Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility

Juan Ignacio Guerrero, Iñigo Monedero, Félix Biscarri, Jesús Biscarri, Rocío Millán, and Carlos León, *Senior Member, IEEE*

***Abstract*—The Endesa Company is the main power utility in Spain. One of the main concerns of power distribution companies is energy loss, both technical and non-technical. A non-technical loss (NTL) in power utilities is defined as any consumed energy or service that is not billed by some type of anomaly. The NTL reduction in Endesa is based on the detection and inspection of the customers that have null consumption during a certain period. The problem with this methodology is the low rate of success of these inspections. This paper presents a framework and methodology, developed as two coordinated modules, that improves this type of inspection. The first module is based on a customer filtering based on text mining and a complementary artificial neural network. The second module, developed from a data mining process, contains a Classification & Regression tree and a Self-Organizing Map neural network. With these modules, the success of the inspections is multiplied by 3. The proposed framework was developed as part of a collaboration project with Endesa.**

***Index Terms*—Data mining, decision tree, neural network, non-technical losses, power utility, text mining.**

I. Introduction

A NON-TECHNICAL loss (NTL) in a power utility is defined as any consumed energy or service that is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of this equipment. NTLs are caused by breakdown or illegal manipulation in customer facilities. These types of losses are very difficult to predict.

Normally, utility companies use massive inspections to reduce NTLs. The main problem is that these companies do not have the necessary technology to carry out a deep processing of this information before carrying out these inspections. Thus, although the utility companies expend effort to detect and correct this type of anomaly, the main focus of their work is other topics as the maintenance of infrastructure.

Manuscript received May 17, 2016; revised September 28, 2016, January 10, 2017, February 28, 2017, and May 14, 2017; accepted June 16, 2017. Date of publication June 29, 2017; date of current version February 16, 2018. This work was supported by the backing of SIIAM project (Reference Number: TEC2013-40767-R), funded by the Ministry of Economy and Competitiveness of Spain. Paper no. TPWRS-00755-2016. *(Corresponding author: Juan Ignacio Guerrero.)*

J. I. Guerrero, I. Monedero, F. Biscarri, J. Biscarri, and C. León are with the Department of Electronic Technology, University of Seville, Seville 41011, Spain (e-mail: juaguealo@us.es; imonedero@us.es; fbiscarri@us.es; jbiscarri@us.es; cleon@us.es).

R. Millán is with Department of Automated Metering Management and Field Works, Endesa, Seville 41004, Spain (e-mail: rocio.millan@endesa.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2017.2721435

A type of regular methodology used by the power utilities to detect NTLs is based on the study of the customers that have null consumption during a certain period. This type of customer has in his null consumption the clearest sign of a non-technical loss. The problem of this methodology is that this customer does not always have an NTL. These cases could be due to private customers' empty houses or a drop in electrical demand in customers with some type of business. Therefore, some additional information about the customer, and not only his consumption, is often helpful in determining the reason for the null consumption. This information includes any data about incidences in the consumer facilities or the type of business in order to know if it is a business the demand for which is currently falling (e.g., now the building construction in Spain). Moreover, inspectors of the Company know that the following types of business are more likely to have drops in energy not due to a possible NTL: wells, lightings, irrigation pumps, water purification and construction (previously mentioned).

Conversely, it is known that data mining techniques [1] are currently being applied to multiple fields, and detection of NTLs is one field in which it has found with success [2]–[7]. These anomalies are frauds in telecommunication and financial sectors; breakdown or fraud in power, water or gas sectors, etc.

Midas Project is the name of a collaborative project between the Endesa Company (the main power utility in Spain) and the University of Seville whose objective is the detection of NTLs by means of artificial intelligence techniques. In this project the authors have been working for 8 years and some results from the project [8]–[10] have already been presented.

The work presented in this article arises from the need of the Endesa Company to improve its detection of NTLs amongst customers with null consumption. Specifically, the need comes from the low percent (approximately 5%) of success in the inspections in-situ carried out by the power utility with its original methodology. These poor results are due to the aforementioned fact that a customer with null consumption does not necessarily have some type of anomaly.

The objective of our work has been to take the methodology that Endesa uses to select inspections of customers with null consumption and to develop two modules based on text mining and data mining techniques in order to improve the results. The first module is a customer filtering based on a text mining process from the Endesa database. The second module was designed from a data mining process of the consumption of the customers and their contract information.

0885-8950 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

II. Bibliographical Review

The NTL problem is very similar to the problems in other sectors because the main objective is the detection of anomalous cases with anomalous information or uncommon data.

In the financial sector, one of the most similar cases is the detection of credit card fraud. Sahin and Duman [11] compares the performance of Artificial Neural Networks (ANN) and logistic regression methods, based on real data set. Seyedhossein and Hashemi [12] describe the problem of fraudsters in e-commerce sales. The authors try to detect the fraud at the transaction level. The authors propose the patterns inherent in the time series of aggregated daily amounts spent on an individual credit card account, decreasing the time just as the fraud occurs and when it is finally detected.

In the telecommunication sector, the cellular clone fraud is one of the more common causes of a mobile communication network, and it is very similar to NTL detection. Mohamed *et al.* [13] propose the use of a detection engine to minimize false cases, using a backpropagation neural network to perform telecommunication interpolation based on local telecommunication network services.

In the intrusion detection field, [14] introduces a basic adaptive boost algorithm and analyzes its drawbacks and then introduces an improved adaptive boost algorithm to classify the detected event as normal or intrusive.

Different data mining techniques are used to detect NTL. Nizar *et al.* [15] present a modern computational technique called Extreme Learning Machine (ELM). The authors propose ELM and Online Sequential-ELM (OS-ELM) algorithms to improve classification performance and to increase the accuracy of the result. A comparison of this approach with other classification techniques, such as the Support Vector Machine (SVM) algorithm, is also undertaken and the ELM performance and accuracy in NTL analysis is shown to be superior. On the other hand, Ramos *et al.* [16] propose a pattern recognition technique called optimum-path forest, including learning and pruning algorithms. dos Angelos *et al.* [17] propose a C-means-based fuzzy clustering for the classification of electricity consumption profiles. Nagi *et al.* [18] extend a previously published SVM [19] with the introduction of a Fuzzy Inference System (FIS). Spirić *et al.* [20] suggest using a rough set theory and gave a general approach to its use.

