An Online Data Streaming Quality Detection Algorithm to Support Digital Twins

J.J. McArthur, jennifer.mcarthur@torontomu.ca
Faculty of Engineering and Architectural Science, Toronto Metropolitan University, Canada

Karim El Mokhtari, elmkarim@torontomu.ca
Faculty of Engineering and Architectural Science, Toronto Metropolitan University, Canada

Abstract
The continuous streaming of data from a sensor network is essential for a digital twin to mirror its physical counterpart in real-time. Missing or interrupted data limits the digital twin functionalities, but can be difficult to detect when data is communicated at a change of value. Such disruptions in data streaming can compromise facility management data uses and lead to overlooked critical alerts. By understanding the unique streaming patterns of each controller, these models are less susceptible to errors, reducing the incidence of false alarms, ensuring more accurate streaming, and notifying users immediately about any significant deviations from the expected data counts. This paper presents the development of such an approach implemented on a Building Automation System for a large multi-use building, which consists of over 14,000 points for the HVAC systems that report whenever a change of value above a defined threshold is observed. Data is collected from thousands of sensors and gathered by controllers wired to the building's private network. Before transmission to cloud services, streaming software structures the records with appropriate formatting. This data is then archived in a time-series database for subsequent analysis. The streaming quality model functions in a two-tiered approach. Initially, an appropriate model is determined for each controller. This is accomplished by fitting the count distribution with known statistical distributions, such as the Poisson or Normal distribution. Algorithms like Expectation-Maximization determine the parameters of these distributions. In the subsequent phase, the calibrated models operate in the cloud, assessing on an hourly basis whether the count aligns with the expected range. If deviations exceed a set threshold, an alert is triggered. This paper presents the development of this approach with a formal use case definition, process map, and information exchange to support its replication using OpenBIM approaches.

Keywords: Data Streaming, Streaming quality, UCM, PPM, information exchange requirements.

1 Introduction
The concepts of 'Smart Buildings' and 'Digital Twins' are increasingly gaining adoption, particularly insofar as they are able to support energy management, decarbonization, and improved facility management practices. In this paper, a 'Smart Building' is defined as a building where sensor data from across multiple systems can be collected and integrated into a common data environment, while 'Digital Twin' refers to a virtual version of an asset capable of acquiring data in near-real-time, storing it, using the data to analyze or optimize the real-world asset performance, and send instructions back to the asset (or a human intermediary) to implement changes. Together, these allow building data to be collected, streamed, and analyzed within the virtual space, where controls optimization and fault detection can occur. Several studies have explored the potential, implementation, and challenges of such "Smart Building Digital Twins" (SBDT), see (Deng, et al., 2021; Ghansah, 2024).

Despite the significant potential of SBDTs, data acquisition can pose a significant challenge. First, such data is complex and heterogeneous, often acquired from a diversity of siloed systems and – in the case of older buildings - drawing from legacy, proprietary systems. Second, the data streams must be robust as missing data can lead to significant events being missed as well as limitations on the accuracy and functionality of the data analytics. While in some cases, data is streamed at predefined intervals, in many cases, a change-of-value (CoV) approach – where data is streamed whenever a reading deviates from the previous one by a threshold amount – can be more helpful to capture system dynamics. In such cases, it can be extremely difficult to determine when data streams are interrupted, resulting in missing data. This paper addresses this latter challenge, presenting a stochastic approach to develop an online data streaming quality algorithm capable of learning normal data streaming patterns from building automation system (BAS) devices so that any interruptions can be quickly identified and resolved.

This paper presents the data streaming quality algorithm development and a case study implementing it in a mixed-use (academic lab/classroom and residence) building. To maximize the value of this contribution, its implementation is contextualized using process maps and information exchanges consistent with the BuildingSMART Information Delivery Manual (IDM).

2 Literature review

A significant body of research exists on data streaming and its quality monitoring. This section begins with an overview of the most common building automation protocols and how these are mapped to OpenBIM concepts to support integration with BIM-based SBDTs. Second, the state-of-the-art regarding data quality streaming research is presented to contextualize this paper's contribution.

2.1 Building Automation, Data Streaming, and OpenBIM Integration Protocols

Data integration protocols arose alongside direct digital controllers in the 1980s to enable data exchanges between these controllers, sensors, and controlled equipment. Within the commercial buildings domain, three significant data streaming protocols have been adopted: LONWorks, MODBUS, and BACnet (ASHRAE, 2001). Of these, BACnet is the most widely used for HVAC system controls and integration, and is governed by ISO 16484-5. It is a client-server protocol allowing bidirectional communication between any devices within a Building Automation System network. MODBUS is an older serial protocol using a master-slave system and while highly efficient for simple data communication between PLCs, it is unable to handle complex data and is rarely found in new commercial buildings. Of the three protocols, LONWorks is the most proprietary protocol, though it uses an open standard. It is primarily used for specialized applications, rather than more broadly across buildings, though with its recent integration with Amazon's Alexa, Google Home, and other home integration systems it is beginning to see adoption in Smart Homes (Chinchero, et al., 2020). Other Smart Home integration protocols are BUSing and KNX, which may be interconnected for broader interoperability. Each of these protocols communicates with either MSTP or IP and can be streamed using either proprietary edge IoT devices or using open protocols, including I4.0 and AWS Greengrass.

