CMPUT 697 Project Report

Unsupervised Anomaly Detection by Analyzing Power Consumption of Individual Households

Xinlei Chen
University of Alberta
Department of Computing Science
xinlei1@ualberta.ca

1 Introduction

Over the past few years, the average global temperature is significantly increasing due to carbon dioxide (CO_2) and other human emissions into the atmosphere. The commercial and residential buildings are one of the largest contributors to global CO_2 emissions. During the building operations, significant energy is wasted due to inefficient utilization of resources or misuse of appliances and equipment in disrepair. Strategies to help increase energy efficiency in buildings are needed.

Anomaly detection is one of the strategies to identify appliances in a state of disrepair or used improperly. Identify these abnormal behaviour can reduce the huge operational costs. It also can create alerts to either repair an appliance or suggest a more optimal use [8] [6].

1.1 Define Anomalies

There is no clear definition of the anomaly. Harkins defines an outlier as "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Anomaly detection techniques are widely used in various areas like fraud detection, intrusion detection, industrial damage detection, personal health, and sensor networks. Anomalies commonly are classified into three categories [5]:

- *Point Anomalies.* A single instance of data is anomalous if it is too far off from the rest.
- *Contextual Anomalies.* The abnormality is context-specific. For example, spending 100 dollars on food during the holiday season is normal, but maybe odd in weekdays.
- *Collective Anomalies.* A set of data instances collectively helps in detecting anomalies.

In this paper, we focus on the detection of contextual and point anomalous energy consumption in individual households.

1.2 Challenges with Anomaly Detection

There are two main challenges in detecting anomalies. The first challenge is the lack of labeled data to train an algorithm for detecting anomalous behaviours. It is expensive and time-consuming to obtain ground truth data since it requires large amount of manual work and hiring experts is costly. Besides, injecting artificial anomalies becomes necessary to evaluate different anomaly detection approaches. In this paper, we choose to use unsupervised anomaly detection techniques due to the lack of labeled data. The second challenge is the high dimensionality of time series data. Commonly used distance functions (e.x. Euclidean distance) can not well represent the actual distance between each pair of points in high-dimensional space.

1.3 Paper Contributions

In this paper, we make the following contributions:

- We investigate different feature construction methods for time series data.
- We compare the performance of different unsupervised anomaly detection methods.

1.4 Paper Organization

Rest of the paper is designed as follows: Section 2 explains the related work. We describe the detail of our work in section 3. The experiment setup and results are discussed in section 4. At the end, we conclude our paper and show some future work.

2 Related Work

Anomaly detection raises more and more attentions in recent years, with surveys appearing covering: anomaly detection [4], novelty detection [15], and outlier detection for temporal data [9]. We now focus on the most related work to our project.

2.1 Power Consumption Analytics

For power consumption analytics, different applications have been researched. Zhenjun et al. [13] focus on cluster analysis strategy to identify typical daily heating energy usage profiles of higher education buildings. Gowtham et al. [2] try to understand campus-scale power consumption. The anomaly detection always is a hot topic in analyzing power consumption. Jakkula and Cook [11] compare statistical with unsupervised clustering-based techniques for detecting periods of unexpected consumption. Jaime and Bo [17] analyze the feasibility of applying outliers detection algorithms for enhancing the security of AMI through the detection of electricity theft in a variety of types. Megha et al. [8] propose two novel methods to generate labeled data for abnormal energy consumption and investigate different performance metrics used in anomaly detection.

2.2 Unsupervised Anomaly Detection

The typical assumption in the unsupervised setting is that outliers represent potential anomalies. In the density-based outlier detection method, a point is identified as an outlier if its density is relatively much lower than that of its neighbours. Breunig et al. [3] introduce a degree of an object being an outlier – local outlier factor (LOF), which measures how isolated the object is with respect surrounding neighbourhood. He et al. [10] propose the cluster-based local outlier factor (CBLOF) which combines distance-based unsupervised clustering and local outlier factor. Liu et al. [12] propose a method called Isolation Forest (iForest), which detects anomalies purely based on the concept of isolation without employing any distance or density measure. Finally, Fan et al. [7] investigate the potential of autoencoders in detecting anomalies in building energy data.

