# Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies

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

***Abstract*—This paper proposes a comprehensive framework to detect non-technical losses (NTLs) and recover electrical energy (lost by abnormalities or fraud) by means of a data mining analysis, in the Spanish Power Electric Industry. It is divided into four section: data selection, data preprocessing, descriptive, and predictive data mining. The authors insist on the importance of the knowledge of the particular characteristics of the Power Company customer: the main features available in databases are described. The paper presents two innovative statistical estimators to attach importance to variability and trend analysis of electric consumption and offers a predictive model, based on the Generalized Rule Induction (GRI) model. This predictive analysis discovers association rules in the data and it is supplemented by a binary Quest tree classification method. The quality of this framework is illustrated by a case study considering a real database, supplied by Endesa Company.**

***Index Terms*—Customer electricity consumption, electric fraud, electricity market, non-technical loss.**

## I. Introduction

In the electricity sector, a non-technical loss (NTL) is defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of the equipment, and also because of administrative errors in the energy invoicing processing. In fact, the NTLs of the company are all losses except in technical losses (the result of the effect of the power dissipation in the electrical network components such as transmission lines, power transformers, measurements systems, etc.). For the electrical distribution business, minimizing NTL is a very important activity because it has a high impact on company profits. We will refer in this paper to fraud detection but we could strictly talk about NTL. Therefore, detection of NTL includes detection of fraudulent users [1]. In Spain, the percent of fraud in terms of energy with respect to the total NTLs round about 35%–45%.

Electrical loss percentages may vary from country to country, even from region to region. This percentage also varies between customers belonging to different economic activities. Not too many authors offer an estimation about these losses. Yap *et al.* [2] estimate distribution losses as 15%, in Sabah State, Malaysia. Filho *et al.* [3] expose a fraud identification per number of in-situ inspection percentage as low as 5%, in Brazil. This rate varies about 5%–10%, according to Cabral *et al.* [4], [5].

The main motivation of this project is to improve the process of Endesa NTL detection using data mining techniques. The main difficulty is the low rate of these in the Power Companies [5]. Besides, in the companies and/or areas with very low percentages of losses, from 1% to %2, it is an inefficient policy to reduce these losses if the companies do not identify them in some way. The enormous cost of inspecting in-situ many customers does not compensate the return of the energy recovered in them. That is why Endesa is investing in this research, exploiting its databases to identify customers with non-invoiced energy in a profitable way. So the number of inspections is reduced only to a small group of customers identified as *anomalous*, with an energy consumption *suspicious* of being different from the amount invoiced, because some company error or customer fraud. In order to detect losses in the group of high consumption customers, an interesting cluster, the careful selection of the feature group used in the methodology is as important as the specific selected methodology. This is because the consumption pattern varies among different groups of customers, strongly related to the economic activity and the tariff contracted. As seen below, potential anomalous high consumption customers can be identified by: abrupt and negative changes in their historical consumption pattern (a 30% drop, for example, as seen in [6]), changes in the consumption pattern compared with changes to the other consumers from the same cluster at the same time (changes detected in the variability on electric consumption [7]), and other non-obvious features, as previous fraudulent activities or anomalous and unstable power factor [8].

The goal of this research is to significantly improve the inspection success and the profitability rate, which is highly dependent on the cluster of customers researched, i.e., the set of features that made the cluster and the class of customers researched (domestic customers, medium, or high consumption customers).

This paper describes a mining framework to detect NTL based on a statistical characterization of the customers' energy patterns [4], [5], [8]. This goal was obtained with a classification procedure, based in the Generalized Rule Induction (GRI) model and the QUEST decision tree. Two innovative statistical

Manuscript received October 29, 2008; revised March 16, 2009; accepted April 23, 2009. Date of publication March 24, 2011; date of current version October 21, 2011. This work was supported by the Endesa Company (since 2005). Paper no. TPWRS-00887-2008.

The authors are with the University of Seville, Seville 41010, Spain.

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.2011.2121350

estimators are also introduced, weighing up variability and trend of customer consumption.

Regarding the proposed methodology and its implementation and use, related Endesa databases organize data as a collection of tables. Information in more than one file can be easily extracted and combined with DataBase Management System (DBMS) like dBase or Oracle. SPSS Clementine is employed as the data mining tool for analysis. The data processing in Clementine is done through the use of nodes which are then connected together to form a stream frame. Designed to support CRISP-DM, which is *de facto* the standard for implementing data mining as a business process, Clementine's open architecture utilizes existing IT investments to enable rapid predictive modeling and high-ROI deployment. The processing time varies greatly depending on the size of the sample. This could take anywhere from a few minutes (hundreds of customers) to several hours (a few million customers). In the case study, that covers roughly 10 000 customers, the data mining processing time is reduced to 15 min, using a standard PC.

The paper is organized as follows.

In Section II, a biographical review has been made. In Section III, the framework is presented, divided into four main steps: in Subsections A and B, Data Selection and Data Preprocessing, the authors insist on the importance of the knowledge of the particular characteristics of the power electric customers. They also describe the main features available in company databases. In Subsection C, Descriptive Section, the authors present two innovative statistical estimators to weigh variability and the trend of the customer consumption, and other sets of features that contribute to information gain. In Subsection D, Predictive Data Mining, a new classification model is presented by a rule set. It is based on discovery association rules in the data and then a binary full classification method. The quality of the framework is also illustrated by a case study considering a real database, supplied by Endesa Company. In Section IV, Conclusion, some concluding remarks are presented.

## II. Bibliographical Review

There are many bibliographical references about main research topics regarding this paper: data mining techniques (see [9]) and fraud detection [1], [10], [11]. The referenced papers present interesting surveys. The methodology to be used has been developed and validated through several practical papers: insurance fraud [12], customer insolvency in telecommunication business detection [13], medical fraud [14], [15], and credit card fraud [16], [17]. Statistical methods and AI tools as back propagation neural network, Kohonen network, and genetic algorithm are nearest neighbors are frequently used to make a descriptive and predictive modeling for classification (fraudulent customer versus non-fraudulent customer).