The ubiquity of 'smart' devices within buildings has increased both the availability and usability of this data, however to map these into OpenBIM, the *IfcBuildingControlsDomain* is critical, notably the *IfcPerformanceHistory* property sets. At a high level, the *IfcBuildingControlsDomain* schema is an extension of the *IfcSharedBldgServicesElements* schema, supporting various controls types such as alarms, instrumentation (sensors, flow instruments), control (controllers, valves and dampers, actuators) and building automation as a concept (BuildingSMART, 2024a). Each of the devices may be assigned an *IfcPerformanceHistory*, which may be used with or without design data, and which uses the *IfcRelAssociatesClassification* to map the network addresses of the relevant device or data point (BuildingSMART, 2024b). The *IfcBuildingControlsDomain* is building automation protocol agnostic but may be mapped to these systems. Key concept templates that can be assigned to *IfcPerformanceHistory* are

classification, aggregation, and control to support the full range of functionality to stream building automation data into OpenBIM.

2.2 Monitoring Data Streaming Quality

Several studies have noted the criticality of reliable sensor data to support data-driven predictive models, for example (Lillstrang, et al., 2022; Bamgboye, et al., 2019; Ghansah, 2024). In a study of two case studies, Lillstrang et al (2022) found significant periodicity, low heterogeneity of patterns, and significant missing data on the order of 11.3-33.7%. They also found significant differences in sampling rates by sensor type and location were also found and they cautioned that overlooking the periodicity and heterogeneity of data could lead to false impressions of accuracy when only homogeneous conditions are predicted (Lillstrang, et al., 2022). This demonstrates the importance of considering sampling rates and their periodicity to avoid overestimating the accuracy of the model. The lack of consistent data collection and the complexity and uncertainty of data remain two of the most significant challenges in the creation of Digital Twins (Ghansah, 2024).

To overcome missing sensor data, many scholars impute values, however, this is insufficient for large bursts of missing data (Bolchini, et al., 2017). Other scholars have sought to analyze the consistency of streamed data. One study (Bamgboye, et al., 2019) applied semantic modeling and reasoning using a sliding window technique to analyze data stream temporal characteristics and explored the effects of different data serializations finding that RDF/XML outperformed NTriple, Turtles, and N3 from a latency perspective. While this study was able to validate data consistency, it was not designed to detect missing data; this latter element remains a significant gap in the literature and one addressed in this paper.

3 Methodology

A mixed-use building consisting of a 16,300m2 (175,000sf) building is used as a case study for this research and consists of an academic podium (lab & academic offices) with 19-storey residence tower housing 332 student rooms in 2- and 4-bedroom apartments. Developed by the university to be a 'living lab', the building is controlled by a typical BAS and enhanced with numerous supplementary measurement points for both hydronic (hot & chilled water) and air flows throughout the academic podium and on selected residence floors. A total of approximately 14,000 data points are thus streamed, each pushed to the IoT device using a CoV strategy with an additional point reported at midnight.

The aim of this work is to model and analyze the streaming rate of data from this case study building, which is. In this approach, updates are sent only when the value of the data point changes beyond a predefined threshold, to minimize the additional network traffic and storage requirements by avoiding the continuous streaming of redundant or unchanged data. This results in a highly variable volume of available data over the course of a day. This paper presents the development of a predictive model to capture the data streaming patterns. When the actual data rate deviates significantly from the predicted model, an alert would be triggered to indicate the anomaly. The overall methodology used is as follows:

1. Gather data from AWS Timestream: Data will be gathered every hour for all devices (NAE-01), noting the timestamped counts. This data will be stored offline in CSV files for analysis.
2. Estimate the value of the parameter λ for a Poisson distribution or (μ, σ) for a Normal distribution, by counting the records received during a time interval Δt = 1 hour. This interval can be adjusted for a quicker response from the alert system.
3. With the estimated λ and (μ, σ), calculate the likelihood of a new hourly recorded count.
4. Build an anomaly detection algorithm that identifies when the likelihood of the count falls below a certain threshold. If we have various values of the parameter λ that depend on the time window (e.g., summer, weekend), then different models should be used, each tailored to its specific case with a different λ.

It is important to note that to ensure that at least one value is recorded per day, there is a daily forced reporting of all data points at midnight.

This research is presented following the Building Smart International UCM framework to permit its reproducibility using OpenBIM approaches. However, it should be noted that the initial research was developed to run using a gltfs extracted from an FM-enabled BIM and connected to a data lake with both relational and time-series databases, rather than mapping the data into a BIM using the IFC protocols discussed previously.

