COMPUTERS & SECURITY 83 (2019) 38–51

Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Analysis of computer user behavior, security incidents and fraud using Self-Organizing Maps

Alberto Urueña López[a], *Fernando Mateo*[b,*], *Julio Navío-Marco*[c], *José María Martínez-Martínez*[b], *Juan Gómez-Sanchís*[b], *Joan Vila-Francés*[b], *Antonio José Serrano-López*[b]

[a] *Department Business Administration, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid C/José Gutiérrez Abascal 2, Madrid 28006, Spain*

[b] *Department of Electronic Engineering, Intelligent Data Analysis Laboratory (IDAL), University of Valencia, Av de la Universidad, s/n, Burjassot, Valencia 46100, Spain*

[c] *Faculty of Economics and Business Administration, Department Business Organization, University of Open Learning (UNED), Paseo Senda del Rey 11, Madrid 28040, Spain*

ARTICLE INFO

Article history:
Received 26 February 2018
Revised 20 December 2018
Accepted 25 January 2019
Available online 30 January 2019

Keywords:
Cybersecurity
Self-Organizing Maps
Data visualization
Survey analysis
Risk assessment

ABSTRACT

This paper addresses several topics of great interest in computer security in recent years: computer users' behavior, security incidents and fraud exposure on the Internet, due to their high economic and social cost. Traditional research has been based mainly on gathering information about security incidents and fraud through surveys. The novelty of the present study is given by the use of Self-Organizing Maps (SOMs), a visual data mining technique. SOMs are applied to two data sets acquired using two different methodologies for collecting data about computer security. First, a traditional online survey about fraud exposure, security and user behavior was used. Second, in addition to surveys, real data obtained from some of the users' computers were also considered. In this way, the answers of the users can be benchmarked with the true situation of their computers. The surveys and the scanning of the computers were conducted in Spain from December 2013 to June 2014 by the National Observatory of Telecommunications and Information Society of the Spanish Ministry of Industry, performing 9181 surveys and 6350 computer scans in total. SOMs were applied to the datasets in their entirety first, and then a local analysis of the most interesting zones was carried out by zooming in on them. This approach allows for more detailed knowledge extraction. We conclude that SOMs enhance insight and interpretability about both data sets by untangling hidden relationships between variables, and could be helpful for similar future studies.

* Corresponding author.
E-mail addresses: alberto.uruena@upm.es (A. Urueña López), fernando.mateo@uv.es (F. Mateo), jnavio@cee.uned.es (J. Navío-Marco), jose.maria.martinez@uv.es (J.M. Martínez-Martínez), juan.gomez-sanchis@uv.es (J. Gómez-Sanchís), joan.vila@uv.es (J. Vila-Francés), antonio.j.serrano@uv.es (A. José Serrano-López).

https://doi.org/10.1016/j.cose.2019.01.009
0167-4048/

1. Introduction

The growth of Internet usage and the number of programs and applications uploads/downloads has been exponential in the recent years. However, new dangers are turning up as the number of advantages grows (Singer and Friedman, 2014). Computer security arises in this environment in order to provide protection against those risks. Computer security, also known as cybersecurity or IT security, can be defined as security applied to devices such as computers and smartphones, as well as computer networks (Knowles et al., 2015). The field includes all the processes and mechanisms by which digital equipment, information, and services are protected from unintended or unauthorized access, change or destruction, and is of growing importance due to the increasing reliance on computer systems in most societies (Donaldson et al., 2015). It includes physical security to prevent theft of devices and information security to protect the data on those devices. Cybersecurity is also essential to maintain the integrity of user information (Gordon et al., 2015). The concepts of cybersecurity and IT, among others, have particular relevance given the exponential growth of connected devices (Internet of Things, IoT) (Jang-Jacard and Nepal, 2014).

The perception of danger in users is a key point in the use of new technologies, as the feeling of mistrust may slow down the adoption of those technologies (Humaidi and Balakrishnan, 2013). It is, therefore, crucial to monitor the feelings and sentiments of users about the risks associated with the new Information and Communications Technologies (ICT) while testing whether those sentiments are founded (Bashir et al., 2017; Davinson and Sillence, 2010; Herrero et al., 2017a; Shillair et al., 2015).

Computer users who engage in risky behaviors have attracted the attention of researchers because they are considered one of the weak links in the field of Information Society (Arachchilage and Love, 2013). The literature on risk taking goes beyond the computer user domain, and available empirical evidence has indeed shown the negative consequences of risky behaviors (Kelly et al., 2004). Quite on the opposite side, the online commerce user community has a greater perception of risk; for example, an e-commerce client has a higher risk consciousness about a transaction in terms of payment and delivery than when making a traditional transaction at a physical shop (Urueña and Hidalgo, 2016).

This research carries out a diagnosis of the status of cybersecurity in Spanish digital households, analyzing the adoption of security measures and the incidence level of situations that may constitute security risks. This study is similar to others already performed (Campbell et al., 2007; Hoonakker et al., 2009; Stanton et al., 2005; Whitty et al., 2015). Nevertheless, two of the key differences in this paper are: (i) the confidence of household users is also assessed by means of computer analysis, which determines the degree of malware infection, and (ii) the selected tool to carry out the analysis mentioned above is the Self-Organizing Map (SOM), a powerful visual data mining method that provides a low-dimensional representation of high-dimensional data, thus enhancing visualization and interpretability for complex pattern recognition (Rossi, 2006).

The overall objective of this study is twofold. On one hand, to carry out an analysis of the status of cybersecurity and digital confidence among Spanish computer users. On the other hand, to contrast the level of incidents suered by the computers with the users' perceptions. For this purpose, user profiles are obtained from a questionnaire and from the analysis of their computers. Internet fraud is specifically addressed on the questionnaire. Fraud refers to the act of taking advantage of others, largely motivated by economic reasons, via various deceptive means.

This paper presents two types of studies to undertake this dual objective. Firstly, a general analysis of a specific questionnaire on cybersecurity was carried out. Secondly, an automated scanning software was used. The goal of using a questionnaire and independent data from scanning software is to reduce the possible bias in users' responses with respect to security perception. This problem, common in all kinds of surveys, is known as optimistic bias (Sharot, 2011), and is especially important when dealing with questions concerning user worries.

Once the data is collected from both sources, it is preprocessed by grouping it into categories and by extracting features of interest. Finally, SOMs are applied to extract knowledge about the interactions between those features.

The rest of this paper is organized as follows. The details of the SOM algorithm are described in Section 2. Aspects related to the questionnaire and the data integration for use in data visualization techniques are described in Section 3. In Sections 4 and 5, the results of both conducted studies (analysis of isolated questionnaires and analysis of combined results from the questionnaires and from the scanning software) are presented. Finally, Section 6 summarizes the conclusions of the present work.

2. Self Organizing Map (SOM)

2.1. *Theoretical framework*

The Self-Organizing Map (SOM) is an artificial model of the brain cortex, proposed in the 1970s by Teuvo Kohonen (Kohonen, 1989; Runz et al., 2012). This model resembles the fact that there is a mapping between the three dimensions of the human body to the two dimensions of the brain cortex, keeping a neighborhood relationship. Therefore, body regions which are close together stimulate nearby regions of the brain cortex. If we generalize this model to more than three dimensions on the original space while keeping a two-dimensional output space, we can map any N-dimensional dataset to a two-dimensional plane keeping the neighborhood relationship which can be directly visualized and analyzed.

This model has two layers of neurons (Fig. 1): an input layer representing the original space with as many neurons as input variables, and an output layer (bi-dimensional space).

Each neuron of the output layer is represented by an N-dimensional synaptic weight vector $\mathbf{m} = [m_1, \ldots, m_N]$, where N is equal to the dimension of the input vector; in this paper, the input space corresponds to the answered surveys and thus the number of input variables corresponds to the number of questions in each survey.

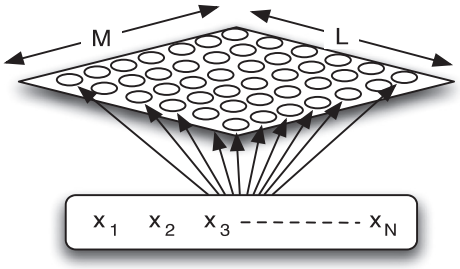


Fig. 1 – Self-Organizing Map scheme.

