FRAUD DETECTION IN LOW VOLTAGE ELECTRICITY CONSUMERS USING SOCIOECONOMIC INDICATORS AND BILLING PROFILE IN SMART GRIDS

Jonatas PULZ
Daimon Eng. and Systems – Brazil
jonatas.pulz@daimon.com.br

Renan B. MULLER
Daimon Eng. and Systems – Brazil
renan.muller@daimon.com.br

Fabio ROMERO
Daimon Eng. and Systems – Brazil
fabio.romero@daimon.com.br

André MEFFE
Daimon Eng. and Systems – Brazil
andre.meffe@daimon.com.br

Álvaro F. GARCEZ NETO
Sulgipe – Brazil
alvaro.garcez@sulgipe.com.br

Aldo S. JESUS
Sulgipe – Brazil
aldo.santana@sulgipe.com.br

ABSTRACT

The Brazilian Association of Energy Distribution Utilities estimates that non-technical losses represent more than 5.5% of the total energy distributed, most coming from fraud and theft. To try to mitigate those losses, the distribution utilities send field crews for the inspection of possible fraudster clients. However, the procedure is expensive and gives no financial return to the utility if it is not focused on areas with high fraud probability. On those locations, there is a correlation between losses and socio-economic indices. Thus, this work proposes a model able to select clients with high fraud probability, which should be visited by the field crews. The smart grid structure, energy consumption data, clients' registration data and socio-economic indices from the 2010 Brazilian Census are used by the model.

INTRODUCTION

The study presented on this paper is part of a Research and Development (R&D) Project of ANEEL (Brazilian Electricity Regulatory Agency), developed jointly by Energy Company of South Sergipe (SULGIPE) and Daimon Engineering & Systems.

The combat against electricity fraud and theft is a great challenge for the distribution utilities in Brazil and other developing countries. The Brazilian Association of Energy Distribution Utilities (ABRADEE) estimates that non-technical losses represent more than 5.5% of the total energy distributed [1]. To mitigate these losses, the distribution utilities send field crews for the inspection of consumers trying to eliminate frauds, deviations on upstream energy meters or other methods used by the consumers to avoid paying for the consumed energy.

Dispatching field crews can be expensive and impacts on the electricity tariff, according to the Brazilian regulatory rules [2]. Thus, the inspection teams must be used efficiently. This efficiency can be achieved by a system able to indicate which clients has high probability of being a fraudster, so that such should be inspected. Thus, the inspections would generate a financial return to the utility with the end of the fraud and a possible repayment [3].

Currently, SULGIPE does not have a smart grid infrastructure. Thus, the R&D project considers the installation of smart energy meters along the grid allowing data to be monitored and acquired remotely, a first step for turning it into a smart grid. Comparing the data acquired by the smart energy meters to the energy charged from the clients, it is possible to identify potential fraudster areas. Within those areas, a classification model can be implemented to identify which consumers should be investigated, using as parameters the energy consumption data, the clients' registration data and socio-economic indices.

Reference [4] mentions on paragraph 49 a study that says the level of losses is associated to a high level of violence, percentage of people living in subnormal domiciles – the definition of subnormal domicile can be found in [5] – and urbanization rate; justifying the using of these data in the classification model.

The socioeconomic information should be regionalized by area, for example, by neighbourhood, to be effective. The proposed model uses these data divided in censual sectors (as it is made available by the 2010 Brazilian Census), which will be defined in the following section.

There are many works that study models for the detection of fraudster consumers. In [6], energy consumption data and a classification method called Optimum-Path Forest are used. In [7], non-payment indices and daily load curves obtained from monthly consumption data are used to train a Support Vector Machine (SVM). In [8], 3 types of models using consumption data are compared: neural network, logistic regression and discriminant analysis.

Other works use the smart grid structure to identify frauds. Research [9] proposes the using of sensors and smart energy meters along the grid allowing the comparison of the energy fed-in and the energy officially consumed and charged from the clients. When it is a low voltage area, a classification model combining SVM and Decision Trees is applied to identify clients with high fraud potential. In [10], there is central energy meter able to monitor the energy fed-in to a certain number of clients, which is used for fraud identification.

The project proposed on this paper uses measurements from smart energy meters available along the grid compared to data of energy charged from the clients to estimate non-technical losses. In low voltage areas with high non-technical losses, a model able to classify the clients from each area will be applied. The model uses SVM, its possible outputs are fraudster and non-fraudster and its inputs are energy consumption data, clients'

registration data and socio-economic indices from the 2010 Brazilian Census.

