Multiscale Data Analytics for Residential Active Power Measurements through Time Series Data Mining

Grigore Stamatescu
Automation and Industrial Informatics
University Politehnica of Bucharest
Bucharest, Romania
grigore.stamatescu@upb.ro

Radu Plamanescu
Faculty of Electrical Engineering
University Politehnica of Bucharest
Bucharest, Romania
radu.plamanescu@upb.ro

Ana-Maria Dumitrescu
Faculty of Electrical Engineering
University Politehnica of Bucharest
Bucharest, Romania
anamaria.dumitrescu@upb.ro

Irina Ciornei
MicroDER Lab, University Politehnica of Bucharest
KIOS Centre of Excellence, University of Cyprus
ciornei.irina@ucy.ac.cy and irina.ciornei@upb.ro

Mihaela Albu
Faculty of Electrical Engineering
University Politehnica of Bucharest
Bucharest, Romania
mihaela.albu@upb.ro

***Abstract*—Modern measurement and automation equipment for energy systems collect, store, process and communicate ever increasing quantities of raw data which can be used to build data-driven prediction and classification models. Directly using these large and unprocessed data sets can be inefficient especially in imbalanced class problems, where the positive class is sparsely represented in the training examples, such as classification of micro-scale transients. This is due to the longer time required for model training, and due to the increased possibility of obfuscating the useful information behind noisy readings. The matrix profile represents a computationally efficient and general purpose time series data mining technique which is suitable for embedded deployment in future generation smart meters, and in embedded energy gateways. Our analysis concerns the application of this technique on two types of residential electric power measurement data sets: a detached single family house and an apartment, with variable reporting rate and subsequence size parametrisation. Quantitative results support the findings that such approaches serve as practical instrument for measurement time series pre-processing in energy analytics.**

***Index Terms*—data analytics, feature extraction, power measurements, time series, pre-processing**

I. Introduction

Traditionally, the power distribution grids were passive by design and no real-time monitoring was required for their operation. However, the expansion of distributed energy resources (DER), especially of renewable sources (RES), together with significant changes in the power profiles of the loads, and expected increase of ad-hoc loads such as electric vehicles, are currently requesting for deployment of advanced metering infrastructure at this part of the grid [1]. Furthermore, enhanced situational awareness is needed for flexible and optimal operation of the current active distribution grids which might incorporate several types of actively controlled entities, such as energy communities [2], microgrids, large charging stations on top of the now classical flexible loads and DER [3]. Next generation smart meters (SM) [4] or micro phasor measurement units (PMUs) [5] are among the advanced metering technologies able to provide high resolution measurements needed for plethora of applications in active distribution grids and smart grids [6]. However, transmitting, processing and analysing these measurements in the raw format, in a central location might be costly, impractical as well as raising privacy concerns [2], [7].

Multi-scale data analytics is an emerging research field which integrates multiscale modelling [8] with multi-scale computing software [9] in several fields of science and engineering, especially for measurements and sensors data coming at different scales. The later helps organizing and store large sets of data in a distributed fashion such that to efficiently exploit and to further apply analytics algorithms on those large data sets, such as multi-scale computing patterns [10], or workload characterization for performance efficiency [11], among many other applications.

Measurements, coming from several types of advanced metering infrastructure, sensors or distributed embedded systems with more and more increasing sampling rates are especially sources of such type of large data sets for power industry applications [5], [4]. Data-driven predictions and classification models are among the machine learning techniques used to extract useful information from high reporting rate metering infrastructure [12]. However, one of the major challenges in this type of solutions is related to imbalanced class problems, where the positive class is sparsely represented in the training samples [13]. This might be the case for anomaly classification models of power consumption in residential locations [14].

Matrix profile is a data mining technique often used in

This work was supported by a grant of the Romanian Ministry of Education and Research, UEFISCDI no. PN-III-P4-ID-PCE-2020-2876 "Advanced Measurement Framework for Emerging Electric Power Systems" (EMERGE).

pre-processing of large time series archives, aiming for fast anomaly detection, classification or labelling of continuous streams of newly arriving data [15]. One of the main advantages of this technique is that it requires low computational resources and it could be easily implemented in distributed low power energy gateways, such as advanced smart meters.

