2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

DATA STREAMING PROPORTIONS: A FRAMEWORK FOR STREAMING DATA MONITORING AND ANOMALY DETECTION

Karim El Mokhtari and J.J. McArthur
Toronto Metropolitan University, Toronto, ON, CANADA

Abstract

Data stream integrity is critical to support robust and secure Smart Building Solutions. A framework for detecting data traffic anomalies is presented using Latent Dirichlet Allocation and the Dirichlet distribution to identify typical signal patterns across different HVAC modes, permitting real-time monitoring of network traffic to provide real-time alerts for both lost signals and potential cyberattacks. This method enables facility managers to monitor system states and diagnose deviations efficiently across multiple states with as little as two months of training data, but requires expert state interpretation. The resultant low-complexity approach provides a practical tool for real-time data streaming monitoring in dynamic environments.

Introduction

Predictive maintenance, fault detection, and online energy management solutions are increasingly developed and show significant promise to support building decarbonization and improved operational efficiency. However, such solutions are compromised by missing data resulting from sensor or network communication faults. Such signal losses can compromise system performance in real time and limit the ability to use data analytics, but are typically only discovered after the fact. Anomaly detection in data streaming is thus essential for maintaining system integrity and quickly resolving critical issues. In addition to detecting missing data, this anomaly detection can also be a valuable from a cybersecurity perspective. The risk of cybersecurity breach has posed a significant barrier to adoption for Smart Buildings (Sándor & Rajnai, 2023)but can be identified from significant spikes in network traffic indicating either direct system attacks or sensor hijacking for third-party cyberattacks.

While there is a wealth of literature on Heating, Ventilation, and Air-Conditioning (HVAC) system anomaly detection that evaluates sensor data *values* to identify potential equipment faults, for example (Chen, et al., 2022; Matetić, et al., 2022; Nassif, et al., 2021), there is a paucity of literature addressing anomaly detection in the data *traffic patterns*, notably (Chen, et al., 2025) and (Safaei, 2017)

For HVAC system data steaming from a Building Automation System (BAS), this type of detection poses unique challenges. Data streams are collected from a highly diverse set of sensors and control points, each of which may have daily, seasonal, or mode-specific reporting patterns. Traditional anomaly detection techniques are designed for static datasets, relying on access to the entire dataset, which is impractical when dealing with data streams that cannot be fully stored in memory, and are thus ill-suited for the continuous transmission and ingestion of streamed data. Further, to maintain system integrity, real-time detection is required, requiring algorithms capable of efficiently processing large volumes of data on an ongoing basis.

Deep learning techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown some promise due to their capacity to learn complex patterns and handle streaming data. However, the effectiveness of such models heavily relies on the availability of sufficient labelled training data, which is a significant challenge (Diro, et al., 2024). A potential approach to overcome this challenge is to leverage techniques like transfer learning, which allows models trained on one dataset to be adapted for another, reducing the reliance on large-labelled datasets. Exploring unsupervised or semi-supervised learning methods could also help mitigate the dependence on labelled data. Incorporating domain knowledge and expert insights can further improve the accuracy and interpretability of anomaly detection models in building-to-cloud streaming data environments (Diro, et al., 2024).

This paper presents a novel approach that integrates domain expertise with a probabilistic framework based on the Dirichlet distribution and Latent Dirichlet Allocation (LDA), designed to efficiently monitor and detect network traffic anomalies. A case study of a mixed-use academic and residential building is presented to demonstrate the value of this work.

Related Work

Within the context of network traffic monitoring, anomaly detection identifies data points or patterns that deviate significantly from expected behavior. This process plays a vital role in cybersecurity, fraud detection, system monitoring, or predictive maintenance (Diro & Chilamkurti, 2018). Types of anomalies include point, contextual, collective, sequential, spatial, and collective contextual anomalies, categorized by their characteristics and data context (Diro, et al., 2021).

Point anomalies involve individual data points that deviate from normal behavior and are detected using statistical methods or machine learning, supporting fault identification (Cauteruccio, et al., 2021). Contextual anomalies consider relationships within data by flagging deviations in specific contexts, such as unusual patterns in time series or network traffic (Nassif, et al., 2021). Collective anomalies detect groups of data points behaving anomalously together and are useful in cybersecurity or sensor networks (Ahmed & Pathan, 2020). Sequential anomalies analyze temporal data by identifying deviations in expected sequences. This is particularly important in sensor data analysis (Oh & Iyengar, 2019). Spatial anomalies focus on spatial datasets where algorithms identify outliers in geographical or image data, such as disease outbreak hotspots or irregular satellite imagery patterns (Song, et al., 2021). Collective contextual anomalies, which combine contextual and collective techniques, effectively identify anomalies in specific contexts or clusters of data and is particularly adapted for analyzing complex datasets (Dou, et al., 2019).