METHODOLOGY

The R&D project considers the installation of 6 smart energy meters along the grid. A specific feeder from SULGIPE's grid was chosen as a pilot area. Figure 1 presents the chosen feeder from Itabaianinha - SE, Brazil as well as the installation locations of the smart energy meters. Thus, with 6 smart energy meters, the feeder is divided into 6 areas (bounded by the red lines in the figure). The installation points were chosen in a way that each circuit area bounded by the meters has almost the same aggregate consumption. The energy measured from the smart energy meters is used to calculate the energy consumed in each area.

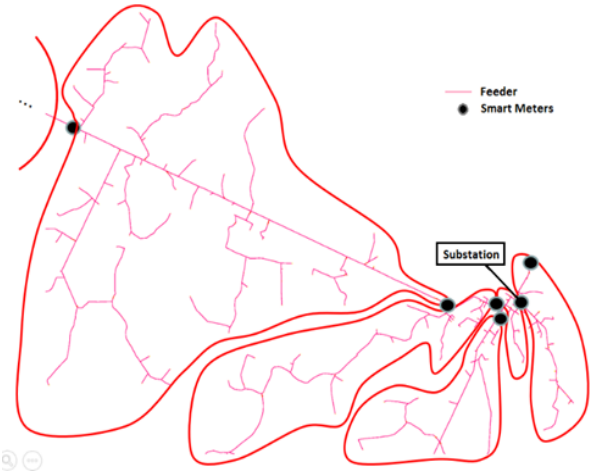

Figure 1. Chosen feeder and installation locations of the smart energy meters.

Figure 2 presents the fluxogram of the methodology that determines the areas with non-technical losses. Knowing the energy charged from the clients and the energy measured from the smart energy meters along the grid, it is possible to estimate the non-technical losses for each of the 6 areas. In an area without non-technical losses, the technical losses plus the energy charged from the clients should be equal to the energy consumed in the area (calculated with the energy measured from the smart energy meters). In case the energy measurements don't match, it means there might be fraud in the area. For the fraudster regions, the classification model is applied, indicating the clients with high fraud probability, which should be inspected.

Over the previous years, each inspection was registered with its results: fraud found or fraud not found, creating an inspection history. Thus, the classification model consists of a SVM trained by the inspection history made by the distribution utility. Its inputs (attributes) are energy consumption data, clients' registration data and socio-economic indices from the 2010 Brazilian Census of each client.

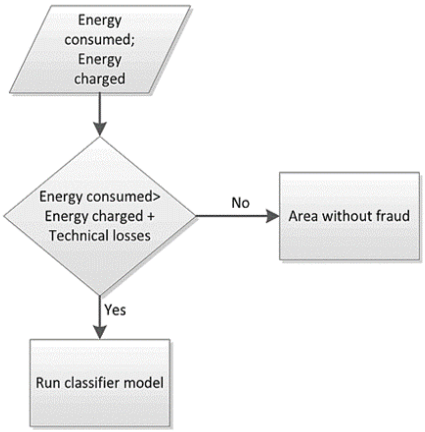

Figure 2. Detection of areas with non-technical losses.

Attributes of the classification model

The attributes used are energy consumption data, clients' registration data and socio-economic indices from the 2010 Brazilian Census of each client.

Socio-economic attributes

The Brazilian Institute of Geography and Statistics (IBGE) is the main provider of data and information about the country. The data are grouped into censual sectors. A censual sector is a territorial unity defined for registration. It is composed of adjacent areas located in an urban or a rural area alone, with dimensions that are supposed to ease the data acquisition process [11].

The censual sectors border data are available on IBGE's site [12] like geographic coordinates in the shape of polygons. The results of the 2010 Brazilian Census are available on [13]. With these two sets of data in hands and the coordinates of each utility consumer, it is possible to know the socio-economic indices of each of them, considering the indices of the censual sector where it is located. Therefore, all clients in the same censual sector will have the same socio-economic indices, representing the situation of the region. In that way, the proposed classification model can identify a socio-economic pattern related to frauds.

The 19 attributes from the 2010 Brazilian Census, initially considered, are presented in the first column in the Table 1. The definitions of inadequate and semi-adequate domiciles are available on [11]. The terms used here for people's color and race were translated directly from IBGE.

Using the information about the clients, a logistic regression was used with data available from the historic of inspections made by Utility over the previous years. The attributes in which the result of the null hypothesis test was 5% or more were eliminated. The attributes used in the classification model is the ones with "YES" in the second column from Table 1.

