Detection of Non-Technical Losses in Smart Meter Data based on Load Curve Profiling and Time Series Analysis

Esther Villar-Rodriguez[a], Javier Del Ser[a,b,c,*], Izaskun Oregi[a], Miren Nekane Bilbao[b], and Sergio Gil-Lopez[a]

[a] *TECNALIA, 48160 Derio, Bizkaia, Spain.*
[b] *University of the Basque Country (EHU/UPV), 48013 Bilbao, Bizkaia, Spain.*
[c] *Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain.*

Abstract

The advent and progressive deployment of the so-called Smart Grid has unleashed a profitable portfolio of new possibilities for an efficient management of the low-voltage distribution network supported by the introduction of information and communication technologies to exploit its digitalization. Among all such possibilities this work focuses on the detection of anomalous energy consumption traces: disregarding whether they are due to malfunctioning metering equipment or fraudulent purposes, strong efforts are invested by utilities to detect such outlying events and address them to optimize the power distribution and avoid significant income costs. In this context this manuscript introduce a novel algorithmic approach for the identification of consumption outliers in Smart Grids that relies on concepts from probabilistic data mining and time series analysis. A key ingredient of the proposed technique is its ability to accommodate time irregularities – shifts and warps – in the consumption habits of the user by concentrating on the shape of the consumption rather than on its temporal properties. Simulation results over real data from a Spanish utility are presented and discussed, from where it is concluded that the proposed approach excels at detecting different outlier cases emulated on the aforementioned consumption traces.

Keywords: Smart Grids; Smart Meter Data; Non-Technical Losses; Outlier Detection.

*Corresponding author: javier.delser@tecnalia.com (Prof. Dr. Javier Del Ser). TECNALIA. P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain. Tl: +34 946 430 50. Fax: +34 901 760 009. E-mail: javier.delser@tecnalia.com.

1. Introduction

According to the official definition introduced by the Energy Independence and Security Act of 2007 [1], Smart Grids can be understood in the wide sense as the technological efforts to modernize and digitize the electricity distribution system of a nation to ensure, in an scalable manner, the improved reliability, security and efficiency of the grid, as well as to guarantee an optimized management of its resources and operation. This definition describes such a term as a comprehensive set of operational resources established to guarantee an efficient power transmission and electricity distribution, paying an special attention to reliability and security.

The advent and progressive deployment of the so-called Smart Grid has unleashed a profitable portfolio of new possibilities for an efficient management of the low and medium voltage distribution network, supported by the introduction of information and communication technologies to exploit its digitalization. In this context, the deployment of the Advanced Metering Infrastructures (AMIs) allows the utilities to acquire fine-grained data about the real consumption of end-users (not based on estimations or monthly measurements), which results essential to acquire deeper insights on how, when and where energy is distributed and consumed through the network [2, 3, 4]. This fact is particularly crucial in regards to the traceability and characterization of electrical losses, which account for the difference between the amount of energy distributed by the electrical distribution company and the amount of energy paid by the consumers. Such losses may be due to two main contributing causes: 1) losses inherent to the transformation and distribution of energy, which are proportional to the squared of electrical current and widely referred to as Technical Losses (TL); and 2) non-technical losses (NTLs), associated to erroneous readings, defected smart meters or fraud [5].

This work focuses on the detection of energy consumption traces which contribute to NTLs: disregarding whether they are due to malfunctioning metering equipment or fraudulent purposes. As mentioned by [6, 7, 8] the amount of energy loss in the distribution grids varies between 7 – 50 % of the total delivered energy (depending of the country and the characteristics of the distribution network), which undoubtedly justifies the strong efforts that utilities are investing towards detecting and inspecting atypical consumption traces to ultimately avoid significant economical losses. As stated in [9], only in US between 1 and 10 billion worth of electricity was stolen in the late 90s, showing an increment between 5-10% in the last two decades with a remark-

able 40% and beyond in Southeast Asia [10]. In addition, the identification of atypicalities provides further profitable advantages beyond fraud assessment: by properly characterizing the statistics of the consumption traces registered over the power grid, the power distribution can be optimized by matching generation to consumption, thereby avoiding network under-dimensioning and electrical surge.

Interestingly for the scope of this work, energy theft accounts for the majority of reasons for the aforementioned non-technical losses. There are indeed very diverse methods by which malicious consumers reduce illegally the consumption monitored by the installed smart equipment, particularly in the last stage of the distribution network. One of the most usual forms of electricity theft is fraud, by which the user deliberately attempts at deceiving the energy supplier (utility) at hand. This can be achieved by diverse means such as meter tampering, by which the meter is forced to register a lower power reading than the real consumption of the user. While other forms of electricity theft prevail across different countries and cultures (e.g. billing irregularities), this work revolves around those fraudulent cases when the non-technical loss may be reflected in a behavioral change of the energy consumption trace registered by the metering device. In this regard, both tampering and electricity theft fall within the scope of this work: they constitute a prioritized target of most utility companies around the world, due to the severe consequences of these phenomena (i.e. higher electricity rates for paying consumers, increased risk of fire or electrocution due to improperly installed bypasses and in general, a reduced grid reliability).

From a data based perspective, a change in the energy consumption profile of a user contributing to NTLs can be understood as a deviating observation in the time series that models such a profile, whose statistics make it quite likely to be generated by another different underlying behavior [11]. However, normal load profiling in the low-voltage network can be produced by the aggregation of different, yet related behavioral components (seasonality, daily and weekly statistical variability, habits changes, among others) that differ from each other in both, amplitude (i.e. amount of energy consumed from the power grid due to different load consumptions) and time domains (correspondingly, the statistical consumption schedule of the set of users' loads along the day or week). It is the dissimilarity of any new consumption trace to any of those previously learned behavioral patterns (load profiling) what should differ both behaviors (normal and NTLs). Furthermore, the detection of anomalous observations allows for the inference of

more robust models by discarding those instances resulted from the strongly irregular samples, which could deviate the models from the representation of statistically significant regular trends in the consumption habits of the user under analysis. This being said, an outlier detection method can be defined as the task of classifying elements as normal or differing with respect to the statistical regularity characterizing a dataset. At this point the concept of *regularity* must be determined by the addressed application scenario.

In this context, a baseline taxonomy of outlier detection algorithms comprises 1) parametric methods that rely on prior hypothesis about the statistical model generating the data; and 2) model-free, non-parametric techniques, which avoid any prior assumption about the underlying distribution of the data or statistical parameter estimates. Among the latter we focus on distance-based unsupervised outlier detection approaches, which generally hinge on local distance measurements (not for accounting behavioral differences) and are capable of efficiently handling large datasets [12]. By properly defining a distance or measure of similarity between samples, subsequent data mining procedures such as cluster analysis can identify group of samples that do not belong to the set of discovered data clusters. This identification can be done based on different distance-based criteria, such as the density of samples within a given distance threshold. In all such cases the selection of the distance metric is a key point for this collection of techniques, since the similarity criterion – which roughly depends on the chosen distance metric – will guide the whole identification process. Therefore, it is clear that the best similarity function must be compliant with the nature of data and the specific particularities of the application.

In this regard, several prior contributions have hitherto dealt with the identification of NTLs in energy consumption traces. To begin with, several contributions have gravitated on the use of machine learning models over supervised datasets, such as Support Vector Machines [13, 14, 15, 16, 17, 18], Neural Networks [19, 20, 21], Extreme Learning Machines [22], Path Forests [23, 24], Decision Trees [25, 26, 27], model ensembles [28], and statistical methods [29, 30]. However, all such previous work builds upon the assumption that supervised datasets capture the entire casuistry of symptomatic anomalies of interest for fraud detection and/or electricity theft, which not only unrealistic in practice but also yields highly imbalanced datasets that subsequently jeopardize the model learning process. By contrast, unsupervised anomaly detection in Smart Grids overrides any need for previously labeled data, yet makes the evaluation and tuning of the model hard to

perform due to the non-utilization of positive examples during the construction of the learner. The literature dealing with electricity fraud using non-supervised learning models has been relatively scarce, with Self Organizing Maps [31] and fuzzy clustering schemes [32] mostly used to date.

This manuscript introduces a novel algorithmic approach for acquiring knowledge of customer's behaviors (load profiling), which allows for the identification of consumption behavioral outliers in Smart Grids based on the hourly measurements provided by the AMIs. The proposed scheme advances over the state of the art by combining probabilistic data mining and time series analysis; we adopt the so-called Dynamic Time Warping (DTW) metric as the measure of similarity between consumption traces registered by the user under analysis, by which such sequences are aligned in a dynamic, non-linear fashion disregarding any shifts or warps along time [33]. This metric is then used within two different distance-based learning models, both relying on density estimations to detect anomalous patterns. A further novel ingredient of this work is a trace encoding strategy that depends on the spanned hourly statistical ranges of every user, which increases the flexibility of the models to avoid false alarms. The performance of the derived schemes is assessed and discussed based on simulation results computed over real AMI data captured by a Spanish utility. Given the obtained scores we conclude that the proposed method accommodates irregularities of the analyzed consumption traces along time by focusing exclusively on their shape.

The rest of the manuscript is structured as follows: Section 2 poses the notation used throughout the manuscript, and formulates the problem of outlier detection contextualized for the application tackled in this manuscript. Section 3 provides an overview of the proposed approach, emphasizing on its constituent elements in subsections therein. Next, Section 4 describes the dataset utilized for performance assessment, justifies the different emulated cases over such data and discusses the obtained results. Finally conclusions are given in Section 5 along with an outline of future research lines.

