# A real application on non-technical losses detection: the MIDAS Project

**J.I. Guerrero[1], C. León[1], Senior Member, IEEE, F. Biscarri[1], Í. Monedero[1], J. Biscarri[2] and R. Millán[2]**
[1]Electronic Technology Department, University of Seville, Seville, Spain
[2]Automated Metering Management and Field Works Department, Endesa, Seville, Spain

**Abstract** - *The MIDAS project began at 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 test phase and it is applied in real cases in company databases.*

**Keywords:** data mining, expert system, text mining, power, utility

## 1 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 increasing of storage capacity and the process capacity allow one to manage large databases. Data mining provides a set of techniques which make information treatment easier. Additionally, there are several 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 of customers. In these databases several maintenance and management processes are performed. In addition, the utility companies have to 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 utility companies invest great quantities of efforts to correct it.

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 new 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, etc. The proposed framework is used in a test stage in Endesa Company.

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.
- Architecture of the framework is described.
- Each module is described.
- Finally the conclusions and references are described.

## 2 Review of current state

### 2.1 Review bibliographic

Generally, there are several similarities between detection of NTL and the detection of anomalies or anomalous patterns. The detection of anomalies or frauds is treated in several references. Specifically, [1] describes several data mining techniques for fraud detection in credit card, telecommunication networks and intrusion detection; [2] applies several algorithms of data mining and artificial intelligence for fraud detection in the financial field; [3] compared neural networks to statistical techniques for fraud in transactional systems; [4] uses classification methods (principal component analysis and bivariate statistics) for fraud detection in mobile communication network. There are

more references in fraud detection ([5],[6],[7]) but there are less references to NTLs detection, specifically: [8] uses the rough sets for fraud detection in electrical energy consumers; [9] proposed neural network with radial basis function (RBF); [10] and [11] proposed the application of support vector machines (SVMs) with genetic algorithms. Other references use predictive techniques for fraud detection, for example: [12] proposes the integration of artificial neural networks, a genetic algorithm to predict electrical energy consumption and [13] proposed the integration of neural network, time series and ANOVA for forecasting electrical consumption. Sometimes, these references include the demand forecasting in short ([14]), medium ([15]) or long ([16]) term.

### 2.2 Utility company

The system proposed in this paper is used in test phase in Endesa company. Endesa is the most important energy distribution company of Spain with more than 12 million of clients, and it is found in European and South American markets.

Traditionally, there are two types of losses: non-technical losses (NTLs) and technical losses. The technical losses are caused by faults in distribution lines. These faults are predictable with a low rate of error. 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 that carry 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.

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.

## 3 The MIDAS Project

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, Sadiel, FIDETIA 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.

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 (Joole effect). The Technical Losses can be forecasted because they are approximate constants, but the NTLs are very irregular and very difficult to forecast.

The MIDAS project follows the prototype life cycle, and gets a new version at each iteration. 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 to make 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 automatic and, normally, they only detect some type of NTLs.

Currently, 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.

## 4 System architecture

The proposed system has an architecture with several modules. Each module is implemented with different techniques. The modules are increased each iteration of life cycle with each prototype. Each prototype is tested with real data of Endesa databases and it is validated with inspections made by Endesa staff.

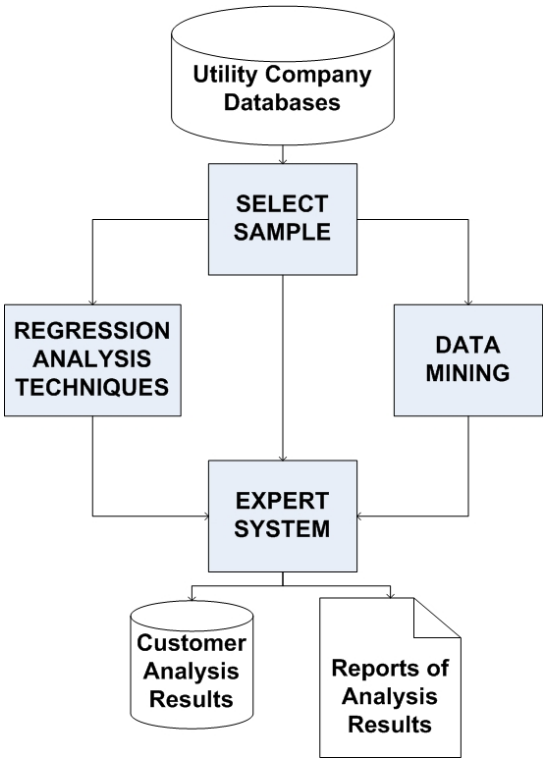


Figure 1. System Architecture

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. This module provides a set of customers who have some anomalies regarding customer self consumption or the other customers consumption. Mainly, this modules works with some parameters: consumption, contracted power, economic activity. In the last place, the knowledge based expert system (KBES) analyzes the rest of information about the customer. The KBES provides the results on databases and reports can be both used by inspectors as an additional source of information. In the following each of these modules is described.