3 Approach

We can formally define the task addressed in this paper as follows:

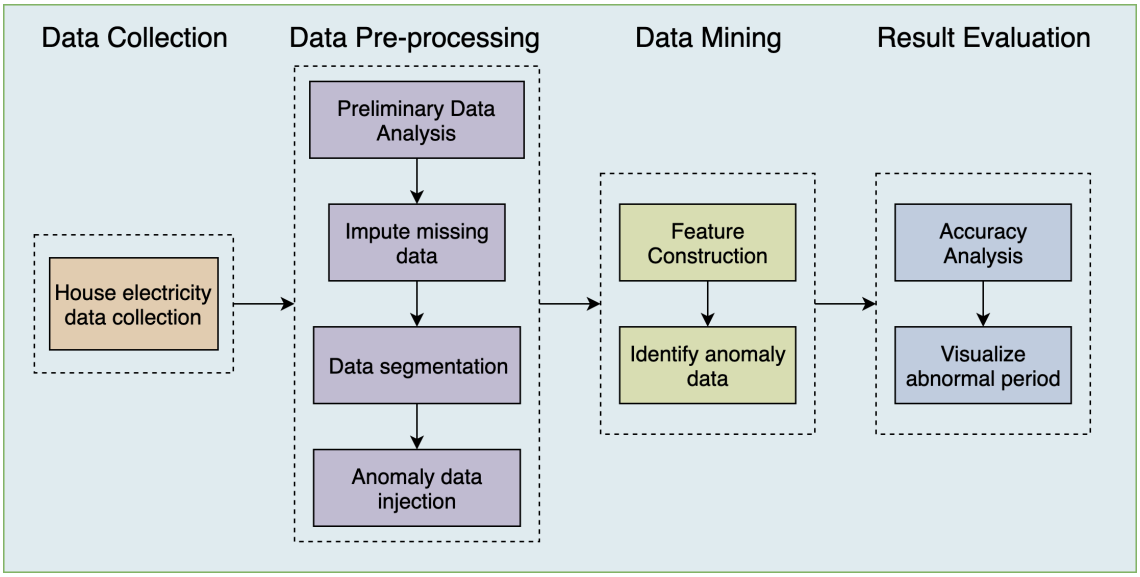


Figure 1: The outline of project.

Given: A time series $T = \{(t_1, v_1), ..., (t_n, v_n)\}$, where t_i is a timestamp and v_i is the power consumption at time t_i, for a single household.

Do: Identify hourly periods of anomalous power consumption in T.
The outline of the whole project is illustrated in Figure 1.

3.1 Dataset

The dataset used in this project is collected from Dataport (Pecan Street) [14]. It consists of 239 houses located in Texas, US. Each house has meter-level data, which are sampled at 1 minute intervals. We are using the whole 2018 year power consumption data from four houses.

3.2 Preliminary Data Analysis

Each household's average energy consumption and the standard deviation are presented in Table 1. The households' average energy consumption range from 1.098 kWh to 1.459 kWh per minute with standard deviations from 0.958 kWh to 1.665

Table 1: Average energy consumption per minute and standard deviation for each household

House	Mean (kWh)	STD (kWh)
370	1.459	1.364
3506	1.098	0.958
4830	1.252	1.665
2814	1.294	1.417

kWh.

Two weeks of energy consumption for four households are shown in Figure 2. We can observe that the power consumption patterns of the households are distinctive. Since different house has a different set of characteristics such as size, location, amount of family members, we choose to perform anomaly detection on a house-by-house basis based on above observation.

The energy consumption of House 370 in the same weekday (Tuesday) is shown in Figure 3. As we can see, these two days' power consumption follow similar usage pattern. For example, they both consume a large amount of energy from 11:00-13:00 while they keep low consumption from 3:00-7:00.

3.3 Preprocessing

This phase includes two steps — imputing missing data and data segmentation.

At the first step, we check the data incompleteness that may arise due to network or sensor failure. We apply linear interpolation to fill these missing values.

