Improving electricity non technical losses detection including neighborhood information

Pablo Massaferro, Henry Marichal,
Matias Di Martino, Fernando Santomauro, Juan Pablo Kosut and Alicia Fernandez.

***Abstract*— Non technical losses (NTL) cause significant damage to power supply companies' economies. Detecting abnormal clients behavior is an important and difficult task. In this paper we analyze the impact of considering customers geo-localization information, in automatic NTL detection. A methodology to find optimal grid sizes to compute a set of local features with a random search procedure is proposed. The number and size of the grids, and other classification algorithm parameters are adjusted to maximize the area under receiver operating characteristic curve (AUC), showing performance improvements in a data set of 6 thousand of Uruguayan residential customers. Comparative analysis with different sub-sets of characteristics, that include the monthly consumption, contractual information and the new local features are presented. In addition, we probe that raw customers' geographical location used as an input feature, gives competitive results as well. In addition we evaluate a entire new database of 6 thousand Uruguayan customers, whom were inspected in-site by UTE experts between 2015 and 2017.**

I. Introduction

Since the nineteenth century the access to electricity has strongly influenced the way we live. In particular, in the past forty years electrical consumption has increased dramatically and access to electricity is a major concern in modern societies. In this context, losses in power grids are a very important problem that every year generates substantial economic losses. These losses can be classified into two categories: technical (TL) and non-technical (NTL). Technical losses are associated with dissipation or failures of the electrical components of the power grid, while NTL are associated with electricity theft, faulty meters or billing errors.

NTLs cause a significant harm to economies, for example, in India NTLs are estimated at $4.5 billion, and in countries such as Brazil, Malaysia or Lebanon NTLs can represent up to the 40% of the total electricity distributed [1], [2]. In the UK and USA non-technical losses are estimated between $1-6 billion [1], [3], [4]. The present work is developed in Uruguay as part of an existing collaboration between the "Universidad de la Republica" (UdelaR) university and UTE (the national company in charge of the power distribution for the whole country). In Montevideo, capital of Uruguay, TL and NTL represented the 7% and 13% respectively of the total energy distributed during the year 2016.

Related Work

Different machine learning approaches have addressed the detection of non-technical losses, both supervised or unsupervised. Leon et al. review the main research works found in the area between 1990 and 2008 [5], and Glauner et al. made a recent survey including the latest work in the field [1]. Several of these approaches consider unsupervised classification using different techniques such as fuzzy clustering [6], neural networks [7], [8], among others. Monedero et al. used regression based on correlation between time and monthly consumption, looking for significant drops in consumption [9]. Supervised approaches on the other hand, build and learn mathematical models that describe the problem based on labeled datasets provided by power distribution companies. For example, many works explored the use of Support Vector Machines (SVM) algorithm [2], [3], [10], [11] or combinations of SVM method with other methods such as Genetic Algorithm [12].

From the point of view of the features used to represent each customer profile, a different path has been followed. Some distribution companies have access to the real-time energy consumption measurements, or smart meters that can monitor the energy consumed with a temporal resolution of minutes or hours [13]. However the most common scenario is to have access to monthly [3], [12], [14]–[16] or bimonthly [17] energy consumptions. These consumption profiles are obtained from different customers, and are used as input features of machine learning systems. It is also common to enhance this feature vector with additional features extracted from the profiles (such as Fourier coefficients, local averages, between many others) [14]. Also, information of customers consumption is sometimes complemented with additional information such as: meter type, history of theft, or credit worthiness rating [3], [18] between other data that could be associated to the customer profile.

More recently, Glanuer et al. [19] included the use of neighborhood local features by splitting the area in which the customers are located into grids of different sizes. For each grid cell they compute the proportion of inspected customers and the proportion of NTL found among the inspected customers. Other recent research [20] uses a Generalized Additive Model to generate a local estimation of NTL, and Markov chains to estimate the future changes of it. To that end, this work makes use of complementary socio-economical variables obtained from the latest national (Brazilian) census.