# 5 Select sample

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

- 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.
- Contractual power.
- Economic Activity. Some economic sectors historically present a high rate of NTLs. The research of data mining is centered in these sectors.
- Consumption range. Sometimes, the consumption range can be used to restrict the quantity of customers.
- Electricity charges.

These parameters allow one to restrict the quantity of customers to analyze. The information of each customer compounds information about: contract, installed equipment, results of inspections realized over the facilities, etc.

The information about consumption is analyzed by data mining, statistical techniques and neural networks, and then, the rest of information is analyzed by KBES.

# 6 Data mining

Initially, techniques used in data mining are the outliers analysis and inherent data variability. These techniques are described in [17]. For this process a sample of homogeneous data which have utility customers with similar characteristics are selected. The temporary and the local components of the individual consumption of customer are eliminated by means of normalization. After this process, 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, outliers are used to guide the inspections.

After the use of these methods, the use of other techniques of data mining was considered. Currently, there exists a framework MIDAS. This framework compounds several techniques related with data mining: descriptive data mining, predictive data mining, etc. All of them allow one to increase the efficiency of detection process using the same information. These methods have two processes in common:

- Data Selection. Although, there is a previous step of Selection Sample this step filters the customer who cannot be treated by the proposed data mining techniques.
- Data Preprocessing.

Normally, these processes are used for normalization and discretization purposes.

## 6.1 Descriptive data mining

This process is described in [18]. Three descriptive 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.

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 model is strongly dependent of the cluster of customers considered and highly changeable amongst different clusters. But the study of the individual trend consumption and also the comparative among trends of customer with similar characteristics is very interesting.

There are some feature levels or some feature relationships quite serious in reference to NTLs detection. Some of them are:

- The hours of consumption at maximum contracted power.
- Minimum and maximum values of consumption in different time zones of the day or time discrimination

bands.
- 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.

### 6.2 Predictive data mining

This process is described in [18], too. This module uses an inference of a rule set to characterize each of two following classes: 'normal' customer or 'anomalous' customer. Each customer is characterized by means of several attributes. The predictive (or classification) model uses supervised learning with the attributes generated by descriptive data mining.

The classification algorithm uses the Generalize Rule Induction (GRI) model. This model discovers association rules in the data. 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. 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 test of the set of rules generates four values, according to the following classifications [19]:

- 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.

### 6.3 Clustering and decision trees

This method is described in [20]. The company in its inspections has developed this method based on the recognition of customers with the same consumption pattern than those NTLs previously detected. This method was held on a process of generation of clusters. Thus, firstly we designed a feature vector that could identify the consumption pattern of each one of the customers. This 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.

In addition, two parameters are added. These parameters were based on the concept of streak, described in previous section.

## 7 Regression analysis

This method is described in [21]. This method identifies the patterns of drastic drop of consumption. 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.

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.

## 8 Expert system

This Rule Based Expert System (RBES) is described in [22]. This system uses the information extracted from Endesa staff. 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 data warehousing.

The RBES can 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 and rules from the statistical techniques and text mining modules.

The system can be used alone or with other modules to provide an additional method to analyze the information. These modules are described in the following sections.

### 8.1 Statistical techniques

The statistical techniques are based in basic indicators such as: maximum, minimum, average and standard deviation. 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 calculi 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.

The creation of these patterns needs to study a lot of customers. 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.

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 isn't correct.

### 8.2 Text mining

Text mining is described in [23] and uses Natural Language Processing (NLP) and neural networks. 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 in consumption study. The text-mining 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.

This technique uses NLP and fuzzy algorithms to extract concepts from inspectors' commentaries. This concept is 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 less frequent concepts. This neural network can be used to classify the new concepts which could appear.

## 9 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 vocational periods. The using of descriptive data mining and expert system allows detect these cases.

## 10 Conclusions

This paper proposes a framework to detect non-technical losses in power utility. Several techniques are used to detect and classify the customers according to the problem found in them. The main contribution of this framework is the possibility to analyze all the information related to the customer. Traditionally, the analysis is restricted to the consumption and some additional parameters, but this framework compounds:

- Analyzing the coherence of information.
- Analyzing of customers' consumption and trend of consumption:
  - Regarding customer self and their characteristics.
  - Regarding other customers and other customers with same or equivalent characteristics.
- Analyzing the characteristics of customers according to the knowledge extracted from Endesa staff.
- Analyzing the commentaries specified by Endesa staff about the customer and their facilities.
- Reporting all results of analysis.

The analysis of all information provides a more efficient response to the staff company. Additionally, this framework provides advanced knowledge and experience to other users of the company.

This framework uses supervised and unsupervised learning methods. These methods allow one to get better results than traditional methods of massive inspections. The data mining techniques and regression analysis techniques allow one to select consumers with a suspicious consumption and rule based expert system is used to analyze, in depth, the rest of information about each consumer.

# 11 References


[1] Yufeng Kou, Chang-Tien Lu, Siriat Sinvongwattana, and Yo-Ping Huang, "Survey of fraud detection techniques". Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control, pp.21-23. Taipei, Taiwan, 2004.

[2] R. Wheeler, and S. Aitken, "Multiple Algorithms for fraud detection" Rev. Knowledge-Based Systems, 13, pp.93-99, 2000.