At the second step, within each house, we subdivide the data T (power consumption of each house) into non-overlapping windows of one hour (i.e., 00:00-01:00, 02:00-03:00, etc.). The reason why we work on the house-level is based on the observation of Figure 2 in the "Preliminary Data Analysis" section. We divide the data within a house into the non-overlapping one hour windows for two reasons. First, the time of day is an important feature that is simplicity captured by this partition. Second, it is a proper unit in which enough data is available for house owners to identify what

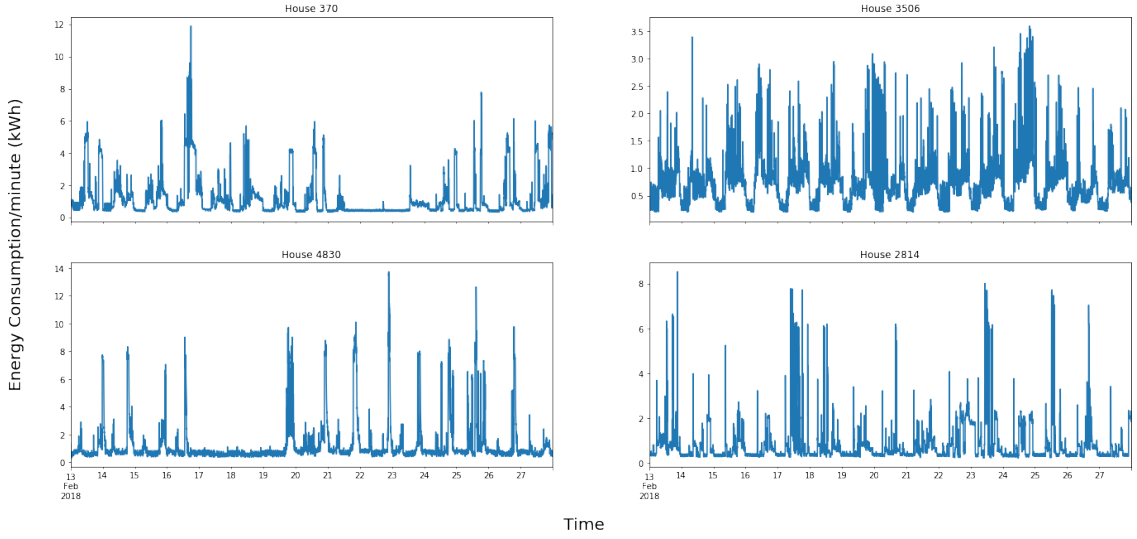


Figure 2: Energy consumption of four different households from February 13-27, 2018.

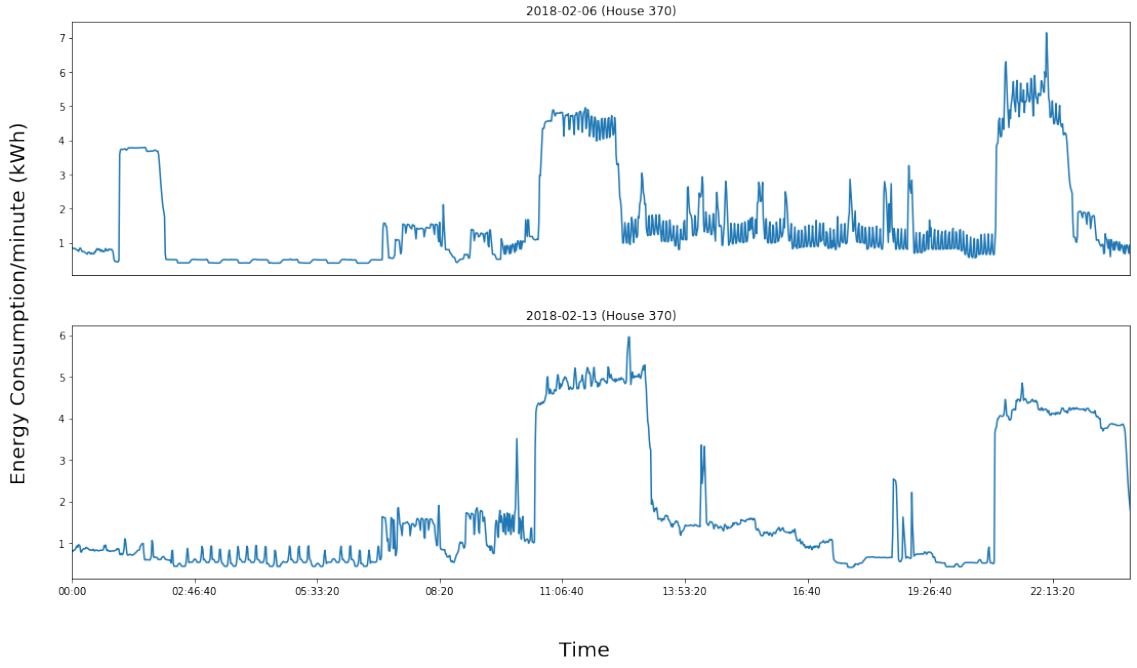
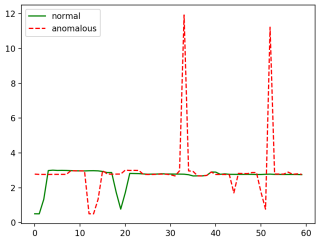
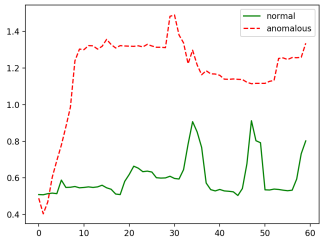