All these approaches address consumption data and some additional information about contract, supply characteristics and inspection results. The proposed solution was integrated with the traditional procedure of Endesa, taking advantage of consumption, contract, and supply information, as well as inspectors' commentaries.

Additionally, the proposed paper addressed a specific case of consumption: a case in which the consumer had null consumption. This pattern of consumption could be due to holidays, abandoned place, closed business, etc. These cases are usually discarded in a "preprocessing" or "customer filtering and selection" stage. [15] classifies this type of behavior as an anomalous consumption (false positive). [18] included a database query in which the consumptions less than 1.5 KWh and with a difference between maximum and minimum consumption was greater than 6.5 KWh. This condition discarded all zero consumption. [19] removes customers having no consumption (0 KWh) throughout the entire 25 month period and removes new customers registered after the first month in the data. [20] removes all customers with three or more consecutive zero consumption readings. [29] filters particular behaviors of the customer, like low demand periods due to holidays.

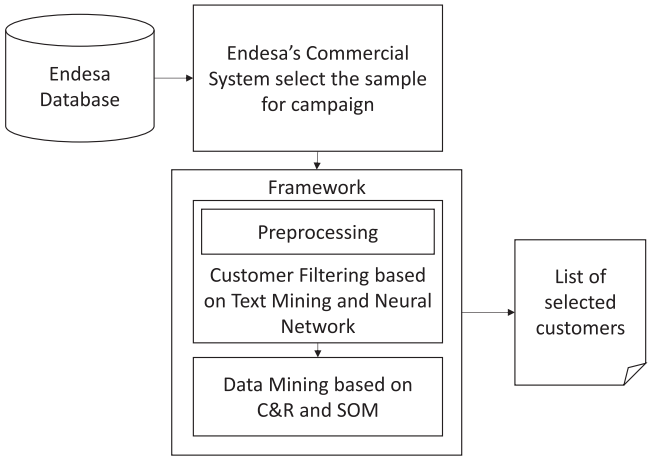


Fig. 1. General view of proposed framework.

III. Overview

A null consumption campaign contains information about consumers with a drop consumption or null one. This consumption behavior is usually treated by the traditional intelligent systems as an abnormality.

The proposed framework is shown in Fig. 1. The distribution company has a system to select customer by simple constraints. The campaign presented in this paper is the null consumption campaign, and it has the 3 or more last null consumptions constraint. The company usually inspects all customers in this selection. This sample is loaded in the framework with the information about contracts, historical consumption, inspectors' commentaries, and supply technical characteristics.

In the framework, the Customer Filtering based on Text Mining and Neural Network mainly analyzes the information of inspectors' commentaries. The preprocessing sub-module filters customers without enough information about customers. The Customer Filtering module filters customers according to the final classification of process.

The Data Mining stage analyzes the historical consumption of the customer, using a model based on Classification & Regression (C&R) and Self-Organizing Map (SOM) [21]. The final result of this module is the list of selected customers for inspection.

IV. Methodology

The proposed framework provided a new methodology to increase the inspections and to reduce NTLs. This methodology is made up of the following steps:

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

1) Check the available information:
   a) Commentaries from inspectors and/or company staff.
   b) Information about consumption.
   c) Contract information: contract power, tariff, economic activity, time discrimination band, billing period, and postal code.
2) The language for concept extraction process is configured.
3) A Detailed Concept Dictionary based on unstructured information (commentaries from inspectors and/or company staff) from whole company database is created.
4) A sample is selected from whole company database, to increase the efficiency of data mining techniques; the analysis is focused on consumers with one or more characteristics in common (geographic location, economic activity, etc.).
5) Customer filtering (Section V) is performed over the selected sample. All customers are classified in different categories. The INCORRECT category usually identifies the customers that should be analyzed by data mining process. Sometimes, the analysis of LOW CONSUMPTION category could provide some results, but the probability to find an NTL is very low.
6) The INCORRECT consumers are analyzed in a data mining module:
   a) The consumption of each consumer is normalized.
   b) A filter by drop consumption is applied (Section IV-B).
   c) A filter by contract use is applied (Section IV-C).
7) The final set of customers are inspected.

This methodology has been applied in the specific case of Endesa databases.

V. Customer Filtering

As mentioned in the introduction, the first module was a customer filtering based on a text mining process from the Endesa database.

A. Preprocessing

In this submodule, the information of customers is checked in order to filter all customers without enough information:

1) At least 24 meter readings. The distribution company has outsourced the readers' staff. This staff is in charge of gathering the meter readings from all the customers who still do not have a smart-metering. If the staff cannot read or get information from the meter, the system automatically generates an estimated meter reading. These estimations are based on the historical consumption of the customer. Thus, the customer's historical consumption must have at least 24 meter readings, without estimated meter readings (this means two years of meter readings).
2) Enough information about the consumer (economic activity, contracted power, operations in the consumer installation, etc.)

The distribution companies gather a great quantity of information from each customer: consumption, contract information, etc. but the companies do not have information about updates or changes in the customer supply. The distribution companies request additional information from staff, in order to use this information in management, maintenance, and monitoring of supply. The inspectors usually take advantage of this information in order to determine the reason of a consumption drop or problems in the consumption. The authors did not find any other reference that took advantage of this information.

The initial data source provided by the distribution company contains the information about contracts, historical consumption, inspectors' commentaries, and supply technical characteristics of 101,215 consumers after the application of these restrictions.