3.1 Data Acquisition and Information Exchange

The building contains many network control (NCE) and automation (NAE) engines, which manage the network traffic from field controllers, which in turn communicate with sensors using the BACnet Protocol. These devices are typically denoted as NAE-XX or NCE-XX depending on their type, where XX is a sequential number. The sensors are attached to a range of assets that are connected to the devices. For instance, device NAE-01 collects data from all HVAC assets on the first and second floors, like FCU-1-1 and FCU-2-6. When a sensor within an asset captures a measurement, the device sends this data to the streaming software upon receipt.

Following the Building Smart International UCM framework, the functional and technical requirements are summarized briefly herein, starting with an identification of the functional parts, using the Information Delivery Manual (IDM) process map, exchange requirement, and functional part definitions as follows.

The overall process begins with the reporting of a change of value from a building sensor, which is communicated to the streaming device, which writes it to the time-series database in the database and adding it to the performance history (*IfcPerformanceHistory)* of the associated element as defined in *IfcRelAssociatesClassification*. On a scheduled basis, the predictive model is run, which checks for any errors or gaps in the data received since the last check; if an error is detected, this is reported as an alarm both flagged in the Digital Twin and – if significant – reported to the facility management team who can investigate the reason for this loss. If there is no error detected in the batch, the process is terminated until the next scheduled time. This process is summarized in Fig 1.

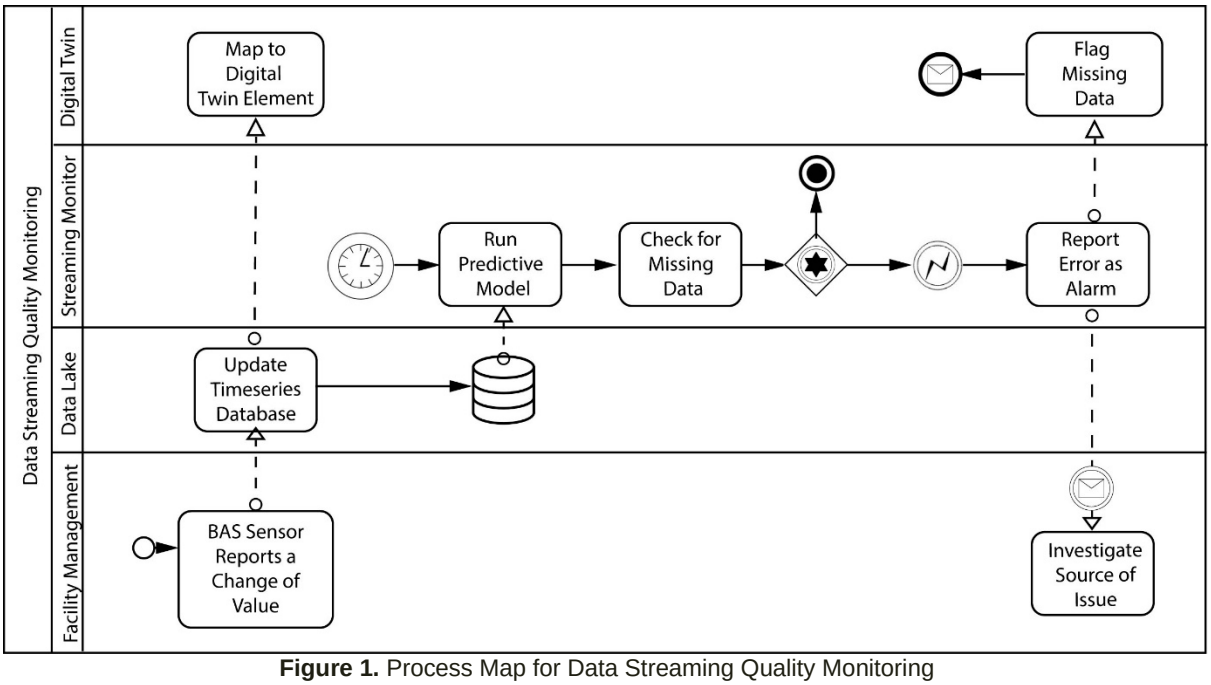

Figure 1. Process Map for Data Streaming Quality Monitoring

To support this process, five data exchanges are required. First, the BAS sensor must report a change of value; this is sent by the edge device to the data lake using the chosen protocol (e.g. BACnet accessed via Greengrass). Within the data lake, this is then appended to the time-series database. From the time-series database, a data exchange is required to add this datapoint to

the *IfcPerformanceHistory* of the relevant element, using the *IfcRelAssociatesClassification* to match the BIM element with the network address of the sensor. Additional data exchanges are much simpler, consisting of the database query to run the predictive model and the reporting of alarms through both the Digital Twin and Facility Management system interfaces.

