Analysis of Quality Measurements to Categorize Anomalies in Sensor Systems

Pedro Merino Laso, David Brosset, John Puentes

Analysis of Quality Measurements to Categorize Anomalies in Sensor Systems

Pedro Merino Laso
Chair of Naval Cyber Defense,
École navale - CC 600 F29240
Brest Cedex 9, France
Email: pedro.merino@ecole-navale.fr

David Brosset
Naval Academy Research Institute,
École navale - CC 600 F29240
Brest Cedex 9, France
Email: david.brosset@ecole-navale.fr

John Puentes
Dpt. Image et Traitement de l'Information
Institut Mines-Telecom;Telecom Bretagne
Lab-STICC UMR CNRS 6285 Équipe DECIDE,
CS 83818 29238
Brest Cedex 3, France
Email: john.puentes@telecom-bretagne.eu

***Abstract*—Sensor networks are becoming ubiquitous, enabling to improve decision-making and reducing human interaction by means of automatic or semi-automatic responses. However, due to deterioration or induced effects, sensors measures can be affected and produce anomalies that could alter decision-making. Most of the existing methods to identify sensors irregularities focus basically on detecting and discarding anomalous values, without looking for complementary information to understand generated anomalies. This paper presents an approach to obtain such complementary information by categorizing sensor anomalies, based on multidimensional quality assessment. It consists of two processing stages: an evaluation of data and information streams to estimate data quality imperfections and information quality dimensions; followed by the determination of agreement limits, compliant with normal states, to identify and categorize anomalies. The case study of discrete and analog sensors system installed in a simulator training platform of fuel tanks is presented, to illustrate an application of the proposed approach, considering 13 experimentally evaluated anomalies.**

***Keywords*—Sensor; Anomaly categorization; Data quality; Information quality; Cyber-physical system**

I. Introduction

Sensor networks, as part of cyber-physical systems, are becoming ubiquitous, because of the control and survey possibilities offered by their exploitation in multiple domains like transport, manufacturing, home automation, and more recently the internet of things. Equivalent systems in the industrial domain, called Supervisory Control and Data Acquisition (SCADA), are composed by sub-systems that make measurements of the surrounding environment with sensors and execute responses using actuators. Besides sensors, control and communication modules also generate critical data streams to support decision makers.

With the growing utilization of sensor networks, potential cost of errors provoked by anomalous sensor responses is becoming increasingly important. Nevertheless, sensor systems are exposed to multiple operational risks. For this reason, sensor systems vulnerability has been examined in other fields like, automated vehicles [1], global positioning system (GPS) [2], [3], and maritime navigation devices [4], [5], [6]. A particular effort is thus required to detect anomalies in cyber-physical systems intended to: estimate automatically the pertinence of sensor data streams regardless of data conditions and deliver contextualized information to assist decision makers.

Conventionally, sensor anomalies are detected to be discarded, without interpreting what was wrong with collected data, or determining if it was possible to extract some information. Yet, anomaly detection could serve to implement complementary analyses to reinforce decisions. For instance: discover evidence of intrusions; identify natural sensor deterioration; recognize or anticipate malfunctions in data acquisition; facilitate decisions about anomalies correction; and determine if malfunctions can be trusted and integrated to the decision support process.

However, before having the possibility of exploiting the latent unknown value of detected anomalies, it is fundamental to rely on a coherent and adapted categorization. Although it is necessary, such categorization has not been specifically defined for sensor systems. On the other hand, multidimensional criteria are necessary to infer the implications of sensor anomalies, independently of system architecture, data types, and processed information. One unexplored alternative in sensor systems is to examine variations of quality measurements, instead of bare anomaly detection. This work addresses therefore the question of how to categorize sensor system anomalies applying quality evaluation, to identify which quality dimensions are the most pertinent for further analysis. The originality of this approach is to extract complementary information despite the detection of anomalous sensor data.

In the rest of the paper, previous related anomaly detection works and quality evaluation principles are summarized in section II, before describing the proposed methodology in section III. To illustrate how the proposed approach could be applied, the use case of a vessel fuel tank prototype is studied in Section IV. It describes and examines an exhaustive group of sensor anomalies, as well as the estimated impacts on data and information quality measures. Discussion and conclusion are presented in Section V.

II. Background

Anomaly detection in sensor systems and quality measurements of sensors data have been investigated separately. The next subsections identify noticeable aspects of both fields that permit to illustrate the possibility of a pluridisciplinary approach.

A. Anomaly Detection

Anomaly detection is a common problem in diverse research domains as statistics, machine learning, data mining, information theory, and spectral theory. Applications of anomaly detection are very large, including among others, the detection of intrusions, frauds, and damages. Anomaly definitions vary depending on the application field. A general definition for cyber-physical systems is [7]:

> Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior.

Notwithstanding considerable research about anomaly detection in sensor networks, there is a lack of standard methods for cyber-physical systems [8]. Given that anomalies have been studied separately for particular systems, developed approaches only fit correctly a specific sub-system and context. Also, some initiatives have examined sensor systems characterized by harsh and uncertain functioning environments. Those closely related to the examined problem are nuclear plants, spacecraft, and ships.

Nuclear plants are critical industrial structures on which anomaly detection studies search to predict system behavior by means of probabilistic methods or sequences models. Making use of hierarchical sensor networks, outliers were detected calculating the distance or the density between neighbor measure points [9]. Also, network activity was characterized by stochastic models to represent the interaction between different components [10]. Indirect anomalies detection applying Bayesian networks was investigated to assist decision-making in critical time for spacecraft, on which it is required to display sensors information in the best possible way [11]. Sensor anomaly detection for spacecraft direction monitoring, compared a normal learned logged telemetry data model to incoming data, assuming that changes of causal associations between components mean arbitrary relationships [12]. Other methods permit to create rules automatically for anomaly detection. Three of the most known algorithms are C4.5 rules, incremental reduced error pruning, and repeated incremental pruning to produce error reductions that have particular performances [13].