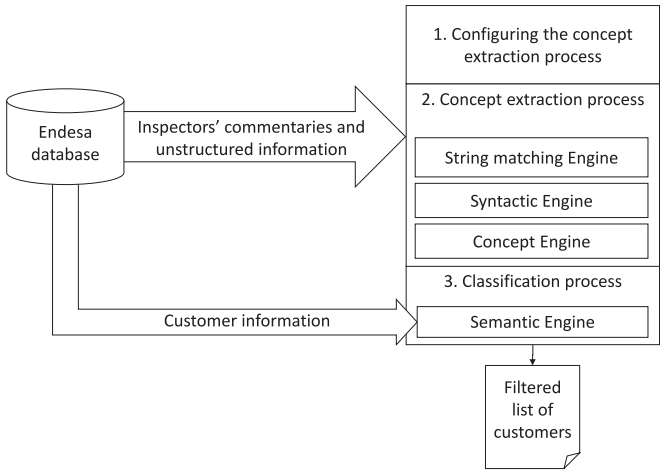


Fig. 2. Customer filtering overview.

B. Process Overview

The customer filtering needs to create a dictionary of classified concepts (the process is described in Fig. 2). After preprocessing, the dictionary modeling is performed in three stages. In the first stage, a configuration process is performed. In the second stage, a text mining method is applied. The text mining method is based on the development of a dictionary of concepts or terms related to several parameters of the consumer contract. This dictionary is extracted from the inspectors' commentaries that were written after inspections or technical interventions. It provides additional information about the consumer facilities, and can be used as a customer filtering method. Finally, the classification of each concept according to the available information from the corresponding customer is performed.

Thus, the dictionary contains approximately 4 million concepts or terms extracted from the inspectors' commentaries. Each concept is classified into one of 5 categories. A simple rule applied in these categories makes it possible to filter the consumers without an NTL, when a customer is analyzed.

The application of this technique provides a dictionary of rules to classify each concept into 5 categories. This dictionary can be periodically updated by applying concept extraction and classification to the greatest number of possible samples.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

C. Configuring the Concept Extraction Process

The first stage of the modeling or creation of the dictionary was the configuration of the concept extraction process. The concept extraction process is based on Natural Language Processing (NLP) techniques [22]. These techniques are very useful for analyzing unstructured information.

The system was performed on the Endesa database. This database contains information in several languages: Spanish, Catalan, Valencian, Majorcan and Aranese. The advantage of having these languages is that they have the same syntactic structure. So instead of creating a dictionary for each language, a synonym dictionary was created establishing synonymously equivalent words in each language, i.e., "fraude" (in Spanish) is set as a synonym for "frau" (Catalan, Valencian and Majorcan) and "fraudaria"/"frauda" (Aranese).

This configuration is performed using a table with several columns (one per each language). The vocabulary is limited to colloquial words, technical terms related to power distribution, business terms, etc.

D. Concept Extraction

The concepts are words or groups of words (syntagms) that represent an idea or action. The concept extraction process is implemented and performed in SPSS Text Analytics and Python. The NLP process has the following four engines (see Fig. 2):

1) String matching engine. This engine is based on fuzzy logic and uses the synonym dictionaries (previously mentioned). A fuzzy ratio is added to each word to identify similar words and mistakes. The mistake correction can be applied according to the lengths of the words. This process was performed combining Text Analytics with Python program.
2) Syntactic engine. This engine assigns a function to each word, according to its position and the previous and following words, in each phrase. Thus, several concepts could be the same function in different phrases or sentences. This process was performed in Python.
3) Concept engine. This engine generates several concepts; the words with the same syntactic function and meaning in the same sentence are grouped into the same concept. This process was performed in Text Analytics and Python.

The problem of different languages was solved with the first engine with string matching and correction of mistakes.

E. Semantic Engine

This engine assigns a semantic function to each concept. The semantic function is represented by a classification in categories. The different semantic categories are defined according to the inspectors' knowledge. This knowledge was provided by inspectors in profiles and commentaries. This process was performed with Python and SPSS Modeler.

After the extraction step, each concept was classified into a semantic category. The semantic category describes knowledge that is related to the NTL detection. In the semantic engine, each concept was associated to its source. In this way, each concept has associated with it the date of the concept, who writes the commentaries, and the result of this commentary. Additionally, the information about consumers is associated:

1) Statistical analysis of the consumption of the date associated to the concept.
2) Contracted power.
3) Time discrimination band.
4) Geographic location.
5) Economic activity.
6) Frequency of occurrence of the concept.

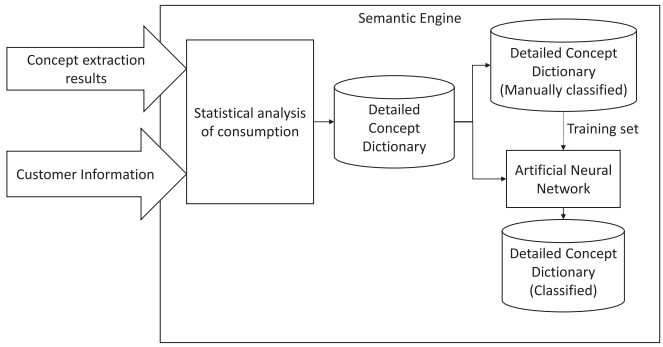


Fig. 3. Semantic engine modelling.

F. Modelling of Classification Process

The classification process is performed in two steps (see Fig. 3). The first step is carried out manually, while the second step is performed automatically by means of an ANN.

In the first step, 5% of the most frequent concepts (roughly 200,000) are manually classified into five categories. This set of concepts used in the second step as a training set:

1) *CLOSED:* Concepts that represent consumers who are: closed, uninhabited, on a holiday, demolished, and so on. These scenarios are usually confused with NTL by the detection of an algorithm because of the consumption pattern.
2) *CORRECT:* This category identifies consumer installations that are correct or without NTLs. This category could prevent false positives.
3) *INCORRECT:* This category identifies consumers whose consumer installation might have an NTL. This category represents concepts that usually identify a measurement problem in the consumer facilities.
4) *LOW CONSUMPTION:* This category identifies consumers who usually have a low or very low consumption, due to their activity. The consumers classified in this category are filtered because the correct consumption is irregular or very low. For example, some consumers with agriculture activity have water pumps, which have irregular and low consumption.
5) *UNUSEFUL:* This category has 101 subcategories, which are not used in the filtering process. These subcategories include the UNKNOWN category, which contains the concepts that could not be classified. Additionally, there are several subcategories that contain information about

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

names, numbers (currency, telephone numbers, address, etc.) and dates. These three subcategories represent 23% of the total number of concepts. This category is excluded from the set of the most frequent concepts.

