Fuzzy Modelling for the Detection of Non-Technical Losses Using Time Variant and Invariant Features

Diogo F. Caridade
IDMEC
Instituto Superior Técnico,
Universidade de Lisboa
Lisboa, Portugal
diogo.caridade@tecnico.ulisboa.pt
November 2018

Joaquim L. Viegas
IDMEC
Instituto Superior Técnico,
Universidade de Lisboa
Lisboa, Portugal
joaquim.viegas@tecnico.ulisboa.pt
November 2018

Susana M. Vieira
IDMEC
Instituto Superior Técnico,
Universidade de Lisboa
Lisboa, Portugal
susana.vieira@tecnico.ulisboa.pt
November 2018

Abstract—**Electricity consumption is steadily increasing every year. Although modern appliances and industries are becoming more efficient, improvements in standards of living and population growth contribute to the growing demand for electricity. The increasing need for electricity worldwide has invariably led to an escalation in instances of electricity theft. Over time, this problem has been recognised by utility companies and many solutions have been tested. This thesis proposes the use of fuzzy modelling for data-based detection of non-technical losses. More specifically, a new clustering scheme, Mixed Fuzzy Clustering, is utilized to leverage both time variant and invariant features (gathered from surveys and electricity consumption records) in the identification of illegal consumers. To evaluate the performance of the developed models, three use cases are developed considering different types of features based on collected consumption data and consumer surveys. For comparison purposes, the Fuzzy C-Means algorithm was also used to derive models. Although this algorithm was not specifically developed for dealing with time-variant data, it has proven suitable/effective in many different applications, such as dealing with consumption data for consumer profiling. The best overall classifiers were computed by applying FCM to the dataset with time variant and invariant features, indicating that the Mixed Fuzzy Clustering algorithm is not best suited to deal with the data features used. The models developed achieved a good performance in detection of non-technical losses, quantified by true positive rate of up to 80% under a false positive rate 21.9%, showing that fuzzy modelling is suited for the task.**

Index Terms—**time variant feature, time invariant feature, Fuzzy C-Means, Mixed Fuzzy Clustering**

I. Introduction

Throughout the entire power grid that starts with the production of electrical energy and ends at the moment of consumption there are power losses expected from any physical process. These can be estimated by computing the difference between the amount of electricity produced and the amount consumed by the end user and they are referred to in the literature as transmission and distribution (T&D) losses. Despite the efficiencies of the components used in these operations already being measured and tabled, the losses they cause, Technical Losses (TL), can only be estimated, as the intensity of load demand, load density and energy patterns fluctuate throughout the day and the capability and configuration of the transmission and distribution system vary among installations [3]. By knowing the amount of power a company produces, the electricity that is billed to the consumer (amount of electricity that the company registered as consumed) and the energy lost from TL even if estimated, it is possible to ascertain the existence of external power losses, also known as Non-Technical Losses (NTLs).

The existence of NTLs is well documented throughout the literature [4, 6, 10], and acknowledged by the industry as well [19, 20]. It has a significant impact in several points around the globe, like Jamaica [6], South Africa [4], India [10], among others.

These high costs lead to several negative consequences. Not only are some of the theft procedures dangerous for the ones executing them (risk of electrocution when bypassing the electric connection to a meter or when rigging a line from the power source) [4], but they will ultimately end up costing to law abiding citizens. Since companies produce energy that is not paid for, they end up charging the difference to legitimate consumers. This forces some to steal electricity because they cannot afford to pay the increased prices. The result is a never ending cycle that perpetuates wrong doing from otherwise upstanding individuals.

As prices cannot rise indefinitely, there will be a saturation point where companies will lose money from NTLs and will not be able to get it back. With less money, maintenance and upgrades to current grids will be affected. Consequently, the overload of the components forming the distribution line, due to unmonitored consumption of electricity, will trigger more frequent blackouts (complete loss of power in a region) and brownouts (voltage supplied to the system falls below the specified operating range but there is no total loss of power) [6].

This thesis focuses on creating models for detection of NTLs built from supervised and interpretable techniques using time variant and invariant information simultaneously (mixed data). These models are then compared with ones built solely from static or temporal features, Static Consumption Indicators (SCI) and Temporal Consumption Indicators (TCI).

As most techniques are not designed to handle both types, a specific algorithm, Mixed Fuzzy Clustering (MFC), is used to cluster both data types. To ascertain its performance MFC is compared with Fuzzy C-Means (FCM), which is often used with static data. While all datasets are applied to FCM, only those with a temporal component with used with MFC. Computing FCM with time variant features results in the temporal component of the data to be disregarded. This clarifies the importance of keeping the temporal nature of the data during clustering and modelling by comparing the performance of Takagi-Sugeno (TS) fuzzy models built from MFC and FCM. Better performing models computed using mixed datasets and MFC would prove the importance of both data types in the detection of NTLs. The following figure, Fig. 1 shows how the distribution of algorithms per datasets was made.

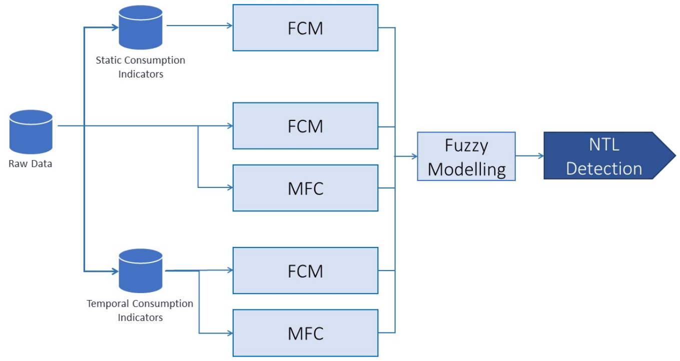
Static Consumption Indicators
Raw Data
Temporal Consumption Indicators
FCM
FCM
MFC
FCM
MFC
Fuzzy Modelling
NTL Detection

Fig. 1. Graph of the methods applied to each dataset

The rest of this document is organized as follows. Section II presents the work that has been done in the field of NTL detection. Section III gives an overview of the theoretical background of MFC, FCM and TS fuzzy modelling. In Section IV, the data processing mechanisms as well as the threat model used are explained. Section V presents the tests done and the respective results and Section VI discusses the final conclusions and future work.

