# Information Extraction Approach for Energy Time Series Modelling

Cristina Nichiforov, Ionut Stancu, Iulia Stamatescu, Grigore Stamatescu
*Department of Automation and Industrial Informatics*
*University Politehnica of Bucharest*
Bucharest, Romania
{cristina.nichiforov, iulia.stamatescu, grigore.stamatescu}@upb.ro, ionut.stancu@stud.acs.pub.ro

***Abstract*—Increased adoption of energy monitoring devices across the energy system has resulted in large quantities of multivariate measurement data sets available for analysis at multi-scale resolutions. For buildings in particular, these can be leveraged to extract relevant information in order to characterize and improve its operation by establishing trends and anticipating faults before they occur. Several time series data mining algorithms have become available for efficient subsequence search and classification which can be adapted for domain-specific load profiling. We present an application of the Matrix Profile (MP) technique to energy time series for large commercial building load modelling. Several results are discussed which concern discord identification, building-specific MP values distributions as well as the effect of the particular distance metrics on the resulting processed input time series. Model free load forecasting can also serve as a suitable baseline for more advanced methods with contextual variables. Working with higher level information pieces leads to a speed up of the analysis and eliminates redundant raw data which makes the processed data suitable for online algorithm implementation and real-time building energy management.**

***Index Terms*—time series, data mining, energy management, smart buildings**

## I. Introduction

Ubiquitous deployment of Internet of Things (IoT) class devices in the energy system results in large quantities of data that have to be efficiently processed for real-time decisions. Such devices can integrate direct measurements of energy parameters such as: voltage, current or frequency, indirect measurements and computed metrics, alongside environment and contextual parameters. The design of such a system is described in [1] which presents the integration within an industrial fieldbus system based on RS-485 and a suitable web interface for data vizualization and programmatic access. Industrial-grade technologies are used in order to provide a robust solutions which is compatible with commercial implementations for long term, reliable operation. Several types of protocols and standards such as OPC-UA and MQTT as well as data formats such as XML, JSON, CSV, have become common to enable convergence between the industrial automation domain (OT - Operational Technology) and the informtation technology domain (IT). These broadly represent implementations of the Industrial Internet of Things (IIoT) in the practical domain, which enables higher level tasks on the collected data: processing, analysis and optimization.

The built environment is a salient application domain of new approaches for instrumenting and optimizing energy use through data processing and IoT systems. A blueprint for designing Building Energy Management Systems (BEMS) is described by [2]. It provides an optimization model for deciding energy use priorities which factors in energy sourcing, economic and environmental constraints. Two real use cases are presented through buildings in Austria and Spain to illustrate the improved performance, while accounting for the dominant role of the heating and cooling subsystems of the energy footprint of the building - as the area most prone to economic optimization given its large share of the total building energy use. [3] investigates the role of local energy storage to assure optimal energy management within a building. The salient finding suggests that thermal energy storage, in the form of heat or cold, provides better results than electrical energy storage through batteries given their particular limitations with regard to cost, cycle lifetime and charging/discharging profile limitations. The working definition for energy management within a building that we consider for the context of this paper includes monitoring and analysing energy usage in order to optimize and conserve energy while maintaining suitable level of service e.g. comfortable conditions for the occupants and reliable power supply to various building subsystems.

An in-depth study for building energy time-series modelling for discord identification in building energy load profiles is presented in [4]. Daily load profiles are constructed for both commercial and residential applications, while anomalies are determined by using statistical tests on the matrix profiles as proxy for typical load patterns. The Kolmogorov-Smirnov (KS) test is used to discriminate unusual daily profiles from "normal" profiles constructed on the aggregated building data at daily timescales. This statistical approach evaluates the null hypothesis $H_0$ that two data instances adhere to the same distribution versus the alternative hypothesis $H_A$ that two data instances do not adhere to the same hypothesis. The $p$-value is computed for the confidence level associated with this test with small values (1% and 5% can be used as cut-off values) indicating the possibility of rejecting the null hypothesis. Daily profiles that do not reject the null hypothesis are flagged for subsequent processing.