2. Notation and Problem Statement

As depicted in Figure 1, we assume that an energy distribution company has deployed a set of N smart meters to monitor the consumption of part of its customer portfolio. Let data samples registered by the n-th smart meter be denoted as $\mathbf{x}^n \doteq \{x_t^n\}_{t=1}^{T_n}$, where t stands for the time dimension discretized as per the granularity t_s^n [minutes] by which the smart meter records data (e.g.

hourly, $t_s^n = 60$ minutes). Here T_n denotes the total number of samples read for the customer at hand, which may vary among different customers due to e.g. the date on which the smart meter was installed in the user premises. We further consider that the minimum decisional unit for the outlier detection model is an entire day (24 hours), for which $\mathbf{x}^n \doteq \{x_t^n\}_{t=1}^{T_n}$ can be reshaped as a matrix $\mathbf{X}^n$, with each column containing the $(24 \cdot 60)/t_s^n$ values that the meter for customer $n \in \{1, \ldots, N\}$ samples during each day. For the sake of simplicity, in foregoing derivations we will force $t_s^n = 60$ minutes $\forall n$, such that $\mathbf{X}^n$ will have 24 readings per every day out of a total of $D^n \doteq \lfloor T^n/24 \rfloor$ days monitored for customer n. Samples for day $d \in \{1, \ldots, D^n\}$ will be expressed as $\mathbf{X}_d^n$, i.e. by the d-th row in $\mathbf{X}^n$.

The aim of an outlier detection model $M_{\boldsymbol{\theta}}^n(\mathbf{X}_{d'}^n; \mathbf{X}^n)$ is to infer, for user n, whether a new daily consumption trace $\mathbf{X}_{d'}^n$ captured by the smart meter of user n follows the same distribution as that characterizing $\mathbf{X}^n$ (declaring it to be an *inlier*) or, instead, differs significantly (correspondingly, is an *outlier*). The latter case serves as a trigger for a further inspection process to confirm whether the behavioral change is due to e.g. fraud. The model is controlled by a set of parameters collected in $\boldsymbol{\theta}$, which permit to balance between the True Positive Rate (TPR, also referred to as sensitivity or recall) and the True Negative Rate of the model (namely, TNR or specificity) [34].

At this point it is important to note that for measuring the TNR and TPR metrics of any outlier detection model we need supervised labels of the test traces over which such metrics are computed. In other words, for assessing the performance of an outlier detection algorithm it is mandatory to know a priori whether the distribution utilized for producing each of the test traces corresponds to that utilized for modeling the outlier prototype that the model should detect.

We refer as $\ell_{d'}^n \doteq M_{\boldsymbol{\theta}}^n(\mathbf{X}_{d'}^n; \mathbf{X}^n) \in \{0, 1\}$ to the predicted label by the model for test trace $\mathbf{X}_{d'}^n$. Bearing this definition in mind, the TPR and TNR scores achieved by model $M_{\boldsymbol{\theta}}^n(\,\cdot\,)$ over a test dataset $\{\mathbf{X}_{d'}^n\}_{d'=1}^{D'}$ are given by $\mathrm{TNR}\,(M_{\boldsymbol{\theta}}^n)$ and $\mathrm{TPR}\,(M_{\boldsymbol{\theta}}^n)$, respectively. Noteworthy is to highlight that these metrics implicitly measure the extent to which the model is adapted to discriminate among the distribution $f_{\mathbf{X}}^1(\mathbf{x})$ followed by outliers within $\{\mathbf{X}_{d'}^n\}_{d'=1}^{D'}$ from that followed by regular traces in $\mathbf{X}^n$ (correspondingly, $f_{\mathbf{X}}^0(\mathbf{x})$). While learning $f_{\mathbf{X}}^0(\mathbf{x})$ is a matter of fitting the model to $\mathbf{X}_n$ on the assumption that all consumption traces therein are legitimate, the casuistry of outliers dictated by $f_{\mathbf{X}}^1(\mathbf{x})$ is driven by the specificities of the application scenario itself. To this end, in this work we focus on four different hypotheses

for the test trace $\mathbf{X}_{d'}^n$ which $M_{\boldsymbol{\theta}}^n(\cdot)$ should declare as an *inlier* or an *outlier*:

1. The test trace $\mathbf{X}_{d'}^n$ belongs to the normal behavioral distribution of customer n, i.e. $\mathbf{X}_{d'}^n \sim f_{\mathbf{X}}^0(\mathbf{x})$ with high likelihood. In this case the model should declare that $\mathbf{X}_{d'}^n$ is an inlier, namely, $\ell_{d'}^n \doteq M_{\boldsymbol{\theta}}^n(\mathbf{X}_{d'}^n; \mathbf{X}^n) = 0$.
2. The test trace $\mathbf{X}_{d'}^n$ falls again within the trace space spanned by the normal behavior of customer n. However, in this case a shape-preserving shift (of $\delta \in [-\Delta_{\max}, \Delta_{\max}]$ hours) in the time domain is present in the test trace to account for exogenous factors affecting the consumption patterns of the user along the time domain. For instance, a domestic user does not necessarily use his/her home appliances at the same time during the week, but it is often the case that such home duties follow a regular pattern in their execution. In this case the model should be elastic enough to accommodate this time variability, focus on purely shape-related characterization of the consumption patterns and predict that $\ell_{d'}^n = 0$.
3. The test trace $\mathbf{X}_{d'}^n$ reflects a subtle energy loss over its time span with respect to a particular legitimate example in $\mathbf{X}_n$. This effect is symptomatic of sophisticated manipulations by which the meter is slowed down regularly in short time intervals (e.g. by installing a circuit inside the device) to halt the recording process and under-register the energy consumed by the customer. Clearly, in this case the model should output $\ell_{d'}^n = 1$ depending on the ratio $\sigma \in (0, 1]$ between the overall energy of the test trace and that of the legitimate consumption trace from where it was produced.
4. Meter tampering, by which the meter is deliberately bypassed so that the device does not record any consumption at all. As a result, abrupt energy losses are obtained in the data traces of the customer, which emerge in the data trace of the day in which the tampering was performed as a series of $Z_{\max}$ zero-valued samples. The model should predict $\ell_{d'}^n = 1$ for this event, and trigger a subsequent manual inspection over the user at hand.

A good outlier detection model should take into account that the goal of the application is to correctly predict test traces falling within any of the above 4 categories. Therefore, the design goal can be formulated as a multi-objective optimization problem where the optimality of the sought set of model parameters is driven by the trade-off between two conflicting objectives: the ratio of confirmed outliers (TPR) and the proportion of correctly identified inliers (TNR) when the model predicts a test set composed by D'

new consumption traces. Mathematically:

$$\boldsymbol{\theta}^{opt} = \underset{\boldsymbol{\theta}}{\arg} \left[\max \mathrm{TNR} \left(M^n_{\boldsymbol{\theta}}(\{\mathbf{X}^n_{d'}\}_{d'=1}^{D'}; \mathbf{X}^n) \right), \max \mathrm{TPR} \left(M^n_{\boldsymbol{\theta}}(\{\mathbf{X}^n_{d'}\}_{d'=1}^{D'}; \mathbf{X}^n) \right) \right],$$

subject to $\mathbf{X}^n_{d'} \sim \{f^{0,\checkmark}_{\mathbf{X}}(\mathbf{x}), f^{0,\delta}_{\mathbf{X}}(\mathbf{x}), f^{1,\sigma}_{\mathbf{X}}(\mathbf{x}), f^{1,z}_{\mathbf{X}}(\mathbf{x})\}\ \forall d' \in \{1, \ldots, D'\}$, wherein by a slight abuse in notation we discriminate the particular hypotheses that each distribution models: normal behavior ($f^{0,\checkmark}_{\mathbf{X}}(\mathbf{x})$), shape-preserving time variability ($f^{0,\delta}_{\mathbf{X}}(\mathbf{x})$), subtle loss ($f^{1,\sigma}_{\mathbf{X}}(\mathbf{x})$) or tampering ($f^{1,z}_{\mathbf{X}}(\mathbf{x})$). In essence: we pursue the best model configuration to detect all classes of inlier and outlier traces in the test set, based on the trace set $\mathbf{X}^n$ for user n.

The above optimization problem models the conceptual, standard model adjustment process in data mining, which can be tackled by using different well-known methodologies such as cross-validation [35]. However, the design challenge goes beyond the numerical refinement of the parameters controlling the learning process of the model itself. Since a design target is to accommodate time shifts in the load curve that are not symptomatic of NTL, we opt for distance-based outlier detectors that leverage a similarity metric between time distances that is not affected by such non-linear variations. Two different outlier detection schemes will be designed based on this similarity measurement, computed not over the original data traces, but rather on their quantized values based on the hourly statistics of $\mathbf{X}^n$. The next section delves into the details of these models, along with the utilized similarity distance and the statistical quantization.