II. Detection of non-technical losses

As mentioned before, NTLs have had noteworthy impact in the industry leading to the acknowledgement of several companies and the attention and interest of researchers. In the last two decades this has culminated in studies focused on techniques and approaches to detect, estimate and analyse NTLs Overall, the research done can be sorted into three distinct categories: theoretical studies, hardware solutions and non-hardware solutions.

Theoretical studies centre on the use of statistical techniques to find the correlation between socio-demographic, economic, market variables and electricity fraud. These techniques have the advantage of producing results that can have a high impact by helping design policy and make decisions to reduce NTLs. However, these same studies can estimate and find the drivers of aggregate NTLs but are not adequate when it comes to finding specific cases of theft and faults in metering or billing.

Hardware solutions primarily focus on characterizing and deploying equipment (e.g. metering hardware and infrastructure, signal generation, sensors) that enables the estimation, detection or even disarming of NTLs. Whether these solutions are to be implemented on a local level (households) or a wider scale (electric grid), the costs associated are generally much higher when compared with the other solutions.

Non-hardware solutions propose computational approaches to detect or estimate the presence of NTLs. Most studies within this field, including this thesis, focus on describing classification techniques that infer the presence of NTLs from electricity consumption or other data. These methods do not carry high costs (unlike hardware solutions) and, if the theoretical information is accurate and extensive, they can detect most types of NTLs. However, the presence of sources of NTLs upon identification is not guaranteed and these methods might not be capable of identifying individual sources. Therefore, these algorithms are better used as an add-on to improve the effectiveness of inspection resources. These types of solutions mainly encompass the following techniques: estimation, game theory and classification.

In [2], a consumption pattern-based energy theft detector is presented which utilizes silhouette plots to distinguish clear distributions in the database and is able to adapt its performance depending on the aim of the application by adjusting the detection delay. One of the main issues within NTLs identification is the data imbalance that is inherent to most experiments done in the field as the number of fraudulent consumers is much smaller than the number of honest customers resulting in standard classifiers to discard the minority class as it is overwhelmed by the remaining one. As it has been stated before, the types of data usually available in the energy research field are static and temporal attributes. Most machine learning techniques focus only on one type, as they are not appropriate when dealing with the complexity of datasets containing different data types. While most studies focus on socio-demographic datasets composed only of static or time invariant features, some, like [11, 12], study the effectiveness of time variant data (electricity consumption over time) in discerning fraudulent from regular behaviour. In [11] the authors propose a novel inspection algorithm which is able to detect malicious meters in one inspection step. In [12], power consumption without the seasonal component is analysed in order to detect fraud and illogical consumption.

Several clustering-based approaches have been proposed to extract fuzzy rules from data, where each cluster represents a region in the product space that contains enough information to support the existence of a fuzzy input/output relationship [13]. The clustering of temporal attributes has seen increased interest as most real life problems are characterized by datasets containing variables that change over time. However, studies on cluster analysis of these features are still limited [14]. So as to introduce a new perspective into the field of NTL detection, a recently proposed method that uses time variant (spatial component) and invariant (temporal component) data as input was chosen as the basis for this work. The algorithm, entitled Mixed Fuzzy Clustering (MFC), applies fuzzy clustering to spatio-temporal data [15] to learn *if-then* rules for identification of TS fuzzy models. MFC introduces a generalization of the spatiotemporal concept to any set of time variant and time invariant features and its extension to the analysis of multiple time-series. This algorithm has been used in medical settings for administration of vasopressors, identification of patients with high risk of mortality and readmissions in intensive care units [5, 1, 7].

This thesis is an extension of the work done in [7] where MFC was used to derive TS fuzzy models and this framework outperformed FCM-based TS fuzzy models and k-Nearest Neighbors classifiers. The present work applies MFC and FCM to datasets of electricity consumption and assess whether or not similar results could be achieved. But while, [7] applied to datasets containing both time variant and invariant features, this thesis also implements datasets containing solely one data type to evaluate if maintaining the temporal component of of time variant attributes is crucial to better performing models.

III. Classification models for detection of Non-Technical Losses

In this chapter the theoretical background of the methodologies used throughout this thesis is presented, more specifically fuzzy clustering and TS fuzzy modelling.

A. Fuzzy clustering

Cluster analysis or clustering is the task of grouping a set of unlabelled objects in such a way that the degree of similarity is higher among objects within the same group or cluster (intra-group similarity) than objects from different clusters (inter-group similarity). It is one of the key tasks in exploratory data mining, and a common technique for statistical data analysis, used in many fields such as machine learning, pattern recognition, among others. This thesis utilizes fuzzy clustering, more specifically, the algorithms Fuzzy C-Means (FCM) and Mixed Fuzzy Clustering (MFC).

1) Fuzzy C-Means: Fuzzy C-Means (FCM) was initially proposed in [16]. This method introduces Fuzzy Logic to the hard clustering framework (K-Means) by allowing each data point $x_j = [x_{j1}, x_{j1}, ..., x_{jR}]$, R being the number of features in the input matrix, to belong to all clusters C by means of a membership degree as opposed to hard clustering where each point can only belong to one cluster.

FCM is an iterative optimization that minimizes the following objective function,

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m d_{ij}^2(x_j, c_i) \tag{1}$$

where N is the number of samples, C the number of clusters, m is the fuzzification parameter which has to range within $[1; \infty[$ and defines how fuzzy or crisp the end clusters will be, u_{ij}^m is the membership degree of element i to the cluster j, and $d_{ij}^2(x_j, c_i)$ is a similarity measure, commonly viewed as a proximity measure and thus handled as a distance variable, between the point x_j and the cluster centre, also known as prototype, c_i.

