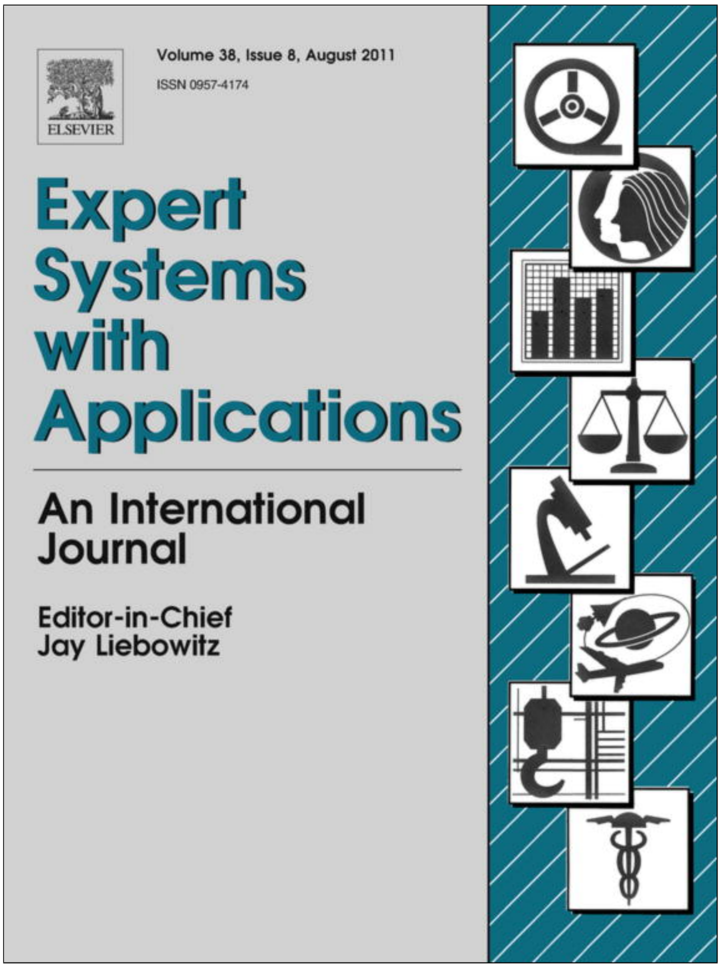

ELSEVIER

Volume 38, Issue 8, August 2011

ISSN 0957-4174

Expert Systems with Applications

An International Journal

Editor-in-Chief
Jay Liebowitz

Expert Systems with Applications 38 (2011) 10274–10285

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Integrated expert system applied to the analysis of non-technical losses in power utilities

Carlos León [a], Félix Biscarri [a], Iñigo Monedero [a], Juan I. Guerrero [a,*], Jesús Biscarri [b], Rocío Millán [b]

[a] School of Computer Science and Engineering, Electronic Technology Department, Av. Reina Mercedes S/N, 41012 Seville, Spain
[b] Endesa, Non Technical Losses Department, Borbolla Building, Av. Borbolla S/N, 41092 Seville, Spain

ARTICLE INFO

Keywords:
Expert system
Data mining
Text mining
Utilities
Power

ABSTRACT

The detection of non-technical losses (NTLs), in most papers, commonly deals with the utilization of the registered consumption for each customer; besides, some researchers used the economic activity, the active/reactive ratio and the contract power. Currently, utility company databases store enormous amounts of information on both installations and customers: consumption, technical information on the measure equipment, documentation, inspections results, commentaries of inspectors, etc. In this paper, an integrated expert system (IES) for the analysis and classification of all the available useful information of the customer is presented. Customer classification identifies the presence of an NTL and the problem type. This IES include several modules: text mining module for analysis of inspector commentaries and extraction of additional information on the customer, data mining module to draw up the rules that determine the customer estimate consumption, and the Rule Based Expert System module to analyze each customer using the results of the text and data mining modules. This IES is used with real data extracted from Endesa company databases. Endesa is the most important power distribution company in Spain, and one of the most significant companies of Europe. This IES is used in the test phase by human experts in the Endesa company. In this phase, the IES is used as a Decision Support System (DSS), as it contains another module which provides a report with additional information about the customer and a summarized result that the inspectors can use to reach a decision.

1. Introduction

Today, in the utility distribution companies, cost reduction is one of the main issues of concern. Therefore, identifying and reducing the non-technical losses (NTLs) are the main objectives. NTL is defined as a non-billed energy due to the existence of irregularities or deviations in the customer facilities. These anomalies or frauds cause an imbalance between the company registered consumption and the real consumption and precipitating serious economic losses, due to the lack of billed consumption.

Utility companies like Endesa solve the problem by conducting massive inspections on a customer's set which satisfies generic conditions. The time and economic costs involved in such inspections are very high. Normally, such large scale inspections achieve 12% success rate at best.

This paper proposes an integrated expert system (IES) which globally analyzes all the available information on Endesa's databases, and differentiates between unnecessary and useful information. The objective of this analysis is classification of customers under one or more different categories, to help Endesa staff identify and categorize NTLs.

This system uses real samples extracted from Endesa's databases. It is currently in the test phase, and used as a Decision Support System (DSS) although it contains a report-generating module that summarizes the result analysis and its classification.

The IES is chiefly composed of several modules: text mining module to collect information on the inspectors' commentaries, data mining module to set up rules to check if the customer consumption is normal, and the Rule Based Expert System (RBES), which uses the generated information by other modules. Besides, it uses a set of parameters (from the domain experts) to classify the customers based on the problems they face. These rules are derived from the inspectors and the Endesa staff.

In most of the earlier published papers, the authors use several features for NTL detection, like consumption data, economic activity, active/reactive rate, and contracted power. The proposed solution uses all the available customer information that the company has like consumption reports, information on the measuring equipment, inspectors' commentaries, documentation, etc. Thus, while the classification process is being performed, time and economic costs are simultaneously reduced. The efficiency accordingly increases due to a cut in false NTLs inspections.

[*] Corresponding author. Tel.: +34 679231193.
E-mail address: juaguealo@us.es (J.I. Guerrero).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.062

The proposed IES is a part of the development project in collaboration with Endesa one of Spain's largest companies, including more than 10 million customers.

In this paper, the proposed IES is explained in great detail, under several sections:

- *The Bibliographical Review.* In this section, authors do a complete review of the papers related with the techniques and technologies used in the IES.
- *The MIDAS Project section explains the project in which this IES is developed.* Also, several concepts and procedures related with the management and control of the customers' consumption are discussed.
- *The System Architecture section reveals the architecture of the proposed IES, and the sections following it describe each of it modules:* data mining, text mining and RBES modules.
- *The Section 9 describe the current results of IES application.* In the Section 10, current researches are briefly presented.

2. Bibliographic review

Frauds and anomalies represent a serious problem for utility companies. Advancement in the different techniques and methodologies related with the Artificial Intelligence has enabled the detection, classification, reduction, and prediction of the problems of frauds and anomalies.

The proposed IES employs several techniques to detect and classify customer abnormalities related with frauds and other anomalies (NTLs). This system combines RBES, data mining and text mining. Several areas of research related with this work are present, although based on pattern detection: for example, the main fields of research include those with high economic risk: finances and economics, telecommunications and others.

2.1. Finances and economics

Sánchez, Vila, Cerda, and Serrano (2009) use the fuzzy association rules for fraud detection in credit card transactions; Quah and Sriganesh (2008) use the auto-organization maps to decipher, filter and analyze the customers' behavior in fraud detection; Kirkos, Spathis, and Manolopoulos (2007) use several data mining techniques (decision trees, neural networks and Bayesian belief networks) analyzing the factors related with Fraudulent Financial Statements (FFS) comparing the efficiency among all the methods. Besides, Richardson (1997) compares statistical and neural network techniques. Wheeler and Aitken (2000) use case-based reasoning to detect frauds in the credit approval process. All these papers propose different techniques, related in large part with data mining, demonstrating their efficiency in fraud detection. Also, Hand and Blunt (2001) have made a significant contribution by an in depth treatment of all the prior information analysis processes for application of data mining techniques. They classify customers on the economic sector in which the transfer is made. There are others subjects within the scope of this research field which use very similar methods, although based on other features, for instance, anti-money laundering (Gao & Xu, 2009), risk evaluation (Yu, Yue, Wang, & Lai, 2010), credit scoring (Huang, Chen, & Wang, 2007), and financial distress (Chen & Du, 2009).