While the majority of required parts and concepts already exist, this paper focuses on the creation and use of the streaming data predictive model to detect anomalous patterns and missing data.

3.2 Streaming Data Model

The exact count of records that we receive from a device isn't precisely known and is treated as a random variable. Our analysis strategy revolves around first determining the count of records received from each device in a specific time interval, denoted as Δt (1 hour in this study). We then calculate the likelihood of this count given our predictive model. If the likelihood falls below a predefined threshold, an alert is triggered.

In theory, we can model the generation of a measurement as a random variable following a Bernoulli distribution, and thus the count of measurements in a time window Δt follows a Poisson distribution with parameter λ (Eq. 1), where λ is the average count of records within the time interval Δt.

$$P(X = k \mid \lambda) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \quad (1)$$

We assume that the parameter λ varies across devices and can even fluctuate within different time windows for the same device. For example, during the weekends, the change dynamics could be slower, possibly leading to fewer CoV events and consequently, lesser data transmission. We verify if this hypothesis holds true. For higher values of λ, the Poisson distribution can be approximated with a Normal distribution with both mean and variance equal to λ: $\mathcal{N}(\lambda, \sqrt{\lambda})$.

A second hypothesis proposes fitting the streaming data distribution to a normal distribution, $\mathcal{N}(\mu, \sigma^2)$ (Eq. 2).

$$\mathcal{N}(X = x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

4 Results

Prior to modeling the data distributions, preliminary data analysis was conducted to understand the frequency of CoV reporting across the building. While each network device (NAE or NCE) had a similar capacity for reporting data points, these varied significantly by device based on the connected equipment. Figure 2 presents the estimated mean of records transmitted hourly for each device (excluding midnight reset). A considerable disparity between devices is observed and explained by the number of controllers connected to each device and the nature of the controlled HVAC equipment or sensor points. NAE-01 controls two floors (Levels 1 and 2), which explains the high number of recorded hourly counts. Meanwhile, NCE-09 and NCE-28, which serve the main mechanical rooms, also show relatively high activity.

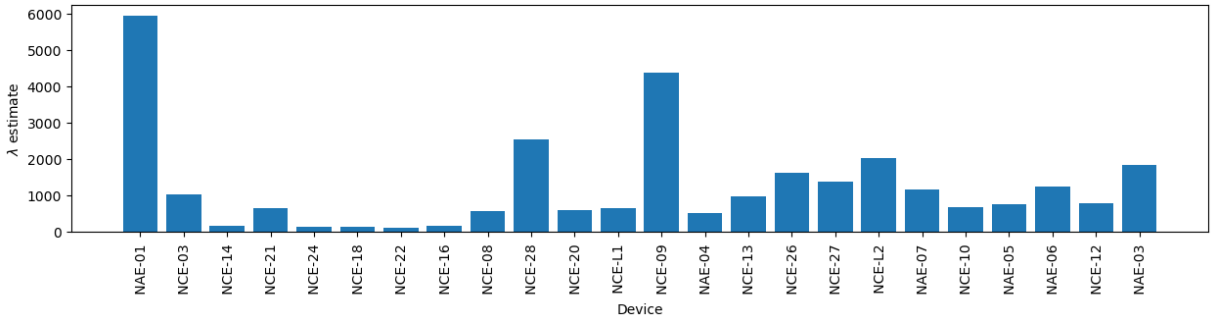

Figure 2. Hourly streaming mean record count per device (excluding the midnight reset).

Visualizing these counts on an hourly basis, provides significant additional information, as shown in Figure 3. The mean value at midnight differs from the mean values observed at other hours. This disparity is attributed to the fact that at midnight, all sensors in a device are queried simultaneously, resulting in a higher rate of data. Modes close to zero are omitted ; they typically occur due to streaming interruptions. The histograms show that most of the devices exhibit a single mode with sometimes a relatively skewed distribution and a mean value remaining consistent throughout the entire period. However, there are exceptions where two modes are present such as for NAE-01, NCE-20, NCE-L1, NCE-10, NAE-05, NAE-03.