In order to deploy a SOM, we need to adjust the following settings (Hassoun, 2004):

1. *Map type*: hexagonal or rectangular grid, which indicates the topology or neighborhood relation; a hexagonal grid was chosen because it has a higher number of neighbors and, therefore, the number of relationships between survey answers can also be higher.
2. *Topology*: number of neurons of the output map: this setting is chosen by the designer. There are different works that define this number as a function of the number and dispersion of the variables (Vesanto et al., 1999).
3. *Initialization of the map*: synaptic weights of the neurons are obtained iteratively starting from an initial value. The most used approaches are Principal Component Analysis (PCA) (Haykin, 2009) and random initialization.
4. *Learning methodology*: there are two approaches for updating the synaptic weights of each neuron: *batch method*, which consists of updating all the weights at once after analyzing all the neurons, or *on-line method*, in which each weight is updated as its neuron is processed. The most used method is *on-line*, which is the one used in this work and is described below.

The update procedure for the on-line methodology is the following:

- A pattern is randomly chosen, whose values are compared with the synaptic weights using a distance metric, normally the Euclidean distance.
- The closest neuron is chosen as the winning neuron, which intuitively corresponds to the most similar one to the chosen pattern.
- The synaptic weights are updated according to the following expression:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t)[x(t) - m_i(t)], \quad (1)$$

where $x(t)$ is the pattern of the surveyed user at time t, and $\alpha(t)$ the learning rate. Here, $h_{ci}(t)$ stands for the neighborhood function, which defines the proportion in which the closest neurons to the winning one are fitted.

Intuitively, the consequence of this update is that the weights of the winning neuron and its neighborhood function are getting closer to the sampled pattern. Therefore, the closest neurons become more and more similar. This update separates the SOM into zones corresponding to different behaviors of the input patterns. Once the map training is finished, the projection of the two-dimensional map onto the different input patterns (component planes) provides qualitative information about how the input variables are related to each other for the data set used to train the map.

2.2. SOM limitations

SOMs are used to perform qualitative visual analysis to get behavior profiles and areas of interest. It is a very powerful tool to draw conclusions and gain insight about a data set in a multidimensional space, taking into account all the variables in a concurrent manner (not only pairs). The limitation of SOM is that the results cannot be used to perform quantitative analysis with statistical significance. Additionally, in many scenarios, SOMs are difficult to interpret and do not help to shed light on the interactions between variables.

3. Material

This section will address the description of: (a) the details of the questionnaires used, (b) the data preprocessing techniques applied (feature engineering) to the questionnaires answers to produce a consistent data and (c) the preprocessing of the variables captured by the scanning software for the same purpose.

3.1. Questionnaire

The data extracted from surveys on a bi-monthly and quarterly basis encompass December 2013 to June 2014 in Spain and were obtained by the National Observatory of Telecommunications and Information Society of the Spanish Ministry of Industry. The surveys were conducted via online questionnaires from a panel of Internet users who were asked to complete them by themselves. The surveys were aimed at Spanish users with at least one monthly access to the Internet from home. The sample design took into account a proportional stratification by habitat type for each Spanish region. Social segment quotas and number of people in the household were considered to that end. A pre-test with 50 users was made and no incidents were found, so the questionnaire was approved for the present study. The user screening criteria were the same as in previously published works that used the same data (Herrero et al., 2017a; 2017b).

First, the user was asked to answer questions describing their profile (demographic variables, operating system, browser, etc.). The rest of the survey was divided into eight different modules, which measure different aspects:

- **Module A: Use of the Internet.** Measures the use of various services offered on the Internet, e.g. select from a list all Internet services that were used in the last quarter.
- **Module B: Measures and Internet security habits.** Measures the actions of users with respect to Internet security, e.g. select security measures or software tools that were used in the last quarter or how frequently the users scan their system or check for antivirus updates.

Table 1 – Information about the surveys conducted in different quarters.

	Questionnaire 1	Questionnaire 2	Questionnaire 3
Minimum user age	19.5	18	18
Maximum user age	61.3	86	87
Mean user age	38.5	41.9	41.9
Median of user age	41.0	41.0	41.0
St. Dev. user age	11.6	11.3	11.4
User sex (% male / % female)	51.1/48.9	50.4/49.6	50.4/49.6
Sample size (surveyed households)	3074	3010	3097
Sample size (scanned equipment)	2127	2092	2131
Time frame	December 2013/January 2014	February 2014/March 2014	April 2014/June 2014
Sampling error of surveys	±1.77%	±1.79%	±1.76%

- **Module C: Security incidents.** Contains questions about security incidents experienced by users on their computers, e.g. malware/adware or virus infections suffered in the last quarter and their gravity.
- **Module D: Fraud.** Contains questions related to fraud experienced by users, e.g. related to phishing or online shopping (non-received or counterfeit goods).
- **Module E: Smartphone security.** Measures the actions of users with respect to smartphone security, e.g. app downloads or Internet services used, backup of important information, use of passwords or PIN numbers, etc.
- **Module F: Wi-Fi Security.** Measures the actions of users with respect to Wi-Fi security, e.g. connection to public Wi-Fi networks and wireless protocols used.
- **Module G: Opinion.** Measures the opinion of the user about confidence in the Internet, e.g. rating how confident the user feels when surfing the Internet, their awareness of responsibility and the measures that should be taken by public administrations to improve cybersecurity.
- **Module H: Behavior.** Checks the user's behavior. It describes whether the user's behavior is appropriate in relation to the use of different Internet services, e.g. compliance with security measures when using online banking, Peer-to-Peer (P2P) downloads or social networks.

Questions are multiple choice and their number of possible answers vary from question to question (between 2 and 16). Therefore, a conversion from this nominal or categorical type into a numerical type is needed to calculate statistics and to be able to use machine learning and visualization techniques. The idea is to scale the variables between 0 and 100. In this way each user is assigned a numerical value in each one of the survey modules depending on their behavior in that module (a high value if the behavior is considered positive and a low one if it is considered negative). The next step is averaging all the scores obtained for questions of the same module. As a result, an averaged score is obtained for each one of the 8 modules. This recoding into 8 variables reduces the computational costs and improves interpretability.

The data presented in this study was extracted from the following sources:

- *Declared data*: obtained from the online surveys aimed at households that defined the study sample.
- *Real data*: obtained from a scanning software, which analyzes systems and the presence of malware on computers.

The scanning software used was iScan (later renamed as Pinkerton), developed by the leading security company Hispasec.[1] This software is based on the transparent and joint use of 50 antivirus engines.

The iScan software was installed on users' computers and was used to analyze them, detecting malware residing on the computers and collecting data from the operating system. Moreover, it analyzed the security tools installed on the computers and their update status. Users had an open telephone support line for any issues they may encounter. The reward system for users consisted in points that could be redeemed as gift vouchers.[2] With respect to data privacy and ethics, in Spain, citizens above 15 years old can participate in this kind of study without consent from their parents. When installing the software, all users are fully informed about the privacy policies with regard to the collected data.[3]

Table 1 provides information about the data contained in each of the surveys. The different number of surveyed households per questionnaire is caused by users who quit the study but are replaced with new users. The number of dropouts for the surveys was 188, but 211 new users were recruited throughout the study. Around 31% of users who took the surveys do not have measurements from the scanning software. The reason for this is that a computer scan takes about 1 h and there are users who do not complete the scanning process. Due to dropouts, the number of missing scans during the study was 124, but 128 new ones were registered from recruited users.

Table 1 shows: (a) the number of users in the different periods of time, (b) the period when the survey was conducted, (c) the number of users whose computer was scanned and (d) the error calculated for the surveys assuming a simple random sampling criterion for dichotomous variables in which the probability of success (p) and the probability of failure

[1] https://hispasec.com

[2] More information about the reward system can be found in the ASKGFK website: https://www.askgfk.es/index.php?id=41

[3] https://hispasec.com/pinkerton/policy.html

(q) are p = q = 0.5, where a confidence level of 95% was used.

3.2. Feature engineering

Due to the size of the database, the presence of noise, inconsistent and redundant data, etc., the use of preprocessing techniques and feature engineering was required.