Another related line of work is "data mining to power consumer behavior", also named "power customer mining". The goal is load profiling and the electrical demand prediction [18]–[21]. These papers offer different methodologies for clustering purposes (multidimensional scaling, MANOVA, K-means clustering, etc.) using the historic load curve of each user, but they do not help to differentiate between "honest" customers, fraudulent users, or equipment failure, i.e., they do not help to detect abnormalities and frauds. In fact, during the data pre-processing phase, abnormal or anomalous data are filtered and removed [18].

Last but not least, our main research topic is "NTL detection in electricity consumers", specifically, "fraud detection in electricity consumers".

### A. Review of Fraud Detection in Power Industry Customers

This section reviews the main research we have founded, from the 1990s to present day.

In 1998, Galván *et al.* [22] presented a methodology to guide inspection campaigns by the characterization of the temporal evolution of customer consumption features. They used a non-supervised classification method: the study of a probability density function (pdf) estimator on some electrical features. Two examples are presented: the study of the power factor in the monthly consumption of customers belonging to the high voltage farm watering sector and the study of the rate between the average power during monthly consumption and the contracted power of customer belonging to the low voltage lodging sector. No results of real inspections after the study are done.

In 2000, Sforna [8], from the ENEL S.A Regional Control Center of Milan, reported another non-supervised classification method: a data mining system based on the application of statistics to add values, calculate new meaningful variables, and the application of a classification technique (a self-organizing Kohonen map) of customer behavior patterns. The aim of this application is to give experts the possibility to improve their knowledge about customers. Results are summarized in terms of automatic identification of anomalous consumption values, according to the hours of usage of the contracted active demand, reactive demand, or monthly power factor.

In 2002, Jiang *et al.* [23] proposed an analysis to identify fraud in Australian electricity distribution networks using wavelet techniques and combining multiple classifiers. An optimal feature vector is selected from an input that includes a feature table containing the average power measurements and customer fraud history. The outputs are fraud reports indicating the suspect customers with corresponding estimated probabilities. They exposed that the classification accuracy reached 70% on the testing data set, with a high number of data profiles (about 1200) and using a relative small amount of data.

In 2004, Reis *et al.* [3] used a decision tree and a database composed of five monthly features, with customers that had undergone inspection in the last year, classified into normal, fraud, or faulty equipment. They exposed a 40% right fraud classification rate.

In 2004 and 2006, Cabral *et al.* [4], [5] proposed an application that used rough sets to classify categorical attributes values in order to detect fraud by electrical energy consumers. The continuous attributes were converted to discrete. The system reached a fraud rightness rate of around 20%. Authors put forward the main difficulty to detect electrical energy profiles is the low "fraudulent customers"/ "normal customers" ratio, around 5%, in the Brazilian electrical energy distribution companies. They also exposed that, to aggravate this advantage, many fraudulent consumers behavior seems like normal behavior.

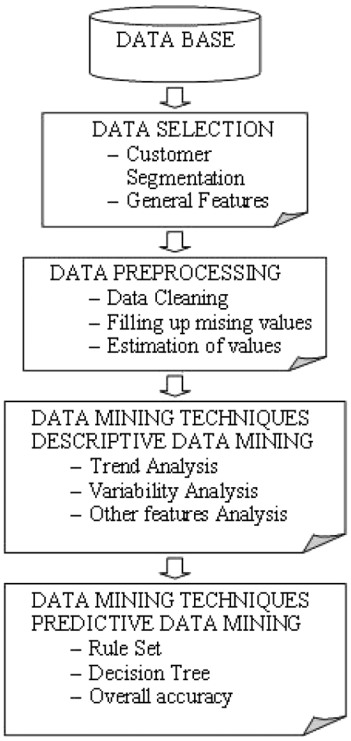


Fig. 1. Study process structure framework.

In 2008, Cabral and Pinto [6] studied high voltage consumers using self-organizing maps (SOM) and a 15-min consumption period sample. They are still waiting for confirmation of suspicions effectively confirmed as fraud.

Lastly, the authors have presented a paper highlighting the importance of the variability of the customer consumption in NTLs detection in power utility companies [7]. A lodging sector customers' example are shown, with 35 customers proposed to be inspected "in-situ" by the automatic detection system, 15 of them inspected by the Endesa staff and 8 of these 15 classified as "anomalous", fraudulent, or with faulty measuring equipment.

## III. Customer Characterization

Fig. 1 shows the proposed study process structure based on knowledge discovery databases (KDD) [24]. This framework is fragmented into different steps with different degrees of complexity and different periods of time.

### A. Data Selection

The customers selected for the mining process have been chosen based on the following feature characterization:

- Period of time of recorded invoices: We use monthly and bimonthly invoices belonging to the sample of customers. Hourly or daily data are not available.
- Geographic area: All customers are located in a Spanish region.
- Contractual power: All customers belong to price code 4.0 and price code 3.0.2. These codes mean these customers consume more than 15 kW during more than 8 h per day.
- Economic activity classification: In Spain, it is called "Clasificación Nacional de Actividades Económicas" (CNAE). Some economic sectors historically present a high rate of NTLs. The research is centered in these sectors.
- Consumption range: At first, the research target is to cover the greatest range of electrical consumption as possible. But customers can only be compared, or studied together, if they have a similar range of consumption. The characterization of customers is highly dependent on consumption. The solution of the problem is to divide the full consumption range into subsets, obtaining subsamples of customers with similar characteristics. Each of these subsamples will be studied independently. For the purpose of this work, the continuous value of the customer consumption is sectioned in ten bins.
- History of customer inspection: Methods used in NTL's detection can be mainly classified into supervised and unsupervised methods. The unsupervised approach allows the discovery of natural patterns in data, detected or undetected before. For this reason, initially we have used an unsupervised approach, based on Kohonen maps and the statistical outlier detection as a classification method. But the verification of the results has been really expensive and time consuming. In order to obtain a statistic of the NTL's right classified rate, all the suspected customers should be inspected "in situ" by the Endesa staff. The results can be interpreted in a very biased way, because it is not often possible to check all the clusters discovered several times. The improvement of the methodology was highly dependent on the success of these inspections.