Figure 3: Tuesday Energy consumption in House 370 from Feb. 06 and Feb. 13.



(a) Type I Anomaly



(b) Type II Anomaly

Figure 4: This figure shows two types of anomaly.

appliances are in disrepair or improperly used.

3.4 Anomaly Data Injection

In order to evaluate different approaches, we manually inject artificial anomaly consumption into our data. Since we construct 24 window groups (one for each one hour interval) within each house, we replace some of normal data in each window group with abnormal consumption data. In each window group, we keep 2% abnormal power consumption and 98% normal power consumption. Above operations are based on assumption that all the origin data are normal data before our injection. There are two types abnormal consumption data.

- **Type I Anomaly**: This type of anomaly which appears because of faulty readings generated by a sensor. In order to simulate this kind of anomaly, we add unexceptional spikes into the one hour windows we choose to become anomalous. The spike means unusual high power usage occurring within short time intervals.

- **Type II Anomaly**: We count a power consumption pattern that rarely or never happens in a specific period as an anomaly in that period (window group). This type of anomalies is contextual anomalies. Since it is complex to simulate this kind of anomalies, we use the one hour windows from different houses

and different time periods to replace the one hour windows we choose to be anomalous.

These two types of anomalies often are observed during the process of data collection. Figure 4 shows the examples of two types of anomaly.

3.5 Feature Construction

Feature construction is a domain-specific task that transfers each one hour window into a feature vector that describes the characteristics of the signal during that particular window. Different classes of features are constructed as follow:

(a) *Summary statistic:* We compute the following eight statistical features: max, min, median value, standard deviation, skewness, kurtosis, and entropy.

(b) *Descriptive features:* We consider three categorical features: the day of the week, the month of the year, and whether a day is a US holiday. These categorical features are dealt with using one-hot encoding.

(c) *Fast Fourier transfrom (FFT):* We compute the frequency spectrum of each one hour window. Given a one hour window $x[n]$, for $n = 1, ..., N$, (For one minute granularity, N will be 60 in our project) its frequency spectrum is computed as

$$X[k] = \sum_{n=1}^{N} x[n] * exp(-j2\pi(k-1)\frac{n-1}{N}),\ 1 \leq k \leq N$$

Let $Y[k] = |X[k]|,\ k = 1, ..., N$ denote its magnitude.

Since traditional distance functions like Euclidean distance function do not work well in high dimensional space, we use principal component analysis (PCA) technique for dimensionality reduction.

In our experiment, unless otherwise specified, the default feature construction method we use is FFT+PCA(3), which means that we first compute the frequency spectrum of one hour window and then reduce dimensionality to 3 by applying PCA.