A variety of statistical-based anomaly detection methods have been developed in the literature, for example (Zhang, 2013). Commonly used distributions in these methods include Gaussian, Poisson, Gamma, and more recently, α-stable distributions, which better capture the heavy-tailed and bursty nature of real network traffic (Simmross-Wattenberg, et al., 2011).

Non-parametric methods, which do not assume a specific distribution, are more suitable for data streams, particularly in single or low-dimensional contexts, though they become ineffective for high-dimensional datasets.

Amini and Wah apply clustering-based methods to group similar data points and detect anomalies in data streams (Amini & Wah, 2013). Amini (2014) proposes a clustering method that uses transitory memory to improve accuracy. While effective for clustering, it is less suited for anomaly detection and struggles with undefined cluster numbers in data streams.

Machine learning techniques have also become highly effective tools for identifying anomalies in data streams. Unsupervised approaches, like clustering and autoencoders, are efficient in identifying patterns within unlabelled datasets (Garg, et al., 2021).

Despite their promise, the deep learning methods face challenges like computational demands, reliance on labeled data, and lack of interpretability (Diro, et al., 2024).

We address this gap by proposing an approach that integrates domain expertise with a probabilistic framework. This method provides interpretable results by focusing on proportion deviations. It requires significantly fewer parameters (20 times) and less data (75 times less) than deep models. It is particularly adapted for edge devices, as it focuses exclusively on monitoring proportions in data streams.

Methodology

Data is streamed in real time from thousands of sensors located in an academic-residential building feeding a cloud-based digital twin. These sensors operate based on the Change of Value (COV) principle where data is transmitted only when a change occurs. This variability in data streaming is closely tied to sensors activity, which is influenced by dynamic changes in their parent HVAC equipment. To better understand these dynamics, it is important to investigate their underlying causes.

Previous work (McArthur & El mokhtari, 2024), explored the modeling of data streaming using a parametric probabilistic method based on transmitting devices (e.g. field controllers). Some devices displayed clear daily or hourly behavior patterns, such as Poisson or Normal distributions, however, field controllers associated with heterogenous equipment lacked consistent patterns. Two primary factors emerged as key drivers of equipment activity:

Weather and seasonal changes: HVAC systems cyclically transition between heating and cooling modes, with shoulder seasons exhibiting modes overlap. These transitions significantly influence system activity.

Occupancy: Increased building occupancy places greater stress on HVAC systems as they work to maintain temperature and humidity setpoints. Occupants activity lead to heat exchange between indoor and outdoor environments, increased solar heat transfer, increased CO2 levels and random changes in setpoints, further driving HVAC activity.

The degree to which these factors impact systems varies. For instance, HVAC systems in individual rooms are more sensitive to occupancy changes, while energy recovery ventilators (ERVs) are more heavily influenced by external weather conditions. ERVs perform most efficiently when there are significant differences between indoor and outdoor temperatures and humidity levels.

The approach presented here focuses on streaming dynamics at the system level rather than the field controller level. We hypothesize that systems exhibit distinct responses to the two primary factors: occupancy and weather conditions. This distinction enables the identification of more meaningful clusters compared to those derived from field controllers.

Our primary hypothesis is that while systems respond differently to these factors, the total daily of records

streamed under normal conditions should exhibit stable proportions across systems, at least within the same season. A significant deviation in these proportions may indicate an anomaly. A comprehensive anomaly detection framework would consider both total and proportions to flag These methods provide customized approaches for various data types and help organizations detect risks, improve decision-making, and uncover meaningful insights (Diro, et al., 2021) (He, et al., 2022).

In statistical-based anomaly detection methods, Zhang (Zhang, 2013) explains that parametric methods rely on predefined distribution models and require a training dataset to construct statistical models. However, in data streams, training datasets may fail in accurate anomaly detection due to the inability to infer a reliable distribution model (McArthur & El mokhtari, 2024). In contrast, non-parametric methods, which do not assume a specific distribution, are more suitable for data streams, particularly in single or low-dimensional contexts, though they become ineffective for high-dimensional datasets.

Machine learning techniques have become highly effective tools for identifying anomalies in data streams. Supervised techniques, such as Random Forest and Support Vector Machines, rely on labeled data for accuracy. Deep models tailored for time-series data such as the Temporal Convolution Network (TCN) introduced by Wang et al. demonstrate many advantages over LSTM-based approaches. Unsupervised approaches, like clustering and autoencoders, are efficient in identifying patterns within unlabelled datasets (Garg, et al., 2021). The challenge with proportions lies in determining the threshold at which a deviation qualifies as anomalous. To address this, we conceptualize the streaming process as a multinomial distribution, representing the building's HVAC systems and their respective contributions to the total data volume.