The work has been funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

Similar application can be found is other domains as well, such as anomaly detection for manufacturing systems using intelligent data processing in [5]. The authors present and use case in pharmaceuticals production where continuous data streams are acquired through plant automation systems and the MP algorithm is used for discord identification on dedicated computing nodes in online operation. This highlights the broad applicability to a wide range of sensor data processing challenges, beyond the current energy use case.

The main contributions of the paper are two-fold:

- Justification of the MP algorithm for domain-specific information extraction from building energy time series;
- Experimental results for exploratory data analysis (EDA), discord analysis and model-free forecasting, on public datasets which allow for benchmarking approaches.

Our previous recent contributions to the field have concerned multi-level models for anomalies that enable local control [6], deep learning using recurrent neural networks for energy time series forecasting [7] and intelligent energy management for buildings integrating state-of-the-art techniques from the recent literature [8].

The rest of the paper is structured as follows. Section II presents a conceptual overview of the Matrix Profile algorithm together with a discussion on discord analysis and the distance metrics that can be used for determining this higher level representation of the collected data. A specific case study and associated results are discussed along three main directions in Section III: evaluation based on examples from a public large commercial building dataset, the application of MP for discord identification and the impact of different distance metrics for constructing the time series profiles and illustration of the use of MP for model-free forecasting with Huber loss regression function. Section IV concludes the paper with outlook on potential developments and future work.

## II. Time Series Data Mining using the Matrix Profile

Matrix Profile (MP) is a relatively new methodology used for time series data mining, introduced by Yeh et al. [9]. Matrix Profile is arguably a dimension reduction approach which requires less training time, data and parameter tuning compared to other data mining methods. A Matrix Profile of a time-series $T$ of length $n$ is a compact time-series that stores the z-normalized Euclidean distance between each subsequence of length $m$ and its nearest neighbor. The two very important use cases of Matrix Profile are: finding similar patterns among a time-series i.e. *motif* detection and anomaly discovery for multivariate time-series i.e. *discord* detection. Various software algorithms implement the method which is suitable for online implementation with high performance.

According to [10] a time-series *motif* is the most similar subsequence pair of a time-series. Considering a time-series $T$ and a two time-series subsequences of length $m$, $\{T_{a,m}, T_{b,m}\}$ is considered a $motif$ pair if:

$$dist(T_{a,m}, T_{b,m}) \leq dist(T_{i,m}, T_{j,m}), \quad (1)$$

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

From an alternative perspective, one of the representative applications for the Matrix Profile is finding *discords* in a time-series. A *discord* is the most unusual subsequence within a time-series. More specific, the subsequence that has the maximum distance to its nearest non-self match neighbor can be interpreted as an unusual subsequence or anomaly. Considering a time-series subsequence $T_{c,m}$ of length $m$ non-self match with $T_{d,m}$ and a subsequence $T_{p,m}$ non-self match with $T_{q,m}$, $T_{c,m}$ is a discord if:

$$min(dist(T_{c,m}, T_{d,m})) > min(dist(T_{p,m}, T_{q,m})), \quad (2)$$

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

For distance calculation between subsequences, Matrix Profile uses the z-normalized Euclidean distance or more general p-norm. The z-normalised Euclidean distance is described by the following formula [11]:

$$D(X, Y) = \sqrt{2m(1 - corr(X, Y)} \quad (3)$$

where $X$ and $Y$ are time-series and $m$ represents the length of a sequence.