3.6 Unsupervised Outlier Detection Algorithms

As we describe data segmentation in the "Preprocessing" section, we construct 24 window groups (one for each one hour interval) within each house. We detect anomalies separately in each window group. The grouping is motivated by the importance of time of day as a factor governing energy consumption. Labeling is always done on the level of window, so construction anomalies are always a multiple of the window length. This is analogous to multi-instance learning: if any behaviour in a window is anomalous, then the window is anomalous.

We compare the performances of anomaly detection in following approaches:

- **LOF** is a density based unsupervised outlier detection technique [3].
- **CBLOF** is a cluster based unsupervised outlier detection technique [10].
- **IF (iForest)** is a method that explicitly isolates anomalies rather than profiles normal instances [12].
- **iNNe** is an anomaly detection technique by isolation using nearest neighbour ensemble [1].
- **kNNo** is a distance-based unsupervised outlier detection technique (in other work also referred to as kNN) [16].

These approaches are selected based on recent extensive empirical evaluations of unsupervised outlier detection algorithms.

4 Experiment

We report the results on the evaluation of the approaches. We answer the questions:
Q1: Which approach does perform best in anomaly detection?
Q2: How do different feature construction methods contribute to the performance?
Q3: How do approaches perform on different types of anomalies?

4.1 Experimental Setup

4.1.1 Experimental setup

The goal of the experiment is to evaluate the performance of different anomaly detection methods. We perform the following experiment for each of the four houses separately. First, we get 24 window groups by preprocessing the power consumption data T (one minute granularity). Second, we inject two different types of anomalous power consumption separately as we describe in the "Anomaly Data Injection" section (we only inject one specific type of anomalies in each time). Third, we use different feature construction methods to convert each one hour window to feature vector. As we mention above in section 3.5, the default feature construction method is FFT+PCA(3). Finally, we apply different approaches to detect anomalies separately in each window group. For each approach, we pick the parameter setting that maximizes area under the ROC curve (AUC).

4.1.2 Evaluation metrics

Each anomaly detection method outputs a ranking over the one hour windows within each window group from most to least anomalous. We evaluate these rankings by computing the Precision, Recall, True negative rate (TNR), False positive rate (FPR), F_1 Score and AUC, which are discussed in [8]. The Precision, Recall, TNR, FPR, F_1 Score are computed based on the setting that the number of anomalous windows is 2% of the total windows within each window group.

4.1.3 Hyperparameters

We use grid search for each approach parameter. LOF and kNNo each have one parameter k: the number of nearest neighbour, which is optimized over interval [1,30]. IF and iNNe both have parameters sub-sampling size ψ and number of estimators t. For IF, we set sub-sampling size to 256 and number of estimators to 100 as the authors suggest [12]. For iNNe, we choose sub-sampling size from set $\{2, 8, 64, 256\}$ and also set number of estimators to 100. CBLOF has three parameters: the number of clusters

n_c which is optimized over [1,30], and two parameters needed by "FindCBLOF" are set to 90% and 5 separately [10].

4.2 Result

Table 2: Average evaluation metric of identify Type I anomalies for each method across all 96 window groups. The best result of each method is shown in bold.

Method	Precision↑	Recall↑	TNR↑	FPR↓	F_1 Score↑	AUC↑
LOF	0.4414	0.5045	0.9875	0.01248	0.4708	0.9003
IF	0.4089	0.4673	0.9868	0.01321	0.4361	0.9108
iNNe	**0.4453**	**0.5089**	**0.9876**	**0.01240**	**0.475**	0.9108
kNNo	0.4180	0.4777	0.9870	0.01301	0.4458	**0.9160**
CBLOF	0.4115	0.4702	0.9868	0.01315	0.4389	0.9130

Table 3: Average evaluation metric of identify Type II anomalies for each method across all 96 window groups. The best result of each method is shown in bold.