Building Description

The building is divided into two zones: the first zone, spanning floors from 1 to 9, is dedicated to academic activities such as classrooms, labs, and student areas, while the second zone, covering floors 10 to 28, serves as a residential area.

The building contains around 14,000 sensor points that stream data daily using the COV method. Additionally, the streaming system is configured to query each sensor once every 24 hours to verify its operational status.

Each data point is characterized by two identifiers. The network identifier follows the structure **<network>:<device>/<field_bus>.<controller>.<input>**. This identifier specifies the controller to be queried or subscribed to for data streaming. In contrast, the logical identifier, structured as **<campus>.<building>.<system>.<object>**, provides a hierarchical and functional classification of points. Based on this schema, the count of records sent by each system can be determined by grouping data points by their system component.

The Dirichlet Distribution

Definition

The Dirichlet distribution is a multivariate continuous distribution that generalizes the Beta distribution. It is the most natural distribution for modeling compositional data and proportions. In Bayesian statistics, it is commonly used as a conjugate prior for the Multinomial distribution (Lin, 2016). The Dirichlet distribution is parameterized by a vector of positive real numbers $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ known as concentration parameters, defined over a K-dimensional simplex (equation (1)).

$$\Delta^{K-1} = \{(x_1, \dots, x_K) \mid x_i \geq 0, \sum_{i=1}^{K} x_i = 1\} \quad (1)$$

The probability density function of the Dirichlet distribution is:

$$f(x_1, \dots, x_K \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \quad (2)$$

where:

$x_i \geq 0, \sum_{i=1}^{K} x_i = 1$,

$B(\boldsymbol{\alpha})$ is the multivariate Beta function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \quad (3)$$

$\Gamma(\cdot)$ is the gamma function.

Role of Concentration Parameters α

Each $\alpha_i > 0$ determines the "weight" or expected value of the proportion x_i.

Larger α_i values indicate that the distribution is more peaked, with proportions clustering closer to their expected values.

The expected value for each component is defined by equation (4).

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j} \quad (4)$$

Detecting Anomalies in Proportions

To detect anomalies, we compare observed proportions $\mathbf{x} = (x_1, x_2, \dots, x_K)$ to the expected proportions derived from the Dirichlet distribution's parameters ($\boldsymbol{\alpha}$). These proportions represent the relative volumes of network traffic from different devices (for example the controllers serving the air handing or heating system equipment) in a building. The likelihood of the observed proportions under the Dirichlet model are defined by equation (5).

$$\mathcal{L}(\mathbf{x} \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \quad (5)$$

Low likelihood values indicate that the observed proportions are unlikely under the expected distribution. Predefined thresholds for likelihood values are used to determine what constitutes an anomaly. To handle extremely small values, log-likelihoods are typically

applied to transform these probabilities into manageable negative values.

$$\log \mathcal{L}(\mathbf{x} \mid \boldsymbol{\alpha}) = -\log\big(\mathrm{B}(\alpha)\big) + \sum_{i=1}^{K} C_i \qquad (6)$$

$$\mathrm{C_i} = (\alpha_i - 1)\log(\mathrm{x_i}) \qquad (7)$$

The contribution of each proportion x_i to the total log-likelihood in equation (6) is denoted C_i – equation (7) – and allows to identify the proportion that has the greatest deviation.

In our model, each component of **x** represents the proportion of daily global record counts contributed by each system, as streamed from the building to the cloud. The parameters of the Dirichlet distribution (**α**) are derived as averages based on historical data.

Latent Dirichlet Allocation

Definition and application to data streaming

LDA is a generative probabilistic model designed for topic modeling, which is the process of discovering abstract topics from a collection of documents. LDA is widely used in Natural Language Processing (NLP), information retrieval, and text mining to uncover hidden thematic structures in text corpora (Blei, et al., 2003).

LDA assumes that documents are mixtures of topics, and topics are distributions over words. In essence, each document exhibits a proportion of topics, and each topic contributes words to the document based on its distribution.

In this paper, we adapt the LDA framework to model the daily data streaming dynamics of HVAC systems. We assume that each HVAC system contributes to the total daily data streaming with a given proportion and that these proportions vary according to the building's state (see Figure 1).

Here, the LDA concept of "topics" corresponds to the building's "states", "documents" represent individual "days", and "words" represent "records" counting the number of data points generated by various HVAC systems. The proportions of topics in documents are analogous to the proportions of states contributing to daily data streaming.