3. Proposed Approach

Figure 2 shows the overall processing flow of the outlier detection methods proposed in this manuscript. Four are the ingredients that lie at the core of the developed techniques, which are described as follows:

3.1. Similarity Measure

As argued in the previous section, a elastic measure of similarity between load profiles will be used to accommodate behavioral changes that do not imply a decrease in the energy consumed by the monitored user (e.g. time warps). To this end we will embrace the so-called Dynamic Time Warping (DTW) measure, by which the similarity between two any given consumption traces $\mathbf{X}^n_d$ and $\mathbf{X}^n_{d'}$ (i.e. traces recorded for user n at days d and d') can be

computed by searching for a minimum-weight optimal path $\mathbf{P}$ between the $(1,1)$ and (N,N) vertices of a rectangular $N \times N$ grid. The weight $w_{i,j}$ associated to vertex (i,j) in this grid correspond to the Euclidean distance between $X^n_{d,i}$ (i.e. the consumption measured for user n, day d and hour i) and $X^n_{d',j}$, namely, $w_{i,j} = \left|X^n_{d,i} - X^n_{d',j}\right|$. The DTW metric between traces of user n corresponding to day d and d' is given by [33, 36]

$$\mathrm{DTW}(\mathbf{X}^n_d, \mathbf{X}^n_{d'}) = \min_{\mathbf{P}\in\boldsymbol{\mathcal{P}}} \sum_{k=1}^{K_\mathbf{P}} w_{\mathbf{p}_k} = \sum_{k=1}^{K_\mathbf{P}} w_{i_k,j_k}, \tag{1}$$

with $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{K_\mathbf{P}}\}$ denoting a $K_\mathbf{P}$-long warping path composed by steps $\mathbf{p}_k = (i_k, j_k)$ ($k \in \{1, \ldots, K_\mathbf{P}\}$), and $\boldsymbol{\mathcal{P}}$ denoting the set of all paths through the grid fulfilling $\mathbf{p}_1 = (1,1)$, $\mathbf{p}_k - \mathbf{p}_{k-1} \in \{(1,1),(0,1),(1,0)\}$ and $\mathbf{p}_{K_\mathbf{P}} = (N,N)$.

When contextualized on the energy application tackled in this manuscript, the DTW metric allows measuring the degree of dissimilarity between two consumption traces by dismissing small behavioral shifts over the time domain and hence focusing strictly on differences in the amplitude of the energy consumed by the customer at hand. The DTW algorithm provides an adapted metric to assess the similarity between two temporal sequences which may vary in speed. A pattern in terms of the daily electric consumption must be flexible enough to cope with time deformations resulting from irregular house habits or different working schedules. Therefore, a concrete consumption pattern does not necessarily correspond to a unique feature vector in terms of both sequence modulation and periodicity – thus considering a constant window spacing and a point-to-point definition – but rather to a shape or a silhouette in a higher-level of abstraction that allows stretching or compressing sections of the series for comparison. In this work we postulate that the DTW properly deals with such an assumption on the similarity between two consumption traces under a more elastic consideration of alignment.

3.2. Statistical Trace Encoding

An optional trace encoding strategy is proposed based on the statistical ranges spanned by the hourly measurements registered for the user at hand. When computing the DTW metric two distinct strategies can be adopted: the first hinges on computing the similarity between data instances $\mathbf{X}^n_d$ and $\mathbf{X}^n_{d'}$ by using directly the numerical values of the hourly energy consumed by the user at hand. However, the straightforward use of unprocessed values

might yield an excessive rate of false positives as a result of the inflexibility of the subsequently developed models to tolerate small amplitude deviations of the hourly consumed energy.

In order to provide the model with a distance distribution capable of flexibly handling statistically negligible deviation in the energy values, a *statistical* encoding method has been also designed. This mechanism has been devised to deliberately group together those samples likely to produce overfitted models with minor contributions to the generalization accuracy. To this end, for each user n, hour h and day d in the training dataset $\mathbf{X}^n$, raw values $X^n_{d,h}$ are first transformed into boxplot symbols $X^{n,\flat}_{d,h} \in \mathcal{B}^n$ according to statistical ranges previously computed over the training set. These ranges are representative of the variability, dispersion and skewness of the energy consumed by user n based on several statistical values: first quartile Q^n_1, median Q^n_2, third quartile Q^n_3, lower limit $Q^n_{low} \doteq Q^n_1 - 1.5(Q^n_3 - Q^n_1)$, upper limit $Q^n_{upp} \doteq Q^n_3 + 1.5(Q^n_3 - Q^n_1)$, absolute minimum $Q^n_{\min} \doteq \min_{d,h} X^n_{d,h}$ and absolute maximum $Q^n_{\max} \doteq \max_{d,h} X^n_{d,h}$. Based on these values computed for user n, a mapping $\lambda^n : \mathbb{R} \mapsto \mathcal{B}^n$ is constructed, with $\mathcal{B}^n$ denoting a discrete alphabet composed by the median values of the ranges bounded by each pair in $\{Q^n_{\min}, Q^n_{low}, Q^n_1, Q^n_2, Q^n_3, Q^n_{upp}, Q^n_{\max}\}$. For instance, if the value of $X^n_{d,h}$ is between Q^n_{low} and Q^n_1, the value of $X^{n,\flat}_{d,h}$ is set to the median value of all samples in $\mathbf{X}^n$ that fall within this range. Once this mapping is constructed, it is applied to the raw energy consumption traces prior to their similarity computation, effectively smoothing the set of raw data traces with statistically coherent boundaries.

3.3. Distance-based Behavioral Pattern Search

After distance computation on either the raw energy traces $\mathbf{X}^n$ or their boxplot-encoded variants $\mathbf{X}^{n,\flat}$, a distance-based processing flow follows with a two-fold aim: to discern behavioral patterns within the consumption traces of the user, and to decide whether a new trace is an outlier or an inlier based on the learned knowledge. To this end a first distance-based clustering approach is included in the proposed scheme to identify those instances falling outside the boundaries of the regions populated by regular patterns. Such isolated observations or small-sized groups are deemed atypical behaviors when their distance generally to any cluster center substantially exceeds the borders confined by the learned pattern [37, 38]. Unfortunately, it is well-known that serious shortcomings spring up with clustering algorithms such as K-means in the presence of outliers due to their extreme sensitiveness to

anomalies. This impacts on the final cluster arrangement and directly affects the rate of false negatives, i.e. outliers declared as false positives.

Density-based methods, however, seek extreme observations or local instabilities with respect to neighboring values, although these observations are not significantly different from the rest of the population. Under these circumstances, it is imperative to discern the notion of local and global in terms of the adopted analytical approach. The former relies on a reference set containing all the data points to assess the *outlierness* of samples, whereas the resolution of the reference set for the latter is a subset of data objects predominantly referred as their neighborhood, thus preventing the model from a misleading dominant influence of anomalous points.

In the proposed scheme we opt for a hierarchical agglomerative distance-based clustering approach to discern behavioral patterns from the similarity measures $\mathrm{DTW}(\mathbf{X}_d^n, \mathbf{X}_{d'}^n)$ for user n and $\forall (d, d') \in \{1, \ldots, D^n\} \times \{1, \ldots, D^n\}$. Here cluster or pattern stands for a group of days within the traces recorded for user n with a consistent, warping-insensitive regularity in their consumption habits as measured by the distance space spanned by the DTW metric. The linkage method used to compute the dissimilarity between two given clusters $\mathcal{D}_c^n \subseteq \{1, \ldots, D^n\}$ and $\mathcal{D}_{c'} \subseteq \{1, \ldots, D^n\}$ such that $\mathcal{D}_c^n \cap \mathcal{D}_{c'}^n = \emptyset$ is given by the maximum value of $\mathrm{DTW}(\mathbf{X}_d^n, \mathbf{X}_{d'}^n)$ $\forall d \in \mathcal{D}_c^n$ and $\forall d' \in \mathcal{D}_{c'}^n$. The hierarchical tree is grown automatically to yield a number of clusters beyond which the average distance between clusters does not increase significantly. The resulting cluster space $\{\mathcal{D}_c^n\}_{c=1}^{C^n}$ comprises different behavioral patterns featured by the user over its historical consumption log. Therefore the proposed scheme should deal with traces belonging to every pattern $\mathcal{D}_c^n$ independently. To this end a new test trace $\mathbf{X}_{d'}^n$ is first mapped to a cluster $\mathcal{D}_c^n$ based on their linkage distance to every cluster in $\{\mathcal{D}_c^n\}_{c=1}^{C^n}$, and thereafter processed through an outlier detection model $M_{\boldsymbol{\theta}}^{n,c}(\,\cdot\,)$ whose parameters $\boldsymbol{\theta}$ are fitted to the characteristics of the cluster at hand.

3.4. Model Construction and Refinement

After a cluster space $\{\mathcal{D}_c^n\}_{c=1}^{C^n}$ has been inferred from the similarity measures computed for all daily traces $\mathbf{X}^n$ – or $\mathbf{X}^{n,\flat}$ – for user n, a outlier detection model $M_{\boldsymbol{\theta}}^{n,c}(\,\cdot\,)$ should be configured and optimized for every discovered cluster $\mathcal{D}_c^n$. For this purpose two different distance-based outlier detection models will be considered:

- Local Outlier Factor [39], hereafter referred to as `LOF`, which is a distance-based algorithm for the detection of anomalies under the locality paradigm.