Previous contributions relating to MP implementation and energy data have been illustrated in [16], [17], [18] which were mostly focused on building energy data with low sampling resolution of 15 minutes and one hour, respectively. This normally cancels many subtle events and anomalies through averaging when focusing on micro-transient regimens in high-sampling rate measurements for power system state estimation. In [19] the objective has been to assess the performance of deep learning forecasting models on power measurements data in terms of various accuracy metrics: Mean Squared Error - MSE, Mean Absolute Error - MAE and Mean Absolute Percentage Error - MAPE, and computational performance at various reporting rates. Our intuition is that a combination of suitable measurement data pre-processing, and feature extraction with state-of-the-art data-driven models based on hybrid convolutional-recurrent neural networks with multiple layers may result in robust models for various applications in energy systems.

To this end, the major contributions of this work are:

- application of the Matrix Profile (MP) time series data mining technique for efficient information extraction from smart meter power measurements;
- comparative results analysis on two residential use cases, single family home and urban apartment, at various timescales;
- discussion on the usage of the extracted features and reduction in the time series representation in a learning framework for steady-state/rvc classification aiming to reduce teh needed computational effort.

Subsequently, we discuss in detail the Matrix Profile (MP) technique in Section II and its application for our energy measurement scenario, along with the available datasets. Section III focuses on a gradual discussion of implementation and results obtained to illustrative the comparative analysis. Section IV presents the conclusion and outlook on future work.

II. Methods and Datasets

Building time series data structures involves associating information on the physical parameters, provided by measurements with reliable time-stamps, using - in most cases - uniform sampling. From the perspective of building data-driven models using machine learning and analytics methods, this type of data includes information both in the actual measurement data, and in the association between the measurement data patterns and their occurence in time. The latter can be used to encode both short-term and long-term dependencies of the underlying monitored phenomenon or process in the model parameters at the training stage. By combining the analysis of recorded data with suitable pre-processing techniques an adaptive system can be achieved that balances accuracy against computationally efficient (re-)training while limiting the inherent noise in large unbalanced datasets used for training. Pre-processing by means of feature extraction and feature selection allows for early pattern discovery and focusing on particular segments of the input data. Popular methods for these tasks such as principal component analysis (PCA) and singural value decomposition (SVD) are well documented in the technical literature and can represent complementary alternatives to MP.

A. Matrix Profile for Time Series Data Mining

Matrix Profile (MP), initially proposed by [15], represents a vector of values, computed by sliding a window of size m over a time series T of size n. Each value in the vector stores the minimum z-normalised Euclidean distance to its neighbors. The use of the Euclidean distance d:

$$d\left(T_a, T_b\right) = \sqrt{\sum_{i=1}^{n} \left(T_{a,i} - T_{b,i}\right)^2} \tag{1}$$

where T_a and T_b are two subsequences of equal length and $T_{a,i}$ and $T_{b,i}$ are the i elements in the respective subsequence, ensures efficient computation, preceeded by the z-normalisation which enables comparable results across various absolute values. Robustness of the method can be assessed by adding synthetic noise traces to the original input time series and quantifying the resulting profiles. Several open-source software implementations: python matrixprofile, stumpy, MATLAB library, and various algorithms: scrimp, scrimp++, stomp, mpx, can be used to compute it for both offline and online usage on streamed data. In this case the new values are gradually incorporated into the resulting profile by sliding the analysis window further, without recomputing the full spectrum of distances.

One advantage of MP lays is the fact that it uses a single parameter for tuning: the subsequence length m, which in our case of analyzing residential active power measurement values can be related to pre-observed daily consumption and appliance use patterns in clusters of homes. Alternatively, several automated and visual inspection procedures are available to identify the optimal subsequence length for a given time series.

Three main features of the technique are leveraged in this work, as follows:

- Derivation of the daily MP vectors for the residential active power measurement data for comparing smart meter traces in conjuction with apriori domain knowledge;
- Identification of measurement time series motifs, as recurring patterns within the measurement time series;
- Identification of time series discords, as highly dissimilar measurement time series patterns which can signal an anomaly for control purposes.

As discussed in [16], given time-series T, two subsequences of length m, $\{T_{a,m}, T_{b,m}\}$ are considered a $motif$ pair if:

$$dist(T_{a,m}, T_{b,m}) \leq dist(T_{i,m}, T_{j,m}), \tag{2}$$

$\forall i, j \in [1, 2, ..., n - m + 1]$ with $a \neq b$, $i \neq j$.

