Detection of Non-Technical Losses:

The Project MIDAS

Juan I. Guerrero
Universidad de Sevilla, Spain

Íñigo Monedero
Universidad de Sevilla, Spain

Félix Biscarri *Universidad de Sevilla, Spain*

Jesús Biscarri *Universidad de Sevilla, Spain*

Rocío Millán
Universidad de Sevilla, Spain

Carlos León
Universidad de Sevilla, Spain

ABSTRACT

The MIDAS project began in 2006 as collaboration between Endesa, Sadiel, and the University of Seville. The objective of the MIDAS project is the detection of Non-Technical Losses (NTLs) on power utilities. The NTLs represent the non-billed energy due to faults or illegal manipulations in clients' facilities. Initially, research lines study the application of techniques of data mining and neural networks. After several researches, the studies are expanded to other research fields: expert systems, text mining, statistical techniques, pattern recognition, etc. These techniques have provided an automated system for detection of NTLs on company databases. This system is in the test phase, and it is applied in real cases in company databases.

INTRODUCTION

The main objective of data mining techniques is the evaluation of data sets to discover relationships in information. These relationships may identify anomalous patterns or patterns of frauds. Fraud detection is a very important problem in telecommunication, financial and utility companies. Currently data mining is one of the most important techniques which are applied to solve these types of problems, joined with: rough sets, neural networks, time series, support vector machines, etc. There are a lot of references about the detection of abnormalities or frauds in a set of data.

The increase of storage capacity and the process capacity allow one to manage large databases. Data mining provides a set of techniques of artificial

intelligence which can be used to increase the efficiency of data mining methods.

The utility companies have large databases which support the management processes. In addition, these companies invest their effort in maintenance of infrastructure and anomaly detection. These anomalies are frauds in telecommunication and financial sectors; breakdown or fraud in power, water or gas sectors; etc.

The non-technical losses (NTLs) in power utilities are defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment. This paper describes advances developed for the MIDAS project. The paper proposes a framework to analyze all information available about customers. This framework uses: data mining, text mining, expert systems, statistical techniques, regression techniques, etc. The proposed framework is actually in the testing phase. It is the main result of the MIDAS Project a collaborative project between the Endesa Company, Ayesa and the University of Seville.

In this paper, a description of the framework is made, following these steps:

- Review of current state about the anomaly detection and NTLs detection. Additionally, the Endesa utility company is described.
- The MIDAS project is explained.
- Each module is described.
- Finally, the conclusions are presented.

REVIEW OF THE CURRENT STATE

Bibliographical Review

The Non-Technical Losses (NTLs) were increasingly regarded as a cause of concern in distribution utility companies. There exists several causes of NTLs and they can affect quality of supply, electrical load on the generating station and tariff imposed on electricity consumed by genuine customers. (Depuru, Lingfeng Wang, Devabhaktuni, & Gudi, 2010) discusses various factors those influence the consumer to make an attempt to steal electricity. There are a lot of methods for detection NTL. The distribution utility companies are interested in analysis of NTLs for detection, location and classification of NTLs, with the objective of reducing them. There are many ways to perform these processes, and can be taken as reference similar methods used for fraud detection in telecommunications, finance, etc. (Yufeng Kou, Chang-Tien Lu, Sirwongwattana, & Yo-Ping Huang, 2004) and (Weatherford, 2002) show different techniques related with data mining for fraud detection, including the most interesting parameters for using with them.

Financial Sector

In the financial sector there are a lot of references with the use of data mining and computational intelligence in the fraud detection. Noteworthy is the use of these techniques in credit card fraud detection. In the nineties, (Ghosh & Reilly, 1994), (Fanning, Cogger, & Srivastava, 1995), (Aleskerov, Freisleben, & Rao, 1997) and (Dorronsoro, Ginel, Sgnchez, & Cruz, 1997) used neural networks to detect ones. Some authors publish researches with other techniques: intelligent hybrid system (Hambaba, 1996), neural networks compared to statistical techniques (Richardson, 1997), neural data mining (Brause, Langsdorf, & Hepp, 1999), distributed data mining (Chan, Fan, Prodromidis, & Stolfo, 1999), etc. In addition, other techniques are used latter, for example: Genetic Algorithm (Özçelik, Işik, Duman, & Çevik, 2010), neural networks and logistic regression (Y. Sahin & Duman, 2011), time series (Seyedhossein & Hashemi, 2010), decision trees (Yusuf Sahin, Bulkan, & Duman, 2013), etc.

In financial sector, there are other areas of interest, for example, based in the theory of Rough Sets (Yezheng Liu, Yuanchun Jiang, & Wenlong

Lin, 2006), (Qian Liu, Tong Li, & Wei Xu, 2009). However, the most often used techniques are neural networks and data mining (Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, & Chao-Hsien Chu, 2007).

Communication Sector

In the communication or telecommunication sector, there is a growing interest in fraud detection. In this sector, several techniques are used, from basic data mining techniques, as feature extraction (Wang Dong, Wang Quan-yu, Zhan Shou-yi, Li Feng-xia, & Wang Da-zhen, 2004), or neural networks (Mohamed et al., 2009) and probabilistic methods (Taniguchi, Haft, Hollmen, & Tresp, 1998), to fuzzy rough sets (Wei Xu et al., 2008) or knowledge-based management (Davis & Goyal, 1992). In fact, this interest has been extended to VoIP (Rebahi, Nassar, Magedanz, & Festor, 2011) and (Seo, Lee, & Nuwere, 2013).

Intrusion Detection

In this subject, there are a lot of techniques which is used as a part of an Intrusion Detection System. Noteworthy is the use of feature analysis (Shuyuan Jin, Daniel So Yeung, Xizhao Wang, & Tsang, 2005), Support Vector Machines or SVM (Liu Wu, Ren Ping, Liu Ke, & Duan Hai-xin, 2011), neural networks (Raghunath & Mahadeo, 2008) and (Lei & Ghorbani, 2012) and several data mining techniques (Ming Xue & Changjun Zhu, 2009) and (Li Han, 2010). This new field of research becomes important due to the proliferation of computer networks and Internet. These facts are important in other research fields, for example Smart Grids. The security of Smart Grids is a very interesting problem and is very related with intrusion detection, because Smart Grids are based in the mix of telecommunication with utility network.

Non-Technical Losses

There are some references about NTLs detection, but wide scope of techniques is used. There are several references about computational intelligence based solutions, but, sometimes, initially the companies start with a strategic plan (Gonzalez & Figueroa, 2006) or implant new follow-up and control mechanisms (Iglesias, 2006), (Mwaura, 2012) of NTL reduction. This paper is oriented in the use of computational intelligence. In this sense, a lot of references can be found, and they will be showed below.

(Nizar, Zhao Yang Dong, & Pei Zhang, 2008) uses detection rules for non-technical analysis, the technique compounds a series of data mining tasks, including feature selection, clustering and classification techniques. (Cabral, Pinto, Martins, & Pinto, 2008) and (Cabral, Pinto, & Pinto, 2009) propose a methodology based on a non-supervised artificial neural network called SOM (Self-Organizing Maps) which is robust on several cases. (Nizar, Dong, Zhao, & Zhang, 2007) proposes two popular classification algorithms, Naïve Bayesian an Decision Tree, extracting the patterns of customers' consumption behavior from historical data and arranging the data in various ways by averaging them yearly, monthly, weekly, and daily. Both techniques are used and compared. Various authors use the Support Vector Machines (SVM) method related with non-technical losses detection. For example: (Aranha Neto & Coelho, 2013), (Depuru, Lingfeng Wang, & Devabhaktuni, 2011), (Nagi, Yap, Sieh Kiong Tiong, Ahmed, & Mohamad, 2010) and (Nagi, Mohammad, Yap, Tiong, & Ahmed, 2008). Each one proposes different ways to make analysis or detection of non-technical losses. In addition, this technique can be combined with any other computational intelligence, for example, with fuzzy inference system (FIS) (Nagi, Keem Siah Yap, Sieh Kiong Tiong, Ahmed, & Nagi, 2011) or with genetic algorithm (GA) (Nagi, Yap, Tiong, Ahmed, & Mohammad,

2008). The use of these mixed methods improves the efficiency of SVM technique.

(Brun, Pinto, Pinto, Sauer, & Colman, 2009) proposes the use of deferential evolution algorithm to find the parameters of a data mining system used to pre-select electrical energy consumers with suspect of fraud, building a pattern recognition system for suspicious behavior detection. The parameter of the pattern recognition system must be well tuned, and that can be modeled as an optimization problem using the available training data.

(Markoc, Hlupic, & Basch, 2011) proposes a new approach based on neural network trained by generated samples.

(Cabral & Gontijo, 2004) describes an application of rough sets in the fraud detection of electrical energy consumers. Rough set is an emergent technique of soft computing that have been used in many knowledge discovery in database applications. From an information system, rough sets concept of reduce was used to reduce the number of conditional attributes and the minimal decision algorithm was used to reduce some values of conditional attributes. The reduced information system derives a set of rules that reaches consumers behavior, allowing the classification rule system to predict many fraud consumer profiles.