Method	Precision↑	Recall↑	TNR↑	FPR↓	F_1 Score↑	AUC↑
LOF	0.4219	0.4821	0.9871	0.01292	0.45	0.8452
IF	0.4284	0.4896	0.9872	0.01277	0.4569	**0.8882**
iNNe	0.4297	0.4911	0.9873	0.01274	0.4583	**0.8882**
kNNo	0.3880	0.4435	0.9863	0.01368	0.4139	0.8209
CBLOF	**0.4414**	**0.5045**	**0.9875**	**0.01248**	**0.4708**	0.8796

a) **Q1:** *Approach Performance:* Table 2,3 show the average evaluation metric for each method across all 96 window groups (4 houses * 24 window groups in each house). Comparing five approaches on table 2, iNNe approach has best performance on identifying Type I anomalies. CBLOF performs best for identifying Type II anomalies. As we can see, IF and iNNe get similar performance on both table 2 and 3. This can be explained that both iNNe and IF use the same isolation forest technique.

We also observe that all approaches only get around 40%-50% recall and precision. This can be explained by that there is some potential anomalous consumption before we inject artificial anomalies and we assume all the original consumption is normal before our injection.

b) **Q2:** *Importance of features:* During the experiment, we test approaches' performance on different combinations of feature categories and report the averaged AUC across all window groups. The result shows in table 4. As we can see, different feature construction methods significantly affect the performance of approaches. For example, by using the combination of statistic and descriptive features, the AUC decrease around 0.2 comparing to FFT+PCA(3). Figure 5 shows how different feature construction methods affect the data distribution of (06:00-07:00) window group in House 370. In order to visualize the result, we reduce the dimensionality to 2. We can observe that FFT feature construction method is more likely to locate anomalies to low-density area.

Table 4: Average AUC for each method across all window groups on different feature construction methods. The best result of each method is shown in bold.

Feature Construction	LOF	IF	iNNe	kNNo	CBLOF
Origin	0.8906	0.8283	0.8282	0.8723	0.8476
FFT	**0.9105**	0.8815	0.8815	0.8994	0.8703
FFT+PCA(3)	0.9003	**0.9118**	**0.9118**	**0.9160**	**0.9185**
Statistic+Descriptive	0.7529	0.7430	0.7430	0.7556	0.6980
Statistic+Descriptive+PCA(3)	0.7504	0.7112	0.7112	0.7562	0.7317

c) **Q3:** *Performance on different types of anomalies:* Based on the metric scores show in table 2,3, we can see all approaches have similar performance on identifying Type I anomalies and Type II anomalies. It also suggests that current outlier detection approaches are all possible to detect both point and contextual anomalies.

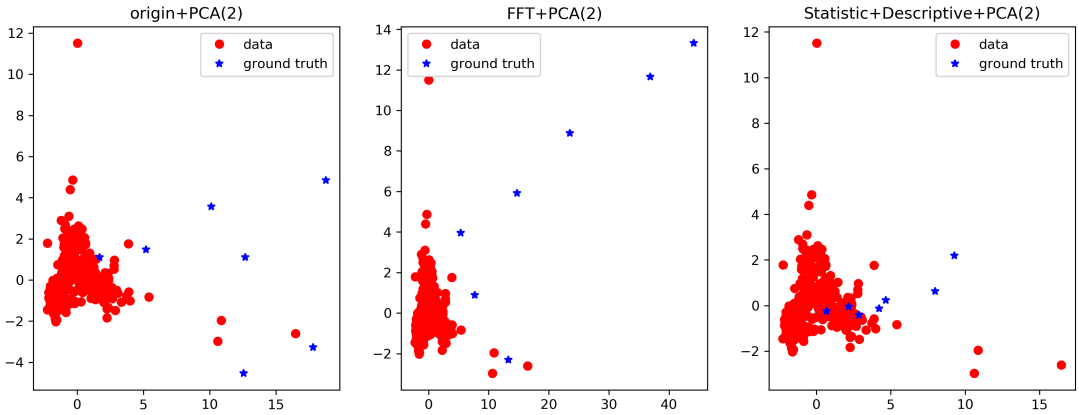


Figure 5: Data distribution of (06:00-07:00) window group in House 370 by different feature construction methods.

5 Conclusions

This paper compares the performance of different unsupervised anomaly detection techniques on real-world residential power consumption data. It also investigates how different feature construction methods affect the performance of anomaly detection methods.