The algorithm starts with the initialization of the partition matrix U which is formed with the membership degrees of every element N to every cluster C.

$$U = \begin{bmatrix} u_{11} & \dots & u_{1C} \\ \vdots & \ddots & \vdots \\ u_{N1} & \dots & u_{NC} \end{bmatrix}$$

This initialization can be performed randomly by giving aleatory belonging values of each point to every cluster. However, the membership degree has to obey the following constraints

$$u_{ij} \in [0,1] \quad \forall i; \qquad 0 < \sum_{j=1}^{N} u_{ij} < N \quad \forall i; \tag{2}$$

$$\sum_{i=1}^{C} u_{ij} = 1 \quad \forall i. \tag{3}$$

Afterwards, through an iterative process the fuzzy partitioning is carried out and the objective function in 1 is optimized. This is done by computing the cluster centres with the equation

$$c_i = \frac{\sum\limits_{j=1}^{N} u_{ij}^m x_j^s}{\sum_{j=1}^{N} u_{ij}^m} \tag{4}$$

also known as prototypes. These are the mean of all points, weighted by their degree of membership to the cluster. Next the partition matrix is updated using equation (5) which calculates the membership degree of element i to cluster j.

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{C} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}} \tag{5}$$

Once this is done a stopping condition is checked and if cleared, the iteration stops. In this project, the cycle halts when $max|u_{ij}^{z+1} - u_{ij}^{z}| < \epsilon$ is verified where z is the iteration step and ϵ is the stopping condition.

In the present study, each sample is assigned to each cluster with a certain degree of membership. This degree is proportional to the distance between the sample and the cluster prototype, which in a general way can be computed as

$$d_{ij}^2(x_j, c_i) = ||x_j - c_i||^2 = (x_j - c_i)^T A_i (x_j - c_i) \qquad (6)$$

where A_i is a positive definite symmetric matrix, usually equal to the identity matrix in the FCM algorithm.

2) *Mixed Fuzzy Clustering:* Mixed Fuzzy Clustering is a novel clustering method based on Fuzzy C-Means [1] which allows the clustering of time variant (remain constant over time) and time invariant (change over time) features simultaneously. This algorithm aims at providing a solution for dealing with longitudinal misaligned data where the length of time variant features is different, and to account for misalignment through the use of Dynamic Time Warping (DTW) distance. This approach clusters the dataset using an augmented form of the FCM [17]. The main difference between them relies on the distance function. In the augmented version, a new parameter λ is calculated, weighting the importance to be given to each feature [18].

Each sample x_i, with $i = 1, ..., n$, is characterized by static features or time invariant, x^s, and by time variant features or time-series, X^t:

$$x_i = (x_i^s, X_i^t), \qquad (7)$$

where x^s is a NxR matrix with N equal to the number of entities and R equal to the number of time invariant features, and X^t a NxP matrix with P being the number of time variant features. Each entry of X^t is an array of values, x_{ip}^t of length Q dependent on p,

$$x_{ip}^t = (x_{i1}^t, x_{i2}^t, ..., x_{iQ(p)}^t), \qquad (8)$$

The time invariant cluster centres l for feature r, also known as the time invariant prototypes j for feature r, v_{jr}^s, and the time variant cluster centres j, v_{jp}^t for feature p are calculated through equations (9) and (10) respectively.

$$v_{jr}^s = \frac{\sum\limits_{i=1}^{N} u_{ij}^m x_{ir}^s}{\sum_{i=1}^{N} u_{ij}^m} \qquad (9)$$

$$v_{jp}^t = \frac{\sum\limits_{i=1}^{n} u_{ij}^m x_{ip}^t}{\sum_{i=1}^{N} u_{ij}^m} \qquad (10)$$

Each cluster j has its own set of feature weights λ, calculated separately for x^t and x^s in every dimension:

$$\lambda_{jr}^s = \frac{1}{\left(\sum\limits_{1<k\leq R} \frac{\sum\limits_{i=1}^{N} u_{ij}^m ||(x_{ir}^s, v_{jr}^s)||^2}{\sum\limits_{i=1}^{N} u_{ij}^m ||x_{ik}^s - v_{jk}^s||^2} + \sum\limits_{R<k\leq R+P} \frac{\sum\limits_{i=1}^{N} u_{ij}^m ||(x_{ir}^s, v_{jr}^s)}{\sum\limits_{i=1}^{N} u_{ij}^m \delta(x_{i(k-R)}^t, v_{j(k-R)}^t)} \right)^{\frac{1}{q-1}}} \qquad (11)$$

$$\lambda_{jp}^t = \frac{1}{\left(\sum\limits_{1<k\leq R} \frac{\sum\limits_{i=1}^{N} u_{ij}^m \delta(x_{ip}^t, v_{jp}^t)}{\sum\limits_{i=1}^{N} u_{ij}^m ||x_{ik}^s - v_{jk}^s||^2} + \sum\limits_{R<k\leq R+P} \frac{\sum_{i=1}^{N} u_{ij}^m \delta(x_{ip}^t, v_{jp}^t)}{\sum\limits_{i=1}^{N} u_{ij}^m \delta(x_{i(k-R)}^t, v_{j(k-R)}^t)} \right)^{\frac{1}{q-1}}} \qquad (12)$$

The variable q offers a degree of feature discrimination and its value ranges from 1 to ∞. According to [18], as q approaches 1, λ will tend to take binary values, meaning one feature will be labelled 1 for being the most relevant in the computation of the distance between samples and the prototypes. On the opposite end of the spectrum, was q to take values approaching infinity, the same feature weights will have the same levels of relevancy thus making the process of feature selection irrelevant.

As previously mentioned, λ alongside a distance function δ was used to compute the distance between an entity and the time invariant and variant prototypes of a cluster j, through equation (13).

$$d_{ji}^2 = \sum_{r=1}^{R} \lambda_{jr}^s ||x_{ir}^s - v_{jr}^s||^2 + \sum_{p=1}^{P} \lambda_{jp}^t \delta^2(x_{ip}^t, v_{jp}^t) \qquad (13)$$

The distance function δ between two vectors a and b of same length M, with $i = 1, 2, ..., M$ and $l = 1, 2, ..., M$, is given by:

$$\begin{cases} \boldsymbol{\delta}^2(\mathbf{a}, \mathbf{b}) = (a_1 - b_1)^2 + ... + (a_M - b_M)^2, & \text{if EUC} \\ \boldsymbol{\delta}^2(\mathbf{a}, \mathbf{b}) = \gamma(M, M), & \text{if DTW} \end{cases} \qquad (14)$$

where $\gamma(i, l) = ||a_i - b_l||^2 + min\{\gamma(i-1, l), \gamma(i-1, l-1), \gamma(i, l-1)\}$ and $\gamma(1, 1) = ||a_1 - b_1||^2$.