  The supervised approach is an interesting working method if we have a very large database to cover many of the NTL cases. The results can be quick and systematically checked. Also, as we have said, we can concentrate our efforts toward clusters of customers with a high rate of historical NTLs. Table I shows a list of CNAEs with a high number of NTLs detected by the Endesa staff. It refers to the time period and to the sample presented in this paper. It shows the number of customers belonging every CNAE (Customer_Count), the number of historical non-technical losses detected (NTL_Count), and the percentage of anomalous customers according to normal customers (%NTL).

TABLE I
NTL Spectrum

| CNAE | NTL_Count | Customer_Count | %NTL |
|---|---|---|---|
| XX001 | 46 | 2407 | 1.91 |
| XX002 | 45 | 4675 | 0.96 |
| XX003 | 38 | 1694 | 2.24 |
| XX004 | 26 | 299 | 8.7 |
| XX005 | 26 | 1639 | 1.59 |
| XX006 | 21 | 1923 | 1.09 |
| XX007 | 15 | 628 | 2.39 |
| XX008 | 13 | 535 | 2.43 |

### B. Data Preprocessing

With respect to data cleaning, the authors avoid rejecting any data from a set. However, customers with less than six monthly

registers per year were eliminated and also customers who had negative values on consumption attributes. There are also several feature limits that give us some information about anomalous customers. As an example, customers with very low consumption are suspected. In the sample considered, customers with a yearly consumption below 100 kWh are recommended to be inspected. Also, customers with a high consumption of reactive energy regarding active energy consumption will be inspected.

On the other hand, reflection exercise about lecture consumption data and billed consumption data are necessary. Normally, the consumption billed is the result of consumption read, but it is not always true. If the company has no access to read the data, and there is no doubt of consumption has been made, company experts estimate the actual consumption, based on the recent historical consumption. Several and continuous differences between read data and billed data show abnormal behavior. The study and the use of statistical estimator based on read data is a new contribution of this paper regarding works cited in the bibliographical review. In this sense, a filling up of missing values has been performed.

### C. Data Mining Techniques: Descriptive Data Mining

We describe three descriptive techniques: one based on the variability of customer consumption, another based on the consumption trend, and a third one that summarizes other feature contributions to NTL detection.

*1) Variability Analysis:* We propose in this section an algorithm that emphasizes customers with a high variability of monthly consumption with respect to other customers of similar characteristics.

The classic approach to the study of the variability classifies data in "normal data" and outliers. Very often, there exists data objects that do not comply with the general behavior of the data. Such data objects, which are grossly different from or inconsistent with the remaining data, are called outliers. Outliers, with regard to consumption feature, can be caused by measurement error or by fraud in customer consumption. But, alternatively, outliers may be the result of inherent data variability. Thus, outliers detection and analysis is an interesting data mining task, referred to as outliers mining.

The statistical approach to outliers' detection assumes a distribution or probability model for the given data set and then identifies outliers with respect to the model using a discordance test [4], [25]. Application of the test requires knowledge of the data set parameters (such as the assumed data distribution), knowledge of the distribution parameters (such as the mean and variance) and, mainly, knowledge of the inherent data variability [26]. Thus, the main task is the estimation of the variance data (or the standard deviation estimation, STD) from a sample.

Once the STD is estimated in the anomaly detection field, the Standard Deviation Chart (S chart) offers a signature for each customer that is in itself the baseline for comparison. In classic research, new consumption for a customer is compared against their individual signature to determine if the user's behavior has changed (Fraud and Intrusion detection [11], [27]–[29]). A significant departure from baseline is a signal that the account may have been compromised.

The research presented in this paper offers another point of view and presents three main differences regarding cited researches:

1) The estimation of the STD, the main task of the variability analysis, is performed in a non-classic way. We use a preprocessed sample in which there are no interactions present between time and space. The temporary component and the local geographical location component have been filtered.
2) Consumptions for a group of customers are compared against their group signature to determine if the behavior of an individual customer is anomalous. In classic research, new consumption for a customer is compared against their individual signature to determine if the user's behavior has changed.
3) The classic approach classifies data into "normal data" and outliers. In order to classify data, a *center line* (CL), the average of the STDs, is estimated. Also an *upper control limit* (UCL) and a *lower control limit* (LCL) are estimated. In a classic way, thresholds of STD (LCL and UCL) are estimated by the mean of STD multiplied by a constant (usually, 1.96 is used, corresponding to a level of significance $\sigma = 0.05$). Data outside control limits are classified as outliers. We do not use the estimated STD to obtain outliers and directly propose them to be inspected by the Endesa staff. We do not establish any control limits. We simply add to each customer a new feature, referred to the estimated $STD_{\Delta_l}$, that will be used as an input for a supervised detection method, showed in the *Predictive Data Mining* section.

The estimation of the $STD_{\Delta_l}$ is described subsequently:

1) Given
   - A data at a set of spatial locations (different customers belong to the same cluster or bin). The whole sample is previously divided in ten bins according the yearly consumption feature. Each bin will be studied independently.
   - Several data acquisitions of the data at each location spaced in time. It is assumed that all the locations are sampled at the same time and are sampled many times.
2) The operating equation is defined as follows: $\text{Data acquired} = D_{lt}$, where $D$ is the actual data point measurement, $l$ is the location of the measurement (customer identifier), and $t$ is the time of the measurement (this is the time at which all the data are recorded at all locations).
3) The next step is to obtain the average at each time across all locations. This is defined by the equation $A_t = \sum_{l=1}^{N}(D_{lt}/N)$, where $A_t$ is the average of all data at time $t$, across all locations, $l$, and $N$ is the number of locations.
4) Next, obtain the differences by comparing the data at each location to the average at that time, that is: $\delta_{lt} = D_{lt} - A_t$, where $\delta_{lt} = \text{the difference between the data at each location, l, and this time, t, average.}$

