Electricity Theft Detecting Based on Density-Clustering Method

Kedi Zheng[1], Yi Wang[1], Qixin Chen[1], and Yuanpeng Li[2]

***Abstract*— Nowadays, the problem of electricity theft and tampered smart meter data is causing widespread concern. Customer load profiles collected from smart meters can help detect abnormal electricity users and identify electricity theft. In this paper, a density-based electricity theft detection method is proposed to find out abnormal electricity patterns. Several malicious types are used to test the validation of the proposed method. Comparisons with *k*-means clustering, Gaussian mixture model (GMM) clustering and density-based spatial clustering of applications with noise (DBSCAN) are also conducted. Numerical experiments show that the proposed method outperforms other methods in almost all the theft types.**

***Index Terms*— Electricity theft, smart meter data, density-based clustering, abnormal detection.**

I. Introduction

The abnormal behaviors of electricity users, especially electricity theft, have been causing huge economic losses to power utilities all around the world. For example, it is roughly estimated that the economic loss of electricity is $15 million [1] in Fujian Province, China. Consumer fraud in the electrical grid is also causing as much as $6 billion loss to providers in the US alone [2]. The research report [3] released by Northeast Group, LLC in January 2017 says that many of the emerging market countries suffer from rampant non-technical losses which are mostly due to electricity theft. The total cost is $64.7 billion each year in lost or unbilled revenue.

With the development of smart grid and rising penetration of smart meters, there have been new kinds of attacks against smart meters. While the energy thefts usually concerned with physically cut-off or damage in the past, they now can have attacker models [4] like Erase Logged Events, Tamper Storage, Intercept Communication, Man in the Middle, etc. These models will change the data recorded by smart meters and help reduce the electricity bills of the fraudsters. It is reported as a case of the Federal Bureau of Investigation (FBI) in 2010 that some former employees of a meter manufacturer alter the smart meters for residential and commercial users and even training others to do so in exchange for money in Puerto Rico [5]. On the one hand, the user profiles received by the power utilities might be disguised sophisticatedly. On the other hand, smart meters record the power consumption data at a rather higher frequency than traditional meters, which provides opportunities to increase the accuracy of anomaly detection.

The common methods for detecting electricity theft can be classified into three categories: system-state based, game-theory based and artificial-intelligence based [6]. The system-state based methods utilize the conflicts between tampered smart meter data and other measurements in the distribution network. The game-theory based methods usually have prepositional assumptions on the customer behaviors. A lot of additional information is required for the first two categories, which is rather difficult to get. The artificial-intelligence based methods use data mining techniques to extract information from basic user profiles, which are more likely to be applied in practice.

Artificial intelligence can help search for data that do not follow expected patterns. These methods have the assumption that the pattern of electricity thefts is different from that of normal users. We classify the artificial-intelligence based methods into three subcategories: classification-based, load-forecasting-based, and clustering-based.

The classification-based methods require a labeled dataset to train a classification model. Classifiers like neural network [7] and support vector machines (SVMs) [8] are applied to detect irregular consumption behaviors. With proper selection of activation function or optimization of parameters, good detection rate can be achieved. However, a labeled dataset for electricity detecting is hard to get in reality.

The load-forecasting-based methods forecast the future load of users and compare the forecast result with the measured load. In [9] a weighted averaging scheme is used to predict the load. Liu et al. applied periodic auto-regression with eXogenous variables (PARX) to predict short-term energy consumption in [10]. Actually, a forecasting model needs to be retrained for each customer and extra information like weather is required to improve accuracy.

The clustering-based methods do not need unlabeled data. These methods extract patterns from a lot of user features and detect the outlier patterns. Júnior et al. applied Optimum-path forest (OPF) clustering in non-technical loss identification [11]. *k*-means, Gaussian mixture model (GMM) clustering and other famous clustering methods are also used as comparison. In [12] the fuzzy C-means clustering is used to detect unusual customer consumption profiles. The abnormality degree of each client can be obtained from the fuzzy membership. In [13] the famous density-based spatial clustering of applications with noise (DBSCAN) is used to detect abnormal consumption data. The effect of the clustering methods usually depends on the selection of

This work was supported by National Key R&D Program of China (No. 2016YFB0900100) and State Grid Shanxi electric power company, Taiyuan power supply company under the project of Research and applicatioin of key technologies (operation mode) of Taiyuan regional energy Internet

[1]K. Zheng, Y. Wang, Q. Chen are with the State Key Lab of Power Systems, Dept. of Electrical Engineering, Tsinghua University, Beijing 100084, China. (E-mail: qxchen@tsinghua.edu.cn)

[2]Y. Li is with Guangdong Experimental High School

978-1-5386-4950-3/17/$31.00

parameters.