Distance Time Warping (DTW) fundamentally differs from the former on how it handles two time series. While Euclidean distance directly compares two points in the same time instance without considering differences in vector size (either from different time samples or different start/end recording) or misalignment between them, DTW takes both vectors and tries to align them so as to minimize their difference thus creating what is called a *warping path*.

Both distance measures are depicted in Fig.2

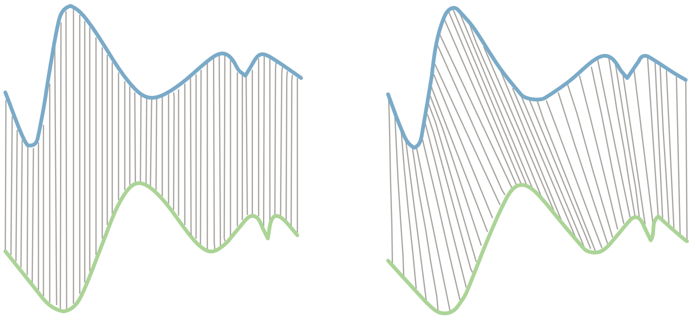

Fig. 2. Graphical representation of the difference between Euclidean distance (on the left) and Dynamic Time Warping (on the right)

One of the inputs to the Mixed Fuzzy Clustering algorithm is the partition matrix $U = [u_{ij}]$ which is built from the degree of membership of each sample i to each and every cluster j. Given N samples and C number of clusters, the matrix will be CxN and each entry is calculated using the equation (15)

$$u_{ij} = \frac{1}{\sum_{g=1}^{C} \left(\frac{d_{ji}^2}{d_{gi}^2} \right)^{\frac{1}{m-1}}} \tag{15}$$

and subject to the following constraints

$$u_{ij} \in [0,1] \quad \forall i; \qquad 0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall i,j; \tag{16}$$

$$\sum_{j=1}^{C} u_{ij} = 1 \quad \forall j; \tag{17}$$

As previously mentioned, this is an augmented version of FCM and as such applies an augmented form of the objective function

$$J = \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^m d_{ji}^2 (v_j^s, v_{jp}^t, x_i) \tag{18}$$

B. Takagi-Sugeno model for detection of NTLs

Takagi-Sugeno fuzzy models are transparent “grey box” that allow the approximation of previously unknown non-linear systems to be modelled using a number of linear and understandable sub-models responsible for distinct sub-domains. Fuzzy models use a training set in order to discover potentially predictive relationships between inputs and outputs, and a test set (invariably smaller than the training set) to validate said relationships. For the binary classification case, each discriminant function consists of rules,

$$\begin{aligned} R_j &: \text{If } x_1 \text{ is } A_{j1} \text{ and ... and } x_M \text{ is } A_{jM} \\ &\text{then } y_j(x) = f_j(x), j = 1, 2, \text{...} \; K \end{aligned} \tag{19}$$

where f_j is the consequent function of rule R_j and M = r+q, the number of features used. The output of the discriminant function $y_j(x)$ can be interpreted as a score (or evidence) for the positive example given the input feature vector x.

The number of rules K of the type R_i and the antecedent fuzzy sets A_{jh} are determined by fuzzy clustering in the product space of the input variables. The consequent functions $f_i(x)$ are linear functions determined by ordinary-least squares (OLS) in the space of the input and output variables.

The degree of activation of the jth rule is given by

$$\beta_j = \prod_{h=1}^{M} \mu_{A_{jh}}(x), \tag{20}$$

where $\mu_{A_{jh}}(x) : \mathbb{R} \to [0;1]$.

In TS models, the overall output is a weighted average of individual rule outputs, β_j being the weight, and the inference is reduced to a simple algebraic expression:

$$y(x) = \frac{\sum_{j=1}^{K} \beta_i f_i(x)}{\sum_{j=1}^{K} \beta_i} \tag{21}$$

The output of the system is often continuous but since it needs to identify classes (in this case two groups, theft and regular consumption) a threshold has to be established. For that, the model uses the training data to set a threshold that creates the most accurate classifier for that specific set of data points. This particular model (with the threshold previously set) is then used with the test set to evaluate its performance.

A sample x_j is considered positive if the score is higher than a certain threshold γ, in other words, $y_j(x_j) > \gamma$.

IV. Data and threat model

A. Databases used

All methods studied in this project used data from the same source. The databases were provided by the Commission for Energy Regulation who conducted surveys and collected data on the electricity consumption of around 5,000 Irish homes and businesses from the Summer of 2009 to the Winter of 2010 and made them available at the Irish Social Science Data Archive (ISSDA). The survey information (number of adults and children, type of house, heating provided through gas, ...) will be referred to as static information or time invariant features while the consumption patterns will be referred to as time series or time variant features. After applying all the preprocessing methods described in the following sections, the number of samples available was 4233.

1) Electricity consumption data: The data gathered from the meters presented many missing entries. In cases where the percentage of missing data of a certain meter exceeded a certain threshold during the two year recording, the information of said meter was completely discarded. In the remaining instances, in order to fill the gaps in the records, Zero Order Hold (ZOH) was used. This method replaces missing values with the value immediately preceding it. Implementing ZOH allowed for the consumption patterns to have the same number of entries.

Despite the fact that MFC was designed to deal with misaligned time series by using DTW distance [5], the computational power needed to compute DTW over Euclidean distance is greatly exacerbated leading to an increase of the time needed to run the experiments. For this reason, Euclidean distance was chosen over DTW distance.

In order to keep confidentiality in check, the data was resampled to 24 points per day, from the initial 48, by adding every two points of the time series. After analysing the consumption data over a weekly period, the difference between patterns from weekends and working days was substantial enough to justify discarding the weekend information from the database. So as to not overload the algorithms with information 18 days from the two year period were picked to apply the threat model. On each day, a time window was applied so that the 5 days prior to the chosen day would also be saved for the analysis.