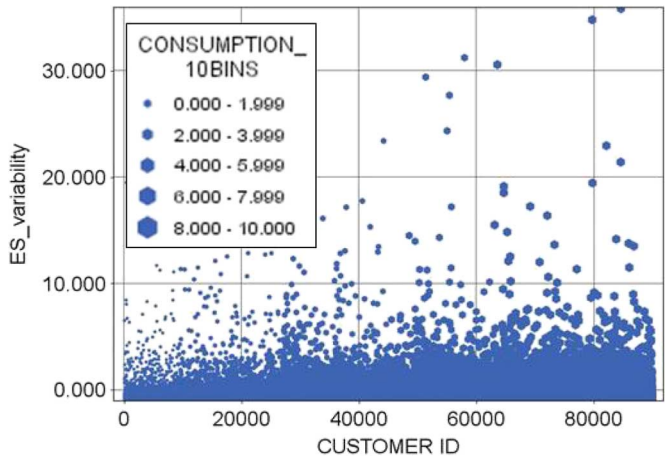


Fig. 2. 10-BINS ES_variability.

5) Now it is necessary to obtain the average of the differences, $\bar{\delta}_l$, at each location across time, that is: $\bar{\delta}_l = \sum_{t=1}^{M}(\delta_{lt}/M)$, where $\bar{\delta}_l$ = the average of all $\delta_{lt}$ at location, $l$, across time, $t$, and $M$ is the number of times averaged.
6) It is then necessary to obtain the differences, $\Delta$, comparing each time difference, $\delta_{lt}$, to its average at location, l, as shown in equation$\Delta_{lt} = \delta_{lt} - \bar{\delta}_l$.
Note that the $\delta_{lt}$ values are the residual electrical consumptions after the linear variations in time and space have been averaged out.

The next step is to calculate the standard deviation associated to each customer with regard to the rest of customers and without inherent variability, $STD_{\Delta_l}$, using equation

$$STD_{\Delta_l} = \sqrt{\sum_{t=1}^{M} \frac{\Delta_{lt} - \bar{\Delta}_l}{M-1}}$$

where

$$\bar{\Delta}_l = \sum_{t=1}^{M} \frac{\Delta_{lt}}{M}.$$

Once the $STD_{\Delta_l}$ is estimated, the following variability estimator is defined:

$$ES\_variability_{li} = \frac{STD_{\Delta_l} - CL_i}{CL_i}$$

where $ES_{variability}$ is a new customer estimated feature, dependent of the customer, $l$, and the yearly consumption bin, $i$. $CL_i$ is the *center line* referring to bin $i$.

To maintain the shape of the variability diagrams and compare the diagrams among them, in consumption pattern terms, each diagram can be normalized. We use the CL of each bin $(CL_i)$ to normalize the sample. Fig. 2 shows the evolution of this estimator through the whole sample studied, through all bins.

The advantages of the proposed algorithm with respect to recent studies are:

- The elimination (or, at least, reduction) of the temporary component and the local geographical location component of the customer consumption. We find outliers can be caused by measurement errors, not those caused by the inherent data variability. Usually, recent works have used the STD for the monthly active energy (see [8]) but the inherent data variability is not filtered.
- The study of the comparative consumption among clients of similar characteristics. This method is based on the idea of fraudsters seldom change their consumption habits [30]. They are closely linked to other fraudsters, but not to the rest of customers.
- The use of the $STD_{\Delta_l}$ estimated as an input to a classification model. Classification methods are particularly useful when a database contains samples that can be used as the basic for future decision making (supervised methods). Thus, researchers have focused on different types of classification algorithms, including nearest neighbor [14], [15], decision tree induction, error back propagation [16], [31], reinforcement learning, and rule learning. The data mining algorithm, originally based on outlier detection, presented in this section is an unsupervised method. This does not require one to be aware about the true classes of the original data used to buil clusters. It can be used as feature input to a complex supervised model.
- The use of a simple tool, developed for mining very large data set.

*2) Consumption Trend: Streak Based Algorithm:* Streaks of past outcomes (or measurements), for example of gains or losses in the stock market, are one source of information for a decision maker trying to predict the next outcome (or measurement) in the series. In the case of gambling, each one is an independent event, so there is no casual mechanism linking outcomes (hence the fallacy). The customer consumption trend has, presumably, an underlying casual model. It depends on the seasonal, economic activity and other hidden features. The discovery of the theoretical consumption model is not the target of this paper. This model is strongly dependent on the cluster of customers considered and highly changeable amongst different clusters. But it is an interesting customer feature that their consumption trend depends of the consumption trend of the other customers in the same cluster.

There are several ways to measure this feature. We show a simple and useful algorithm, based on the six-month lagging moving average of customer consumption, described subsequently:

1) The input data are, for each customer from each cluster, 24 monthly consumptions, billed data. The cluster characterization is described in the *Data Selection* section of this paper.
2) We calculated the six-month simple moving average for each customer consumption.
3) We counted how many times the consumption line is over the mean line (positive streaks, $po_s$) and how many times the consumption line was below the mean line (negative streaks, $ne_s$). The whole number of streaks is

$$Ns = po_s + ne_s.$$

The proposed algorithm does not distinguish positive from

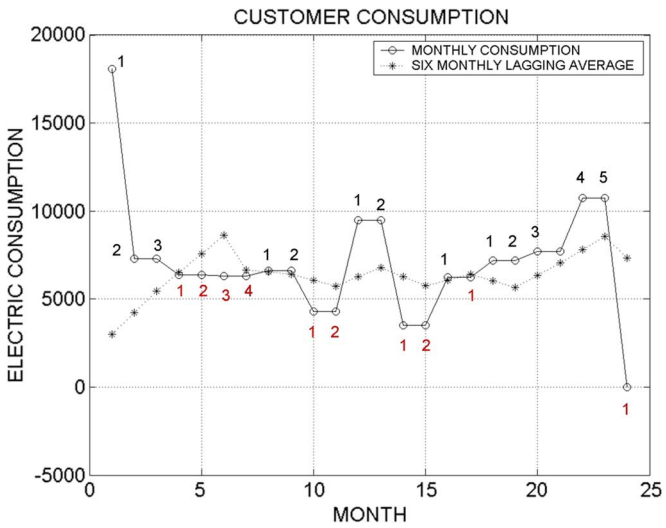


Fig. 3. Consumption trend. Customer with short and numerous streaks.