Rodriguez et al. put forward a new density-based clustering method in [14]. For convenience, we call it densityClust which is the name of its implementation package in R [15]. It calculates the density features of a dataset without any preset parameters. The features can be used to find core points and abnormal points effectively. When applied to electricity theft detection, it can adapt to large datasets and does not require any additional data except the customer load profiles. A general model that need not be retrained for different customers can be built. In this paper, we proposed an electricity detecting method based on the theory of densityClust. Several evaluation criteria are introduced and the method is tested on a synthetic dataset. The results are compared with other unsupervised learning techniques to show the effect.

The rest of this paper is organized as follows. Section II presents our methodology of electricity theft detecting. Section III details the evaluation and comparison of the methodology. Section IV shows a case study and its results. Finally, Section V gives the conclusions of this paper.

II. Methodology

Density-based clustering methods have been widely adopted in anomaly detecting. Compared with *k*-means and other partition based clustering methods, density-based clustering can deal with clusters with an arbitrary shape. However, in traditional density-based clustering methods like DBSCAN, one needs to choose the radius of the neighborhood and the density threshold, which is usually non-trivial. Our method tries to overcome the disadvantages using the new densityClust theory.

In densityClust, two values are defined for each data point p_i: its local density ρ_i and its distance δ_i from points of higher density. Both values depend on the distances d_{ij} between the data points. Eq. 1 shows the definition of ρ_i:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

Where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is the cut-off distance. Since the local density ρ_i is discrete in Eq. 1, a Gaussian kernel is used to estimate ρ_i as shown in Eq. 2 to avoid conflicts:

$$\rho_i = \sum_{j \neq i} e^{-(\frac{d_{ij}}{d_c})^2} \tag{2}$$

As the cut-off distance d_c changes, ρ_i defined in Eq. 2 changes more smoothly for small datasets than in Eq. 1. The definition of δ_i is shown in Eq. 3:

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \tag{3}$$

For data points with the highest local density, δ_i is conventionally written as in Eq. 4:

$$\delta_i = \max_j d_{ij} \tag{4}$$

The cut-off distance d_c is exogenous in the definitions. It can be defined by the users or automatically chosen by a rule of thumb, in which d_c is chosen so that the average of ρ_i is around 1 to 2% of the total number of points. From the definition above, we can see that data points with global or local maximum ρ_i usually have much larger δ_i. These points can be recognized as cluster centers.

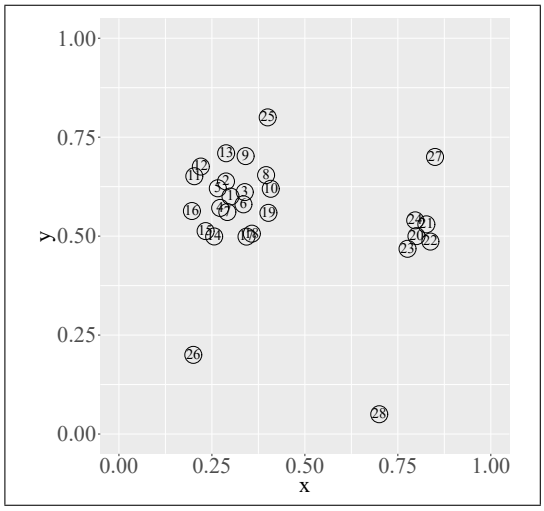

Fig. 1. An example of point distribution

We use the data points in Fig. 1 as an example. It is clear that the points labeled with 1 and 20 are cluster centers and that the points labeled with 26 to 28 are abnormal points. We perform densityClust with the Gaussian kernel and plot the (ρ_i, δ_i) of each point in the coordinate system as in Fig. 2, which is called the decision graph.