The subsequence with the maximum distance to its nearest non-self match neighbor can be interpreted as an unusual subsequence or anomaly and is denoted as a *discord*. Given the time-series subsequence $T_{c,m}$ of length m non-self matched with $T_{d,m}$ and the subsequence $T_{p,m}$, non-self matched with $T_{q,m}$, $T_{c,m}$ we label a *discord* if:

$$min(d(T_{c,m}, T_{d,m})) > min(d(T_{p,m}, T_{q,m})), \quad (3)$$

with $c \neq d$, $p \neq q$ and d a z-normalized Euclidean distance function.

B. Description of the two datasets and measurement context

It is recognized that residential customers have a large variability in their electricity use e.g. short periods of time with peak demand and lower demand during most of the time during the day. Despite this, most of network impact studies for RES at LV grid still rely on hourly aggregated load or, in the best case, on 15 minutes data aggregation. It is acknowledged in [20] that near real-time information and data communication is a critical technical requirement for the operation of LV grids. To highlight the energy analytic approach of pre-processsing active power time series, for the energy consumption side, it has been made use of 1s-reporting rate load profiles derived from high reporting measurements using the so-called Unbundled Smart Meter (USM) concept [21], required for achieving an optimal operation of the proposed emerging system. We have collected two datasets of residential energy meters over the course of several weeks as experimental deployment. The first dataset represents a single-family house, while the second is an apartment dwelling from Bucharest, Romania. Data has been collected in shoulder season, corresponding to the month of September 2021 and 2020 respectively, which has a neutral impact on the observed patterns through the absence of considerable cooling and/or heating loads.

Figure 1 illustrates the daily active power profile derived from measurements provided by a three phase smart meter installed in the house, with raw data made available every 1 second. A similar pattern is presented in Figure 2 with the difference being that, in this case, the power profile is obtained by mediating the raw data on a 900-samples window, equivalent with linear aggregation over the 15 minute reporting interval, thereby obfuscating both noise and potentially relevant signal variability.

III. Results

We present the results of the analysis using the MP technique on the above mentioned datasets. The implementation has been carried out using the Python programming language and the open source *matrixprofile*[1] library. The library provides software methods to analyze, compute and visualize the profile of the input time series. A pan-matrix profile data structure is also provided that allows the selection and review of multiple MP records, useful for comparing various parametrisation

[1]https://github.com/matrix-profile-foundation/matrixprofile

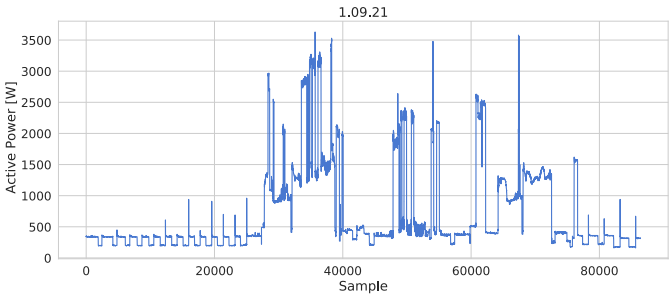


Fig. 1: Daily power profile from measurements with 1 frame/s reporting rate -example

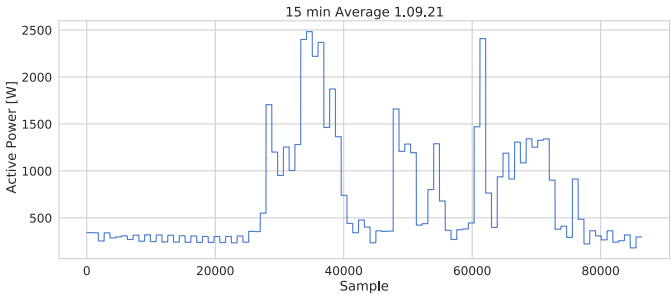


Fig. 2: Daily power profile from emulated meters with 4 frames/h reporting rate, using linear averaging

options such as: window size, noise levels, type of algorithm. We use Google Colab as online development environment.

In Figures 3 and 4 the MP values vector for each of the two daily power measurement traces introduced in the previous section are illustrated. The most suitable time window is automatically calculated for each of the examples.

