# An interactive error-correcting approach for IoT-sourced event logs

Mohsen Shirali[1], Zahra Ahmadi[2], Carlos Fernández-Llatas[3,4], Jose-Luis Bayo-Monton[3], and Gemma Di Federico[5]

[1]Computer Science and Engineering, Shahid Beheshti University, Tehran 19839-63113, Iran (m_shirali@sbu.ac.ir)
[2]Research Centre for Information Systems Engineering (LIRIS), KU Leuven, Brussels 1000, Belgium (zahra.ahmadi@kuleuven.be)
[3]Process Mining 4 Health Lab–SABIEN-ITACA Institute, Universitat Politècnica de València. Valencia 46022, Spain (jobamon@itaca.upv.es, cfllatas@itaca.upv.es)
[4]Department of Clinical Sciences Intervention and Technology (CLINTEC), Karolinska Institutet, Stockholm 17177, Sweden
[5]Technical University of Denmark, Kgs. Lyngby, 2800, Denmark(gdfe@dtu.dk

Corresponding author: Mohsen Shirali (m_shirali@sbu.ac.ir)

**Abstract**

Although Internet of Things (IoT) systems are widely used in various industries, they are prone to data collection errors due to device limitations and environmental factors. These errors can significantly degrade the quality of collected data and the event log extracted from raw sensor readings, impact data analysis and lead to inaccurate or distorted results. This article emphasizes the importance of evaluating data quality and errors before proceeding with analysis. The effectiveness of three error correction methods, a rule-based method and a Process Mining (PM)-based method which are adjusted for a smart home use case, and their combination was also investigated in resolving log errors. The study found that understanding different types and sources of errors, and adapting the error correction algorithm based on this knowledge of error sources, can greatly improve the algorithm's efficiency in addressing various error types.

**Keywords. IoT, Data quality, Error Correction, Process Mining, Rule-Based Correction, Noise, Missed Events**

## 1 Introduction

The Internet of Things (IoT) has become a part of people's everyday lives and as smart devices become ubiquitous, they will continue to integrate into our day-to-day activities [1]. From a logical viewpoint, an IoT system can be depicted as a collection of low-cost smart sensors that can continuously collect different types of data, from ambient environments such as room temperature, light or humidity to personal data such as basic physiological data [2]. These smart sensors could be embedded within the devices people wear (like wristbands, smartwatches or smart glasses), the devices are used in a home or installed in an environment (including security cameras, smart assistants and smart locks or even simpler devices like smart thermostats and movement detectors) [3].

The IoT technology has the potential to create a new "cyber-physical" world, in which "things" can directly operate, act and influence the physical world. Hence, the continuous flow of data from the physical to the digital world, provided by IoT devices, will extend the situational awareness of computers, thus, gaining the ability to act on behalf of humans through ubiquitous services [4]. Indeed, the data collected by IoT devices, when mined using data mining techniques and algorithms, gives insights about a given phenomenon, person or entity which can be used to provide intelligent services [5,6]. Thus, data is the main valuable asset of every IoT system and its quality

is of great importance. Using data with degraded quality, due to the occurrence of errors[1], may complicate further analysis and lead to misleading results or wrong decisions [7,8].

Incomplete data and erroneous (or incorrect) data are two main subtypes of errors that mostly occur during data collection by IoT devices. Both of these types of errors can compromise the quality of the collected data. These errors can be caused by various factors, such as damaged or depleted sensor batteries, data loss or corruption at the network level, leading to unstable or incomplete sensor readings [4, 9, 10]. Additionally, environmental factors such as local weather conditions, improper device placement, range limitations, and device malfunctions can all affect sensor operation and result in incorrect data [7]. In these cases, high-quality sensors such as industry-grade sensors could help, however, due to the need for deployment of a large number of sensors in a network, which is the case for many IoT applications, it is not always feasible to use expensive industry-grade sensors and most IoT applications use low-cost sensors, at the expense of losing data quality. Furthermore, sometimes retransmitting packets to make up for lost data caused by wireless communication failure is not feasible. Reasons like the complexity of identifying lost packets for retransmission, the impracticality of deferring decision-making in delay-sensitive IoT applications, and power limitations prevent retransmission [7,11].

Moreover, when sensors' raw data is collected, they should be prepared before starting any analysis. Typical sensor data preparation [12] or pre-processing steps are cleaning, formatting and event abstraction to derive a meaningful event log from sensor data. In this way, various types of errors in the sensor log, mainly stemming from the intrinsic characteristics of IoT sensors, could be propagated or amplified into the event log by event abstraction. Therefore, these event log quality issues (e.g., incorrect events) are mostly inevitable and they greatly hamper the mining techniques [8].

Many studies express that sensor data quality plays a vital role in IoT applications and state the importance of data quality for data mining purposes [4,13,14]. The potential degraded quality issues stress the need for through data quality assessment, *i.e.* determining whether data quality issues are present in the event log. Awareness regarding the prevailing data quality issues can help to take initiatives to alleviate them [15]. In this way, to take the most out of IoT technology's full potential, before starting any analysis and applying mining techniques on data, its quality and accuracy should be ensured to prevent misguided decisions [8]. If it is possible, the errors should be detected or quantified and removed or corrected in order to improve sensor data quality [7].

Several methods have been proposed to detect, remove or correct the errors before applying data mining algorithms. However, the challenge is that the existing methods, proposed to solve errors in sensor data, cannot be compared due to different evaluation processes. In addition, the relation between the sources of errors in the log, their impact on deviating analyses and how the correction techniques can be designed and adjusted to error sources in order to properly correct the errors are not clear. Therefore, this motivates us to investigate the impact of errors within a log collected in a real-life scenario and try to improve the log quality by correcting the errors.

To be more specific, we have focused on IoT usage in healthcare applications, which also suffer from data quality issues. Despite the great potential of mining algorithms to analyse healthcare data, the reliability of their outcomes ultimately depends on the quality of the input data used by the algorithms. Applying mining techniques to low-quality data can lead to counter-intuitive or even misleading results [16]. For our study, we have chosen a smart home case study in healthcare and used a dataset collected by ambient binary sensors in a solo-resident house. This particular scenario was selected to provide a consistent platform for comparing the effectiveness of multiple error correction techniques in addressing various types of errors. We aim to investigate the impact of adjustments made to existing methods in addressing errors and examine our proposed hybrid approach for inspecting a log to identify possible errors.