Feature engineering is one of the most important tasks in the *Knowledge Discovery in Databases* (KDD) (Liu and Motoda, 1998). The aim is to obtain a data set of such quality that it may be used to develop models, patterns or rules of higher quality. The importance of data preparation is reflected in the following three aspects (Han et al., 2011):

- The real-world data can be incomplete, inconsistent, or noisy.
- The preprocessing step generates data sets that are smaller than the original set, which can significantly improve the efficiency of data mining algorithms.
- The preprocessing step results in better quality data due to the recovery of incomplete instances, correction of mistakes or resolution of conflicts.

The first step of the data preprocessing consists of the integration of several questions as a single block. In this step, a total of 8 blocks were constructed from the original questions, corresponding to each one of the modules present in the polls. We also used 6 descriptive variables from each user (Gender, Age, Occupation, Education level, Operating System and Browser). It should be noted that, for the final database, only those questions, and therefore modules, matching throughout all quarters were taken into account. This approach was used because the main objective was to carry out an overall study of all surveys conducted in the different time periods. Thus, from all the variables listed in the original data set (320 variables), 14 new ones were ultimately obtained in the feature engineering step, namely, *Gender, Age, Occupation, Education level, Operating System, Browser, Use of the Internet, Measures and Internet security habits, User behavior, Security incidents, Smartphone security, Fraud, Wi-Fi Security*, and *Opinion about confidence in the Internet*. The integration of the data mentioned above can be divided, broadly, into two major tasks:

- **Numerical re-coding.** To integrate the database, the first decision is to carry out a re-coding of values. This database contains mostly responses to a questionnaire in which the user has several choices to answer. Those options range from 2–16 possible answers depending on the question. Moreover, the responses are not always ordered in the same way (i.e., the first option is the most positive answer and the last one the most negative). Additionally, the nature of the responses is different. For example, one question may evaluate a certain security issue in which the response may have 3 choices. Among many other types, there are questions in which users must select the Internet services or antivirus tools that they use from a list. Therefore, it was crucial to convert nominal or categorical variables to numeric variables to compute averages and then enable the use of data visualization techniques. The idea is to scale the variables (corresponding to each question) from 0 to 100. In this way, a numerical value is assigned to each user in each of the modules of the survey depending on their behavior in that module. A high value corresponds to a good performance/score and a low value to a bad one, regardless of whether the block evaluates a positive or negative aspect. After the coding process, all variables were grouped according to the block to which they belonged.
- **Average of variables.** Once the numerical re-coding is carried out, each user is described by one numerical value (between 0 and 100) for each of their variables (responses from each module). This value indicates how good or bad the user behavior is on a module, according to the responses given. Therefore, after obtaining this numerical matrix, the next step was to group all variables/questions according to their membership to each of the 8 blocks mentioned above: (*Use of the Internet, Measures and Internet security habits, User behavior, Security incidents, Smartphone security, Fraud, Wi-Fi Security* and *Opinion about confidence in the Internet*). To carry out this grouping, all the numerical values (calculated in the numerical re-coding stage) for each variable/question belonging to a specific module for each user were averaged. Thus, a considerable reduction in the number of variables was achieved, obtaining new ones that were able to describe the user behavior more clearly. Finally, a numerical value that indicates the overall performance of a user in each of the modules was obtained.

3.3. Integration of the data from the scanning software

This section describes the coding performed for variables obtained from the software used to automatically analyze the systems of the survey respondents. Of all the variables captured by the scanning software, the ones ultimately used were *Risk* and *Total Infections* because they summarize all the variables acquired by the software and describe the real user behavior. As its name suggests, the *Risk* variable is an index calculated from other variables from the scanning software, and indicates the level of risk. On the other hand, *Total Infections* is an index related to the total number of actual infections that are present in the analyzed computer. With regard to the variable *Risk*, note that it was coded between 0 and 100 to maintain consistency with the remaining variables of the survey. A value of 0 indicates no risk and a value of 100 indicates maximum risk. Regarding the variable *Total Infections*, it was recalculated using a logarithmic scale:

$$log_{10}\,(Total\ Infecctions + 1) \tag{2}$$

This decision was made because some users presented very extreme values (reaching 106 infections), but the vast majority of the remaining users presented lower values (an order of magnitude below). Therefore, the logarithmic scale allows the observation, in the case of SOM, of all values without excessive bias in the color map. Note that before applying the *log*10, 1 is added to the number of infections; the new recoding remains 0 if the number of infections is nonexistent. It is noteworthy that, for the variables concerning the scanning software: (*Risk* and *Total Infections*), high values mean a

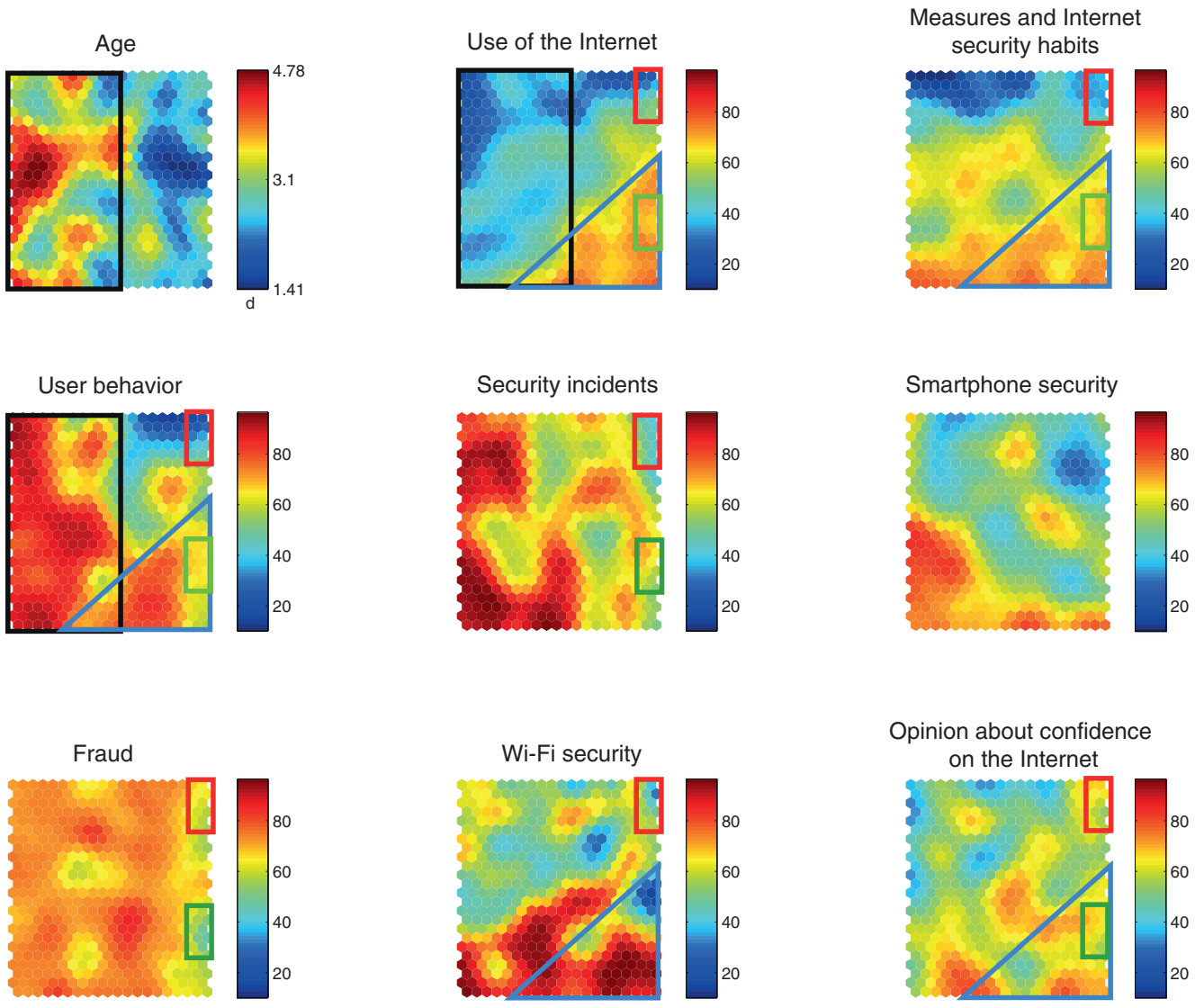


Fig. 2 – Component maps obtained after training SOM with the data from survey after the "fusion". Warm tones indicate a good behavior in that component and cold tones a bad one.

negative situation (high risk or high number of infections), as opposed to the variables obtained from surveys.