$corr(X, Y)$ is the covariance of the two variables $(X, Y)$ divided by the product of their standard deviations or shortly, Pearson's Correlation Coefficient and is described by:

$$corr(X, Y) = \frac{\sum_{i=1}^{m}(X_i Y_i - m\mu_X \mu_Y)}{m\sigma_X \sigma_Y}, \quad (4)$$

where $\mu$ is the mean and $\sigma$ is the standard deviation.

$$\mu_X = \frac{\sum_{i=1}^{m} X_i}{m}, \mu_Y = \frac{\sum_{i=1}^{m} Y_i}{m} \quad (5)$$

and

$$\sigma_X^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu_X)^2, \sigma_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \mu_Y)^2. \quad (6)$$

For the current research two more distance metrics have been proposed. The Matrix Profile algorithm has been adapted using Manhattan and Chebyshev distances respectively, in order to have comparison results and to see which of the distance metrics helps obtaining the best results.

The Manhattan distance is described by the equation 7 and Chebyshev distance formula is described by equation 8.

$$D(X, Y) = \sum_{i=1}^{m} |X_i - Y_i| \quad (7)$$

$$D(X, Y) = \max_i (|X_i - Y_i|) \quad (8)$$

From a computational point of view, the Manhattan distance is faster to compute while the Chebyshev distance takes longer to compute than the standard Euclidean distance metric. Introducing a selection of the computed distance metric as a parameter in the MP calculation allows for flexibility in the implementation.

Combining motif and discord discovery obtained with the MP, classification models can be developed to efficiently learn patterns that can be integrated into higher level decision schemes. Efficiency is closely related to the significant dimension reduction achieved by labelling the relevant components of the energy time series while avoiding full processing of the datasets. As a side benefit, this inherently eliminates most of the redundant information contained in the input data.

## III. Results

For the purpose of our study, we use the Building Data Genome (BDG) database [12] to test our information extraction approach for consumer-side building energy time series. This provides a curated collection of energy readings from more than 500 (mostly academic) buildings from various regions, climates and with a good mix of dominant power usage. Active power readings are sampled at 1h for a full year, with context e.g. local temperature and meta-information e.g. usage profile and building size. The database serves as a good support for interdisciplinary research in this area in what concerns, models and algorithms for energy system control and optimization.

### A. Exploratory Data Analysis for Building MP Series

We select from the BDG dataset, three representative buildings with different dominant usage profile: office, laboratories and classroom. All buildings are from an academic campus located in Zurich, Switzerland, which mitigates external environmental factors to be compensated in the analysis. We first compute the MP vectors for each of the buildings, that we use for further analysis. The z-normalisation embedded into the MP method eliminates the difference in absolute energy use, partly determined also by the difference in the building size. The reference buildings are subsequently identified through their short names, as identified in the raw input files: "Travis", "Tracy", "Teri".

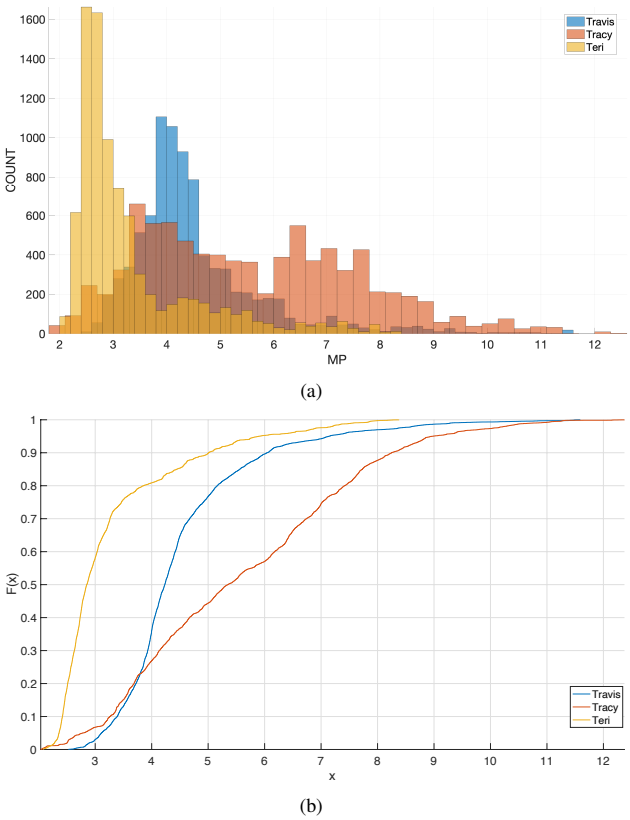
In Table I, the main statistical indicators are highlighted based on the pre-computed MP values for each building. The minimum, maximum, average $\mu$ and standard deviations $\sigma$ are reported. In addition to these basic metrics, we also report the skewness and kurtosis indicators. There provide additional information with regard to the shape of the underlying probability distribution of the computed values, as compared to an ideal gaussian distribution. Skewness is computed as $s = E(y-\mu)^3 \backslash \sigma^3$ with negative and positive values indicating left and right unabalancing of the data around the sample mean. With regard to kurtosis $k = E(y-\mu)^4 \backslash \sigma^4$, lager values e.g. higher than three indicate increased tailedness of an error prone underlying probability distribution.