We initially processed the event log to identify the error types that existed in the dataset and the most problematic sensors and events and employed a rule-based approach tailored to our case study. Then, a modified Process Mining-based method and a proposed approach by combining both of these error correction methods are used to detect and correct the errors in the log. We compare the outcomes of each method and the combination of them based on the types of errors that are resolved.

The contributions of this study are: Firstly, we consider a use case of IoT in healthcare and a single dataset to provide a fair comparison between different methods. The healthcare area is selected due to the increasing need for development of IoT-based services in future, and the importance of addressing errors in this field to provide reliable analysis. Secondly, as the main

[1]The definition of error in this article is mainly close to the definition provided in articles like [7], and we are not considering other definitions related to data quality in database-level or high-level applications.

novelty of the paper, we leverage the experts' knowledge of sensors' limitations and identified error sources to adapt and enhance the correction algorithms to improve accuracy. Thirdly, we involve experts' knowledge in the interactive development of error correction methods to modify the methods according to the required data quality demands and to ensure the correctness of the changes in the log. Finally, with the proposed approach, we contribute to combining multiple error corrections to form a hybrid method to address different types of errors effectively.

The rest of this article is organized as follows: Background knowledge section provides the required information to familiarise readers with the AAL systems by looking into their applications, declares the definition of error, its types and the importance of addressing errors, and reviews the existing error correction methods in the literature. In addition, the dataset, which is used in this research and its features are presented in this section. Next, Error Correction methods section describes the error correction techniques and how they are adapted to our use case for addressing errors. Then, in the Results and Discussion section, the results and performance of error correction techniques and the combination of them in the hybrid approach are evaluated and we will discuss the concerns related to the data quality issues in IoT systems and error correction techniques. Finally, section Conclusions finalise the article and sums up the article's findings.

# 2 Background knowledge

To have a common understanding of IoT systems and their applications in the healthcare domain, and Ambient Assisted Living (AAL) systems as the selected case study in this manuscript, we first introduce these systems in Section 2.1 to familiarise the readers with these systems and existing sensors for data collection. Then, in Section 2.2 the error definition and its various types, which are applicable to the selected AAL scenario, are briefly defined to avoid any confusion and to establish a common ground for later usage. The significance of errors and the concerns related to why and when it is important to investigate and address them is also discussed in Section 2.3. Section 2.4 shortly overviews the existing related approaches for error correction. Finally, in Section 2.5 the details of the used dataset in this research, which is used for the evaluation of our proposed approach, is given.

## 2.1 Ambient Assisted Living Systems

The health sector is one of the main domains experiencing a transformation through the IoT revolution and its traditional systems and services are currently changing to improve healthcare quality and accessibility [17]. With the advent of ambient sensors embedded into environments and wearable technologies measuring physiological parameters, now it is possible to recognise people's activities, analyse individual behaviour [18–20] or monitor and predict people's health conditions [21]. Therefore, a large quantity of human-centric technologies such as smart environments and Ambient Assisted Living (AAL) solutions have been proposed in the literature to assist with services like health monitoring, behavioural modelling and interventions [22–25].

For instance, these systems proposed to take advantage of tracking and positioning in indoor locations to support activity recognition of humans in smart homes [26], monitor the daily activity of elderly people in their comfort zone [27], create a mobility pattern of users in areas equipped with sensors [28] and so on. The sensors capture readings when people perform daily routines inside the smart environments in order to gain insights on human activities, movements, and gestures which are generally known as human daily behaviours [29, 30]. Then, this behavioural information can be used, for instance, to assess the functional ability of elderlies for living independently [31, 32] or aftercare monitoring for post-surgery treatments [25].

The AAL systems are classified into three different categories based on the types of their data collection devices [25, 32, 33]. The first group use vision-based methods, like cameras to capture a series of images or videos. The second category is wearable methods that need subjects to wear devices with sensors like accelerometers or carry devices like Radio Frequency Identification (RFID) tags. Lastly, in the third solution of ambient monitoring, the sensors are installed in the residential environment to provide awareness about the resident context (location, preferences, and activities), the physical context (lighting and temperature), and the time context (hour of the day, day of the week, season, and year) [29, 34, 35]. The systems using WiFi triangulation techniques [36] or binary sensors such as plug-in power meters [37, 38], Passive Infra-Red (PIR) motion detectors [18, 25, 39, 40] and magnetic switches are listed in this category.

In this research, we used a dataset from a smart home based on the third category of AALs which is designed to detect subject movements as they occur in different locations of a solo-

resident home. The AAL systems like the one we are going to use are typically employed in medical applications for health assessment by determining the activity level of the elderlies [25], behaviour modelling [18], supporting individual behaviour analysis, detecting behavioural changes and offering a human-understandable view of the real changes of a user [28].

AAL systems use various sensors to measure physical quantities such as temperature, humidity, and motion. These measurements, along with any changes captured by the sensors, are recorded in **data logs** (also known as **sensor logs**). An **event log**, on the other hand, records specific occurrences, activities or incidents that happen in the under-control environment. While events are happening, multiple entries are added to the data log by the sensors to capture any changes in their measurements. One or more sensor measurements can then be converted into different events in the event log, which can be inferred from the raw sensor data using techniques like abstraction or aggregation. Indeed, an event log provides a timeline of actions or incidents taken by users, systems, or devices, with every single event occurring at a given point in time. While sensor logs focus on recording raw data, event logs focus on significant events, incidents, and activities extracted or abstracted from the sensor raw readings in the sensor log. To understand these terms better, consider an example where a motion detector sensor in area $A$, samples for movement every $t$ time interval. It generates a record for all samples collected between $T1$ and $T2$ ($\frac{T2-T1}{t}$ samples or records in the sensor log). As a result, the **actual event** or the event that happens in reality (e.g. moving inside area $A$) is observed by the sensor in that area, and multiple sensor readings are collected for this movement. These records are then turned into a single event; for instance, an event with the *"moving in Area A"* label, with a start time of $T1$ and an end time of $T2$. Therefore, the sensor readings or records are converted into a single event (known as **captured event** or observed event) inserted into the event log.