2.2. Telecommunications

In this field of research several other methods and tools related with data mining (Daskalaki, Kopanas, Goudara, & Avouris, 2003; Wang, Wang, Zhan, Li, & Wang, 2004), Support Vector Machine (SVM) (Wang et al., 2009), neural networks and fuzzy rules (Estévez, Held, & Pérez, 2006) and Expert Systems (Hilas, 2009) are applied. These techniques identify various types of trouble solving, though they will need to use the domain expert knowledge at a certain level.

2.3. Others

There are several other research fields in which these techniques are applied for pattern detection: medicine (Cierniakoski, De, & May, 1991; He, Wang, Graco, & Haukins, 1997; Yang & Hwang, 2006), Intrusion Detection or IDS (Depren, Topallar, Anarim, & Ciliz, 2005; Hernández-Pereira, Suárez-Romero, Fontenla-Romero, & Alonso-Betanzos, 2009; Kim, Im, & Park, 2010), transport (Chen, Qu, & van Zuylen, 2010), etc. All these papers use different search methods to detect anomalous of fraudulent patterns, ruling out the others by any of the means, except in the case of Richardson (1997). He establishes a classification system on several groups based on the problems involved in each case.

Unlike these papers, the electrical energy customer usually displays more than one type of pattern that indicates the same NTL. Therefore, the IES classifies customer employing several categories.

2.4. NTL detection

There are several papers focusing on NTL detection:

Galván, Elices, Muñoz, Czernichow, and Sanz-Bobi (1998) proposed a general methodology based on using Radial Basis Function Networks (RBFN), with the following steps: (1) variable selection, (2) data filtering, (3) Model fitting, (4) Model analysis, and (5) Model evaluation. The third step, the RBFN input, is taken of the variables monthly periods of each annual consumption pattern and active/reactive consumption. For instance, the methodology is applied in two economic sectors: low-voltage lodging sector and high-voltage farm watering sector.

Cabral, Pinto, Onofre, Gontijo, and Filho (2004) proposed an application that used rough sets to classify the categorical attribute values to detect fraud in customer electrical energy use. The continuous attributes were converted to discrete. The system achieved a fraud rightness rate around 20%. The authors expressed that the main difficulty to detect electrical energy profiles was the low 'fraudulent customer'/'normal customer' ratio, around the 5%, in Brazilian electrical energy distribution companies. They also admitted that to add to their problems, many fraudulent customer behaviors appeared as normal behavior. Cabral, Pinto, Linares, and Pinto (2006) added Knowledge Discovery Databases (KDD) to improve the success of fraud pattern detection. In the previous version, they performed a KDD process by selecting 12 attributes (2 string and 10 numerical).

Unlike these papers, the proposed IES does not include a variable selection process, as this selection is performed in a previous data mining step, according to the domain expert knowledge. Besides, it uses several modules to adapt unstructured and unusual information. It enables the creation of a set of rules to improve and fine tune the NTLs detection results.

Therefore, all the reviewed papers reveal several ideas in common:

- Using different techniques related with the data mining and the pattern detection.
- In the papers directly related with the NTL detection in energy consumption, only a few key indicators are used: the energy consumption, the economic activity, the contracted power and the active/reactive ratio. Much of the interesting information, for instance the reports of the NTL inspectors, is rejected.

The present paper proposed a new integrated expert system, which includes all the available information in Endesa's databases, using real inspection results to test the system. The domain expert knowledge is obtained from the Endesa staff. To incorporate the knowledge three techniques are used:

- The static knowledge is included in the RBES.
- The data mining technique involves different statistical techniques to create rules related with the estimated consumption that a customer without NTL would have.
- The text mining technique comprises knowledge of the customers' facilities provided by Endesa's inspectors. The text mining objective is unstructured information, using natural language, and it can be used to generate new information (Yang & Lee, 2004), to extract information (Sung & Chang, 2004), to summarize information (Aliguliyev, 2009), etc.

On the contrary, Schutzer (1990) suggested applying a business expert system for fraud detection. Liao (2005) submitted references for fraud management, and Rahman and Lauby (1993) proposed the use of expert systems in Power System Planning.

These papers advanced the usefulness of an expert system in the utility distribution field, to implement the domain expert knowledge. In recently published papers, authors have not produced papers related with electrical NTLs detection containing a RBES. However, there are papers relating to fraud detection in other research fields. For example, Cierniakoski et al. (1991) for processing medical insurance claims; Bowen (1994) for police investigators of economic crimes; and Hilas (2009) for fraud detection in private telecommunications networks.

3. The MIDAS Project

The aim of the MIDAS Project is NTL detection by utilizing the data mining techniques and other Artificial Intelligence (AI) techniques over Endesa's databases. Initially, the project began with the NTL pattern detection. These advancements are published in various papers (Biscarri et al., 2008; Biscarri et al., 2009), where different techniques of data mining and neural networks are used to detect the consumption pattern by which the NTLs they identify with are proposed. Supervised and non-supervised techniques too are applied.

The information flow diagram is shown in Fig. 1 depicting the research methodology and the IES role in the project. The steps of this cycle are:

- *Sample Selection.* A set of customers is selected. Contracted power, economic sector and geographic location features are used in this step.
- *Sample extraction from corporate databases.* The main the difficulty in this step is the large number of customers. Two extraction methods may be used: designing of an extraction system for batch processing or designing of an intermediate database. The first method is used at night or during inactive database periods. The second method, used in the IES, works with the off-line information and it is actualized periodically using a batch process.
- *Application of studies based on data mining and AI using customer consumption information.* The output is a list of customers 'suspected' of NTL. These studies are described in detail by Biscarri et al.. (2008), Biscarri et al.. (2009).
- *Analysis of customers.* In this phase, the remaining available information on the customers is analyzed, to determine if those with a 'suspected' anomaly are wrongly classified. Normally, this process demands much time and effort from the inspectors or domain experts, as it is necessitates a review of all the customer information manually, one at a time. This step will be replaced with the proposed IES that offers an automatic and accurate analysis.
- *Customer review.* The final conclusions, obtained through the application of studies and the customer analysis are verified with 'in situ' inspections.
- *Review of the results.* The results obtained by the inspectors are checked. This information improves the future studies and the IES. Fraud identification per number of 'in-situ' inspection percentage as reported (15–20%, or higher depending on customer characteristics).

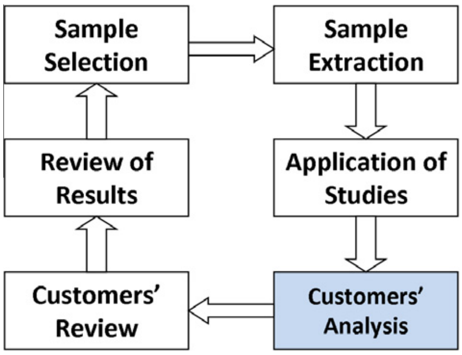

Fig. 1. The information flow diagram of the MIDAS Project.

Actually, after analyzing the 'in-situ' inspections conducted by the Endesa staff, the authors conclude that the utilization of additional techniques becomes necessary to detect the false-positives, i.e., the customers who present an NTL pattern but who are not fraudulent or anomalous. We had earlier experienced that the number of false-positives was high. This is a serious problem in all the utility companies; and there is no easy solution as customer consumption depends on several factors.

Also, the high cost associated with 'in-situ' inspections poses a great limitation. The inspections conducted to identify NTL are more expensive than a standard revision or equipment maintenance, as more qualified inspectors are needed.

The proposed IES attempts to meet this need by adding new verifications to the selected customers utilizing any of the data mining methods, to minimize the false-positive cases.