Two different types of anomaly are defined and injected into the data respectively and used for evaluating different approaches. The result of experiment shows that current outlier detection algorithms are possible to detect abnormal behaviour, but requires further study.

There are multiple research directions for future work. First, we plan to investigate overlapping windows instead of non-overlapping windows of one hour. Second, we will also investigate if the addition of water consumption and weather data can improve anomaly detection performance.

References

[1] T. R. Bandaragoda et al. "Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble". In: *2014 IEEE International Conference on Data Mining Workshop*. Dec. 2014, pp. 698–705. DOI: 10.1109/ICDMW.2014.70.

[2] Gowtham Bellala et al. "Towards an understanding of campus-scale power consumption". In: (2011), pp. 73–78. DOI: 10.1145/2434020.2434043.

[3] Markus M. Breunig et al. "LOF: Identifying Density-based Local Outliers". In: *SIGMOD Rec.* 29.2 (May 2000), pp. 93–104. ISSN: 0163-5808. DOI: 10.1145/335191.335388. URL: http://doi.acm.org/10.1145/335191.335388.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Comput. Surv.* 41.3 (July 2009), 15:1–15:58. ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. URL: http://doi.acm.org/10.1145/1541880.1541882.

[5] Pramit Choudhary. *Introduction to Anomaly Detection*. 2017. URL: https://blogs.oracle.com/datascience/introduction-to-anomaly-detection (visited on 12/05/2019).

[6] M. Fahim and A. Sillitti. "An Anomaly Detection Model for Enhancing Energy Management in Smart Buildings". In: *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. Oct. 2018, pp. 1–6. DOI: 10.1109/SmartGridComm.2018.8587597.

[7] Cheng Fan et al. "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data". In: *Applied Energy* 211 (2018), pp. 1123–1135. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2017.12.005. URL: http://www.sciencedirect.com/science/article/pii/S0306261917317166.

[8] M. Gaur et al. "Performance Evaluation of Techniques for Identifying Abnormal Energy Consumption in Buildings". In: *IEEE Access* 7 (2019), pp. 62721–62733. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2915641.

[9] M. Gupta et al. "Outlier Detection for Temporal Data: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 26.9 (Sept. 2014), pp. 2250–2267. ISSN: 2326-3865. DOI: 10.1109/TKDE.2013.184.

[10] Zengyou He, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers". In: *Pattern Recognition Letters* 24.9-10 (2003), pp. 1641–1650. ISSN: 0167-8655. DOI: 10.1016/s0167-8655(03)00003-5.

[11] Vikramaditya Jakkula and Diane Cook. "Outlier Detection in Smart Environment Structured Power Datasets". In: *Proceedings of the 2010 Sixth International Conference on Intelligent Environments*. IE '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 29–33. ISBN: 978-0-7695-4149-5. DOI: 10.1109/IE.2010.13. URL: https://doi.org/10.1109/IE.2010.13.

[12] F. T. Liu, K. M. Ting, and Z. Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. Dec. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[13] Zhenjun Ma, Rui Yan, and Natasa Nord. "A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings". In: *Energy* 134 (June 2017). DOI: 10.1016/j.energy.2017.05.191.

[14] O. Parson et al. "Dataport and NILMTK: A building data set designed for non-intrusive load monitoring". In: *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Dec. 2015, pp. 210–214. DOI: 10.1109/GlobalSIP.2015.7418187.

[15] Marco A.F. Pimentel et al. "A review of novelty detection". In: *Signal Processing* 99 (2014), pp. 215–249. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2013.12.026. URL: http://www.sciencedirect.com/science/article/pii/S016516841300515X.

[16] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. "Efficient Algorithms for Mining Outliers from Large Data Sets". In: *SIGMOD Rec.* 29.2 (May 2000), pp. 427–438. ISSN: 0163-5808. DOI: 10.1145/335191.335437. URL: http://doi.acm.org/10.1145/335191.335437.

[17] J. Yeckle and B. Tang. "Detection of Electricity Theft in Customer Consumption Using Outlier Detection Algorithms". In: *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. Apr. 2018, pp. 135–140. DOI: 10.1109/ICDIS.2018.00029.