## 2.2 Error definition and types

The term error refers to the discrepancy or difference between the measured or observed value of a particular data point and the true or expected value of that data point. Different terms related to the error (sometimes not uniformly) have been used in the literature to describe its various types such as outliers, missing data, bias, drift, noise, constant value, etc. [7]. However, in this section, we focus on the errors that can occur and exist in the log collected by AAL systems equipped with ambient binary sensors, as our target scenario (interested readers who want to study other error types, in more detail, can refer to other studies such as [7, 15]).

The binary sensors including PIRs and contact switches, can only provide two values: *zero* and *one*. Thus, in a smart home with binary sensors, only specific kinds of errors can occur during data collection and decrease the quality of the sensor log. We categorise the possible errors for an AAL system with binary sensors into two different types:

- **Type 1) Missing events (incomplete data).** These are events that occurred in reality but were not recorded in the sensor/event log [41]. Indeed, the under-control environment is changed in such a way that sensors should be triggered, but sensors do not capture the actual event, or the event is captured but the transmission failed, hence the event is not added to the log. According to [11], unstable wireless connection due to network congestion or environmental interference *e.g.* human blockage, walls, and weather conditions, sensor device outages due to its limited battery life and malicious attacks are the most frequently mentioned reasons for these errors. In addition, the sensor coverage (sensing area or field of view) is limited and the sensors can only detect changes in a specific zone. Thus, if sensors are not properly installed, their coverage might not encompass the whole area where the activities could take place. This issue leads to missing events that happened beyond the sensor's reach. The sampling rate of sensors is limited as well and inadequate sampling rates can cause missing events in the event logs [13, 25].

- **Type 2) Noises (erroneous events).** Event entries which are present, but for which the recorded values do not reflect reality [15]. The noises happen when the sensors do not behave as expected and are present in the sensor data due to issues during logging (*e.g.* transmission of packets and writing the reported data in the dataset) or due to the presence of conditions affecting the phenomenon measured by the sensors [8]. Incorrect or inexactness of events' timing is one of the main subcategories of these errors. When there is a discrepancy between the moment at which an event took place in reality and the moment at which it is recorded in the system, the incorrect timing error occurs [42]. Another reason for incorrect events is the noise in localization systems which can caused by unpredictable reflections, refractions, or absorptions of signals due to the heterogeneity of walls and furniture of rooms, environmental

factors, or even humans, depending on their positions [43]. In addition, an inaccurate sensor placement can also lead to erroneous events. When multiple sensors, which are installed to cover different areas, cover the same area or phenomenon (*e.g.* the sensing area of multiple sensors overlap with each other), they create multiple records for a single event -occurred inside the overlapping area- with same timestamps and different labels, so that one of them is correct, while the others are noises. For instance, movement detection sensors installed in adjacent locations will detect a single movement simultaneously and generate multiple erroneous events with different labels in the log [25].

The other subtypes of errors like drifts that refer to changes in the reported values by sensors (*e.g.* constant changes in the measured value, for instance, always reporting a value higher than the actual value) are not applicable to binary sensors. In addition, some sub-types are impossible, undiscoverable or at some points equivalent to the considered types. For instance, drifted values are impossible because we only have two outputs and reporting a constant value is equivalent to either undetected events, when the reported value is zero, or noisy detection if the value equals one.

The missing event errors can sometimes be addressed by performing appropriate actions, like re-transmission of sensor readings to compensate for the loss of packets containing data for captured events and accurate installation of sensors to avoid undiscoverable areas. A post-evaluation of the log quality to find possible error sources and adjust the sensors' field-of-view and sampling rates is another effective solution to prevent missing event-type of errors. On the other side, using more accurate sensors, modifying the placement and coverage of sensors to remove overlapping areas, and filtering out events that are obviously abnormal (for instance too short or too long) are possible solutions to handle noises.

However, modifying the configuration of the IoT system for data collection to improve the quality is not always possible and even high-accuracy sensors might have noises. In addition, in some cases, data mining techniques are used to analyse offline datasets, which are collected before and cannot be changed. Thus, it is mandatory to have solutions that can address errors in the sensor and event log after data collection. Error detection and correction methods can be used in these situations to improve the quality of the data.

## 2.3 Why errors are important?

The analysis of data collected by IoT systems is primarily targeted at non-expert users such as healthcare providers and industry professionals. These individuals often lack a technical background in data analytics, necessitating the development of models and analyses that are easily understandable and interpretable. However, the presence of errors (missing events and noises) in the collected data introduces artificial undesired variability, complicating the extraction of the meaning from data. Thus, such variability hardens human interpretability and can adversely affect data mining algorithms in generating a model to describe data [44]. The algorithms are compelled to account for a broader range of states to accommodate all the discrepancies, which increases their complexity and computational cost. In some instances, these challenges can also compromise the accuracy of the generated models.

As an example, let's take a look at the field of Process Mining (PM), since we will later use its technique for error corrections. PM is a discipline that provides a set of tools to discover human-understandable models from event logs [45]. The goal of PM is to turn event data into insights and its algorithms are able to maximise the understandability of the models inferred [44]. The discovery techniques in PM are used to identify different activities and flows among them in an event log based on their time precedence. This helps in creating a model to visualise and explore the dynamic nature of the processes that existed within the event log. However, the presence of errors in the collected data often leads to a common challenge in process mining literature, known as the "spaghetti effect" [45]. This term metaphorically describes complex models that represent processes with numerous transitions, resulting in visually unreadable and difficult-to-comprehend models [46]. On the other hand, "Lasagna" processes are characterised by well-structured, relatively simple models where most cases are handled in a prearranged manner. These structured processes are particularly valuable for detailed performance analysis, deviation detection and predictions. However, the complexity and lack of structure in spaghetti models shift their added value to providing insights and generating ideas for a better understanding of processes [45].

As a result, the undesired variability introduced by errors dramatically decreases the human understandability of models [44]. This, in turn, limits the practicality of advanced analytical techniques such as predictive modelling and degrades the quality of analysis [45]. Nevertheless,

if the complexity of models is mainly caused by errors, addressing and resolving these errors can provide simpler and more readable models [47]. This fact makes the transformation of "Spaghetti-like" processes into more "Lasagna-like" processes highly beneficial. It enables advanced analysis, including history-based predictions and recommendations, and is applicable across various data mining and analysis techniques [45]. Therefore, it underscores the importance of error correction in improving the quality and utility of analytical outcomes.