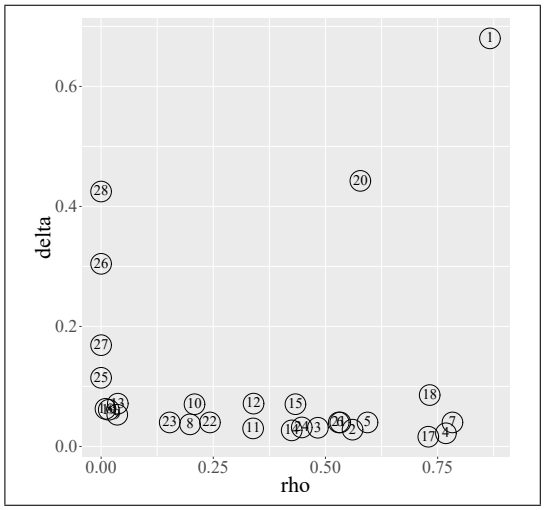

Fig. 2. The decision graph of the points

The data points with large $\gamma_i = \rho_i \delta_i$ is very likely to be the cluster centers and those with small ρ_i and large δ_i is very likely to be abnormal points. We define that $\zeta_i = \delta_i / \rho_i$ represents the degree of abnormality. In occasions when the dataset is large enough and the Gaussian kernel is not necessary, ζ_i can be defined as in Eq. 5 to avoid infinity.

$$\zeta_i = \frac{\delta_i}{\rho_i + 1} \tag{5}$$

The flow chart of our methodology is shown in Fig. 3. We first normalize the customer load profiles for every customer i and for every day j (K is the number of days), because our method focuses on the shape of the load curve. The normalized load profiles are the input of densityClust and the abnormality degree ζ_{ij} is calculated. For every customer i, the number of his abnormal days M_i is calculated. It is believed that stealing electricity is a continuous process, so if M_i is larger than a threshold, the customer will be labeled as electricity theft.

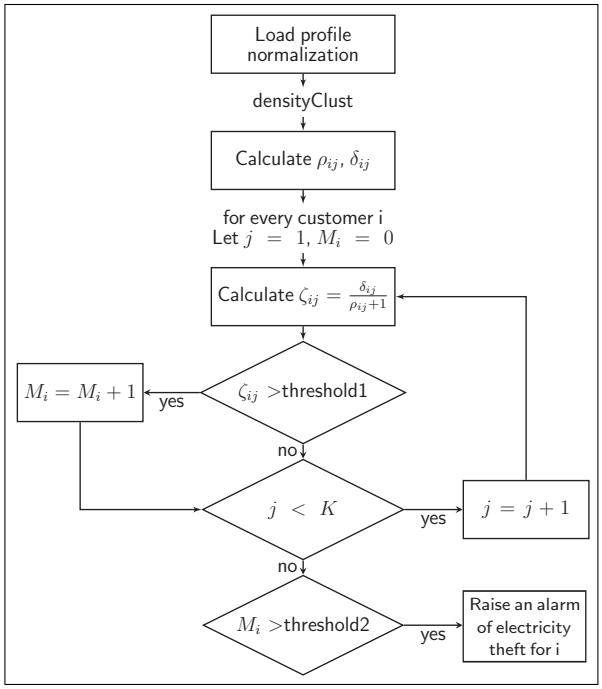

Fig. 3. The flow chart of the proposed methodology

III. Evaluation Criteria and Comparisons

In this section, we will introduce three widely-used evaluation criteria for electricity theft detecting methods, or binary classifiers more generally. Also, several comparison methods including *k*-means, GMM clustering, and DBSCAN will be briefly introduced.

A. Confusion Matrix and Evaluation

The confusion matrix divides the whole dataset into four parts: true positive (TP), false positive (FP), false negative (NP) and true negative (TN). TP, FP, FN, and TN are defined as the numbers of positives correctly predicted as positives, negatives incorrectly predicted as positives, positives incorrectly predicted as negatives, and negatives correctly predicted as negatives respectively. An example of the confusion matrix is shown in Table I.

Several evaluation criteria can be derived from the confusion matrix. The true positive rate (TPR) also known as the hit rate or recall rate is defined as the proportion of positives that are correctly identified as positives. The false positive rate (FPR) is defined as the proportion of positive results that are false positive. The positive predictive values (PPV) also known as the precision are defined as the proportion of positive results that are true positive.

TABLE I
Confusion Matrix of a Binary Classifier

		Predicted	
		Electricity theft	Normal user
Actual	Electricity theft	TP	FN
	Normal user	FP	TN

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

The accuracy (ACC) of the classifier is defined as the proportion of the correct results. The F1 score is defined as the harmonic mean of TPR and PPV, which is useful while dealing with imbalanced datasets.