In the second step, the set of the most frequent concepts (manually classified) were used as a training set in an ANN. The training method is based on the evolutionary strategy [22] and the SPSS Modeler [24]. The evolutionary strategy is a method for parametric optimization. Several neural networks are trained in parallel, starting in an ANN with 22-22-22-4. In each generation, the best ANN is selected. Additionally, one more ANN is randomly selected. Each selected ANN is pruned [25] and increased in one neuron. The algorithm increases the size of ANN until a maximum of 20 neurons in the first hidden layer and a maximum of 24 neurons in the second hidden layer by default. In this case, due to the number of existing entries the limit was modified up to 60 neurons for each hidden layer. The best ANN will be selected according to what the criteria stablishes in [26]. The criteria of selection is based on Cross-Entropy. The ANN is trained using the Softmax function [27]. This training algorithm uses 30% of the training sample to prevent overfitting. This means the 1.5% of the whole sample (roughly 60,000).

The best trained neural network has two hidden layers, and its structure is 20-28-24-4. The first and last layers are the input and output layers, respectively. The input layer has a high quantity of inputs due to the number of input parameters. The output layer has four outputs that identify each category.

As mentioned previously, the validation process of the ANN was performed with 30% of the sample (1,5% of the whole sample); additionally, the results were validated with a manual analysis on a well-known sample. It was not possible to check all the concepts because it would require examining all customers.

The application of the proposed neural network produced successful classification in approximately 27% of the extracted concepts. Thus, 45% of the extracted concepts were in the UNKNOWN semantic subcategory.

G. Filtering Process

The module of filtering starts with the extraction process applied to the commentaries associated to each customer in the databases. The extracted concepts are compared with the dictionary (see Fig. 4). The customer is filtered If we find in the commentaries of the last six months a concept classified in the CLOSED, CORRECT, INCORRECT, or LOW CONSUMPTION categories.

The application of the described text mining techniques decreases the number of customers to 51,204 (50.6% of the preprocessing sample).

In order to carry out the data mining process and reduce the volume of inspections (because this is a test stage), Endesa Company provided us the database of a particular area of Spain (Catalonia) corresponding to the data of last inspections of customers with null consumption. Basically, the criterion that Endesa uses to select these customers in their inspections depends on the number of the last consecutive bills with null consumptions that the customers register.

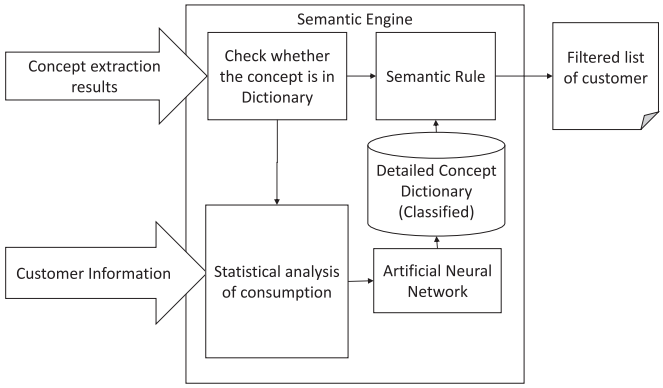


Fig. 4. Semantic engine analysis.

TABLE I
RESULTS OF THE INSPECTIONS OF THE COMPANY

Result of inspection	Number	Percentage
Correct	2,284	65.07%
Not done	455	12.96%
Anomaly without NTL	607	17.30%
Anomaly with NTL	164	4.67%

The data from these inspections were dated on April 2012 and included the contracts selected for the inspection as well as the inspection results.

After preprocessing stage, this sample set included 15,853 customers with the information referring to their consumption during the last two years (from April 2010 to March 2012), contracted power and the specific use of that contract (private or the type of business). This sample was compared with the customer filtering results, finding 3,510 customers in the set of 51,204 customers. Thus, the final data source had 3,510 customers. The customer filtering process decreased the sample size to 77.9%.

VI. DATA MINING

A. Data Source After Customer Filtering

As was described in the previous section, the customer filtering provides a final sample of 3,510 customers.

Additionally, the result of the previous inspections carried out for each customer is known. The classification of these inspections and the number of customers in each set are presented in Table I. The anomaly without the NTL is defined as an NTL without energy losses, this means, although there is an NTL the company knows how much energy was consumed. The anomaly with the NTL is defined as NTL with energy losses which the company should estimate and bill to the consumer.

These results included: 1) inspections without any anomaly (Correct), 2) inspections that could not be performed because it was not possible to access the meter (Not done), 3) inspections that detected some type of anomaly but without energy loss (Anomaly without NTL), and 4) inspections that detected some type of anomaly with energy loss (Anomaly with NTL). Endesa is interested in identifying only the customers in the last group, the success of the inspections was less than 5%.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

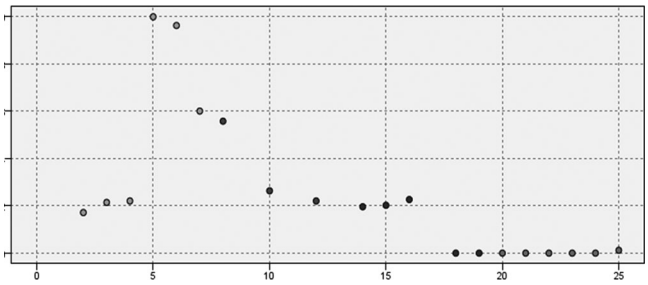


Fig. 5. Pattern of drop consumption.