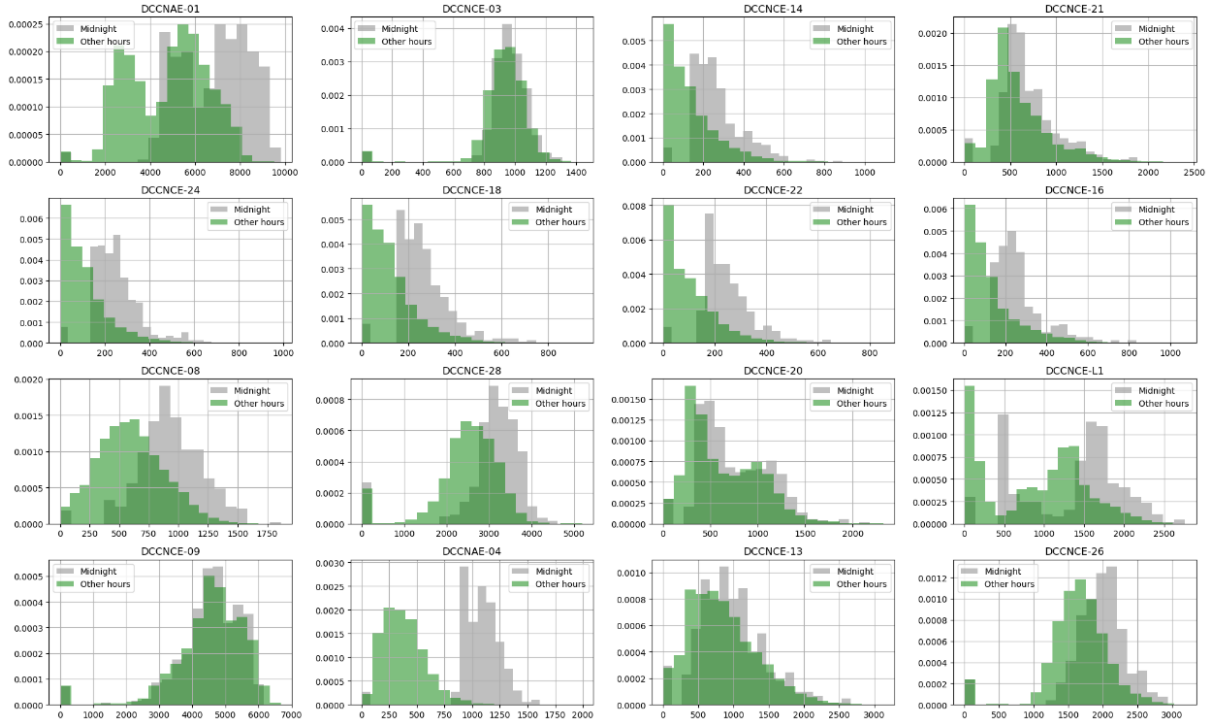

Figure 3. Streaming histogram in normal hours and midnight (streaming reset) for selected devices

4.1 Distribution fitting

The Expectation-Maximization (E-M) algorithm was used to fit distributions, along with the Kolmogorov–Smirnov normality test for normal distributions. Figure 4 shows examples of real distributions, excluding midnight data, compared to Poisson and a distribution with a standard deviation $\sigma = 10\sqrt{\lambda}$ where λ is the estimated mean. This value of σ is observed to be a good fit for many devices. But in the analysis, the standard deviation σ associated with each device is estimated using the E-M algorithm. The fitting result is shown in Table 1.

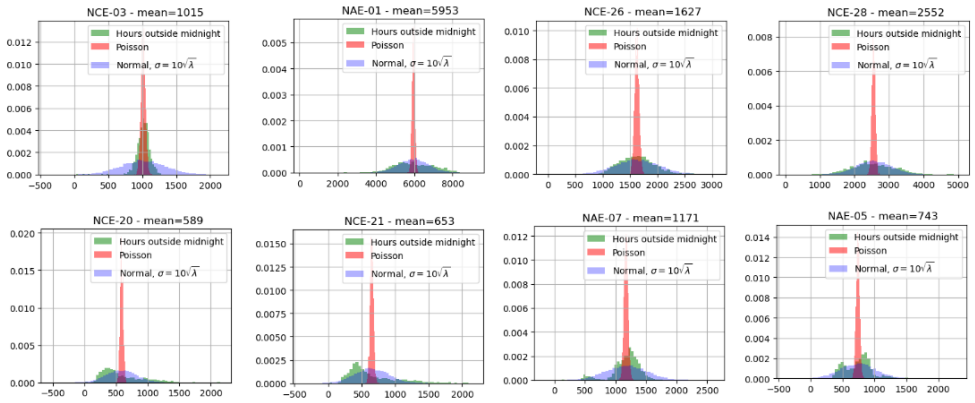

Figure 4. Distribution Fitting for selected devices fitting to a Poisson or Normal distribution (Row 1) or skewed or bi-modal distribution (Row 2). References for Poisson and Normal distributions are also shown.

Table 1. Distribution Summary – First round

Poisson	Normal	Skewed	Bi-modal
NCE-03	NAE-01	NAE-03	NAE-05
	NAE-04	NCE-14	NAE-07
	NAE-06	NCE-16	NCE-08
	NCE-09	NCE-18	NCE-10
	NCE-26	NCE-20	NCE-12
	NCE-28	NCE-21	NCE-13
	NCE-L2	NCE-22	NCE-27
		NCE-24	NCE-L1

4.1.1 Skewed distributions

To address the skewness problem observed for many devices, the logarithms of hourly record counts are fitted to a normal distribution, thereby creating a log-normal distribution. The new classification results after fitting are shown in Figure 5 and Table 2.