4. Results of the survey on cybersecurity

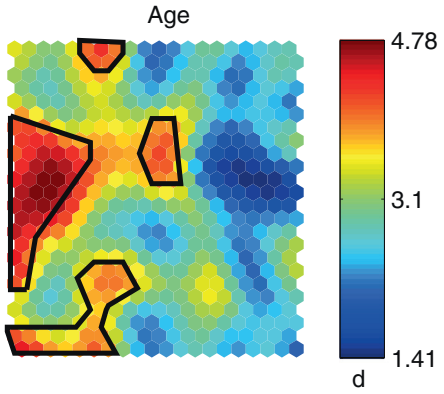
4.1. *Global results*

This section presents the results obtained from the cybersecurity survey after applying the SOMs to the dataset described in Section 3. Once the final dataset was built after the data fusion process explained in Section 3.2, several SOMs were trained. For the training, different options of the tuning parameters of the SOM algorithm were tested, combining all the possibilities (*Weight Initialization*, *Neighborhood Function* and *Type of Training*). Furthermore, the random initialization was fulfilled 100 times for each combination of parameters. Finally, the SOM that showed the least topographic error was selected (Kiviluoto, 1996). The topographic error measures the topology preservation between the original feature space and the final feature space.

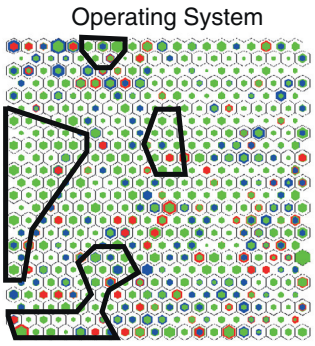
Fig. 2 shows the components of the SOM map obtained after performing the training algorithm, as explained in Section 2. In order to interpret the maps, it is worth noting that the data patterns to analyze fall in the same spatial zone for all component maps of the SOMs.

Different areas of interest have been highlighted to assist in the interpretation of the results. After observing this figure, the following conclusions can be drawn:

- Taking into account the area delimited by the black rectangle, it can be observed that older users (warm tones in the *Age* map) are usually the least likely to use Internet services (cold tones in the *Use of the Internet* map).
- Also in this area, older users (enclosed by the black rectangle) presented the best behavior (warm tones in the *User*



(a) *Age* component extracted from Figure 2.



(b) Winners' map labeled by colors according to the Operating System. Windows (green), Unix/Linux (red), Macintosh (blue).

Fig. 3 – Relationship between the age of the users and the Operating System they use.

behavior map). Regarding young users, there are both good and bad profiles.

- The area bounded by the blue triangle gathers users with a greater use of the Internet (see *Use of the Internet* map). These users obtain better security scores both on the Internet and on Wi-Fi, with the exception of a small area of interest that will be explained below. Furthermore, these users have good behavior and think that the Internet is trustworthy (they have a good opinion about confidence in the Internet).
- The smaller area defined by the green rectangle shows the worst scores in the *Fraud* map (higher level of fraud). This area gathers users with the highest use of the Internet, with good security habits and good behavior. Despite the bad scores on the *Fraud* component, they present medium and high scores in *Security incidents* map. Although they present fraud, these users do not have a bad opinion about their confidence in the Internet. This observation may suggest that fraud is not really associated with security habits, but it is with confidence in the Internet. That means that being a victim of fraud can be related to their overconfidence in the Internet regardless of their use of security tools to protect their computer.
- Similarly, the red rectangle limits a small area of great interest because it is the area where the worst scores regarding security incidents are found (greater number of incidents). Users located in this area present a profile of low use of the Internet, poor security measures on the Internet and Wi-Fi and bad behavior. They also had intermediate values in the *Fraud* map. However, they do not have a very bad opinion about confidence in the Internet. This finding may be explained by the shared use of equipment in several households, which results in a questionnaire answer representing the opinion of just one of them. Users falling in this profile may need to become more aware of their need for better protection, as they tend to have a good opinion about security despite having some real problems. Moreover, they usually underestimate the probability of fraud because they are infrequent users, but the reality is that the problem is related to the use of the correct security tools rather than the amount of time spent on the Internet.

It is remarkable that all of these qualitative conclusions can be directly visualized because of the special characteristics of SOMs, and would otherwise be harder to interpret from other numerical and statistical analysis tools, as it creates low-dimensional, highly interpretable maps. Moreover, this analysis methodology reduces the subjectivity of the conclusions drawn.

Besides the study of the component maps obtained after training the SOMs, the winner neurons map was used to gather further information. In this map, the colored area inside each hexagon is proportional to the number of input patterns (users surveyed) that are most similar to this neuron. If multiple *hits* are drawn in different colors, it is possible to compare the different patterns associated with different classes by the distribution of their *hits* on the map. In this way, a quantitative idea of the number of input vectors belonging to each neuron for each class is obtained. Thus, it is possible to observe how the data associated with different classes are distributed in the SOM. Fig. 3b shows the distribution, in different colors, of the users surveyed by the SOM depending on the operating system they used. That distribution, analyzed together with the map of components, shows the relationships between user behavior, their concept of informatics security and the operating system used. Thus, if there are clear trends (isolated groups), their characterization or behavior can be assessed by observing the values of each variable in such areas of the component maps.

Fig. 3 focuses on drawing conclusions about the operating system used in relation to user age. As observed, the vast majority of older users typically use operating systems in the Windows family (shown in green) instead of Unix/Linux (red) or Macintosh family (blue). It is also clear that the predominant operating system used by the users surveyed belongs to the Windows family.

Besides the study of the operating system on the SOM, colored winner maps of the variables *Gender, Occupation, Education level* and *Browser* were obtained. These figures are not shown in this paper because they did not lead to conclusive results. All classes were completely overlapped all over the map, so a specific profile for those classes did not show up in the map.

The analysis of the overall survey results obtained by the SOM suggests that there are certain areas of interest to study