The data in the table is grouped by categories, associated to the MP subsequence length parameter $m$ used for obtaining the profiles for each building: D - daily subsequences of length 24, W - weekly subsequences of length 168 and M - monthly subsequences of length 720. The total size of the time series is 8760 corresponding to a full year of hourly readings.

TABLE I
MP Values Statistical Indicators

| | **MP** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Teri* | | | *Traci* | | | *Travis* | | |
| | D | W | M | D | W | M | D | W | M |
| MIN | 5.8 | 2.9 | 2.5 | 7.7 | 3.3 | 2 | 6.6 | 2.1 | 2.1 |
| MAX | 9.3 | 9.3 | 9.8 | 8.7 | 8.7 | 9.5 | 8.2 | 11.8 | 11.8 |
| $\mu$ | 7.9 | 4.4 | 4.4 | 8 | 4.7 | 5.5 | 7.6 | 3.8 | 3.4 |
| $\sigma$ | 1.4 | 1.7 | 1.1 | 0.2 | 1.7 | 1.8 | 0.6 | 2 | 1.2 |
| s | -0.3 | 1.7 | 1.5 | 1 | 1.1 | 0 | -0.3 | 1 | 1.4 |
| k | 1.4 | 5.1 | 5.7 | 3.1 | 2.4 | 1.9 | 1.3 | 2.3 | 4.5 |

Figure 1 shows the histogram for the three analysed buildings alongside the empirical cumulative distribution function (ECDF). Main finding is that we observe distinguishing shapes of the MP values histogram based on dominant usage profile. In this case the classroom building has most of the MP values concentrated at lower levels which indicate a recurring subsequence similarity given the deterministic usage schedule of such spaces according to predetermined. In contrast, the laboratory building is more prone to distinguishing energy use patterns and anomalies. These are reflected by a somewhat uniform distribution of the MP values, including larger values. The office building profile is in between the two with a smaller peak, shifted to the right of the classroom building.




(a)

(b)

Fig. 1. Analysis of MP Values: (a) Histogram (b) ECDF

The ECDF plot illustrates how 90% of the values are below 5, 6 and 8.2 for the classroom (Teri), office (Travis) and laboratory (Tracy) buildings respectively. The increased steepness of the curves for the Travis and Teri buildings are directly associated with the increased density of the values in different areas of the histogram at 2.5 and 4.5 values. The smoother progression of the laboratory building can be also determined by somewhat inconsistent and aperiodic variations of the local energy use which can be caused by sporadic use of particular energy consuming laboratory equipment.

### *B. Discord Analysis and Effect of Distance Metric Selection*

Figure 2 illustrates the active power readings in kW for a half year period in the case of the Travis office building. MP is computed with a weekly subsequence length parameter $m = 168$ and presented in the bottom part of the figure. The red dots mark the identified top-k discords with $k = 3$. In order of importance these occur on the following dates: December 27th, December 2nd and September 6th. For all time series that were investigated one of the discords resulting from the MP vector was placed in the period between Christmas and New Year's Eve which can indicate the fact that, from the perspective of the active power consumption, this period is highly dissimilar to any other period during the year. This is valid for all types of dominant building usage including offices, laboratories and classrooms and can be clearly assigned to the winter holiday. Other discords exhibit a strong correlation with either certain periods of the academic year and with the building dominant usage type. These are however strongly dependent on local particularities e.g. the university scheduling, country, etc. and other contextual information such as weather. Identifying discords can be a promising bootstrapping solution to obtain labelled energy datasets that can be useful for higher level supervised learning tasks. In this situation a model can be built to anticipate inconsistent energy forecasts and adjusts energy management schemes dynamically.

