Computers & Security 88 (2020) 101609

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Improving information security risk analysis by including threat-occurrence predictive models

Pedro Tubío Figueira, Cristina López Bravo*, José Luis Rivas López

Rúa Maxwell s/n, Escola de Enxeñaría de Telecomunicación, Universidade de Vigo, Vigo 36310, Spain

ARTICLE INFO

Article history:
Received 20 February 2019
Revised 29 July 2019
Accepted 7 September 2019
Available online 9 September 2019

Keywords:
Information security
Risk analysis
SVM regression
Logistic regression
Predictive models
Magerit
Secitor

ABSTRACT

Protecting information is a crucial issue in today society, in both work and home environments. Over the years, different tools and technologies have contributed to safeguarding information, including risk analysis methodologies developed to evaluate the risk of threat materialization despite security measures. Traditional risk analysis methodologies base risk computation on, among other parameters, the frequency of occurrence of threats, which is gathered from available historical data. However, as new safeguards are implemented, and vulnerability potential changes, threat frequencies may also change.

To take into account the current state of an organization's system as well as historical data, we propose to substitute past threat frequency by the probability of a threat occurring in the future. To compute this future threat probability, we use regression models, validated by a risk analysis for a Spanish SME based on Magerit (Spanish adaptation of ISO/IEC 27005). The results show that the future probability of each threat can be calculated with accuracy, precision, sensitivity and specificity rates above 70%.

Obtaining a more realistic risk estimate (reflecting to the current state of vulnerabilities) is translated into the adoption of better and more efficient safeguards that reduce losses and improve information security in a business.

1. Introduction

The information and communication technologies (ICTs) are essential resources for our society nowadays, and, with vast amounts of data saved or sent through the Internet daily, protection becomes a priority. This issue concerns all kinds of organizations, including the home and work environments, where information is crucial to the proper development of business activities. Files with confidential information, and the media where they are stored or through which they are sent are critical points for the safeguarding of assets.

Different information security risk analysis methodologies have been developed to study and evaluate the security measures used to protect data and how different events could affect information assurance (Fredriksen et al., 2002; Peltier, 2010; Shameli-Sendi et al., 2016; Suh and Han, 2003; Yazar, 2002). Traditional methodologies base their risk calculations on historical data, using threat-occurrence frequency as one of the input parameters. However, as new safeguards are implemented and the vulnerability potential changes, previously frequent threats may cease to be so. Hence, an interesting approach would be to explore the use of predictive algorithms to estimate threats' frequency (and, hence, risk levels), i.e., to focus on what could happen in the future rather than review what has happened in the past.

The main goal of this work is to include a threat-occurrence predictive module in the risk analysis process that takes into account the current state of the system- in particular, the current state of vulnerabilities affecting the system- in order to improve risk computation, and so identify the most critical risks. The aim is to develop better and more efficient safeguards that can reduce losses to businesses by improving information security, once the most risky assets are identified.

The rest of the paper is organized as follows: in Section 2 we discuss the background of this research, in Sections 3 and 4 we describe and evaluate the proposed risk analysis method, and finally, in Section 5 we conclude the paper.

* Corresponding author.
E-mail addresses: ptubio@gti.uvigo.es (P. Tubío Figueira), clbravo@gti.uvigo.es (C. López Bravo), jlrivas@secitor.es (J.L. Rivas López).

https://doi.org/10.1016/j.cose.2019.101609
0167-4048/

2. Related works

Over the years, in the information security risk analysis field, multiple studies have been undertaken with different approaches and objectives, but with the main goal of providing some kind of information about the risks that could impact on an organization's assets. The different studies have focused on specific contexts, such as the location of vulnerability propagation paths (Feng et al., 2014), data acquisitions systems (Cherdantseva et al., 2016), legislative requirements and limitations (Massaccia et al., 2005), and operational continuity (Suh and Han, 2003), among others. In addition, given its importance, risk analysis is supported by security-related international standards and national guides such as ISO-27005 (International Organization for Standarization, 2008), Mehari (Mehari, 2007) CRAMM (Yazar, 2002), NIST 800-30 (NIST and SP800-30, 2002), and Magerit (MAGERIT V.3, 2014) along with the corresponding software tools, such as PILAR (Procedimiento Informático-Lógico para el Análisis de Riesgos, 2019) and CRAMM-Manager (Yazar, 2002).

CORAS (Coras.sourceforge.net., 2018; Fredriksen et al., 2002), one of the first information security risk analysis projects, was based on a step-by-step philosophy that begins with a definition of the analysis objectives and ends with a treatment of the risks. Following the same philosophy as CORAS is ISRAM (Karabacak and Sogukpinar, 2005); among other novelties, ISRAM introduced managers and staff in the risk analysis process through the conduct of surveys whose results were analyzed so that, through different mathematical computations, the cost of risks was calculated. CORAS and ISRAM represent two ways of approaching the risk analysis problem: qualitatively and quantitatively. The qualitative approach uses mathematical and statistical tools to estimate the risk of materialization of a threat to which particular assets are exposed (Fredriksen et al., 2002; Rajbhandari and Snekkenes, 2013). The analysis provides information about the particular exposure level categorized as low, medium, or high. The quantitative approach (Bojanc and Jerman-Blažič, 2013; Karabacak and Sogukpinar, 2005; Yazar, 2002) yields information on cost and loss that could be incurred in reputation, logistics and economic terms from the materialization of a threat.

Both approaches have advantages, but also disadvantages (Lee and Ming-Chang, 2014; Xu and Zhao, 2011), which is why a third way is gaining prominence (MAGERIT V.3, 2014; Yazar, 2002). This new approach -a mixed qualitative-quantitative approach- aims to improve risk analysis results by exploiting the fact that the weakness of one approach is compensated for by the strengths of the other. For instance, while qualitative methods quickly and inexpensively determine the areas of greatest risk, quantitative methods yield a more accurate image of risk in those previously identified areas.

However, to take full advantage of the benefits of the mixed approach, data have to be carefully selected and processed. The data used for the application of security guidelines are usually obtained from expert knowledge, which implies a subjective component that would affect the reliability of the analysis and the precision of the results. It is necessary to mitigate this risk to ensure results as independent as possible of the expert. This can be done by endowing the methodology with additional information obtained through more objective sources and tools, or by introducing fuzzy elements in the analysis. In any case, it must be remembered that the very concept of risk implies subjectivity, which is why it is important to invest efforts in developing a methodology where risks can be measured objectively.

Several approaches can be followed to gather additional information. Online resources such as *Common Vulnerabilities and Exposures* (CVE) (Cve.mitre.org., 2018), provide information on known public vulnerabilities and threats. Known public vulnerabilities are summarized in a numerical score known as *Common Vulnerability Score System* (CVSS) (Common Vulnerability Scoring System, 2019), which reflects their principal characteristics and their severity. This numerical score can be translated into a qualitative representation (such as low, medium, high or critical) to help evaluate and prioritize vulnerabilities in vulnerability management processes derived from risk analysis. In terms of support locating vulnerabilities, commercial tools can be used, e.g., Nessus (Tenable., 2018), a vulnerability scanner that provides reliable information on an organization's state of security.