However, addressing errors to eliminate artificial variations while preserving the genuine dynamic and dispersion inherent in data and processes poses a significant challenge. Distinguishing between real variations and those induced by errors is crucial for accurate analysis [48]. This task becomes particularly complex in applications that aim to identify actual variations and change points in data flows, such as human behaviour analysis [49–51]. Human behaviours inherently evolve over time, introducing natural shifts and changes that must be distinguished from errors to ensure high quality analysis. For instance, in [19], clustering techniques were employed to group behavioural patterns based on similarity, thereby illustrating the complexity of subjects' behaviours based on the number of derived clusters. But, when errors are present in the data, they introduce artificial discrepancies that diminish the observed similarities among behavioural patterns. Consequently, mining techniques fail to identify genuine similarities, yielding misleading results that inaccurately depict a user as exhibiting complex behaviour. It is also consistent with the notion "Garbage In, Garbage Out"; applying mining algorithms, like process mining, to low quality data can lead to counter-intuitive or even misleading results [15, 16].

Additionally, utilising erroneous data more than compromising the quality of analysis, can significantly decrease the reliability of analysis outcomes and the trust in systems. In particular, in health-related use cases, using erroneous data can potentially lead to incorrect diagnoses, treatments, and monitoring procedures, ultimately disrupting or deviating health-sensitive decisions and interventions. Thus, medical staff and end-users must have confidence in the decision-making systems powered by IoT devices, ensuring that these systems are consistently fed with correct and high-quality data. This trust in data quality is essential for the acceptance and adoption of insights derived from IoT systems in clinical and healthcare settings, enhancing their utility and impact in monitoring and treatment procedures. This underscores the critical importance of conducting thorough data analysis to identify and rectify errors, thereby ensuring the quality and reliability of the analytical outcomes [19].

Therefore, it is essential to assess the prevalence and impact of errors on the data quality before proceeding with any analysis. Prioritising error correction and reduction becomes imperative, especially when the frequency of errors is high, as unchecked errors can compromise the confidence and accuracy of the models and insights derived from the data [48]. Failure in data validation in terms of errors prior to analysis can result in misleading outcomes, making them invaluable and necessitate the repetition of data collection and analysis processes [52]. However, the decision to invest in error correction should also consider the associated costs in terms of time, energy, and resources. If the cost of error correction exceeds the expenses and efforts required to redesign the system and re-execute the data collection process or the accuracy gain would be negligible, it may be more pragmatic to reconsider the data collection strategy rather than focusing solely on error correction. Therefore, selecting appropriate tools and strategies to effectively address errors and undesired variations is crucial to ensure data quality and the reliability of analytical outcomes [44].

In summary, the quality and accuracy of data collected by sensors in IoT systems are pivotal for conducting precise analyses. If data errors are left untreated and inaccurate data is used for further analysis, the results and interpretations of the analysis become invalid and untrustworthy [4]. Consequently, this article aims to address the critical issue of data quality in smart home systems, providing insights, strategies, and solutions to ensure the reliability and trustworthiness of data collected from IoT devices. Data quality assurance facilitates more accurate, reliable, and actionable insights that empower healthcare providers and users to make informed decisions and interventions based on IoT-generated data.

## 2.4 Existing methods for error detection and correction

Several solutions have been proposed to quantify or detect errors in the literature. In a study in [7] different types of sensor data errors and solutions for detecting and correcting them were systematically reviewed, irrespective of the IoT architecture layer. Another possible solution to tackle data quality issues is using data cleaning heuristics, as suggested in [53]. However, these heuristics are limited by strong assumptions and are designed to address specific event log quality problems. This can hinder their practical application and may lead to incorrect and misleading conclusions if the assumptions do not hold. To ensure the validity of assumptions and determine

whether the proposed corrections make sense in practice, domain expertise is required. This stresses the potential of interactive data cleaning approaches, where professionals have complete control over the process. In the following, Section 2.4.1 briefly provides an overview of ideas and approaches to address errors in general, while in Section 2.4.2, we will delve deeper into techniques developed for correcting errors in IoT settings, specifically in smart spaces, to familiarize ourselves with the existing methods that are applicable to our case study.

### 2.4.1 Approaches for error correction

Data mining practitioners often overlook infrequent events to avoid incurring additional costs associated with error correction [19]. These infrequent events are seen as noise and removed to reduce variability and limit the spaghetti effect. On the other hand, there are also de-noising methods (such as [54]) that attempt to remove the noise associated with the measured value. These methods usually come in the form of building a normal behaviour model of the system and comparing the new observed values with the normal model. When the observed data significantly differs from the estimated value, it is identified as an outlier or anomaly, and it is either removed or corrected using an estimated value from the model.

This discarding or replacement of detected anomalies decreases the variability, leading to cleaner models that produce better understandable solutions. However, it's not always advisable to discard infrequent or anomalous behaviors as they may contain critical patterns or variations that offer valuable insights in certain contexts. They provide a view of the flow of non-standard cases, which could be interesting for experts in some domains. In domains like healthcare, where experts are familiar with standard cases, showing them the standard model may not provide any new knowledge. However, presenting non-standard cases to health professionals can help identify patients who require special attention. To detect anomalies, approaches based on statistics, machine learning, clustering and ontology have been proposed in the literature. However, standard error reduction or correction techniques have challenges in addressing infrequent behaviours [44].

There are also missing data imputation methods that are used to estimate the missing sensor measurement values or the corresponding events. Two such methods are clustering and k-nearest neighbour, which are used to estimate missing sensor values. Association rule mining is also a rule-based algorithm that can detect errors and fill in for missing data. In association rule mining, the most frequent patterns are discovered to determine how items are associated with each other, and the occurrence of a specific item is predicted based on the existence of other items [7, 55, 56].

### 2.4.2 Error corrections for data from smart spaces

As discussed before, missing and erroneous events are inevitable in data and event logs collected by IoT systems. In cases where the data collection system settings can be adjusted, some actions such as fine-tuning sensor sampling rates or adjusting sensor placement and field of view as recommended in [25] and [8] can be helpful. In these cases, a post-hoc solution is required to identify weaknesses in the data collection system and the most problematic areas. Otherwise, data cleaning techniques must be applied to address errors.