In each figure, the red point marks the location of the top level discord i.e. the most dissimilar subsequence in the original data. One salient observation is that, for the 1s sampled data, a brief power spike is identified in the evening of the respective day. In the second situation, when processing 15 minute averaged data, the brief spike is eliminated through averaging and thus the MP procedure then marks an earlier and wider spike of power. This can be extended to the top n discords that mark multiple anomalies in the data which can be analyzed through deeper inspection of the original measurement sample subsequences.

In addition a heatmap profile is shown on top of each figure with lower (blue) values in the second part of the morning, after a brief period of increased values (orange). In this case, the rectangular patterns at the beginning of the day for both examples are labeled as time series motifs e.g. corresponding to the sole periodic power draw of a home appliance such as a refrigerator in the absence of other consumers in an unoccupied house.

A histogram showing the overlayed empirical distributions of the matrix profile values for the two cases is presented in Figure 5. Clustering and thresholding MP values can point out areas of interest in the original time series where

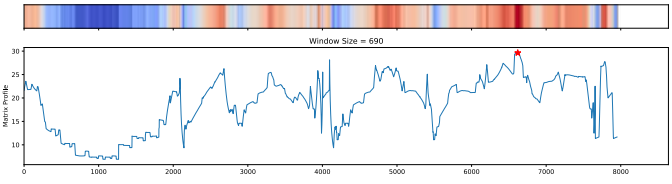


Fig. 3: Matrix profile for daily power measurements trace (1 frame/s reporting rate)

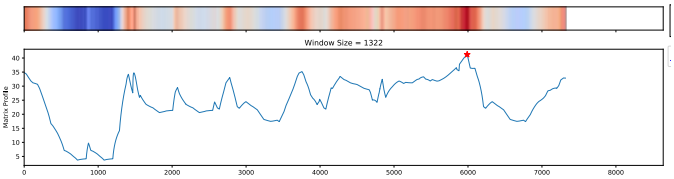


Fig. 4: Matrix profile for daily power measurement trace (averaged)

appropriate measures can be taken with regard to the adaptive (increased) sampling of the measurement and control actions. z-normalisation of the time series ensures comparability across multiple domains. The higher MP values for the second power measurement time series can be explained though the steeper changes on sparser discrete power levels produced during the averaging procedure.

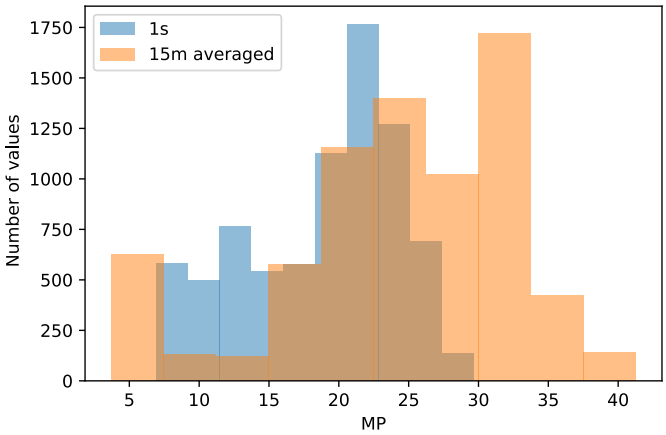


Fig. 5: Empirical distributions of the MP values

Data structure representation of the analysis function for a single MP vector computation is presented below, corresponding to the 1s sampled measurements. It allows programmatic access to both the input values and the computed profile values and also records the parametrisation options during the procedure. In this case the mpx algorithm corresponds to a fast implementation of the MP that does not use the Fast Fourier Transform (FFT).


```
    {'algorithm': 'mpx',
 'class': 'MatrixProfile',
 'data': {'query': None,
  'ts': array([341.99, 341.99, 341.99, ...,
  296.44, 296.44, 296.44])},
 'ez': 0,
 'join': False,
 'lmp': None,
 'lpi': None,
 'metric': 'euclidean',
 'mp': array([34.81832531, 34.80253447,
 34.786677  , ..., 32.89044937,
        32.88824545, 32.88603246]),
 'pi': array([ 360,  361,  362, ...,
 6774, 6775, 6776]),
 'rmp': None,
 'rpi': None,
 'sample_pct': 1,
 'w': 1324}
```


We further include the apartment dataset in the analysis. Figure 6 presents one days of active power measurements for each of the residential dwellings, taken during the same day in September. A lower baseline power consumption can be observed as well as narrower consumption peaks. In the middle of day an unusual pattern is represented which can correspond to a particular appliance. This can thus be considered as a potential candidate for flagging of anomalous behaviour.