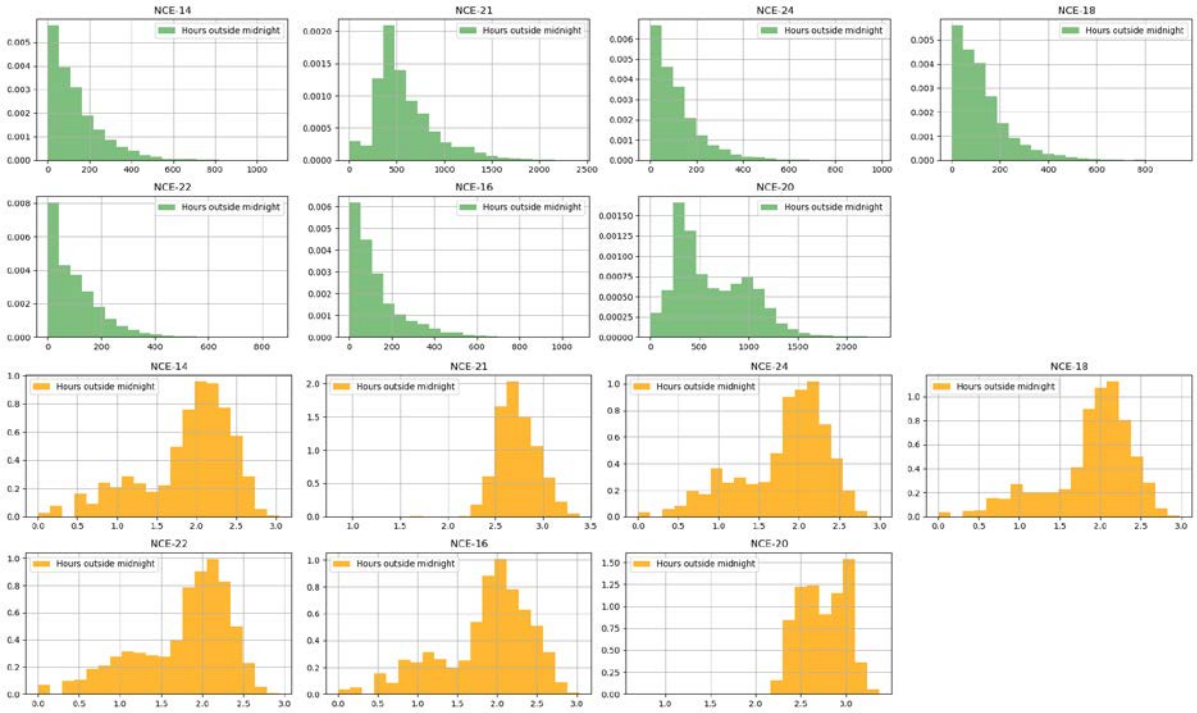

Figure 5. Applying a Logarithmic Transformation to Skewed Distributions. The top two rows represent the original distributions, and the bottom two rows represent the log-transformed distributions.

Table 2. Distribution Summary – Final results

Poisson	Normal	Skewed	Bi-modal	Log Normal
NCE-03	NAE-01	NCE-14	NCE-08	NCE-20
	NAE-04	NAE-03	NCE-L1	NCE-21
	NAE-06	NCE-24	NCE-13	
	NCE-09	NCE-18	NCE-27	
	NCE-26	NCE-22	NAE-07	
	NCE-28	NCE-16	NCE-10	
	NCE-L2		NAE-05	
			NCE-12	

4.1.2 Remaining skewed and bi-modal distributions

A bimodal or skewed distribution indicates the presence of different data streaming patterns that cannot be effectively modeled by treating all hours equally. There are several approaches to address this. One method is to create clusters of days based on criteria such as working days, holidays, weekends, or seasons, like the cooling and heating seasons. Alternatively, each hour of the day could be modeled as an independent distribution, resulting in 24 distinct distributions for the same device. This process was replicated for the remaining devices. For each hour, a test of normality is conducted to determine which hours conform to a normal distribution. Subsequently, the Expectation-Maximization (EM) algorithm is employed to identify the optimal fit for the normal distribution. Finally, logarithmic transformations are applied to address skewness in hourly distributions. The final classification is shown on Table 3.