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$F1 = \frac{2 \times PPV \times TPR}{PPV + TPR} \tag{10}$$

As the discrimination threshold of a binary classifier varies, the TPR and FPR will change consistently. The track of (FPR, TPR) is a curve connecting $(0, 0)$ and $(1, 1)$, which is called the receiver operating characteristic (ROC) curve. The area under the curve (AUC) measures the effectiveness of the classifier. An example of the ROC curve is shown in Fig. 4.

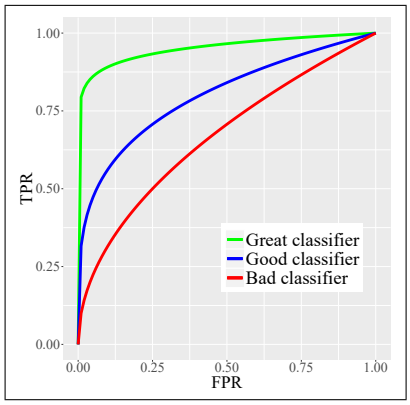

Fig. 4. The ROC curve of three classifiers

The three evaluation criteria we use in this paper are ACC, F1, and AUC. The ROC curve will also be used as an intuitive comparison.

B. Comparisons

Since our method is based on an unsupervised learning technique, we use three other famous unsupervised learning techniques as follows:

- k-means clustering: after the clustering procedure, those points that are far away from their clustering centers are considered abnormal. k is chosen from 5 to 20 with an interval of 5 to demonstrate the effectiveness.
- GMM clustering: a matrix containing the posterior probability that every element belongs to each Gaussian will be obtained. The abnormal degree is calculated according to the matrix. The number of Gaussians is chosen from 5 to 20 with an interval of 5.
- DBSCAN: the classic density-based clustering method will directly output the abnormal data. The neighborhood radius ε is chosen in $\{0.3464, 0.5542, 0.6928\}$ and the density threshold $minPts$ is chosen in $\{3, 5, 10\}$.

IV. Numerical Experiments

To verify that our method can actually find out the electricity theft, we create a dataset which is close to reality and test the method as well as the comparisons.

A. Dataset

Since a real dataset with electricity theft labels is difficult to obtain, we use a synthetic dataset in which the abnormal load profiles are generated from the six malicious types mentioned in [16] and the benign load profiles come from the Irish Smart Energy Trial [17]. The Irish dataset contains the smart meter data of over 5,000 Irish residential and commercial users for 535 days, and the sample rate is 48 S/day. Let $\boldsymbol{x} = \{x_1, \cdots, x_{48}\}$ be the real load profile of a certain customer, the six malicious types are as follows [16]:

- $h_1(x_t) = \alpha x_t$, $\alpha = \text{random}(0.1, 0.8)$;
- $h_2(x_t) = \beta_t x_t$,

$$\beta_t(x_t) = \begin{cases} 0 & \text{for a period of time } t \\ 1 & \text{else} \end{cases};$$

- $h_3(x_t) = \gamma_t x_t$, $\gamma_t = \text{random}(0.1, 0.8)$;
- $h_4(x_t) = \gamma_t \text{mean}(\boldsymbol{x})$, $\gamma_t = \text{random}(0.1, 0.8)$;
- $h_5(x_t) = \text{mean}(\boldsymbol{x})$;
- $h_6(x_t) = x_{49-t}$.

Note that the first type does not change the shape of the load profile, so the performance for the first type is not tested. We choose one-month period load profiles of the 391 small and medium-sized enterprises (SMEs) recorded in the survey of the dataset as our test set. So we have 11,730 load profiles as input. Among the 391 SMEs, 100 certain users are suspected to commit fraud under our assumed scenario. Thus, a part of their corresponding 3,000 user profiles will be processed by the malicious type functions mentioned above. Actually, in type 5 the tampered curve is a straight line, and in type 6 the curve of an SME is almost unchanged. So type 5 and 6 will produce tampered curves with a rather normal shape, which are hard to detect using unsupervised learning methods.

B. Numerical Results

We test all the methods in two cases: the fixed types and the random type. In fixed types, 5 fraud users are randomly chosen in the 100 suspected users. Each fraud user will be assigned to a fixed type. 15 out of the 30 load profiles of a fraud user will be tampered according to his malicious type. For each type 100 scenario are generated to avoid randomness of results and the average performance is evaluated. In the random type, each fraud user will be assigned to a random type from type 2 to type 6, which is more realistic. Also, 100 scenarios are tested.