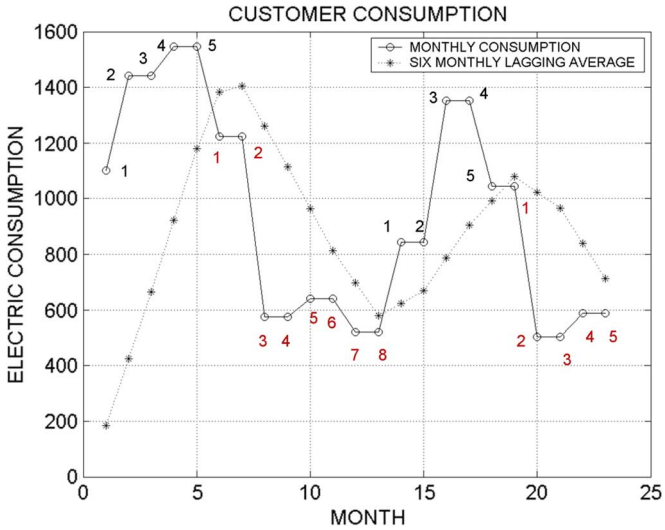


Fig. 4. Consumption trend. Customer with long and few streaks.

negative streaks. It simply counts its own. The number of measurements in each streak $(n_j)$ is also registered. The number of streaks for each customer offers interesting information about their consumption behavior but it is also interesting to know the weight of each streak.

4) Finally, we sum up all the information in one cuadratic estimator:

$$ES\_streak_l = \frac{\sqrt{\sum_{t=1}^{Ns}(n_t)^2}}{Ns}$$

where $l$ is the customer identifier, $Ns$ is the number of streaks of this customer, and $n_t$ is the number of measurements of the streak $t$.

The advantages of the proposed algorithm are:

- We use the $ES_{streak_l}$ estimator as a new customer feature, as a new input to a classification model. The data mining algorithm presented in this section, originally based on a trend following study, is an unsupervised method, that allows us to classify customers in order of their $ES_{streak_l}$ level. This does not require one to be conscious about the true classes of the original data used to build clusters. It can be used as a feature input for a complex supervised model.
- The study of the individual trend consumption and also the comparative among trends of customers with similar characteristics.
- The use of a simple tool, developed for mining very large data sets.

Figs. 3 and 4 show two different consumption behaviors of two customers from the same cluster. The number of measurements in each streak are counted. Table II summarizes and models their trend behavior.

TABLE II
STREAKS STUDY

| Streaks | Customer XX01 | Customer XX02 |
|---|---|---|
| $n_1$ | 5 | 3 |
| $n_2$ | 8 | 4 |
| $n_3$ | 5 | 2 |
| $n_4$ | 5 | 2 |
| $n_5$ | - | 2 |
| $n_6$ | - | 2 |
| $n_7$ | - | 1 |
| $n_8$ | - | 1 |
| $n_9$ | - | 5 |
| $n_{10}$ | - | 1 |
| $Ns$ | 4 | 10 |
| $ES\_streak_l$ | $\frac{\sqrt{5^2+8^2+5^2+5^2}}{4} = 2.95$ | $\frac{\sqrt{3^2+4^2+4*2^2+2*1^2+5^2+1^2}}{10} = 0.83$ |

*3) Other Features:* There are some feature levels or some feature relations quite serious in reference to NTL's detection. We describe some of them used in our framework:

- The hours of consumption at maximum contracted power (HMP). It is the rate between the daily consumption (DC) and the contracted power (CP). For example, if CP $=$ 15 kW and DC $=$ 150 kWh, then HMP $=$ 10 h.
- Minimum and maximum values of consumption in different time zones of the day. In some tariffs, the day is divided into different time zones with different rates. Our sample is divided into three zones: cheap (zone 1), normal (zone 2), and expensive (zone 3). Suspect fraudulent customers have relatively very low consumption in normal and expensive time zones.
- The number of valid consumption lectures (NL). Usually, when there is not a valid lecture value and the company is sure that consumption existed, the consumption is estimated and billed.
- The active/reactive consumption rate (ARR). The power company penalizes low ARR with an added invoice.

TABLE III
DESCRIPTION OF THE RULES

| Model | Rule | Description |
|---|---|---|
| GRI | R1 | ES_variability_i>0.95 and ES_streak_l>2.13 and HMP_1>0 and Maximum_3>490 and NL<12 |
| GRI | R2 | Maximum_2>63400 and Minimum_3<2400 |
| GRI | R3 | HMP_3<4.60 and ES_variability_l<0.10 and Minimum_2>76 |
| QUEST | R4 | ES_variability_l>2.6 |
| QUEST | R5 | ES_variability_l≤2.6 and ES_variability_i>0 and NL>7 |

TABLE IV
RULE SET FOR NTL DETECTION

| |
|---|
| IF CUSTOMERS ∈ (SAMPLE1∩R1) THEN 'SUSPECT' |
| IF CUSTOMERS ∈ (SAMPLE1∩(R1∪R2)) THEN 'SUSPECT' |
| IF CUSTOMERS ∈ (SAMPLE1∩(R1∪R2∪R3)) THEN 'NORMAL' |
| IF CUSTOMERS ∈ (SAMPLE2∩R4) THEN 'SUSPECT' |
| IF CUSTOMERS ∈ (SAMPLE2∩R5) THEN 'SUSPECT' |

### D. Data Mining Techniques: Predictive Data Mining

The major goal of the predictive module is the suggestion of a rule set to characterize each of two following classes: "normal" customer or "anomalous" customer. We characterized each consumer by means of the attributes described in the previous section.