(Caio C. O Ramos, Papa, Souza, Chiachia, & Falcao, 2011) shows the importance of using feature selection in non-technical losses detection, in fact, there are several authors which use this technique (Nizar, Jun Hua Zhao, & Zhao Yang Dong, 2006). Also in [44], a characterization of customer using evolutionary-based feature selection method.

(de Oliveira, Boson, & Padilha-Feltrin, 2008) proposes a statistical analysis of relationship between load factor and loss factor using the curves of a sampling of consumers in a specific company, these curves are summarized in different bands of coefficient k. Then, it is possible determine where each group of consumer has its major concentration of points.

Even, there exist computational techniques specifically developed for NTL detection, for example, (dos Angelos, Saavedra, Cortés, & de Souza, 2011) proposes a computational technique for the classification of electricity consumption profiles based on fuzzy clustering and Euclidean distance.

Other references are based on load profiling calculation. These are additional point of view, they use remote management systems and smart metering. These systems provide more information about client consumption. (Nizar, Dong, & Zhao, 2006) and (Nizar, Dong, Jalaluddin, & Raffles, 2006) propose a study for detection the best load profiling methods and data mining techniques to classify, detect and predict NTLs in the distribution sector, due to faulty metering and billing errors, as well as to gather knowledge on customer behavior and preferences so as to gain a competitive advantage in the deregulated market. (Nizar & Dong, 2009) and (Nizar, Dong, & Wang, 2008) propose Extreme Learning Machine (ELM) and online sequential-ELM (OS-ELM) algorithms which are used to achieve an improved classification performance and to increase accuracy of results. A comparison of this approach with other classification techniques, such as the SVM algorithm, is also showed. (C. C.O Ramos, Souza, Papa, & Falcao, 2009) and (C. C.O Ramos, de Sousa, Papa, & Falcão, 2011) propose Optimum Path Forest (OPF) classifier for a fast non-technical losses recognition, they show a comparison with neural networks and SVM, getting best results with OPF than neural networks and similar results than SVM. However, (Depuru, Lingfeng Wang, Devabhaktuni, & Nelapati, 2011) proposes a hybrid neural network, which implements a neural network model and suggests a hierarchical model for enhanced estimation of the classification efficiency. In addition, this paper proposes and encoding a new technique that can identify illegal consumers. (Yi Zhang, Weiwei Chen, & Black, 2011) presents a method to accurately identify anomalous days for individual premises so that they can be removed from the premise data.

The new technologies related with Smart Grids, provides more information and control over the consumption and demand. Smart Grids provides new technologies to improve the reduction of NTL (Abaide, Canha, Barin, & Cassel, 2010). For example, smart metering (Openshaw, 2008) or Hall Effect based electrical energy metering devices (Wilks, 1990). In this sense, there are some references which are based on increasing of metering infrastructure capabilities. (Alves, Casanova, Quirogas, Ravelo, & Gimenez, 2006), (Kerk, 2005) and (Nagi, Yap, Nagi, et al., 2010) propose an Advanced Metering Infrastructure based on remote management of equipment, provides more information about consumption and events which could happen in consumer installation. These new information compounds: more information about consumption (quarter-hourly), information about events (illegal manipulation, inspector operations, alarms, etc.), etc. Some references also use this information with computational intelligence techniques, for example, (Depuru, Lingfeng Wang, & Devabhaktuni, 2011) proposes a SVM which take information from smart metering infrastructure.

Normally, works that use more information have best results. The references which use load profiling or data from smart metering have more information about consumer and optimal pattern consumption can be established. Currently, the companies have a lot of clients. Smart Metering provides a new research field, but it is necessary to establish the methods and techniques of NTL detection in scenarios when limited information in consumption are available. For example, there exist a lot of clients with monthly measurements, because they don't have smart meters. Although there exists smart metering infrastructure, the company databases have a lot of additional information about client, which could be used for NTL detection. It is necessary to determine methods for analysis of present and future situations (Gemignani, Tahan, Oliveira, & Zamora, 2009), which compounds monthly and quarter-hourly periods, taking advantage of the rest of company databases information. A complete framework for NTL detection is proposed in this paper. This framework is based on data mining, statistical techniques, text mining, neural network and expert system, which gather all information about client to get a classification of the client, according to the problem which the client's facility shows. Most references specified above are based on consumption, contracted power, tariff, economic sector and geographic location. In the corporate databases much more information exists, for example, results of inspections, inspectors' feedback information, etc. The proposed framework takes advantage of all information stored in corporate databases.

Do Companies Fight Against NTLs

The system proposed in this paper is actually operating in the testing phase in the Endesa Company. Endesa is the most important Spanish energy distribution company with more than 12 million clients in Spain, and more than 73 million clients in European and South American markets.

Traditionally, companies identify two different types of losses: non-technical losses (NTLs) and technical losses. The NTLs are caused by breakdown or illegal manipulation in customer facilities. These types of losses are very difficult to predict. Normally, utility companies use massive inspection to reduce NTLs. These inspections are performed on the customer who carries out a series of conditions, as example: customers who have measure equipment without transformers and it is located in a limited geographic zone. These conditions reduce the volume of number of customers to inspect. Utility companies are very interested in the detection of NTLs.

The Technical Losses represents the rest of the losses which is produced by distribution problems (Joule effect). The Technical Losses can be forecasted because they are approximate constants, but the NTLs are very irregular and very difficult to forecast. The technical losses are caused by faults in distribution lines. These faults are predictable with a low rate of error.

When the inspector finds an NTL, the company has to be notified. The inspector stores all information about the problem when it is detected until it is solved. This information is named proceeding.

THE PROJECT MIDAS

The objective of the MIDAS Project is the detection of Non-Technical Losses (NTLs) using computational intelligence over Endesa databases. This project is the collaboration between Endesa, Ayesa Tecnología and Electronic Technology Department of University of Seville. This project began at 2006 with the study of a little set of customers, and getting good results.

In this project a lot of lines are researched: data mining, statistical techniques, neural networks, expert systems, text mining, pattern recognition, etc.

Traditionally, the utility companies used massive inspections to avoid the NTLs, but this method is very expensive both in time and in money. Currently the utility companies use more advanced systems that allow the selection of clients who carry out some simple conditions. This type of system allows one to reduce the economic and time cost, increasing the efficiency. But these simple conditions aren't automatically selected and, normally, they only detect some type of NTLs.

As it is said, the prototype developed is in test stage and is tested with Endesa databases. This system has provided better results than the traditional system of inspection.

SYSTEM ARCHITECTURE

The proposed system architecture contains several modules. Each module is implemented with different techniques. Each module can increase their capabilities with each new prototype, using the previous results to make better modules. Each prototype is tested with real data of Endesa databases and it is validated with inspections made by Endesa staff.

The system architecture is shown in Figure 1. In this architecture the different steps of the process are applied in an ordered way. In the first place, a sample of customers is selected using the data stored in utility company databases. In the second place, several artificial intelligence and statistical techniques are applied. The regression analysis techniques and data mining modules provide a set of customers whose have some anomalies regarding customer's own consumption or the other customers' consumption. Mainly, these modules work with some parameters: consumption, contracted power, economic activity. In the last place, the integrated expert system analyzes the rest of information about the customer. The integrated expert system provides the results on databases and reports that they can be both used by inspectors as an additional source of information. In the following sections each of these modules is described.

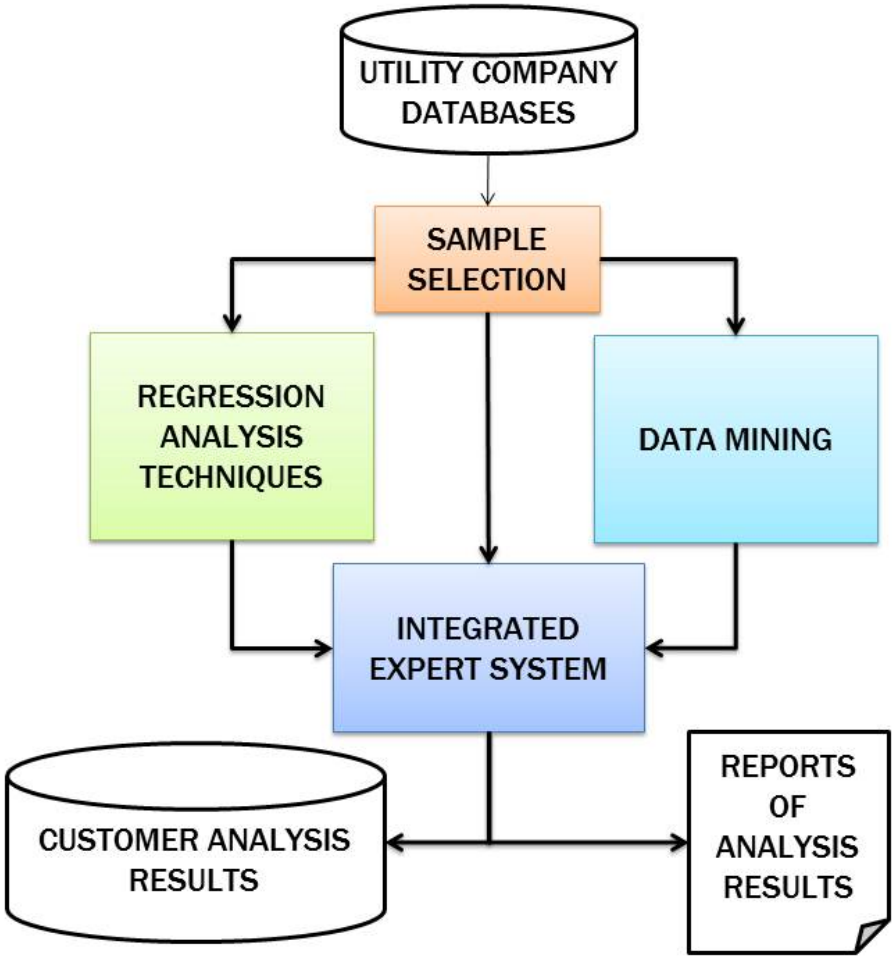
SELECT SAMPLE

The sample selection uses several data sources. Each data source provides information about different aspects about the customers:

- **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 localization**
- **Economic Activity:** Some economic sectors historically present a high rate of NTLs.
- **Consumption range:** Sometimes, the consumption range can be used to restrict the quantity of customers.
- **Electricity charges**

These parameters allow restricting the quantity of customers to analyze. The information of each customer compounds information about: contract,

Figure 1. System architecture



installed equipment, results of inspections realized over the facilities, etc.

The information about consumption is analyzed in different ways by each module. Additionally the rest of information is analyzed by integrated expert system.

DATA MINING

First of all, and as previous step to the data mining, two processes from which taken the data all the detection methods was carried out:

- **Data Selection:** In this step, a set of the sample to process was selected by the proposed data mining techniques.
- **Data Preprocessing:** A pre-processing of the data, during which data was carried out sets were prepared (generating new tables, filtering wrong data, etc) for the mining process.

Mainly, the objective of these processes was to normalize and to discrete the sample set for the set of data mining models developed in the project.

Initially, the first techniques developed for the data mining process were the outliers' analysis

and inherent data variability. These techniques are described in (Biscarri et al., 2008). For this process a sample of homogeneous data which have utility customers with similar characteristics were selected. The temporary and the local components of the individual consumption of customer were removed by means of normalization. After this step, the probability distribution of the transformed sample, for the normal operating condition, as Gaussian is considered. The threshold of the sample variance is calculated and adjusted. Finally, the detected outliers are used to guide the inspections. This process would be fired in the first step in the framework of data mining. The resulting customers (not detected by this process) would be those processed by the other data mining techniques.

Thus, after the development of these methods, the inclusion of other techniques of data mining was carried out. In this way, the framework of MIDAS integrates several methods related to different data mining techniques. These methods were classified in different modules depending on the type of technique used for detection, in purely statistical or evolved models (referring to association and segmentation). All of them allow increasing the efficiency of detection process by means of complementary detections using the same information.

Statistical Methods Based on Comparison with Similar Customers

These three statistical techniques are used in this module: one based on the variability of customer consumption, another based on the consumption trend and a third one that summarizes other feature contributions of NTL detection. This module was described in (Biscarri, Monedero, León, Guerrero, & Biscarri, 2009).

The variability analysis provides an algorithm that emphasizes customers with a high variability of monthly consumption in comparison to other customers of similar characteristics. The classic approach to the study of variability classifies data in 'normal data' and outliers. The proposed variability analysis uses the standard deviation estimation (STD) to associate to each customer a new feature that will be used as an input for a supervised detection method, showed in the Predictive data mining section.

The consumption trend uses a streak-based algorithm. Streaks of past outcomes (or measurements) are one source of information for a decision maker trying to predict the next outcome (or measurement) in the series.

This set of techniques is strongly dependent of the cluster of customers considered and highly changeable amongst different clusters. The study of the individual trend consumption and also the comparative among trends of customer with similar characteristics is very interesting and it contributed to detect NTLs relative to customers with a different behavior to their similar environment.

Statistical Methods Based on an Individual Analysis of the Customers

This method is used to identify the customers with pattern of drastic drop of consumption. It is because according to the Endesa inspectors and the studies of consumption, the main symptom of a NTL is a drop in billed energy of the customers. The detection methods referring to this module are described in (Hutchison et al., 2010).

This method compounds several algorithms: based on regression analysis, based on the Pearson correlation coefficient and based on a windowed linear regression analysis. These algorithms are based on a regression analysis on the evolution of the consumption of the customer. The aim is to search for a strong correlation between the time (in monthly periods) and the consumption of the customer. The regression analysis makes it possible to adjust the consumption pattern of the customer by means of a line with a slope. This slope must be indicative of the speed of the drop of the consumption and, therefore, the degree of

correlation. These algorithms identify with a high grade of accuracy two types of suspicious (and typically corresponding to NTL) drops.

Thus, this module provided a set of effective and robust techniques to detect cases with a particular manifestation of the NTLs (consumption drops).

Evolved Models Based on Clustering and Decision Trees

In this type of detections, as well as those ones with association rules, the objective was to find new NTLs searching for customers with similar behavior to those NTL detected in the past. For it, first of all, we featured the type of contract and consumption pattern of the customer. After featuring the customers, the different techniques search for similar customers in the sample set. The feature vector included the following patterns:

- Number of hours of maximum power consumption.
- Standard deviation of the monthly or bimonthly consumption.
- Maximum and minimum value of the monthly or bimonthly consumptions.
- Reactive/Active energy coefficient.
- The number of valid consumption lectures. Usually, when there is not a valid lecture value and the company is sure that consumption existed, the consumption is estimated and billed.

In addition, two parameters are added to this set. These parameters were those ones based on the concept of streak (and generated in the work described in [63]).

Concretely, the use of these two techniques (clustering and decision trees) is described in (Monedero et al., 2009). Thus, this work included a process of generation of clusters by means of the K-Means algorithm and, in parallel, an algorithm that generates decision trees.

Both algorithms carry out a clustering of the customers (one by clusters and other by branches) and those clusters with a higher rate of NTLs identified by Endesa Company in the past were studied. Two techniques with the objective of that both ones searched the same thing by two complementary ways were used.

Evolved Models Based on Association Rules

The module uses an inference of a rule set to characterize each of two following classes: 'normal' or 'anomalous' customer (depending if there was been detected as NTL in the past by Endesa Company in its inspections). Each customer is characterized by means of the attributes previously described. The association uses supervised learning that by means of a set of input attributes search of NTLs. This module is described in (Biscarri et al., 2009).

In particular, the algorithm uses the Generalize Rule Induction (GRI) model. It discovers association rules in the data.

The test of the set of rules generated four values, according to the following classifications (Cabral et al., 2008):

- **True positives (TP):** Quantity of test registers correctly classified as fraudulent.
- **False positives (FP):** Quantity of test registers falsely classified as fraudulent.
- **True negatives (TN):** Quantity of test registers correctly classified as non-fraudulent.
- **False negatives (FN):** Quantity of test registers falsely classified as correct.

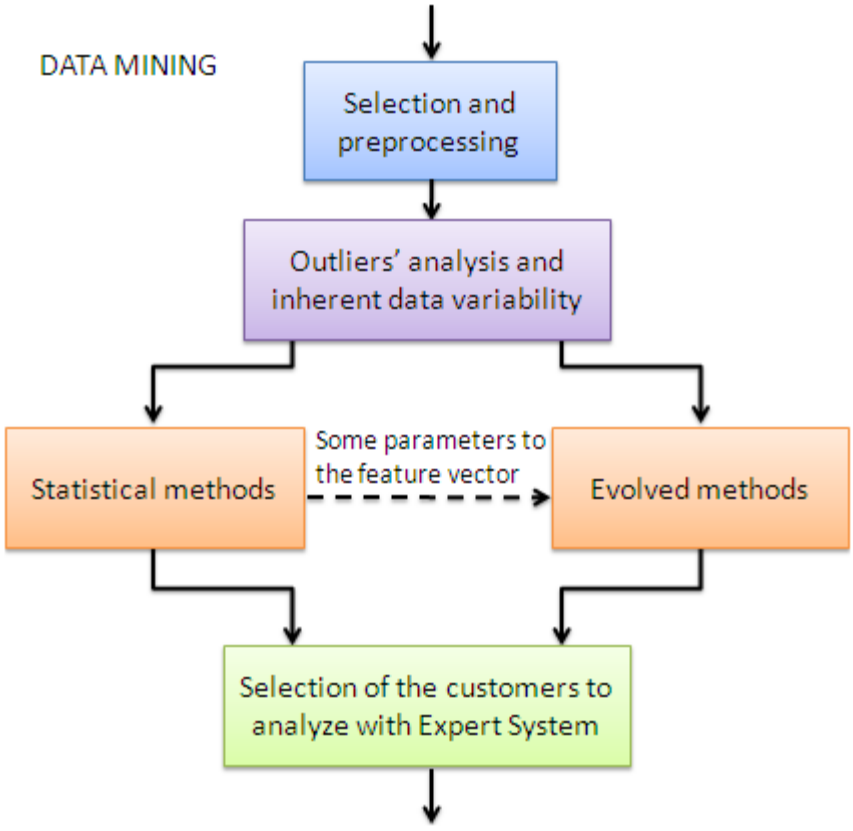
A decision tree algorithm GRI extracts rules with the highest information content based on an index that takes both the generality (support) and accuracy (confidence) of rules into account. GRI can handle numeric and categorical inputs, but the target must be categorical.