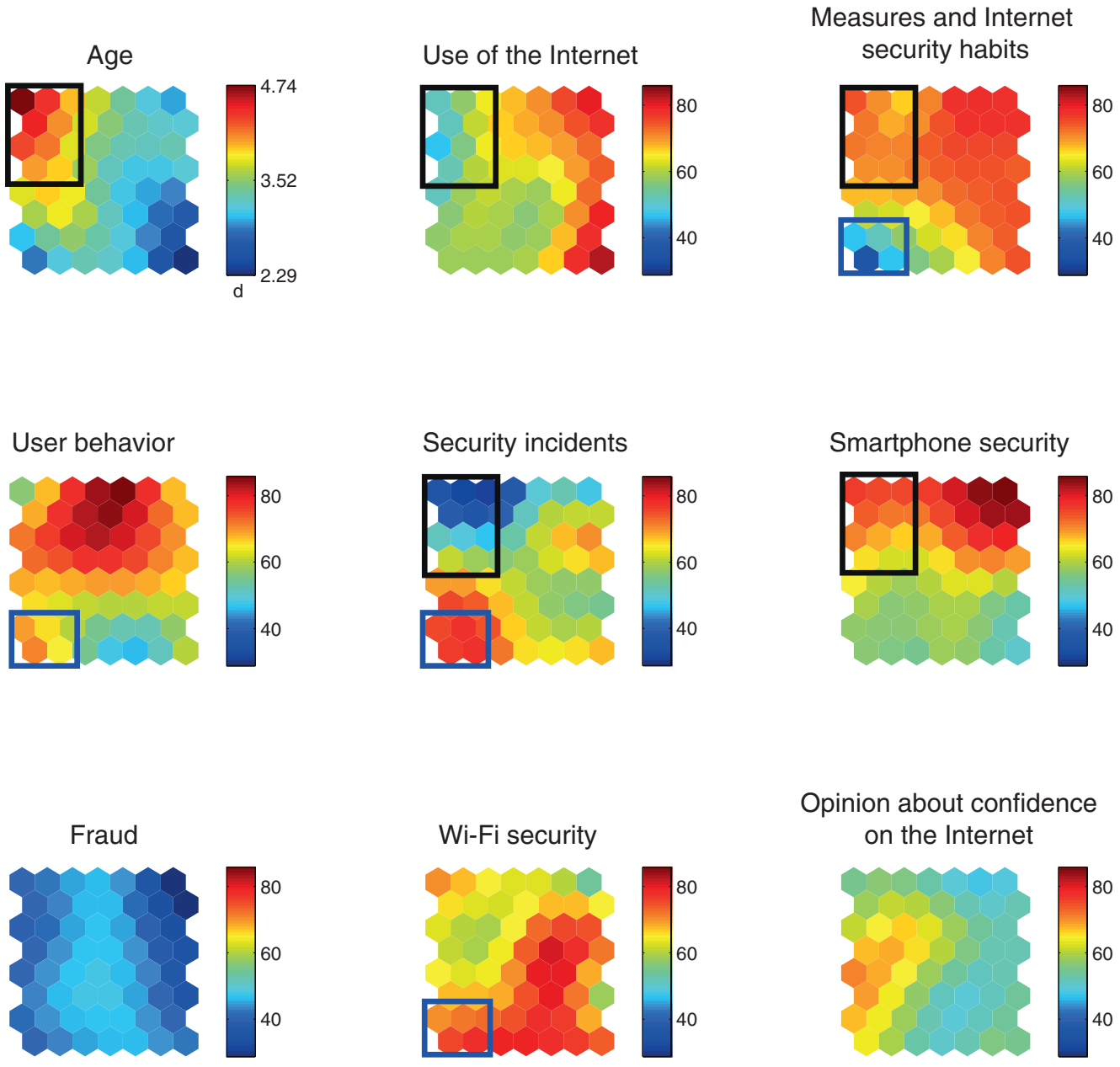


Fig. 4 – Component maps obtained after training SOM with the subset of data corresponding to the area of highest fraud. Warm tones indicate a good behavior in that component and cold tones a bad one.

in depth as they are directly related to computer security risk exposure. These areas correspond to areas defined by the green and red rectangles in Fig. 2. As mentioned above, these areas are of particular interest because they represent the lowest scores in the blocks corresponding to *Fraud* and *Security incidents*. To provide better insight, the next goal was to extract users located in each one of those specific areas for a new training with the SOM algorithm. Section 4.2 presents the results obtained from this procedure.

4.2. *Local results: areas of interest.*

These local results are focused on those users who presented the worst scores in either fraud (most fraud across all surveys, i.e. the green rectangular area) or in terms of security incidents (most security incidents, i.e. the red rectangular area).

4.2.1. *Analysis of the area of maximum fraud levels.*

The area of the SOM that presented the highest levels of fraud (lowest scores) appears in the lower right region of the maps. Fig. 4 shows the component maps of the SOM obtained after performing the training algorithm on this data subset. When observing this figure, the following conclusions can be drawn:

- The zone marked by the black square represents older users with a low use of Internet services.

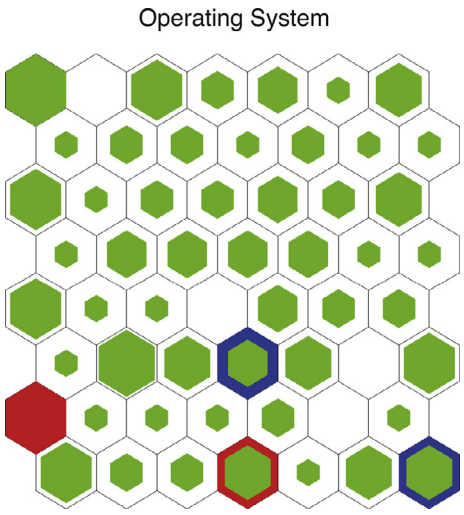


Fig. 5 – Winners map labeled by colors according to the Operating System for the local study corresponding with the area of highest fraud. Windows (green), Unix/Linux (red), Macintosh (blue).

- We observe that older people also present the worst scores in security incidents. This may be due to the lack of awareness of new technologies and, in particular, of the Internet.
- It is worth pointing out the components *Security incidents and Measures* and *Internet security habits*; as seen in the area framed by the blue square, a bad score (cold tones) in security habits entails the best scores in security incidents (small number of incidents). This finding has several possible explanations. On one hand, the computer could be used by several people, with one of them answering the questionnaire and others being responsible for maintaining computer security.
  It is noticeable, however that *Behavior* and *Wi-Fi security* present good scores in this area. The same situation, but reversed, occurs for the area enclosed by the black square: a good score in security involves a bad score in infection. This finding is noteworthy since it is expected that security incidents and security measures are somehow correlated. As discussed above, this may be caused by the lack of knowledge of the users about their level of infection or their security measures, or simply because they did not have all the necessary information to provide correct answers in the survey.

As in the case of the global study of the surveys, the winners map was used to obtain further information. Fig. 5 shows the distribution of different users depending on the operating system used in different colors in the SOM. As observed in the global study, the vast majority of users located in the area of the highest fraud used operating systems from Windows family (green), especially older users.

4.2.2. *Analysis of the area of maximum security incidents.* This section focuses specifically on the area of the SOM that presented the worst scores in security incidents; this area corresponds to the upper right part of the maps.

Fig. 6 shows the component maps of the SOM obtained after training the algorithm on this data subset. From the inspection of the Figure, the following conclusions can be drawn:

- The vast majority of young people show a greater use of Internet services (see the black square in the first two components).
- In this area, young people present the best behavior (*User behavior* component).
- In the area delimited by the red rectangle, the users with the lowest scores in security incidents (fifth component) have high scores on measures and security habits (third component) and vice versa. This may be because they are actually not aware of the fact that their computers are infected. As a consequence, their answers do not reflect reality. They are users with an optimism bias (Matheson et al., 2008; Sharot, 2011) whose answers tend to be more positive than the true facts.
- In spite of the fact that the users located in this area under study present a higher number of *Security incidents*, almost none of them suffered fraud (high scores reflected by red in the entire components plane corresponding to *Fraud*). It could be for this reason that they have a relatively high confidence in the Internet.

As in the previous cases, the winners map was used to obtain further information. Fig. 7 shows the distribution of different users depending on the operating system used in different colors in the SOM. Again, it can be observed that the vast majority of users within the area of the highest number of incidents used operating systems from the Windows family (green). None of them used Unix/Linux.

5. Results of the survey on cybersecurity together with the variables from scanning software

5.1. *Global results*

This section presents new results of the SOM after including new variables in the training stage obtained from a scanning software on the users' computers. Specifically, the same database as in the previous study was used, but adding two variables from such software. These two variables summarize all variables measured by the scanning software, namely, *Risk* and *Total infections*. In this way, it is possible to compare the responses of users with the actual status of their computers.