In the present work we improved our previous works [14], [18] presenting a more effective and robust automatic method for NTL detection. Inspired by the recent work of Glauner et al. [19] we define a new set of features and extend some of the ideas there presented. The main contributions of the present work are: (i) We performed thousands of in site inspections to

obtain an updated database of Uruguayan customer behavior. This database was never reported or tested before and we performed several experiments to test different methods in the actual context of the city of Montevideo. (ii) We propose a procedure to find optimal sets of neighborhood features that prove to be effective improving detection of non-technical electrical losses, we probe that this technique can improve previous features proposed in the literature. (iii) We compare engineered features extracted using neighborhood information with the use of raw geographical coordinates.

II. NTL DETECTION SCHEME

In the present article we analyze the effectiveness of different sets of novel features for the detection of Non-Technical losses. The general structure of the implemented solution consists of two main blocks: (a) features computation from customers' input data, and (b) customer classification (as suspicious or normal). For the classification task, we consider Random Forest algorithm as an standard classification algorithm on pattern recognition. The first step, features representation, is where we put the focus on this work and will be described in detail in the following.

A. Performance metrics

One key aspect of the design of automatic solutions using pattern recognition theory, is definition of the performance criteria. This is how the effectiveness of the proposed algorithms and techniques is defined.

NTL detection can be typically stated as an *imbalanced problem* [14], because the number of customers that belong to the Normal (or negative) class, greatly exceeds the number of customers that belong to the Fraud (or positive) class. Hence, standard performance measurements such as the Classification Error or Accuracy are not suitable on the context of NTL detection and alternative indicators need to be defined. For example, reference [19] used the Area Under ROC Curve (AUC) as the performance criteria. AUC can be defined in binary classification problems by taking into account the true positive rate and true negative rate as follows,

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (1)$$

where TP denotes the total number of Positive samples correctly classified, TN the number of negative samples correctly classified, while FN/FP represents the number of false Negative/Positive obtained. In addition, other useful indicators in the context of imbalanced problems are:

$$Recall = \frac{TP}{TP+FN}, \qquad Precision = \frac{TP}{TP+FP}, \quad (2)$$

and

$$F_\beta = (1+\beta^2).\frac{Recall.Precision}{\beta^2 Recall + Precision}. \quad (3)$$

TABLE I
ADDITIONAL FEATURES SELECTED FOR USE

Cod	Feature	Description
01	Actual Readings Proportion	Ratio between readings carried out by UTE employees over total readings.
02	Power Paid For	Maximum power usable by the client, limited by a thermal switch
03	Number of Irregularities	Number of irregularities recorded for the period.
04	Days Since Last Inspection	Days passed since the last on field inspection.
05	Days Since Update	Days passed since the meter was updated or relocated.
06	Default	Days of late payment during the past year.
07	Days Since the contract is in force	Days passed since the actual contract was signed.
08	Contract Status	Contract status: active, inactive

B. Features proposed

We focused on commercial and residential customers for which the following data was available: (i) monthly consumption (typically one measure per month) (ii) additional billing information (such as: the contracted power, the contract status: active/inactive), (iii) customers geographical location, and (iv) additional information generated by on-site inspections performed by UTE employees.

Two years of monthly consumptions were selected from the entire customer consumption profile, the starting month was the same for all clients in order to avoid the impact of climate stationarity.

In addition, a set of additional features were also analyzed, that showed to be discriminative in previous related works e.g. [18]. The features shown in table I were selected from a larger set available.

Neighborhood Features. As first set of features we wanted to analyze the eight characteristics proposed in [19], in order to match the physical grid dimensions to the ones described, we define the set of discrete grids over the geographical extension of the city of Montevideo. Then, the ratio of irregularities Eq. (4) and the inspection ratio Eq. (5) are calculated on each cell of each grid.

$$inspected\_ratio = \frac{\#inspected}{\#customers} \quad (4)$$

$$NTL\_ratio = \frac{\#NTL}{\#inspected} \quad (5)$$

Figure 1 illustrates the value of features inspected_ratio and NTL_ratio over the finest grid (24×24 cells, each cell has an area of $0.4km^2$).