The objective that we searched with association rules was the same one that with clustering and decision trees: Detecting new NTLs identifying a similar pattern to those ones detected in the past. In contrast, the advantage is that the association rule algorithm over the more standard decision tree algorithms is that associations can exist between any of the attributes. This made it possible to detect different customers to those ones detected with the methods of previous sections and therefore, to do new complementary detections.

The detections carried out by each of these modules were later analyzed by the Integrated Expert System (it is described in C section) in order to perform a deeper study (with other parameters not only related to consumption pattern).

Summarizing the overall data mining process, Figure 2 shows the framework of this process.

Figure 2. Flow chart of the data mining process



INTEGRATED EXPERT SYSTEM

The Integrated Expert System has a core based on a Rule Based Expert System (RBES). Although this RBES was described in (León et al., 2011), this paper presents new advances in some modules and it is used as part of a complete framework. This system uses the information extracted from Endesa staff and inspectors. The RBES has several additional modules which provide dynamic knowledge using rules. The expert system has additional modules which uses different techniques: data warehousing (it is used as a preprocessing step), text mining, statistical techniques and neural networks.

In the proposed framework, the RBES may be used as additional methods to analyze the rest of information about the customer. The company

databases store a lot of information, including: contract, customers' facilities, inspectors' commentaries, customers, etc. All of them are analyzed by RBES using the rules extracted from Endesa staff, inspectors and rules from the statistical techniques and text mining modules. This information is not analyzed by the previously described modules. The system provides the point of view of the Endesa staff and inspectors.

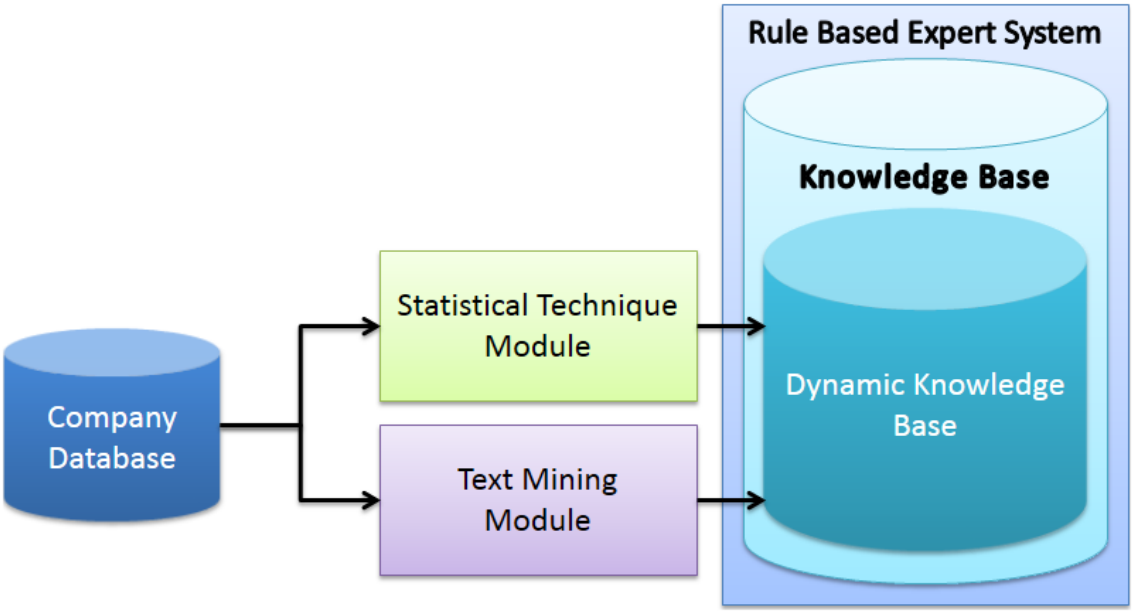
Integrated Expert System Architecture and Application

The integrated expert system is compounds of several modules; each module has an information type as objective:

- **Statistical techniques module.** This module is used to make patterns of correct range and trend of consumption. This module generates a series of values which they are used in dynamic rules of a dynamic knowledge base.
- **Text mining and neural network module.** This module is used to treat the information provided by inspectors' commentaries. This module generates a series of dictionaries with characterized concepts which they are used in dynamic rules of a dynamic knowledge base.
- **Integrated expert system.** This module uses the rules (extracted from inspectors and staff of Endesa, generated by statistical techniques module and generated by text mining and neural network module) for analyzing the clients.

The integrated expert system is applied in two steps. The first step or learning step the modules of statistical techniques and text mining and neural network are applied in a database of clients as large as possible. This fact allows make a reference for dynamic knowledge base. In Figure 3, this step is shown. This step is performed only once per month in case of statistical module or once per year in case of text mining and neural network module.

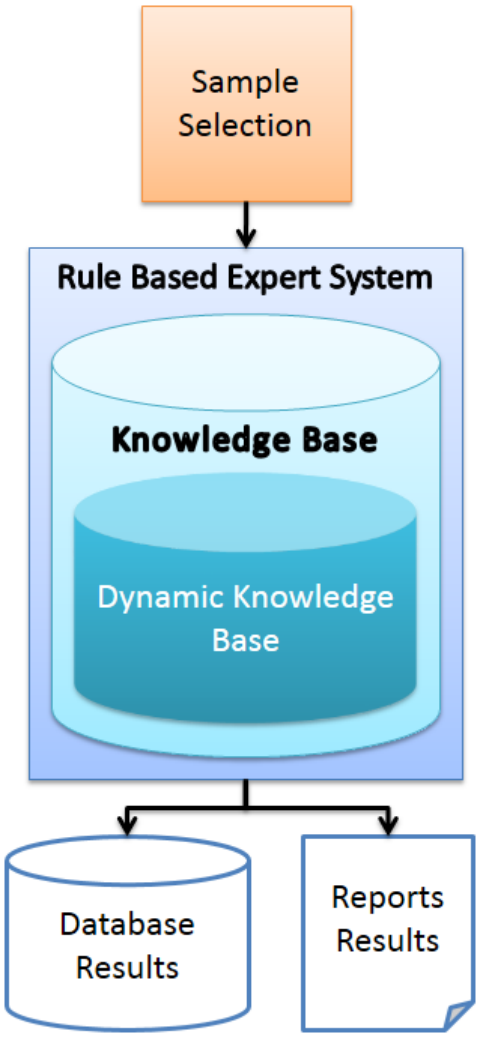
Figure 3. First step or learning step of application process of integrated expert system



In the second step or analysis step, the integrated expert system uses the results of the first step or learning step to make the analysis of the information about the clients. In Figure 4 this step is shown. The analysis is applied over sample selection; this sample may compound the clients selected by other type of analysis methods that they showed in previous sections.

The statistical techniques and text mining and neural network modules are described in the following sections.

Figure 4. Second step or analysis step of application process of integrated expert system



Statistical Technique Module

The statistical techniques are based in basic consumption indicators such as: maximum, minimum, average and standard deviation of consumption. These indicators are used as patterns to detect correct consumption. Additionally, the slope of regression line is used to detect the regular consumption trend. Each of these techniques is made for different sets of characteristics. These characteristics are: time, contracted power, measure frequency, geographical location, postal code, economic activity and time discrimination band. Using these characteristics it is possible to determine the patterns of correct consumption of a customer with a certain contracted power, geographic location and economic activity. These groups are described in Table 1.

It is necessary a learning step in which a lot of clients are used to apply all statistical calculations. In this study all customers are not used because the anomalous consumption of the customers with an NTL is filtered. This idea allows the elimination the anomalous consumption getting better results. This step is made before the client analysis, because this process provides the correct reference of consumption which it is stored in dynamic knowledge base.