As mentioned earlier, introducing fuzzy theory is one way of reducing expert subjectivity. A qualitative extension of the Magerit methodology (Vicente et al., 2014)—based on fuzzy linguistics computational models—is a proposal that enables vague and imprecise information to be used as model input parameters when experts are not able to provide accurate values. Values for dependent relationships between assets, value of assets (losses), degradation and threat frequency are represented in the system by fuzzy linguistic labels. The results depend on the capacity of experts, based on their past experience, to assign an appropriate linguistic label to each input.

The fact that most common risk analysis methodologies are based solely on historical data that potentially undermines the obtained results, as an event that occurred in the past may not occur in the future if the vulnerability has been addressed. Moreover, no methodology takes into account black swans,[1] which result in ignoring a part of the risk that assets are exposed to. The inclusion of prediction in risk analysis process, therefore, would allow organizations to widen their knowledge of the contexts of their assets.

Therefore, any new methodology to compute risks should offer the most accurate vision possible of the state of an asset's context. In (Jindong et al., 2013) a risk prediction method based on game theory is presented. It takes into account the cost of attacks and defense, the revenues from attack and defense strategies, and the probability (calculated from a large amount of historical data) that an attack strategy succeeds when faced with a particular defense strategy. Calculating the Nash Equilibrium between attackers and defenders' strategies yields a probability vector of attackers' strategies, resulting in meaningful information that enables the best countermeasures to improve information security to be selected. This methodology is applied in a very specific assumed scenario, namely, that all threats come from an attack. However, this is not always the case, as sometimes threats come from natural disasters or from internal failures in an application. Below we describe a generic risk analysis methodology that is applicable in all scenarios.

3. Proposed information security risk analysis model

We describe how to include a prediction component in an information risk analysis methodology, specifically in Magerit (MAGERIT V.3, 2014), due to its proximity to the environment where the research was conducted. This methodology, developed by the Spanish government, is applied in the case study described in this paper. Magerit computes two types of risk: (i) potential risk, and (ii) residual risk. Potential risk is a theoretical risk that applies to situations in which no safeguards have been deployed, whereas residual risk is the risk after the implementation of safeguards.

Bellow we explain the computational basis underpinning Magerit and how to include and apply a predictive component in the risk computations.

[1] Unpredictable highly improbable events, which, when they do occur, have extraordinary and unpredictable consequences

3.1. Magerit computational model

As previously mentioned, Magerit (MAGERIT V.3, 2014) is a mixed qualitative-quantitative risk analysis methodology, which, to estimate the quantitative risk of likely damage to a system, computes potential risk (when no safeguards have been deployed) and residual risk (when safeguards have been deployed). The potential and residual risks of an asset exposed to a threat are calculated as follows:

$$\text{Potential Risk} = \text{Frequency} \times \text{Potential Impact} \quad (1)$$

$$\text{Potential Impact} = \text{Value} \times \text{Degradation} \quad (2)$$

where:

- *Frequency* represents how often the threat appears (calculated from historical data and considering how many times the threat has materialized in the evaluated period, usually a year).
- *Value* represents how important the asset is to the organization (the loss incurred if the asset is no longer available) and is assigned on a scale from 0 (not significant) to 10 (very significant) by the risk analyst (based on their expertise, knowledge of the SME's assets, vulnerabilities, and threats, and the importance of the asset to managers).
- *Degradation* is a percentage that represents the damage that a threat can cause to the asset (0% means no asset degradation, and 100% means the asset is no longer available).

$$\text{Residual Risk} = \text{Frequency} \times \text{Residual Impact} \quad (3)$$

$$\text{Residual Impact} = \text{Value} \times (\text{Degradation} \times \text{Mitigation}) \quad (4)$$

where *Mitigation* is a percentage that measures the reduction in asset degradation after safeguards are implemented.

To calculate the final risk value for a particular asset, these formulas are first applied to each threat to which the asset is exposed and then the final risk value is computed as the highest obtained risk value for all the considered threats.

As is shown in Eqs. (1) and (3), one value that directly determines risk is the frequency of occurrence of a threat. In the most common commercial tools implementing Magerit, e.g., PILAR (Procedimiento Informático-Lógico para el Análisis de Riesgos, 2019) and Secitor (Secitor.com., 2018), frequency is obtained by the security team (from the recorded incidents). Since it reflects what has happened along the year, it is partly biased by past events.

As we will see, our predictive model also uses past samples of threats to compute the risk, but instead of computing frequency directly, a regression model computes the future probability of materialization of a threat. From this future threat probability value, we calculate a new frequency that reflects the current state of the system. The regression model thus considers the current state of vulnerabilities as computed by the risk analyst.

3.2. Proposed computational model

In order to compute the risk, we replace Eqs. (1) and (3) with Eqs. (5) and (7), where the original frequency has been substituted by a new frequency (P_{TH}_Frequency) based on the probability of a threat occurring in the future:

$$\text{Potential Risk} = P_{TH}\text{\_Frequency} \times \text{Potential Impact} \quad (5)$$

$$\text{Potential Impact} = \text{Value} \times \text{Degradation} \quad (6)$$

Table 1
Equivalence between frequency and probability values.

Description	Frequency range	Probability range
Very high potential	(5, 10]	(0.70, 1]
High potential	(2, 5]	(0.55, 0.70]
Medium potential	(0.50, 2]	(0.25, 0.55]
Low potential	(0.10, 0.50]	(0.05, 0.25]
Very low potential	[0, 0.10]	[0, 0.05]

$$\text{Residual Risk} = P_{TH}\text{\_Frequency} \times \text{Residual Impact} \quad (7)$$

$$\text{Residual Impact} = \text{Value} \times (\text{Degradation} \times \text{Mitigation}) \quad (8)$$

Therefore, we first need to compute the probability of materialization of each threat, and we then need to link the calculated probabilities with P_{TH}_Frequency, in order to compute the risk.

To calculate the probabilities several machine learning solutions are available, e.g., logistic regression (Harrell and Frank, 2015), decision tree algorithms (Safavian and Landgrebe, 1991) or support vector machines (Hearst et al., 1998). We chose to use logistic regression (Harrell and Frank, 2015), and support vector regression (Hearst et al., 1998), given their advantages for this work.

Both logistic regression and SVM regression fit with the objectives of the work scenario (i.e., modeling the probability of occurrence of a variable from a set of independent values). In this case, the dependent variable will be threats, and the independent variables will be vulnerabilities. Thus, each model will represent the behavior of a particular threat.