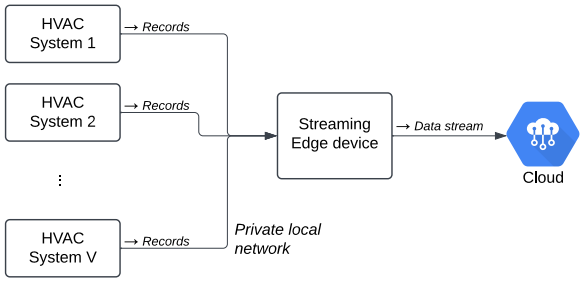

Figure 1: *Data streaming process*

In the LDA generative process for NLP depicted in Figure 2, we consider M documents of N words drawn from a vocabulary of size V, and K topics. Each document has a topic proportion distribution θ that specifies the contribution of each topic. The actual words are then drawn from the vocabulary using the topic-specific word distribution β. This process generates a document represented as a vector of words **r** (the letter **w** was used instead in (Blei, et al., 2003)).

Similarly, in the data streaming process modeled with LDA, we consider M days, V systems, and K states. Each day is characterized by a proportion of states, defined by θ, which governs the assignment of states to individual records. Based on the assigned state of each record, the β distribution determines the HVAC system responsible for generating the record. This process results in a vector of N records for each day, with each record generated by a specific system.

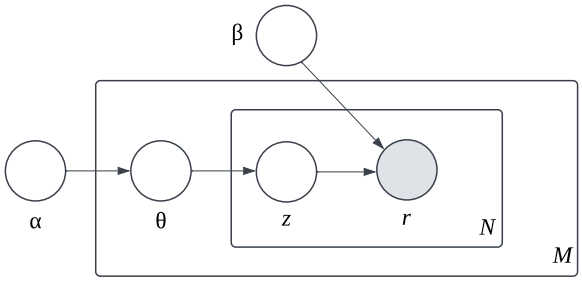

Figure 2: *Graphical model representation of LDA where boxes are plates representing replicates* (Blei, et al., 2003). *The output r represents the records originating from HVAC systems*

Mathematical formulation

To model the data streaming dynamics, we use the Latent Dirichlet Allocation (LDA) framework to infer the hidden distributions: the state proportions per day (θ) and the system distributions per state (β). These are derived from the observed distribution of records (r), where $r_{d,n}$ represents the n^{th} record of day d and its corresponding system.

The joint probability distribution for the observed and latent variables is expressed in equation (8), where:

$P(\beta_k \mid \eta) = Dirichlet(\beta_k; \eta)$: System distribution for state k.

$P(\theta_d \mid \alpha) = Dirichlet(\theta_d; \alpha)$: State proportion for day d.

$P(z_{d,n} \mid \theta_d) = Multinomial(z_{d,n}; \theta_d)$: State assignment for record n in day d.

$P(r_{d,n} \mid \beta_{z_{d,n}}) = Multinomial(r_{d,n}; \beta_{z_{d,n}})$: System assignment for record n based on state $z_{d,n}$

$$P(r,z,\theta,\beta \mid \alpha,\eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{M} P(\theta_d \mid \alpha) \prod_{n=1}^{N} P(z_{d,n} \mid \theta_d) P(r_{d,n} \mid \beta_{z_{d,n}}) \quad (8)$$

Experiment

The daily record count was collected from all building sensors over a period of 34 months (January 2022 to October 2024). Each record is mapped to one of the HVAC building systems listed in Table 1, based on the sensor's logical identifier. The average daily of records sent by each system is presented in **Error! Reference s ource not found.**, which highlights the predominance of five of the systems: BTU, electricity E-MTR, room sensors RM-A and RM-R, and hot water systems HWS-A and HWS-R.

Figure 4 illustrates the proportions of each system in the daily data and reveals a general trend of stability over the experimental period, with a few exceptions. Notably, the dominant systems account for nearly 90% of the total data volume. Figure 4 highlights several key patterns in the streaming. ERV-A shows a notable reduction in activity in January 2022, while HWS-A exhibits decreased activity during the summer of 2023. BTU activity displays distinct seasonal trends, with significant drops at the end of August 2022 and July 2023, which reflects reduced heating/cooling operations during these periods (Academic vacations). Conversely, higher BTU activity occurs during the winter months, particularly in January 2023 and January 2024.

Table 1: Building systems

System	Description
BTU	BTU meters (heating/cooling energy)
E-MTR	Electricity meters
RM-A/R	Rooms HVAC systems Academic / Residential
ERV-A/R	Energy Recovery Ventilators – Academic/Residential
HWS-A/R	Hot Water System Academic/Residential
AFS	Air Flow System
VAV	Variable Air Volume
CHWS	Chilled Water System
LAB	Labs HVAC systems
MISC	Miscellaneous
ACBCHWS	Active Beams Chilled Water System
Other	Other systems with low streaming rates

LDA application

The LDA algorithm was applied to the dataset, and the best number of states was determined to be three. Using only two states produced suboptimal results, while selecting more than three states led to duplicate states or states with very low probabilities.