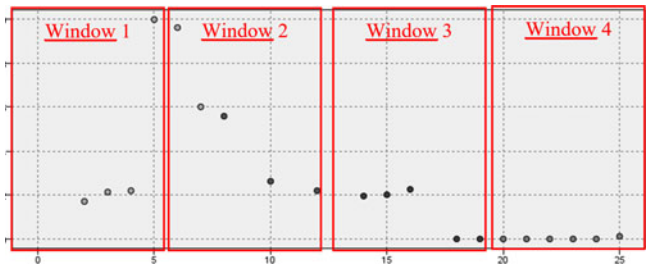


Fig. 6. Windows in the consumption pattern.

There were two main problems at the time of our data mining:

1) The paucity of available data for each customer (only its contracted power, its consumption and the type of business for the contract).
2) The paucity of case series in consumption patterns of customers (because most of them were relative to customers without consumption within two years of the analysis interval).

Once data and problems founded are analyzed, authors focused on data mining process on order to develop some type of detections (rule set) to raise the current percentage of success. It can be considered two modules for two types of detections:

1) Detection of customers that have stable consumption for two years, but suddenly dropped. So, the authors would try to discard accounts associated with empty flats, closed business, etc. That is, those contracts with actual null consumption.
2) Detection of contracts of businesses for which the index of NTLs (which embedded frauds) was high.

In order to carry out the data mining process as well as the generation of the models corresponding to this second module, a powerful piece of software called IBM SPSS Modeler 15, which [23] extended into the field of data mining, is used. This software provides quick access to the databases and many libraries for the generation of models such as clustering processes, decision trees or neural networks.

B. Filter by Drops of Consumption

The pattern that the authors attempted to find with this first model of the module is shown in Fig. 5. This figure shows one example of consumers with some meter readings which show a large period of null consumption. That is, customers with consumptions during a time and, later, a sudden drop to null consumption.

In order to detect this type of pattern the following processing was carried out:

Divide the analysis time frame (two years) for each contract into 4 windows (6 months each). The division of the previous example is shown in Fig. 6.

Carry out a normalization of the consumption of each customer. This was performed by dividing each value of consumption by the contracted power consumption of the contract. This process was carried out in order to equate all the customers in the analysis.

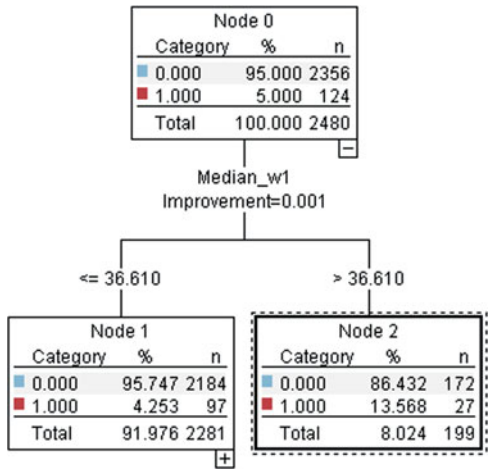


Fig. 7. C&R tree for filtering.

Value	Proportion	%	Count
0		87.13	237
1		12.87	35

Fig. 8. Results of first C&R tree.

For each window the following values were calculated: maximum, minimum, mean, median and standard deviation.

Once they obtained these new window parameters for each customer, the authors applied algorithms that, using these as inputs, obtain filtering rules.

The first algorithm was a decision tree, specifically C&R [28]. The node classification and regression tree (C&R) is a prediction and classification method based on trees. Similar to C4.5, this method uses binary partitioning to repeatedly divide training records into segments with similar values in the output field.

By means of C&R the first simple tree in Fig. 7 was obtained. Fig. 7 shows in each node two categories, each category represents the percentage and total number of consumers in each branch of corresponding node. The sample set was divided depending on the median of the window 1 (Median_w1).

Considering that the training algorithm uses 30% of the data to prevent overfitting, the final results are shown in Fig. 8. That is, with the whole sample and the rule Median_w1 $>$ 36.61 was obtained a confidence of 12.87% (35 out of 237) with a support of 7.8% (272 out of 3510) as shown in Fig. 8. The 1 value represents the consumers with a confirmed NTL. The 0 value represents the consumers with a possible NTL, because of the similarity of their consumption pattern.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

Value	Proportion	%	Count
0		90.72	645
1		9.28	66

Fig. 9. Results of third C&R tree.

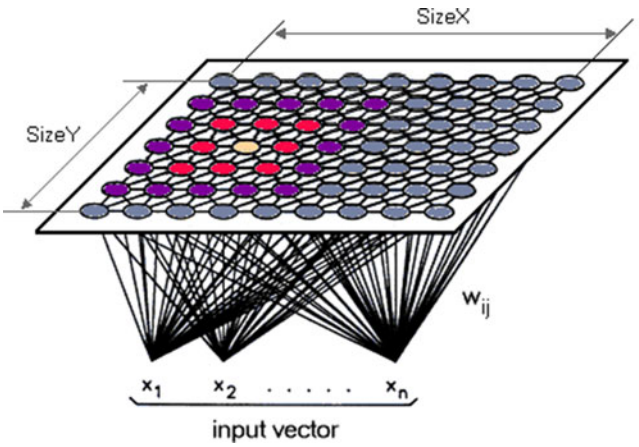


Fig. 10. SOM neural network.

As it is possible to observe in Fig. 8, this rule greatly improves greatly on the original sample set with a 5% of success. The fact is that this rule has very little support. Thus, searching for a higher support by means of C&R, another rule (Median_w1 $>$ 36.6 or Mean_w2 $>$ 22.9) with a higher support (12.5%) and a good confidence (10.8%) was obtained.

Finally, to obtain an even higher support, another rule (Median_w1 $>$ 36.61 or (Mean_w3 $<=$ 0.04 and Median_w1 $>$ 0.63)) with a support of 21% and a confidence of 9.28% was generated. The results of this rule applied in the whole sample are shown in Fig. 9.