For the variables concerning the scanning software, i.e. (*Risk* and *Total Infections*), high values mean a negative situation (high risk or high number of infections), unlike variables from the surveys. In the case of variables from surveys, scores for each module were computed. The higher the score, the better the performance, regardless of whether the block evaluates a negative aspect. That is, a value of 100 in the module corresponding to *Fraud* leads to a positive occurrence in that module (the user did not suffer fraud) and not the opposite. Furthermore, it is worth noting that around 31% of users who conducted the surveys did not have any measurements

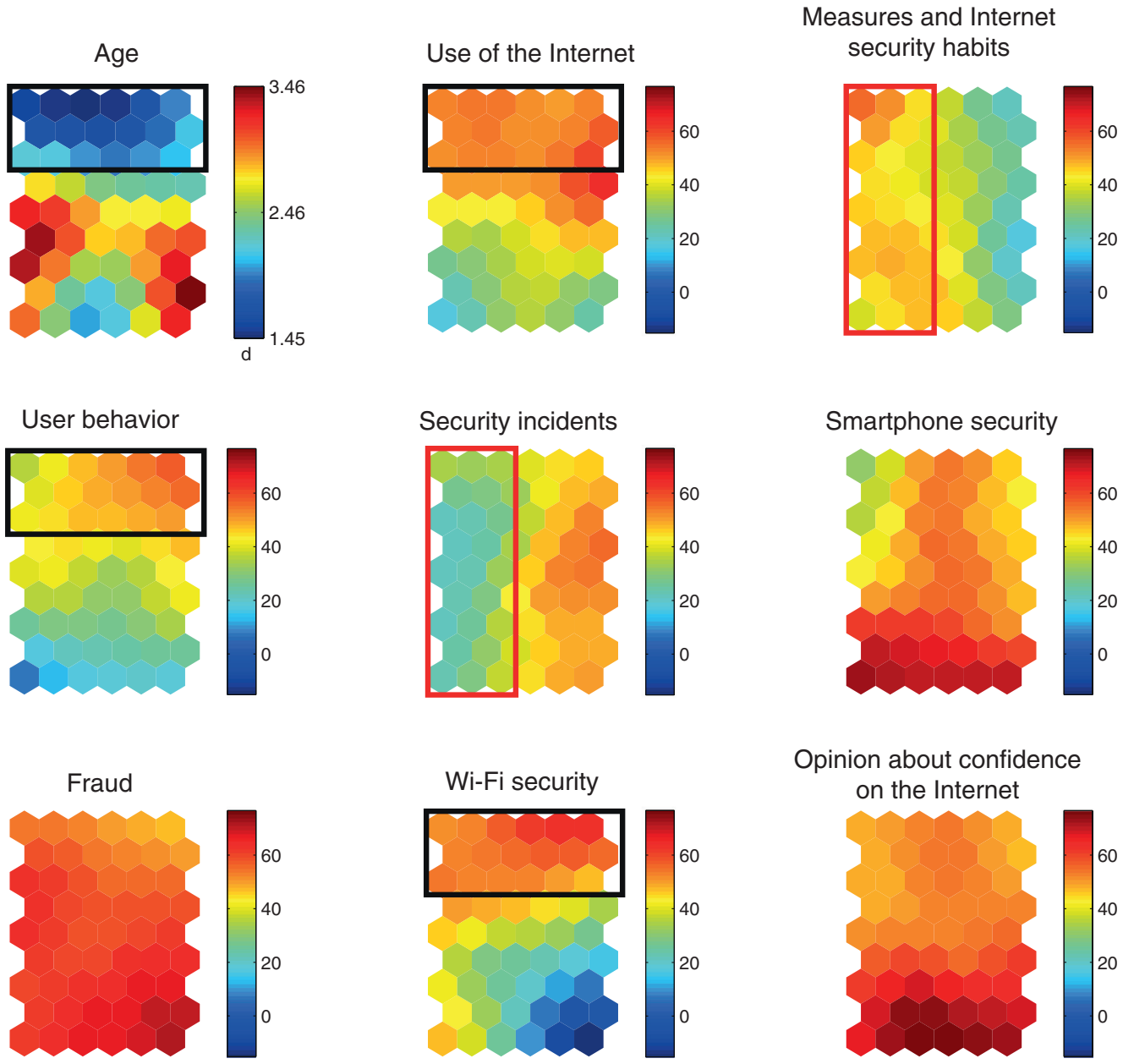


Fig. 6 – Component maps obtained after training SOM with the subset of data corresponding to the area with the highest number of incidents. Warm tones indicate a good behavior in that component and cold tones a bad one.

associated with the scanning software. An additional advantage of using SOM stems from the fact that it allows for a study of all users but without considering the missing values of these two variables for those users.

After training the SOM algorithm with this new set of data, the resulting component maps were represented (Fig. 8). From this figure, the following conclusions can be drawn:

- We observed that there are no distinguishable profiles for users presenting high risk and infections according to the scanning software. That is, in the remaining variables (which represent the age and the score associated with each module of the survey) the full range of values can be found in the black frame at the top. Therefore, all sorts of scores (high and low) are found for all the modules so no conclusion or concrete profile can be drawn for these users (who presented high number of infections and high risk from the scanning software). In some cases, the answers of the users do not correspond exactly to reality since there is no correlation between surveys and the software (e.g.: a bad score on behavior or security modules reflected by the survey does not always lead to high risk or infection in the scanning software).
- However, we noticed that the worst fraud scores (green spots representing intermediate values, delimited by a red rectangle) are found in the area of high risk according to the scanning software (top of the map). This justifies a further local analysis of this area.

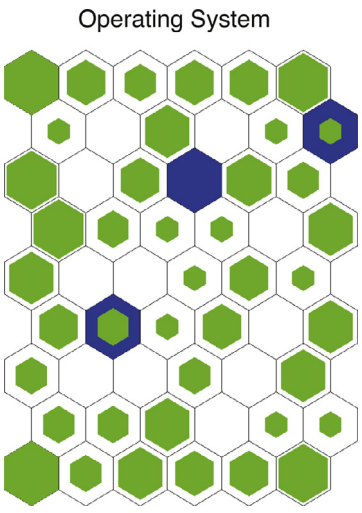


Fig. 7 – Winners map labeled by colors according to the Operating System for the local study corresponding with the area of highest security incidents. Windows (green), Unix/Linux (red), Macintosh (blue).

- The figure reveals that users with the worst score in the *Behavior* block of the survey are found in the zone of high risk and and medium infection according to the scanning software (top part of the map). In particular, those users are located in the upper left corner of the *Behavior* component where very low scores are seen in this block (cold tone). The number of infections of these users is not very high, which might be caused by their low use of the Internet (the top left corner of *Use of the Internet* map contains the lowest scores). However, the zone of high risk according to the software (top part of the map) also located users who, according to them, had very good behavior (warm tone). Furthermore, the figure shows some users presenting bad behavior (according to the survey) in the area of no risk and no infection according to the scanning software (bottom of the map), specifically in the colder tones area located in the left-center part of the *Behavior* component.
- With respect to *Age* and *Total infections*, young users have a higher number of infections than older ones (the area bounded by the red rectangle is dominated by blue tones in the *Age* map). Moreover, young users presented the lowest scores in fraud (medium level of fraud) as reflected by the three green spots delimited by the red rectangle on the