2) Survey data: With respect to the surveys, these consisted of 234 questions made to both private and corporate consumers. However certain questions were directed specifically to one of these groups. This, allied with a possible relative weight between the questions to better characterize a consumer, justified the use of Feature Selection. Applying a rating from 0 (completely irrelevant for consumer profiling) to 3 (of the highest importance) to all 234 questions, choosing the ones with score above 1 and removing the ones that showed a skewed distribution in the answers (questions where over 85% gave the same answer) left 39 variables eligible to be used for the study.

B. Feature engineering

The following sections explain how the datasets SCI and TCI were built.

1) Static consumption indicators: Rather than just keeping the same characterizing variables from the original set of time variant and invariant features, indicators were computed to make future models as transparent and interpretable as possible. These new features reflect changes from past consumption patterns and from other consumers with similar characteristics when comparing answers to the survey. The computed dataset presents the following information:

- Meter : Serial number that identifies each meter
- Date : The date in which the data was recorded
- Attack : The type of attack each sample is meant to represent ($None$ for normal consumption and h_i for the attack i)
- I_1 : Indicator of consumption variation. Ratio between current and past consumption;
- I_2^e and I_2^c : Indicators of hourly consumption pattern change.
- I_3 : Indicator of consumption difference in comparison to consumers with similar characteristics.
- I_4^e and I_4^c : Indicators of hourly consumption pattern difference in comparison to consumers with similar characteristics

This dataset displayed null values, NaN, in certain indicators, namely I_c^2 and I_c^4. Three approaches to this problem were considered: turn all values of NaN into 0; turn all NaN into the average value of the corresponding feature; turn only the NaN entries that characterize the attack models 5 and 50 into 0 and the rest into the average values. In Section V-A these approaches were evaluated to understand which yielded the best results.

2) Temporal consumption indicators (Temporal CI): The third dataset was a combination of the last two, it had indicators computed from raw data but displaying a temporal evolution as opposed to being static. This scheme used the principle behind the SCI (developing indicators of electricity consumption instead of feeding raw information directly into the clustering algorithms) but still contained the temporal nature that is the focal point of this thesis. As a result, besides the date of each registry and the meter and attack identification, it had the temporal evolution of the same 4 indicators presented in the first database, I_1, I_2^e and I_2^c, I_3 and I_4^e and I_4^c. The recorded evolution spanned from 9 days prior to the selected day (day of the attack in the case of the non zero-day threat model) to the day itself. The dataset used the threat model with 6 attacks with both zero-day and non zero-day scenarios for 1000 meters.

C. Threat model

Despite the different approaches to the input data, the same attack vectors were computed onto the data points. For this project, the provided information by ISSDA was treated as data from legitimate consumers and was used to apply a threat model which represents distinct ways a consumer might steal from electricity providers. The idea behind the application of these attack vectors is to create a flexible model with information regarding the same consumer interfering with his records through a variety of methods all the while keeping the unaltered consumption pattern for theft identification. The threat model used was taken from the work [9] which in turn was based off of [2].

According to [9], the attack procedures can be grouped in two different categories: attacks that started during the meters data gatheringz (non zero-day attacks) and attacks which had already started by the time the information was collected (zero-day attacks). The different attacks and scenarios (zero-day and non zero-day) are presented below. The equations use the following notation: m$_i^{d,t}$ are the meter consumption readings from consumer i in day d for hour t. This results in m$_i^d$ = $(m_i^{d,1}, m_i^{d,2}, ..., m_i^{d,24})$ representing the 24 hour vector of metered data of consumer i on day d.

- h_1; h_{10} : constant random reduction of consumption.

$$h_1(m_i^{d,t}) = \alpha m_i^{d,t} \quad , \qquad \alpha = random(0.1, 0.8) \tag{22}$$

- h_2; h_{20} : registering zero consumption for a random period of the day; zero day scenario.

$$\begin{aligned} & h_2(m_i^{d,t}) = \beta^h m_i^{d,t} \\ & \beta_t = \begin{cases} 0, & t_{start} < t < t_{end} \\ 1, & \text{else} \end{cases} \\ & t_{start} = random(0, 19) \\ & \delta = random(4, 24) \\ & t_{end} = t_{start} + \delta \end{aligned} \tag{23}$$

- h_3; h_{30} : random hourly reduction of consumption; zero day scenario.

$$h_3(m_i^{d,t}) = \gamma_t m_i^{d,t} \quad , \qquad \gamma_t = random(0.1, 0.8) \tag{24}$$

- h_4; h_{40} : random hourly consumption pattern with reduced average consumption; zero day scenario.

$$h_4(m_i^{d,t}) = \gamma_t \; \mu(m_i^d) \quad , \qquad \gamma_t = random(0.1, 0.8) \tag{25}$$

- h_5; h_{50} : constant hourly consumption equal to the average; zero day scenario.
  The entire daily consumption pattern is replaced by the average of the electricity usage in that same day.

$$h_5(m_i^{d,t}) = \mu(m_i^d) \tag{26}$$

- h_6; h_{60} : reversed hourly consumption; zero day scenario.

$$h_6(m_i^{d,t}) = m_i^{d,24-t} \tag{27}$$

- h_7; h_{70} : shift of consumption from peak hours to the rest of the day, turning peak hours consumption to the average; zero day scenario.

$$\begin{aligned} h_7(m_i^{d,t}) &= \begin{cases} m_i^{d,t} - \delta\, m_i^{d,t}, & p_{start} < t < p_{end} \\ m_i^{d,t} + \epsilon\, /21, & \text{else} \end{cases} \\ p_{start} &\text{ is the starting hour of the highest consumption three hour period} \\ p_{end} &= p_{start} + 3 \\ \epsilon &= \sum_{j=1}^{3} m_i^{d,p_{start}+j-1} \end{aligned} \tag{28}$$

- h_8; h_{80} : shift the consumption data to one of a legitimate consumer with lower average of electricity consumption; zero day scenario.

$$\begin{aligned} &h_8(m_i^{d,t}) = m_r^{d,t} \\ &r \text{ is a random consumer with } \mu(m_r^{d,t}) < \mu(m_i^{d,t}) \end{aligned} \tag{29}$$

V. Results

The simulations done on the three databases were evaluated using the following criteria: Area Under the Receiver Operating Characteristics Curve (AUC); Accuracy; True Positive Rate (TPR); False Positive Rate (FPR); Difference between TPR and FPR. More specifically, for the dataset of static indicators, the threshold determination method was studied using indexes such as Youden Index, minimum distance and testing every possible threshold until one yields the lowest difference between specificity and sensitivity. The Fig. 3 shows the difference between Youden Index, minimum distance and the meaning behind the values they take.