*Classification Method:* The predictive (or classification) method uses supervised learning. The attributes representing each customer, from the preliminary descriptive step, are shown in the Appendix. A feature named "suspect" is added. If $\text{'suspect'} = 1$, the customer had a non-technical loss during the period of study. We should clarify that our experiment refers to real cases, including more than 10 000 customers, and the available database has a significant limitation: although $\text{'suspect'} = 0$, it is possible that the customer had a non-detected NTL. As often occurs in power companies, it is not very realistic to assume that all the customers from a large sample are inspected.

The classification method uses, first of all, the Generalized Rule Induction (GRI) model to obtain rules R1 to R3. After, it uses the QUEST decision tree to obtain the remaining rules R4 and R5. First, the GRI model discovers association rules in the data. The advantage of the association rule algorithm over the more standard decision tree algorithms is that associations can exist between *any* of the attributes. The GRI algorithm extracts rules with the highest information content based on an index that takes both the generality (support) and accuracy (confidence) of rules into account. GRI can handle numeric and categorical inputs, but the target must be categorical: $\text{'suspect'} \in \{0;1\}$. Afterward, the QUEST model provides a binary, simple, and complete classification model.

Table III shows the description of the obtained rules. Table IV shows the application of the rule set.

The structure of this classification module is the following: the full sample studied is composed by 10 279 customers, 188 of them with detected NTL in the period of study (feature $\text{'suspect'} = 1$) and 10 091 "normal" or not detected, with $\text{'suspect'} = 0$. First, Rule 1 applies to 102 customers, 78 of them classified with $\text{'suspect'} = 0$ and 24 with $\text{'suspect'} = 1$. Customers included in Rule 1 are removed from the rest of the sample.

The remaining sample is made up of from 10 177 customers, 164 of them with detected NTLs in the period of study (feature $\text{'suspect'} = 1$) and 10 013 with $\text{'suspect'} = 0$. Rule 2 applies to 117 customers, 103 of them classified as $\text{'suspect'} = 0$ and 14 with $\text{'suspect'} = 1$.

Rule 3 applies to 3364 customers of the remaining sample (10 060 customers), obtaining 3348 classified with $\text{'suspect'} = 0$ and 16 with $\text{'suspect'} = 1$. GRI model does not offer much more room for improvement.

The remaining sample, 6696 customers, 6562 classified with $\text{'suspect'} = 0$, and 134 with $\text{'suspect'} = 1$, will be analyzed by the next algorithm, the QUEST decision tree. This model contributes to the process with a binary classification method for building decision trees, designed to reduce the processing time required for large C&TR analysis, while also reducing the tendency found in classification tree methods to favor predictors that allow more splits. All splits are binary. The QUEST tree adds two meaningful rules: Rule 4 applies to 61 customers, 58 of them with $\text{'suspect'} = 0$, and 3 with $\text{'suspect'} = 1$. Rule 5 applies to 692 customers, 665 of them with $\text{'suspect'} = 0$ and 27 with $\text{'suspect'} = 1$. The obtained rules are simple, with straightforward interpretation, and are integrated in the framework.

The test of the set of rules generates four values, according to the following classifications [5]:

1) True positives (TP): quantity of test registers correctly classified as fraudulent.
2) False positives (FP): quantity of test registers falsely classified as fraudulent.
3) True negatives (TN): quantity of test registers correctly classified as normal.
4) False negatives (FN): quantity of test registers falsely classified as normal.

Table V summarizes the described test and adds support and confidence data. Usually, the support can be defined according to the entire rule, the antecedents, the consequents, etc. In this paper, it is defined as the proportion of the complete population for which the antecedents are true and it is represented as a percentage. For example, given Rule 1, $support = ((24 + 78)/10\,279) \times 100 = 1\%$. The confidence is defined as the ratio of the rule support to antecedent support. This indicates the proportion of records with the specified antecedent(s) for which the consequent(s) is/are also true. For example, given Rule 1, $confidence = (24/(24 + 78)) \times 100 = 23.5\%$.

Results in Table V can be interpreted in a practical way. This classification can be used to assign new customers to existing classes and/or to inspect customers that had not been previously inspected but that belong to a class with a high rate of historical NTL. In this last sense, Endesa staff action is required. It is necessary to add new information to "suspect" customers: those who have initiated, changed, or canceled their contract in the period of study, those who have payment problems and other qualitative features. The Endesa staff, due to the extremely high cost of the in-situ inspection for this class of customers, usually only

TABLE V
TEST OF THE SET OF RULES

| Rule | Support | Confidence | TP | FP | TN | FN |
|---|---|---|---|---|---|---|
| R1 | 1.0% | 23.5% | 24 | 78 | - | - |
| R1∪R2 | 2.1% | 17.3% | 38 | 181 | - | - |
| R1∪R2 ∪R3 | 34.8% | 17.3% | 38 | 181 | 3348 | 16 |
| R1∪R2 ∪R3∪R4 | 35.4% | 14.6% | 41 | 239 | 3348 | 16 |
| R1∪R2 ∪R3∪R4 ∪R5 | 42.1% | 7.0% | 68 | 904 | 3348 | 16 |

revises and inspects small samples (100 or so medium-high consumption customers). This is a real restriction of the obtained results. Depending on the available number of customers to be inspected, Endesa staff can select a set of few rules, with a high rate of correct fraud identification (confidence around 20%), or a set of many rules, with less confidence but more support and more TPs.

Both algorithms create a robust model and the time for training large data set is reduced. A divide and conquer strategy is used [24]. These algorithms work by splitting the sample, based on attributes that provide the maximum information gain. Each subsample defined by the first split is split again, based on different attributes or in different classes of the same attributes, and the process is repeated until the subsamples cannot be split with a significant information gain.