Thanks to its simplicity, logistic regression allows us an initial rapid approximation to the problem. In addition, a solution based on logistic regression would allow non-machine learning experts (if it is the case of a company securit analylist) to use the proposed solution, since no parameters have to be tuned.

SVM regression allows more complex relationships between vulnerabilities and threats to be captured, and also enables a focus on the most important vulnerabilities.

We established equivalence between frequencies and probabilities using a qualitative scale. The scale represented the potential of a threat, varying from "very high potential" to "very low potential". Then, each level of the scale was associated with a range of frequencies and with a range of probabilities, getting them linked. For instance, the level "very high potential" was associated with very high frequently threats (those that happened between 5 to 10 times in the analyzed period) and with very high probability of materialization threats (those between 0.7 and 1). The whole equivalence is described in Table 1.

The frequencies ranges are defined by the security analyst during the analysis of the system. Steps to define the probability ranges are as follows:

- For each threat we collect a set of samples (as described in Section 4.1). Those samples include a feature called `HAPPENED` that indicates whether or not the threat is active.
- For each threat we estimate the probability from the relative frequency of the event "`HAPPENED=1`" in the corresponding dataset (P_{TH}). That is, P_{TH} will be given by:

$$P_{TH} = f_{TH} \pm Z_{\alpha/2}\sqrt{\frac{f_{TH}(1-f_{TH})}{N}} \quad (9)$$

where
- f_{TH} is the relative frequency of event "`HAPPENED=1`",
- $1-\alpha$ is the confidence interval,
- $Z_{\alpha/2}$ is the value of the normal distribution at $\alpha/2$, and
- N is the number of samples.

Table 2
Equations to calculate equivalence between probability and frequency values. x represents the probability of a threat and y represents the equivalent frequency.

Description	Slope formula
Very high potential	$y = (5/0.3)x - 20/3$
High potential	$y = (3/0.15)x - 9$
Medium potential	$y = (1.5/0.3)x - 0.75$
Low potential	$y = 2x$
Very low potential	$y = 2x$

Table 3
Overview of the SME analyzed in the case study.

SME characteristics	
Founded	1999
Number of sites	2
Number of employees	27
Number of user equipment	27
Third party relationships	Yes
Network infrastructure	Segmented networks, DMZ, VPN

Confidence intervals are set to 1-α = 0.96.

- Probability ranges are defined from the frequency ranges, starting with the set of threats with frequencies in the "very high potential" range, the lowest probability of the set determines the lower bound of the "very high potential" range of probabilities. We repeat this process with the remaining ranges.

For the calculus we considered all the threats datasets available from the analyzed SME, i. e., not only those analyzed in the case study presented in this article.

From the previous information, it was possible to obtain equations that relate frequencies and probabilities, as shown in Table 2. These formulas are the dot-slope equations of the lines described with bounds in Table 1.

Using these data and formulas, we could calculate new equivalent values for frequencies and calculate risks using the new methodology.

With this information, which implicitly includes the numeric probability of occurrence of a threat, the organization will be able to effectively focus on the most important risk because:

- They will have reliable information on the state of the assets and of potential threats that could damage them.
- They will be able to anticipate possible incidents, because they will know what threats are more likely based on the real state of the organization's systems.

4. Results

4.1. Case description

The case study of risk analysis performed using the proposed methodology was conducted at an SME in Spain. Table 3 summarizes the principal features of the SME (minimum details are provided to preserve confidentiality and anonymity). To define the scenario -identify assets and their dependencies, threats and vulnerabilities, and the corresponding relationships, as well as the safeguards in place- we conducted personal interviews with system administrators and managers, and reviewed facilities and systems documentation. The scenario was further depicted using a database containing historical data on threats. Database samples included information about the threat, evaluation values of vulnerabilities related to that specific threat, and whether the threat was active, at the moment of sampling. The evaluation of vulnerabilities represented the state of each vulnerability, once evaluated by the system's risk analyst.

Table 4
List of key assets.

Asset ID	Description	Potential risk	Residual risk
7	Internet Connection	12.67	20.00
11	DMZ	16.89	26.67
26, 27	Palo Alto 2050 IPS	33.33	33.33
45	Development Server	26.67	26.67
149	Client Portal	26.75	14.80
198	Client Data	36.35	46.89

We refer the reader to Figueira (2019) for a detailed description of the SME's identified assets, threats, and vulnerabilities. Here, to simplify the analysis and validate the proposed methodology, we use a subset of key assets in terms of the high associated risk, listed in Table 4. The potential and residual risks (see Eqs. (1) and (3)) were calculated using Secitor (Secitor.com., 2018), a commercially available application designed for integral management of information security that includes a risk analysis module based on the Magerit methodology (MAGERIT V.3, 2014). Secitor also includes a module that reflects dependencies between assets. The existing dependencies between the key assets (Table 4), reflected in Fig. 1, were determined by the risk analysts (based on their own expertise) in an evaluation of the system, supported by Nessus (Tenable., 2018) and CVSS (Common Vulnerability Scoring System, 2019). Note that the residual risk depends on the implemented safeguards, which means that it will decrease or increase with regard to potential risk if the safeguards turn out to be unsuitable, as happened for instance, with assets 7 (Internet Connection) and 198 (Client Data).

Once the set of assets was identified, they had to be contextualized by determining which vulnerabilities and threats affected each. From the analysis carried out on the SME, we extracted the set of vulnerabilities listed in Table 5. In choosing the set of vulnerabilities, an extra effort to measure each independently was done; thus, instead of considering "Absence of security guidelines" as a whole, we broke it down into seven different vulnerabilities (VUL_57, VUL_58, VULN_68, VULN_71, VULN_74, VULN_77, VULN_78 and VULN_81) that reflected specific aspects of the initial vulnerability. Although vulnerabilities are defined as independent (because there is no dependence between the factors that cause them), it is important to highlight the fact that vulnerabilities could still be related to each other, whether by the context in which they are evaluated or by the internal processes of the SME.

The evaluation parameter describes the state of each vulnerability on a scale 0 to 10, where 0 indicates that a vulnerability has been resolved, and 10 that a vulnerability remains fully exploitable. As for dependencies, those values were calculated by risk analysts based on their expertise. Vulnerabilities are considered as risk triggers, as they are evaluated individually considering their real state. In this way it is possible to reflect, for instance, whether a guideline has been correctly developed (corresponding to a 0 vulnerability value), if it is well developed but could be improved (a 3-4 vulnerability value), if it is well developed but is defective (a 7-8 vulnerability value), or if it has not been deployed at all (a 10 vulnerability value).

Regarding threats, we found twenty different crucial threats to the identified assets and vulnerabilities:

- TH_004 - Line eavesdropping.
- TH_005 - Unauthorized use of IT systems.
- TH_006 - Unauthorized use of remote maintenance connections.
- TH_013 - Improper use of administrator privileges.
- TH_015 - Malware.