Similarly to what is assumed in density-based clustering procedures such as DBSCAN, points belonging to a regular behavior tend to populate delimited regions characterized by a strong measure of intra-similarity (neighborhoods), whereas atypical patterns usually lie on disconnected areas far from the aforementioned neighborhood convention. `LOF` avoids taking any global assumption on the shape of the regions populated by inliers, or the maximum distance imposed as a threshold to classify a point as an outlier. Instead, this method permits judging the outlierness level of a sample $\mathbf{X}_{d'}^n$ – its local outlier factor $\mathrm{LOF}(\mathbf{X}_{d'}^n)$ – according to the *local reachability* of their neighbors. This means that if adjacent counterparts of the test sample are reachable by their own adjoined instances (under a parametric notion of closeness), it is reasonable to expect that such a object p is correspondingly accessible at an equivalent radius distance. This permits to infer a numerical degree – the local outlier factor – that reflects the level of isolation of the sample with respect to a restricted neighborhood surrounding it. The parameter set $\boldsymbol{\theta}_c^{n,\texttt{LOF}}$ of this method for cluster c and user n consists of two main parameters: 1) the number of nearest neighbors used to define the local neighborhood of the sample; and 2) a threshold $\gamma_c^{n,\texttt{LOF}} \in [0,1]$ under which the test sample $\mathbf{X}_{d'}^n$ is declared to be an inlier ($\mathrm{LOF}(\mathbf{X}_{d'}^n) \leq \gamma_c^{n,\texttt{LOF}}$, yielding $\ell_{d'}^n = 0$) or an outlier (corr. $\mathrm{LOF}(\mathbf{X}_{d'}^n) > \gamma_c^{n,\texttt{LOF}}$, hence $\ell_{d'}^n = 1$).

- Least Squares Approach [40], hereafter referred to as `LSA`, which is a probabilistic, nonparametric method for anomaly detection. Kernel models have paved the way towards more flexible and robust procedures proving useful in fields such as pattern recognition, denoising or dimensionality reduction. The main goal is to infer intrinsic relations in datasets by generating new representations capable of turning raw instances into a collection of more discriminative data samples. `LSA` hinges on such an approach to deduce the class-conditional probability of the test dataset under an analogous assumption of density to the one considered in `LOF`: regular data samples should occur in high-probability regions, whereas anomalous data points occur in low-probability regions. `LSA` leverages this assumption and approximates the conditional probability $p_{L|\mathbf{X}}(\ell|\mathbf{x})$ as $p_{L|\mathbf{X}}(\ell|\mathbf{x}) \propto \boldsymbol{\xi}_\ell \boldsymbol{\Theta}(\mathbf{x})$, where $\boldsymbol{\xi}_\ell$ is a vector of real-valued, class-dependent coefficients, and $\boldsymbol{\Theta}(\mathbf{x}) \doteq (\Phi(\mathbf{x}, \mathbf{x}_1), \ldots, \Phi(\mathbf{x}, \mathbf{x}_Y))^\top$ is a vector of $Y \leq D$ kernel functions, with $\mathbf{x}_y$ denoting a training sample. As was first proven in [41] and reviewed in [40], the optimum value $\boldsymbol{\xi}_\ell^{opt}$ of $\boldsymbol{\xi}_\ell$ for $\ell \in \{0,1\}$

is given by the minimization of a regularized squared loss with a penalty coefficient ρ that can be adjusted via cross-validation. Once this optimum vector has been computed, an estimate of the probability of a point belonging to class $\ell \in \{0, 1\}$ can be computed as

$$p_{L|\mathbf{X}}(\ell|\mathbf{x}) \approx \frac{\boldsymbol{\xi}_{\ell}^{opt}\boldsymbol{\Theta}(\mathbf{x})}{\boldsymbol{\xi}_{0}^{opt}\boldsymbol{\Theta}(\mathbf{x}) + \boldsymbol{\xi}_{1}^{opt}\boldsymbol{\Theta}(\mathbf{x})}. \tag{2}$$

The work in [40] extended the above concept to the scenario where not all classes are represented in the training data but can be present in the test data, hence being *outliers* with respect to the training samples. In essence the approach proposed in this work declares that a given test sample belongs to an anomaly class whenever it occupies a region in the with low density in the training data; to this end, the high-density regions of the feature space is characterized by means of a optimally weighted kernel model (following a regularized loss function similar to the one adopted in [41]), resulting in an estimated outlier conditional probability:

$$p_{L|\mathbf{X}}(1|\mathbf{x}) \approx \max\left\{0, 1 - \boldsymbol{\xi}_{0}^{opt}\boldsymbol{\Theta}(\mathbf{x})\right\}, \tag{3}$$

which takes a value close to 0 when the instance is within a high-density region (*inlier*) and close to 1 otherwise (*outlier*). The parameter set controlling `LSA` is $\boldsymbol{\theta}_c^{n,\texttt{LSA}} = \{\rho_c^n, \tau_c^n, \gamma_c^{n,\texttt{LSA}}\}$, where ρ_c^n is the regularization penalty of the loss function, τ_c^n stands for the free parameter in the selected kernel function (e.g. the standard deviation in a Gaussian kernel $\Phi(\mathbf{x}, \mathbf{x}') = \exp(||\mathbf{x} - \mathbf{x}'||^2/2\,(\tau_c^n)^2)$), and $\gamma_c^{n,\texttt{LSA}}$ acts as a threshold on $p_{L|\mathbf{X}}(1|\mathbf{x})$ similarly to $\gamma_c^{n,\texttt{LOF}}$ on $\text{LOF}(\mathbf{X}_{d'}^n)$.

3.5. Algorithm Description

The training, validation and test procedures of the proposed scheme when applied to the smart meter data of user n are described in Algorithm 1, whose lines have been split into blocks depending on the processing stage to which they belong: similarity computation, statistical trace encoding, hierarchical pattern search, model construction and refinement, and prediction of the test sample. Once the DTW similarity metric of every pair of daily consumption traces (either in a raw or a precomputed fashion) is calculated in Lines 1 to 3 (following Subsections 3.1 and 3.2), behavioral patterns in the form of clusters are inferred over the distance space spanned by such similarities (Line 4,

as per Subsection 3.3). The membership to any cluster is thus delegated to the DTW metric describing the similarity as a wider concept in terms of flexibility and resilience to slight warps os local phase deviations. Following the conventional F-fold cross-validation strategy to ensure generalizable models, the selected algorithm (either `LOF` or `LSA` as per Subsection 3.4) is fed with data belonging to each cluster, which eventually yields that all per-cluster models are individually fitted by means of the parameters set by the fitness metric evaluation. The prediction is accomplished by first mapping the test sample to its closest cluster using a complete linkage with the DTW as its inner distance (Line 19), followed by the application of the model associated to the cluster at hand (Line 20).

4. Experiments and Discussion

In order to assess the performance of the proposed models, several computer experiments have been carried out over real smart meter data provided by a Spanish utility. The dataset is composed of $N = 84$ consumers with their individually recorded consumption traces, containing active energy load curves obtained by AMIs for one year and a half discretized in an hourly basis. Unfortunately, the provided dataset lacks any supervision in regards to confirmed NTL cases. Furthermore, to the best of the authors' knowledge there is no collection with samples labeled as inlier/outlier publicly available for research purposes. Consequently, we have attempted at both keeping the model from being biased and at quantitatively assessing our contribution, for which a validation (targeted to the grid search for the optimization of the parameters) and a test set have been emulated. These partitions have been synthetically generated by following the same procedure: the first part of the datasets has been populated with real data traces, the second one with real but warped samples (with time shifts of up to $\Delta_{\max} = 4$ hours to evaluate the insensitiveness of the models to warps by virtue of the DTW metric. The third segment consists of an equivalent number of real data traces instances modified with subtle drops in their load profiles, drawn uniformly at random from the range $\sigma \in [0.8, 0.9]$. In other words, hourly measures recorded by the smart meter become affected by random, slight decreases in their amplitude. Finally, the last part of the validation and test dasets comprises real data traces with $Z_{\max} = 3$ zeros randomly inserted through the meter as a quantification drift or a sharp NTL event.

Algorithm 1: Proposed outlier detection approach over AMI traces.

Input: Data traces $\mathbf{X}^n$, test sample $\mathbf{X}^n_{d'}$, number of folds F, maximum expected shift $\Delta_{\max}$, maximum number $Z_{\max}$ of zero-valued samples expected in a confirmed outlier, ratio σ of the expected energy decrease of a NTL event, inner outlier detection model $M_{\boldsymbol{\theta}}(\cdot)$.

Output: Label $\ell^n_{d'} \in \{0, 1\}$ of test sample.

`Statistical Trace Encoding` (Section 3.2, optional)

Compute bounding statistics $\{Q^n_{\min}, Q^n_{low}, Q^n_1, Q^n_2, Q^n_3, Q^n_{upp}, Q^n_{\max}\}$ over $\mathbf{X}^n$.

Encode $\mathbf{X}^n$ to $\mathbf{X}^{n,\flat}$ via mapping λ^n defined by the previously computed boundaries through $\mathcal{B}^n$.

`Distance Matrix Computation` (Section 3.1)

Compute DTW similarities $\text{DTW}(\mathbf{X}^n_d, \mathbf{X}^n_{d'})$ (corr. $\text{DTW}(\mathbf{X}^{n,\flat}_d, \mathbf{X}^{n,\flat}_{d'})$) $\forall d, d' \in \{1, \ldots, D^n\} \times \{1, \ldots, D^n\}$.

`Hierarchical Pattern Search` (Section 3.3)

Extract C^n clusters $\{\mathcal{D}^n_c\}_{c=1}^{C^n}$ via hierarchical clustering using the computed pairwise similarities as a core distance metric for the linkage.