These results obtained with the previous rules were considered highly satisfactory. It should be noted that the application of additional metrics like the confusion matrixes ([30]) was not possible in our case because the results about the consumers who were not inspected was not available.

As it is possible to deduce by observing their conditions, all the previous rules are aimed at detecting the pattern of Fig. 5. That is, these rules identify customers with consumption in the first windows and null in the last ones. Depending on the number of clients that the Endesa Company wants to inspect (that is, the support of the rule), it would be necessary to apply one or the other.

Later, another type of data mining algorithm was applied: clustering. For the clustering process, first of all, all the outliers of the sample set were filtered: 283 customers. Later, K-Means algorithm and Self-organizing maps (SOM) was tested.

The best results were obtained with an SOM [29] with a neuronal structure of 4x3. This structure obtained the best result in the silhouette coefficient (the main metric in this type of algorithms [31]). This coefficient has a range of $[-1, 1]$; a value near 1 indicates a strong structure and a value near -1 a no substantial structure. Specifically, the obtained value was 0.6 which is considered a good result for this type of problems.

Value	Proportion	%	Count
X=0, Y=0		10.66	344
X=0, Y=1		0.46	15
X=0, Y=2		72.27	2332
X=1, Y=0		0.53	17
X=1, Y=1		2.39	77
X=1, Y=2		3.56	115
X=2, Y=0		3.13	101
X=2, Y=1		0.19	6
X=2, Y=2		2.7	87
X=3, Y=0		2.32	75
X=3, Y=1		0.28	9
X=3, Y=2		1.52	49

Fig. 11. Results of SOM neural network.

Value	Proportion	%	Count
0		85.25	104
1		14.75	18

Fig. 12. Results merging C&R tree and SOM network.

Fig. 11 shows the NTLs grouped by neuron. The X and Y values corresponding to a coordinates, as is described in Fig. 10. Additionally, the proportion column in Fig. 11 is a bar chart representation of the cluster according to the sample. The percent and count column indicate the percentage and number of consumer classified in the corresponding X and Y values.

Note that there are two clusters: X $=$ 2, Y $=$ 0 and X $=$ 3, Y $=$ 0 with a higher rate of NTLs. Specifically, if the two clusters are joined, a confidence of 12.5% with a support of 5% will be obtained.

If the customers detected with both algorithms are merged, corresponding to the third rule from the C&R tree and the SOM algorithm, it is possible to observe that the confidence increased to 14.75% for a support of 8%. These results are shown in Fig. 12. The junction of both algorithms could serve as method to validate the results.

C. *Filter by Type of Contract use (Private or Business)*

On the other hand, furthermore, customer consumption, an additional parameter called CNAE[1] or list of economic activities is available. This parameter defines if the customer is private or a business (as well as the type of business).

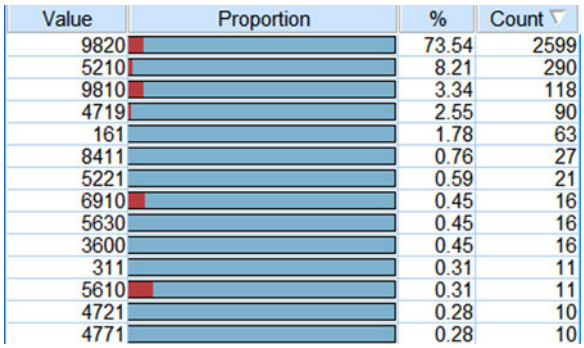
With this part of the module the authors tried to see the influence of the type of business, knowing that businesses such as the following ones innately have drops in consumption: wells, lightings, irrigation pumps, water purification and construction.

A first analysis of the distribution of the main CNAEs (each one is coded with a numeric value) with its corresponding distribution of NTLs is shown in Fig. 13. The column Proportion indicates the normalized proportion of NTLs for each CNAE. On the other hand, the columns % and Count quantify the number and percentage of each type of economic activity in the sample set with respect to other activities.

As it is possible to observe knowing the meaning of the values of the CNAE, the private customers (values 9820 and 9810), as well as warehousing and storage (value 5210) approximately encompass the values reached through inspection. Conversely,

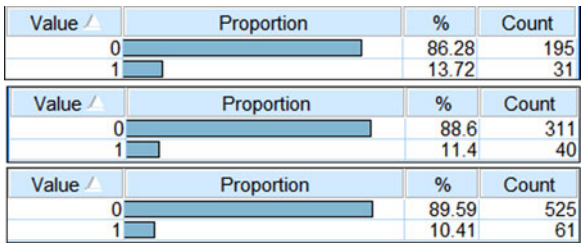
[1]http://www.cnae.com.es/lista-actividades.php

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.



Value	Proportion	%	Count
9820		73.54	2599
5210		8.21	290
9810		3.34	118
4719		2.55	90
161		1.78	63
8411		0.76	27
5221		0.59	21
6910		0.45	16
5630		0.45	16
3600		0.45	16
311		0.31	11
5610		0.31	11
4721		0.28	10
4771		0.28	10

Fig. 13. Results merging C&R tree and SOM network.

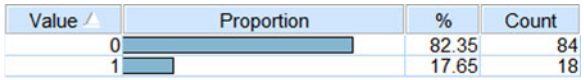


Value	Proportion	%	Count
0		86.28	195
1		13.72	31

Value	Proportion	%	Count
0		88.6	311
1		11.4	40

Value	Proportion	%	Count
0		89.59	525
1		10.41	61

Fig. 14. Results of C&R with selection of CNAEs.



Value	Proportion	%	Count
0		82.35	84
1		17.65	18

Fig. 15. Results of merging C&R/SOM with CNAEs.

the higher rates of NTLs are found in private customers, legal activities (value 6910) and restaurants (value 5610).

Thus, if the customers with CNAE code 9820, 9810, 6910 and 5610 are extracted from the rules of the C&R algorithm, their confidences improve with the same order on the support. The results for the rules are shown in Fig. 14.