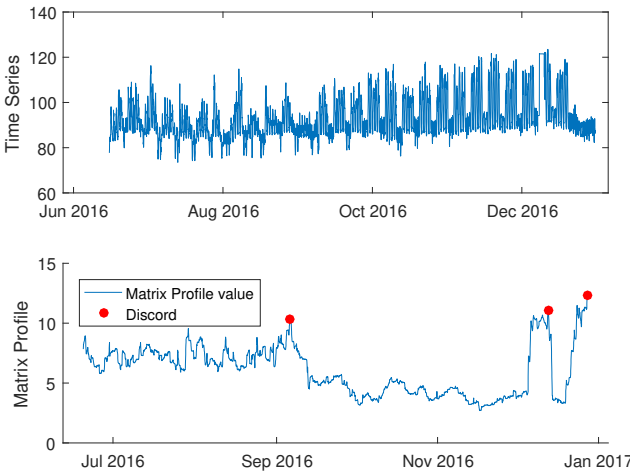



Fig. 2. Sample Input Data and Associated MP with Discords

Figure 3 shows a boxplot graphic of the MP values of a building grouped by weekdays versus weekends. This could suggest a difference between the two groups with higher average values during weekends, 4.2 versus 2.8. This can be explained through the periodic structure of the workdays in what concerns power consumption. The weekend values have a wider interval of variation and it can be observed as well the large number of outliers that fall outside the 1.5 times the Inter Quartile Range (IQR). This can be explained through factors that produce larger absolute variations of the total power consumption of the building throughout the year which produces highly dissimilar weekly subsequences. This is considered to be a characteristic of commercial building types as compared to residential ones.

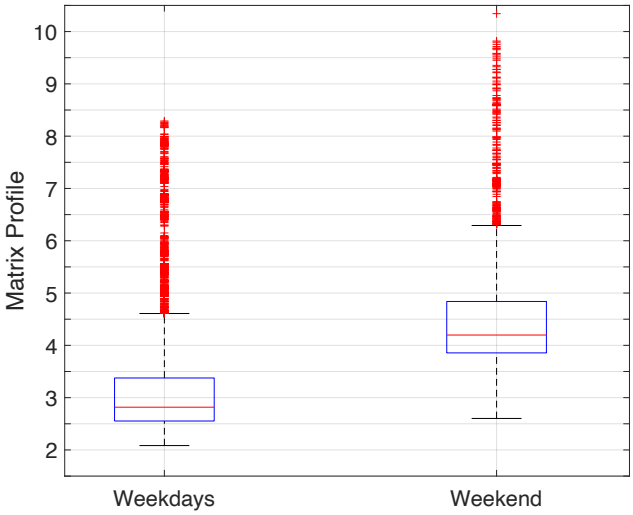



Fig. 3. Sample of Weekdays vs. Weekend boxplot (Travis)

Figure 4 shows a boxplot graphic of the MP values of a building grouped each day of the week. The distribution is on average similar during every day of the week. Fridays can be observed to have a wider variation range which can be assigned to longer weekends and shorter schedules in different times of the year at the end of the work week. Further investigations can be carried out by relating the MP values to typical daily schedules with lower subsequence length which can provide fine grained insights. The probabilities distribution built based on these values serve as a reference parametric model for certain classes of buildings.