The algorithm's output infers two latent distributions: θ and β. The θ distribution (Figure 5) reveals the mix of

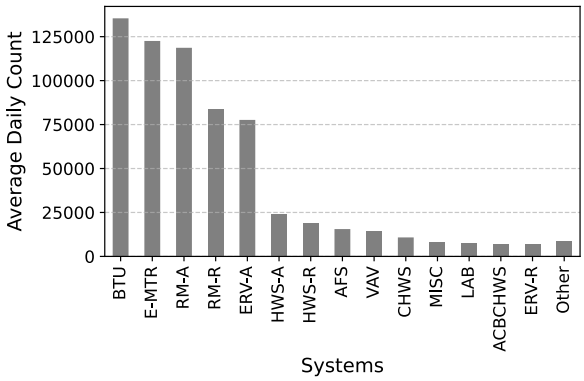

Figure 3: Average daily record count by system

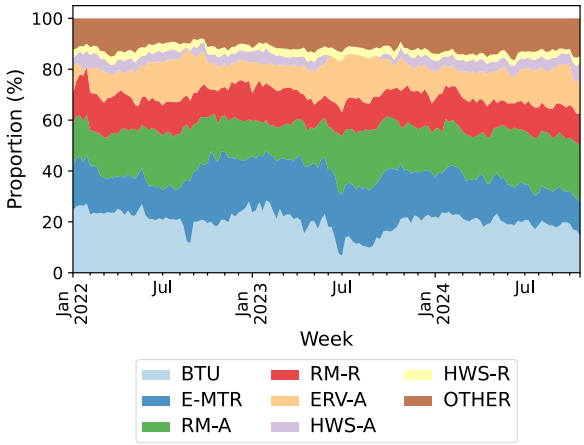

Figure 4: Weekly System record streaming proportions

states and the predominant state for each day, allowing for seasonal trends to be visualized, while the. β distribution (Figure 6) identifies the dominant systems in each state and enables experts to assign labels to the states through detailed analysis. Considering these two figures together, three distinct states are evident:

State 0 captures the **seasonal switchover** of the facility, peaking during the annual periods in early summer and late fall when neither heating nor cooling are available. Its β-distribution approximately reflects the long-term averages observed in Figure 3 that mirror the contributions of each system to the total record count.

State 1 corresponds to the **cooling-dominant unscheduled** mode. This aligns with both the non-academic term (May-August) when the residence is underutilized and very few classes are scheduled but research labs remain active and offices are occupied on an unscheduled basis. From a β-distribution perspective, it is

dominated by HVAC systems in the academic areas (RM-A and ERV-A), which remain occupied, the chilled water system, and electricity meters.

distribution. The results, shown in Figure 7, reveal a cluster of anomalies detected in January 2022, along with a few others sporadically appearing over 2023 and 2024. Cross-referencing the broader facility management data

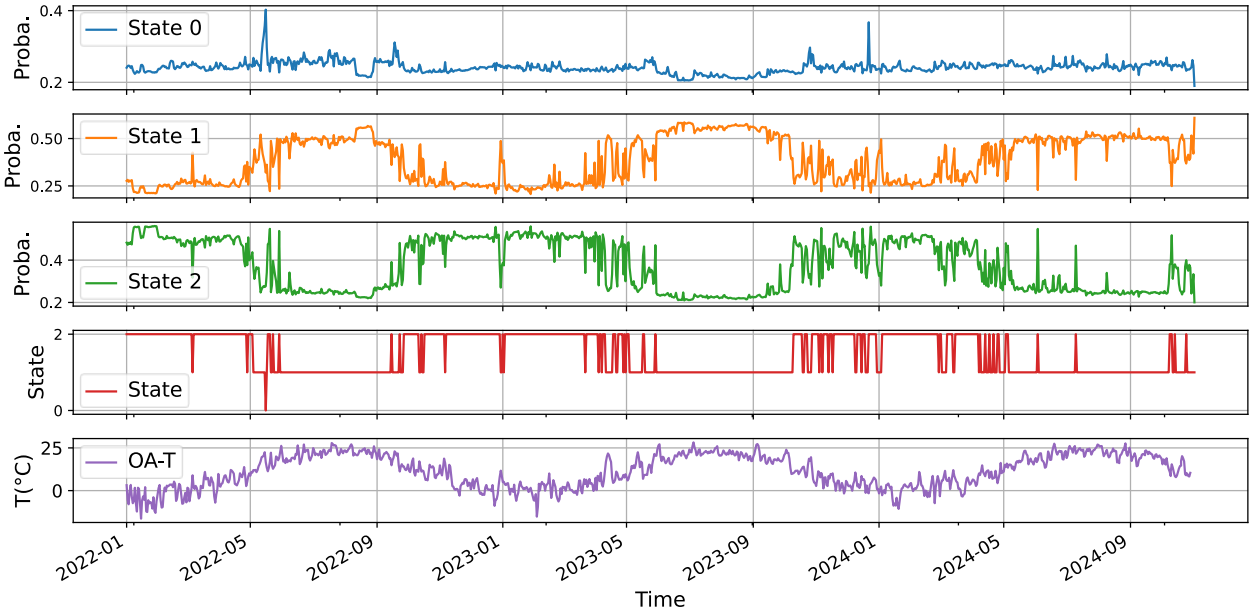

Figure 5: State probabilities (θ distribution in the LDA model) over time, and outdoor temperature (OA-T).