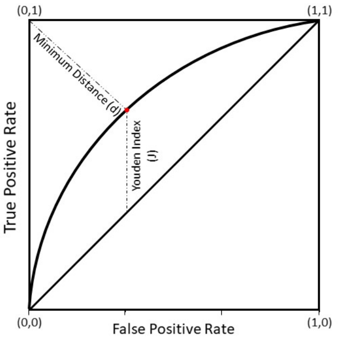

Fig. 3. Graphical representation of the performance evaluation methods.

A. Static Consumption Indicators

There were four main parameters to be chosen in the tests done to this database: optimal number of clusters, C, and fuzzification parameter, m, skewed class distribution, best method to deal with invalid entries in the SCI dataset.

To determine which values of C and m computed the best classifier, a grid search was performed where the tested values were $[2:19]$ and $[1.4:0.3:3.8]$ for C and m respectively.

Since the threat model was applied to each consumer, the end result was 16 samples of attack to 1 sample of unaltered consumption. This yields a distinct disparity between the positive and negative cases of fraud to train the algorithms. To understand the impact of this fact, different ratio were tested to understand which resulted in better performing models. Besides the original set with 5% (1 regular sample to 16 fraudulent ones), 20%, 36% and 47% were also tested

To correct the invalid data in variables I_c^2 and I_c^4 three methods were tested: turn all values of NaN into 0; turn all NaN into the average value of the corresponding feature; turn only the NaN entries that characterize the attack models 5 and 50 into 0 and the remaining NaN into the average values. With the intention of simplifying the document these techniques will be referred to as Method 1, Method 2 and Method 3, respectively.

Lastly, to determine which method (Youden Index, minimum distance or testing all possible thresholds) would result in the best classifier, all three were applied to each pair of Method/Balance and the accuracy, TPR and FPR were analysed and compared.

The results from these experiments showed that low values of m and high values of C yielded the best classifiers. Considering the results and the need to choose a small number of clusters (the higher the value the higher the computational power need to compute the respective model), $C = 9$ and $m = 1.4$ were the best combination, as seen from the Table I which summarises the entire grid search test.

TABLE I
Table with best results from the search grid done on SCI

Parameters Modelling	AUC	Accuracy	TPR	FPR	TPR-FPR
[11 1,7]	0,783	0,727	0,712	0,247	0,465
[9 1,4]	**0,786**	**0,725**	**0,722**	**0,268**	**0,454**
[10 1,7]	0,787	0,729	0,716	0,247	0,469
[11 1,4]	0,799	0,737	0,720	0,235	0,486

The model with the parameters $C = 9$ and $m = 1.4$ was then used to evaluate balancing ratios, NaN solutions and threshold specification approaches. Results showed that both the original dataset and the one with 36% benign data consistently yielded the best results. Since there is an increase in computational power needed to run larger datasets, applying no balancing technique was chosen as the better option. As for the NaN solution, methods 1 and 3 tied for the most effective but the difference with Method 2 was negligible. With respect to threshold calculation, all three methods consistently found similar points in the Receiver Operating Characteristic (ROC) curve, as seen in the example of Fig. 4. Since computing all thresholds to get the best classifier implied an increase in costs when compared to the other two methods (increased computational power) and Youden Index computed slightly less accurate models, minimum distance was chosen as the standard method to determine the threshold.

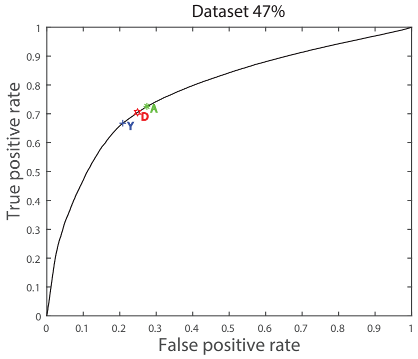

Fig. 4. ROC curve for dataset with a 47% distribution and Method 3 applied

With all of this said, the best combination of parameters was Method 1 applied to a database with no redistribution procedure implemented and with the threshold being computed through the use of minimum distance. This combination resulted in a model with $AUC = 0,79$. In the next two datasets, minimum distance will be used by default as the threshold study proved no significant difference between the three methods.

B. Time variant and invariant features

For this dataset several tests and approaches were taken to assess how MFC varied its outcome and which framework better suited the algorithm. At the end, FCM was also used to compare with the results gathered from SCI. Throughout these tests, the dataset is consistently divided into two groups, one with 200 meters and the other with 1000. This is done to determine the impact of different sample sizes in the clustering and modelling processes. Throughout these experiments, these two datasets will be referred to as 200M and 1000M. Every test performed grid search on C, m and q.

1) Experimental analysis of fuzzy model parameters, $C = 2$: The first experimental test consisted in evaluating the effects of varying m and q on the models' performance. For this test C was kept at 2. The original dataset was divided into *All Attacks*, *Attack 5 / Attack 50* and *Attack 50*. As the names imply, the last two frameworks only had attacks 5 and 50 along with the legitimate sample so as to simplify the clustering algorithm even further since these two attacks would, in theory, be the easiest to identify (they turned the normally irregular consumption pattern into a constant one equal to the daily average consumption).

The results showed *Attack 50* on the 200M dataset to yield better classifiers than the other configurations with $AUC = 0,644$. However this was less than what transpired from using FCM on SCI.

2) Balance data in preprocessing: Similarly to what had been done with SCI, balance ratios were also tested with this dataset. Two ratios were used, 3 regular samples and 6 regular samples in total over the 16 attacks. The attack vectors were divided into three sets, *Zero-Day*, *Non Zero-Day* and *All Attacks*.