`Cluster-wise Model Construction and Refinement` (Section 3.4)

foreach $c = 1$ *to* C^n *(clusters)* **do**

  Build a grid of ϑ possible values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\vartheta\}$ for the parameter set $\boldsymbol{\theta}$ of model $M^{n,c}_{\boldsymbol{\theta}}(\cdot)$. Let $\text{TNR}(\boldsymbol{\theta}^c_{sel}) = \text{TPR}(\boldsymbol{\theta}^c_{sel}) = 0$ (*selected*).

  foreach $v = 1$ *to* ϑ *(parameter set values)* **do**

    foreach $f = 1$ *to* F *(folds)* **do**

      Train a model $M^{n,c}_{\boldsymbol{\theta}_v}(\cdot)$ over a random subset $\{\mathbf{X}^n_d\}_{d \in \mathcal{D}^{n,f}_c}$, where $\mathcal{D}^{n,f}_c \subset \mathcal{D}^n_c$ and $|\mathcal{D}^{n,f}_c| = 0.7 \cdot |\mathcal{D}^n_c|$ (70% training partition).

      Build a validation set composed by the remaining samples $\mathcal{D}^n_c - \mathcal{D}^{n,f}_c$ plus three replica of this subset with randomly emulated time shifts, energy decreases and zeroed values driven by $\{\Delta_{\max}, \sigma, Z_{\max}\}$.

      Predict labels of the validation set via the trained model, and compute $\text{TNR}_f(\boldsymbol{\theta}_v), \text{TPR}_f(\boldsymbol{\theta}_v)$ based on their true labels.

    Compute mean scores $\text{TNR}(\boldsymbol{\theta}_v), \text{TPR}(\boldsymbol{\theta}_v)$ averaged over $f \in \{1, \ldots, F\}$.

  Parameter set $\boldsymbol{\theta}^c_{sel}$ for cluster c is $\boldsymbol{\theta}^c_{sel} = \underset{v \in \{1,\ldots,\vartheta\}}{\arg\max} \min\{\text{TNR}(\boldsymbol{\theta}_v), \text{TPR}(\boldsymbol{\theta}_v)\}$.

`Test Sample: Cluster Mapping and Prediction`

Assign a cluster $c^* \in \{1, \ldots, C^n\}$ to $\mathbf{X}^n_{d'}$ based on its DTW-based linkage distance to every cluster in $\{\mathcal{D}^n_c\}_{c=1}^{C^n}$.

Predict the label $\ell^n_{d'}$ of the test example as $\ell^n_{d'} = M^{n,c^*}_{\theta^{c^*}_{sel}}(\mathbf{X}^n_{d'})$.

The primary goal of the experimental benchmark is to provide an empirically validated response to the following questions:

Q1: Do all encoding-model combinations (i.e. `LOF`, `LSA`, `LOF-box`, `LSA-box`) perform reasonably well with respect to the targeted casuistry for NTL events? Which dominates? In terms of which metric? (TNR/TPR)

Q2: When opting for encoding traces based on their statistical boundaries (`LOF-box`, `LSA-box`), does it yield an enhanced robustness against false positives? (i.e. a higher value of TNR). What is the downside in return?

Q3: How are misclassified traces distributed over the different parts comprising the test dataset? Is there any link to the regularity of the user?

Q4: Is there any chance for increasing the performance scores in a practical implementation of this scheme?

To this end macroscopic performance score statistics have been computed based on the results obtained over after a previous data cleansing stage comprising corrupted data discarding. The parameter grid $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\vartheta\}$,over which models for every discovered cluster were refined via cross-validation, are, for `LOF`, $\{1, 2, \ldots, 20\} \times \{0, 0.1, \ldots, 1.9, 2\}$, where the first term corresponds to the number of neighbors and the second one stands for the decision threshold $\gamma_c^{n,\texttt{LOF}}$. As for models based on `LSA`, the parameter grid is $\{0, 0.1, \ldots, 0.9, 1\} \times \{0, 0.1, \ldots, 0.9, 1\} \times \{0.5\}$, corresponding to ρ_c^n, τ_c^n and $\gamma_c^{n,\texttt{LSA}}$ for alleviating the computational complexity of the cross-validation process. To this end the kernel estimation within `LSA`-based approaches was further restrained to a maximum of 50 points instead of resorting to the whole training set of the cluster at hand. Those representative points can be emulated by the $\min(|\mathcal{D}_c^n|, 50)$ medoids computed by a hierarchical clustering model, where we recall that $|\mathcal{D}_c^n|$ is taken as the number of samples the training set of the cluster c. The number of folds is $F = 10$ in all cases.

As previously stated in Algorithm 1, the fitness function quantifying the optimality of a parameter set during the cluster-wise cross-validation process is $\max\{\min\{\text{TNR}, \text{TPR}\}\}$ for both `LOF` and `LSA` approaches. This combined metric prevents any of the involved metrics from becoming dominated by the other, hence forcing the model to achieve a high score in one of the two pursued criteria to the detriment of the other.

4.1. Results and Discussion

In response to Q1, we begin our discussion by analyzing Figure 4, which depicts a scatter plot comprising the test TNR/TPR scores attained by the proposed methods for every user in the dataset. Also are included in the

plot fitted Gaussian distributions for every score and technique via Kernel density estimation with a bandwidth parameter equal to 1 in all cases. A first look on the results plotted in this figure reveals that indeed both `LSA` and `LOF` benefit from the optional statistical encoding approach (Subsection 3.2) when the focus is placed on maximizing the number of true negative scores. This is specially notable in the case of `LOF`, where the average TNR increases from 0.62 (`LOF`) to 0.77 (`LOF-box`). This, as expected, comes along with a severe penalty in the number of detected positives, with a decrease in average TPR from 0.70 (`LOF`) to 0.36 (`LOF-box`). This particular result evinces the trade-off between both scores, for which the inclusion of algorithmic design options as the statistical encoding scheme is crucial to achieve performance scores aligned with the operational requirements. For instance, the operator might conservatively prioritize a low number of false positives due to internal budgetary/resource constraints for inspection tasks, hence opting for the aforementioned encoding scheme.

Comparisons between techniques can be better analyzed by redrawing the results in Figure 4 as a series of violin plots, i.e. an enhanced version of the conventional boxplot with extended information about the shape of a kernel distribution fitted to the data samples. Such plots are provided in Figure 5 along with conventional boxplots overlaid over each case. In light of these results and linking to question Q2, it can be inferred that the naive `LOF` and `LSA` schemes in general outperform their statistically encoded counterparts in terms of outlier detection (TPR), since they essentially yield a fine-grained adjusted model capable of discriminating slight deviations from the regular consumption patterns of the user. However, for users with more chaotic or unsteady patterns the parameter search procedure of the overall model fails to find a proper balance between sensitivity (TPR) and specificity (TNR). Due to the fact that a portion of the validation set (and accordingly another part of the test set) is produced by emulating minor fluctuations in legitimate consumption traces, the new data traces are likely to fall in high-density regions already populated by legitimate user traces, hence being eventually infeasible to draw boundaries for binary classification. At this point it is interesting to remark that the `LSA-box` scheme seems to be more resilient to the TPR degradation expected when including the statistical encoding within the outlier detection flow, with 70% of the overall set of analyzed users with TPR scores kept above 0.6 for this scheme.

The discussion follows by addressing question Q3; in this regard, Figure 6 depicts the distribution of the accuracy metric (i.e. the proportion of true

estimations – both positive and negative – with respect to the total number of samples processed for each user) over the different parts in which the test set is divided: Region 1 (original legitimate test traces of the user), Region 2 (original traces with random shifts in the time domain), Region 3 (subtle random perturbations in the hourly consumption value of the user) and Region 4 (sharp zeroing of the consumption trace). For the sake of space and clarity results are only shown for the `LSA` and `LSA-box` schemes. Expectedly the use of an elastic measure of similarity at the core of the classifier design implies that the score statistics between Regions 1 and 2 are similar to each other, thus evincing that the overall model is capable of accommodating occasional behavioral changes in the consumption habits of the user that other conventional similarity metrics (e.g. pairwise Euclidean distance) would declare as a false positive. When focusing on Regions 3 and 4 the obtained results confirm the intuition that subtle variations in Region 3 are significantly more challenging to detect as outliers than the zeroed data traces composing Region 4. Interestingly, accuracy scores of `LSA-box` for Region 4 are lower than those of the naive `LSA` scheme, due to the fact that small deviations may fall within the computed statistical boundaries driving the trace encoding strategy of `LSA-box`. By contrast, zeroed samples playing the role of malfunctions in the power quantification or tampering (namely, Region 4) are better detected by the `LSA-box` scheme, with accuracy scores above 0.8 for 80% of the total set of users in the experimental setup.

The rationale for the different performance patterns found between techniques over the regions of the test data traces can be also understood in connection to the regularity of the user in his/her energy consumption patterns. When translating raw values of the consumed energy to a reduced yet statistically meaningful alphabet, the overall dataset of the user at hand can be explained more likely by a reduced set of patterns. A byproduct of this simplification is a better discrimination of outliers when they are characterized by severe amplitude drops, as distances become enlarged by virtue of the range discretization to their median values. We exemplify this observation in Figure 7, which shows a boxplot of the hourly energy measurements for two different users in the dataset considering the DTW alignment between the data traces and the average consumption habit of every customer. As opposed to the consumption irregularity characterizing User A, User B features relatively more stable consumption patterns, yielding significantly better predictive scores than those obtained for user A (i.e. average TNR/TPR scores equal to 0.95/0.93 versus 0.83/0.58 for `LSA-box`).

We end the discussion by elaborating on the implementation of the proposed detectors in practice (question Q4). In this context it is important to remark that scores so far have reported for isolated daily predictions, i.e. TNR/TPR values correspond to decisions made over one single day. This, however, lays at an unrealistic extreme with respect to the practical implementation of the proposed detectors, in which the operator would enforce the inspection department to investigate the equipment installed at certain user's premises only after a number consecutive positives have been detected on his/her data traces.