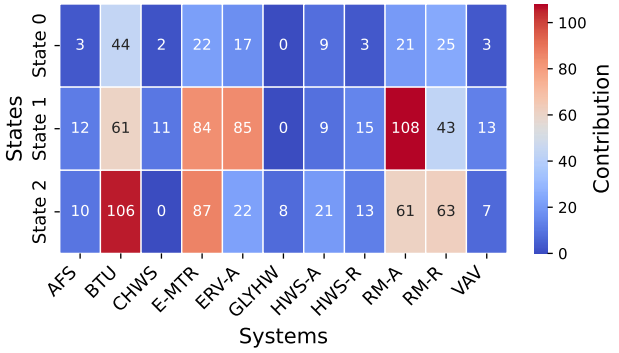

Figure 6: Heatmap of the top systems (columns) contributing to each LDA state (rows).

State 2 corresponds to the **heating-dominant scheduled mode**, aligning with the academic year (September-April) and heating season. Its β-distribution assigns higher weights to heating systems, such as GLYHW and HWS-A, while down-weighting the cooling system CHWS. This period has much higher occupancy than the other states, which explains the higher BTU and electrical meter signals and room behaviour in both academic and residence zones.

While the original algorithm was trained on three years of data, this was repeated with two months – January 2022 for heating and July 2023 for cooling – and similar patterns and distributions to those shown in Figures 5 and 6, respectively, were obtained. This shows the power of this approach with limited training data.

Anomaly detection with the Dirichlet distribution

A second experiment was conducted to detect anomalies in data streaming proportions using the Dirichlet

as well as the ground truth regarding the sensor reporting frequencies, ERV-A was found to transmit intermittently in January and May 2022, as shown in Table 2, with several days showing zero records despite other systems reporting normally. While the overall record count remained relatively stable in Figure 7, the change in proportions between systems allowed the Dirichlet distribution to highlight the anomaly.

Table 2: Reported record counts for January 2022

Date	BTU	…	ERV-A	RM-A
2022-01-06	180607	…	83014	108523
2022-01-07	170373	…	31670	113285
2022-01-08	167863	…	0	106240
2022-01-09	148479	…	0	102238
2022-01-10	172858	…	0	108659
2022-01-11	214589	…	0	112778
2022-01-12	164319	…	0	106513
2022-01-13	151279	…	0	102321
2022-01-14	141104	…	35387	107596
2022-01-15	209017	…	80764	114359
…	…	…	…	…
2022-05-20	180962		82	125565
2022-05-21	135016		110604	127436

We adopt a score-based anomaly detection framework, as commonly described in the literature (Chandola, et al., 2009), and following conventions where thresholds are often set empirically or through domain expertise (Ahmed, et al., 2016), we selected the second percentile

of log-likelihood values to conservatively flag the lowest-likelihood observations as anomalies in streaming proportions. We note that these proportion anomalies are not aligned with spikes or drops in the total daily record count, which is also shown in the figure. This indicates that spikes or drops in record count reflect changes in HVAC activity levels but do not necessarily affect the proportional contributions of individual systems to the overall data volume. By analyzing the contribution of each system to the log-likelihood, using Scipy's *dirichlet.logpdf* function, which applies the transformation in Eq. 5-7 to calculate log-likelihood distribution probabilities, the system with the highest deviation can be identified, as illustrated in Figure 8. This figure highlights that the Energy Recovery Ventilator in the academic area (ERV-A) is the source of these anomalies and an examination of the data streaming records (which form the 'ground truth' for validation) reveal that ERV-A had fewer change-of-value records reported, attributable to maintenance activities. Again, this was repeated for reduced datasets and results were replicated with as little as one week, so the two months of data required in the first part are more than sufficient for data traffic anomaly (e.g. intrusion or loss of signal) detection.

Conclusion

In this paper, we presented a framework for monitoring data streaming and detecting anomalies in HVAC data streams by analyzing data streaming proportions. The core premise is that these proportions remain generally stable over time, and clusters of proportions can reveal the operational states of the HVAC systems. Using Latent Dirichlet Allocation, we identified three distinct states corresponding to cooling, heating, and overall streaming stability. The model enables the prediction of a day's state based on observed proportions, thus offering insights into system behavior. Anomalies were detected using the Dirichlet distribution, which identifies extreme deviations in proportions as potential issues.