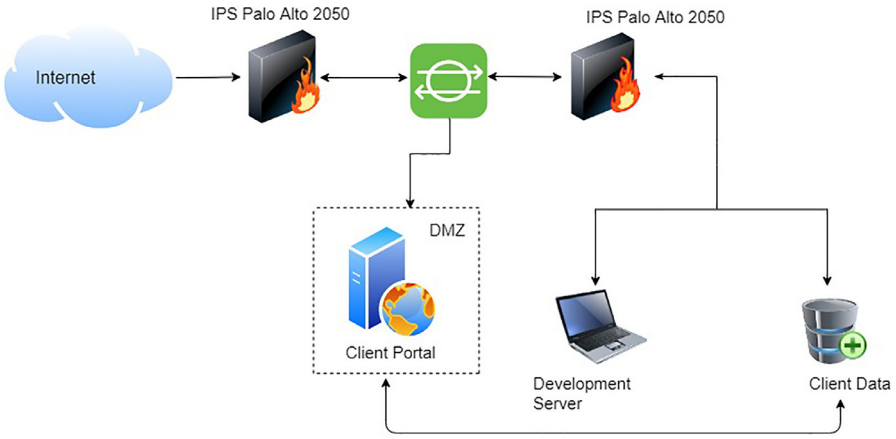

Fig. 1. Asset dependencies.

Table 5
List of vulnerabilities considered for the SME.

Vulnerability	Name	Evaluation
VULN_05	Absence of an efficient configuration change control	5
VULN_08	Unprotected storage	6
VULN_09	Access Control not deployed	10
VULN_10	Uncontrolled copies	7
VULN_12	Known software vulnerabilities	9
VULN_13	Users do not logout when they leave their workplace	7
VULN_15	Absence of audit signs	5
VULN_16	Poor assignation of access permissions	7
VULN_17	Widely distributed software	6
VULN_23	Absence of adequate identification and user authorization mechanisms	6
VULN_24	Password tables without protection	5
VULN_25	Inadequate management of passwords	6
VULN_27	Immature or very new software	4
VULN_29	Absence of effective change control	4
VULN_30	Download and installation of uncontrolled software	3
VULN_34	Absence of assurance on sending and receiving messages	4
VULN_35	Communication lines unprotected	4
VULN_36	Sensitive network traffic unprotected	3
VULN_38	Single points of failure	1
VULN_39	Absence of identification and authorization of sender and receiver	8
VULN_40	Network architecture unsecured	8
VULN_41	Password transmissions in clear text	3
VULN_42	Sensitive information transmission in clear text	3
VULN_43	Inappropriate network management	8
VULN_44	Connections with public network unprotected	8
VULN_46	Inadequate recruitment procedures	2
VULN_47	Insufficient security training	7
VULN_48	Hardware or software misuse	6
VULN_52	Absence of guidelines for the correct use of telecommunications	6
VULN_57	Absence of guidelines for the addition of new users	4
VULN_58	Absence of guidelines for supervising rights access	7
VULN_61	Absence of audit and regular supervision	6
VULN_62	Absence of risk identification procedures	6
VULN_63	Absence of incident or failure reports in the logs of operators and administrators	9
VULN_65	Absence of change management procedures	6
VULN_68	Absence of guidelines to allow users access to information	7
VULN_71	Absence of guidelines related to the use of corporate e-mails	4
VULN_72	Absence of software installation procedures in different OS	5
VULN_73	Absence of operator and administrator records	7
VULN_74	Absence of guidelines related to classified information management	5
VULN_77	Absence of guidelines related to disciplinary procedures in case of security incidents	2
VULN_78	Absence of guidelines related to mobile device management	6
VULN_80	Absence of clean screens and tables policies	6
VULN_81	Absence of authorization to access information processing devices	9
VULN_82	Absence of monitoring mechanisms to avoid theft or incidents	10
VULN_86	Absence of a security policy	7
VULN_87	Absence of developed safety regulations	7
VULN_88	Absence of surveillance in the building	3

Table 6
Threats and the related vulnerabilities in the SME.

Threat	Related vulnerabilities
TH_004	VULN_35, VULN_36, VULN_40, VULN_42, VULN_44
TH_005	VULN_09, VULN_23, VULN_39, VULN_48, VULN_58, VULN_82, VULN_86
TH_006	VULN_15, VULN_44, VULN_73
TH_013	VULN_13, VULN_15, VULN_46, VULN_53, VULN_57, VULN_58, VULN_61, VULN_73, VULN_82
TH_015	VULN_12, VULN_17, VULN_30, VULN_47, VULN_62, VULN_71, VULN_72, VULN_86
TH_016	VULN_24, VULN_25, VULN_34, VULN_36, VULN_39, VULN_41, VULN_58, VULN_62, VULN_80, VULN_82
TH_023	VULN_13, VULN_80, VULN_86, VULN_87, VULN_88
TH_024	VULN_10, VULN_16, VULN_17, VULN_27, VULN_30, VULN_48
TH_025	VULN_38, VULN_43, VULN_44
TH_026	VULN_05, VULN_29, VULN_40, VULN_65, VULN_82
TH_029	VULN_08, VULN_25, VULN_29, VULN_30, VULN_39, VULN_40, VULN_48, VULN_62, VULN_78, VULN_82
TH_044	VULN_09, VULN_15, VULN_16, VULN_58
TH_045	VULN_09, VULN_81
TH_046	VULN_36, VULN_40, VULN_41, VULN_42, VULN_43, VULN_44
TH_047	VULN_35, VULN_39, VULN_40, VULN_81
TH_048	VULN_57, VULN_58, VULN_68, VULN_74
TH_145	VULN_09, VULN_16, VULN_78
TH_146	VULN_08, VULN_09, VULN_40, VULN_44, VULN_81
TH_147	VULN_12, VULN_43
TH_148	VULN_09, VULN_16, VULN_52, VULN_71, VULN_77

- TH_016 - Impersonation.
- TH_023 - Capturing information using video.
- TH_024 - Introduction of malicious code.
- TH_025 - Denial of service due to a hacker attack.
- TH_026 - Deliberate alteration of system configuration data.
- TH_029 - Backdoor access.
- TH_044 - Unauthorized use of access rights: use of credentials without prior authorization to access SME data.
- TH_045 - Uncontrolled resources usage: use of SME resources without authorization or control.
- TH_046 - Insufficient protection of network connection: poor protection of SME communication networks.
- TH_047 - Uncontrolled use of telecommunications: use of SME communication lines without any control over the communications.
- TH_048 - Inappropriate management of access permissions: handling of credentials and access identifications to the SME systems in an inappropriate manner.
- TH_145 - Inappropriate remote authentication system: wrong remote user authentication in access to the SME systems.
- TH_146 - Hacking: occurrence of an attack against the IT infrastructures of the SME, with the attacker trying to access both systems and stored data.
- TH_147 - Wi-Fi vulnerability: wrong Wi-Fi configuration or vulnerable Wi-Fi protocols allowing access to the IT systems of the SME without authorization.
- TH_148 - Unauthorized access to distribution lists: unauthorized access to the distribution lists of the SME.