In ships, anomaly detection is usually limited to demanding sub-systems. Compared to spacecraft, features to be examined are more manageable, *i.e.* the functioning environment is accessible and most of physical components can be inspected. In addition, a specialized crew may detect part of these anomalies, associated to secondary decisions. Nevertheless, a guiding system like GPS is sensible to spoofing attacks, for which six detection techniques were identified [14]: Amplitude discrimination, time-of-arrival discrimination, consistency of navigation inertial measurement unit cross-check, polarization discrimination, angle-of-arrival discrimination, and cryptographic authentication. Another important sub-system in maritime navigation is the automatic identification system (AIS). It allows sharing navigation information between vessels to avoid collisions, permitting also to detect anomalous behaviors, potentially associated to illicit activities as [3]: deviation from standard routes, unexpected AIS activity, unexpected port arrival, close approach, and zone entry. Additionally, sensor systems were conceived to cope only with monitoring needs, without taking into account security [15]. Nevertheless, current technology and operational trends are leading to develop a growing interest in sensor system security [16], [17], [18], [19]. Among various alternatives, anomalies detection emerges as an approach that could also be used to address the cyber-attacks detection problem [20].

Described works, mostly identify anomalies to make sure that decisions taken by human experts, rely exclusively on cleaned data. Hence, data and information quality is not studied to analyze sensor anomalies in these cases. Furthermore, neither the source nor the kind of anomaly, along with the consequences for data and information quality analysis, are characterized.

B. Quality Evaluation

Although data and information are two different concepts, they are frequently used indistinctly or in a confusing manner in the literature [21]. To avoid such misunderstanding, we base our data and information definitions on the well-known DIKW - Data, Information, Knowledge, and Wisdom - pyramid, *i.e.* know-nothing, know-what, know-how, and know-why, respectively [22]. This work focuses on the data and information entities.

To adapt the DIKW definition to cyber-physical systems, **data** are the streams of bits with no comprehensible sense (know-nothing) *i.e.* binary data and multidimensional signals; whereas **information** corresponds to data with semantic sense in a context (know-what). Taking into account the characteristics of the system, data and information definitions can be thus defined for cyber-physical systems applying:

$$Information = Data + Context_{sub\text{-}system} + Context_{system} \quad (1)$$

Where we assume that the context is correctly defined by the respective sub-system and system specifications. For subsystems, the context is commonly available in one or several data-sheets. For systems, the context is given by global specifications and the environment characteristics. Global specifications and the environment characteristics are also considered as variables of particular system specifications, *i.e.* where and how it is installed, in addition to its composition in terms of a set of fixed and changeable attributes and components.

There is a consensus to define data quality, which can be characterized according to some key imperfections as follows [23]. Data are **erroneous** when values are different from the true data. Data are **incomplete** when not totally supplied. Data are **imprecise** when denoted as a set of possible values, among which the real value can be found, but without knowing how. Data are **uncertain** when values cannot be stated with absolute confidence. Data are **unavailable** when the system cannot obtain a value because of its limitations or due to missing measurements.

On the other hand, although widely studied, works on information quality have not reached until now, a general agreement about its definition. In the absence of a global consensus on basic methodological elements to measure information quality, we concentrate on few known works to define an adapted approach for sensor systems [24]. Information quality has been previously analyzed in other domains

like Management Information Systems [25], Web Information Systems [26], and Information Fusion Systems [27]. Even if none of the previously defined approaches can be directly and completely applied to sensor systems in a specific context, these works define an ensemble of suitable information quality aspects that can be examined. Hence, some can be adapted to quality categorization studies. On the other hand, information quality dimensions are conventionally classified in four groups: intrinsic, contextual, representational, and accessibility. Additionally, in our case it is necessary to identify the source of anomalies. We categorize information dimensions according to the assessed entity: sensor, measure, or system. This will allow identifying the anomaly source depending on the concerned dimensions.

Also, existing approaches are not adapted to cyber-physical systems in general and particularly to sensor systems, because humans are considered just as data consumers. However, regardless of the existence of a wide range of automatic processes, cyber-physical networks relate to humans as decision-makers. Furthermore, even if information is part of different tasks, quality evaluation can be completely independent of those tasks.

On the other hand, existing anomaly detection methodologies focus on very specific systems to identify a subset of behaviors. For instance, timeliness or coherence are not considered as signs of anomalies, as it could be the case when a network is attacked, or when a sensor is not calibrated [9], [28]. Moreover, although it is possible to detect multiple anomalies, no complementary information is processed after the detection [9], [10], [11], [12], [13], [28]. In none of these approaches, data and information quality are neither part of the detection approach, nor used later to understand anomalies.

Compared to previous works, the proposed approach does not aim at discarding data or information when the corresponding estimated quality is found to be poor. Instead, its main objective is to categorize meaningful quality variations, once anomalies have been detected. This feature informs decision-makers about detected anomalies, including the impacted data and information quality dimensions.

III. Categorization of Quality Measures

The proposed approach searches to categorize in a detailed manner detected anomalies, according to a suitable strategy. In this section the main components of it are described, along with anomaly types categorized depending on source location and inducing factors.

A. Quality Assessment

Despite their inadequacy if applied directly to a sensor system, some elements of the previously summarized works about data and information quality definition can be revised to define a quality categorization model. Specifically, we make use of Total Information Quality Management intended to improve business data warehousing and raise benefits [29]; and a methodology for Information Quality Assessment that evaluated and benchmarked information quality [30]. It is important to note that although these two studies handle information quality, the question about anomalies detection in cyber-physical systems is not adressed.