This framework provides facility managers with tools to monitor the HVAC building state, compare it to expected operational mode, and detect deviations in proportions. Additionally, it highlights systems with extreme deviations for further investigation and diagnosis. However, the framework has some However, the framework has some limitations. First, it requires sufficient data to accurately fit the model, ideally at least one year, to capture seasonal HVAC patterns. Second, the number of states may vary between buildings, and expert knowledge is essential to interpret the states and extract

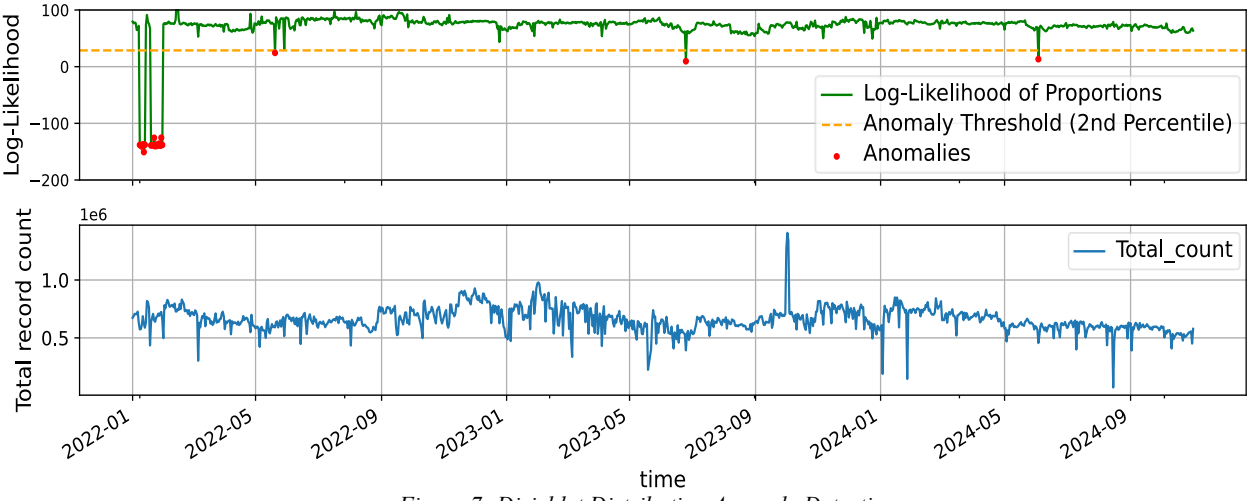

Figure 7: Dirichlet Distribution Anomaly Detection

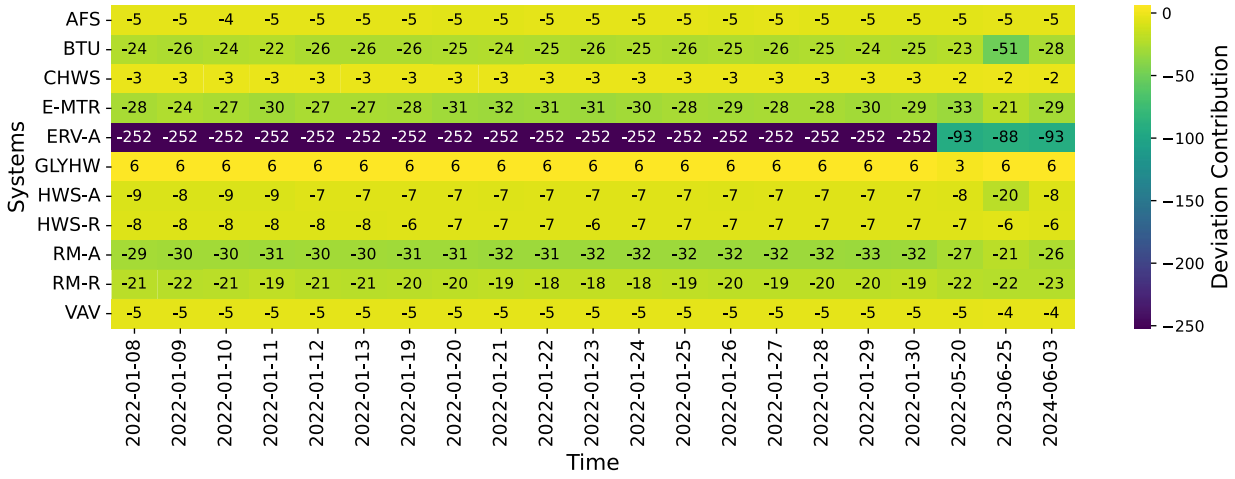

Figure 8: Contribution of the primary system deviations to the 21 detected anomalies

actionable insights. Finally, the framework focuses only on proportions, which identifies deviations in specific systems but does not detect anomalies in the total of streamed data. Integrating this approach with other anomaly detection methods could expand its capabilities to address broader streaming issues.