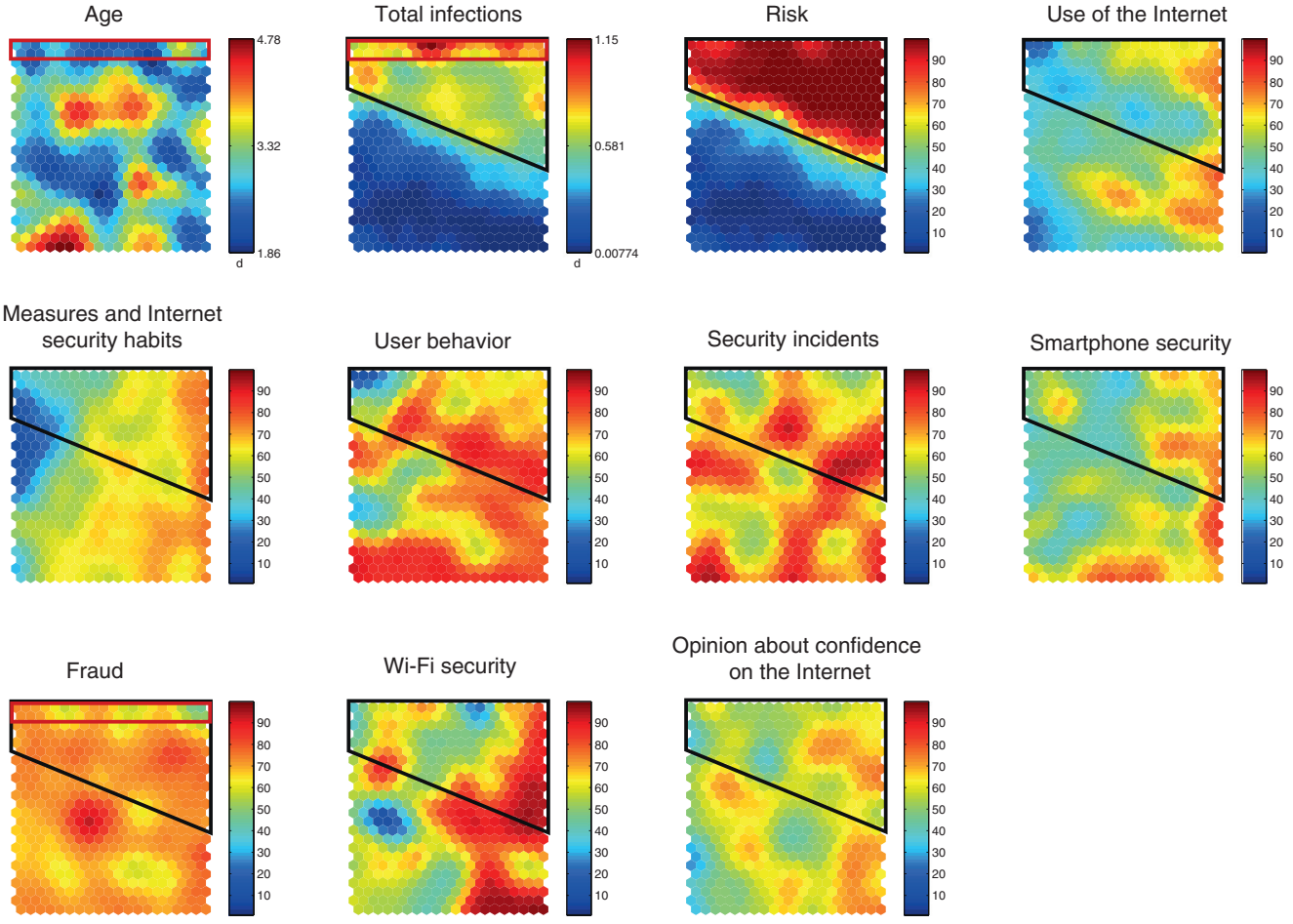


Fig. 8 – Component maps obtained after training the SOM with the data from survey along with the variables acquired from the scanning software. Warm tones indicate a good behavior in that component and cold tones a bad one.

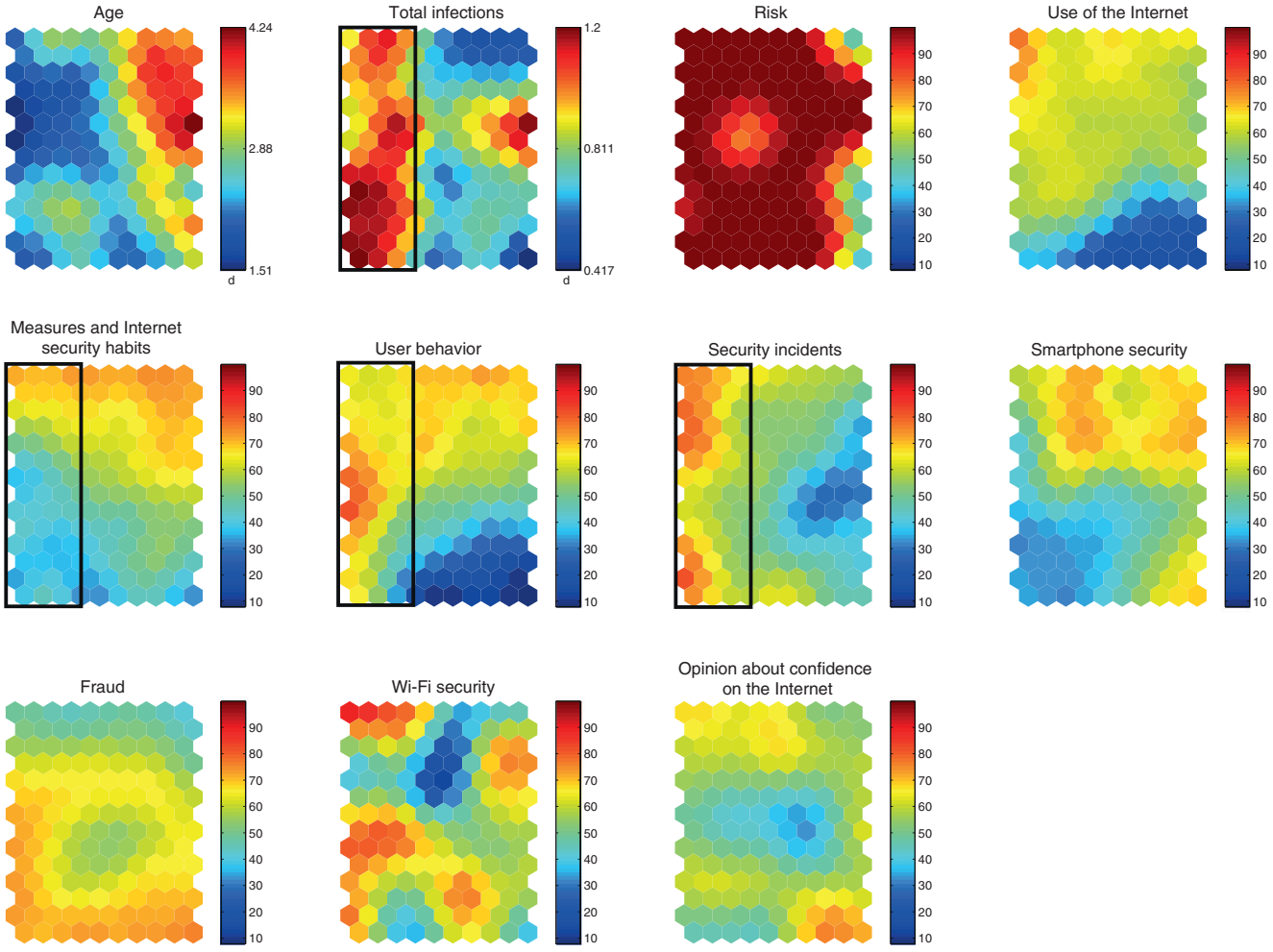


Fig. 9 – Component maps obtained after training SOM with the subset of data corresponding to the area of the highest number of infections according to the scanning software. Warm tones indicate a good behavior in that component and cold tones a bad one.

Fraud component. This fact reflects that there is not always a correlation between *Total infections* and *Fraud* due to: a) there is a lack of protection for browsing but not for other operations; b) youngsters hardly carry out any transactions.

5.2. *Local results*

A local analysis of the high Risk area (black rectangle) is not useful as it is only directly correlated with a high number of infections, which is logical. On the other hand, the maximum number of infections (red rectangle) is open for deeper interpretations that we analyze below.

5.2.1. *Analysis of the area of maximum number of infections.* This section focuses specifically on the area of the SOM that presented the largest number of infections.

Fig. 9 shows the component maps of the SOM obtained after performing the training algorithm on this data subset. After observing this figure, the following conclusions can be drawn:

- By observing the area delimited by the black rectangle, it can be confirmed that the vast majority of users who were most infected according to the scanning software (*Total infections* component) are young people. Additionally, they were those with the lowest number of security incidents according to the surveys (high scores on the *Security incidents* map). This may seem illogical since the higher the number of actual infections measured by the scanning software, the worse the scores extracted from the surveys in the security incidents module are. This might be caused by users lacking the necessary information to provide a correct answer in the survey, and hence, not being objective or ignorant of the level of infection found in their computer.
- Those users showed good behavior in the surveys, which was unexpected either because of the high level of actual infection. However, most of them have a low score in the module *Measures and Internet security habits*, which coincides with the fact that the total number of infections is high.

6. Conclusion and future work

This paper has proposed the use of Self-Organizing Maps (SOM) to analyze the current state of digital confidence among Spanish Internet users and, at the same time, to contrast the level of incidents suffered by their computers together with the users' perceptions. SOM is an artificial neural model that allows the combined visualization of *N*-dimensional patterns and is therefore particularly suitable for this task. The proposed tool has allowed drawing qualitative conclusions from a highly subjective data set, because of the ability of SOMs to extract hidden information from multivariate, complex data.

Regarding the methodology, two sources of information have been used: an online questionnaire and the malware scan of computer equipment. This approach has been used to contrast reality with the users' assumptions about security. A limitation about both the data collected from the questionnaire and from the computer scan is that they were evaluated in 2014, and users' behavior and scan results could have changed since.