Based on the studies presented in Section II-B, the evaluation of data and information quality is defined for each imperfection and dimension, respectively. Even if multiple data imperfections and information quality dimensions could be assessed in a given case, not all are necessarily mandatory, while others are inappropriate. In order to define the searched categorization, data quality evaluations are represented in a multidimensional vector called DQV (Data Quality Vector). For a case where N imperfections (I_i) are evaluated, DQV is defined as:

$$DQV = \{I_1, I_2...I_N\} \tag{2}$$

In the same manner, for information quality, IQV (Information Quality Vector) is defined for M dimensions (D_j) as:

$$IQV = \{D_1, D_2...D_M\} \tag{3}$$

When a categorization is carried out, the specific data imperfections and data quality dimensions are defined in advance, taking into account the variety of possible values ranges, units, and operational conditions. In the rest of the paper, DQV and IQV, are considered as a result of the initial measure of quality .

B. Detection of Anomalies

Two main modules compose the defined approach (Figure 1): anomaly detection and definition of agreement levels. Automatic quality measures of data and extracted information streams are made first, before comparing values to expected characteristics. The study of dimensions like source precision, confidence, coherence, or timeliness, additional data provided by sensors and actuators are required. In either case, multiple values are assigned to vectors DQV and IQV to be filtered by the Agreement levels (ALs) module, on which anomalies are detected. ALs are pre-defined quality threshold values or intervals that the elements of DQV and IQV must comply with to be considered as normal. Otherwise, non-compliant elements are considered as anomalous. Afterwards, anomalous quality measures are categorized depending on the identified data imperfection or information dimension.

ALs are separately defined for the different states of each system module, to represent what is considered to be the expected or normal behavior of multiple components that produce data streams, like sensors, network, and humans interacting with the system. All ALs are determined according to the corresponding identified data imperfections and information dimensions. Since the ALs calculation is semi-automatic, expert support is needed to make the set ups. To this end, data and information streams stocked in quarantine are manually identified as normal or anomalous. Normal data can be used to set ALs automatically. If values show a Gaussian distribution, the ALs can be defined based on 68-95-99.7 rule. Further manual verification is necessary to define ALs representing anomalous entities, avoiding the assignment of several anomaly types to a given AL, in order to decrease the probability of false detections.

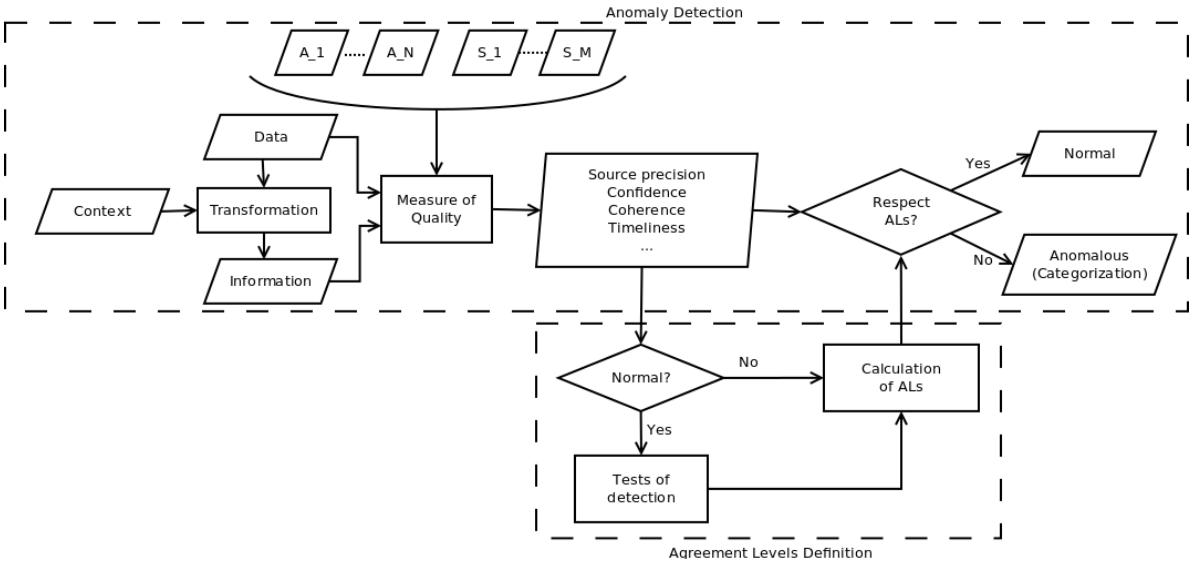

Fig. 1. Proposed method for anomaly detection on cyber-physical systems.

C. Anomaly sources and inducing factors

Anomalies have multiple characteristics that allow building corresponding classifications [7]. Based on cyber-physical system behavior, *i.e.* networked sensors, synchronous and asynchronous measures, sensors deterioration, types of anomalies, and error cost, we consider that two anomalies classifications are suitable. A new categorization is introduced according to the anomaly origin, namely: Sensor (S), Measure (M), or Network (N). An anomaly is also categorized depending on who or what produced it. Thus, we propose a new categorization of anomaly types according to an inducing factor represented in Table I.

TABLE I. CATEGORIZATION OF ANOMALY TYPES ACCORDING TO AN INDUCING FACTOR.

ID	Description
1	Natural deterioration
1.a	Fault / Damage detection
1.b	External factor
2	Caused deterioration
2.a	Sabotage (physical access)
2.b	External attack

This second categorization makes the difference between natural factors, intrinsic or environmental (which deteriorate sensors progressively), and provoked anomalies like network attacks. Meaningful detected anomalies can be therefore properly detailed within a structured scheme, expected to allow giving a faster and more pertinent response.