Project development performed by the SPSS Clementine environment is commonly used in the commercial development of the data mining process. Liao and Wen (2007) and Liao, Hsieh, and Huang (2008) presented an interesting review of some features of this software, besides proposing utilization examples.

4. The energy distribution management and the expert system

Endesa, similar to other power utilities companies, implements several procedures to manage the energy distribution. For example, this company establishes contract procedures which lay the groundwork to create a new contract for the customer. Likewise, the company inspectors routinely check customer installation. These inspections follow established guidelines, depending on the requests by the company.

Normally, the inspectors keep adding on new information to the corporative databases or modify the existing ones on customer installation. This information includes inspector comments on any thing or any event which the inspector has observed. However, in the case of NTL search, these comments summarize inspector observations and the procedural steps taken to rectify it. When the company inspectors identify any abnormality in the customer measurement system, they notify the company. All the information related with identified NTL are reported and stored in Endesa's databases. Thus, the automatic processing of this information, involving an adequate expert system, is clearly very useful in determining NTL patterns, and consequently to reduce false NTL 'in-situ' inspections.

Endesa staff measure customer consumption by installed measurement registers. The measurement period could either be once per month (monthly) or one per 2 months (bimonthly). These data are carefully stored in corporate databases for billing purposes. Sometimes, the readings cannot be reported due to a reading error or causes outside the scope of the reader; for example, the worker cannot access the customer register. In such a scenario, the Endesa system automatically calculates an estimate of this measure which the company terms 'estimated consumption', and these are usually lower than the real consumption. For the NTL detection process, the estimated consumptions need to be carefully studied, as they could actually hide an NTL. Authors noted several cases with zero real consumption, resulting from an abnormality or a fraud, and a low billed estimated consumption, for several reasons. Therefore, for the expert system, the number of customers estimated readings becomes highly significant data.

All measurement actions are carried out at the customer's installation and the data are stored in Endesa's databases. As Endesa has more than 10 million customers, such large numbers of customers need large databases and complex hardware architecture.

When future customers request the electric energy service, they need to provide the company with personal information and further details about his economic activity, his electric supply requirement and his selected contractual power rate. The contractual rate establishes the power range and voltage to be installed on the customer property. Besides, the contractual rate provides the measuring and control equipment features, by which the price of the billed energy can be established, depending on the time band consumption and the power contracted. RBES checks customer consumption relative to the contractual information. For example, there are some customers for whom the consumption is assembled in particular time band discriminations.

5. Expert system architecture

The proposed IES is the result of a research of the all information available in company databases. Each type of information calls for its own technique based on the information type and search objective. The techniques and technologies listed below are used in the following manner:

- *RBES module.* This technology is used to include the domain expert knowledge. This is static knowledge, and normally, excludes the learning processes. This technology is added to the system as the main module. It uses the information generated by the other modules to include the dynamic knowledge or mining learning.
- *Data mining module.* This technique includes several statistical techniques, like basic statistical indicators and regression techniques. It is used to determine whether the consumption is correct based on the geographic location, contracted power, billing frequency, time band discrimination, postal code, and economic activity. This module includes a learning process (described in Section 6).
- *Text mining module.* This technique permits the inclusion of the inspectors' knowledge of certain customers. This information is chiefly used to determine if false-positives exist. It enhances the efficiency of the classification process. This module too, requires a learning process (described in Section 7).

The IES includes several modules which interact with the knowledge base. Fig. 2 shows the basic architecture of the system.

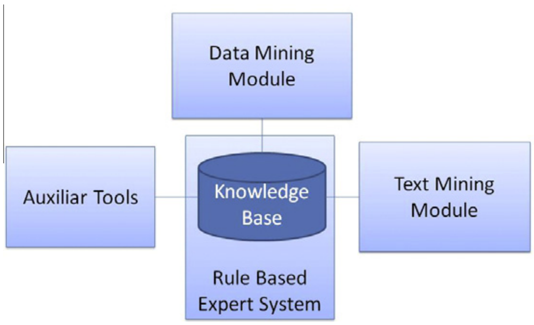

Fig. 2. System architecture.

The Auxiliary Tools perform several tasks, to check on NTL's 'suspicious' customers and to summarize the accumulated results. This module is not used in the classification process, but is very useful for the company to create summary reports.

The core of the expert system, text mining and data mining modules are done using SPSS Clementine. The knowledge base and sample storage are done using MySQL. The connection is maintained through ODBC drivers.

The data mining and text mining modules update a certain part of the knowledge base, because of which, they must be applied in anticipation of customers' studies. These modules require only periodic update. The data mining module particularly needs to update its knowledge base, either monthly or bimonthly. The updating time process depends on the number of customers. For example, with all customers in the low voltage sector (this sample is termed 'Low Voltage Sample' or 'LV Sample') this process could take 3 h (the features of the computer test are specified in Section 9). The text mining module needs to annually update its knowledge base or when there is an important change in the company. Following the earlier example, this process was executed in 2 h.

6. The data mining module

The data mining module has to conduct either a monthly or bimonthly learning process, as described in this section. Then, the information generated by this module is used in the analysis process, by the IES. The data mining model of this module thus helps to establish relations among customers' consumption, and it categorizes them using statistical techniques by estimation of the consumption curve in customers' sets, fulfilling several conditions.

The process begins with filtering customers who present a specific abnormality in consumption. This module establishes the normal consumption range for various customers' sets. The filters applied are:

- Elimination of customers possessing a high number of estimated measurements.
- Elimination of customers with three or more consecutive zero consumption readings.
- Elimination of customers showing highly dispersed consumption. For example, a customer with three or more consumption peaks greater than twice the standard deviation of consumption.

Normally, the electric consumption is highly dependent on a few contract features. The system must select the relevant customer consumption information, available in the corporate databases. The selected parameters are:

- *Contracted power.* First, the customers are divided into two main groups, the low- and high-contracted power types, based on several practical studies. These studies show that the majority of customers have contracted electric power less than 300 kW. However, customers with higher contracted power reveal different consumption behaviour.

Next, each main group is divided into 40 different subgroups using 2 vingtiles, each containing an identical number of customers. The vingtiles are numbered from 1 to 20 for the 0 to 300 kW range, and 21–40 for customers with contracted power higher than 300 kW.

Usually, customers with similar contracted power display a similar range of consumption. Domestic customers, the most numerous of the clusters, contract low power, normally between 3 and 13 kW. Small businesses, like pubs, restaurants and shops, contract a power greater than 13 kW. Normally, the interval ranging between 0 and 3 kW is assigned to stores, warehouses or auxiliary support. Contracted power higher than 300 kW is associated with industrial or distribution activities.

- *Geographic location and postal code.* These data indicate the customer location enabling the formation of geographic zones including customers with the same climatic and administrative conditions.
- *Economic sector.* This feature is also strongly related to customer consumption. Similar to what Hand and Blunt (2001) proposed, a clear relationship does exist between economic activity and consumption.
- *Time band discrimination.* This parameter enables the formation of customer sets based on the time band consumption. The aim is to determine a pattern of consumption for each time band.
- *Billing frequency.* This determines measurement as a regular recurrence. It allows tuning the statistical studies circumventing the need to make interpolations. However, these interpolations are necessary if the cycles of the readings are not the same for the customers. Normally, the time spans are either monthly or bimonthly.

