11] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu, "Cyber-physical security testbeds: Architecture, application, and evaluation for smart grid," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 847–855, Jun. 2013.

[12] B. Zhu and S. Sastry, "SCADA-specific intrusion detection/prevention systems: A survey and taxonomy," in *Proc. 1st Workshop Secure Control Syst. (SCS)*, vol. 11, Apr. 2010, pp. 1–16.

[13] B. Zhu, A. Joseph, and S. Sastry, "A taxonomy of cyber attacks on SCADA systems," in *Proc. Int. Conf. Internet Things (iThings/CPSCom), 4th Int. Conf. Cyber, Phys. Soc. Comput.*, Oct. 2011, pp. 380–388.

[14] Y. Mo *et al.*, "Cyber–physical security of a smart grid infrastructure," *Proc. IEEE*, vol. 100, no. 1, pp. 195–209, Jan. 2012.

[15] R. M. Clark and S. Hakim, Eds., *Cyber-Physical Security: Protecting Critical Infrastructure at the State and Local Level*, vol. 3. Cham, Switzerland: Springer, 2017.

[16] M. Yampolskiy, P. Horváth, X. D. Koutsoukos, Y. Xue, and J. Sztipanovits, "A language for describing attacks on cyber-physical systems," *Int. J. Critical Infrastruct. Protection*, vol. 8, pp. 40–52, Jan. 2015.

[17] S. Huang, C.-J. Zhou, S.-H. Yang, and Y.-Q. Qin, "Cyber-physical system security for networked industrial processes," *Int. J. Autom. Comput.*, vol. 12, no. 6, pp. 567–578, 2015.

[18] J. E. Kim, T. Barth, G. Boulos, J. Yackovich, C. Beckel, and D. Mosse, "Seamless integration of heterogeneous devices and access control in smart homes and its evaluation," *Intell. Buildings Int.*, vol. 9, no. 1, pp. 23–39, 2017.

[19] U. Salama, L. Yao, X. Wang, H.-Y. Paik, and A. Beheshti, "Multi-level privacy-preserving access control as a service for personal healthcare monitoring," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Jun. 2017, pp. 878–881.

[20] B. Genge, I. N. Fovino, C. Siaterlis, and M. Masera, "Analyzing cyber-physical attacks on networked industrial control systems," in *Proc. Int. Conf. Critical Infrastruct. Protection*. Berlin, Germany: Springer, Mar. 2011, pp. 167–183.

[21] P. Uchenna, D. Ani, M. Hongmei, and H. A. Tiwari, "Review of cybersecurity issues in industrial critical infrastructure: Manufacturing in perspective," *J. Cyber Secur. Technol.*, vol. 1, no. 1, pp. 32–74, 2017.

[22] R. Waslo, T. Lewis, R. Hajj, and R. Carton, *Industry 4.0 and Cybersecurity—Managing Risk in an Age of Connected Production*. U.K.: Deloitte Univ. Press, Mar. 2017.

[23] A. A. Cárdenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Proc. Workshop Future Directions Cyber-Phys. Syst. Secur.*, Jul. 2009, pp. 1–7.

[24] L. Cavallaro and D. Gollmann, Eds., *Information Security Theory and Practice. Security of Mobile and Cyber-Physical Systems*, vol. 7886. Heraklion, Greece: Springer, May 2013.

[25] D. Gollmann and M. Krotofil, "Cyber-physical systems security," in *The New Codebreakers* (Lecture Notes in Computer Science), vol. 9100. Berlin, Germany: Springer, 2016, pp. 195–204.

[26] A. Moore and M. S. Lee, "Cached sufficient statistics for efficient machine learning with large datasets," *J. Artif. Intell. Res.*, vol. 8, pp. 67–91, Mar. 1998.

[27] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *Proc. Conf. Vis. Celebrating Ten Years*. San Francisco, CA, USA: IEEE Computer Society Press, Oct. 1999, pp. 43–50.

[28] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining: Closed sequential patterns in large datasets," in *Proc. SIAM Int. Conf. Data Mining Soc. Ind. Appl. Math.*, May 2003, pp. 166–177.

[29] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain, "A multiresolution Gaussian process model for the analysis of large spatial datasets," *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 579–599, 2015.

[30] B. Bordel, A. B. Orúe, R. Alcarria, and D. Sánchez-De-Rivera, "An intra-slice security solution for emerging 5G networks based on pseudo-random number generators," *IEEE Access*, vol. 6, pp. 16149–16164, 2018.

[31] M. Grigoriu, "Reduced order models for random functions. Application to stochastic problems," *Appl. Math. Model.*, vol. 33, no. 1, pp. 161–175, 2009.

[32] J. E. Warner, M. Grigoriu, and W. Aquino, "Stochastic reduced order models for random vectors: Application to random eigenvalue problems," *Probabilistic Eng. Mech.*, vol. 31, pp. 1–11, Jan. 2013.
[33] J. E. Warner, W. Aquino, and M. D. Grigoriu, "Stochastic reduced order models for inverse problems under uncertainty," *Comput. Methods Appl. Mech. Eng.*, vol. 285, pp. 488–514, Mar. 2015.
[34] S. Sarkar, J. E. Warner, W. Aquino, and M. D. Grigoriu, "Stochastic reduced order models for uncertainty quantification of intergranular corrosion rates," *Corrosion Sci.*, vol. 80, pp. 257–268, Mar. 2014.
[35] B. B. Sánchez, R. Alcarria, D. Sánchez-de-Rivera, and A. Sánchez-Picot, "Predictive algorithms for mobility and device lifecycle management in Cyber-Physical Systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, p. 228, Sep. 2016.

BORJA BORDEL received the B.S. and M.S. degrees in telecommunication engineering from the Technical University of Madrid in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree in telematics engineering with the Telecommunication Engineering School. His research interests include cyber-physical systems, wireless sensor networks, radio access technologies, communication protocols, and complex systems.

RAMÓN ALCARRIA received the M.S. and Ph.D. degrees in telecommunication engineering from the Technical University of Madrid in 2008 and 2013, respectively. He is currently an Assistant Professor with the E.T.S.I Topography, Technical University of Madrid. He is involved in several research and development European and national projects related to future Internet, Internet of Things, and service composition. His research interests are service architectures, sensor networks, human–computer interaction, and prosumer environments.

TOMÁS ROBLES received the M.S. and Ph.D. degrees in telecommunication engineering from the Technical University of Madrid in 1987 and 1991, respectively. He is currently a Full Professor of telematics engineering with the E.T.S.I. Telecommunication, Technical University of Madrid. His research interests are focused on advanced applications and services for wireless networks, also on blockchain-based infrastructures.

ÁLVARO SÁNCHEZ-PICOT received the M.S. degree in telecommunication engineering from the Technical University of Madrid in 2014, where he is currently pursuing the Ph.D. degree with the Department of Telematics Systems Engineering, E.T.S.I. Telecommunications. His thesis entitled "Contribution to the Integration of Network Simulators and Social Simulation Environments for the Modeling of Environments and Smart Devices." His research interests are focused on sensor networks, simulation of network communications, wireless communications, and Web development.