# Collect, Compare, and Score: A Generic Data-driven Anomaly Detection Method for Buildings

Haroon Rashid, Pandarasamy Arjunan, Pushpendra Singh, Amarjeet Singh
IIIT-Delhi
{haroonr, pandarasamya, psingh, amarjeet}@iiitd.ac.in

## ABSTRACT

Buildings are one of the largest energy consumers around the world. Several studies show that degraded and misconfigured devices waste upto 30% of energy in commercial buildings. In this paper, we propose Collect, Compare, and Score (CCS), a *generic* anomaly detection method that can be used across buildings following different energy usage patterns. CCS is a density based approach. We evaluate CCS using a real-world dataset, consisting of 16 weeks of data from commercial and residential buildings. We find average Area Under Curve (AUC) value of 0.92 for CSS and 0.78 for baseline method.

## 1. INTRODUCTION

In India and the US, buildings consume 47% [1] and 41%[1] of energy respectively. Major energy consuming devices within buildings include HVAC, lights, fans, refrigerators. It is found that on average buildings waste 30% of energy[2]. This wastage can be reduced by identifying misconfigured and malfunctioning devices, which result in abnormal energy usage. The abnormal power usage, i.e., any energy usage which significantly differs from previous patterns is commonly referred as an *anomaly*.

It is shown that threshold based techniques for anomaly detection result in high false positive rate (FPR) [2]. High FPR makes the anomaly detection method obsolete. Further, it is difficult to find threshold for different buildings following different usage patterns. The aim of this work is to develop a generic anomaly detection method for different buildings following different energy usage patterns. In this paper, we propose *Collect, Compare, and Score* (CCS) – a generic, unsupervised anomaly detection approach to detect point anomalies. A point anomaly refers to significantly higher power consumption for a certain duration of a day. CCS is based on Local Outlier Factor (LOF) – a proximity based outlier detection approach[3]. LOF uses a concept of local density, which for each data instance need the distance of its $k$ nearest neighbours. $k$ nearest neighbours define the locality of each considered data instance. Large distances result in low-density regions for anomalous data instances as compared to normal data instances.

We use a real-world dataset to evaluate the performance of the proposed CCS method. This dataset, at a sampling rate of hourly average, contains power consumption of both the residential and commercial buildings of IIIT-D campus. The dataset contains 16 weeks of data starting from August $1^{th}$ 2015 and ending on November $29^{th}$ 2015. Furthermore, we compare the performance of CCS with a baseline anomaly detection approach (BADA) [3]. We found that CCS provides an increase of 17.94% in average AUC value over the baseline method. The increase in AUC value due to low FPR makes CCS suitable for real deployments.

**Algorithm 1:** Collect, Compare, and Score (CCS) method for anomaly detection

**Input**: $X[M]$: A time series of power-time curve for $M$ days
**Output**: $A[M]$: Anomaly score [0-1] for $M$ days

1. Aggregate power consumption hourly for each $X_d^i$ as vector $y_i[1:24]$, where $1-24$ represent hours of day
2. Compute Discrete Fourier Transform (DFT) of each power-time curve for all the days $Y[M] = DFT(X[M])$.
3. Compute dissimilarity matrix $\Delta_{<M,M>}$ using Euclidean distance measure for all pairs of power-time curves $M$ days, where $\Delta_{<i,j>} = [\sum_{k=1}^{M}(Y[i] - Y[j])^2]^{1/2}$ .
4. Reduce the dimensionality of $\Delta$ from $M$ to 2 using Multi Dimensional Scaling (MDS) algorithm. $\widehat{\Delta}_{<M,2>} = MDS(\Delta_{<M,M>})$
5. Compute the Local Outlier Factor (LOF), $L_k[M]$ using $\widehat{\Delta}_{<M,2>}$ for different values of $k$, i.e., number of neighbours
6. Compute final LOF, $\widehat{L}[M]$ as maximum of $L_k[M]$
7. Normalise $\widehat{L}[M]$ between $[0-1]$ to compute the anomaly score for each day as $A[M] = Normalise(\widehat{L}[M])$

## 2. METHODOLOGY

The proposed CCS method takes hourly readings of several days from a single meter as input and outputs the

[1]http://www.eia.gov/forecasts/aeo/
[2]https://goo.gl/yWdaRO
[3]https://en.wikipedia.org/wiki/Local_outlier_factor