To comprehensively define the scenario, we asociated threats with vulnerabilities that could induce their materialization, as depicted in Table 6. Establishing the relationships between threats and vulnerabilities is a costly and complex process, since there is a large number of variables to consider in each scenario. Currently, several regulations such as Magerit (or ISO 27005) has already done this work to a high level of reliability. In this article, we part from Magerti regulation. For instance, through the years, security experts have considered that the vulnerabilities "single point of failure" (VULN_38), "inappropriate network management" (VUL_43) and "unprotected connections to public networks" (VULN_44) are some of the main vulnerabilities we need to face in order to deal with a Denial of Service (TH_025). This has been recorded in Magerit, and therefore we have used it in our work.

	VULN36	VULN40	VULN41	VULN42	VULN43	VULN44	HAPPENED
1	1	2	0	1	2	1	0
2	1	1	2	2	2	1	0
3	0	2	2	2	1	2	0
4	1	2	0	1	0	2	0
5	1	2	2	1	1	0	0
6	2	0	0	0	2	1	0
7	1	0	1	0	0	2	0
8	2	2	0	1	2	0	0
9	0	0	2	1	2	1	0
10	1	1	2	0	1	1	0
11	2	1	2	1	1	2	1
12	1	1	1	1	1	1	0
13	0	1	2	2	0	0	0
14	0	2	1	0	1	0	1

Fig. 2. Threat TH_046 samples.

Note that the identification of assets, threats and vulnerabilities, the determination of dependencies and the discovery of relationships, are common tasks for both Magerit and our proposal.

Finally, in order to apply our approach, we need two kinds of additional data for the evaluated period (one year): the frequency of materialization of each threat, and a set of samples for each identified threat.

Samples are colleted by a platform that, with a periodicity depending on each system, obtains the evaluation value of each vulnerability. The time of sampling is also recorded. Afterwards, when a security incident is detected by a company security analyst, a security auditor or a third party, all the samples of the threat(s) involved in the incident are marked active during the time period in which systems and assets were compromised.

Each sample in a threat dataset contained the following information:

- A list of all the vulnerabilities that apply to a particular threat.
- The evaluation value of each listed vulnerability, as obtained by the risk analyst (based on his/her knowledge and expertise) or by the platform at the time of sampling. Evaluation values reflect the state of the system at that particular moment, considering among other things, the implemented safeguards.

Table 7
Number of samples: total, training and testing.

Threat	Total no. of samples	No. training samples	No. testing samples	Threat	Total no. of samples	No. training samples	No. testing samples
TH_004	902	720	181	TH_005	673	538	135
TH_006	780	624	156	TH_013	705	564	141
TH_015	535	428	107	TH_016	1623	1298	325
TH_023	595	476	119	TH_024	727	581	146
TH_025	3044	2435	609	TH_026	606	484	122
TH_029	1813	1450	363	TH_044	1650	1320	330
TH_045	826	660	166	TH_046	2400	1920	480
TH_047	1321	1056	265	TH_048	1981	1584	397
TH_145	1981	1584	397	TH_146	3241	2592	649
TH_147	1006	804	202	TH_148	2820	2256	564

- A feature (called HAPPENED) indicating whether or not the threat was active at the time of sampling, and considering the current state of the system (vulnerabilities and their evaluation values).

Fig. 2 shows a subset of samples corresponding to threat TH_046.

To collect enough data to develop a valid model, the scenario was replicated in multiple locations, and between 500 and 3200 samples were collected for each threat (depending on its nature and the related events in the evaluated time period).

4.2. Obtaining threat models

To obtain the models for each threat, we used the R packages Classification And Regression Training (Caret,[2] v.6.0–84) and Kernel-Based Machine Learning Lab (Kernlab,[3] v.0.9-27).

For each threat, we used the 80% of the available samples for training, and the remaining 20% for testing. During the training phase, we use k-fold cross-validation (Stone, 1974) to find the best models. Table 7 shows the number of available samples of each threat, and the number of samples used for training and testing.

4.2.1. Logistic regression

The logistic regression method determines the best model that describes the relationship between a dependent variable of interest and a set of independent variables that influence the variable of interest. As a result, logistic regression yields the probability of occurrence of the dependent variable, calculated as follows:

$$p = \frac{1}{1 + e^{1-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}} \tag{10}$$

where x_i represents each of the independent variables and where β_i are the coefficients generated by the logistic regression. Applying this algorithm to our scenario, it is possible to read β_i as a description of each vulnerability, and the model as representing the behavior of a particular threat. As we will see in the validation section, a threat is considered to occur if p is higher than a given threshold.

In order to determine the number of folds for the implementation of cross-validation, we tested different values of k ranging from 3 to 12. The best models -in this case, those with the highest AUC- were achieved for $k = 7$.

Table 8
SVM tuning parameters.

Threat	C_{best}	γ_{best}	k_{best}	Threat	C_{best}	γ_{best}	k_{best}
TH_004	0.080	2^{-5}	6	TH_005	0.500	2^{-5}	6
TH_006	0.070	2^{-5}	6	TH_013	4.300	2^{-13}	6
TH_015	7.500	2^{-13}	6	TH_016	8.300	2^{-14}	6
TH_023	0.200	2^{-6}	6	TH_024	0.120	2^{-7}	6
TH_025	0.007	2^{-3}	6	TH_026	0.400	2^{-7}	6
TH_029	0.005	2^{-6}	6	TH_044	1.100	2^{-12}	4
TH_045	0.030	2^{-4}	6	TH_046	0.09	2^{-9}	6
TH_047	0.300	2^{-10}	4	TH_048	5.500	2^{-15}	6
TH_145	0.003	2^{-3}	6	TH_146	0.002	2^{-4}	6
TH_147	0.030	2^{-5}	6	TH_148	9.800	2^{-17}	6

4.2.2. SVM regression

SVM learning is based on mapping samples into a high-dimensional space in order to obtain an optimal separating hyperplane that maximizes the sum of the distances between the two classes considered in this space. In our scenario, for each threat, SVM is used to obtain the hyperplane that separates samples in which the HAPPENED flag equals "one" (the threat occurs), from samples in which the HAPPENED flag equals "zero" (the threat does not occur).

In this paper we used ϵ-supported vector regression, with $\epsilon = 0.1$ and a Gaussian Radial Basis Function (RBF) kernel (Vapnik, 1995). The value of ϵ defines the margin of tolerance where no penalty is associated with points predicted within a distance ϵ from the actual value. In order to select the SVM parameters (C, γ), we use cross-validation and grid search.