Grid optimization. In addition to the features proposed in [19], in the present work we analyze the impact of learning

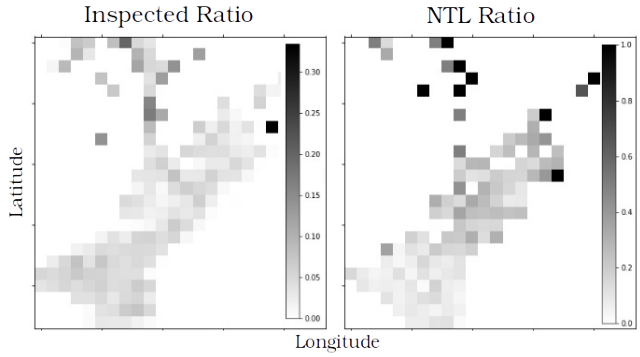

Fig. 1. Inspected ratio and NTL ratio values on the finest grid over the city of Montevideo.

an optimal grid topography from the data collected through inspecting customers across Montevideo city.

We learn from the data: (i) which are feasible grid structures that capture the desired neighborhood information, and (ii) from these set of possible grids which is the optimal subset of features that we must keep.

In order to create different grid structures that can be adapted in a better way to the geographical structure of the data, we need to establish a set of basic partition units. For example, considering the sizes of the regular grids presented in the previous section, we can start by considering (3×3), (6×6), (12×12) and (24×24) as the basic partition units, then we can generate new sets of grids by combining these set of basic partitions sizes. For example, we can consider grids of dimensions (3×12), (24×6), (9×3) or any other possible combination obtained from the basic size partitions $\{3, 6, 12, 24\}$ as it is illustrated in Fig.2.

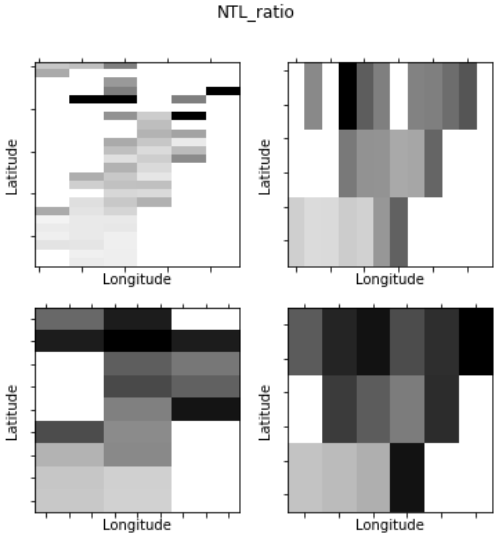

Fig. 2. Examples of some possible random grids generated form the basic grid sizes $\{3, 6, 12, 24\}$

Once new grids are defined, it is possible to compute two different features (*inspected_ratio* and *NTL_ratio*) on each of these new grids. Then from this large set of features, we can find which are the subsets of more discriminative ones. Let us illustrate the previous with a simple example, assume that we generate new arbitrary grids using as basic sizes $\{s_1, \cdots, s_N\}$ and we compute two new features on each of those new grids. We have in total N^2 possible grids, so we will have $2 \times N^2 \stackrel{def}{=} M$ possible features. Moreover, from those new defined features, we can define $\sum_{i=1}^{M-1} \frac{M!}{i!(M-i)!} + 1$ different subsets, for example, if $N = 4$, we have in total about 10^9 possible features subsets.

In the present work, we find suitable subsets of neighborhood features defined in non-homogeneous grids by evaluating random subsets of features. To that end, we follow six simple steps:

1) Set the grid basic partition units, i.e. which are the unitary sized to be considered.
2) Define a set of random grids for features NTL_ratio and Inspected_ratio.
3) Compute NTL_ratio and Inspected_ratio on their respective set of grids.
4) Update the feature set by including the new (computed) featured to the set of consumptions and additional features.
5) Train Random Forest classifier on this new feature space using a data set.
6) Evaluate the previous model on a test data set and compute the AUC.[1]
7) Repeat this procedure N times and define the optimal grids set as the set that achieved the highest AUC.