To resolve errors caused by improper sensor sampling rates or placement, events that are too short or too long can be filtered out. The literature suggests using a threshold duration (e.g., one minute in [19, 57] and 24 seconds in [58]), which can be refined based on the sensors' sampling rates or the physical condition of the location being monitored by the sensors. An Inductive Miner infrequent (IMf) discovery algorithm with an adjustable threshold is also used in [59] and [60] to determine the level of infrequent behaviour to be included in the mined model mined and the events which should be left out of the model.

Regarding missing events, both [15] and [61] have pointed out that control-flow discovery algorithms can be used to detect missing events caused by blind spots in smart spaces. The errors can be identified and corrected using a model that represents the system's valid behaviour and this model can also be adjusted by experts in an iterative and interactive manner, as described in [62]. For instance, [63] presented a method to identify the most probable missing events based on path probabilities of process models developed by domain experts and a method is proposed to add these missed events. In [53] also an interactive discovery-based framework is developed to evaluate data quality problems by visualising incorrect relationships between activities in the available data.

Additionally, [25] and [43] are two works which employed a discovery-based approach using process mining on IoT-sourced event logs for data quality assessment and error correction. In [25], a process mining technique is used to create a human-readable model for transitions between different

areas of a house equipped with a PIR-based movement detection system. The discovered model highlighted many invalid transitions caused by detection errors, enabling the system designers to identify the most problematic sensors. By redeploying noisy sensors, the quality of the logged data was significantly improved. Furthermore, [43] proposed a post-hoc process mining-based solution to investigate noises[2] in a log collected by real-time location systems based on the characteristics of the physical environment. The authors used process mining to compare the current model with a pre-determined valid model, identify deviations, and correct the traces (*e.g.* a sequence of events from the event log occurred within a specific time interval) with minimum changes to the log. This method is used and adjusted as one of the solutions in the current paper, with more details provided in Section 3.1.

In addition to improving the quality of logs, error correcting techniques can also be utilised in certain analysis applications. [64] have proposed a method that uses these algorithms to provide a more comprehensive overview of compliance and identify the most common types of deviations. They provide a conformance-checking approach that employs error-correcting methods to measure the degree of deviation from clinical guidelines in the treatment process of patients, by integrating healthcare professionals' knowledge. In another study on human behaviour modelling [19], the error correcting method from [43] is used to determine the distances between various workflows and group them based on a predefined similarity ratio.

There are also hybrid approaches, which incorporate more than one type of method in detecting and correcting sensor data errors, such as the sensor validation technique in [65], which uses predictive polynomial filters and fuzzy rules together.

## 2.5 The case study dataset

This study uses a dataset [66] collected from a house with a 60-year-old female resident. It contains data from ambient sensors installed within a house, wristband data, smartphone information on the usage of mobile applications and daily psychological reports for a period of 147 days. However, we only use data captured by ambient sensors in this study.

Ambient sensorial information includes multiple PIR sensors showing the presence of the subject in different areas of the house. A power usage sensor indicating TV usage, contact sensors highlighting the opening and closing of the Bathroom, WC and closet doors, and a gas detection sensor that detects cooking activity. The ambient data is pre-processed to extract information about the subject's presence in different areas of the house. Through this process, we have been able to detect events related to the visited home locations for the entire dataset period.

We evaluate this dataset, which we call it FR-dataset hereafter, in terms of errors. In addition, we applied multiple approaches including a process mining-based error correction, a rule-based error filtering method and a combination of both on the location event log to investigate how these algorithms can handle the error problems.

### 2.5.1 Locations and duration of events

To determine the exact location of the resident inside the home using ambient sensorial information, we can look at the locations in which sensors are mounted and use the timestamp of logged data to infer the person's location at different times [25, 67, 68]. For instance, a record captured by the bedroom's closet sensor indicates that the subject was in the bedroom at that time. Then, by correlating each sensor to its installation location, every captured record can be mapped to a location. Therefore, if we have the whole sequence of captured records for a specific duration, then events related to the person's presence in different areas of the house and even the sequence of they movements can be inferred [18].

The FR-Dataset is collected by using *15* ambient sensors in total, which were installed in *six* different areas of a house. The sensor locations and their types are shown in Figure 1a.

The pre-processing step of our study uses a set of simple rules over the records generated by the PIR and contact sensors readings to discover the subject's presence in *seven* different areas of the house and assigns the following labels to the discovered events at each place: Kitchen, LivingRoom, Bedroom, Corridor, WC, Bathroom and Entrance[3]. The duration of each event is also calculated

[2] In process mining terminology, the noise term is used to describe incomplete, abnormal or inaccurate events, and it is equivalent to the error term, which is used in pattern recognition and the IoT community as a general word to describe all types of incorrect/incomplete measurements or events. In this article, as defined in Section 2.2, we followed the terminology of error in IoT.

[3] The specific location names of the areas in the FR-dataset are written with capital letters, like Kitchen or LivingRoom. However, small letters are used to refer to places in general.

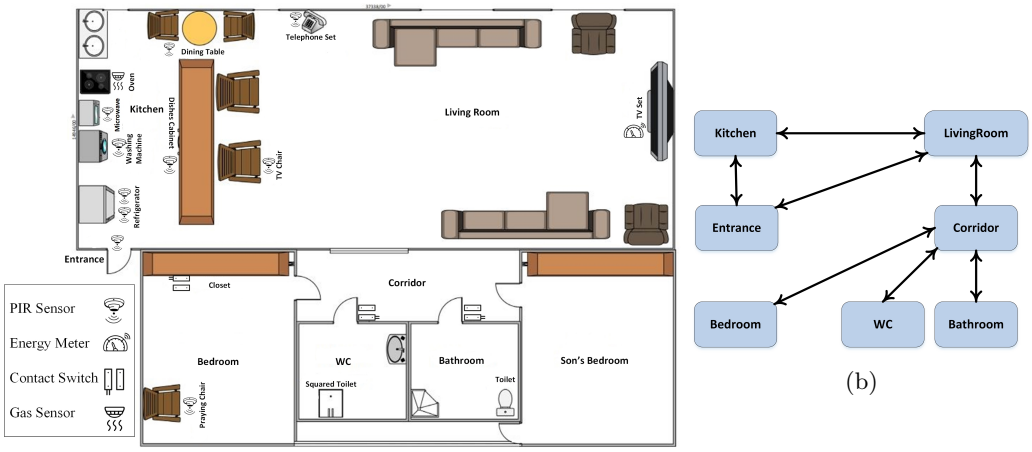



Figure 1: (a) The house floor plan, location and types of ambient sensors (b) Possible transitions between various areas of the house.

by considering the time difference between the first sensor reading in that location and the time of the first reading by sensors in another area.