The parameters presented allow the establishment of customer sets with similar behavior, by applying statistical indicator. Four sets are established to fix the normal levels of the indicators, which would show possible customers' statistical abnormalities. These sets are formed by an aggregation of the following parameters:

- The geographic location, the contracted power and the billing frequency (Group A). These parameters establish the usual consumption, in a particular geographic location. Through this group, the system extracts group consumption patterns of the customers based on these features in the 'normal' consumption range. Fig. 3 shows the monthly consumption average graph versus that of the contracted power, for a region in the north of Spain clearly indicating the relationship between customer consumption and contracted power.
- *The geographic location, contracted power, billing frequency and postal code (Group B).* In the IES-conducted analysis this set allows the detection of seasonal behaviours, for example, the customers related the increase in their consumption in summer to tourism. Fig. 4 shows the common contracted power in a particular postal code and the range of consumption within this zone. This postal code corresponded to a particular place in northern Spain.
- *The geographic location, contracted power, billing frequency, and economic sector (Group C).* This group establishes the dependence of the customers' consumption on its economic activity. In the analysis performed by the proposed IES, this information is used to establish the estimated consumption curve for a particular economic sector. The system assumes that the customers' economic activity data from Endesa's databases, is true, but it also implements processes to check the veracity. Fig. 5 shows the common contracted power in a particular economic sector and the range of consumption within this sector. This sector corresponded to restaurants and pubs in northern Spain.
- *The geographic location, contracted power, billing frequency and time band discrimination (Group D).* Studying this group is important as it helps to establish the consumption range within a time band, according to the features mentioned above. In the analysis performed by the proposed IES, it is combined with the results from the other sets. Figs. 6(a) and (b) show the difference between customer consumption within or without the dis-

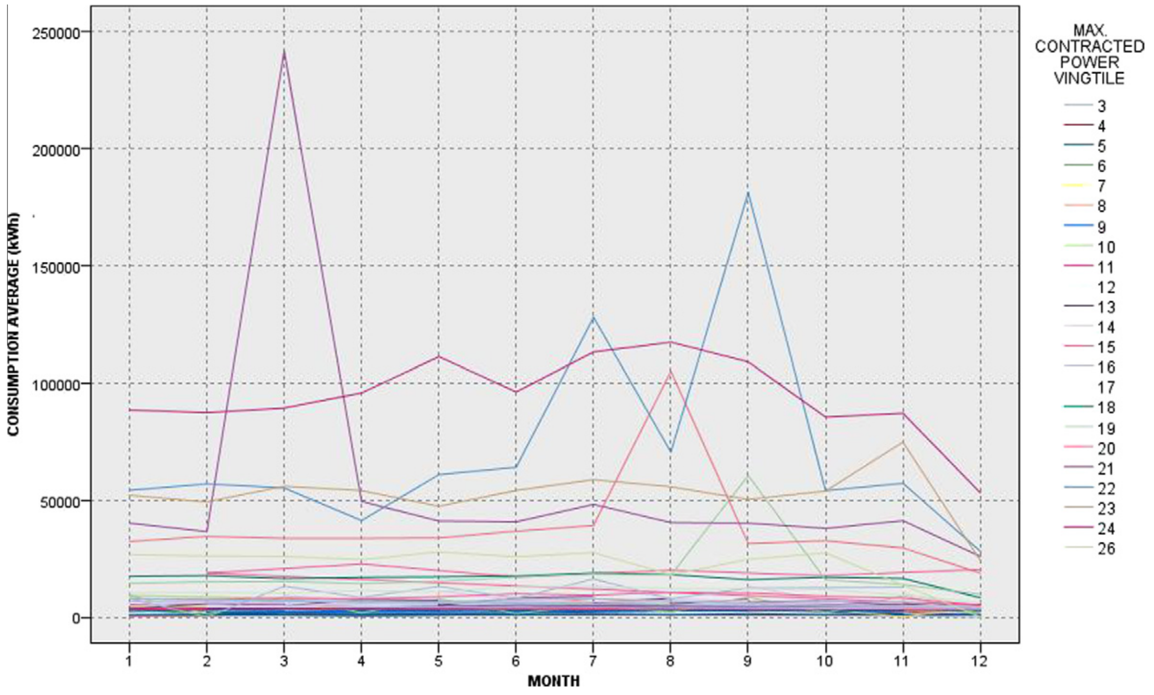

Fig. 3. Graph showing the average monthly consumption (kWh) of a geographic location (in northern Spain).

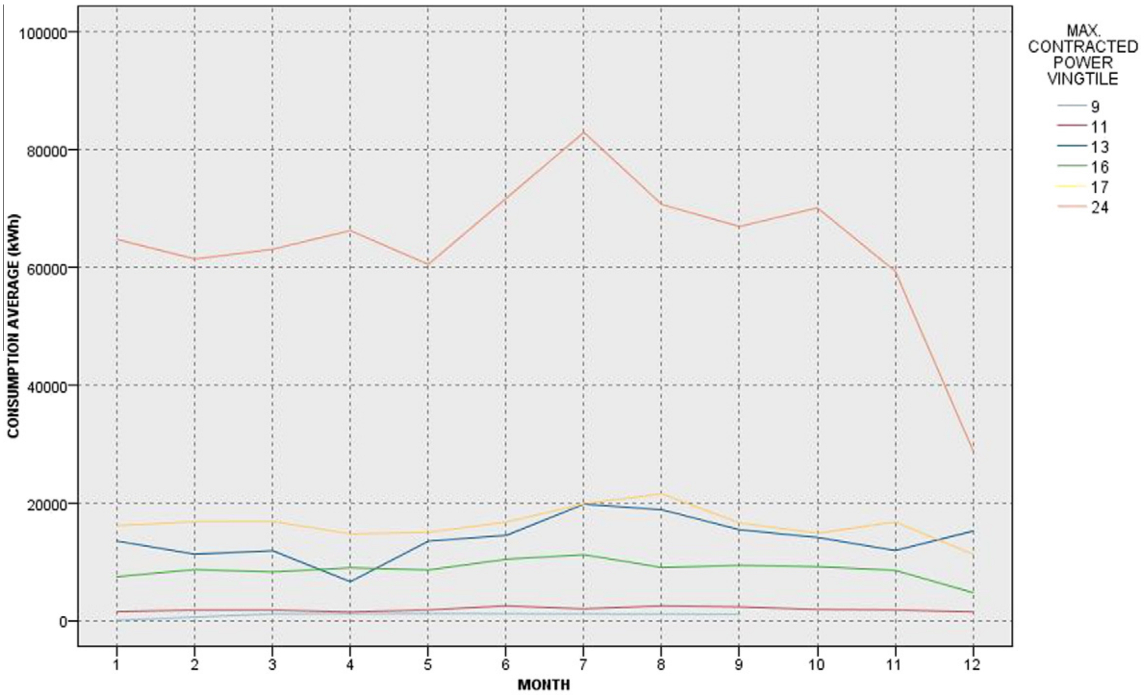

Fig. 4. Graph showing the average monthly consumption in the postal code XXX03.

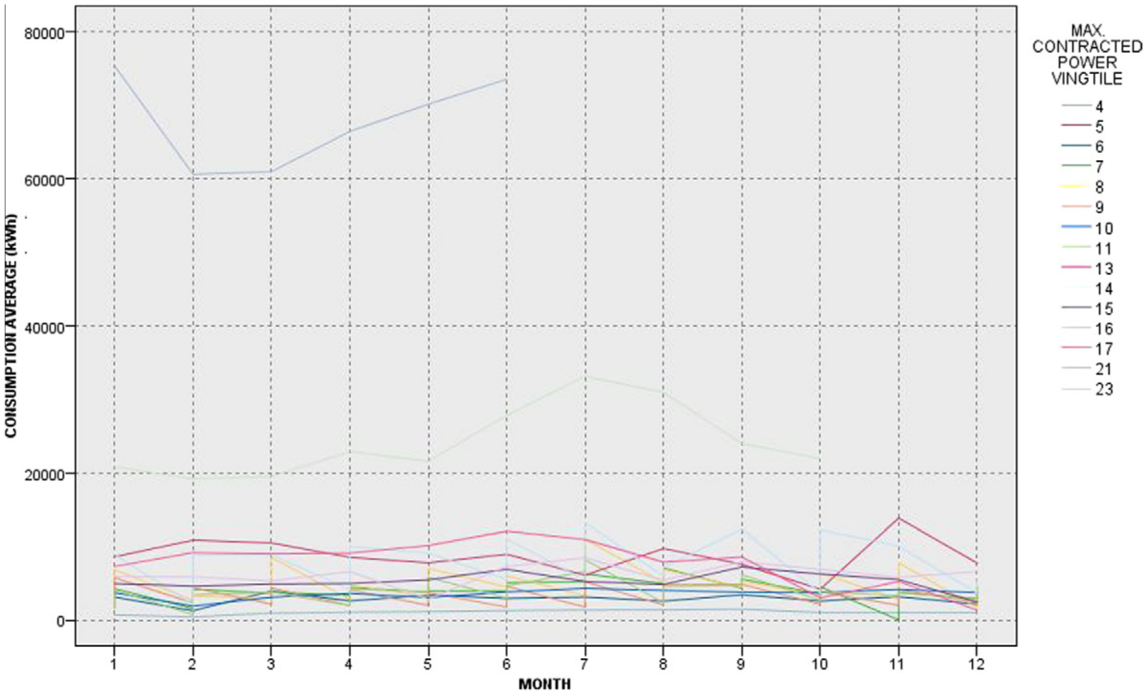

Fig. 5. Graph showing the average consumption of kWh in the economic sector of restaurants and pubs.

crimination band. The consumption curve shows that when several discrimination bands are present, a high consumption range will be seen.