Finally, if the results of merging the C&R tree are filtered by CNAEs (9820, 9810, 6910 and 5610) and the SOM network, a very high confidence (almost 18%) in the resulting rule will be obtained. This result almost multiplies by four the original results obtained by the Endesa Company in its inspections. The results for this case are shown in Fig. 15.

D. Validation With Other Samples

As an additional validation criteria the proposed framework was applied in two samples. These samples are the results of other campaigns of null consumption. There are several considerations to evaluate these results:

1) The original samples were bigger, but they suffer some filter processes. The original data set and the details of filter processes are unavailable.
2) The results of analysis are limited by the information of each campaign; the success of the proposed framework only can be evaluated in relation to the results of each campaign.

TABLE II
REAL INSPECTIONS VS. FILTERING WITH PROPOSED SOLUTION

NULL CONSUMPTION CAMPAINGS		C.1	C.2
REAL	NTL	2413	80
	WITHOUT NTL	24037	1134
	SUCCESS RATE	10.04%	7.05%
PROPOSED SOLUTION	NTL	1724	70
	WITHOUT NTL	13407	804
	SUCCESS RATE	12.86%	8.71%
	INSPECTION REDUCTION RATE	57.21%	71.99%

3) The original campaigns have high number of failure inspections. Thus, some supplies cannot be inspected due to the lack of information, empty place, etc. This information is not included because it is not possible to validate, due to some features of the data set are not available.
4) The results only show NTL with energy loss, they have not included the NTL without energy loss. An NTL without energy loss could be, for example, a lack of measures because of opaque pane. In this case, although the company cannot bill the real consumed energy, the energy is registered by the measurement unit and the problem can be corrected replacing the pane.

In Table II the results of this comparison process are shown. The comparison was performed in two campaigns of null consumption with different success rate. The total number of inspection is the summation of "NTL" and "WITHOUT NTL", so for example, in C.1 in real campaign, 26450 inspections were performed. But in the proposed framework (or solution) the total number of inspections would be 15131. The number of inspections were reduced in a 57.21%, increasing the success rate to 12.86% and decreasing the economic cost, because 10630 unnecessary inspections were discarded.

VII. CONCLUSION

For the electrical distribution business, detecting NTLs is a highly important task, because for instance, it is estimated in Spain that the percentage of fraud in terms of energy with respect to the total NTLs is approximately 35%-45%.

One of the methodologies used by the power utilities is the search for and inspection of customers with null consumption during a certain period. The main problems with this methodology are:

1) The low percentage of success in this type of inspection (approximately 5%), which is because null consumption does not necessarily indicate a NTL and many times additional information is necessary to confirm it.
2) The requirement for a large number of inspectors and, therefore, the high cost to the Company.
3) Additionally, traditional techniques are based on data mining only applied in consumption are not useful in this case, because all cases are null or steep drops.

The Endesa Company is the most important Spanish energy distribution company with more than 12 million clients in Spain. For 8 years, the authors have been carrying out a collabora-

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

tive project with the Company and, in the last stage, they have been working to improve the inspections of customers with null consumption.

A framework comprising two modules was developed. The first module performs a customer filtering based on text mining and an artificial neural network. The second module is based on rules from a data mining process for the improvement of the results of the inspections. These rules are generated from two algorithms: a C&R (a type of decision tree) and a SOM (a type of neural network for clustering).

The main contributions of this paper were:

1) The development and deployment of a methodology to increase the efficiency of inspections.
2) The successful application of proposed techniques in a real case.
3) The application text mining and neural network to analyze information, that it is not traditionally treated in any other reference.
4) The model of relation between inspectors' commentaries and customer consumption.
5) The improvement of traditional methods used by a power distribution company.
6) The application of data mining techniques to detect NTLs in samples with null consumptions. The proposed data mining methods modelled the abnormal consumption in a sample in which all the customers have consumption drops.

Thus, the whole framework consists of a set of different rules that filter a number of customers depending on the support (percentage filtered) that Endesa wants to reach in its inspections. Both modules have been developed and validated with the database of the results of a real campaign of inspections by the company. The results of our module triple the success rate of the inspections (specifically from approximately 5–14.75%). This module is currently being used by the Endesa Company.

Finally, the proposed solution were evolved and tested in Smart Grid ecosystems with AMI infrastructures [32], providing the possibility of real time analysis of all available information in company databases.

REFERENCES

[1] N. Padhy, P. Mishra, and R. Panigrahi, "The survey of data mining applications and feature scope," *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 3, pp. 43–58, Jun. 2012.

[2] S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *Proc. 2010 Int. Conf. Intel. Computat. Tech. Autom.*, Changsha, China, 2010, pp. 50–53, doi: 10.1109/ICICTA.2010.831.

[3] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y. Y. o. -P. Huang, "Survey of fraud detection techniques," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, 2004, vol. 2, pp. 749–754.

[4] M. Weatherford, "Mining for fraud," *IEEE Intell. Syst.*, vol. 17, no. 4, pp. 4–6, Aug. 2002.

[5] S. Ghosh and Y. D. L. Reilly, "Credit card fraud detection with a neural-network," in *Proc. 27th Hawaii Int. Conf. Syst. Sci.*, 1994, vol. 3, pp. 621–630.

[6] E. Aleskerov, B. Freisleben, and Y. B. Rao, "CARDWATCH: A neural network based database mining system for credit card fraud detection," in *Proc. IEEE/IAFE Comput. Intell. Financial Eng.*, 1997, pp. 220–226.

[7] J. R. Dorronsoro, F. Ginel, C. Sgnchez, and Y. C. S. Cruz, "Neural fraud detection in credit card operations," *IEEE Trans. Neural Netw.*, vol. 8, no. 4, pp.. 827–834, Jul. 1997.

[8] I. Monedero *et al.*, "Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees," *Int. J. Elect. Power Energy Syst.*, vol. 34, no. 1, pp. 90–98, 2012.

[9] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, and J. Biscarri, "Integrated expert system applied to the analysis of non-technical losses in power utilities," *Expert Syst. Appl.*, vol. 38. no. 8, pp. 10274–10285, 2011.