C. Classification algorithm

After the features space is set, it is necessary to proceed with a segmentation of it, identifying the regions of the feature space associated to each of the classes. To that end, in the present work Random Forest classifier is considered. This very robust and popular classifier consists on a combination of simple classification trees applied to subsets of data generated by randomly selecting features. The result of each tree is used to define the prediction by majority voting. This method is robust to noise and the generation of data subsets decreases potential effects of overfitting. Parameters commonly used to optimize the training model are: the number of trees, and the number of random features per tree. A detailed description and analysis of Random Forest method can be found in reference [21]

III. Experiments

A. Data

The database analyzed in the present work was generated from a set of 6029 clients that were inspected on cite by UTE experts during the year 2015 in Montevideo city. The

[1]Steps 5 and 6 can be performed on a single data base using Cross-validation.

geographical region in which the clients where selected has an extension of $25.2km^2$ and a total of $115K$ active clients. The inspections performed allowed to label each of the analyzed customers as abnormal (label 1) or normal (label -1), in total 11.6% of the customers inspected present irregularities associated to some kind of NTL. Thanks to a joint work with the Uruguayan UTE company, in the present work we were able to test and evaluate the presented ideas in a new and updated dataset. This data was collected after months of in-place inspections during the past two years (2015-2017).

B. Results

As we stated in the previous section, the focus on the present article is the analysis of different features for the sake of fraud detection. Specifically, we want to explicitly evaluate the impact of including geographical customers information. To that end, five experiments were designed and tested over the database described above. Each experiment was performed using 10-fold cross validation, and that procedure was in addition repeated 10 times (mixing randomly the database). Random Forest was used as the classifier algorithm. The five set of featured evaluated are:

1) set "C": 24 monthly consumptions.
2) set "C + A": set C plus 8 additional features.
3) set "NF": set C + A plus 8 neighborhood features computed over a pre-defined regular grid.
4) set "ONF": set C + A plus the optimal set of neighborhood features (on irregular optimal grids).
5) set "GC": set C + A plus the raw geographical coordinates (latitude and longitude).

Each experiment was performed imposing five different NTL proportions on the training set, which helps to deal with the data unbalance problem. For each experiment, AUC, Accuracy, Recall, Precision and F_1 are reported. Tables II - III and IV show the results obtained using the set of features "C", "NF" and "ONF" respectively. Figure 3 shows the AUC results obtained for each of the feature sets evaluated, for a NTL ratio of 50%. Each experiment was repeated 10 times (each time performing 10-fold cross-validation), the minimum, maximum, mean and standard deviation AUC is represented on the plot.

TABLE II
RESULTS USING "C" FEATURES.

NTL prop.	AUC	Accuracy	Precision	Recall	F_1
12%	0,521	0,884	0,443	0,050	0,090
20%	0,542	0,875	0,354	0,110	0,168
30%	0,575	0,837	0,263	0,234	0,248
40%	**0,607**	0,750	0,208	0,423	0,279
50%	0,601	0,594	0,162	0,610	0,256

C. Discussion

The analysis of the experiments here presented disclose interesting conclusions. First, the results obtained using neighborhood features ("NF") are on the same order of magnitude of the ones obtained in [19], this is a interesting result as we

TABLE III
RESULTS USING "NF" FEATURES.

NTL prop.	AUC	Accuracy	Precision	Recall	F_1
12%	0,541	0,880	0,401	0,100	0,161
20%	0,567	0,860	0,315	0,187	0,235
30%	0,597	0,824	0,264	0,303	0,282
40%	0,627	0,759	0,226	0,456	0,302
50%	**0,633**	0,642	0,184	0,620	0,284

TABLE IV
RESULTS USING "ONF" FEATURES.