The statistical indicators used in these studies are the average and the standard deviation of customer consumption calculated for each group mentioned. Also, a time division is made for each group. Temporal parameters depending on the measurement period are used to make the temporal divisions. These divisions are absolute (all the electricity consumption during the study period), yearly, seasonally, and monthly consumption. A consumption regression study is made to identify the consumption trends. Trends can explain the reduction in consumption in the customers analyzed, if the energy demands in a particular geographic location and economic sector decrease.

7. Text mining module

The text mining module learning process is discussed in this section. Following the learning process, the IES uses the information

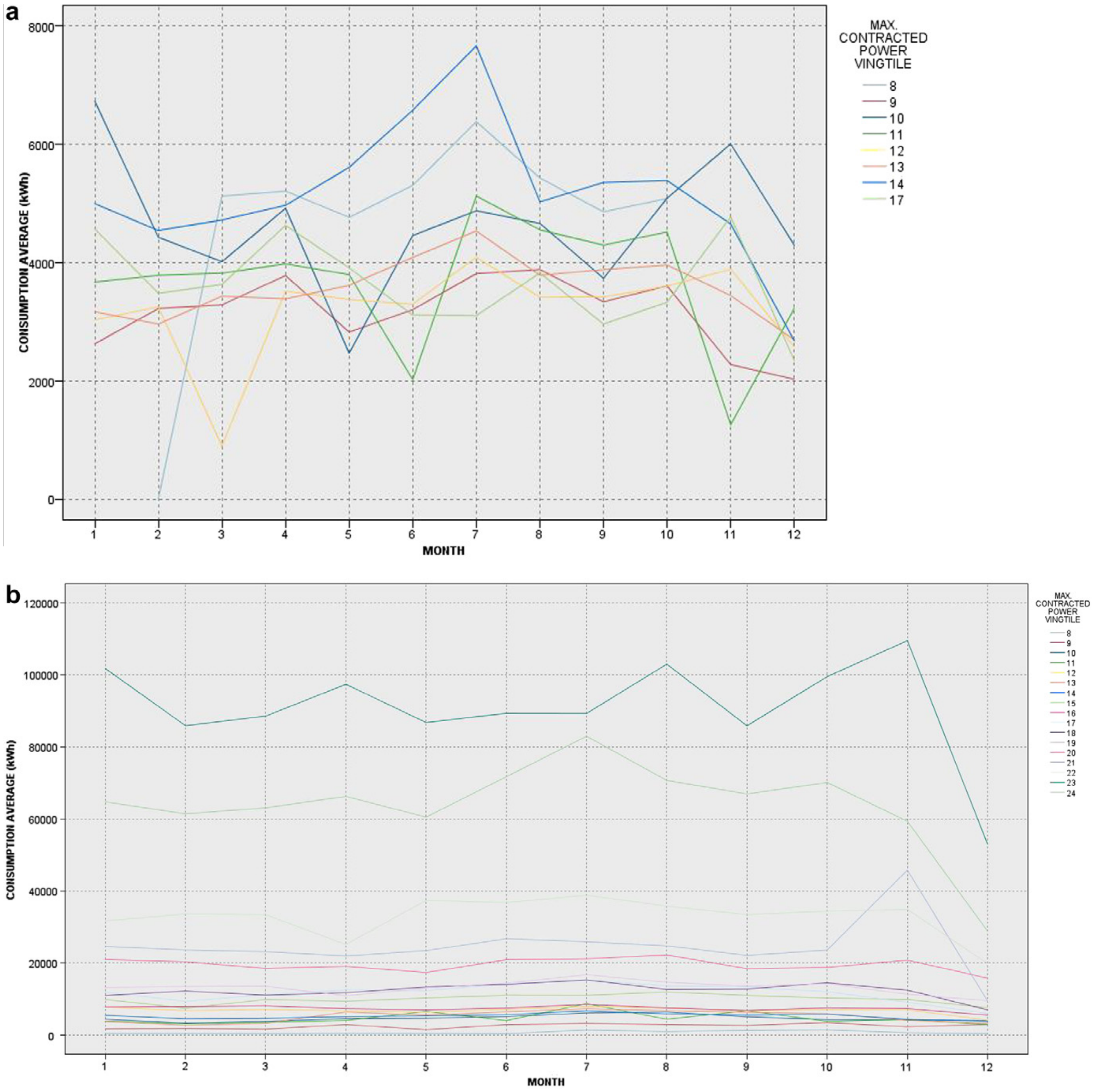

Fig. 6. Graph showing the average consumption of kWh on customers (a) without time band discrimination and (b) with three time bands discrimination.

generated in the analysis process. The main objective of this text mining module is the extraction of the information from the inspector commentaries and from the company documents on customer facilities. This is non-structured information, couched in natural language. Usually, the power company inspectors check the customer installations from time to time, and enter their results in the corporate databases. The proposed IES uses this very interesting information to improve on the analysis of the NTL research. A common term dictionary has been compiled to be used on a set of rules that execute the following actions:

- *Economic sector checking.* It is necessary to determine whether the customer informed activity sector concurs with the real information. For instance, if is possible that the customer has specified the economic sector, which has changed or not been correctly informed possibly due to a contract change. Such a scenario would make the analysis of the consumers difficult; therefore, it becomes useful to detect this abnormality and identify the customer's economic sector.
- *Characterization of the customers' consumption.* Inspectors' reports or technical repair reports can justify certain circumstances that can occur in the customer's installation: for example, the estimated measures, the low or even zero consumption or the work inspector's orders, repetitively, cancelled. We have detected cases ranging from very little to frequent cases with inaccessible or dangerous access; that explains the infrequent reading measurements. Inspectors report these situations.
- *Other information checking.* These include concepts extracted from the commentaries compiled by the technical staff members, which in turn facilitates collecting additional information

on the customer's installation, such as the existence of a capacitor bank, which explains the low reactive energy consumption or information about when the measurement installation is to occur.