Figure 1a also illustrates the walls, doorways and furniture of the house, which dictate the possible transitions of the subject between adjacent areas inside the house. Indeed, using the house's floor plan, we can easily determine the valid transitions between different areas of the house (Figure 1b). Later, these possible transitions can be used to check if the sequence of captured events is reasonable and to verify the correctness of observed events.

### 2.5.2 Dataset errors and their sources

Error correction methods are helpful for cleaning data from unavoidable errors caused by IoT systems intrinsic characteristics and improving the data quality and subsequently the analysis accuracy. But, cleaning errors from a log is only possible by discovering their origins. Indeed, to discriminate actual events from noisy captured events, we should first understand why the errors are generated and then remove the events, that were possibly added to the dataset owing to those error sources.

As mentioned, behaviours are composed of sequences of routines and series of activities at the lower level [69]. Regarding the FR-dataset, in fact, the purpose of the system is to detect specific activity times and durations to map them into behaviour patterns. Hence, the sensors are installed in different areas of the house to highlight the time of using some home's utilities or discovering some specific activities, such as cooking, watching TV, hygiene activities, or doing home chores.

In this study, we use the available sensor readings from the FR-dataset, corresponding sensor locations and their triggering timestamps to extract the location information and create a log which we refer to as the *location event log*. However, due to the lack of sensors in doorways, it is not possible to precisely determine the exact time of resident's transitions between the areas. Also, there are many undiscoverable areas on the home floor plan, which are not covered by any sensors, like the Corridor, son's bedroom and most part of LivingRoom (out of the sensing range of the two existing PIRs). Thus, the applied location discovery algorithm may produce some unintended errors in the identified locations, the timestamps and sequence of events and the transitions between places. This study's objective is to address these issues and make the necessary corrections to provide a reasonable sequence of events.

To demonstrate what kind of errors are expected to be generated by the pre-processing algorithm, consider the example illustrated in Figure 2. It shows a part of the sensor log that contains all sensor reports on the mentioned timestamps, the extracted corresponding location event log, and three sample paths drawn on the map. Each *one* value in the sensor log shows a data record (*e.g.* an actual event or activity) captured by a sensor, and the timestamp of that row indicates the occurrence time of the sensor measurement, which in fact expresses the exact time of the subject's movement within the sensing area or her interaction with a utility in the home.

Assuming the data log and discovered locations in the event log in Figure 2, the first timestamp (1/8/2020 23:04:33) shows the time of the subject's presence in the Bedroom and the second timestamp (1/8/2020 23:05:39) shows the time when the subject opened the Bathroom door and

| Date | TV | Telephone | TV Chair | Dining Chair | Bathroom | WC | Closet | Entrance | PrayingChair | Refrigerator | Fridge | Washing Machine | Microwave | Stove | Kitchen Cupboard |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1/8/2020 23:04:33 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/8/2020 23:05:06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/8/2020 23:05:39 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1/8/2020 23:07:21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| Event/Room | StartTime | EndTime | Duration |
|---|---|---|---|
| Bedroom | 1/8/2020 23:04:33 | 1/8/2020 23:05:39 | 00:01:06 |
| Bathroom | 1/8/2020 23:05:39 | 1/8/2020 23:07:21 | 00:01:42 |
| Corridor | 1/8/2020 23:07:21 | 1/8/2020 23:07:32 | 00:00:11 |
| Kitchen | 1/8/2020 23:07:32 | 1/8/2020 23:07:45 | 00:00:13 |
| Bedroom | 1/8/2020 23:07:45 | 1/9/2020 0:50:20 | 01:42:35 |

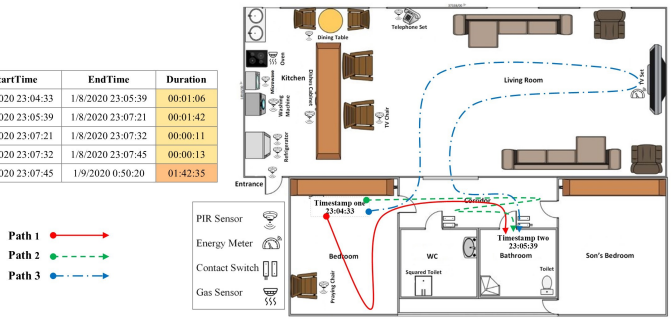



Figure 2: An example piece of actual log and the corresponding detected events.

entered the Bathroom. The Corridor is one of the undiscoverable areas of the house, hence, given the two timestamps mentioned in this particular case, it's difficult to precisely determine when the subject left the Bedroom and entered the Corridor. The subject could spend more time in the Bedroom and finally go to the Bathroom by passing the Corridor (following Path 1), go to the Corridor and spend some time there (Path 2), or even it is possible for the subject to leave the Bedroom and visit other undiscoverable spaces such as the LivingRoom and then enter the Corridor again and Bathroom at timestamp two (drawn as Path 3).

The specific times of entering and leaving certain locations like Bedroom and LivingRoom cannot be determined with certainty, except for the WC and Bathroom which have micro-switch sensors. As a result, the required location event log must rely on estimations based on previous and subsequent log entries. Despite efforts made by the pre-processing algorithm to generate accurate predictions based on captured sensor readings, there may be instances where some events are missed or incorrectly identified.

Nevertheless, it is expected that these predictions and undiscoverable areas could lead to a high amount of errors in the captured events. Subsequently, the missed transitions result in some mismatches between the discovered sequence of events and the possible transitions defined based on the floor plan of the house (Figure 1b). Every pair of consecutive events that indicates an impossible direct transition between two areas of the home is labelled as an error (such as direct transitions between the Bedroom and Kitchen). The validation of the transitions in the location event log showed that 4269 out of 12494 transitions (34.16%) are invalid.

According to this preliminary analysis of event log quality, it contains many errors, and we have to try to remove or correct them (or at least part of them) by applying an error correction algorithm. In the following section, the proposed and implemented error correction methods are described.