The value of threshold1 is chosen as the top 5% of the abnormal degrees of each method (i.e. the top 150 of the 3,000 load profiles are recognized as abnormal). The threshold2 is changed from top 0% to 100% so we can plot a ROC for each method. To calculate the ACC and F1 score of each method, threshold2 is fixed at the top 5% (i.e. users who have the top 5% number of abnormal load profiles are classified as electricity theft). The evaluation results of all the methods are shown in Table II. For methods with a series of parameters, we only present the best results. The ROC curves of the fixed types and the random type are shown in Fig. 5 to Fig. 10. We only plot the curve with max AUC for each method. DBSCAN reaches its max AUC when $\varepsilon = 0.6928$ and $minPts = 3$ for all the types, and the parameters are omitted in the figures.

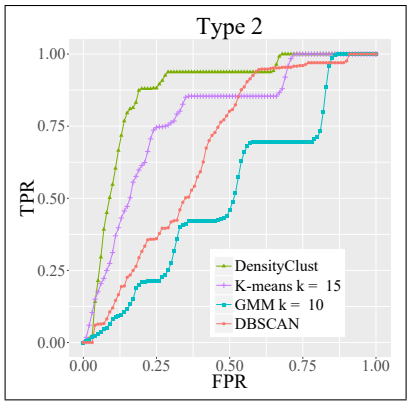

Fig. 5. The ROC curves of type 2

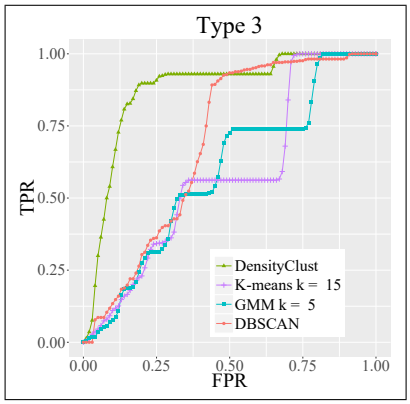

Fig. 6. The ROC curves of type 3

TABLE II

THE EVALUATION RESULTS OF THE METHODS

Type		AUC (densityClust)	AUC (k-means)	AUC (GMM)	AUC (DBSCAN)	ACC (densityClust)	ACC (best of others)	$F1$ (densityClust)	$F1$ (best of others)
Fixed types	2	**0.867**	0.776	0.514	0.662	**0.927**	0.922	**0.274**	0.22
	3	**0.874**	0.576	0.587	0.687	**0.933**	0.911	**0.328**	0.112
	4	**0.962**	0.844	0.629	0.719	**0.995**	0.926	**0.948**	0.262
	5	**0.473**	0.457	0.395	0.369	**0.904**	0.903	**0.044**	0.032
	6	**0.457**	0.456	0.4	0.377	0.903	**0.905**	0.028	**0.052**
Random type		**0.743**	0.667	0.53	0.606	**0.932**	0.92	**0.322**	0.204