[10] F. Biscarri, C. León, I. Monedero, J. I. Guerrero, and J. Biscarri, "Variability and Trend-Based generalized rule induction model to NTL detection in power companies," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 1798–1807, Nov. 2011.

[11] Y. Sahin and E. Duman, "Detecting credit card fraud by ANN and logistic regression," in *Proc. Int. Symp. Innovations Intell. Syst. Appl.*, 2011, pp. 315–319.

[12] L. Seyedhossein and M. R. Hashemi, "Mining information from credit card time series for timelier fraud detection," in *Proc. 5th Int. Symp. Telecommun.*, 2010, pp. 619–624.

[13] A. Mohamed *et al.*, "Telecommunication fraud prediction using back propagation neural network," in *Proc. Int. Conf. Soft Comput. Pattern Recog.*, 2009, pp. 259–265.

[14] L. Wu, R. Ping, L. Ke, and D. Hai-xin, "Intrusion detection using SVM," in *Proc. 7th Int. Conf. Wireless Commun., Netw. Mobile Comput.*, 2011, pp. 1–4.

[15] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[16] C. C. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcão, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.

[17] E. W. dos Angelos, O. R. Saavedra, O. A. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Trans. Power Del.*, vol. 26, no. 4, pp. 2436–2442, Oct. 2011.

[18] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving SVM-Based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1284–1285, Apr. 2011.

[19] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.

[20] J. V. Spirić, S. S. Stanković, M. B. Dočić, and T. D. Popović, "Using the rough set theory to detect fraud committed by electricity customers," *Int. J. Elect. Power Energy Syst.*, vol. 62, pp. 727–734, 2014.

[21] "IBM SPSS modeler 15 algorithms guide," IBM [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf, p. 386, 2012.

[22] S. Linkels and C. Meinel, *Natural Language Processing, in: E-Librarian Service*. Berlin, Germany: Springer, 2011, pp. 61–79.

[23] M. A. Javarone, "An evolutionary strategy based on partial imitation for solving optimization problems," *Physica A: Statistical Mechanics Its Appl.*, vol. 463, pp. 262–269, Dec. 2016.

[24] 2017. [Online]. Available: http://www.ibm.com/software/products/es/es/spss-modeler/

[25] M. R. Silvestre and L. L. Ling, "Pruning methods to MLP neural networks considering proportional apparent error rate for classification problems with unbalanced data," *Measurement*, vol. 56, pp. 88–94, Oct. 2014.

[26] IBM SPSS Modeler 16 Algorithms Guide. IBM Press, 2013.

[27] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, F. F. Soulié and J. Hérault, Eds. Berlin, Germany: Springer, 1990, pp. 227–236.

[28] P. Perner, "Recent advances in data mining," *Eng. Appl. Artif. Intell.*, vol. 19, no. 4, pp. 361–362, 2006.

[29] S. Valero, M. Ortiz, C. Senabre, A. Gabaldón, and F. García, "Classification, filtering and identification of electrical customer load pattern through the use of self-organizing maps," *IEEE Trans. Power Systs.*, vol. 21, no. 4, pp. 1672–1682, Nov. 2006.

[30] F. Provost and R. Kohavi, "On applied research in machine learning," *Mach. Learn.*, vol. 30, pp. 127–132, 1998.

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[32] J. I. Guerrero *et al.*, "Intelligent information system as a tool to reach unaproachable goals for inspectors—High-performance data analysis for reduction of non-technical losses on smart grids," in *Proc. 5th Int. Conf. Intell. Syst. Appl.*, 2016, pp. 83–87.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.

Juan Ignacio Guerrero received the B.Sc. and Ph.D. degrees in computer science from the University of Seville, Seville, Spain, in 2006 and 2011, respectively. He is an Assistant Professor with the Department of Electronic Technology at the University of Seville. His research areas include big data analytics, data mining, knowledge based systems, computational intelligence, high-performance data analysis, and machine learnning in Smart Grids.

Iñigo Monedero received the B.Sc. and Ph.D. degrees in computer science from the University of Seville, Seville, Spain, in 1994 and 2004, respectively. Since 2001, he has been a Professor in the Department Electronic Technology at the University of Seville. His research areas include artificial intelligence, data mining, and software engineering in aeronautics.

Félix Biscarri received the B.Sc. degree in electronic physics and the Ph.D. degree in computer science from the University of Seville, Seville, Spain, in 1991 and 2001, respectively. He is currently a Coordinating Professor of Power Electronic in the Polytechnic University School of Seville. His research areas include electricity markets, electrical customer classification, and fraud detection in the power electric industry.

Jesús Biscarri received the B.Sc. and Ph.D. degrees in electronic physics from the University of Seville, Seville, Spain, in 1982 and 2001, respectively. He has been working in Endesa since 1985 at IT, Measure and Non-Technical Losses Control Areas. Since 2010, he has been responsible for Smartmetering Project and Operations of Endesa Distribucion. He also collaborates as an Associate Professor at the Polytechnic University School of Seville.

Rocío Millán received the B.Sc. and Ph.D. degrees in economics and business administration from the University of Seville, Seville, Spain, in 1985 and 1996, respectively. She was a Professor of economic theory and finance in this university for more than ten years and is working for Endesa, Seville, as the Metering Control Deputy Director. Her research areas include public deficit, energy futures markets, and NTLs detection in electricity companies.

Carlos León (SM'10) received the B.Sc. degree in electronic physics and the Ph.D. degree in computer science from the University of Seville, Seville, Spain, in 1991 and 1995, respectively. He is currently a Full Professor of electronic engineering and computer science at the University of Seville. His research areas include knowledge-based systems, computational intelligence, big data analytics, and machine learning focus on Utilities System Management. Dr. León is a Senior Member of the IEEE Power Engineering Society.

Authorized licensed use limited to: Imperial College London. Downloaded on February 21,2020 at 17:00:18 UTC from IEEE Xplore. Restrictions apply.