The text mining module is applied to every database field that contains information in natural language. This process is performed prior to the IES analysis, as it is necessary to update the knowledge base with the information collected by the text mining module. The following steps are necessary to correctly realize this process:

1. *Data Preprocessing.* Inconsistent, erroneous or missing information must be corrected. Besides, both data integration and data transformation are performed.
2. *Concepts' Extraction.* It attempts to elicit the text field concepts, structured or otherwise. A concept can comprise one or more words. The concept may be a syntagm or a word which represents an entity (action, event, etc.). Natural Language Processing (NLP) methods are used, to extract linguistic (words, phrases, etc.) and non-linguistic (dates, numbers, etc.) concepts. An interesting review of this technique and its use in the information system management is proposed by Métais (2002). The following set of functionalities are included:
   a. *Recognition of punctuation errors.* These types of mistakes include the incorrect use of the tilde, the period, the comma, the point and comma, the dividing bar, etc. Frequently, the text fields contain commentaries in very colloquial language, with less attention being paid to the correct placement of these punctuation signs.
   b. *Recognition of spell errors.* A grouping fuzzy technology is applied. When concepts of the text are extracted, words with similar spelling (referring to the letters that compose it) or that are closely related are classified together. By applying this algorithm, mistakes of omission of letters, duplication of letters or permutation of letters are corrected.
   c. *Dictionaries.* A dictionary of technical words is compiled, as well as a dictionary of synonyms and abbreviations to help the system recognize the concepts of a more sophisticated form. Also, dictionaries of undesirable concepts are established, to determine the words or concepts that are rejected in the recognition process. In Spain, several dialects exist, which complicate extraction of the concepts.
3. *Categorization.* In this step, each concept is qualified and classified based on the inspectors' knowledge. This process classifies the key concepts or words concurrent with the functions previously specified:
   a. *Identification of a possible change of economic sector.* This classification allows the detection of 90 different economic sectors.
   b. *Characterization of the customer consumption.* Customer consumption characterization is mainly focused on detecting the justification of the anomalous measurements or of the consumption anomalies. Another utility is to eliminate minor concepts, categorizing and classifying them within groups that are designated as sets of little interest, which prevent them from being overlapped or confused with others.
   c. *Diverse information of the customer installation.*

8. Rule based expert system module

The RBES controls and monitors the analysis process. RBES employs a set of rules to determine the customers' problems. In the proposed IES, this knowledge is represented by rules of type IF-THEN-ELSE. The representation of the objects and results is done with a dynamic table, in which each row represents a contract of a different customer. The result of applying a rule is entered in a column that identifies its origin. More than 500 rules (including the rules that generated the text mining and the data mining modules), help in classifying customers under seven different groups, based on their intentions:

- *Generation rules.* These rules deal with the generation of new information from the existing customers. They are used to update erroneous information and to preprocess the database information. These rules include the rules generated by the text mining process. For example, the cycle consumption calculation (a cycle is the period of time between taking a measurement and the follow up) is performed by applying the rule, as shown in Table 1. This rule calculates the corresponding consumption between a measure (*measure*) and the previous one (*previous_measure*) which are registered in the company databases. The information generated is applied to the other groups of rules.
- *Classification rules for contract mistakes.* These rules deal with the detection of information mistakes in the customer's contract information. In Table 1, as shown. This rule, for example, allows the classification of customers with very low contracted power. These customers cannot be analyzed in the same manner as the normal or high contracted power customers are studied. They are analyzed using the text mining rules.
- *Classification rules for facility problem.* The objective is to detect information problems related to the measuring equipment. In Table 1, an example of a rule is shown, that selects customers using obsolete measuring equipment.
- *Classification rules for consumption problem.* The application of this rule set is based on the results of the previous rules, along with the information generated in data mining; the customer is classified according to his electrical consumption. Also, it establishes the limits of low consumption, excessive consumption or null consumption. The initial values for these parameters are determined during the knowledge acquisition process. Low consumption particularly, usually indicates the possibility of an NTL. Detection of low consumption customers employs the rules generated with data mining as well as the text mining process. There are several rules to determine a low consumption case, regarding the history of customer consumption or considering customer's cluster. An example, as shown in Table 1. This rule compares customer consumption (*customer_cycle_consumption*) with the results of statistical studies in which the

Table 1
Some RBES rules.

Group of rules	Rules
Generation rules	*CYCLE CONSUMPTION=IF measure>previous_measure THEN measure-previous_measure ELSE* (10$^{maxlmum\_number\_of\_digits}$-*previous_measure-1+measure*)
Consumption problem classification rules	*LOW_CONSUMPTiON=IF range_con_month_group_a > customer_cycie_consumption and range_con_month_group_b > customer_cycle_consumption and range_con_month_group_c > customer_cycte_consumption and range_con_month_group_d > customer_cycle_consumption THEN TRUE ELSE FALSE*
Contract power classification rules	*POWER_VERY_LOW=IF customer_contracted_power < 1.5 THEN TRUE ELSE FALSE*
Facility problem classification rules	*LOW_NUMBER_WHEELS=IF number_wheels<=4 THEN TRUE ELSE FALSE*

consumption range is established (*range_con_month_group_a*, *range_con_month_group_b*, *range_con_month_group_c* and *range_con_month_group_d*). In this case, the customer's consumption is compared with the low consumption range of each group described in Section 6.

- *Rules of customer selection for inspection.* These sets of rules propose a list of customers to be inspected, identified from the previous rule conclusions. This selection classifies customers under several categories, which indicate the risk of NTL. It is possible to solve the NTL without making an 'in-situ' inspection, where the problem can be solved by updating customer information in the company databases.
- *Verifying rules.* Due to the irregularities in the information available in the corporate databases, rules to check the veracity of the systems results become an absolute necessity. These rules determine if the customer has been correctly analyzed, and also if the customer cannot be reported due to too much incoherent information collected.
- *Explanation rules.* This system strengthens the reports using these IF-THEN-ELSE rules, adding interesting information on the customer or his installation. The reported information includes:
  - Contractual information like billing frequency, time band discrimination, contracted power, postal code, economic activity, etc.
  - Problems related with contract as wrongly applied rate, incorrectly contracted power, incoherent information, etc.
  - Problems related with measuring equipment: old register, warn register, lack of obligatory power control switch, etc.
  - Information and problems related with consumption: characterization of customer's consumption, unexpected low consumption, etc.
  - Other consumption operations, as the average consumption on any consumption time band.

9. Experimental results

The tests were run on a computer with a double processor AMD Opteron dual core (1, 7 Ghz), 3Gbytes of RAM and 100 Gb of hard disk space.

This IES is part of a project named MIDAS, as shown in Fig. 1. The IES have been designed to analyze the samples obtained by the application of other detection studies, based on Artificial Intelligence, statistical studies, and neural networks.

The IES was applied over a set of contracts 'suspected' of fraud or abnormalities that had been obtained from studies described in Biscarri et al.. (2008), Biscarri et al.. (2009). This 'suspected customer set' consists of 134 contracts.

These contracts are selected as they reveal an anomalous consumption pattern. The IES analyzes this set and configures it in the most restrictive manner. All the selected contracts will be inspected, and the number of inspections is highly restricted because they are very expensive. In this sense, the IES selected:

- Thirty two contracts with a possible NTL.
- Sixteen contracts with incoherent information or including some problem which could be solved without inspection.
- Eighty six contracts without a clear NTL, as they had been solved previously or those where the anomalous consumption can be explained. These contracts were classified as false NTLs.

This set of 32 contracts was inspected 'in situ'. The results of these inspections were:

- Nine contracts have an NTL.
- Two contracts have measurements problems.
- Fourteen contracts have a non-NTL.
- Two contracts are not in force, and not included in Endesa databases as yet.
- Five contracts cannot be inspected, because the customers' businesses had closed down, and it had not been entered in the Endesa databases as yet.

The IES thus filtered 86 cases of false NTLs. A review of four filtered cases is shown below. These cases have an NTL pattern, although they have several data which explain the anomalous consumption.