We have further implemented the Manhattan distance, as alternative to the reference MP algorithm, and Figure 5 presents the comparative view in reference to the default Euclidean distance for the period associated with the first discord in the month of December. It can be seen hot the Manhattan MP calculations result in a noisier time series compared to the standard version. Smoothing out this time series would results in a highly similar distance metric profile with an advantage in the computational time required.

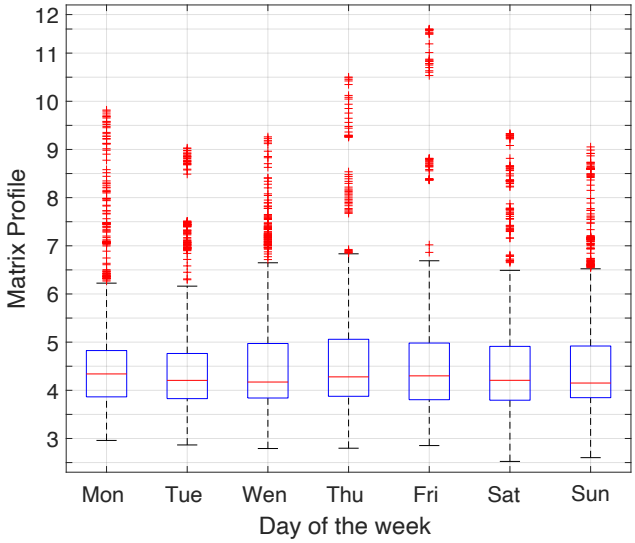



Fig. 4. Sample of Day of the week boxplot (Travis)

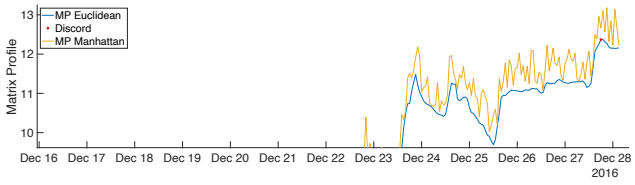



Fig. 5. Manahattan and Euclidean Distance MP

### C. Model free load forecasting using MP

Common metrics to evaluate prediction performance for regression tasks include the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). These two metrics are computed as follows:

$$MSE = \sum_{1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \tag{9}$$

$$MAE = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \tag{10}$$

with $y_i$ the real sample $i$, $\hat{y}_i$ the estimated or forecasted value and $n$ the number of samples. More specifically, MSE quantifies the squared bias of the estimates plus the variance, while MAE only accounts for positive variations from the real values. MSE penalizes larger prediction errors comparative to MAE. In order to revert to the same units of measure (non-squared) the square root of the MSE can also be considered in the form of RMSE.

Depending on the way that extreme prediction errors should be handled, either by emphasizing them as in the case of MSE or discounting them for MAE, a suitable trade-off can be achieved by implementing the Huber loss function [13]. Huber loss is expressed as:

$$L_\delta = \begin{cases} 1/2(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - 1/2\delta^2, & \text{otherwise} \end{cases} \tag{11}$$

where $y$ is the actual value and $\hat{y}$ is the predicted value. The metric is parametrised through the value of $\delta$ where values near zero lead to a quadratic formulation of the metric, similar to the MSE, and a large values lead to the MAE. In this manner domain specific objectives for building automation can be accounted for, for example in the case of high charges for peak loads versus constant accumulation of prediction errors in estimating local consumption.

Figure 6 presents the result of model-free MP-based load forecasting for the Travis building at hourly sampling rates. This type of approach can be compared as baseline to new methods such as SARIMA models and Long Short Term Memory neural networks for time series. The prediction performance in this case is lower but for some use cases it can be compensated by online forecasts without computationally intensive model selection and training.