IV. CASE STUDY

Ships represent a strategic infrastructure for international commerce and military activity. While nearly 9.6 billion tons were delivered by cargos during 2013 [31], more than 50% of container ships transport over 5000 twenty-foot equivalent units around the world. Vessel governance is thus an example of strong synergy between automatic responses and decision-aid: numerous sub-systems that produce voluminous data streams, support crew decisions concerning the identification of efficient and secure routes.

To attain this objective, on board vessel sensor systems provide navigation and vessel monitoring data and information on a permanent basis, having external access to be monitored and controlled from distant computers. Moreover, as in other domains, naval cyber-physical systems are developed constantly to optimize functionality, improving performance and simplifying systems' use. Nevertheless, maritime vessels are highly sensitive to sensor system anomalies, which imply that decision-making based on wrongly understood anomalies can be potentially catastrophic.

SCADA systems installed on naval vessels usually present multiple anomalies, due to natural and external factors. Anomalies are provoked among others, by the hostile maritime environment that deteriorates cyber-physical systems' components prematurely, maintenance deficiencies, and negligent human intervention. Given the considerable increase in modern ships of sensors data requiring automatic and assisted responses, cyber-physical systems should be capable of correctly discard anomalous data and correct data that can be recovered, to prevent wrong operational decisions. Among multiple tanks on a vessel that stock different liquids, including to ensure weight balance, a particularly critical system is fuel storage.

To test the proposed quality-based anomaly categorization approach, a comparable and slightly simplified training platform of two fuel tanks was used (Figure 2). Two tanks compose this platform: a main tank and a secondary tank. While the secondary tank fills the main tank using the pump A_1, the main tank (smaller than the secondary) provides the liquid to other sub-systems, like the fuel to vessel engines. A valve placed at the bottom simulates this consumption. To simulate when the vessel is refilling the tanks, the secondary tank can be filled with the pump A_2 from an external source. In order to predict vessel autonomy, two types of sensors generate data: four discrete sensors (S_1-S_4) and one ultrasonic sensor (S_5) placed on the secondary and main tanks, respectively. Data provided by the four discrete sensors are multiplexed in a register. All these sensors and actuators are connected to a remotely controlled IP network, accessible through Modbus

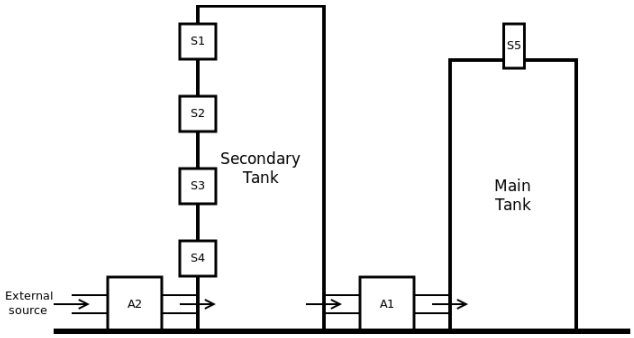

Fig. 2. Scheme of the training platform that simulates a vessel fuel system of two tanks.

sentences (protocol largely used in SCADA systems).

A. Quality Evaluation and Agreement Levels

A control and monitoring computer performs system analysis. The system, that is connected to the platform network, receives IP packets containing Modbus sentences. Binary sequences used by communication protocols are not directly comprehensible and hence called **data**. **Information** is defined when a protocol decodes the data and the context gives a meaning to values, making use of the protocol' specifications, the sensors' data specifications and tanks' size. Resulting information is the measured level of liquid, represented by a comprehensible value. To obtain this information a measure of distance generated by the ultrasonic sensor is transformed to a measure of volume, according to the tank dimensions.

For the case study, two data imperfections and eight information dimensions are considered as the most representative quality variables. The chosen data imperfections are **Erroneous** (I_{err}) that indicates if the CRC (Cyclic Redundant Code) is verified and **incomplete** and (I_{inc}) that detects if all the data fields are supplied. The used information quality dimensions are described as follows:

- **Source precision** (D_{sp}) and **real precision** (D_{rp}) are the noise produced by the sensor and by the measure, respectively.
- **Confidence** (D_{con}) is the number of anomalies detected in the past.
- **Erroneous** (D_{err}) quantifies if information is a possible value for the system.
- **Timeliness** (D_{tim}) is the difference between arrival times.
- **Coherence** (D_{coh}) is the difference between a theoretical behavior and measured valued.
- **Incomplete** (D_{inc}) verifies that all information fields are filled.
- **Uniqueness** (D_{uni}) verifies that information is unique.

In this manner, DQV and IQV are defined for the case study as:

$$DQV = \{I_{err}, I_{inc}\} \tag{4}$$

$$IQV = \{D_{sp}, D_{rp}, D_{con}, D_{err}, D_{rp}, D_{tim}, D_{coh}, D_{inc}, D_{uni}\} \tag{5}$$

I_{sp} and I_{rp} are obtained filtering the signal produced by S_5 with a high pass filter. Additionally, when noise is represented along with local values, it is also filtered to show its global trend. This two dimensions are evaluated together, except when one of the tanks is empty and the liquid does not produce waves.

According to operational experience, states of normality are defined for quality analysis and any state of normality can be restricted by the ALs. Sensor cycles are logged to identify these limits. An evidence of anomaly can be presumed when measures do not comply with those AL, which are defined depending on different inputs. Being at the methodology definition stage, such limits are for the moment manually defined.

In the described case study multiple ALs have been identified. An example of the AL used for the filtered noise, which changes depending on the measured level x, is represented as:

$$AL_1(x) = \begin{cases} D_{sp} + D_{rp} < 0 & \text{when } x = 0 \\ D_{sp} + D_{rp} < 25 & \text{when } 0 < x < 3000 \\ D_{sp} + D_{rp} < 8 & \text{when } 3000 < x \end{cases} \tag{6}$$

Once a set of limits is defined for all data quality imperfections and information quality dimensions, the impact of different anomalies is studied. The goal is to detect if anomalies can be correctly detected and categorized, as well as to find out if initial limits should be more restrictive to improve detection accuracy. This analysis can lead to identify additional limits or modify the existing ones.