The regularization parameter C represents a trade-off between training error and model complexity. The search range for C was [0.001,10], in 0,001 steps. The width kernel parameter γ can be interpreted as the inverse of the radius of influence of the training samples selected by the model as supported vectors. The search range for γ was $[2^{-20},1]$, multiplying by 2 in each step. To determine the number of folds for the cross-validation, we conducted a search from $k = 3$ to $k = 12$. The best models, in terms of Root Mean Square Error (RMSE), were obtained for $k = 4$ and $k = 6$.

Table 8 shows a summary of the parameter values used to generate models for each threat.

4.3. Validating threat models

In order to validate the results obtained for the threat models, the following indicators were considered:

- Accuracy: ratio for true positives and true negatives using all the samples.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Number\ of\ samples} \tag{11}$$

[2] Max Kuhn (2008). Caret package. Journal of Statistical Software, 28(5), https://github.com/topepo/caret/.

[3] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. http://www.jstatsoft.org/v11/i09/.

Table 9
Logistic regression: validation parameters.

Logistic regression						
Threat	NIR	Accuracy	Precision	Recall	Specificity	F-score
TH_004	0.5894	0.8619	0.7414	0.8113	0.8828	0.7747
TH_005	0.5260	0.9185	0.9000	0.8823	0.9405	0.8911
TH_006	0.6789	0.8141	0.6889	0.6739	0.8727	0.6813
TH_013	0.5186	0.8369	0.8194	0.8551	0.8194	0.8369
TH_015	0.5070	0.8130	0.7719	0.8627	0.7678	0.8148
TH_016	0.5166	0.8123	0.8121	0.8448	0.7748	0.8282
TH_023	0.6842	0.8487	0.8182	0.6923	0.9250	0.7500
TH_024	0.6172	0.8561	0.8113	0.7963	0.8913	0.8037
TH_025	0.6125	0.8177	0.7368	0.7333	0.8621	0.7350
TH_026	0.6446	0.8678	0.8000	0.8000	0.9012	0.8000
TH_029	0.5321	0.7741	0.7318	0.7939	0.7575	0.7616
TH_044	0.5651	0.7818	0.7686	0.7153	0.8333	0.7410
TH_045	0.5575	0.8303	0.8286	0.7838	0.8681	0.8055
TH_046	0.6730	0.8542	0.8042	0.7301	0.9140	0.7654
TH_047	0.5056	0.8409	0.8500	0.8095	0.8695	0.8293
TH_048	0.5549	0.8131	0.7696	0.8059	0.8186	0.7873
TH_145	0.5233	0.8358	0.8350	0.8308	0.8408	0.8329
TH_146	0.5027	0.8395	0.8557	0.8131	0.8654	0.8339
TH_147	0.5758	0.8209	0.8333	0.7143	0.8974	0.7692
TH_148	0.5426	0.8450	0.8272	0.8380	0.8509	0.8326

Table 10
SVM regression: validation parameters.

SVM						
Threat	NIR	Accuracy	Precision	Recall	Specificity	F-score
TH_004	0.5894	0.8564	0.7288	0.8113	0.8750	0.7679
TH_005	0.5260	0.8963	0.8246	0.9216	0.8810	0.8704
TH_006	0.6789	0.8205	0.6667	0.7826	0.8364	0.7200
TH_013	0.5186	0.8723	0.8072	0.9710	0.7778	0.8816
TH_015	0.5070	0.8224	0.7424	0.9608	0.6964	0.8376
TH_016	0.5166	0.8092	0.7800	0.8965	0.7086	0.8342
TH_023	0.6842	0.8571	0.8235	0.7179	0.9250	0.7671
TH_024	0.6172	0.8493	0.7758	0.8333	0.8587	0.8036
TH_025	0.6125	0.8227	0.7277	0.7762	0.8471	0.7512
TH_026	0.6446	0.8595	0.7556	0.8500	0.8642	0.8000
TH_029	0.5321	0.7438	0.6714	0.8545	0.6151	0.7520
TH_044	0.5651	0.8152	0.7823	0.7986	0.8280	0.7904
TH_045	0.5575	0.8424	0.8333	0.8108	0.8681	0.8219
TH_046	0.6730	0.8771	0.7833	0.7616	0.9300	0.7958
TH_047	0.5056	0.8333	0.8254	0.8254	0.8406	0.8254
TH_048	0.5549	0.8106	0.7387	0.8647	0.7699	0.7967
TH_145	0.5233	0.8459	0.8317	0.8615	0.8308	0.8463
TH_146	0.5027	0.8364	0.8413	0.8255	0.8471	0.8333
TH_147	0.5758	0.8308	0.8378	0.7381	0.8974	0.7848
TH_148	0.5426	0.8546	0.8094	0.8858	0.8290	0.8459

- Precision: ratio for true positives and the sum of true and false positives, interpretable as the positive value ratio.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{12}$$

- Recall/sensitivity: ratio between true positives and the sum of true positives and false negatives, interpretable as the predictive positive value ratio.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{13}$$

- Specificity: ratio for true negatives and the sum of true negatives and false positives, that is, the portion of actual negatives that are correctly identified as such.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{14}$$

- F-score: harmonic average for precision and recall.

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

- No Information Rate (NIR): is the largest proportion of the observed classes.

$$NIR = \max\left(\frac{True\ Positives + False\ Negatives}{Number\ of\ samples}, \frac{True\ Negatives + False\ Positives}{Number\ of\ samples}\right) \tag{16}$$

To validate the models we used the 20% of the samples, reserved for testing.

4.3.1. Logistic regression validation

A threat sample is considered to be positive if the probability obtained by the corresponding model is equal to or higher than 0.5. Once the sample is classified, it is compared with the real behavior as recorded in the dataset. That is, a threat sample is considered a "True Positive" if it is classified as positive and the corresponding `HAPPENED` feature in the dataset equals one. Accordingly, a threat sample is considered a "True Negative" if it is classified as negative and the corresponding `HAPPENED` feature in the data set equals zero.

Table 9 shows the results for each validation indicator. It can be observed that all validation parameters (taking NIR off) are above 70% (except for precision in TH_006), and most are above 80%. No clear trend is observed regarding whether failure occurs more with positive or negative classification: for some of the models recall is better, for some others specificity is better, and in all cases the F_score is above 70%. Accuracy is above 80% in all the cases, and surpasses the NIR by at least 13.52 percentage points. For those reasons, we believe that logistic regression models are reliable.

4.3.2. SVM regression validation

As in the previous section, a threat sample is considered to be positive if the probability obtained by the corresponding model is equal to or higher than 0.5, and it is considered a "True Positive" ("True Negative") if it is classified as positive (negative) and the corresponding `HAPPENED` feature in the dataset equals one (equals zero).