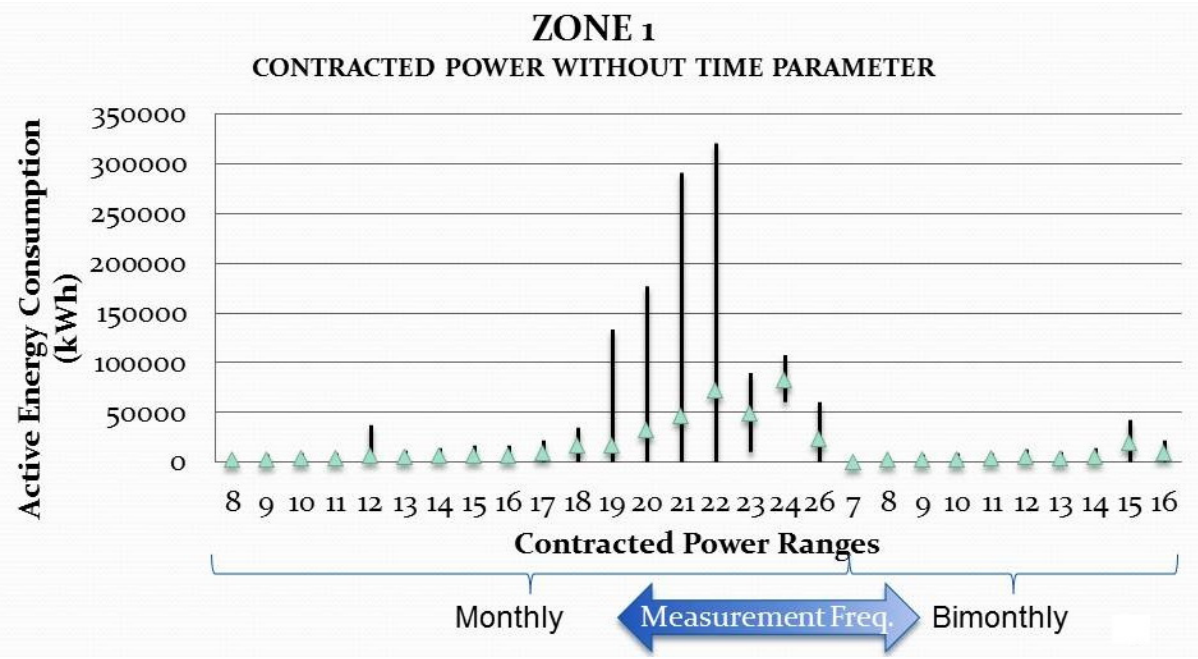
Several tables of data are generated as a result of this study. These data are used to create rules which implement the detected patterns. If a customer carries out the pattern, this means that the customer is correct. But if a customer does not carry out the pattern, this does not mean that the customer could be correct.

In Figure 5 only contracted power, zone and measurement frequency are used, this case there are several intervals, in 23 and 24 ranges, in which the correct consumption limits are defined. Each of these ranges represents identifiers of intervals, for example, the24 range represents the customers with the interval of contracted power between 366 kW to 455 kW. These intervals can be obtained for others contracted power ranges if more

Table 1. Groups of consumption characteristics

Consumption Characteristics	Description
Basis Group (A)	This group provides consumption patterns by general geographical location: north, south, etc.
Basis Group and Postal Code (B)	This group provides patterns useful for cities with coastal and interior zones.
Basis Group and Economic activity (C)	The granularity of geographical location is decreased, in this way, the economic activity takes more importance, but the geographical location cannot be eliminated, because, as for example, a bar has not the same consumption whether is in interior location or coastline location.
Basis Group and time discrimination band (D)	There are several time discrimination bands each band register the consumption at different time ranges. This group provides consumption patterns in different time discrimination bands, because there exists customers who makes their consumption in day or night time.

Figure 5. Ranges of correct consumption for a group without time parameter (Group A according to Table 1)



characteristics are added, for example, in Figure 6, the intervals are more specific and provide patterns for correct consumption. Additionally, in Figure 7, the intervals are more specific if the year characteristic is added.

This module is applied in samples as large as possible, because it is necessary to get a better statistical reference. This learning process only is made monthly or bimonthly.

The information generated by statistical techniques module is stored in a database and it is automatically translated to dynamic knowledge base of the integrated expert system, using several classes according to the group of consumption characteristics.

Text Mining and Neural Network Module

The text mining method is based on Natural Language Processing (NLP). The neural network is based on a multi-layer neural network. This method is used to provide a method to analyze the inspectors' commentaries. When an inspection in customer's location is made, the inspector has to register their observations and commentaries. This data is stored in company databases.

This information is not commonly analyzed, because the traditional models are based on consumption study. This module uses the rest of important information, because the inspectors' commentaries provide real information about the client facilities, which may be different from the stored in database.

The process begins with the application of text mining method. This technique uses NLP and fuzzy algorithms to extract concepts from inspectors' commentaries. Concept is a word or group of words which has own meaning. In addition, the fuzzy algorithms and the utilization of synonyms' dictionaries allow the interpretation of different language and dialects. These concepts are classified initially according to their frequency of appearance. The more frequent concepts are classified manually according to their meaning. Additionally, consumption indicators, date of commentary, number of measures (estimated and real), number of proceedings, source of commentary, frequency of appearance, time discrimination band and some others are associated to each concept. This data is used in a neural network, which is trained with data of the more frequent concepts and is tested with the concepts which were not classified manually. In the test process, the correct classification of concepts was verified manually. This neural network can be used to classify the new concepts which could appear.

The neural network is trained by means of a multiple method. This method creates several neural networks of different topologies. At the end of training, the model with the lowest Root Mean Square or RMS error is presented as the final neural network. The trained neural network assigns an importance value to each feature. SoftMax transfer function was used as a punctuation method. The trained neural network has two hidden layers and its structure is 22-28-26-4, due to the quantity of inputs. The first and last layers are the input and output layers, respectively.

Initially, the utilization of neural networks in non-structured language learning could seem very complex, not only the synonyms' dictionaries and the concepts' characterization make easy the process, but the inspectors have to make a lot of inspections in a day, and they must to store all information and commentaries in the company database. Due to this they use a very brief and concrete language.

This process generates dictionaries with characterized concepts. These dictionaries are stored in a database and it is automatically translated to dynamic knowledge base of the integrated expert system, using a class which allows to analysis the commentaries.

The first version of text proposed text mining and neural network module is described in (Guerrero et al., 2010).

Figure 6. Ranges of correct consumption for zone 2, several contracted power range, measurement frequency (monthly) and service sector. The grasfts show the limits of consumption for different contracted power ranges without the influence of time parameter (Group C according to Table 1)

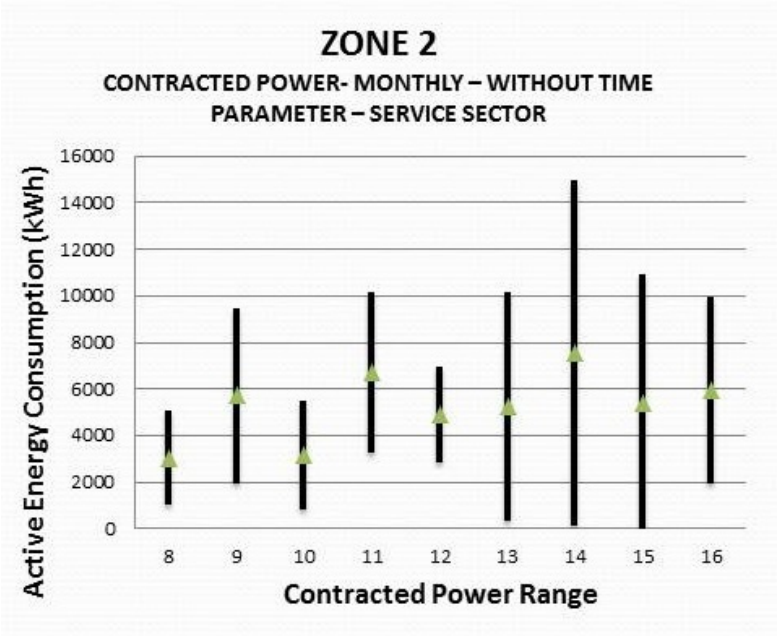
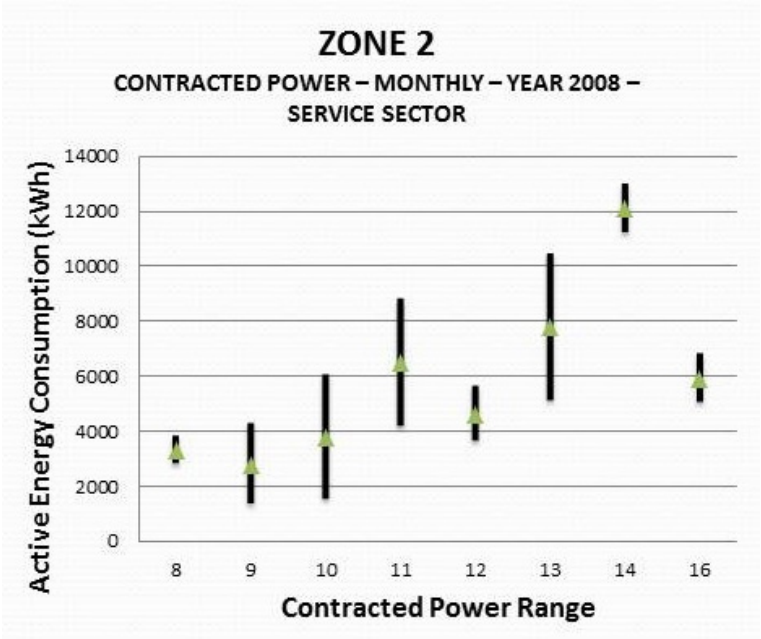


Figure 7. Ranges of correct consumption for zone 2, several contracted power range, measurement frequency (monthly) and service sector. The grasfts show the limits of consumption for different contracted power ranges with the influence of time parameter (Group C according to Table 1)



EXPERIMENTAL RESULTS

Several studies are made over several real cases, testing the efficiency and accuracy of the proposed framework. In this paper, only the last study is showed, because is the only one in which the framework is completely tested.