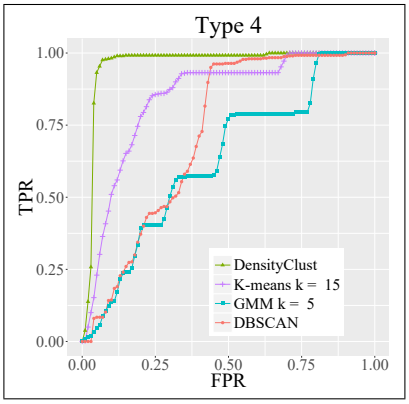

Fig. 7. The ROC curves of type 4

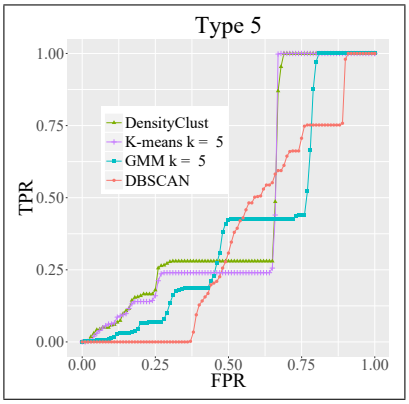

Fig. 8. The ROC curves of type 5

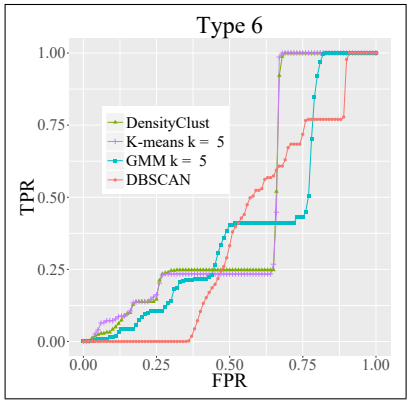

Fig. 9. The ROC curves of type 6

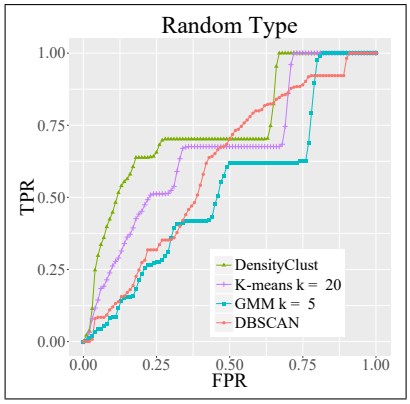

Fig. 10. The ROC curves of the random type

From Table II we can see that our proposed method does a good job in dealing with type 2, 3 and 4. It outperforms the other comparisons in all the evaluation criteria. While dealing with type 5 and 6, the performances of all the methods are bad, because the tampered load profiles of type 5 and 6 still look normal somehow. The results can also be seen from the ROC curves. In Fig. 5, 6 and 7, the proposed method is clearly a great classifier for the certain types. In Fig. 8 and 9, we can see that type 5 and 6 are hard to identify for all methods. As we mentioned above, the abnormal detecting methods cannot directly detect them based on the shape of the load profiles when the tampered curves still have a normal shape.

The ROC curves in Fig. 10. has a horizontal part of it in the middle, which is also due to the fact that the methods are better dealing with type 2, 3, and 4 and almost cannot deal with type 5 and 6.

V. CONCLUSIONS

This paper proposes a density-based abnormal detecting technique for identifying electricity thefts using smart meter data. The abnormal degree of user profiles is calculated according to their distance matrix. To demonstrate the effectiveness of the technique, comparisons with other unsupervised learning methods including k-means clustering, GMM clustering, and DBSCAN are conducted. Results show that the proposed technique can precisely detect electricity thefts based on their abnormal load profiles. For malicious types that do not produce load profiles with an abnormal shape, the technique is helpless.

REFERENCES

[1] J. Liu, Y. Hou, and N. Liu, "Abnormity detection method of intelligent electricity consumption for nontechnical loss," *East China Electric Power*, no. 04, pp. 650–656, 2014, 31-1479/TM.

[2] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security & Privacy*, vol. 7, no. 3, 2009.

[3] Northeast Group, LLC. (2017) Emerging markets smart grid: Outlook 2017. [Online]. Available: http://www.northeast-group.com/reports/Brochure-Emerging%20Markets%20Smart%20Grid%20Outlook%202017%20-%20Northeast%20Group.pdf

[4] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in *International Workshop on Critical Information Infrastructures Security*. Springer, 2009, pp. 176–187.

[5] FEDERAL BUREAU OF INVESTIGATION. (2010) Cyber intelligence section: Smart grid electric meters altered to steal electricity. [Online]. Available: https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/

[6] A. Fragkioudaki, P. Cruz-Romero, A. Gmez-Expsito, J. Biscarri, M. J. de Tellechea, and . Arcos, "Detection of non-technical losses in smart distribution networks: A review," in *Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection*. Springer, 2016, pp. 43–54.

[7] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 946–955, 2008.

[8] 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.

[9] H. Janetzko, F. Stoffel, S. Mittelstdt, and D. A. Keim, "Anomaly detection for visual analytics of power consumption data," *Computers & Graphics*, vol. 38, pp. 27–37, 2014.

[10] X. Liu and P. S. Nielsen, "Regression-based online anomaly detection for smart grid data," *arXiv preprint arXiv:1606.05781*, 2016.

[11] L. A. P. Júnior, C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. Da Costa, and J. a. P. Papa, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016.

[12] T. V. Babu, T. S. Murthy, and B. Sivaiah, "Detecting unusual customer consumption profiles in power distribution systemsapspdcl," in *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–5.

[13] L. Tian and M. Xiang, "Abnormal power consumption analysis based on density-based spatial clustering of applications with noise in power systems," *Automation of Electric Power Systems*, no. 05, pp. 64–70, 2017, 32-1180/TP.

[14] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[15] T. L. Pedersen and S. Hughes, "densityclust: Clustering by fast search and find of density peaks," 2016, r package version 0.2.1.

[16] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[17] Irish Social Science Data Archive, "Data from the Commission for Energy Regulation (CER) - smart metering project," 2012. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/