A naive albeit insightful scheme modeling a more realistic implementation hinges on voting by majority a number of consecutive predictions for every user. Results shown in Figure 8 for 3 consecutively voted outcomes of the model buttress this hypothesis: predictive scores are improved notably by adopting this practical approach over those obtained by the model predicting on an individual sample basis (included also in the plot for comparison). Remarkably, `LSA-box` achieves TNR/TPR scores above 0.9 for at least 75% of all users, promisingly paving the way for the deployment and operation of this model in real smart grid scenarios.

5. Concluding Remarks and Future Research Lines

This manuscript has elaborated on the detection of NTL events in energy consumption profiles captured by AMIs in Smart Grids. In particular we have proposed a portfolio of techniques incorporating several novel ingredients over the related literature. First, a elastic measure of similarity between consumption traces has been adopted so as to accommodate the eventual temporal variability of the consumption patterns featured by the user under analysis, thus enforcing the overall detector to rather focus on shape patterns within the consumption traces disregarding the time support over which they occur. Second, we have defined an optional encoding strategy relying on boundaries driven by the statistics of the load curves of the user, conceived as a means to provide flexibility to the overall detector against minor amplitude fluctuations and consequently, to detect true negatives more reliably.

A data mining flow has been built upon two different distance-based learning mechanisms (`LOF` and `LSA`) that can be adopted as its inner classification model, incorporating further elements (e.g. distance-based clustering and cross-validation) aimed at a proper characterization of the user in regards to the casuistry of NTL events targeted in the paper. The combination

of distance-based learning algorithms and the optionality of the encoding strategy has given rise to 4 different schemes – namely, `LOF`, `LSA`, `LOF-box` and `LSA-box` –, which have been described in detail throughout the article and compared to each other over a dataset comprising real data traces of a Spanish utility company. Results obtained therefrom have been analyzed macroscopically by assessing how each scheme balances the trade-off between sensitivity and specificity when detecting emulated events reflecting different effects of NTL events in the load curves. The observed performance scores for each technique in the benchmark confirms the postulated hypotheses: the use of an elastic measure of similarity between time series reduces the rate of false alarms due to the eventual variability of legitimate consumption traces along time, whereas the inclusion of an statistical encoding approach prior to distance computation enhances the reliability of the detector when predicting legitimate traces (higher true negative rate), at the cost of a degraded discriminability of confirmed NTL events (lower true positive rate). Nevertheless, the ultimate decision concerning the selection of one model or another (accepting possibly optimal models and discarding suboptimal ones) is essentially a business-related matter depending on both the availability of inspection resources and the interest of the utility company to trigger manual inspection campaigns. Frequently, in real environments a misclassification involves considerable inspection costs derived from checking *in situ* the reasons for the predicted NTL event, hence turning the rate of false alarms into the most critical objective. Among the methods compared in our experiments, `LSA-box` stands out as the one achieving the best balance between the rate of true positives and the rate of true negatives.

Finally, we have presented a more practical detection scheme based on majority voting consecutive predictions of the proposed NTL detection algorithms, which has been shown to enhance the performance scores significantly for all techniques in the benchmark, with values above 0.9 for 75% of the users for `LSA-box` with just three votes in the decision. This last result is specially encouraging for the practical deployment and operation of the proposed scheme, to which research efforts will be invested in the near future. Other aspect in the research agenda related to this work will gravitate on the alleviation of the computational complexity characterizing the cluster-wise parameter setting by selecting the cluster samples over which models are subsequently trained and optimized. Practical policies to periodically reschedule the overall detector based on the prediction accuracy statistics and the feedback from inspection campaigns will be investigated. The applicability of the

proposed method to other energy-related scenarios (e.g. sub-metering, user profiling, demand-side management) will be also examined.

Acknowledgments

This work has been partially supported by the Basque Government under the ELKARTEK program (BID3ABI project, grant ref. KK-2015/0000080), as well as by the Spanish Ministerio de Energía y Competitividad under the RETOS program (OSIRIS project, grant ref. RTC-2014-1556-3).

Bibliography

[1] US Energy Independence and Security Act (EISA) (2007): Energy Independence and Security Act of 2007. 110th United States Congress.

[2] Liu, X., Nielsen, P. S. (2016): A Hybrid ICT-Solution for Smart Meter Data Analytics. Energy 115 (3): 1710-1722.

[3] Trindade, F. C., Ochoa, L. F., Freitas, W. (2016): Data Analytics in Smart Distribution Networks: Applications and Challenges. IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia): 574-579.

[4] Beckel, C., Sadamori, L., Staake, T., Santini S. (2014): Revealing Household Characteristics from Smart Meter Data. Energy 78: 397-410.

[5] McLaughlin, S., Podkuiko, D., MacDaniel, P. (2009): Energy Theft in the Advanced Metering Infraestructure. International Workshop on Critical Information Infrastructures Security, CRITIS, 176-187.

[6] Refou, O., Alsafasfeh, Q., Alsoud, M. (2015): Evaluation of Electrical Energy Losses in Southern Governorates of Jordan Distribution Electric System. International Journal of Energy Engineering 5(2): 25-32.

[7] Antmann, P. (2009): Reducing Technical and Non-Technical Losses in the Power Sector. Background Paper for the World Bank Group Energy Sector Strategy.

[8] Smith, T. B. (2004): Electricity Theft: A Comparative Analysis. Energy Policy 32: 2067-2076.

[9] Nesbit, B. (2000): Thieves Lurk – the Sizeable Problem of Stolen Electricity. Electrical World T&D.

[10] Nagi, J., Yap, K. S., Nagi, F., Tiong, S. K., Koh, S. P., Ahmed, S. K. (2010): NTL Detection of Electricity Theft and Abnormalities for Large Power Consumers in TNB Malaysia. IEEE Student Conference on Research and Development (SCOReD): 202-206.

[11] Hawkins, D. M. (1980): Identification of outliers. Chapman and Hall 11.

[12] Knorr, E., Ng R., Tucakov V. (2000): Distance-based Outliers: Algorithms and applications. VLDB Journal: Very Large Data Bases 8(3-4): 237-253.

[13] Ahmad, T. (2017): Non-technical Loss Analysis and Prevention using Smart Meters. Renewable and Sustainable Energy Reviews 72: 573-589.

[14] Jokar, P., Arianpoo, N., Leung, V. C. (2016): Electricity Theft Detection in AMI using Customers' Consumption Patterns. IEEE Transactions on Smart Grid 7(1): 216-226.

[15] Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., Nagi, F. (2011): Improving SVM-based Nontechnical Loss Detection in Power Utility using the Fuzzy Inference System. IEEE Transactions on Power Delivery 26(2): 1284-1285.

[16] Depuru, S. S. S. R., Wang, L., Devabhaktuni, V. (2011): Support Vector Machine based Data Classification for Detection of Electricity Theft. IEEE/PES Power Systems Conference and Exposition 1-8.

[17] Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., Mohamad, M. (2010): Nontechnical Loss Detection for Metered Customers in Power Utility using Support Vector Machines. IEEE Transactions on Power Delivery 25(2): 1162-1171.

[18] Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., Mohammad, A. M. (2008): Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines. IEEE TENCON Conference, 1-6.

[19] Ford, V., Siraj, A., Eberle, W. (2014): Smart Grid Energy Fraud Detection using Artificial Neural Networks. IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG), 1-6.

[20] Markoč, Z., Hlupić, N., Basch, D. (2011): Detection of Suspicious Patterns of Energy Consumption using Neural Network trained by Generated

Samples. ITI International Conference on Information Technology Interfaces, 551-556.

[21] Monedero, Í, Biscarri, F., Leon, C., Biscarri, J., Millan, R. (2006): MIDAS: Detection of Non-Technical Losses in Electrical Consumption using Neural Networks and Statistical Techniques. International Conference on Computational Science and Its Applications, 725-734.

[22] Nizar, A. H., Dong, Z. Y., Wang, Y. (2008): Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method. IEEE Transactions on Power Systems 23(3): 946-955.

[23] Ramos, C. C., Souza, A. N., Chiachia, G., Falcão, A. X., Papa, J. P. (2011): A Novel Algorithm for Feature Selection using Harmony Search and its Application for Non-Technical Losses Detection. Computers & Electrical Engineering 37(6): 886-894.

[24] Ramos, C. C. O., de Sousa, A. N., Papa, J. P., Falcão, A. X. (2011): A New Approach for Nontechnical Losses Detection based on Optimum-Path Forest. IEEE Transactions on Power Systems 26(1): 181-189.

[25] Cody, C., Ford, V., Siraj, A. (2015): Decision Tree Learning for Fraud Detection in Consumer Energy Consumption. IEEE International Conference on Machine Learning and Applications (ICMLA), 1175-1179.

[26] Nizar, A. H., Zhao, J. H., Dong, Z. Y. (2006): Customer Information System Data Pre-processing with Feature Selection Techniques for Non-Technical Losses Prediction in an Electricity Market. International Conference on Power System Technology, 1-7.

[27] Filho, J. R., Gontijo, E. M., Delaiba, A. C., Mazina, E., Cabral, J. E., Pinto J. P. O. (2004): Fraud Identification in Electricity Company Customers using Decision Trees. IEEE International Conference on Systems, Man and Cybernetics 4: 3730-3734.

[28] Muniz, C., Figueiredo, K., Vellasco, M., Chavez, G., Pacheco, M. (2009): Irregularity Detection on Low Tension Electric Installations by Neural Network Ensembles. IEEE International Joint Conference on Neural Networks, 2176-2182.