# 3 Error Correction methods

As previously mentioned in Section 2, when collecting data using IoT devices, errors such as missing events and noises are inevitable and our case study log, *location event log* also had some data quality issues. To address these issues and prevent a decrease in accuracy for future data analysis, we have adapted two approaches from existing error correction techniques to be suitable for our scenario. We tailored a PM-based correction technique, inspired by the one proposed in [43], and a rule-based error correction technique to correct misdetected noises in the location event log. We then proposed a novel interactive and hybrid method by combining both techniques.

In our smart home scenario, we first discover the reason for capturing erroneous events by analysing the intrinsic specifications of the installed sensors, such as sensing range and sampling rate. To achieve this, the validity of travelled paths and the sequence of visited areas in the house are evaluated to discover the errors and identify the most problematic sensors. A travelled path is invalid if it includes impossible direct transitions according to the house map. We have utilised a process mining technique to extract a model from the pre-processed events and checked the conformance of this model with an existing reference model with valid transitions to find the deviations.

In the next step, we used error correction techniques adjusted for our case study to correct

the discovered errors. We have contributed to using the knowledge about the most frequent and problematic error sources to highlight events with a higher likelihood of being errors, then inspect and correct them in a heuristic and semi-supervised manner. The output of this step, which we called the rule-based approach, is then fed to a modified version of a PM-based correction technique, which is implemented in [43]. The modified PM-based correction tries to make the event sequences reasonable and valid by considering two constraints: without deleting any event record or just deleting events from specific areas with a higher probability of being errors. More details on the correction techniques and the process of applying them, along with an explanation of their adjustment and the proposed hybrid approach are provided in this section.

## 3.1 Preliminary PM-based error correction technique [43]

The PM-based correction technique [43] detects all impossible transitions and defines a penalization score for the possible correction options. The penalization score is a value, which the algorithm calculates for each single correction option to indicate the amount of changes that are needed based on their impact on the actual log. After measuring the penalization score for all possible correction options, the algorithm selects the correction option with the minimum penalization score and corrects the actual trace accordingly.

Long-duration events are typically encompass multiple sensor records in the sensor log, abstracted in a long-duration event, which means that there are fewer chances for them to be errors. The reason is that the probability of generating multiple noisy reading by multiple sensors or even a single sensor during a long interval is very low. For example, if a person spends an hour in the kitchen and uses appliances like the refrigerator, washing machine, or microwave, all the sensor readings will be combined into a single event with a "Kitchen" label (See Figure 4). The PM-based correction technique aims to correct a log with minimal changes in terms of duration. This approach involves assigning a penalty score to each change based on its duration, and this score is then multiplied by the square of the duration of the selected event for correction. In this way, the penalization score for long-duration events is high, and they should only change if no other correction with less penalty exists. The PM-based technique corrects the discovered errors by adding or removing transitions to create a valid trace with the minimum penalty score.

To ensure a timely and efficient correction of the process model, a pre-determined time-out (5 seconds in this case) is considered in the implemented algorithm. If the time-out elapses and no solution has been found to correct the trace, the erroneous trace will remain unaltered. It's worth noting that longer pre-defined time-outs lead to the evaluation and measurement of the penalty score for exponentially increasing higher possibilities and taking more processing time. In addition, different penalization scores can be determined for adding a node within a trace, deletion of a node and fusion of the events with the same label (we have used 100 for add and delete scores and *zero* for fusion).

Moreover, the PM-based method needs a reference model, which includes valid transitions (such as Figure 1b) to inspect the actual traces and correct the errors. By adjusting the algorithm parameters, like the penalty scores or the reference model, the correction method will produce different outputs. Finally, the output is a sub-optimal log with the best possible traces (*i.e.* probably it is possible to get better corrected logs by spending more time for correction, but it is not efficient in terms of processing). Figure 3 shows the modified output of the PM-correction technique method derived from [43] for the provided piece of the location log. Since the Corridor-to-Kitchen and Kitchen-to-Bedroom direct transitions are impossible, the event with the Kitchen label is deleted by the PM-correction technique, and the current trace is logically acceptable.

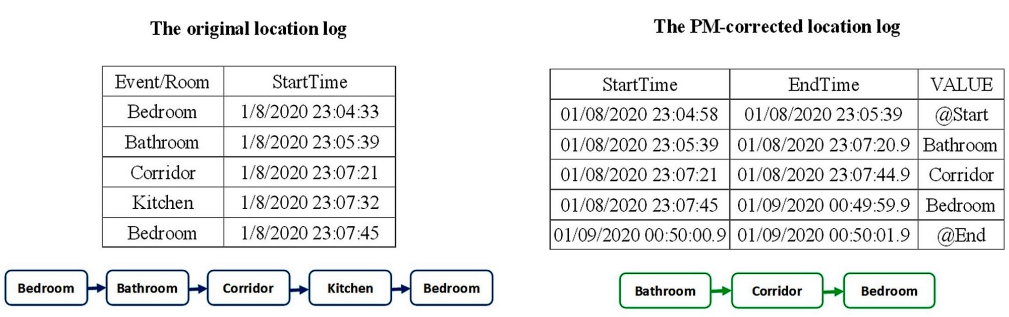



Figure 3: The output of preliminary PM-correction method.

## 3.2 False corrections of preliminary PM-based error correction

The decision to not remove long-duration events to prevent too much change in the log in some complex situations can be inaccurate and lead to the removal of some valuable parts of the information in the log. In these cases, the PM-based correction will try to delete short-duration events to correct the trace, consequently, it may falsely remove some transitions (which probably happened in reality).

In the example shown in Figure 4, the PM-based correction algorithm removes the Kitchen events and combines the Bedroom events because the total duration of the Bedroom events is longer than that of the Kitchen events, and the Bedroom-Kitchen transition and vice versa are not valid. However, the actual log on top, clearly indicates multiple sensor detections in the Kitchen, making it highly unlikely that these consecutive detections from different sensors were false. Thus, it is more reasonable to assume that Kitchen events occurred with greater certainty, while transitions in the undiscoverable areas of the LivingRoom and Corridor were not captured.

Nevertheless, this example shows that it is not always correct to consider shorter events as the most probable and reasonable sources of errors, and removing these events will somehow change the semantics of the log. In particular, when the purpose of analysing logs is to detect and identify changes and anomalies, every visited location, no matter how brief, provides valuable information. As such, it is important that the correction algorithm does not alter or remove these locations. Ultimately, the accuracy and certainty of events should be taken into account when deciding whether or not to remove them from the log.