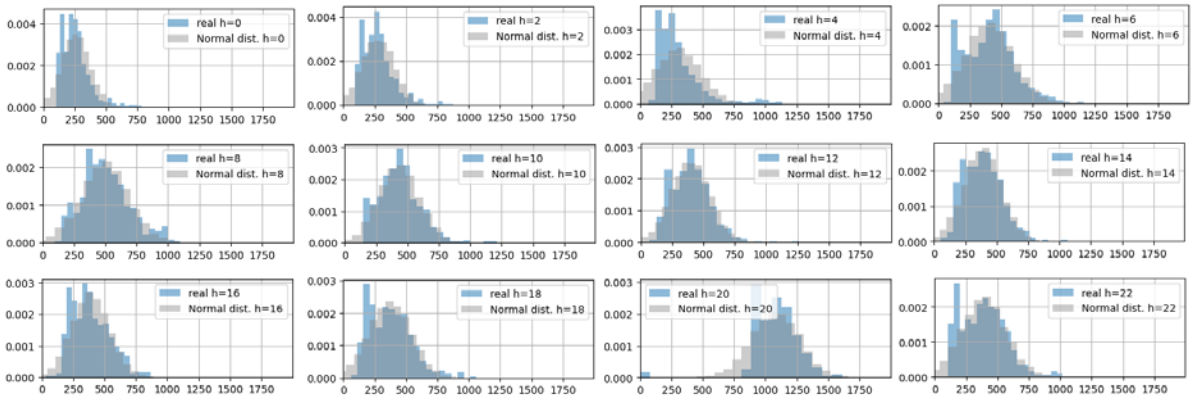

Figure 6. Hourly Record Count Distribution for NCE-08: Both Real and Predicted Distributions After Test of Normality and E-M Algorithm Application for Each Hour of the Day – Only even hours are shown.

Table 3. Distribution Summary – Third round

Poisson	Normal	Log Normal	Hourly Normal	Hourly Log Normal	Bi-modal
NCE-03	NAE-01	NCE-20	NAE-03	NCE-12	NAE-05
	NAE-04	NCE-21	NAE-07	NCE-14	NCE-10
	NAE-06		NCE-08	NCE-16	NCE-13
	NCE-09		NCE-27	NCE-18	NCE-L1
	NCE-26			NCE-22	
	NCE-28			NCE-24	
	NCE-L2				

The remaining bi-modal distributions require further modeling that may consider factors such as occupancy or seasonal changes.

4.2 Anomaly detection process

For devices that follow either a Poisson or a Normal distribution (with or without the logarithmic transformation), an anomaly detection algorithm can be implemented using the negative log likelihood of the observed hourly record count. A predefined threshold is used to generate alerts when the negative log likelihood exceeds that threshold. In the case of a Normal distribution, the threshold is set to $Th_N = -L(\mu - 3\sigma)$, and for the Poisson distribution, it is $Th_P = -L(\lambda - 3\sqrt{\lambda})$. Only lower-bound thresholds are considered, as the focus is on detecting drops in record counts.

The negative log likelihood for Poisson and Normal distributions are defined respectively as:

$$-L(x|\lambda) = \lambda - x\log(\lambda) + \log(x!) \quad (3)$$

$$-L(x|\mu,\sigma^2) = \frac{1}{2}\log(2\pi) + \log(\sigma) + \frac{(x-\mu)^2}{2\sigma^2} \quad (4)$$

Figure 7 and Figure 8 illustrate the use of the calculated threshold to detect streaming issues in a real setting for both a Poisson distribution (e.g. device NCE-03) and a Normal distribution (e.g. device NAE-01). Alerts are generated when the record counts fall below the thresholds set previously for each distribution. The method allows for the detection of streaming interruptions and major deviations from the distribution mean.

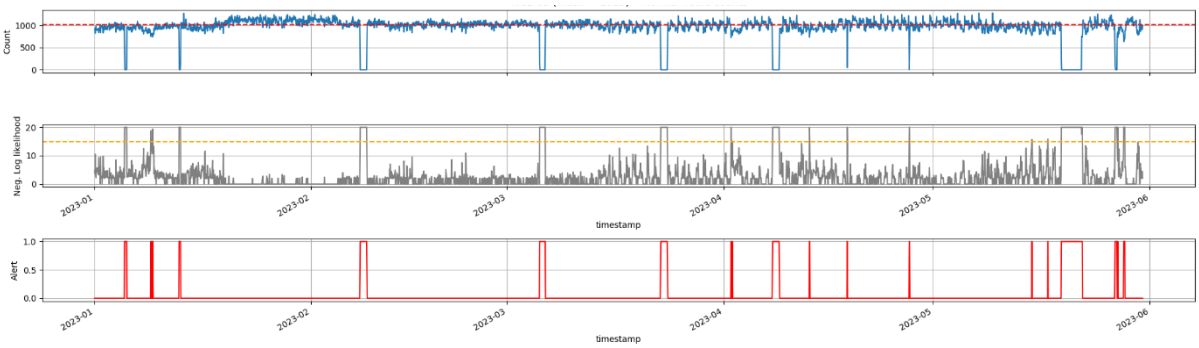

Figure 7. Anomaly detection with a Poisson-distributed Streaming Data (NCE-03). Top: Count, Middle: Negative log-likelihood; Bottom: Alert Signal.