[29] Fourie J. W., Calmeyer J. E. (2004): A Statistical Method to Mini-

mize Electrical Energy Losses in a Local Electricity Distribution Network. IEEE AFRICON Conference 2: 667-673.

[30] Nizar A. H., Dong Z. Y., Jalaluddin M., Raffles M. J. (2006): Load Profiling Non-Technical Loss Activities in Power Utility. First International Power and Energy Conference (PECON) 1: 82-87.

[31] Cabral, J. E., Pinto, J. O., Pinto, A. M. (2009): Fraud Detection System for High and Low Voltage Electricity Consumers based on Data Mining. IEEE Power & Energy Society General Meeting, 1-5.

[32] Angelos, E. W. S., Saavedra, O. R., Cortes, O. A. C., de Souza, A. N. (2011): Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. IEEE Transactions on Power Delivery 26(4): 2436-2442.

[33] Berndt, D. J., Clifford, J. (1994): Using Dynamic Time Warping to find Patterns in Time Series. KDD workshop 10(16): 359-370.

[34] Fawcett, T. (2006): An introduction to ROC analysis. Pattern Recognition Letters 27(8): 861-874.

[35] Arlot, S., Celisse, A. (2010): A survey of cross-validation procedures for model selection. Statistics surveys 4: 40-79.

[36] Fu, T. C. (2011): A review on time series data mining. Engineering Applications of Artificial Intelligence 24(1), 164-181.

[37] Shekhar S., Lu C. T., Zhang P. (2002): Detecting Graph-Based Spatial Outlier. Intelligent Data Analysis 6(5): 451-468.

[38] Acuna E., Rodriguez C. A. (2004): Meta Analysis Study of Outlier Detection Methods in Classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez.

[39] Breunig, M. M., Kriegel, H. P., Ng, R. T., Sander, J. (2000): LOF: Identifying Density-based Local Outliers. ACM SIGMOD record 29(2): 93-104.

[40] Quinn, J. A., Sugiyama, M. (2014): A Least-Squares Approach to Anomaly Detection in Static and Sequential Data. Pattern Recognition Letters 40: 36-40.

[41] Sugiyama, M. (2010): Superfast-Trainable Multi-class Probabilistic

Classifier by Least-Squares Posterior Fitting. IEICE Transactions on Information and Systems 93: 2690-2701.

List of Figures

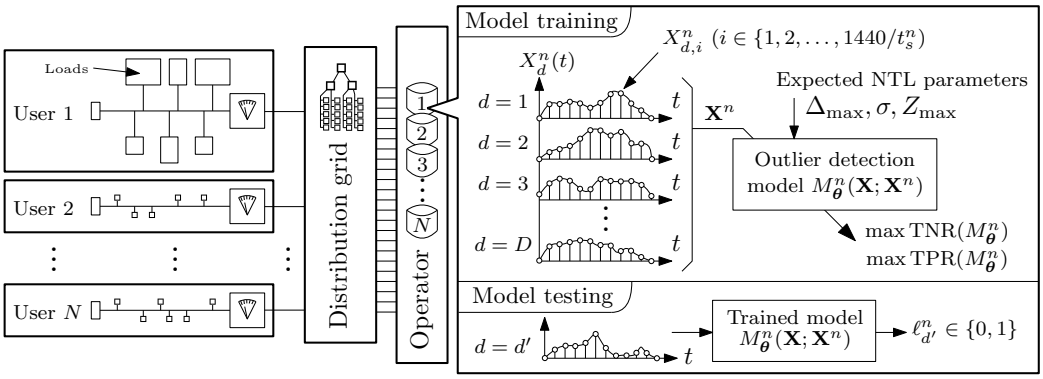

Figure 1: Schematic diagram of the scenario tackled in this work. The energy operator processes consumption traces $\mathbf{X}^n$) registered for every user towards building an outlier detection model $M^n_{\boldsymbol{\theta}}(\mathbf{X}^n_{d'}; \mathbf{X}^n)$ in accordance with the expected statistics of the NTL events to be detected as outliers.

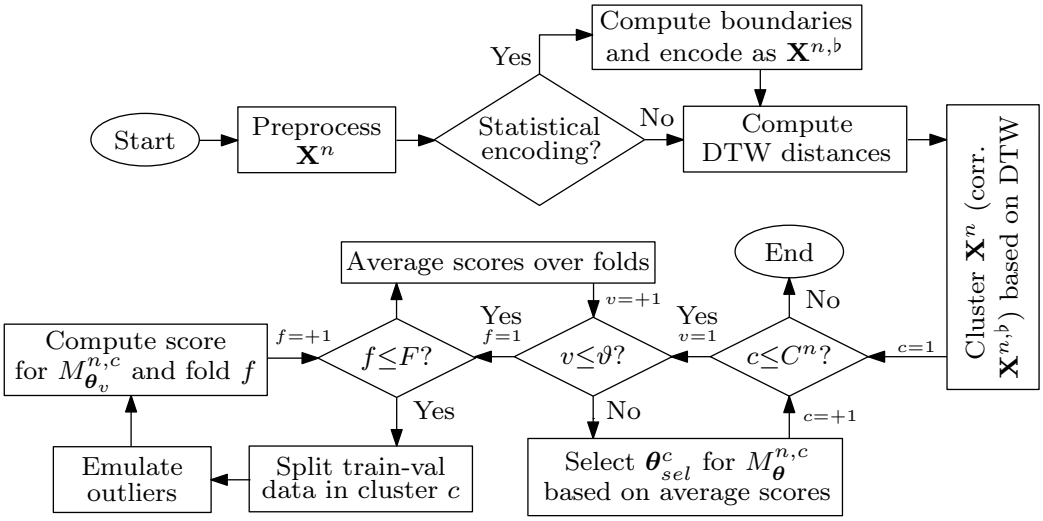

Figure 2: Flow diagram of the training procedure to construct the proposed models.

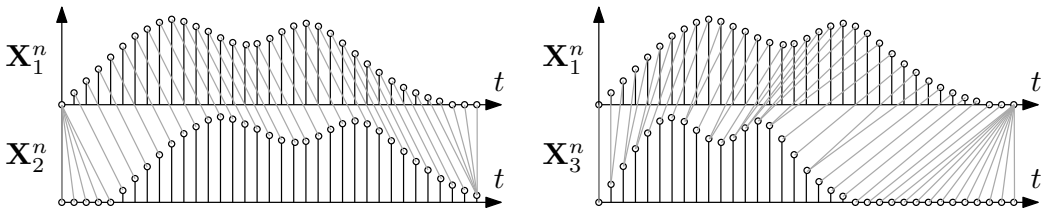

Figure 3: Three data traces (original, shifted, warped) whose pairwise DTW metric is close to zero disregarding the nonlinear variations among them along the time dimension.

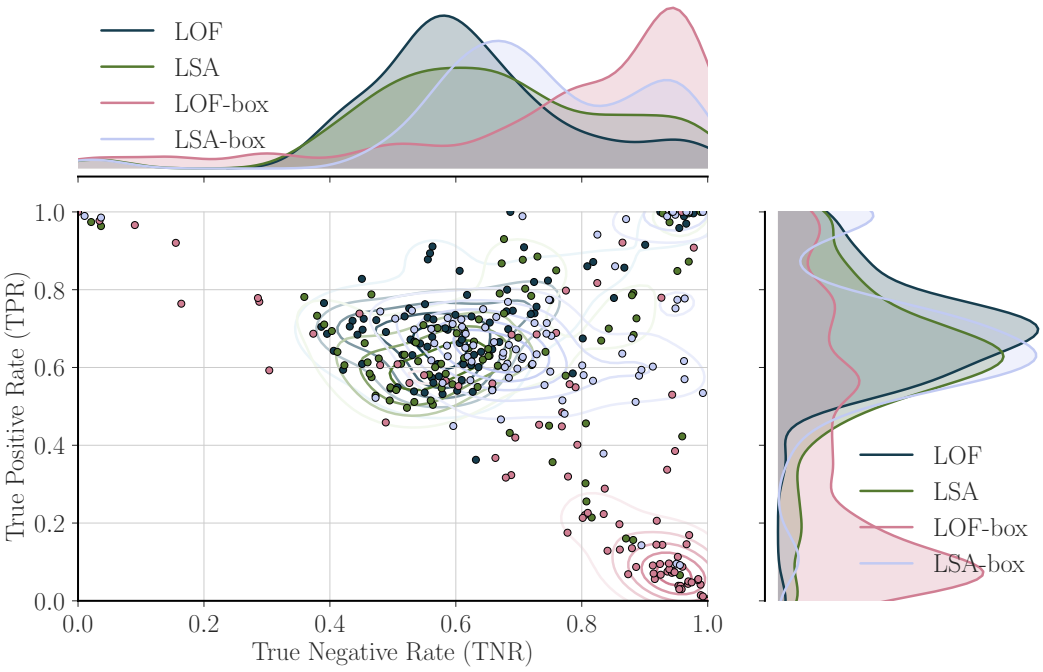

Figure 4: Scatter plots depicting the test TNR/TPR values obtained for each user in the dataset and the four model variants proposed in this work. For a better macroscopic understanding of the underlying distributions the output of a Gaussian kernel density estimator with bandwidth 1 is included for every performance score and technique.