B. Anomalies Analysis Based on Quality

A set of 13 potential anomalies has been identified and studied in order to determine their impact on quality elements. A description of each anomaly and its detected impact is presented in this section. Examples show analogical data provided by S_5, converted to digital values in 10000 uniformed steps. This sensor needs a calibration to adjust its range of measure. Additionally, the period between samples in log files is set to 0.1s for all components.

Environmental: Sometimes, environmental reasons generate anomalous sensor measurements. These anomalies are caused by natural effects like dust on sensors or capacitive charges, and usually appear as outlier values with respect to neighboring values. In Figure 3, an environmental anomaly example is depicted for S_5, showing its impact on filtered noise. Since this noise is measured as the addition of D_{sp} and D_{rp}, it is not feasible to determine exactly if it is caused by internal or external factors.

These anomalies appear momentarily and can be discarded therefore without significant consequences for the system. Nevertheless, whenever the number of these anomalies increases in a given period of time, a detection alarm should be activated to give a pertinent response. This is managed by D_{con} that decreases with each detected anomaly.

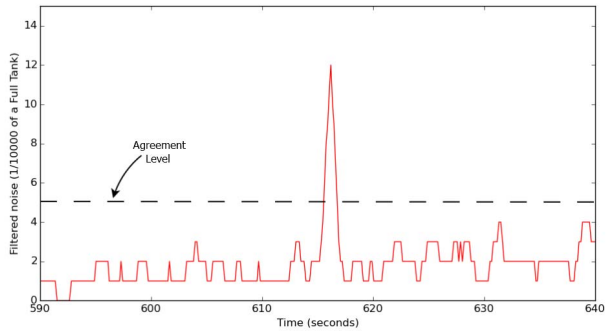

Fig. 3. S_5' environmental anomaly impact on filtered noise.

Wet sensor: Because of surrounding evaporation, moisture can later accumulate on the surface of a sensor. Nonetheless, small drops condensed over the ultra-sounder surface do have a significant impact on measurements. Moisture on the sensor doubles D_{sp} values with respect to a dry sensor situation. Since source precision values are very small, the detection of a wet sensor is only possible when the tank is empty and D_{rp} becomes zero.

Bad connection: Sometimes, network or supply wires are disconnected or do not have a correct connection. This problem evidently produces system anomalies, having a direct impact on D_{tim} when it produces loss of data packages or delays. Whenever connections do not produce data packages loss but are sources of error, a bad connection quality is detected by I_{err}. Depending on its impact, this anomaly can be relatively simple to detect, resulting in data preservation when it is detected early.

Bad calibration: One of the most frequently produced anomalies is a bad calibration of the system. Given that the ultra-sounder needs a calibration to function properly, a bad calibration can produce liquid overflow or leave the vessel adrift, because of wrong autonomy estimation. An inadequate calibration of the ultra-sounder can be produced by wrong user manipulations of mechanical vibrations.

If the measured level is higher than the corresponding overflow reference, although it is an unlikely event, it becomes possible when the overflow slot is blocked. Also, a measured liquid level lower than an empty tank, although physically impossible, can be provoked by a wrong or non-existent calibration and therefore impacts D_{err}.

Damaged sensor: A damaged sensor can have unpredictable behavior, so its probable impact depends on the damage degree. As a consequence, all imperfections and dimensions are potentially modifiable. Every anomaly of the ultra-sounder can be considered then as a damaged sensor, given the resulting irregular behavior. On the other hand, damage of a discrete sensor like the ones of the secondary tank, can be detected in a rather simple manner.

Blocked measure: To simulate a blocked measure attack in the ultra-sounder sensor, a sheet of paper is used to block it, producing a constant value. When this sabotage is produced, a variation in the filtered noise is detected. Since the surface of a paper sheet does not present waves and the reflection of the ultrasound pulse is better in paper than water, the measured values for D_{rp} and D_{sp} are almost zero. An example is shown in Figure 4. This anomaly can be detected with an AL that fixes a minimum for filtered noise. Depending on the data used to create ALs, the respective AL can be set manually.

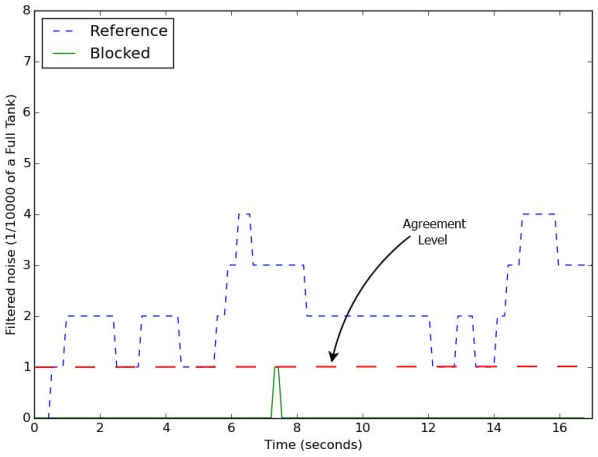

Fig. 4. Filtered noise of S_5 when the measure is blocked.

Liquid replacement: Due to accidents, errors, or sabotage, other liquids can be introduced in the tanks. Even if sensor measures are done exactly in the same way as for a normal case, an impact on quality can be observed in the first tank. However, anomalies were not detected in the secondary tank, except when the new liquid produces other anomalies on its own.