Increasing the regular to illegitimate sample ratio did not improve the results significantly. Although there was a negligible increase in AUC, this was not much higher than the standard deviation and therefore can be thought of as fluctuations in the final results rather than an actual increase in performance. However, a distinct difference between the three attack vectors was found where *Zero-Day* consistently outperformed *All Attacks* and especially *Non Zero-Day*. This could be explained by the fact that *Zero-Day* applied the attacks from the first day of recording, providing more information for the algorithm to cluster different patterns. Since *All Attacks* is a combination of the other two it is logical that it would present values of AUC averaging both extremes. It is expected that this difference in threat models will remain during the following tests.

3) Threat model experiments: The threat model proposed in [2] was then used to understand if introducing less information into the system would yield better results. Aside from this, a new balancing strategy was adopted where for each regular sample only one attack would be computed, 50% ratio. The datasets were also split into *Zero-Day*, *Non Zero-Day* and *All Attacks*.

In general, *Non Zero-Day* continued to yield the worst models as opposed to *Zero-Day*. The imbalanced sets achieved better results using 200 meters instead of 1000, where the difference in AUC reached $0,07$ for *Zero-Day*. The balanced set, tested only for 1000M, yielded better results when compared to the imbalanced counter part. The attack vector containing all approaches continued to showed similar values of AUC as in the last test.

4) Practical case scenario: Since it is common for fraudulent consumers to initiate their behaviour with new contracts with utility providers, companies do not have much information on them. Clustering with 18 time series did not accurately represented that reality. For that reason, the number of time series was reduced to 1 and both iterations of the threat model (12 attacks and 16 attacks) were also used for comparison.

Across the board, the results worsened. Aside from *Non Zero-Day* which maintained its poor results, both *Zero-Day* and *All Attacks* had their AUC lowered by $0,1$. Comparing both threat models, although the AUC had similar values, the accuracy differed, increased accuracy for the threat model with 12 attacks. As for the size of the dataset, 1000M showed better results even if by a small margin, as seen in Table II. In order to have a compact table, Attack Vectors and Threat Model are abbreviated to AV and TM, respectively.

5) Modelling using FCM from raw datasets: The need to compare the effects of using indicators as opposed to the original variables and the unsatisfactory results from using MFC led to the implementation of FCM to this dataset. However, applying FCM to these datasets meant that the temporal component present in the data would be lost. Each

TABLE II
TABLE WITH THE BEST MODELS FROM 200M APPLIED TO THE 1000M DATASET USING ONLY ONE TIME SERIES

Dataset	Attacks	TM	AUC	Accuracy	TPR-FPR
200M	Zero Day	6 AV	0,558	0,607	0,111
		8 AV	0,535	0,298	0,055
	Non Zero Day	6 AV	0,519	0,457	0,033
		8 AV	0,511	0,124	0,015
	All attacks	**6 AV**	**0,529**	**0,731**	**0,052**
		8 AV	0,527	0,720	0,015
1000M	Zero Day	6 AV	0,548	0,854	0,039
		8 AV	0,541	0,733	0,052
	Non Zero Day	6 AV	0,511	0,370	0,023
		8 AV	0,505	0,144	0,004
	All attacks	6 AV	0,530	0,568	0,057
		8 AV	**0,555**	**0,421**	**0,110**

entry of the time series would be regarded by the clustering algorithm as a value for another time invariant feature. Given that no significant change came from using the 12 attack vector, the threat model used had the original 16.

Going back to the tests done on Static CI, using FCM on raw data yielded similar results, as seen in Table III. In comparison with MFC, FCM produced much better classifying models. While MFC would repeatedly create worse models when using higher volume of data points, this did not happen when using FCM. With 1000M, models showed better values across all criteria, having the biggest difference being registered in AUC. Another point of contrast between the two methodologies is the performance improvement of the *Non Zero-Day* scenario with respect to the other two. Unlike before, this scenario proved to yield accurate classifiers using both set of meters (the best results out of the three scenarios for 200M). Not only did it have better AUC but the remaining metrics were also higher.

TABLE III
TABLE WITH THE BEST MODELS WHEN APPLYING FCM TO RAW DATA OF 200 AND 1000 METERS

Dataset	Attacks	AUC	Accuracy	TPR-FPR
200M	Zero Day	0,771	0,778	0,444
		0,795	0,785	0,473
	Non Zero Day	0,803	0,786	0,431
		0,804	0,758	0,460
	All attacks	0,747	0,760	0,465
		0,747	**0,760**	**0,465**
1000M	Zero Day	0,861	0,799	0,570
		0,871	0,798	0,582
	Non Zero Day	0,863	0,806	0,576
		0,863	0,806	0,577
	All attacks	0,842	0,792	0,538
		0,844	**0,791**	**0,539**

C. Temporal consumption indicators

Similarly to the datasets in the previous section, TCI also contain a temporal component making it eligible for MFC. And as a way to compare with the first dataset, SCI, and evaluating the impact of the temporal component in this particular dataset, FCM was also applied.

1) MFC with the Temporal Consumption Indicators: The use of temporal indicators, despite also possessing a time variant nature, proved to accelerate the simulation time when compared to using original data directly.

Aside from *Non Zero-Day* which showed the worse results so far, applying MFC to temporal indicators proved to yield slightly better results than using time variant and invariant features with AUC reaching values of $0,680$ in the best case recorded. The only reason the accuracy of *Non Zero-Day* was so high relied on the fact that these models classified every single sample as ”fraudulent” which lead to 16 of the 17 samples per meter being correctly classified.

2) FCM with the Temporal Consumption Indicators: To serve as a base for comparison not only with the last test but also with the other experiments done using FCM, TCI dataset was also used to create FCM fuzzy models. This particular experiment is relevant since it will provide a means of comparison between static and temporal indicators and, overall, if in fact there is an advantage in keeping the time variant nature of certain features.

Unlike what happened when this algorithm was applied to time variant and invariant features, the results this time did not show good performances across all scenarios. In fact, *Non Zero-Day* attacks had classification scores identical to the ones where MFC was used on TCI. As a consequence, *All Attacks* evaluation values worsened as it combines *Zero-Day* attacks which computed well performing models and *Non Zero-Day* attacks which computed very bad ones.