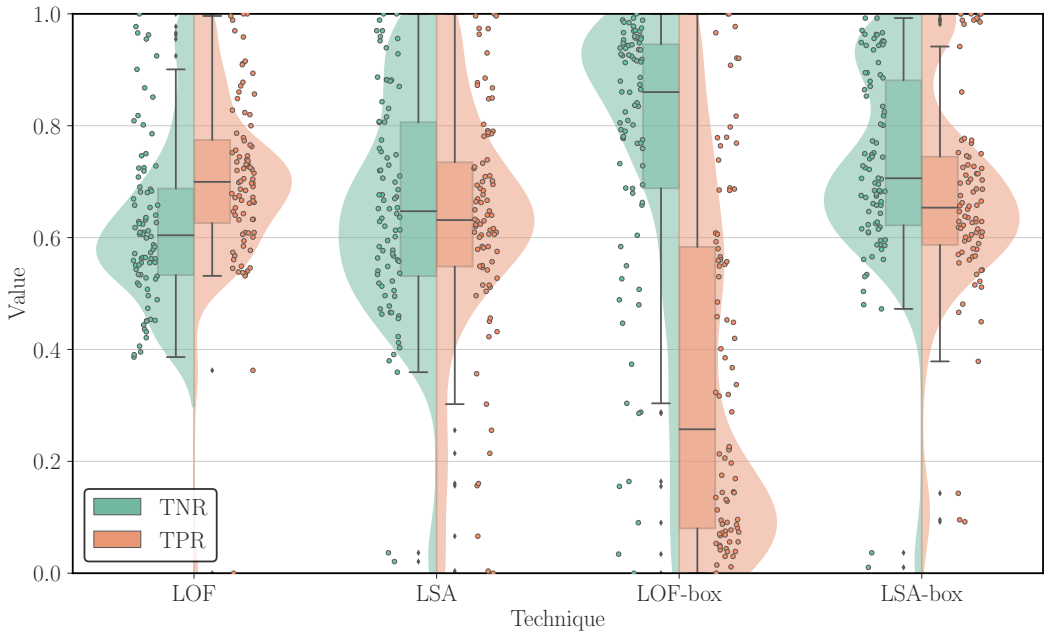

Figure 5: Violin plot of the TNR-TPR statistics for every technique in the benchmark. The `LOF-box` is severely affected by the statistical trace encoding strategy, with the Pareto between TNR and TPR severely unbalanced in favor of the latter. By contrast, TNR stats of `LSA-box` enhance slightly, yet keeping the TPR score still at admissible levels.

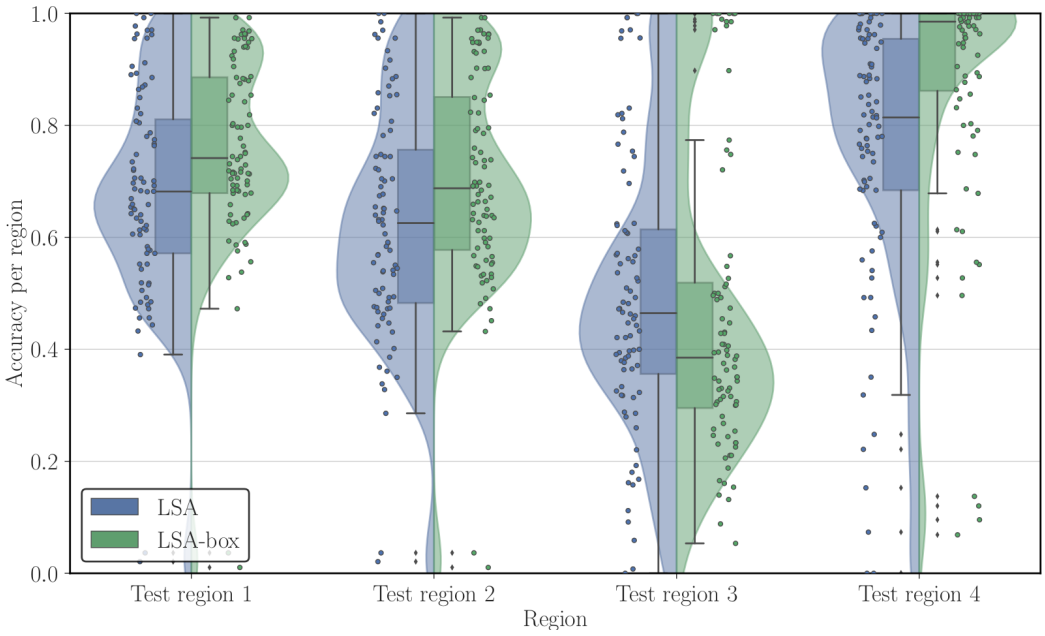

Figure 6: Distribution of errors in every region of the dataset for `LSA-raw` and `LSA-box`: region 1 corresponds to original data traces that should be labeled as *inliers*, similarly to those in region 2 where original data traces are warped along time for a maximum shift of $\Delta_{\max} = 4$ hours. Regions 3 and 4 should be declared as outliers since they emulate sharp (zeroing, as could happen in tampering) and subtle (small decreases of the recorded energy) NTL events, respectively. Expectedly, scores are significantly lower in region 3, where the effect NTL event is less severe over the test data than in the rest of regions.

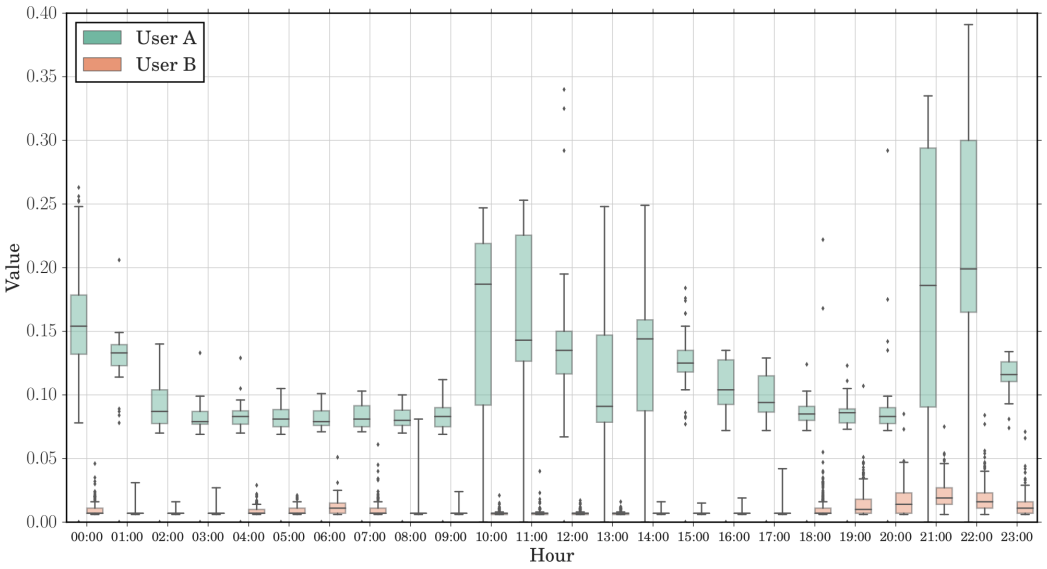

Figure 7: Hourly boxplot exemplifying the regularity and irregularity of two consumers in what regards to his/her energy consumption habits. Data samples used for computing the boxplot at hour $h \in \{0, 1, \ldots, 23\}$ are composed by those hourly measurements along $\mathbf{X}^n$ (with $n \in \{A, B\}$) matched, via DTW alignment, to the h-th hour of the average consumption habit of the customer (computed over $\mathbf{X}^n$).

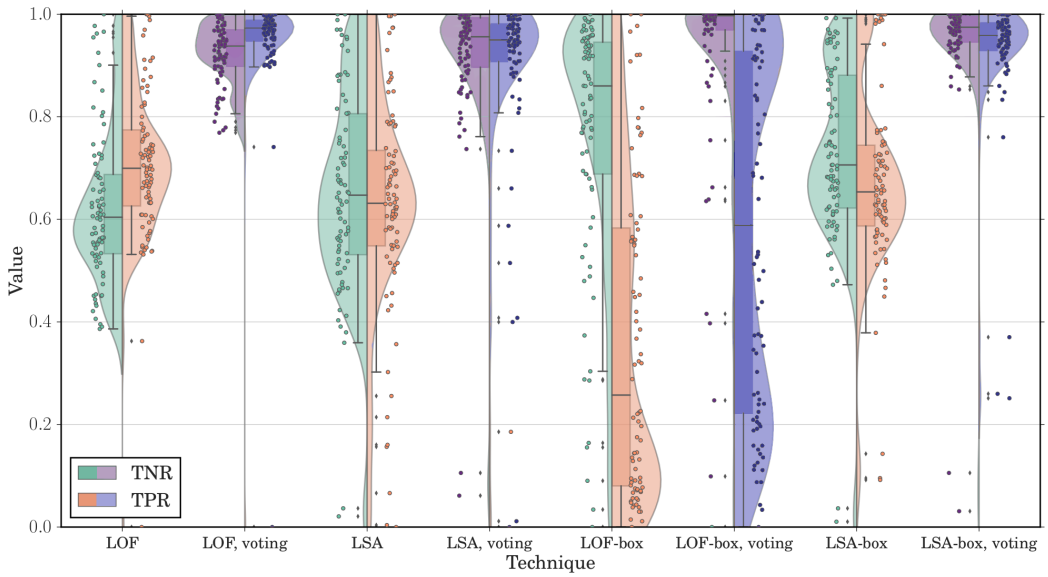

Figure 8: Boxplots corresponding to the TNR/TPR scores obtained for each technique (retrieved from Figure 5), and those scored by voting by majority three consecutive outcomes of the model.