The model evaluation is performed using ten-fold cross validation [32]. This kind of evaluation was selected to train the algorithms using the entire data set and to obtain a more precise model. This will increase the computational effort but improves the model's capacity for generating different data sets. The evaluation is performed by splitting the initial sample in ten subsamples in order to fill consumption range. The model is trained using 9/10 of the data set and tested with the 1/10 left. This is performed ten times on different training sets, and finally, the ten estimated errors are averaged to yield an overall error estimate. The overall accuracy obtained is around 80%.

## IV. CONCLUSION

This paper deals with the characterization of customers in power companies in order to detect consumption NTLs. A bibliographical review has been made, and a new framework is presented, to find relevant knowledge about the particular characteristics of the electric power customers and to describe the main features available in the companies databases. The authors present two innovative statistical estimators to weigh variability and trend of the customer consumption, also considering other features that contribute to information gain. The final classification model is presented by a rule set. The model creates a characterization of customer's classes based on the most relevant attributes previously selected. This classification can be used in two ways: to assign new customers to existing classes and to inspect customers that had not previously been inspected but that belong to a class with a high rate of historical NTLs: abnormalities and fraud in customer consumption.

The framework presented follows two specific ideas:

- To highlight the special importance of the use of proposed variability and trend estimators in NTL detection, not published on the literature before.
- To present a robust rule set with a high rate of correct NTL identification, based in the most relevant customer's attributes available on the utility database. This rule set allows a high reduction in the number of customers to be revised by the Endesa experts. For example, in the case study, rule R1 suggests to check 102 customers from a sample of 10 000 customers.

The quality of this framework is illustrated by a case study that uses a real database, supplied by the Endesa Company. The results obtained were satisfactory considering the limitation of a real database, demanding bigger effort in the preprocessing and data consolidation steps. Only 188 of 10 279 customers (less than 2%) of the selected registers for mining present results of NTLs inspected. Regardless of the difficulty to study real data instead of simulated data, rate of correct fraud identification (between 7% to 20%) significantly improved previous company detection campaigns, referring to medium-high consumption customers.

Usually, all the information about billing process is strictly confidential. Utility companies do not offer data about their NTL detection results and the quantitative comparison between our method and other methods is difficult. Nevertheless, the following comments show the real (economic) value of the implementation of the detection methodology:

- Statistical studies indicate that, at the origin, in the distribution networks, the non-technical losses in Spain are around 1.5% of the distributed energy. Normally they are distributed as 40%–45% of frauds and 55%–60% of abnormality.
- Detected NTLs can be recovered (rebilling the customers) so that the process is very important and profitable for Electric Distribution Companies. In the last five years, the Control of Non Technical Losses Process of Endesa Distribucion (Spain) has detected and reduced its NTLs 800–900 GWh per year approximately. The framework presented in this paper is now in the testing phase but it offers promising results. In the last two years, and only for a few small samples from low voltage customers (about 15 kWh), on the order of 1 GWh of energy has been rebilled. In any case, it also reveals some NTL cases not usually detected by means of regular company inspections.

## APPENDIX
## MAIN ATTRIBUTES IN NTL DETECTION

$ES\_variability\_i$: Proposed variability estimator, calculated from invoices.

$ES\_variability\_l$: Proposed variability estimator, calculated from lectures.

$ES\_streak\_i$: Proposed consumption trend estimator, calculated from invoices.

$ES\_streak\_l$: Proposed consumption trend estimator, calculated from lectures.

$HMP\_i$: Hours of consumption at maximum contracted power, to refer to the time zone $i$ ($i = \{1, 2, 3\}$).

$Maximum\_i$: Maximum value of consumption, in kWh, in time zone $i$ ($i = \{1, 2, 3\}$).

$Minimum\_i$: Minimum value of consumption, in kWh, in time zone $i$ ($i = \{1, 2, 3\}$).

$NL$: Number of valid consumption lectures in the period of study ($NL = \{1, 2, \ldots, 24\}$).

## ACKNOWLEDGMENT

The authors would like to thank colleagues M. A. López and F. Godoy for their valuable assistance in the project. The authors also would like to thank J. I. Cuesta, T. Blazquez, and J. Ochoa for their help and cooperation to extract the data from Endesa.

## REFERENCES

[1] R. Wheeler and S. Aitken, "Multiple algorithms for fraud detection," *Knowl. Based Syst.*, no. 13, pp. 93–99, 2000.

[2] K. S. Yap, Z. Hussien, and A. Mohamad, "Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm," in *Proc. 3rd IASTED Int. Conf. Advances in Computer Science and Technology*, Phuket, Thailand, Apr. 2–4, 2007.

[3] J. Filho, "Fraud identification in electricity company costumers using decision tree," in *Proc. IEEE/PES Int. Conf. Systems, Man and Cybernetics*, The Hague, The Netherlands, 2004.

[4] J. Cabral, J. Pinto, E. M. Gontijo, and J. Reis, "Fraud detection in electrical energy consumers using rough sets," in *Proc. 2004 IEEE Int. Conf. Systems, Man and Cybernetics*, 2004.

[5] J. Cabral, J. Pinto, K. Linares, and A. Pinto, "Methodology for fraud detection using rough sets," in *Proc. 2006 IEEE Int. Conf. Granular Computing*, 2006.

[6] J. Cabral, J. Pinto, E. Martins, and A. Pinto, "Fraud detection in high voltage electricity consumers using data mining," in *Proc. IEEE/PES Transmission and Distribution Conf. Expo. T&D*, Apr. 21–24, 2008.

[7] F. Biscarri, I. Monedero, C. León, J. Guerrero, J. Biscarri, and R. Millán, "A data mining method based on the variability of the customers consumption," in *Proc. 10th Int. Conf. Enterprise Information Systems ICEIS2008*, Barcelona, Spain, Jun. 12–16, 2008.

[8] M. Sforna, "Data mining in power company customer database," in *Electric Power Systems Research*. London, U.K.: Elsevier, 2000, vol. 55, pp. 201–209.