Table 10 shows the results for each validation indicator. The results are quite similar to those for logistic regression, even slightly better if we consider average accuracy, recall and F_Score, although slightly worse if we consider average specificity. However, it should be noted that, in this scenario, false negatives are less harmful than false positives (in the sense that is more important to consider a non-dangerous threat than fail to consider a dangerous one).

4.4. Probabilities and frequency computing

Once the regression models were defined, they were applied to all the threats arising in the scenario described in Section 4.1, and, using the vulnerability evaluation values given in Table 5 -as the inputs of the models, the probability of the materialization of each threat was calculated. Results are shown in Table 11.

The original frequency values used for the recalculated frequency values were obtained as the degree of materialization of a threat in the previous year. The method used to recalculate frequencies is described in Section 3.2. The original frequency values and the new frequency values are also shown in Table 11.

Note that the recalculated frequencies may be lower or higher than the original frequencies based only on historical data. However, the new values more precisely reflect the current scenario, because the current state of vulnerabilities is considered in calculating them. That is, the goal of the proposed methodology is not to reduce the risk, but to ensure that the calculated risk better reflects the current state of the system.

Table 11
Original frequency values, probabilities and recalculated frequency values of each identified threat.

Threath ID	SVM Probability	LR Probability	Original Frequency	SVM P_{TH}_Frequency	LR P_{TH}_Frequency
TH_004	0.5627	0.6369	0.50	2.25	3.74
TH_005	0.9460	0.9942	0.50	9.03	9.90
TH_006	0.7357	0.7026	0.50	5.60	5.04
TH_013	0.8249	0.8598	0.05	7.08	7.66
TH_015	0.8020	0.8833	10.0	6.70	8.06
TH_016	0.7653	0.8412	10.0	6.09	7.35
TH_023	0.5644	0.5004	0.10	2.29	1.75
TH_024	0.6598	0.6556	0.05	4.20	4.11
TH_025	0.5521	0.4776	0.05	2.04	1.64
TH_026	0.5684	0.4618	0.01	2.37	1.56
TH_029	0.8298	0.8470	2.00	7.16	7.45
TH_044	0.6908	0.6136	10.0	4.82	3.27
TH_045	0.9590	0.9045	10.0	9.32	8.41
TH_046	0.4597	0.3806	1.00	1.55	1.15
TH_047	0.7280	0.7004	10.0	5.47	5.01
TH_048	0.5303	0.4843	5.00	1.90	1.67
TH_145	0.7791	0.7469	2.00	6.32	5.78
TH_146	0.8526	0.7831	5.00	7.54	6.39
TH_147	0.8287	0.7860	10.0	7.15	6.43
TH_148	0.4590	0.3855	3.00	1.55	1.18

Table 12
Risks for key SME assets.

Asset ID	Description	Potential risk (historical)	Residual risk (historical)	Potential risk (SVM prediction)	Residual risk (SVM prediction)	Potential risk (LR prediction)	Residual risk (LR prediction)
7	Internet Connection	12.67	20.00	17.27	27.27	18.35	28.98
11	DMZ	16.89	26.67	43.97	69.61	46.73	73.97
26, 27	Palo Alto 2050 IPS	33.33	33.33	5.03	5.03	5.35	5.35
45	Development Server	26.67	26.67	17.27	17.27	21.41	21.41
149	Client Portal	26.75	14.80	14.39	7.96	15.30	8.46
198	Client Data	36.35	46.89	68.60	88.49	72.89	94.03

Note also that threats with the same original frequency value obtained different recalculated frequency values; this is a consequence of the application of probability values obtained in the regression models. In conclusion, the model was capable of adapting to current and future conditions of the evaluated context based on the obtained probabilities.

4.5. Discussion

Table 12, which shows the risk values for the key assets, compares the results obtained by the predictive model and by an historical-based model.

It can be observed that the risk for all the key assets changed, increasing for some (7, 11, 198), and decreasing for others (26, 27, 45, 149). This is the result of a clearer vision of the current state of the system obtained once prediction was incorporated in the scenario. When data were calculated only using historical data, the methodology was based on biased information, because possible actions (fixes or safeguards) addressing vulnerabilities were not considered in threat frequency calculations. However, when data are calculated using the predicted probabilities, the current state of the vulnerabilities is considered. As a consequence, the calculated risk responds to a more updated scenario. Calculating such tighter risk values will allow organizations to focus their efforts and investments on threats with higher probabilities of future materialization.

5. Conclusions

In this paper we have proposed an alternative predictive model for risk analysis methodologies, particularly for Magerit, the open-source risk analysis and management methodology developed by the Spanish government. The proposal is based on a modifying risk calculation by substituting past (historical) threat frequencies with future threat probabilities taking into account current system vulnerabilities.

Forecasting threat occurrence probabilities instead of compiling historical data, in shifting the focus from the past to the future, ensures better knowledge of the system and produces results for the current state that are more accurate and therefore more valuable. These better results mean that organizations can focus on the most dangerous threats and so implement better targeted and more efficient safeguards, which, in turn, will reduce damage and losses and improve information security overall. Compared to traditional risk analyses based solely on historical data, this new approach facilitates the adaption to system changes, since the calculated probabilities vary as the status of vulnerabilities varies.

To calculate the probabilities, we proposed two well-known approaches, namely, logistic regression and SVM regression models. Although both provide good results, logistic regression models would appear to be a good alternative, as models are simpler to tune, and less time consuming.

Since the proposed methodology can be easily automatized, particularly in the case of logistic regression (with no parameters to tune), it can feasibly be integrated into risk analysis tools such as PILAR and Secitor.

The proposed methodology was validated by a case study carried out in a business environment and using real data. The obtained results were satisfactory opening the door to further analyses using more sophisticated machine learning techniques to improve the calculation of probabilities.

Declaration of Competing Interest

None.

Acknowledgment

The authors would like to thank Eduardo Cunha Rodríguez (CTO Secitor and Goblal CISO Andbank) and Olga Solís Navarro (graduated in Mathematics and Cybersecurity Consultant at Deloitte Spain) for their comments that greatly improved the manuscript. Ailish Maher revised the English in a version of the manuscript. This research has been partially supported with MINECO grant TEC2016-76465-C2-2-R and Xunta de Galicia grant GRC2018/53.

References

Bojanc, R., Jerman-Blažič, B., 2013. A quantitative model for information-security risk management. Eng. Manage. J. 25 (2), 25–37. ISSN 1042-9247, https://doi.org/10.1080/10429247.2013.11431972.

Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., Stoddart, K., 2016. A review of cyber security risk assessment methods for SCADA systems. Comput. Security 56, 1–27. ISSN 0167-4048, https://doi.org/10.1016/j.cose.2015.09.009.

Coras.sourceforge.net.. The CORAS method. 2018. [online] available at: http://coras.sourceforge.net/ [Accessed 19 Oct. 2018].