NTL prop.	AUC	Accuracy	Precision	Recall	F_1
12%	0,538	0,885	0,485	0,087	0,148
20%	0,570	0,873	0,379	0,178	0,242
30%	0,605	0,837	0,294	0,305	0,299
40%	0,634	0,769	0,237	0,460	0,313
50%	**0,646**	0,645	0,190	0,649	0,294

TABLE V
MEAN AUC RESULTS FOR 10 TIMES 10 FOLDS EXPERIMENTS

NTL prop.	C	C+A	set NF	set ONF	set GC
12%	0,521	0,526	0,541	0,538	0,542
20%	0,542	0,547	0,567	0,570	0,574
30%	0,575	0,587	0,597	0,605	0,607
40%	**0,607**	0,621	0,627	0,634	0,639
50%	0,601	**0,623**	**0,633**	**0,646**	**0,646**

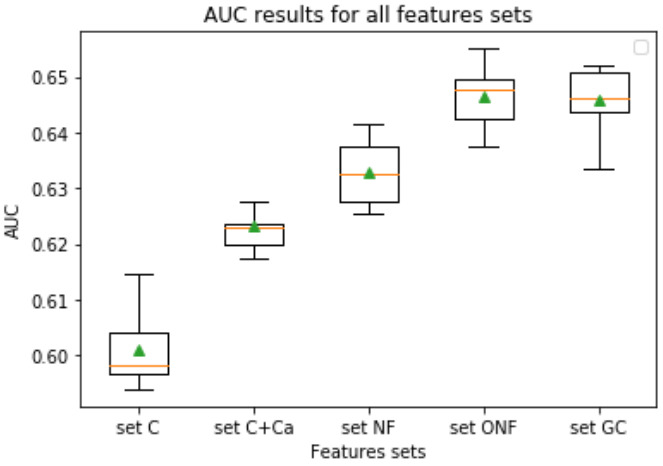

Fig. 3. AUC for different set of features evaluated. The box extends from the lower to upper quartile values of the data, with a line at the median and a triangle at the mean value. The whiskers extend from the box to show the range of the data obtained on the 10 runs of the 10-fold cross-validation for each experiment.

TABLE VI
ALL METRICS ALL FEATURES SETS

	C	C+A	set NF	set ONF	set GC
AUC	0,607	0,623	0,633	**0,646**	**0,646**
F_1	0,279	0,273	0,284	0,294	0,294
Recall	0,423	0,644	0,620	0,649	0,650
Precision	0,208	0,173	0,184	0,190	0,190
Accuracy	0,750	0,608	0,642	0,645	0,643

evaluated these features on a completely independent dataset, on a different country. Secondly, we verify that the inclusion of additional features, significantly improve NTL detection rates, compare for instance, the results obtained using "C" and "C+A" features. Thirdly, we probe that neighborhood features can be improved by defining optimal grids topographies, which allows us to capture in a more flexible way the geographical structure of a city. Fourthly, we show that if we consider customers' geographical location (i.e. their address latitude and longitude) very competitive results can be obtained. This last observation is very important from a practical point of view, since the inclusion of this type of information can be done in a very easy way.

IV. Conclusions and future work

We improved our previous works [14], [18] presenting a more effective feature set for NTL detection, where, inspired in the recent work of Glauner et al. [19] we defined a new set of local features using customers geo-localization. Moreover we proposed a methodology to find optimum grids sizes in order to compute local features and compared the results with other methods and the use raw geographic coordinates. The results obtained allowed us to conclude that the inclusion of additional features, significantly improve NTL detection rates; and that neighborhood features are enhanced when optimal grids are defined. In addition, we probed that raw customers' geographical location used as an input feature, gives very competitive results. In future work we plan to use a grid search methodology to perform a smarter normalization of other features derived from the monthly consumption record. We are currently testing the NTL strategy presented by performing new on-site inspections across the country, and we plan to conclude and report these set of on-field validations within two years. In the present work, we followed the AUC score as optimization criteria, we will investigate the impact of using different metrics (e.g. F_β) to set, and optimize the NTL algorithms here presented.

Acknowledgments

The authors would like to thank UTE for providing datasets and share fraud detection expertise. This work was supported by "Proyecto ANII Fondo Sectorial de Energía 14038".