Since the reflection surface of the ultrasound pulses is different, it can produce quality modifications. The reflected pulse could be less readable and then produce an impact on the D_{sp}. Otherwise, a change of viscosity modifies the size of surface waves. Different levels on the D_{rp} can identify this variation. A bigger viscosity produces lower values of real precision, while lower viscosity produces higher ones. Ease of detection varies as a function of the difference between the original liquid and the replacement.

Solid foreign objects: When this anomaly is produced in the secondary tank, it would only have an impact if a foreign object produces another anomaly. When these objects are present in the main tank, several quality measurement variations can be imagined. After foreign objects appear in the main tank, an impact can be observed on D_{rp} and in a lesser extent on D_{sp}. Foreign objects change the movements of the liquid's surface and the ultrasound pulse reflects differently.

To simulate the varying effects of these objects, different amounts of foam squares were introduced in the tank. The impact of this anomaly is represented by resulting filtered noise in Figure 5. A reference measure shows when the system works normally. Three ALs values are defined for each functioning state of the system. Initially, the pump does not work, the tank is empty, and the noise is considerable higher than normal. Afterwards, when the pump produces perturbations because the level of liquid is lower than its expected level, the impact is important for seven objects and reduced but easily detectable for two objects. Once the pump is covered, the perturbation diminishes, but it is always detectable.

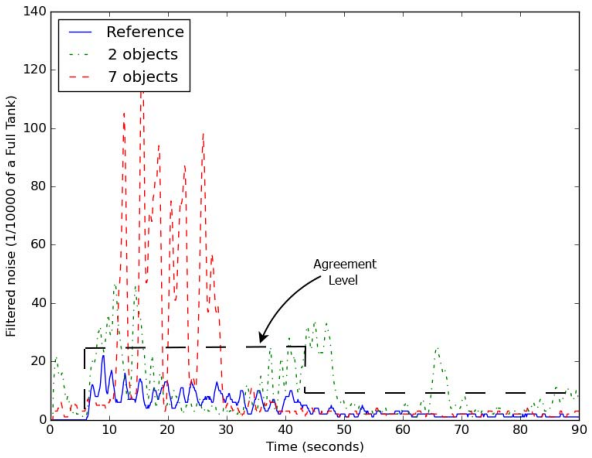

Fig. 5. Impact on S_5' filtered noise produced by solid foreign objects.

As seen in Figure 5, the identified ALs are appropriate to detect this anomaly, but can be improved if more functioning states are defined.

Hits on the tanks: Several material attacks can be detected by variations in quality, too. One of the easiest material attacks to detect is whenever a tank is hit to produce liquid leaks or to block the fuel supply. Consequently, hits could have an identifiable impact on quality measures.

A hit on the tanks produces measures anomalies identified as perturbations in D_{rp}. The momentarily or occasional activation of a discrete sensor produced by a hit, causes erroneous values and impossible transitions, affecting D_{err} and D_{coh}.

Leak: Another fatal scenario is a leak in a tank. Depending on the leak volume, its impact on quality dimensions can vary. The scenario is tested using two valves, one on each tank. The leaking volume per second changes according to the volume of liquid contained in the tank. Identified impacts on quality concern the D_{coh} and D_{rp}.

Assuming that the measured liquid level is correct, D_{coh}, *i.e.* the difference between the current level and the theoretical level, can be an indication of the ratio between the leak volume and its position. When the leak is lower than the liquid level, its position can be estimated when the detected effect has disappeared.

Activated or blocked sensors: Discrete floating sensors of the secondary tank can be accidentally or intentionally activated or deactivated, indicating for instance the presence of a given fuel level that is considerably lower or higher that the real level. This anomaly excludes anomalies produced by damaged sensors.

Due to the discrete sensors configuration (Figure 2), only five activation values are possible: 0000, 0001, 0011, 0111, and 1111. These values indicate four bidirectional ordered transitions of the sensors states, from an empty to a full tank. Any value different from these five, or a modification in the transition sequence of sensors' states affects D_{err} and D_{coh}. Transitions are important too, since two discrete sensors can be activated at the same time providing a correct state, which is physically incoherent and therefore anomalous. In that case, the measured values are correct but not the transition and then D_{tim} is altered.

Denial-of-Service: Denial-of-Service (DoS) anomalies are quite common attacks, consisting basically on an overflow of network resources to block the system. Currently, due to the use of standard computers, SCADA naval systems can be easily infected by a virus or attacked using external connections. As a consequence, every component of the system could perform a DoS attack.

Imperfection of data sequences affects I_{inc}, because when the network is blocked some packets are discarded. But the most crucial information dimension to detect these attacks is $Dtim$, which is impacted by transmission delays.

In Figure 6, the impact of a DoS attack on D_{tim} is shown. After $10s$ that correspond to more or less 90 samples, a DoS attack is executed. Firstly, some small perturbations are detected but around the 160^{th} sample, the network is completely blocked. When the attack is stopped, the network recovers its normal operative state, but still with some perturbations that gradually disappears through time.

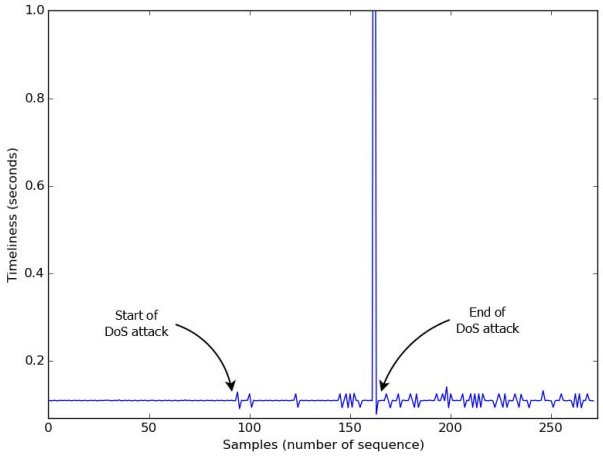

Fig. 6. Timeliness of the main tank.