The first case has an anomalous consumption for several reasons. Between 2005 and 2007 there was a period of decrease in consumption. In 2007, the company started a proceeding, which

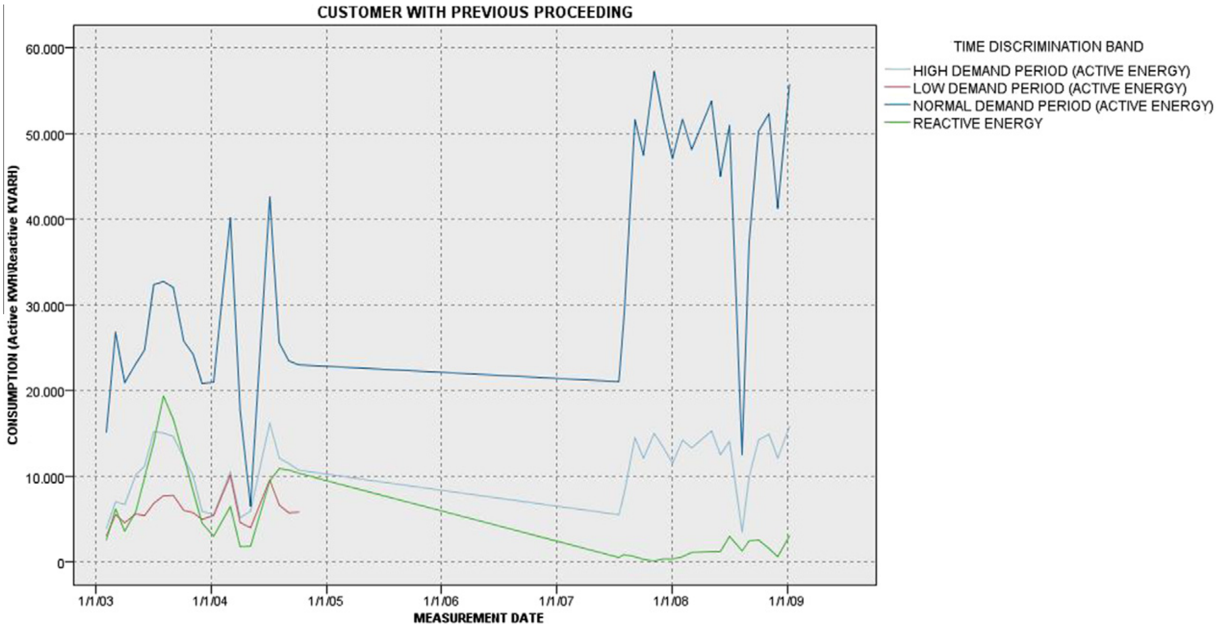

Fig. 7. Consumption graph of a customer with previous proceeding.

solved the problem. Later, there were some cycles of low consumption, which were explained in the inspectors' commentaries. They specified that the place had closed and the measurement was estimated. Fig. 7 shows the consumption of the customer in a different time discrimination band.

The second case is a customer with a contract for a fountain and irrigation engine. The consumption of this type of activity is very irregular, and usually involves a high rate of reactive energy. The IES classified the customer as a non-NTL case. The inspectors' commentaries specified that it was not possible to access and measure the equipment because it was padlocked and damaged. Besides, these types of customers show very irregular consumption as they depend on the vagaries of the climatic conditions, and the IES has no information on these. Fig. 8 shows the consumption of this customer in a different time discrimination band. The reactive energy is sufficiently high due to the latency period of the engine.

The third case is a customer with a hotel. Usually, such clients have five time discrimination bands making the analysis more difficult. Fig. 9 shows this customer's consumption in a different time discrimination band. The rules generated by the Data and Text Mining modules are very important to correctly classify this case. Using these rules, several reasons were detected: the customer's consumption is established by the tourist industry (normally, it is seasonal); and the period of zero consumption was explained by the inspectors' commentaries, as a result of a problem with

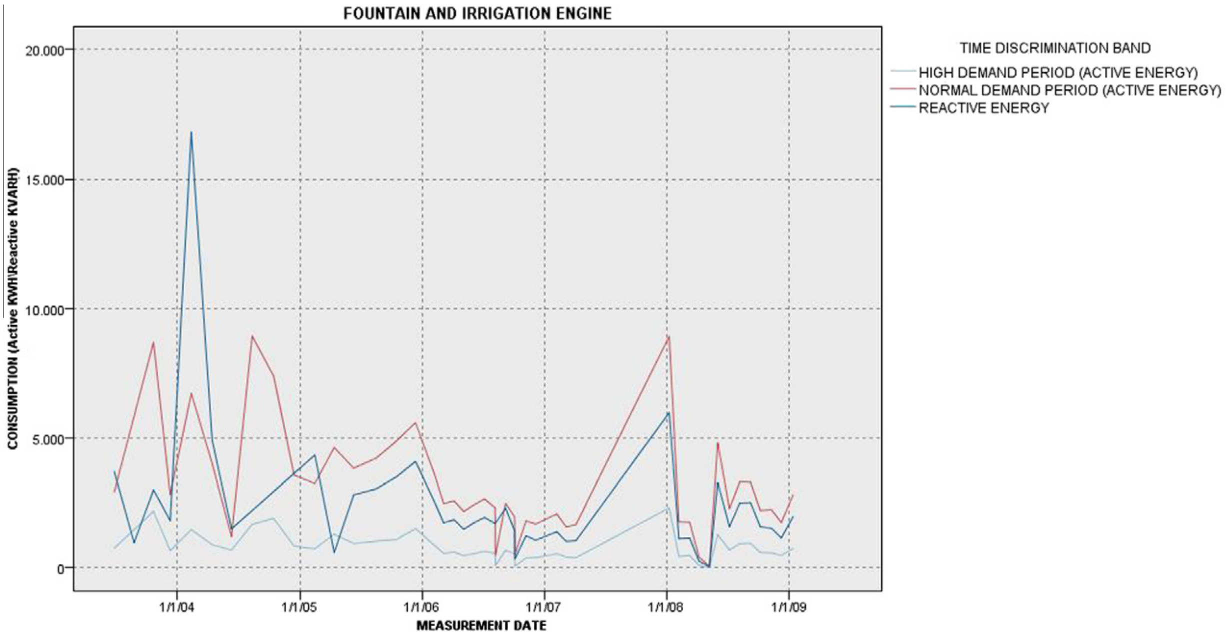

Fig. 8. Consumption graph of a fountain and irrigation engine.

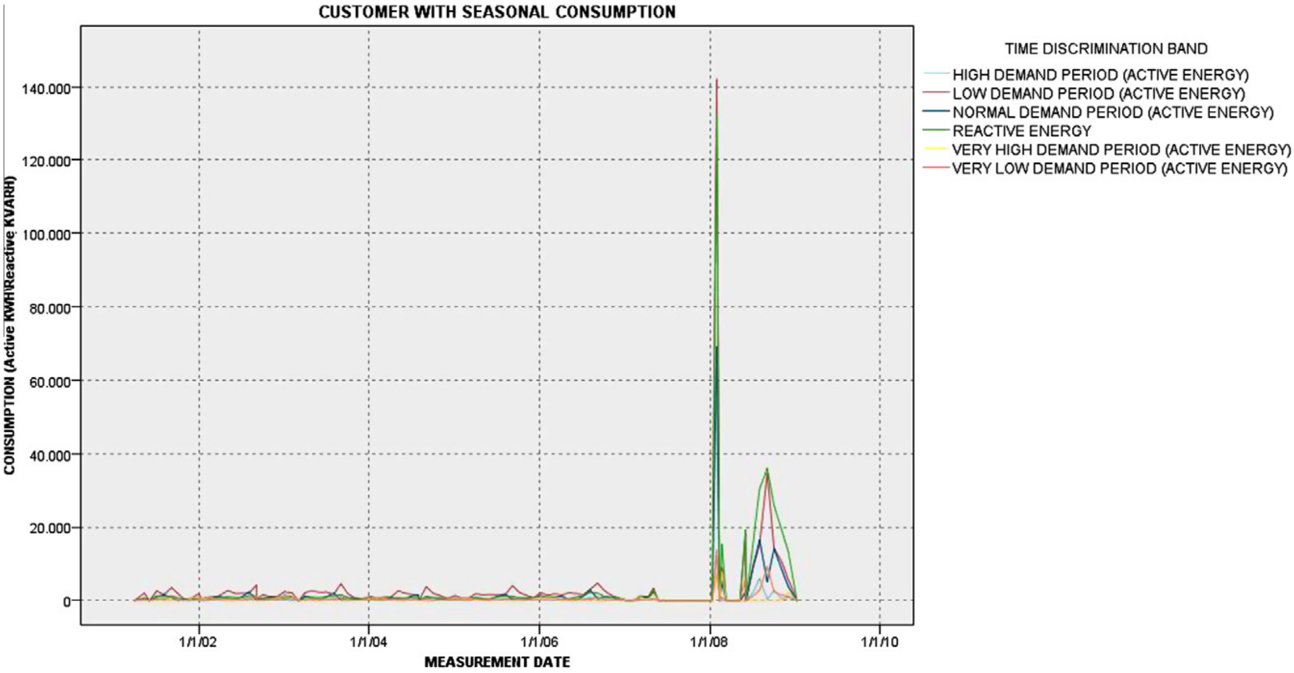

Fig. 9. Consumption graph of a seasonal case.

the measuring equipment and was successfully solved. The IES classified this as a non-NTL case.