References

[1] P. Glauner, J. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *International Journal of Computational Intelligence Systems 10.1 (2017): 760-775.*, 2017.

[2] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and R. C. Green, "High performance computing for detection of electricity theft," *International Journal of Electrical Power & Energy Systems*, vol. 47, pp. 21–30, 2013.

[3] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, 2010.

[4] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving svm-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Transactions on power delivery*, vol. 26, no. 2, pp. 1284–1285, 2011.

[5] C. Leon, F. X. E. L. Biscarri, I. X. F. I. Monedero, J. I. Guerrero, J. X. F. S. Biscarri, and R. X. E. O. Millan, "Variability and trend-based generalized rule induction model to ntl detection in power companies," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 1798–1807, 2011.

[6] E. dos Angelos, O. Saavedra, O. Cortes, and A. De Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 4, pp. 2436–2442, 2011.

[7] Z. Markoc, N. Hlupic, and D. Basch, "Detection of suspicious patterns of energy consumption using neural network trained by generated samples," *Proceedings of the ITI 2011 33rd International Conference on Information Technology Interfaces*, pp. 551–556, 2011.

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

[9] I. Monedero, F. Biscarri, C. Leon, J. Guerrero, J. Biscarri, and R. Millan, "Using regression analysis to identify patterns of non-technical losses on power utilities," in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, R. Setchi, I. Jordanov, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2010, vol. 6276, pp. 410–419.

[10] C. C. O. Ramos, A. N. De Souza, D. S. Gastaldello, and J. P. Papa, "Identification and feature selection of non-technical losses for industrial consumers using the software weka," pp. 1–6, 2012.

[11] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Support vector machine based data classification for detection of electricity theft," *2011 IEEE/PES Power Systems Conference and Exposition*, pp. 1–8, 2011.

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

[13] S.-j. Chen, T.-s. Zhan, C.-h. Huang, J.-l. Chen, and C.-h. Lin, "Non-technical Loss and Outage Detection Using Fractional-Order Self-Synchronization Error-Based Fuzzy Petri Nets in Micro-Distribution Systems," *IEEE Transaction on smart grid*, vol. 6, no. 1, pp. 411–420, 2015.

[14] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "Improving electric fraud detection using class imbalance strategies." in *International Conference on Pattern Recognition and Methods, 1st. ICPRAM.*, 2012, pp. 135–141.

[15] P. Glauner, A. Boechat, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Large-scale detection of non-technical losses in imbalanced data sets," in *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*. IEEE, 2016, pp. 1–5.

[16] C. Muniz, K. Figueiredo, M. Vellasco, G. Chavez, and M. Pacheco, "Irregularity Detection on Low Tension Electric Installations by Neural Network Ensembles," no. March 2016, 2009.

[17] F. Biscarri, I. Monedero, C. Leon, J. I. Guerrero, J. Biscarri, and R. Millan, *A data mining method based on the variability of the customer consumption - A special application on electric utility companies*. Inst. for Syst. and Technol. of Inf. Control and Commun., 2008, vol. AIDSS, pp. 370–374.

[18] J. P. Kosut, F. Santomauro, A. Jorysz, A. Fernández, F. Lecumberry, and F. Rodríguez, "Abnormal consumption analysis for fraud detection: Ute-udelar joint efforts," in *Innovative Smart Grid Technologies Latin America (ISGT LATAM), 2015 IEEE PES*. IEEE, 2015, pp. 887–892.

[19] P. Glauner, J. Meira, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Neighborhood features help detecting electricity theft in big data sets," in *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies.*, 2016.

[20] L. T. Faria, J. D. Melo, and A. Padilha-Feltrin, "Spatial-temporal estimation for nontechnical losses," *IEEE Transactions on Power Delivery*, vol. 31, no. 1, pp. 362–369, 2016.

[21] L. Breiman, "Random forests machine learning. 45: 5–32," *View Article PubMed/NCBI Google Scholar*, 2001.