Cve.mitre.org.. CVE -common vulnerabilities and exposures (CVE). 2018. [online] Available at: https://cve.mitre.org/ [Accessed 19 Oct. 2018].

Feng, N., Wang, H.J., Li, M., 2014. A security risk analysis model for information systems: causal relationships of risk factors and vulnerability propagation analysis. Inf. Sci. 256, 57–73. ISSN 0020-0255, https://doi.org/10.1016/j.ins.2013.02.036.

Fredriksen, R., Kristiansen, M., Gran, B.A., Stølen K. Opperud, T.A., Dimitrakos, T., 2002. The CORAS Framework for a model-based risk management process. Computer Safety, Reliability and Security. SAFECOMP 2002. Lecture Notes in Computer Science, 2434. Springer, Berlin, Heidelberg. 94–105.

Harrell, Frank, E., 2015. Ordinal logistic regression. In: Regression Modeling Strategies. Springer, Cham, pp. 311–325. ISBN 978-3-319-19424-0, https://doi.org/10.1007/978-3-319-19425-7_13.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. in IEEE Intell. Syst. Appl. 13 (4), 18–28. doi:10.1109/5254.708428.

International Organization for Standarization. ISO/IEC 27005:2008, Tecnologías de la información. Técnicas de Seguridad. Gestión de riesgos de Seguridad de la Información. ISO 27000. 2008.

Jindong, W., Jian, Z., Na, W.y., Yu, C., 2013. Risk prediction method of information system based on Bayesian game. In: 3rd International Conference on Computer Science and Network Technology (ICCSNT), pp. 191–195. Zhengzhou, China.

Karabacak, B., Sogukpinar, I., 2005. ISRAM: information security risk analysis method. Comput. Security 24 (2), 147–159. ISSN 0167-4048, https://doi.org/10.1016/j.cose.2004.07.004.

Lee, M.C., 2014. Information security risk analysis methods and research trends: AHP and fuzzy comprehensive method. Int. J. Comput. Sci. Inf. Technol. 6 (1), 29–45.

MAGERIT V.3 (English version). Methodology for information systems risk analysis and management. Ministerio de Hacienda y Administraciones Públicas, 2014. España. - NIPO: 630-14-162-0.

Massaccia, F., Prestb, M., Zannonea, N., 2005. Using a security requirements engineering methodology in practice: the compliance with the italian data protection legislation. Comput. Standards Interfaces 27, 445–455. ISSN 0920-5489, https://doi.org/10.1016/j.csi.2005.01.003.

Mehari. Overview, Club de la Securité de l'Information Français (CLUSIF). 2007.

NIST, SP800-30, 2002. Risk Management Guide for Information Technology Systems. National Institute of Standards Technology, Gaithersburg.

Peltier, T.R., 2010. Information Security Risk Analysis, 3rd edition Auerbach Publications. ISBN 1439839565.

Procedimiento Informático-Lógico para el Análisis de Riesgos. Centro criptológico nacional. centro nacional de inteligencia. Ministerio de Presidencia. España. 2019.

Rajbhandari, L., Snekkenes, E., 2013. Using the conflicting incentives risk analysis method. Security Privacy Prot. Inf. Process. Syst. 405. ISBN 978-3-642-39217-7, https://doi.org/10.1007/978-3-642-39218-4_24.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology, in IEEE transactions on systems. Man Cybern. 21, 660–674. doi:10.1109/21.97458.

Secitor.com. SECITOR. 2018. [online] Available at: http://www.secitor.com/[Accessed 19 Oct. 2018].

Shameli-Sendi, A., Aghababaei-Barzegar, R., Cheriet, M., 2016. Taxonomy of information security risk assessment (ISRA). Comput. Security 57, 14–30. ISSN 0167–4048, https://doi.org/10.1016/j.cose.2015.11.001.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. 36 (1), 111–147. B

Suh, B., Han, I., 2003. The IS risk analysis based on a business model. Inf. Manage. 41 (2), 149–158. ISSN 0378-7206, https://doi.org/10.1016/S0378-7206(03)00044-2.

Team, C. Common Vulnerability Scoring System v3.0 : Specification Document (Qualitative Severity Rating Scale). First.org. 2015.

Tenable. Nessus Professional. 2018. [online] Available at: https://www.tenable.com/products/nessus/nessus-professional[Accessed 19 Oct. 2018].

Tubío Figueira P. Estudio y desarrollo de una metodología de análisis de riesgos utilizando algoritmos de predicción abiertos, similares a Cortana. 2019. Available at: http://castor.det.uvigo.es:8080/xmlui/handle/123456789/126.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. New York, Springer.

Vicente, E., Mateos, A., Jiménez-Martín, A., 2014. Risk analysis in information systems: a fuzzification of the MAGERIT methodology. Knowl.-Based Syst. 66, 1–12. ISSN 0950–7051, https://doi.org/10.1016/j.knosys.2014.02.018.

Xu, N.y., Zhao, D., 2011. The research of information security risk assessment method based on AHP. Adv. Mater. Res. 187, 575–580. https://doi.org/10.1016/www.scientific.net/AMR.187.575.

Yazar, Z., 2002. A Qualitative Risk Analysis and Management Tool-CRAMM. SANS InfoSec Reading Room White Paper.

Pedro Tubío Figueira received his Telecommunications Engineering Degree in 2017. He has been a Security and Systems Consultant at Hermes Sistemas in 2017. At this moment, he works as a Senior Cybersecurity Consultant in Infrastructure & Cloud Protection department at Deloitte Spain. His fields of interest are mainly focused on network security issues, in both physical and cloud environments, and the risks associated with both technologies, such as exposure to DDoS attacks. In addition, he has also focused his specialization on OT networks, endpoints protection, and Active Directory audits and securization.

Cristina López Bravo received her Telecommunications Engineering and Ph.D. degrees in 2000 and 2004, respectively. She has been an associate professor with the Departamento de Enxeñaría Telemática, Universidade de Vigo, since 2008. She has authored or coauthored over 30 papers in international refereed journals and international conferences, in the fields of telecommunications and computer science. She has participated in several relevant European (EphotonOne, Bone) and national projects (ARPAq,CAPITAL, CALM, COINS, AIMS). Her fields of interest are mainly focus on network performance evaluation, security in wireless environments.

José Luis Rivas López, Systems and Telematics engineer, graduated in Telecommunications technologies with two master's degrees, first in Free Software and second one, in Security of Information and Communication technologies. He has authored or coauthored over 12 books, in the fields of telecommunications, computer science and cybersecurity. He is a member of the Technical Committee of AENOR (Spanish Association for Standardization and Certification) Normalizer on standard 19701 and 197010 (Forensics Reports). He has been CEO of Secitor, since 2010. His fields of interest are mainly focused on networking and security technologies.