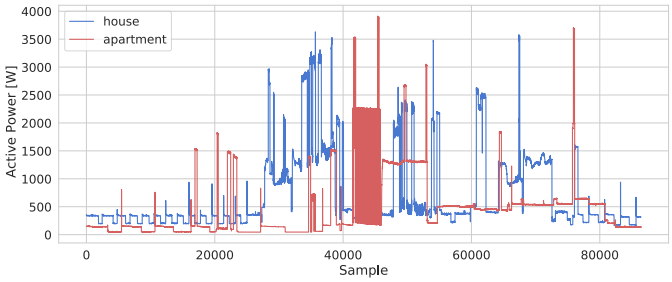


Fig. 6: Daily active power measurements: house and apartment

The corresponding MP vectors are illustrated in Figure 7. We observe a larger variability in the apartment profile corresponding to larger minimum distances between the subsequences of the values which can be explained through a large amplitude of the changes relative to the lower baseline. The motif pattern which has been previously observed is not replicated in this second dataset given several spikes in the data at the beginning of the series.

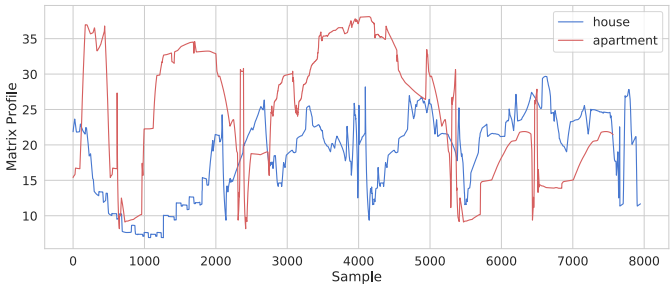
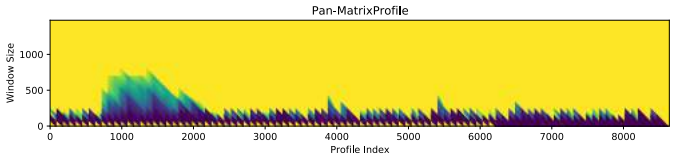


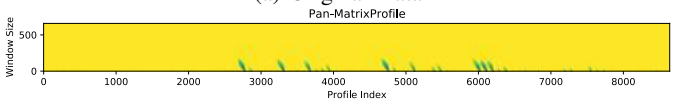
Fig. 7: Comparative profiles between two dwellings

The robustness of the method is tested by leveraging the 'add noise' parametrisation option for evaluation of the computed profiles. We validate the fact that by inserting Additive White Gaussian Noise (AWGN) to the original time series the fundamental shape of the profile for the same window size does not change. The standard deviation $\sigma = \sqrt{1/n \sum (x_i - \bar{x})^2}$ with and n the number of samples is computed and presents similar values for both series.

In the case of the pan-matrix profile, where the MP vector is computed for various subsequence length on the averaged data from, the differences are shown in Figure 8. The data has been decimated for tractable computation and the window size ranges from 1 to the full length of the decimated series (8640). This stands to validate the fact that on flat surfaces noise has a disproprotionate effect on the MP values. In this case the noise has been added to averaged data set. The denser colored areas in the figure show the similarity of the motifs detected for various subsequence lengths. The lighter colored areas show that the motifs are not replicated across matrix profiles computed at different window sizes in the presence of noise, while they are fairly consistent across multiple values of the window size in the original data. The effect of noise is amplified for short subsequences.



(a) Original Data



(b) Original Data with Noise

Fig. 8: Effect of noise on Pan-Matrix Profile Data Structure

IV. Conclusion

The MP technique serves as a robust tool for pre-processing and feature extraction in a data analytics framework for measurements in the energy domain. Given the efficient computational performance and multiple options for implementation MP is a reliable tool in a machine learning pipeline. An important characteristic is also the fact that exactness can be traded off for speed by choosing a faster yet inexact algorithm for computation. This can be useful on extremely high reporting rates encountered in current and future advanced metering technologies. An optimisation problem can be thus formulated to dynamically adjust the processing parameters according to the dynamics of the underlying process. Future work is focused on validating the approaches presented at scale and testing the implementation on embedded edge devices on online collected power measurements data. Bootstrapping time series labels for transient analysis can present a relevant use case.