The sample compounds clients which carried out the following conditions:

- Contracted Power greater than 15 kW.
- Clients of north of Spain.

Although, the study compounds consumers with several economic sectors, the inspectors' experience and the results of other studies showed that the service sector is the one which have greater number of NTL cases.

The experimental results are described using architecture described in Figure 1. The first step of the process was Select Sample. In this step 540 consumers was selected, using different consumers' characteristics. In this step, only two possible groups are considered, the customers getting by data mining and regression analysis techniques. 119 customers were selected by the system and were inspected. Henceforth this group was referred as Group S1. Really, in these methods is possible to add some factors which determine probability of NTL. The quantity of selected customers was defined by utility company according to this factor.

The Expert System module can be used alone or joined with other methods. In case of the use of Expert System alone, the select sample step can include more customers, in this case, 81385 customers were selected for analysis in expert system. After analysis of customers, the number of selected customers with any detected problem was 3215. Henceforth, this group was referred as Group S2. At the same time, the Expert System determined that 63411 customers were not presented any NTL. The rest of customers, 14759, did not have enough information, because:

- They did not have enough information.
- They had few measurements.
- They were recently inspected.

The Group S1 is included in Group S2. Finally, Group S1 was selected by company and was send to inspectors. There are cases in which the expert system rule selected customers by the methods of data mining and regression analysis techniques, in this case, it is discarded if the customer is classified as correct by the text mining rules of the Integrated Expert System.

After inspections, the following results are obtained:

- Inspections could not be performed: 28 customers. These inspections could not be performed for various reasons, usually because it is indoors or with customers whose refuse, therefore, cannot ensure the existence of an NTL.
- Without NTL: 54 customers.
- With NTL: 37 customers.

In this way, the correctness was 40,66%. This result is better than massive inspections campaigns carried out by companies. Furthermore, the time spent in the analysis is 90% lower than traditional techniques.

The framework provided reports about analysis about each customer, in which it is possible to check:

- Problems or incidents found in the facilities or consumption of customer. These problems can mask an NTL.
- The method and rules used in analysis of each client.
- The conclusions of analysis.
- Statistical information about analysis process.

Highlight Cases

The proposed framework has been more efficient in analysis. There are some cases which traditionally were very difficult to detect. Concretely, two cases are treated in this section.

The first case is a client with an irrigation activity. The consumption of this type of client is strongly influenced by climate. The consumption of this client is very irregular, and difficult to analyze. These clients decrease their consumption when rainfalls increase. In this system, data about climate are not available, and only use the information about client. Sometimes, variations of climate conditions make that the data mining or regression analysis techniques select this type of clients. This client is analyzed by expert system, and normally it is dismissed according to the elapsed time since the last inspection.

The second case is the client with seasonal consumption. This type of clients is very difficult to detect with traditionally methods. The consumption of these clients shows one or two great peaks, which can be classified as a fraud. This type of clients can be hotels in coast line, which only has consumption in month with a good climate or in holiday periods. The using of descriptive data mining and expert system allows detecting these cases.

FUTURE RESEARCH DIRECTIONS

Future research fields are addressed to improve the knowledge about non-technical losses and extract knowledge from companies' databases. These objectives are translated in different research areas:

- Researching on the knowledge of other inspectors, and trying to extend the possibilities of detection.
- Application of the techniques and models developed in other utility companies, not only power distribution but also gas and water distribution.
- Adaptation to new technologies, such as smart metering or smart grids environments.
- Researching on the optimization of inspection routes.
- Improving these results with other techniques of data mining, computational intelligence and statistical inference.
- Testing with larger sample size and greater number of supplies to inspect.
- Researching on clients of medium and low voltage. This area is very similar to the fraud detection with smart metering facilities, because these clients used to have advanced measurement equipment.

CONCLUSION

The MIDAS project proposes an integrated framework which provides several methods to obtain better results in NTL detection. This framework is being used in Endesa company to make campaigns of massive inspections which includes computational intelligence in the analysis process. The process includes computational intelligence based on statistical techniques and data mining. Additionally, it includes an expert system based on knowledge of the inspectors of the company. This variety of modules provides different detection spectrum, i.e. regression analysis and data mining modules, provide NTL detection methods which are traditionally not detected, enriched by the knowledge of inspectors. Moreover, the expert system provides an automated method for traditional NTL detection.

The developed framework shows better results than traditional techniques of massive inspection campaigns. These methods provide additional intelligence to customer selection for inspection. These ones take advantage of all the stored information to make a decision about the customer. By

automating this framework is achieved by making available to inexperienced staff of Endesa, with the possibility of their use in training.

Additionally, the contribution of this work with respect to previous work is the integration into a single framework of the knowledge of inspectors with knowledge extracted from information. Thus, this framework is able to detect non-technical losses obtained by supervised learning techniques and provides new information about non-technical losses previously not easy to detect. This framework not only classifies the different types of consumers suspected of having a non-technical loss but also classifies consumers without non-technical losses.

This framework is used by Endesa and it is in the testing process. Furthermore, it is researching the use of this framework in other utilities.

REFERENCES

Abaide, A. R., Canha, L. N., Barin, A., & Cassel, G. (2010). *Assessment of the smart grids applied in reducing the cost of distribution system losses*. Paper presented at the Energy Market (EEM). London, UK. doi:10.1109/EEM.2010.5558678

Aleskerov, E., Freisleben, B., & Rao, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997*. IEEE. doi:10.1109/CIFER.1997.618940

Alves, R., Casanova, P., Quirogas, E., Ravelo, O., & Gimenez, W. (2006). *Reduction of non-technical losses by modernization and updating of measurement systems*. Paper presented at the Transmission & Distribution Conference and Exposition: Latin America, 2006. doi:10.1109/TDCLA.2006.311590

Aranha Neto, E. A. C., & Coelho, J. (2013). Probabilistic methodology for technical and non-technical losses estimation in distribution system. *Electric Power Systems Research*, *97*, 93–99. doi:10.1016/j.epsr.2012.12.008

Biscarri, F., Monedero, Í., León, C., Guerrero, J. I., & Biscarri, J. (2009). A mining framework to detect non-technical losses in power utilities. In J. Cordeiro & J. Filipe (Eds.), *Proceedings of the 11th International Conference on Enterprise Information Systems* (pp. 96–101). IEEE.

Biscarri, F., Monedero, I., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2008). A data mining method based on the variability of the customer consumption - A special application on electric utility companies. In J. Cordeiro & J. Filipe (Eds.), *Proceedings of the Tenth International Conference on Enterprise Information Systems* (pp. 370–374). Academic Press.

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*. IEEE. doi:10.1109/TAI.1999.809773

Brun, A. D., Pinto, J. O., Pinto, A. M. A., Sauer, L., & Colman, E. (2009). Fraud detection in electric energy using differential evolution. In *Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems*. IEEE. doi:10.1109/ISAP.2009.5352917

Cabral, J. E., & Gontijo, E. M. (2004). Fraud detection in electrical energy consumers using rough sets. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. doi:10.1109/ICSMC.2004.1400905

Cabral, J. E., Pinto, J. O., Martins, E. M., & Pinto, A. M. (2008). Fraud detection in high voltage electricity consumers using data mining. In *Proceedings of the Transmission and Distribution Conference and Exposition*. IEEE/PES. doi:10.1109/TDC.2008.4517232

Cabral, J. E., Pinto, J. O., & Pinto, A. M. A. (2009). Fraud detection system for high and low voltage electricity consumers based on data mining. In *Proceedings of the IEEE Power & Energy Society General Meeting*. IEEE. doi:10.1109/PES.2009.5275809

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and their Applications, 14*(6), 67–74. doi:10.1109/5254.809570

Davis, A. B., & Goyal, S. K. (1992). Knowledge-based management of cellular clone fraud. In *Proceedings of the Third IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE. doi:10.1109/PIMRC.1992.279930

De Oliveira, M. E., Boson, D. F., & Padilha-Feltrin, A. (2008). A statistical analysis of loss factor to determine the energy losses. In *Proceedings of the Transmission and Distribution Conference and Exposition: Latin America*. IEEE/PES. doi:10.1109/TDC-LA.2008.4641691

Depuru, S. S. S., Wang, L., & Devabhaktuni, V. (2011). Support vector machine based data classification for detection of electricity theft. In *Proceedings of the Power Systems Conference and Exposition* (PSCE). IEEE/PES. doi:10.1109/PSCE.2011.5772466

Depuru, S. S. S., Wang, L., Devabhaktuni, V., & Gudi, N. (2010). Measures and setbacks for controlling electricity theft. In *Proceedings of the North American Power Symposium* (NAPS). IEEE. doi:10.1109/NAPS.2010.5619966

Depuru, S. S. S., Wang, L., Devabhaktuni, V., & Nelapati, P. (2011). A hybrid neural network model and encoding technique for enhanced classification of energy consumption data. In *Proceedings of the 2011 IEEE Power and Energy Society General Meeting*. IEEE. doi:10.1109/PES.2011.6039050

Dong, W., Quan-yu, W., Shou-yi, Z., Feng-xia, L., & Da-zhen, W. (2004). A feature extraction method for fraud detection in mobile communication networks. In *Proceedings of the Fifth World Congress on Intelligent Control and Automation, 2004*. IEEE. doi:10.1109/WCICA.2004.1340996

Dorronsoro, J. R., Ginel, F., Sgnchez, C., & Cruz, C. S. (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks*, *8*(4), 827–834. doi:10.1109/72.595879 PMID:18255686

Dos Angelos, E. W., Saavedra, O. R., Cortés, O. A., & de Souza, A. N. (2011). Detection and identification of abnormalities in customer consumptions in power distribution systems. *IEEE Transactions on Power Delivery*, *26*(4), 2436–2442. doi:10.1109/TPWRD.2011.2161621

Fanning, K., Cogger, K. O., & Srivastava, R. (1995). Detection of management fraud: A neural network approach. In *Proceedings of the 11th Conference on Artificial Intelligence for Applications*. IEEE. doi:10.1109/CAIA.1995.378820

Gemignani, M., Tahan, C., Oliveira, C., & Zamora, F. (2009). Commercial losses estimations through consumers' behavior analysis. In *Proceedings of the 20th International Conference and Exhibition on Electricity Distribution - Part 1*. IET.

Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. In *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*. IEEE. doi:10.1109/HICSS.1994.323314

Gonzalez, G., & Figueroa, L. (2006). Strategic plan for the control and reduction of non-technical losses applied in C.A. Energia ElÃƒÂ©ctrica de Valencia. In *Proceedings of the Transmission & Distribution Conference and Exposition: Latin America, 2006*. IEEE/PES. doi:10.1109/TDCLA.2006.311491

Guerrero, J. I., León, C., Biscarri, F., Monedero, I., Biscarri, J., & Millán, R. (2010). Increasing the efficiency in non-technical losses detection in utility companies. In *Proceedings of the MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*. IEEE. doi:10.1109/MELCON.2010.5476320

Hambaba, M. L. (1996). Intelligent hybrid system for data mining. In *Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering, 1996*. IEEE. doi:10.1109/CIFER.1996.501832

Han, L. (2010). Research and implementation of an anomaly detection model based on clustering analysis. In *Proceedings of the 2010 International Symposium on Intelligence Information Processing and Trusted Computing* (IPTC). IEEE. doi:10.1109/IPTC.2010.94

Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., & Mitchell, J. C. … Millán, R. (2010). Using regression analysis to identify patterns of non-technical losses on power utilities. In R. Setchi, I. Jordanov, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based and intelligent information and engineering systems* (Vol. 6276, pp. 410–419). Berlin, Germany: Springer. Retrieved from http://www.springerlink.com/content/43m1340538478854/

Iglesias, J. M. (2006). Follow-up and preventive control of non-technical losses of energy in C.A. Electricidad de Valencia. In *Proceedings of the Transmission & Distribution Conference and Exposition: Latin America, 2006*. IEEE. doi:10.1109/TDCLA.2006.311381

Jin, S., So Yeung, D., Wang, X., & Tsang, E. C. (2005). A feature space analysis for anomaly detection. In *Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. doi:10.1109/ICSMC.2005.1571706

Kerk, S. G. (2005). An AMR study in an Indian utility. In *Proceedings of the Power Engineering Conference, 2005*. IEEE. doi:10.1109/IPEC.2005.206894

Kou, Y., Lu, C.-T., Sirwongwattana, S., & Huang, Y.-P. (2004). Survey of fraud detection techniques. In *Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control*. IEEE. doi:10.1109/ICNSC.2004.1297040

Lei, J. Z., & Ghorbani, A. A. (2012). Improved competitive learning neural networks for network intrusion and fraud detection. *Neurocomputing*, *75*(1), 135–145. doi:10.1016/j.neucom.2011.02.021

León, C., Biscarri, F., Monedero, I., Guerrero, J. I., Biscarri, J., & Millán, R. (2011). Integrated expert system applied to the analysis of non-technical losses in power utilities. *Expert Systems with Applications*, *38*(8), 10274–10285. doi:10.1016/j.eswa.2011.02.062

Liu, Q., Li, T., & Xu, W. (2009). A subjective and objective integrated method for fraud detection in financial systems. In *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*. IEEE. doi:10.1109/ICMLC.2009.5212307

Liu, Y., Jiang, Y., & Lin, W. (2006). A rough set and evidence theory based method for fraud detection. In *Proceedings of the Sixth World Congress on Intelligent Control and Automation, 2006*. IEEE. doi:10.1109/WCICA.2006.1712608

Markoc, Z., Hlupic, N., & Basch, D. (2011). Detection of suspicious patterns of energy consumption using neural network trained by generated samples. In *Proceedings of the ITI 2011 33rd International Conference on Information Technology Interfaces* (ITI). IEEE.

Mohamed, A., Bandi, A. F., Tamrin, A. R., Jaafar, M. D., Hasan, S., & Jusof, F. (2009). Telecommunication fraud prediction using backpropagation neural network. In *Proceedings of the International Conference of Soft Computing and Pattern Recognition, 2009*. IEEE. doi:10.1109/SoCPaR.2009.60

Monedero, Í., Biscarri, F., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2009). *New methods to detect non-technical losses on power utilities*. Paper presented at the IASTED - Artificial Intelligence and Soft Computing. Palma de Mallorca, Spain.

Mwaura, F. M. (2012). Adopting electricity prepayment billing system to reduce non-technical energy losses in Uganda: Lesson from Rwanda. *Utilities Policy*, *23*, 72–79. doi:10.1016/j.jup.2012.05.004

Nagi, J., Mohammad, A. M., Yap, K. S., Tiong, S. K., & Ahmed, S. K. (2008). Non-technical loss analysis for detection of electricity theft using support vector machines. In *Proceedings of the Power and Energy Conference, 2008.* IEEE. doi:10.1109/PECON.2008.4762604

Nagi, J., Siah Yap, K., Kiong Tiong, S., Ahmed, S. K., & Nagi, F. (2011). Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. *IEEE Transactions on Power Delivery*, *26*(2), 1284–1285. doi:10.1109/TPWRD.2010.2055670

Nagi, J., Yap, K. S., Kiong Tiong, S., Ahmed, S. K., & Mohamad, M. (2010). Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery*, *25*(2), 1162–1171. doi:10.1109/TPWRD.2009.2030890

Nagi, J., Yap, K. S., Nagi, F., Tiong, S. K., Koh, S. P., & Ahmed, S. K. (2010). NTL detection of electricity theft and abnormalities for large power consumers. In *Proceedings of the 2010 IEEE Student Conference on Research and Development* (SCOReD). IEEE. doi:10.1109/SCORED.2010.5704002

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohammad, A. M. (2008). Detection of abnormalities and electricity theft using genetic support vector machines. In *Proceedings of the TENCON 2008 - 2008 IEEE Region 10 Conference*. IEEE. doi:10.1109/TENCON.2008.4766403

Nizar, A. H., & Dong, Z. Y. (2009). Identification and detection of electricity customer behaviour irregularities. In *Proceedings of the Power Systems Conference and Exposition, 2009*. IEEE. doi:10.1109/PSCE.2009.4840253

Nizar, A. H., Dong, Z. Y., Jalaluddin, M., & Raffles, M. J. (2006). Load profiling method in detecting non-technical loss activities in a power utility. In *Proceedings of the Power and Energy Conference, 2006*. IEEE. doi:10.1109/PECON.2006.346624

Nizar, A. H., Dong, Z. Y., & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. *IEEE Transactions on Power Systems*, *23*(3), 946–955. doi:10.1109/TPWRS.2008.926431

Nizar, A. H., Dong, Z. Y., & Zhao, J. H. (2006). Load profiling and data mining techniques in electricity deregulated market. In *Proceedings of the IEEE Power Engineering Society General Meeting, 2006*. IEEE. doi:10.1109/PES.2006.1709335

Nizar, A. H., Dong, Z. Y., Zhao, J. H., & Zhang, P. (2007). A data mining based NTL analysis method. In *Proceedings of the IEEE Power Engineering Society General Meeting, 2007*. IEEE. doi:10.1109/PES.2007.385883

Nizar, A. H., Hua Zhao, J., & Yang Dong, Z. (2006). Customer information system data preprocessing with feature selection techniques for non-technical losses prediction in an electricity market. In *Proceedings of the International Conference on Power System Technology, 2006*. IEEE. doi:10.1109/ICPST.2006.321964

Nizar, A. H., Yang Dong, Z., & Zhang, P. (2008). Detection rules for non technical losses analysis in power utilities. In *Proceedings of the 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*. IEEE. doi:10.1109/PES.2008.4596300

Openshaw, D. (2008). Smart metering an energy networks perspective. In *Proceedings of the 2008 IET Seminar on Smart Metering - Gizmo or Revolutionary Technology*. IET.