Table 1. Attributes extracted from the 2010 Brazilian Census.

Attributes	Used in the classification model
Average number of people by domicile;	YES
Variance of the number of people by domicile;	YES
Average income of the head of the domicile;	NO
Variance of the income of the head of the domicile;	NO
Percentage of rented domiciles;	YES
Percentage of people who live in rented domiciles;	YES
Percentage of domiciles with water supply;	YES
Percentage of domiciles with private restroom and sewage treatment;	YES
Percentage of domiciles with waste collection;	YES
Percentage of inadequate domiciles;	NO
Percentage of semi-adequate domiciles;	YES
Percentage of white people;	NO
Percentage of black people;	NO
Percentage of yellow people;	YES
Percentage of multiracial people;	YES
Percentage of people living in inadequate domiciles;	YES
Percentage of people living in semi-adequate domiciles;	NO
Percentage of literate people aged 18 and over; and	YES
Percentage of people aged 10 and over receiving one minimum wage or less.	NO

Attributes extracted from the monthly energy consumption and the clients' registration data

The following attributes were extracted from the clients' registration data:

i. Input power type (three-phase, two-phase and single-phase, encoded as -1, 0 and +1, respectively); and
ii. Energy meter type (electromechanical and electronic, encoded as -1 and 1, respectively).

The following attributes refers to monthly energy consumption data of the last 12 months:

i. Consumption coefficient of variation (standard deviation over mean);
ii. Consumption standard deviation;
iii. Average consumption of the sector;
iv. Average consumption of the client over the average consumption of the census sector;
v. Maximum of the absolute value of the first difference series of the last 12 months of consumption over the average consumption of the client, and;
vi. Ratio between the modules of the coefficient of the first harmonic and the coefficient of index 0 of the Discrete Fourier Series of the last 12 months of consumption.

The understanding of attributes i and ii is direct. Attribute iii is the mean of the consumption of all the clients of a census sector. Attribute iv is the ratio between the average consumption of a client over the last 12 months and the average consumption of its census sector over the same period. Attribute v comes from a difference series: given the 12 monthly consumption values, its first difference series is shown in (1)

$$c(2)-c(1),\ c(3)-c(2),\ \ldots c(12)-c(11), \quad (1)$$

where c(n) the energy consumption on month n, n = 1, 2, …, 12.

It is possible to note that the first difference series has 11 elements. Then, the absolute value is calculated for each element and the greatest is chosen, resulting in attribute v when divided by the average consumption of the client. In attribute vi, the coefficients of the Discrete Fourier Series are calculated, considering the consumption of the last 12 months as a periodic signal. From this series, it is calculated the ratio between the modules of the coefficient of the first harmonic (with a period of 12 months) and the coefficient of index 0 (average consumption). This last attribute is used because of the energy consumption periodicity, once a domicile usually presents a repetitive consumption pattern over the period of a year. The periodicity is related to the seasons and the repetition of annual habits like travelling in the end of the year.

The data used for model training are based on the historic of inspections made by the utility and, to obtain the attributes related to the consumption of a costumer inspected, it must be used the 12 months before the inspection.

Classification Model

To indicate which clients will be inspected, a classification model using a SVM is proposed.

In possession of the socio-economic data from the 2010 Brazilian Census, the clients' registration data and the consumption data, the attributes are extracted, as showed in the 'Attributes of the classification model' Section, resulting in 20 attributes for each client (12 from the Census, 2 from registration and 6 related to consumption). Thus, each client has an attribute vector **x**, given as input to the SVM.

$$\mathbf{x}=[x_1\ x_2 \cdots x_m \cdots x_{20}], \quad (2)$$

where x_m the m-th attribute of the attribute vector of a client.

The historic of inspection made by the utility is used for training and testing the model. An attribute vector **x** is generated for each visited client as well as its class: fraudster, when a fraud was found, and non-fraudster, when a fraud was not found. The model's output is y, which is the probability of the client to belong to the fraudster class.

RESULTS

Until now, in the R&D project, only the classification model was developed and tested. The identification of areas with non-technical losses using the smart energy meters along the grid will be implemented later in the project.

The concept of cross-validation – k-fold method – was used to test the classification model. More details about the process are available on [14]. This method can find a result less dependent on the chosen test sample, once it is possible to calculate the average result of the m implemented tests and the its variance.

Table 2 is the contingency table of the proposed model. It evaluates the outcomes relating the model's output and the sample's real condition.