[9] Editorial, "Recent advances in data mining," *Eng. Appl. Artif. Intell.* vol. 19, 2006, pp. 361–362.

[10] Y. Kou, C.-T. Lu, S. Sinvongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Proc. 2004 IEEE Int. Conf. Networking, Sensing and Control*, Taiwan, Mar. 21, 2004, pp. 89–95.

[11] T. Fawcett and F. Provost, "Adaptative fraud detection," *Data Mining Knowl. Discov.*, vol. 1, pp. 291–31, 1997.

[12] M. Artís, M. Ayuso, and M. Guillín, "Modeling different types of automobile insurance frauds behavior in the Spanish market," in *Insurance Mathematics and Economics*. New York: Elsevier, 1999, vol. 24, pp. 67–81.

[13] S. Daskalaki, I. Kopanas, M. Goudara, and N. Avouris, "Data mining for decision support on customer insolvency in the telecommunication business," *Eur. J. Oper. Res.*, vol. 145, pp. 239–255, 2003, Elsevier.

[14] H. X. He, J. C. Wang, W. Graco, and S. Hawkins, "Application of neural networks to detection of medical frauds," *Expert Syst. Appl.*, vol. 13, no. 4, pp. 329–363, 1997, Elsevier.

[15] H. He, W. Graco, and X. Yao, "Application of genetic algorithm and k-nearest neighbors in medical fraud detection," in *Lecture Notes in Computer Science*. New York: Springer, 1999, vol. 1585, LNCS, pp. 74–81.

[16] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *Proc. 11th IEEE Int. Conf. Tools With Artificial Intelligence*, 1999.

[17] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Syst. Appl.*, vol. 32, pp. 995–1003, 2007.

[18] 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 Syst.*, vol. 21, no. 4, pp. 1672–1682, Nov. 2006.

[19] S. Ramos and Z. Vale, "Data mining techniques application in power distribution utilities," in *Proc. IEEE/PES Transmission and Distribution Conf. Expo. T&D*, Apr. 21–24, 2008.

[20] A. Azadeh, S. Ghaderi, S. Tarverdian, and M. Saberi, "Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption," *Appl. Math. Comput.*, no. 186, pp. 1731–1741, 2007.

[21] A. Nizar, Z. Dong, and J. Zhao, "Load profiling and data mining techniques in electricity deregulated market," in *Proc. IEEE/PES Power Eng. Soc. General Meeting*, Jun. 18–22, 2006.

[22] J. Galván, E. Elices, A. M. Noz, T. Czernichow, and M. Sanz-Bobi, "System for detection of abnormalities and fraud in customer consumption," in *Proc. 12th IEEE/PES Conf. Electric Power Supply Industry*, Nov. 2–6, 1998.

[23] R. Jiang, H. Tagiris, A. Lachsz, and M. Jeffrey, "Wavelet based features extraction and multiple classifiers for electricity fraud detection," in *Proc. IEEE/PES Transmission and Distribution Conf. Exhib. 2002: Asia Pacific*, Oct. 6–10, 2002.

[24] V. Figuereido, F. Rodrigues, Z. Vale, and B. Gouveia, "An electric energy characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[25] M. I. Barao and J. A. Tawn, "Extremal analysis of short series with outliers: Sea-levels and athletic records," *J. R. Statist. Soc. Series C-Appl. Statist.*, vol. 48, pp. 67–81, 1999.

[26] M. Kantardzic, *Data Mining: Concepts, Models Methods and Algorithms*, 1st ed. Cambridge, MA: AAAI/MIT Press, 1991.

[27] P. Burge and J. Shawe-Taylor, "Detecting cellular fraud using adaptative prototypes," in *Proceeding on AI Approaches to Fraud Detection and Risk Management*. Menlo Park, CA: AAAI, 1997, pp. 9–13.

[28] D. Denning, "An intrusion-detection model," *IEEE Trans. Softw. Eng.*, vol. 13, no. 2, pp. 222–232, Feb. 1987.

[29] T. Lunt, "A survey of intrusion detection techniques," *Comput. Security*, vol. 12, pp. 405–418, 1993.

[30] J. McCarthy, "Phenomenal data mining," *Commun. ACM*, vol. 43, no. 8, pp. 75–79, Aug. 2000.

[31] P. L. Brokett, X. Xia, and R. A. Derrig, "Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud," *J. Risk Insur.*, vol. 65, no. 2, pp. 245–274, 1998.

[32] I. Witthen and E. Frank, *Data Mining—Practical Machine Learning Tools and Techniques With Java Implementations*. New York and San Mateo, CA: Morgan Kaufmann, Academic, 2000.

**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 has been a Professor of electronic engineering at the University of Seville since 1991 and currently is the CIO of the University of Seville. His research areas include knowledge based systems and computational intelligence focus on Utilities System Management.

Dr. León is a Member of the IEEE Power Engineering Society.

**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 with the Polytechnic University School of Seville. His research areas include electricity markets, electrical customer classification, and fraud detection in the power electric industry.

**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.

He joined the Automatics and Robotics Department for two years and he has been a Professor in the Electronic Technology Department of the University of Seville since 1998. His research areas include artificial intelligence, expert systems, and data mining in the power electricity industry.

**Juan Ignacio Guerrero** received the B.Sc. degree in computer science from the University of Seville, Seville, Spain, in 2006, where he is pursuing the Ph.D. degree in the Electronic Technology Department.

He is currently an Assistant Professor with the University of Seville. His research areas include artificial intelligence, neural networks, expert systems, and data mining focus on Utilities System Management.

**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. Currently, he is also collaborating as an Associate Professor at the Polytechnic University School of Seville.

**Rocío Millán** received the B.Sc. degree and the Ph.D. degree in economics and business administration from the University of Seville, Seville, Spain, in 1985 and 1996, respectively.

She worked as Professor of economic theory and finance in this university for more than ten years and is working for Endesa as Metering Control Deputy Director. Her research areas include public deficit, energy futures markets, and NTLs detection in electricity companies.