Spoofing: Spoofing anomalies consist on injecting fake data in the sensor system. These injected data can be generated according to specific knowledge about the concerned system or based on past records, as Stuxnet did [32].

If both systems, the legitimate system and the fake system are working simultaneously, the most relevant effect is produced on D_{tim} because it is reduced, also in D_{uni} when information is repeated, and in I_{err} when the sequence ID is altered. Other variations of quality measures can be detected, depending system components ignored by the attacker. For example, if the illegitimate injected information is a constant value, D_{sp} and D_{rp} are zero, hence aberrant. Otherwise, if the whole system is spoofed, another way to detect this attack is looking at the coherence between sub-systems.

C. Results

Numerous experiments were carried out to characterize the impact of anomalies on quality components, *i.e.* data imperfections and information dimensions, adjusting operational

limits to refine evaluations. A summary of analyzed vessel fuel tank platform anomalies categorization is described for the ultrasound sensor that monitors the main tank (Table II), the discrete sensors that monitor the secondary tank (Table III), and the network (Table IV). These three tables represent the anomaly reason (first column), relating it to the anomaly origin and anomaly types introduced in Section III-C (second and third columns, respectively). The ease of detection estimated based on observations (fourth column) is included, along with the expected error cost (fifth column) and directly affected data and information quality analysis components (ten last columns respectively). Columns are marked with 'x' when the quality component is always affected and '(x)' when it can be potentially affected.

Described tables permit to understand how each anomaly can impact different data quality imperfections and information quality dimensions. The proposed analysis allows categorizing anomalies to assist decision-makers, by providing complementary information. These tables also indicate that the relative importance and pertinence of imperfections and dimensions, vary depending on the studied system.

V. Discussion and Conclusion

Error costs associated to anomalies vary from fake measures to a vessel left adrift and risk of fire, including damaged material, lost information, and wrong decisions. Anomalies in sensors networks may put at risk personnel, infrastructure, economic activity, and systems security.

An approach to caracterize anomalies that produce meaningful data and information quality variations has been proposed. It is based on definition of agreement levels, two categorizations of anomalies -according to the source and inducing factors-, as well as the selective identification of data quality imperfections and information quality dimensions.

Experimental tests were conducted to identify anomalies' reasons in the sensors system of a two fuel tanks training platform. Results show that the impact on quality elements allows categorizing quality measurements. Found anomalies are related to their respective origin, factor, occurrence probability, ease of detection, error cost, and affected quality components. Obtained results indicate that the ultrasound sensor of the main tank is the most vulnerable component of the system from a quality analysis perspective, since it is prone to ten potential anomalies, compared to five for the discrete sensors of the secondary thank, and two for the network.

Out of 13 identified anomalies, only four (bad connection, damaged sensor, hits on the tanks, and leak) are common to the sensors of the main and secondary tanks. Besides confirming the vulnerability of the ultrasound sensor, this fact illustrates that anomalies causes in the sensing and network contexts are different and should be studied independently. From a global point of view, measurement anomalies are predominant with respect to network and sensor anomalies, being likely induced particularly by external factors, or a combination of external and natural factors. Finally, affected quality components are mostly associated to particular anomalies that allow making a first categorization of a detection. However, all data and information quality elements can be affected when the anomaly is a damaged sensor or a spoofing attack.

Obtained preliminary results show that an approach based on variations of quality measures is likely to implement anomaly categorization. The main reason is that analysis elements are independent of data and information types. Further works will concern the total automatic definition of agreement levels and anomalies propagation trough different systems, considering their respective individual and combined impact.

References

[1] J. Petit and S. Shladover, "Potential cyberattacks on automated vehicles," in *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–11, 2014.

[2] J. Volpe, "Vulnerability assessment of the transportation infrastructure relying on the global positioning system," Office of the Assistant Secretary for Transportation Policy U. S. Department of Transportation, Tech. Rep., 2001.

[3] A. Grant, P. Williams, N. Ward, and S. Basker, "Gps jamming and the impact on maritime navigation," *Journal of Navigation*, vol. 62, no. 02, pp. 173–187, 2009.

[4] M. Balduzzi, A. Pasta, and K. Wilhoit, "A security evaluation of ais automated identification system," in *Proceedings of the 30th Annual Computer Security Applications Conference*. ACM, 2014, pp. 436–445.

[5] C. Iphar, A. Napoli, and C. Ray, "Detection of false ais messages for the improvement of maritime situational awareness," in *Oceans' 2015*, 2015.

[6] C. Iphar, A. Napoli, C. Ray, E. Alincourt, and D. Brosset, "Risk analysis of falsified automatic identification system for the improvement of maritime traffic safety," in *Proceedings of the European Safety and Reliability Conference (ESREL)*, 2016.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[8] I. Garitano, R. Uribeetxeberria, and U. Zurutuza, "A review of scada anomaly detection systems," in *Proceedings of the 6th International Conference SOCO 2011 on Soft Computing Models in Industrial and Environmental Applications, .* Springer, 2011, pp. 357–366.

[9] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006, pp. 187–198.

[10] J. Rrushi and R. Campbell, "Detecting cyber attacks on nuclear power plants," in *Critical Infrastructure Protection II*. Springer, 2008, pp. 41–54.

[11] E. Horvitz and M. Barry, "Display of information for time-critical decision making," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 296–305.

[12] R. Fujimaki, T. Yairi, and K. Machida, "An approach to spacecraft anomaly detection problem using kernel feature space," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 401–410.

[13] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115–123.

[14] T. E. Humphreys, B. M. Ledvina, M. L. Psiaki, B. W. O´Hanlon, and P. M. Kintner Jr, "Assessing the spoofing threat: Development of a portable gps civilian spoofer," in *Proceedings of the ION GNSS international technical meeting of the satellite division*, vol. 55, 2008, pp. 56–68.