The fourth case is a customer with zero consumption, as he had several contracts in the same place. The inspectors' commentaries specified that the supply was used as a supplement and for occasional use only. The IES classified this case as a non-NTL case.

10. Conclusions

The present paper includes the investigation of the expert system in the power utility consumption subject, and its combination with other technologies to further enhance the efficiency.

This paper particularly significantly contributes to an as-yet little exploited subject – the automatic analysis of available customer information of the utilities, on NTL classification. The main area of complexity of this research field was the enormous quantity of information required and the great variety of casuistry that is present in the customer analysis. As evident in the bibliographic review section, customer analysis is based on customer consumption. When the inspector's knowledge is included, new techniques need to be added to treat it.

The main contributions done by the IES in this paper are:

- Identification and classification of the casuistry of utility distribution on customer analysis.
- This system increases the availability of the quantity of useful information on the customer.
- Increasing of the efficiency, regarding massive inspections. The utilization of additional available information about the customer (in addition to customer consumption) helps to greatly increase the success ratio.
- *Classification of the normal and NTL cases.* The normal detection methods are dedicated only to select the anomalous cases: the present IES makes a classification of the normal and NTL cases. The proposed IES is a real framework, actually in the test phase, in Endesa.
- The proposed IES can be used as an additional method to detect NTL increasing the efficiency of the studies.

In reality, we researched new techniques and technologies to improve the efficiency of this IES. From this viewpoint, we are researching the use of the following techniques.

- *Text mining.* In the proposed IES, the inspector's knowledge in the categorization process is used. The customer's consumption will be added to this process to automatically make the categorization process. Thus, each rule created by text mining has several associated concepts, one associated action and several features related with consumption.
- *Data mining.* An additional improvement in curve consumption calculation will be researched. This new feature will include the time series and Bayesian networks with the use of events or intervention fields like special independent fields that will be used to model the effects of external occurrences.
- *Data warehousing.* The data mining and text mining techniques will be completed by employing data warehousing technology. This technology will improve the efficiency of the mining process.
- *Real-time analysis.* A new module to reduce the analysis time will be added. This module will determine when a customer will have new information available to analyze it.

References

Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications, 36*, 7764–7772.

Biscarri, F., Monedero, I., León, C., Guerrero, J., Biscarri, J., & Millán, R. (2008). A data mining method based on the variability of the customers consumption. In *10th international conference on enterprise information systems, ICEIS2008, June 12–16, Barcelona, Spain*.

Biscarri, F., Monedero, I., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2009). A mining framework to detect non-technical losses in power utilities. In *11th international conference on enterprise information systems, ICEIS2009, May 6–10, Milano, Italy*.

Bowen, J. E. (1994). An expert system for police investigators of economic crimes. *Expert Systems with Applications, 7*(2), 235–248.

Cabral, J. E., Pinto, P., Onofre, J., Gontijo, E. M., & Filho, J. R. (2004). Fraud detection in electrical energy consumers using rough sets. In *2004 IEEE international conference on systems, man and cybernetics* (Vol. 4, pp. 3625–3629).

Cabral, J. E., Pinto, J. O., Linares, K. S. C., & Pinto, A. M. A. (2006). Methodology for fraud detection using rough sets. In *2006 IEEE international conference on granular computing*. IEEE Press.

Chen, W.-S., & Du, Y.-K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications, 36*, 4075–4085.

Chen, S., Qu, G., & van Zuylen, H. (2010). A comparison of outlier detection algorithms for ITS data. *Expert Systems with Applications., 36*(8), 10976–10986.

Cierniakoski, J. J., De, R., & May, J. H. (1991). MEDIN: An expert system for processing medical insurance claims. *Expert Systems with Applications, 2*, 211–218.

Daskalaki, S., Kopanas, I., Goudara, M., & Avouris, N. (2003). Data mining for decision support on customer insolvency in telecommunication business. *European Journal of Operational Research, 145*, 239–255.

Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer network. *Expert Systems with Applications, 29*, 713–722.

Estévez, P. A., Held, C. M., & Pérez, C. A. (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications, 31*, 337–344.

Galván, J. R., Elices, A., Muñoz, A., Czernichow, T., & Sanz-Bobi, M. A. (1998). System for detection of abnormalities and fraud in customer consumption. In *12th conference on the electric power supply industry. November 2–6, Pattaya, Thailand*.

Gao, S., & Xu, D. (2009). Conceptual modelling and development of an intelligent agent-assisted decision support system for anti-money laundering. *Expert Systems with Applications, 36*, 1493–1504.

Hand, D. J., & Blunt, G. (2001). Prospecting for gems in credit card data. *IMA Journal of management Mathematics, 12*(2), 173–200.

He, H., Wang, J., Graco, W., & Haukins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications, 13*(4), 329–336.

Hernández-Pereira, E., Suárez-Romero, J. A., Fontenla-Romero, O., & Alonso-Betanzos, A. (2009). Conversion methods for symbolic features: A comparison applied to an intrusión detection problem. *Expert Systems with Applications, 36*, 10612–10617.

Hilas, C. S. (2009). Designing an expert system for fraud detection in private telecommunications networks. *Expert Systems with applications, 36*, 11559–11569.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, 847–856.

Kim, H. R., Im, H. K., & Park, S. C. (2010). DSS for computer security incident response applying CBR and collaborative response. *Expert Systems with Applications., 37*(1), 852–870.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*, 995–1003.

Liao, S.-H. (2005). Expert system methodologies and applications – A decade review from 1995 to 2004. *Expert Systems with Applications, 28*, 93–103.

Liao, S.-H., Hsieh, C.-L., & Huang, S.-P. (2008). Mining product maps for new product development. *Expert Systems with Applications, 34*, 50–62.

Liao, S.-H., & Wen, C.-H. (2007). Artificial neural networks classification and clustering of methodologies and applications – Literature analysis from 1995 to 2005. *Expert Systems with Applications, 32*, 1–11.

Métais, E. (2002). Enhancing information systems management with natural language processing techniques. *Data & Knowledge Engineering, 41*, 247–272.

Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications, 35*, 1721–1732.

Rahman, S., & Lauby, M. (1993). Identification of potential areas for the use of expert systems in power system planning. *Expert Systems and Applications, 6*, 203–212.

Richardson, R. (1997). Neural networks compared to statistical techniques. computational intelligence for financial engineering (CIFEr). In *Proceedings of the IEEE/IAFE 23–25 March 1997, pp. 89–95*.

Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J. M. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications, 36*, 3630–3640.

Schutzer, D. (1990). Business expert systems: the competitive edge. *Expert Systems with Applications, 1*, 17–21.

Sung, N. H., & Chang, Y. S. (2004). Business information extraction from semi-structured webpages. *Expert Systems with Applications, 26*, 575–582.

Wang, D., Wang, Q.-Y., Zhan, S.-Y., Li, F.-X., & Wang, D.-Z. (2004). A feature extraction method for fraud detection in mobile communication networks. In

Fifth world congress on intelligent control and automation, WCICA 2004 (Vol. 2, pp. 1853–1856).

Wang, S.-J., Mathew, A., Chen, Y., Xi, L.-F., Ma, L., & Lee, J. (2009). Empirical analysis of support vector machine ensemble classifiers. *Expert Systems with Applications, 36*, 6466–6476.

Wheeler, R., & Aitken, S. (2000). Multiple algorithms for fraud detection. *Knowledge-Based Systems, 13*(2–3), 93–99.

Yang, W.-S., & Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications, 31*, 56–58.

Yang, H.-C., & Lee, C.-H. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications, 27*, 645–663.

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert System with Applications., 37*(2), 1351–1360.