Table 2. Contingency table of the classification model.

		Sample's real condition	
		Fraudster	***Non-fraudster***
Model's output	***Fraudster***	True Positive (TP)	False Positive (FP)
	Non-fraudster	False Negative (FN)	True Negative (TN)

Some indices are defined for the evaluation of the model's performance. The Accuracy is defined as follows

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The True Positive Ratio (TPR), also called sensitivity, is

$$TPR=\frac{TP}{TP+FN} \quad (4)$$

The False Positive Ratio is (FPR) is

$$FPR=\frac{FP}{FP+TN} \quad (5)$$

Finally, the Precision is

$$\text{Precision}=\frac{TP}{TP+FP} \quad (6)$$

The model's output is the probability of a client to be fraudster, that is, a value between 0 and 1. A threshold must be chosen as a minimum value for the client to be considered fraudster. If the clients' probability y is greater than it, then it is classified as fraudster. Choosing the threshold is not trivial and is a challenge for any binary classification problem. The Receiving Operating Characteristic (ROC) curve usually helps in this kind of problem. In this case, it is a TPR by FPR plot for many threshold values. In this type of curve, the greater its area or the closer it is to the vertical axis, the better is the model. A common reference of ROC curve is the diagonal, which represents a random model. More details about ROC curves are available on [15].

In order to train and test the model, the data set was balanced between fraudster and non-fraudster, that is, 50% of the samples are fraudster and 50% are not. Figure 3 to Figure 5 present the results of the cross-validation. The first chart, Figure 3, shows the Accuracy curve as function of the threshold and the Precision bars as function of a threshold range, that is, for each Precision bar, it was considered a threshold interval with minimum and maximum values in which the client would be fraudster. The second chart, Figure 4, presents the results of the model's test in histograms of the clients in relation to the y value (model's output); the blue are the fraudster clients and the orange are the non-fraudster; the brown bars represent the superposition of the two histograms. The third chart, Figure 5, presents the ROC curve of the proposed model; the diagonal is drawn just for reference (representing a random model).

The result indicates that the classification model is able to detect the frauds, once the fraudster clients from the inspection history are concentrated on y values closer to 1 and the non-fraudster are closer to 0 as seen on Figure 4. It is also confirmed by the ROC curve, Figure 5,which is always above the diagonal.

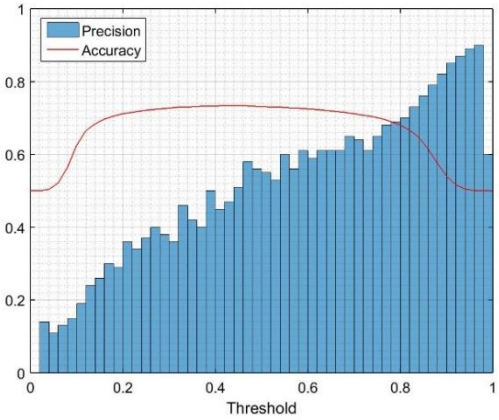

Figure 3. Precision and Accuracy of the classification model by threshold.

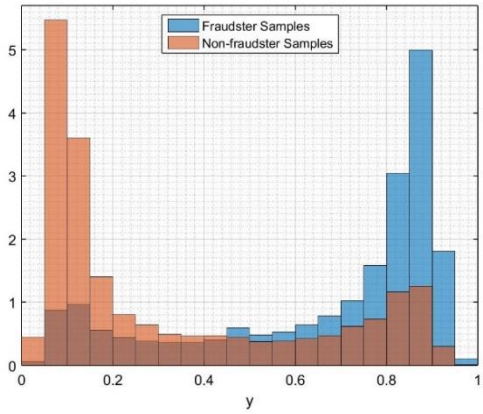

Figure 4. Normalized histograms of the samples used for test the model related to its output y.

CONCLUSION

This work proposes a model able to identify potential low

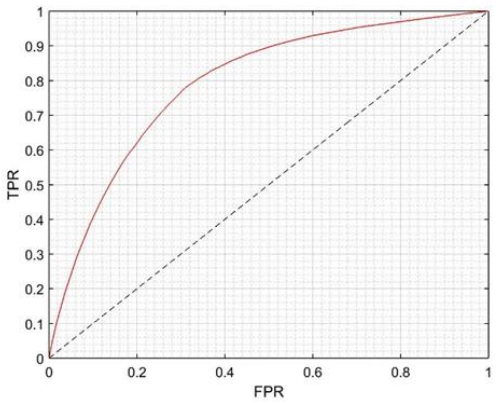

Figure 5. ROC curve of the classification model obtained by cross-validation.

voltage fraudster clients by detecting areas with high non-technical losses and then applying a classification model that indicates the fraudster clients. The inputs of the classification model are: regionalized socio-economic attributes from the 2010 Brazilian Census, attributes extracted from energy consumption and attributes extracted from the clients' registration data.