[15] A. Perrig, R. Szewczyk, J. D. Tygar, V. Wen, and D. E. Culler, "Spins: Security protocols for sensor networks," *Wireless networks*, vol. 8, no. 5, pp. 521–534, 2002.

[16] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[17] X. Li, X. Liang, R. Lu, X. Shen, X. Lin, and H. Zhu, "Securing smart grid: cyber attacks, countermeasures, and challenges," *IEEE Communications Magazine*, vol. 50, no. 8, pp. 38–45, 2012.

TABLE II. CATEGORIZATION OF IDENTIFIED ANOMALIES IN THE ULTRASOUND SENSOR

Reason	**Type: Origin**	**Type: Factor**	**Ease of detection**	**Error-cost**	I_{err}	I_{inc}	D_{sp}	D_{rp}	D_{con}	D_{err}	D_{tim}	D_{coh}	D_{inc}	D_{uni}
Environmental	M-S	1.b	It depends (on effect)	It depends (on damage degree)			x	x	x					
Wet sensor	M-S	1.b	Difficult	Lower time of life			x							
Bad connection	N	1.b/2.a	Easy	Lost of information	x						x			
Bad calibration	N	1.a/2.a	It depends (on measure)	Wrong information						x				
Damaged sensor	S	1	It depends (on damages)	Lost of information - Fake measure	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)
Blocked measure	M	2.a	Easy	Lost of information - Fake measure			x	x						
Liquid replacement	M	2.a	It depends (on liquid)	Vessel adrift - Damaged components			x	x						
Solid foreign objects	M	1.b/2.a	It depends (on object)	Wrong lectures - Damaged components			x	x						
Hits on the tanks	M	2.a	Easy	Damaged tanks				x		x		x		
Leak	M	1/2.a	It depends (on size)	Vessel adrift and risk of fire				x				x		

TABLE III. CATEGORIZATION OF IDENTIFIED ANOMALIES IN THE DISCRETE SENSORS

Reason	**Type: Origin**	**Type: Factor**	**Ease of detection**	**Error-cost**	I_{err}	I_{inc}	D_{sp}	D_{rp}	D_{con}	D_{err}	D_{tim}	D_{coh}	D_{inc}	D_{uni}
Bad connection	N	1/2.a	Easy	Lost of information	x						x			
Damaged sensor	S	1.a/2.a	It depends (on damages)	Lost of information - Fake measure	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)
Activated or blocked sensors	M	2.a	It depends (on measure)	Decisions based on fake information						x		x		
Hits on the tanks	M	2.a	It depends (on strength)	Damaged tanks				x		x		x		
Leak	M	1/2.a	It depends (on size)	Vessel adrift and risk of fire				x				x		

TABLE IV. CATEGORIZATION OF IDENTIFIED NETWORK ANOMALIES

Reason	**Type: Origin**	**Type: Factor**	**Ease of detection**	**Error-cost**	I_{err}	I_{inc}	D_{sp}	D_{rp}	D_{con}	D_{err}	D_{tim}	D_{coh}	D_{inc}	D_{uni}
DoS	N	2.b	Easy	Lost of information		x					x			
Spoofing	N	2.b	It depends (on complexity)	Decisions based on fake information	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)

[18] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[19] C. Alcaraz, R. Roman, P. Najera, and J. Lopez, "Security of industrial sensor network-based remote substations in the context of the internet of things," *Ad Hoc Networks*, vol. 11, no. 3, pp. 1091–1104, 2013.

[20] J. Raiyn *et al.*, "A survey of cyber attack detection strategies," *International Journal of Security and Its Applications*, vol. 8, no. 1, pp. 247–256, 2014.

[21] C. Zins, "Conceptual approaches for defining data, information, and knowledge," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 4, pp. 479–493, 2007. [Online]. Available: http://dx.doi.org/10.1002/asi.20508

[22] M. Zeleny, *Human Systems Management: Integrating Knowledge, Management and Systems*, W. Scientific, Ed. World Scientific Publishing Co. Pte. Ltd., 2005.

[23] A. Motro and P. Smets, *Uncertainty management in information systems: from needs to solutions*, A. Motro and P. Smets, Eds. Springer Science & Business Media, 1996.

[24] P. Merino Laso, D. Brosset, and J. Puentes, "Monitoring Approach of Cyber-physical Systems by Quality Measures," in *Proceedings of International Conference on Sensor Systems and Software*, 2016, pp. 1–12.

[25] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 3 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=1189570.1189572

[26] F. Naumann, "From databases to information systems-information quality makes the difference." in *IQ*, 2001, pp. 244–260.

[27] I.-G. Todoran, L. Lecornu, A. Khenchaf, and J.-M. Le Caillec, "Information quality evaluation in fusion systems," in *Proceedings of 16th International Conference on Information Fusion (FUSION), 2013* , 7 2013, pp. 906–913.

[28] L. Wei, N. Kumar, V. N. Lolla, E. J. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Assumption-free anomaly detection in time series." in *SSDBM*, vol. 5, 2005, pp. 237–242.

[29] L. P. English, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, R. M. Elliot, Ed. New York, NY, USA: John Wiley & Sons, Inc., 1999.

[30] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "Aimq: A methodology for information quality assessment," *Inf. Manage.*, vol. 40, no. 2, pp. 133–146, 12 2002. [Online]. Available: http://dx.doi.org/10.1016/S0378-7206(02)00043-5

[31] CNUCED, "Review of maritime transport 2014," United Nations, Tech. Rep., 2014.

[32] N. Falliere, L. O. Murchu, and E. Chien, "W32. stuxnet dossier," *White paper, Symantec Corp., Security Response*, vol. 5, p. 69, 2011.