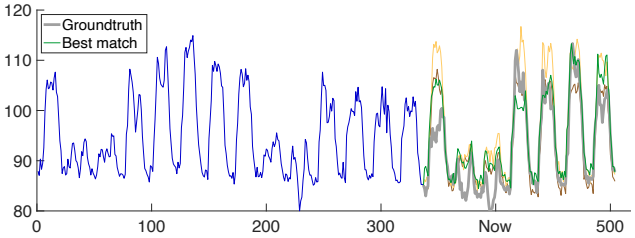



Fig. 6. Load Forecasting using MP

## IV. Conclusion

The paper presented an approach to extract useful information from large commercial building energy traces by using a state of the art technique. We have presented results which concern exploratory data analysis on MP value vectors, identification of discords and the effect of distance metric selection as well as using the MP for model-free load forecasting as baseline technique. Given fast performance of the algorithms, the approach is suitable for implementation in embedded building energy management techniques for real-time local control. As many public datasets of commercial building energy traces are now publicly available, the method is scalable to analyze them for more robust conclusions.

Future work will be focused on the integration of this approach in decision support systems for local (micro-)grids [14] with focus on the active role of large commercial buildings in demand response and peak shaving strategies. We plan to scale up the modelling to provide a generalized methodology for new richer datasets with increased contextual information [15]. New open-source libraries that implement the MP technique allow the integration of the methods into common data science environments and workflows with larger adoption that enables productive applications for industry and energy.

## REFERENCES


[1] H. Luan and J. Leng, "Design of energy monitoring system based on iot," in *2016 Chinese Control and Decision Conference (CCDC)*, 2016, pp. 6785–6788.

[2] P. Rocha, A. Siddiqui, and M. Stadler, "Improving energy efficiency via smart building energy management systems: A comparison with policy measures," *Energy and Buildings*, vol. 88, pp. 203 – 213, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778814010469

[3] Z. Xu, X. Guan, Q. Jia, J. Wu, D. Wang, and S. Chen, "Performance analysis and comparison on energy storage devices for smart building energy management," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 2136–2147, 2012.

[4] J. Y. Park, E. Wilson, A. Parker, and Z. Nagy, "The good, the bad, and the ugly: Data-driven load profile discord identification in a large building portfolio," *Energy and Buildings*, vol. 215, p. 109892, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778819332578

[5] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert, "Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings," *Sensors*, vol. 19, no. 24, p. 5370, Dec 2019. [Online]. Available: http://dx.doi.org/10.3390/s19245370

[6] 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.

[7] C. Nichiforov, G. Stamatescu, I. Stamatescu, V. Calofir, I. Fagarasan, and S. S. Iliescu, "Deep learning techniques for load forecasting in large commercial buildings," in *2018 22nd International Conference on System Theory, Control and Computing (ICSTCC)*, 2018, pp. 492–497.

[8] C. Nichiforov, G. Stamatescu, I. Stamatescu, I. Făgărăşan, and S. S. Iliescu, "Intelligent load forecasting for building energy management systems," in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*, 2018, pp. 896–901.

[9] 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.

[10] C.-C. M. Yeh, "Towards a near universal time series data mining tool: Introducing the matrix profile," *ArXiv*, vol. abs/1811.03064, 2018.

[11] D. De Paepe, D. Nieves Avendano, and S. Hoecke, *Implications of Z-Normalization in the Matrix Profile*, 01 2020, pp. 95–118.

[12] C. Miller and F. Meggers, "The building data genome project: An open, public data set from non-residential building electrical meters," *Energy Procedia*, vol. 122, pp. 439 – 444, 2017.

[13] J. Niu, J. Chen, and Y. Xu, "Twin support vector regression with huber loss," *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 6, pp. 4247–4258, 2017.

[14] I. Stamatescu, N. Arghira, I. Făgărăşan, G. Stamatescu, S. Iliescu, and V. Calofir, "Decision support system for a low voltage renewable energy system," *Energies*, vol. 10, no. 1, p. 118, Jan 2017. [Online]. Available: http://dx.doi.org/10.3390/en10010118

[15] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J. Y. Park, Z. Nagy, P. Raftery, B. W. Hobson, Z. Shi, and F. Meggers, "The Building Data Genome Project 2 – Energy meter data from the ASHRAE Great Energy Predictor III competition," *arXiv e-prints*, p. arXiv:2006.02273, Jun. 2020.