Acknowledgments

This research was funded by the Natural Science and Engineering Research Council of Canada (RGPIN-2018-04105).

References

Ahmed, M., Mahmood, A. N. & Hu, J., 2016. A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, pp. 19-31.

Ahmed, M. & Pathan, A.-S. K., 2020. Deep learning for collective anomaly detection. International Journal of Computational Science and Engineering, 21(1), pp. 137-145.

Amini, A., 2014. An adaptive density-based method for clustering evolving data streams. University of Malaya (Malaysia).

Amini, A. & Wah, T. Y., 2013. Requirements for clustering evolving data stream. In 2nd International Conference on Soft Computing and its Applications 5ICSCA, 2013 Sept. pp. 25-26.

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp. 993-1022.

Cauteruccio, F. et al., 2021. A framework for anomaly detection and classification in Multiple IoT scenarios. Future Generation Computer Systems, 114, pp. 322-335.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), pp. 1-58.

Chen, D., Sun, Q. Z. & Qiao, Y., 2025. Defending against cyber-attacks in building HVAC systems through energy performance evaluation using a physics-informed dynamic Bayesian network (PIDBN). Energy, p. 135369.

Chen, J. et al., 2022. A review of computing-based automated fault detection and diagnosis of heating, ventilation and air conditioning systems. Renewable and Sustainable Energy Reviews, 161, p. 112395.

Diro, A. A. & Chilamkurti, N., 2018. Distributed attack detection scheme using deep learning approach for Internet of Things. Future Generation Computer Systems, 82, pp. 761-768.

Diro, A., Chilamkurti, N., Nguyen, V.-D. & Heyne, W., 2021. A comprehensive study of anomaly detection schemes in IoT networks using machine learning algorithms. Sensors, 21(24), p. 8320.

Diro, A. et al., 2024. Anomaly detection for space information networks: A survey of challenges, techniques, and future directions. Computers & Security, 139, p. 103705.

Dou, S., Yang, K. & Poor, H. V., 2019. Pc²A : predicting collective contextual anomalies via LSTM with deep generative model. IEEE Internet of Things Journal, 6(6), pp. 9645-9655.

Garg, A. et al., 2021. An evaluation of anomaly detection and diagnosis in multivariate time series. IEEE Transactions on Neural Networks and Learning Systems, 33(6), pp. 2508-2517.

He, J. et al., 2022. Application of sparse representation method based on K-SVD-ADMM in anomaly detection of satellite telemetry. 2022 Global Reliability and Prognostics and Health Management 5PHM, Yantai, pp. 1-7.

Lin, J., 2016. On the dirichlet distribution. Department of Mathematics and Statistics, Queens University, 40.

Matetić, I., Štajduhar, I., Wolf, I. & Ljubic, S., 2022. A review of data-driven approaches and techniques for fault detection and diagnosis in HVAC systems. Sensors, 23(1), pp. 1-37.

McArthur, J. & El mokhtari, K., 2024. An Online Data Streaming Quality Detection Algorithm to Support Digital Twins. Proceedings of the 41st CIB W78 Conference, Marrakesh.

Nassif, A. B., Talib, M. A., Nasir, Q. & Dakalbab, F. M., 2021. Machine learning for anomaly detection: A systematic review. IEEE Access, 9, pp. 78658-78700.

Oh, M.-h. & Iyengar, G., 2019. Sequential anomaly detection using inverse reinforcement learning. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & data mining, pp. 1480-1490.

Safaei, A. A., 2017. Real-time processing of streaming big data. Real-Time Systems, 53, pp. 1-44.

Sándor, B. & Rajnai, Z., 2023. Smart Building IoT Cybersecurity: A Review of Threats and Mitigation Technique. In 2023 IEEE 21st Jubilee International Symposium on Intelligent Systems and Informatics (SISY), pp. 000321-000326.

Simmross-Wattenberg, F. et al., 2011. Anomaly detection in network traffic based on statistical inference and alpha-stable modeling. IEEE Transactions on Dependable and Secure Computing, 8(4), pp. 494-509.

Song, X. et al., 2021. Spectral-spatial anomaly detection of hyperspectral data based on improved isolation forest. IEEE Transactions on Geoscience and Remote Sensing, 60, pp. 1-16.

Zhang, J., 2013. Advancements of outlier detection/ A survey. EAI Endorsed Transactions on Scalable Information Systems, 1(1).