In order to detect the areas with non-technical losses, it is necessary to allocate smart energy meters along the grid to provide data to be compared to the energy charged from the clients. The installation of that meters is the first step towards a smart grid.

According to data provided by SULGIPE, in 2015, frauds were found in just 54% of all the inspections made. Naming this percentage as inspection effectiveness index, it corresponds to the Precision (6) of a model that indicates the clients to be inspected. That is, a model that presents a high Precision percentage will provide to the utility a high inspection effectiveness index. As indicated by Fig. 5, the model proposed by this project can have an average inspection effectiveness index of 70%. However, previous experiences indicate that this theoretic effectiveness index is not verified in the real inspections. Commonly, the real effectiveness index is substantially lower than the theoretic one, mainly because of the crew's capacity to detect frauds, the crew's honesty in not accepting bribery and others aspects.

REFERENCES

[1] T. Resende. (2013, Oct.) Perdas na distribuição: baixa tensão, altos prejuízos – Reportagem Especial Canal Energia. Associação Brasileira de Distribuidores de Energia Elétrica (ABRADEE). [Online]. Available on: http://www.abradee.com.br/imprensa/artigos-e-releases/1018-perdas-na-distribuicao-baixa-tensao-altos-prejuizos-reportagem-especial-canal-energia

[2] Procedimentos de Regulação Tarifária: Submódulo 2.5 - Fator X v2.0, Agência Nacional de Energia Elétrica - ANEEL.

[3] Agência Nacional de Energia Elétrica – ANEEL. "Resolução nº 414, de 9 de setembro de 2010. Estabelece as Condições Gerais de Fornecimento de Energia Elétrica de forma atualizada e consolidada," September 2010.

[4] Agência Nacional de Energia Elétrica – ANEEL. "Nota técnica nº 342, 11 de novembro de 2008. Metodologia de tratamento regulatório para perdas não técnicas de energia elétrica," November 2008.

[5] Aglomerados subnormais, Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2010, Rio de Janeiro, 2010.[Online]. Available on: http://biblioteca.ibge.gov.br/visualizacao/periodicos/552/cd_2010_agsn_if.pdf

[6] C. O. Ramos, A. N. Sousa, J. P. Papa, and A. X. Falcão, "A New Approach for Nontechnical Losses Detection Based on Optimum-Path Forest," IEEE Trans. on Power Systems, vol. 26, no. 1, February 2011.

[7] J. Nagi, A. M. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, "Non-Technical Loss Analysis for Detection of Electricity Theft using Support Vector Machines," in Proc. of IEEE International Conference on Power and Energy.

[8] A. S. Penin, "Combate, Prevenção e Otimização das Perdas Comerciais de Energia Elétrica," Ph.D. dissertation, Polytechnic School of University of São Paulo , São Paulo, 2008.

[9] Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and M. Sukumar, "Decision Tree and SVM-based Data Analytics for Theft Detection in Smart Grid," IEEE Trans. on Industrial Informatics, 2016.

[10] W. Han and Y. Xiao, "NFD: A practical scheme to detect non-technical loss fraud in smart grid," in Proc. of 2014 IEEE International Conference on Communications (ICC).

[11] Características urbanísticas do entorno dos domicílios, Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2010, Rio de January, 2010. [Online]. Available on: http://biblioteca.ibge.gov.br/visualizacao/periodicos/96/cd_2010_entorno_domicilios.pdf

[12] Malha Digital de Setores Censitários, Instituto Brasileiro de Geografia e Estatística. [Online]. Available on: ftp://geoftp.ibge.gov.br/malhas_digitais/censo_2010/setores_censitarios/

[13] Resultados Agregados por Setores Censitários, Instituto Brasileiro de Geografia e Estatística. [Online]. Available on: ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_ Censitarios/

[14] C. S. Jensen and R. T. Snodgrass, Encyclopedia of Database Systems, 1st ed., L. Liu and M. T. Özsu, Eds. Springer US, 2009.

[15] N. A. Macmillan and C. D. Creelman, Detection theory: a user's guide, 2nd ed. Psychology Press, 2004.