Özçelik, M. H., Işik, M., Duman, E., & Çevik, T. (2010). Improving a credit card fraud detection system using genetic algorithm. In *Proceedings of the 2010 International Conference on Networking and Information Technology* (ICNIT). IEEE. doi:10.1109/ICNIT.2010.5508478

Raghunath, B. R., & Mahadeo, S. N. (2008). Network intrusion detection system (NIDS). In *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology, 2008*. IEEE. doi:10.1109/ICETET.2008.252

Ramos, C. C. O., de Sousa, A. N., Papa, J. P., & Falcão, A. X. (2011). A new approach for nontechnical losses detection based on optimum-path forest. *IEEE Transactions on Power Systems*, *26*(1), 181–189. doi:10.1109/TPWRS.2010.2051823

Ramos, C. C. O., Papa, J. P., Souza, A. N., Chiachia, G., & Falcao, A. X. (2011). What is the importance of selecting features for non-technical losses identification? In *Proceedings of the 2011 IEEE International Symposium on Circuits and Systems* (ISCAS). IEEE. doi:10.1109/ISCAS.2011.5937748

Ramos, C. C. O., Souza, A. N., Papa, J. P., & Falcao, A. X. (2009). Fast non-technical losses identification through optimum-path forest. In *Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems, 2009*. IEEE. doi:10.1109/ISAP.2009.5352910

Rebahi, Y., Nassar, M., Magedanz, T., & Festor, O. (2011). A survey on fraud and service misuse in voice over IP (VoIP) networks. *Information Security Technical Report*, *16*(1), 12–19. doi:10.1016/j.istr.2010.10.012

Richardson, R. (1997). Neural networks compared to statistical techniques. In *Proceedings of the Computational Intelligence for Financial Engineering (CIFEr)*. IEEE.

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916–5923. doi:10.1016/j.eswa.2013.05.021

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. In *Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications* (INISTA). IEEE. doi:10.1109/INISTA.2011.5946108

Seo, D., Lee, H., & Nuwere, E. (2013). SIPAD: SIP–VoIP anomaly detection using a stateful rule tree. *Computer Communications*, *36*(5), 562–574. doi:10.1016/j.comcom.2012.12.004

Seyedhossein, L., & Hashemi, M. R. (2010). Mining information from credit card time series for timelier fraud detection. In *Proceedings of the 2010 5th International Symposium on Telecommunications* (IST). IEEE. doi:10.1109/ISTEL.2010.5734099

Taniguchi, M., Haft, M., Hollmen, J., & Tresp, V. (1998). Fraud detection in communication networks using neural and probabilistic methods. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. doi:10.1109/ICASSP.1998.675496

Weatherford, M. (2002). Mining for fraud. *IEEE Intelligent Systems*, *17*(4), 4–6. doi:10.1109/MIS.2002.1024744

Wilks, A. J. (1990). Hall effect based electrical energy metering device with fraud detection and instantaneous voltage, current and power outputs. In *Proceedings of the Sixth International Conference on Metering Apparatus and Tariffs for Electricity Supply*. IET.

Wu, L., Ping, R., Ke, L., & Hai-xin, D. (2011). Intrusion detection using SVM. In *Proceedings of the 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing* (WiCOM). IEEE. doi:10.1109/wicom.2011.6040153

Xu, W., Pang, Y., Ma, J., Wang, S.-Y., Hao, G., Zeng, S., & Qian, Y.-H. (2008). Fraud detection in telecommunication: A rough fuzzy set based approach. In *Proceedings of the 2008 International Conference on Machine Learning and Cybernetics*. IEEE. doi:10.1109/ICMLC.2008.4620596

Xue, M., & Zhu, C. (2009). Applied research on data mining algorithm in network intrusion detection. In *Proceedings of the International Joint Conference on Artificial Intelligence, 2009*. IEEE. doi:10.1109/JCAI.2009.25

Yue, D., Wu, X., Wang, Y., Li, Y., & Chu, C.-H. (2007). A review of data mining-based financial fraud detection research. In *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE. doi:10.1109/WICOM.2007.1352

Zhang, Y., Chen, W., & Black, J. (2011). Anomaly detection in premise energy consumption data. In *Proceedings of the 2011 IEEE Power and Energy Society General Meeting*. IEEE. doi:10.1109/PES.2011.6039858

ADDITIONAL READING

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer. doi:10.1007/978-1-4614-3223-4

Banchs, R. E. (2013). *Text Mining with MATLAB*. Springer. doi:10.1007/978-1-4614-4151-9

Berry, M. W. (2004). *Survey of Text Mining I: Clustering, Classification, and Retrieval*. Springer.

Berry, M. W., & Castellanos, M. (2008). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer.

Berry, M. W., & Kogan, J. (2010). *Text Mining: Applications and Theory*. John Wiley & Sons.

Dawid, P., Lauritzen, S. L., & Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer.

Eberhart, R. C., & Shi, Y. (2011). *Computational Intelligence: Concepts to Implementations*. Elsevier.

Elder, J., Hill, T., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Academic Press.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Gallant, S. I. (1993). *Neutral network learning and expert systems*. MIT Press.

Gibbons, J. D., & Chakraborti, S. (2003). Nonparametric Statistical Inference, Fourth Ed.: Revised and Expanded. CRC Press.

Giudici, P., & Figini, S. (2009). *Applied Data Mining for Business and Industry*. John Wiley & Sons. doi:10.1002/9780470745830

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Springer. doi:10.1007/978-3-642-19721-5

Gottlob, G., & Nejdl, W. (1990). *Expert Systems in Engineering: Principles and Applications*. Springer. doi:10.1007/3-540-53104-1

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Hayes-Roth, F., & Lenat, D. B. (1983). *Building expert systems*. Addison-Wesley Pub. Co.

Ibrahim, M., Küng, J., & Revell, N. (2000). *Database and Expert Systems Applications: 11th International Conference, DEXA 2000 London, UK, September 4-8, 2000 Proceedings*. Springer.

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer. doi:10.1007/978-1-84628-754-1

Klahr, P., & Waterman, D. A. (1986). *Expert systems: techniques, tools, and applications*. Addison-Wesley Pub. Co.

Konar, A. (2005). *Computational Intelligence: Principles, Techniques and Applications*. Springer.

Krishnamoorthy, C. S., & Rajeev, S. (1996). *Artificial Intelligence and Expert Systems for Engineers*. CRC PressINC.

Liebowitz, J. (1988). *Introduction to Expert Systems*. Mitchell Publishing.

Liebowitz, J. (1998). *The Handbook of Applied Expert Systems*. CRC PressINC.

Lucas, P., & Gaag, L. V. D. (1991). *Principles of Expert Systems*. Addison-Wesley.

Marik, V., Retschitzegger, W., & Stepankova, O. (2003). *Database and Expert Systems Applications: 14th International Conference, DEXA 2003, Prague, Czech Republic, September 1-5, 2003, Proceedings*. Springer.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques* [electronic resource]. Springer.

Pace, L., & Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. World Scientific.

Parthasarathy, S. (2010). *Enterprise Information Systems and Implementing IT Infrastructures: Challenges and Issues*. IGI Global Snippet. doi:10.4018/978-1-61520-625-4

Rohatgi, V. K. (2003). *Statistical Inference*. Courier Dover Publications.

Rutkowski, L. (2008). *Computational Intelligence: Methods and Techniques*. Springer.

Segura, J. M., & Reiter, A. C. (2011). *Expert System Software: Engineering, Advantages and Applications*. Nova Science Publisher's, Incorporated.

Shafer, G. (1996). *Probabilistic Expert Systems*. SIAM. doi:10.1137/1.9781611970043

Shapiro, A. D. (1987). *Structured Induction in Expert Systems*. Turing Inst. Press.

Slatter, P. E. (1987). *Building expert systems: cognitive emulation*. Ellis Horwood.

Sol, H. G., Takkenberg, C. A. T., & Robbé, P. F. de V. (1987). Expert Systems and Artificial Intelligence in Decision Support Systems. Springer.

Srivastava, A., & Sahami, M. (2010). *Text Mining: Classification, Clustering, and Applications*. CRC Press.

Tommelein, I. D. (1997). *Expert Systems for Civil Engineers: Integration Issues*. ASCE Publications.

Turban, E., & Frenzel, L. E. (1992). *Expert Systems and Applied Artificial Intelligence*. Macmillan Publishing Company.

Turban, E., & Watkins, P. R. (1988). *Applied expert systems*. North-Holland.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. Springer. doi:10.1007/978-1-84996-226-1

Williams, G. J., & Simoff, S. J. (2006). *Data Mining: Theory, Methodology, Techniques, and Applications*. Springer.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier.

KEY TERMS AND DEFINITIONS

Data Mining: Data mining is the discovery of interesting, unexpected or valuable structures in large datasets.

Neural Network: Neural networks imitate the brain's ability to sort out patterns and learn from trial and error, discerning and extracting the relationships that underlie the data with which it is presented.

Non-Technical Losses: The non-technical losses (NTLs) in power utilities are defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment.

Power Utility: Industry dedicated to the power distribution.

Rule-Based Expert Systems: An expert system based on a set of rules that a human expert would follow in diagnosing or analysis problems.

Statistical Inference: Statistical inference is the process of drawing conclusions from data that is subject to random variation.

Text Mining: Text mining deals with the machine supported analysis of text, it uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of Knowledge Discovery in Databases (KDD), data mining, machine learning and statistics.