D. Overall results

Table IV presents a summary of the results gathered from the tests made throughout the thesis. Time series is abbreviated to TS in the table. In each dataset, the *All Attacks* threat model is highlighted as it represents all possible attack vectors.

From this global perspective, the difference in performances becomes clear when using MFC or FCM to cluster data points before applying Takagi-Sugeno modelling. Aside from some outliers, using Fuzzy C-Means, even in data containing a temporal component, results in more proficient classifying models. It is then plausible to conclude that maintaining the temporal nature of time variant features at the clustering stage does not directly translate into more accurate models.

VI. CONCLUSIONS AND FUTURE WORK

The goal of this thesis was to identify electricity theft not only through static features (demographic information) but also through time variant features (power consumption patterns). This was done by assessing the performance of classifiers built with algorithms that took into consideration

TABLE IV
SUMMARISED TABLE

Dataset	Model	Parameters		AUC	TPR-FPR
SCI	FCM FM	Method 1 No Balance		0,790	0,469
Raw Data	MFC FM	Zero Day	18 TS	0,725	0,323
			1 TS	0,558	0,111
		Non Zero Day	18 TS	0,543	0,075
			1 TS	0,519	0,033
		All Attacks	**18 TS**	**0,620**	**0,189**
			1 TS	0,555	0,110
	FCM FM	Zero Day	1 TS	0,871	0,581
		Non Zero Day	1 TS	0,863	0,577
		All Attacks	**1 TS**	**0,844**	**0,565**
TCI	MFC FM	Zero Day		0,646	0,206
		Non Zero Day		0,500	0,000
		All Attacks		**0,680**	**0,251**
	FCM FM	Zero Day		0,867	0,594
		Non Zero Day		0,500	0,000
		All Attacks		**0,692**	**0,280**

the time varying component of crucial information, such as electricity consumption of households over a period of time.

Across the three datasets, the best results came from applying FCM to time variant and invariant features. Keeping the temporal nature of the data might not be necessary for high performing classifiers and instead doing so may hinder their performance as it might overload the algorithms with unnecessary information.

Since MFC was designed to operate with DTW distance and this approach was not used in this thesis, applying it to future works in this field might yield better results. The algorithm SVM has proven to compute accurate classifiers, so using it in a similar framework to what was used in this thesis might result in better classifiers than the ones found.

REFERENCES

[1] M. C. Ferreira et al. "Fuzzy modelling based on Mixed Fuzzy Clustering for health care applications". In: (2015), pp. 1–5.

[2] P. Jokar, N. Arianpoo, and V. Leung. "Electricity Theft Detection in AMI Using Customers' Consumption Patterns". In: *IEEE Transactions on Smart Grid* 7.1 (2015), pp. 216–226.

[3] M. Mahmood et al. "Real Time Study on Technical Losses in Distribution System". In: *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 3 (2014), pp. 131–137.

[4] T. B. Smith. "Electricity theft: a comparative analysis". In: *Energy policy* 32.18 (2004), pp. 2067–2076.

[5] C. M. Salgado, M. C. Ferreira, and S. M. Vieira. "Mixed Fuzzy Clustering for Misaligned Time Series". In: *IEEE Transactions on Fuzzy Systems* 25.6 (2017), pp. 1777–1794.

[6] F. B. Lewis. "Costly 'Throw-Ups': Electricity Theft and Power Disruptions". In: *The Electricity Journal* 28.7 (2015), pp. 118 –135.

[7] C. M. Salgado et al. "Takagi–Sugeno fuzzy modeling using mixed fuzzy clustering". In: *IEEE Transactions on Fuzzy Systems* 25.6 (2017), pp. 1417–1429.

[8] J. L. Viegas and S. M. Vieira. "Clustering-based novelty detection to uncover electricity theft". In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (2017), pp. 1–6.

[9] J. L. Viegas, P. R. Esteves, and S. M. Vieira. "Clustering-based novelty detection for identification of non-technical losses". In: *International Journal of Electrical Power & Energy Systems* 101 (2018), pp. 301–310.

[10] J. P. Navani, N. K. Sharma, and S. Sapra. "Technical and Non-Technical Losses in Power System and Its Economic Consequence in Indian Economy". In: *International Journal of Electronics and Computer Science Engineering* 1.2 (2012), pp. 757–761.

[11] X. Xia et al. "BCGI: a fast approach to detect malicious meters in neighborhood area smart grid". In: *2015 IEEE International Conference on Communications*. 2015, pp. 7228–7233.

[12] J. V. Spirić, M. B. Dočić, and S. S. Stanković. "Fraud detection in registered electricity time series". In: *International Journal of Electrical Power & Energy Systems* 71.C (2015), pp. 42–50.

[13] R. Babuška and H. B. Verbruggen. "Constructing fuzzy models by product space clustering". In: *Fuzzy model identification*. 1997, pp. 53–90.

[14] T. W. Liao. "Clustering of time series data — a survey". In: *Pattern Recognition* 38.11 (2005), pp. 1857–1874.

[15] S. Kisilevich et al. "Spatio-temporal clustering". In: *Data mining and knowledge discovery handbook*. 2009, pp. 855–874.

[16] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57.

[17] H. Izakian, W. Pedrycz, and I. Jamal. "Clustering spatiotemporal data: An augmented fuzzy c-means". In: *IEEE Transactions on Fuzzy Systems* 21.5 (2013), pp. 855–868.

[18] H. Frigui and O. Nasraoui. "Unsupervised learning of prototypes and attribute weights". In: *Pattern recognition* 37.3 (2004), pp. 567–581.

[19] Siemens. *More revenue by fighting non-technical losses*. URL: https://www.siemens.com/customer-magazine/en/home/energy/agility-in-energy/more-revenue-by-fighting-non-technical-losses.html.

[20] Forbes. *Electricity Theft: A Bigger Issue Than You Think*. URL: https://www.forbes.com/sites/peterdetwiler/2013/04/23/electricity-theft-a-bigger-issue-than-you-think/#56a6c0425ed7x.