References

[1] S.-C. Huang, C.-N. Lu, and Y.-L. Lo, "Evaluation of ami and scada data synergy for distribution feeder modeling," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1639–1647, 2015.

[2] M. Sanduleac, V. I. Ciornei, L. Toma, R. Plamanescu, A.-M. Dumitrescu, and M. Albu, "High reporting rate smart metering data for enhanced grid monitoring and services for energy communities," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2021.

[3] A. K. Marvasti, Y. Fu, S. DorMohammadi, and M. Rais-Rohani, "Optimal operation of active distribution grids: A system of systems framework," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1228–1237, 2014.

[4] A. Yassine, A. A. Nazari Shirehjini, and S. Shirmohammadi, "Smart meters big data: Game theoretic model for fair data sharing in deregulated smart grids," *IEEE Access*, vol. 3, pp. 2743–2754, 2015.

[5] V. S. Kumar, T. Wang, K. S. Aggour, P. Wang, P. J. Hart, and W. Yan, "Big data analysis of massive pmu datasets: A data platform perspective," in *2021 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2021, pp. 1–5.

[6] EURELECTRIC, "Electricity for europe - active distribution grids," 2018.

[7] R. Mello and R. A. Martins, "Can big data analytics enhance performance measurement systems?" *IEEE Engineering Management Review*, vol. 47, no. 1, pp. 52–57, 2019.

[8] M. A. Ferreira and H. K. Lee, "Multiscale time series," *Multiscale Modeling: A Bayesian Perspective*, pp. 113–143, 2007.

[9] D. Groen, J. Knap, P. Neumann, D. Suleimenova, L. Veen, and K. Leiter, "Mastering the scales: a survey on the benefits of multiscale computing software," *Philosophical Transactions of the Royal Society A*, vol. 377, no. 2142, p. 20180147, 2019.

[10] S. Alowayyed, T. Piontek, J. L. Suter, O. Hoenen, D. Groen, O. Luk, B. Bosak, P. Kopta, K. Kurowski, O. Perks *et al.*, "Patterns for high performance multiscale computing," *Future generation computer systems*, vol. 91, pp. 335–346, 2019.

[11] M. Malik, K. Neshatpour, S. Rafatirad, and H. Homayoun, "Hadoop workloads characterization for performance and energy efficiency optimizations on microservers," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, no. 3, pp. 355–368, 2018.

[12] M. Ferdowsi, A. Benigni, A. Monti, and F. Ponci, "Measurement selection for data-driven monitoring of distribution systems," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4260–4268, 2019.

[13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[14] K. Mahmud, J. Ravishankar, M. J. Hossain, and Z. Y. Dong, "The impact of prediction errors in the domestic peak power demand management," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4567–4579, 2020.

[15] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, A. Dau, D. Silva, A. Mueen, and E. Keogh, "Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," 12 2016, pp. 1317–1322.

[16] C. Nichiforov, I. Stancu, I. Stamatescu, and G. Stamatescu, "Information extraction approach for energy time series modelling," in *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*, 2020, pp. 886–891.

[17] C. Nichiforov, G. Stamatescu, I. Stamatescu, and I. Făgărăşan, "Learning dominant usage from anomaly patterns in building energy traces," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 548–553.

[18] G. Stamatescu, R. Entezari, K. Römer, and O. Saukh, "Deep and efficient impact models for edge characterization and control of energy events," in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, 2019, pp. 639–646.

[19] G. Stamatescu, I. Ciornei, R. Plamanescu, A.-M. Dumitrescu, and M. Albu, "Reporting interval impact on deep residential energy measurement prediction," in *2021 IEEE 11th International Workshop on Applied Measurements for Power Systems (AMPS)*, 2021, pp. 1–6.

[20] Y. Yuan and Z. Wang, "Mining smart meter data to enhance distribution grid observability for behind-the-meter load control: Significantly improving system situational awareness and providing valuable insights," *IEEE Electrification Magazine*, vol. 9, no. 3, pp. 92–103, 2021.

[21] M. Sanduleac, L. Pons, G. Fiorentino, R. Pop, and M. Albu, "The unbundled smart meter concept in a synchro-scada framework," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 2016, pp. 1–5.