*e-Energy'16 June 21-24, 2016, Waterloo, ON, Canada*

ACM ISBN 978-1-4503-4417-3/16/06.
DOI: http://dx.doi.org/10.1145/2939912.2942354

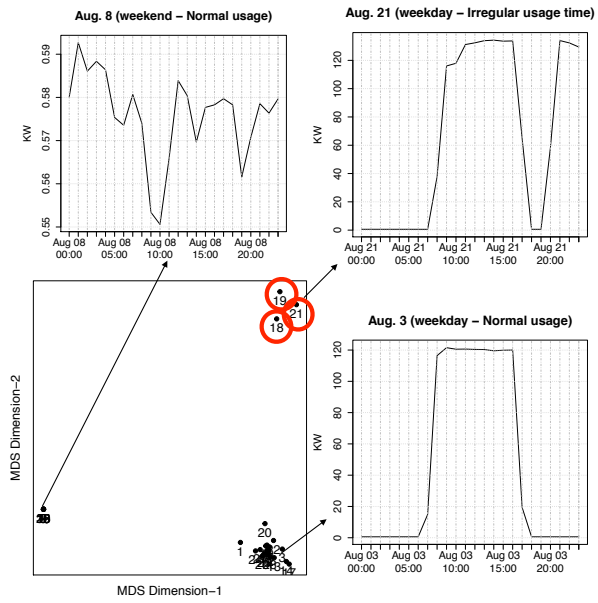


**Figure 1: Multidimensional scaling (MDS) plot of HVAC chiller and energy consumption pattern on selected days**

anomaly score for each day. The anomaly score is in the range of $0-1$. The value 0 being non-anomalous and 1 being highly anomalous. We refer power meter data collected over a 24 hour period from a single meter as power-time curve (similar to the notation used in [3]). In this work, we use a power-time curve at hourly resolution, *i.e.*, each reading of a power-time curve represents the hourly average of power consumption. CCS works in three steps: *Collect* - In this step, we collect hourly energy usage data of several days, *Compare* - Next, we compare the energy usage of several days using Euclidean distance, *Score* - Next, we assign an anomaly score to each day energy consumption using LOF. Algorithm 1 outline steps in detail.

# 3. EVALUATION

We evaluate CCS using data of different residential (Flat_1, Flat_2) and commercial buildings (HVAC chiller, LectureBlock). Further, we compare its performance with BADA. For the CCS, k (Algorithm 1, Step 5) value is set between 4 - 8 and the final anomaly score represents the maximum of anomaly scores corresponding to different $k$. For BADA, we set $k$ value to 6, calculated using the formula mentioned in the original paper.

Figure 1 shows Multidimensional scaling (MDS) plot, i.e., energy consumption distribution of HVAC chiller on the different days for the month of August. Also, it shows the energy consumption pattern on 3 different types of days. In Figure 1, we find two dense clusters corresponding to the power consumption of weekdays and weekends. Also, we observe days 18, 19, and 21 far from both the clusters and hence possibly represent the anomalous days. On analysing the power consumption of day 21, we found that Chiller was operational during night hours as shown in the subplot of Figure 1. The same reason was found for the days 18 and 19. Detecting these anomalous instances save a good amount of energy as chiller is high power consuming device. During discussions with the HVAC operator, we came to know that

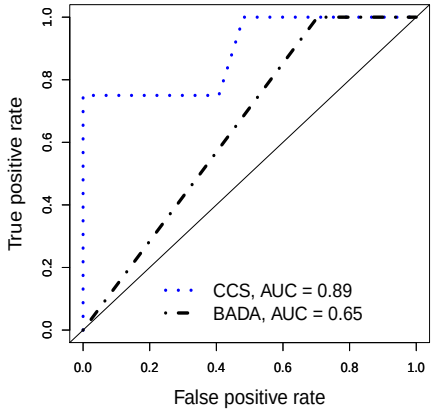


**Figure 2: ROC curve showing AUC value for HVAC Chiller with CCS and BADA methods**

| Method | Chiller | LectureBlock | Flat_1 | Flat_2 |
|---|---|---|---|---|
| CCS | 0.89 | 0.83 | 1.00 | 0.98 |
| BADA | 0.65 | 0.67 | 0.87 | 0.95 |

**Table 1: AUC values for different buildings with CCS and BADA methods**

on campus there are two chillers, which operate in alternate intervals (day and night) usually. In exceptional cases, both chillers operate simultaneously, and same reason was found for specified days.

Figure 2 shows AUC value for HVAC chiller with CCS and BADA. Table 1 shows AUC values for all the buildings with CCS and BADA methods. The low value of AUC in BADA is due to high FPR. The reason for high FPR in BADA is that it computes anomaly score for all the days with respect to a single observation having highest density, while as CCS computes anomaly score locally, i.e., anomaly score for each observation is computed with respect to neighbour observations. Therefore, in BADA a single observation dominates the anomaly scores while as no such effect is found in CCS.

# 4. CONCLUSION

CCS reduces FPR significantly thus making it suitable for real deployments. For real data collected from both commercial and residential buildings, we observed an average AUC value of 0.92 for CCS and 0.78 for the baseline method, BADA.

# 5. REFERENCES

[1] M. Evans, B. Shui, and S. Somasundaram. *Country Report on Building Energy Codes in India*. Pacific Northwest National Laboratory, 2009.

[2] B. Narayanaswamy, B. Balaji, R. Gupta, and Y. Agarwal. Data Driven Investigation of Faults in HVAC Systems with Model, Cluster and Compare (MCC). In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 50–59. ACM, 2014.

[3] G. Bellala, M. Marwah, M. Arlitt, G. Lyon, and C. E. Bash. Towards an Understanding of Campus-scale Power Consumption. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 73–78. ACM, 2011.