In summary, this study has allowed the collection and interpretation of real indicators of infection/fraud both from objective and subjective data sources, which is a novelty in this kind of work. The analysis has proved the correlation between some features but has also revealed the inconsistencies between users' conception about their security and risk and what they are actually exposed to. In addition, the use of a visual data mining tool such as SOM has revealed qualitative relationships between the questionnaires and the level of infection present in the users' computers, which is another novelty of this work.

Given the usefulness of SOMs for this kind of study, a logical future line of work is the application of the same methodology in mobile phone security, as more people are using their smartphones to connect to the Internet.

REFERENCES

Arachchilage NAG, Love S. A game design framework for avoiding phishing attacks. Comput Hum Behav 2013;29(6):706–14.

Bashir M, Wee C, Memon N, Guo B. Profiling cybersecurity competition participants: self-efficacy, decision-making and interests predict effectiveness of competitions as a recruitment tool. Comput Secur 2017;65:153–65.

Campbell J, Greenauer N, Macaluso K. Unrealistic optimism in internet events. Comput Hum Behav 2007;23:1273–84.

Davinson N, Sillence E. It won't happen to me: promoting secure behaviour among internet users. Comput Hum Behav 2010;26:1739–47.

Donaldson SE, Siegel S, Williams C, Aslam A. Enterprise cybersecurity how to build a successful cyberdefense program against advanced threats. Apress; 2015.

Gordon LA, Loeb MP, Lucyshyn W, Zhou L. The impact of information sharing on cybersecurity underinvestment: a real options perspective. J Account Public Policy 2015;34(5):509–19.

Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Morgan Kauffman; 2011.

Hassoun MH. Fundamentals of artificial neural networks. Cambridge, MA, USA: MIT Press; 2004.

Haykin S. Neural networks and learning machines. 3rd. Prentice Hall; 2009.

Herrero J, Urueña A, Torres A, Hidalgo A. My computer is infected: the role of users sensation seeking and domain-specific risk perceptions and risk attitudes on computer harm. J Risk Res 2017a;20(11):1466–79.

Herrero J, Urueña A, Torres A, Hidalgo A. Smartphone addiction: psychosocial correlates, risky attitudes, and smartphone harm. J Risk Res 2017b:1–12. Published on-line doi: 10.1080/13669877.2017.1351472.

Hoonakker P, Bornoe N, Carayon P. Password authentication from a human factors perspective: results of a survey among end-users. In: Proceedings of the human factors and ergonomics society 53rd annual meeting San Antonio. SAGE; 2009. p. 459–63.

Humaidi N, Balakrishnan V. Exploratory factor analysis of user's compliance behaviour towards health information system's security. J Health Med Inform 2013;4(2):2–9.

Jang-Jacard J, Nepal S. A survey of emerging threats in cybersecurity. J Comput Syst Sci 2014;80(5):9763–993.

Kelly AE, Schochet T, Landry CF. Risk taking and novelty seeking in adolescence. Ann New York Acad Sci 2004;7:27–32.

Kiviluoto K. Topology preservation in self-organizing maps. In: Proceedings of the 1996 IEEE international conference on neural networks (Cat. No. 96CH35907); 1996. p. 294–9.

Knowles W, Prince D, Hutchison D, Disso JFP, Jones K. A survey of cyber security management in industrial control systems. Int J Crit Infraestructure Protect 2015;9:52–80.

Kohonen T. Self-organization and associative memory: 3rd ed. New York, NY, USA: Springer-Verlag New York, Inc.; 1989.

Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Springer; 1998.

Matheson SM, Asher L, Bateson M. Larger, enriched cages are associated with 'optimistic' response biases in captive european starlings (sturnus vulgaris). Appl Animal Behav Sci 2008;109(2–4):374–83.

Rossi F. Visual data mining and machine learning. In: Proceedings of European symposium on artificial neural networks (ESANN); 2006. p. 251–64.

Runz C, Desjardin E, Herbin M. Unsupervised visual data mining using self-organizing maps and a data-driven color mapping. In: Proceedings of 16th international conference on information visualisation; 2012. p. 241–5.

Sharot T. The optimism bias. Current Biology 2011;21(23):R941–R945.

Shillair R, Cotten S, Tsai H, Alhabash S, Rifon N. Online safety begins with you and me: Convincing internet users to protect themselves. Comput Hum Behav 2015;48:199–207.

Singer P, Friedman A. Cybersecurity and cyberwar: what everyone needs to know. Oxford University Press; 2014.

Stanton J, Stam K, Mastrangelo P. Analysis of end user security behaviors. Comput Secur 2005;24:124–33.

Urueña A, Hidalgo A. Successful loyalty in e-complaints: FsQCA and structural equation modeling analyses. J Bus Res 2016;69(4):1384–9.

Vesanto J, Alhoniemi E, Himberg J, Kiviluoto K, Parviainen J. Self-organizing map for data mining in Matlab: the SOM toolbox. Simul News Europe 1999(25):54.

Whitty M, Doodson J, Creese S, Hodges D. Individual differences in cyber security behaviors: an examination of who is sharing passwords. Cyberpsychol Behav Soc Netw 2015;18(1):3–7.

Alberto Urueña, PhD, is associate professor at the Universidad Politecnica de Madrid. He is interested in the analysis of personality and behavioral problems. Alberto has published on computer risk-taking, computer use and security and is author of several books and papers about technology management published in different international journals.

Fernando Mateo obtained a degree in Telecommunication Engineering from the Polytechnic University of Valencia in 2005, and a

Ph.D. in Electronics Engineering from the same University in 2012. At present, he works as an Assistant Professor and data scientist at the Intelligent Data Analysis Laboratory, at University of Valencia. His research focuses on data mining and preprocessing, feature selection, machine learning models both for regression and classification, clustering and time series forecasting.

Julio Navio-Marco holds a M.Sc. in Telecommunications Engineering; BA and PhD in Economics and Business Administration at the UNED; and is Postgraduate in IESE Business School. Julio NavÃo is Professor of Business Organization, Economics of Telecommunications and Entrepreneurship at the UNED in Spain. Guest speaker for Digital Economy, Regulation and Telecommunications Policy in Carlos III University of Madrid and Polytechnic Univ. Madrid. Dr Navio is Deputy Dean of the Spanish College of Telecommunication Engineers and Vice-President of the Spanish Association of Telecommunication. Dr Navio is also expert for the EC Directorate-General for Regional and Urban Policy (DG REGIO) and H2020.

José Martínez Martínez holds a Ph.D. Degree related to Data Visualization from the University of Valencia (Spain). He also received the B.Eng. degree on Electronic Systems for Telecommunication (2006), the M.Eng degree in Electronics (2009) and the M. degree in Electronic Engineering (2011), from the same University. He is currently researching at the Intelligent Data Analysis Laboratory, University of Valencia. His research interests are machine learning methods, data mining and data visualization. He has been working as a data scientist on various national and European research projects.

Juan Gómez-Sanchis received a B.Sc. degree in Physics (2000) and a B.Sc. degree in Electronics Engineering from the University of Valencia (2003). He joined at the Public Research Institute IVIA in 2004, developing is Ph.D. in hyperspectral computer vision systems applied to the agriculture. He joined to the Department of Electronics Engineering at University of Valencia in 2008, where he currently works as Senior Lecturer in pattern recognition using neural networks.

Dr. Joan Vila Francés has studied Technical Engineering in Telecommunication and Electronics Engineering, both finished with first class honors, at the University of Valencia (Spain). In 2009 he received a Ph.D. Degree in Electronics Engineering from the same university, where he currently holds a Senior Lecturer position. His teaching is focused on Programmable Digital Systems. He has skills on Industrial Systems, Digital Electronics, Computing, Database Management and Web development.

Antonio J. Serrano-Lopez received a BS degree in Physics in 1996, a MS degree in Physics in 1998 and a Ph.D. degree in Electronics Engineering in 2002, from the University of Valencia. He is currently an Associate Professor at the Electronics Engineering Department in this same university. His research interest is machine learning methods for biomedical signal processing. Currently, he teaches courses of analog electronic design and predictive analytics in healthcare.