[3] R. Richardson, "Neural networks compared to statistical techniques". Computational Intelligence for Financial Engineering (CIFEr). Proceeding IEEE/IAFE (1997).

[4] Wang Dong et al., "A feature extraction method for fraud detection in mobile communication networks". Proceeding of the 5 World Congress on Intelligent Control and Automation, Hangzhou, P. R. China, June 15-19, 2004.

[5] David J. Hand, "Prospecting for gems in credit card data". IMA Journal of Management Mathematics 12, pp. 172-200. 2001.

[6] R. Bolton, and D. Hand, "Statistical fraud detection: a review". Statistical Science. Volume 17, Issue 3, pp. 235-255. 2002.

[7] G. K. Palshikar, "The hidden truth – frauds and their control: a critical application for business intelligence". Intelligent Enterprise, Volume 5, Issue 9, pp. 46-51. 2002, 28 May.

[8] J. Cabral, J. O. P. Pinto, E. Gontijo, and J. Reis Filho, "Fraud detection in electrical energy consumers using rough sets". IEEE International Conference on Systems, Man and Cybernetics, Volume 4, pp. 3625-3629, The Hauge, Netherlands, 2004.

[9] J. R. Galván, A. Elices, A. Muñoz, T. Czernichow, and M. A. Sanz-Bobi. "System for detection of abnormalities and fraud in customer consumption". 12th Conference on the Electric Power Supply Industry. Pattaya, Thailand. November 2-6, 1998.

[10] J. Nagi, A. M. Mohammad, K. S. Yap, J. K. Tiong, and S. K. Ahmed. "Non-Technical loss analysis for detection of electricity theft using support vector machines". 2nd IEEE International Conference on Power and Energy (PECon 08), Johor Bahru, Malaysia. December 1-3, 2008.

[11] J. Nagi, S. K. Yap, S. K. Tiong, S. K.Ahmed, and A. M. Mohammad. "Detection of abnormalities and electricity theft using genetic support vector machines". TENCON 2008, IEEE Region 10 Conference. Pp. 1-6. 2008, 19-21 Nov.

[12] A. Azadeh, S. F. Ghaderi, S. Tarverdian, and M. Saberi, "Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption". Applied Mathematics and Computation 186, pp. 1731-1741. 2007.

[13] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, "Forecasting electrical consumption by integration of neural network, time series and ANOVA". Applied Mathematics and Computation 186, pp. 1753-1761. 2007.

[14] B. F. Hobbs, U. Helman, S. Jitprapaikulsarn, S. Konda, and D. Maratukulam. Artificial neural networks for short-term energy forecasting: accuracy and economic value. Neurocomputing 23, pp. 71-84, 1998.

[15] M. Gavrilas, I. Ciutea, and C. Tanasa, "Medium-term load forecasting with artificial neural network models". CIRED2001, Conference Publication No. 482. 2001 June.

[16] K. Padmakumari, K. P. Mohandas, and S. Thiruvengadam, "Long term distribution demand forecasting using neuro fuzzy computations". Electrical Power and Energy systems, 21, pp. 315-322, 1999.

[17] F. Biscarri, I. Monedero, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, "A data mining method based on the variability of the consumer consumption". ICEIS 2008: 10th International Conference on Enterprise Information Systems. Barcelona, Spain. Proceedings, Volume 2, pp. 370-374. June 2008.

[18] F. Biscarri, I. Monedero, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, "A mining framework to detect non-technical losses in power utilities". ICEIS 2009: 11th International Conference on Enterprise Information Systems. Barcelona, Spain. Pp- 96-101. May 2009.

[19] J. Cabral, J. Pinto, E. Martins, and A. Pinto, "Fraud detection in high voltage electricity consumers using data mining". In IEEE Transmission and Distribution

Conference and Exposition T&D. IEEE/PES. April 21-24, 2008.

[20] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, "New methods to detect non-technical losses in power utilities". ASC 2009: 13th IASTED International Conference on Artificial Intelligence and Soft Computing. Palma de Mallorca (Spain). Proceedings, pp. 7-13. September 2009.

[21] Iñigo Monedero, Félix Biscarri, Carlos León, Juan I. Guerrero, Jesús Biscarri, and Rocío Millán, "Using regression analysis to identify patterns of non-technical losses on power utilities". KES 2010. Knowledge-Based and Intelligent Information and Engineering Systems, 14th International Conference, Cardiff, UK. September 8-10, 2010.

[22] Carlos León, Félix Biscarri, Iñigo Monedero, Juan I. Guerrero, Jesus Biscarri, and Rocío Millán, "Integrated expert system applied to the analysis of non-technical losses in power utilities". Expert System with Applications, in press [doi: 10.1016/j.eswa.2011.02.062].

[23] J. I. Guerrero, Carlos León, Félix Biscarri, Iñigo Monedero, Jesús Biscarri, and Rocío Millán, "Increasing the efficiency in non-technical losses detection in utility companies". MELECON 2010, 15th IEEE Mediterranian Electromechanical Conference. Pp. 136-141. Valleta, Malta. 25-28 April, 2010.