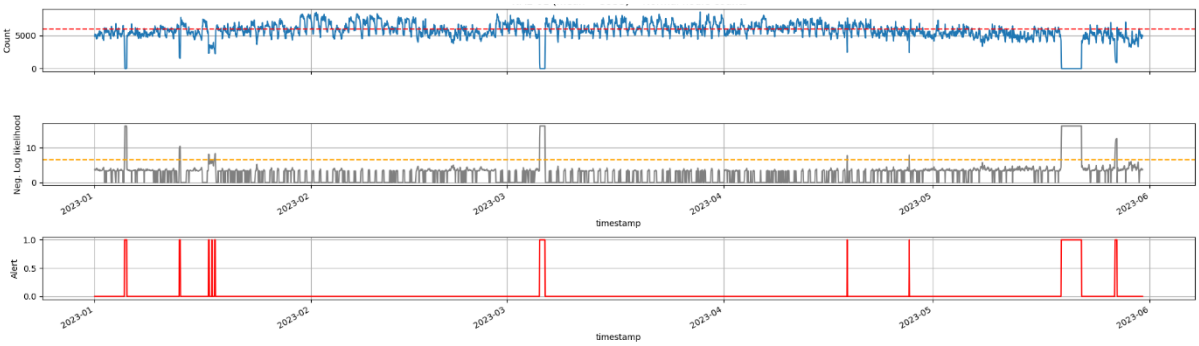

Figure 8. Anomaly detection with a Normally-distributed Streaming Data (NAE-01) Top: Count, Middle: Negative log-likelihood; Bottom: Alert Signal.

5 Conclusions

This paper discusses the challenges of monitoring data streaming quality in Building Automation Systems by testing various statistical distributions: Poisson, Normal, Log Normal, Hourly Normal, and Hourly Log Normal. Each device was fit to one of these distributions after conducting individual fitting tests. Each model has proven effective in detecting anomalies in data streams, which is crucial not only for identifying issues in streaming and investigating the source of the problem but also for ensuring the digital twin accurately reflects the building state.

Our analysis shows that while the Poisson distribution was fit to only one device under the assumption that the mean remains consistent throughout the entire analysis time window, Normal and Log Normal distributions were suitable for nine devices. These distributions efficiently handle general trends and skewed data, which frequently occurs in building systems. For ten devices, data was distributed either normally or log-normally by hour. The remaining four devices exhibited patterns that require more advanced modeling to accurately capture their data distributions.

The method was successfully applied to detect streaming interruptions or deviations from normal operations which significantly reduced the rate of false alerts that result from fixed thresholds, not accounting for the distribution of streaming data. Moreover, this algorithm has the advantage of simplicity, and can be deployed on edge devices within the building or on IoT devices with minimal computational overhead. Future work will aim to enhance these models to

improve their predictive accuracy and adapt to more complex temporal patterns in variable rate streaming for digital twins.

6 References

ASHRAE, 2001. Standard 135-2001: BACNet-a data communication protocol for building automation and control networks, Atlanta, Georgia, USA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Bamgboye, O., Liu, X. & Cruickshank, P., 2019. Semantic stream management framework for data consistency in smart spaces. Milwaukee, IEEE, pp. 85-90.

Bolchini, C., Geronazzo, A. & Quintarelli, E., 2017. Smart buildings: A monitoring and data analysis methodological framework. Building and environment, 121( ), pp. 93-105.

BuildingSMART, 2024a. IFC Building Controls Domain. [Online] Available at: https://standards.buildingsmart.org/IFC/RELEASE/IFC2x3/TC1/HTML/ifcbuildingcontrolsdomain/ifcbuildingcontrolsdomain.htm [Accessed 22 May 2024].

BuildingSMART, 2024b. IFC Performance History. [Online] Available at: https://standards.buildingsmart.org/IFC/RELEASE/IFC4/ADD1/HTML/schema/ifccontrolextension/lexical/ifcperformancehistory.htm
[Accessed 22 May 2024].

Chinchero, H., Alonso, J. & Ortiz T, H., 2020. LED lighting systems for smart buildings: a review. IET Smart Cities, 2(3), pp. 126-134.

Deng, M., Menassa, C. & Kamat, V., 2021. From BIM to digital twins: A systematic review of the evolution of intelligent building representations in the AEC-FM industry. Journal of Information Technology in Construction, Volume 26.

Ghansah, F., 2024. Digital twins for smart building at the facility management stage: a systematic review of enablers, applications and challenges. Smart and Sustainable Built Environment, Volume Ahead-of-print.

Ghansah, F., 2024. Digital twins for smart building at the facility management stage: a systematic review of enablers, applications and challenges.. Smart and Sustainable Built Environment, ( ), p. (in press).

Lillstrang, M. et al., 2022. Implications of properties and quality of indoor sensor data for building machine learning applications: two case studies in smart campuses. Building and Environment, 207( ), p. 108529.

Acknowledgements

This research was funded by the Natural Science and Engineering Research Council (RGPIN-2018-04105 and ALLRP 544569-19) and FuseForward Solutions Group. The support of Toronto Metropolitan University Facility Management Department is also gratefully acknowledged.