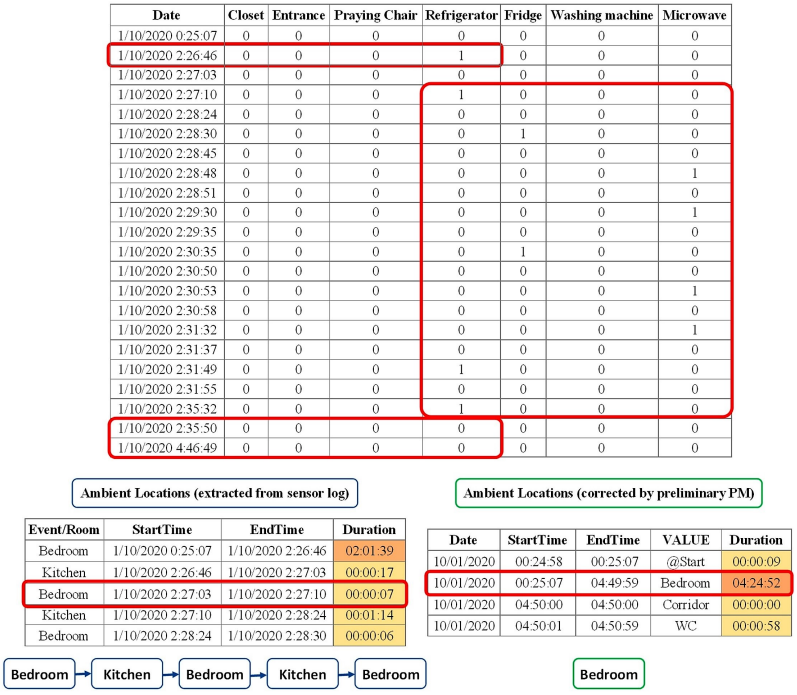



Figure 4: Example of false correction by PM-based algorithm.

To improve error correction in the PM-based method, the algorithm is modified to prioritize adding actions over deletions. This is done by assigning lower penalization for adding, which results in correcting the traces by adding missed transitions instead of removing events. This modified version of PM-based correction method limits the use of deletion changes for events, especially in areas like the Bathroom and WC, where more reliable micro-switch sensors are installed. Additionally, we have adapted and developed a rule-based correction approach (described in Section 3.3) to remove incorrect events from the log based on the certainty of events (i.e., the number of sensor records for each event) before applying PM-based correction.

## 3.3 Rule-based error correction

To correct the errors in a log using a rule-based approach in a semi-supervised way, a set of rules should be heuristically defined to highlight possible errors in the entire log for further inspection, and as explained in section 2.4.1, it has been used in multiple studies so far. Our contribution is to integrate the knowledge of IoT devices, their challenges, and the reasons for capturing errors

(as described in Section 2.2) into the rule-determination process. First, we must identify the most problematic areas of the house and sensors, and then determine what type of errors can be generated by sensors in each individual room. Once we have this information, we can establish one or multiple rules for each location to detect doubtful events.

These doubtful events are then analysed by experts, who consider the predecessor and successor events of that trace and determine the most reasonable action to correct the trace. For instance, experts can check the certainty of highlighted location events by counting the number of sensor records that occurred within the time interval of those events and decide to remove them if the events are recognised as possible errors. Therefore, if accurate rules are defined based on the expected behaviour of the system and applied to the log, this rule-based approach can find unusual and incorrect events, and then clean the dataset from noisy and false detections. In addition, the PM-based approach can complete the rule-based approach and the correction process by recovering missed transitions. Both approaches, together and when they run back-to-back form a hybrid error correction method.

Based on the experiences of IoT systems experts who observed the behaviour of ambient binary sensors over a long period and recommendations found in the literature, the following issues regarding the usage of PIR sensors in smart homes can lead to possible errors, and they are carefully considered in the determination of rules.

- Using PIR sensors in areas without walls such as open-plan kitchens or living rooms could record some unnecessary detections due to their wide sensing angles [25]. For instance, in the house used for the FR-dataset collection, the PIR sensors in the Kitchen (such as the sensors mounted on the microwave and the refrigerator) at some points observe the movements in the LivingRoom because they can see the movements from the top of the cabinets. Hence, there are some unnecessary captured events in the Kitchen and transitions between the Kitchen and LivingRoom in the event log, which should be removed. These consecutive Kitchen and LivingRoom events have unusually short durations and can be highlighted by considering a minimum duration threshold.

- Adjacent sensors might observe a shared area and movements, and if these sensors belong to different areas (such as the sensor in the Entrance and the refrigerator sensor in the Kitchen), the overlap between them will also create unnecessary events. Therefore, ping-pong-like transitions are created for the duration of the subject's movement within the overlapping area. These captured events can be differentiated again due to their short durations.

- As mentioned in Section 2.2, the sensors may not detect some fast movements due to their sampling rate restrictions. Hence, we may have some events with long durations because of these undetected transitions. As a result, a threshold on the maximum value of the duration at each area could also be useful for selecting suspicious events of this type for further inspection.

In addition to the aforementioned issues, we should again emphasize that in the house used for collecting the FR-dataset, the existence of multiple undiscoverable areas leads to many missed transitions that we ignore at this step and have left them to be handled by the modified PM-based error correction method. To summarize, we have contributed to the error correction problem by using a rule-based correction method. The rules are defined based on recognised weaknesses of IoT systems, as identified by the experts, to correct noise-type erroneous events. Additionally, we'll use a modified PM-based correction method to address and correct the missed transitions and some noises that are not handled by the rule-based method.

By taking the described issues into account, two types of rules are determined in the proposed rule-based error correction approach:

- **Type 1)** Events with unusual long-duration. For the places, which we know the subject uses for a short time (WC and Bathroom) or just for transition and reaching other areas (Corridor) a threshold for the maximum duration is used for highlighting potential noises. In addition, for other areas, the maximum threshold is also helpful for finding very long-duration events.

- **Type 2)** Events with duration less than usual. A minimum duration threshold can highlight the unusual events of each place for further inspection. We have adapted the rule-based approach to leverage a minimum threshold just for the places where we expect the subject to be there for a long time, *i.e.* LivingRoom, Kitchen, Entrance and Bedroom. Short-duration events for other areas are reasonable; hence, the minimum threshold for them is not beneficial.

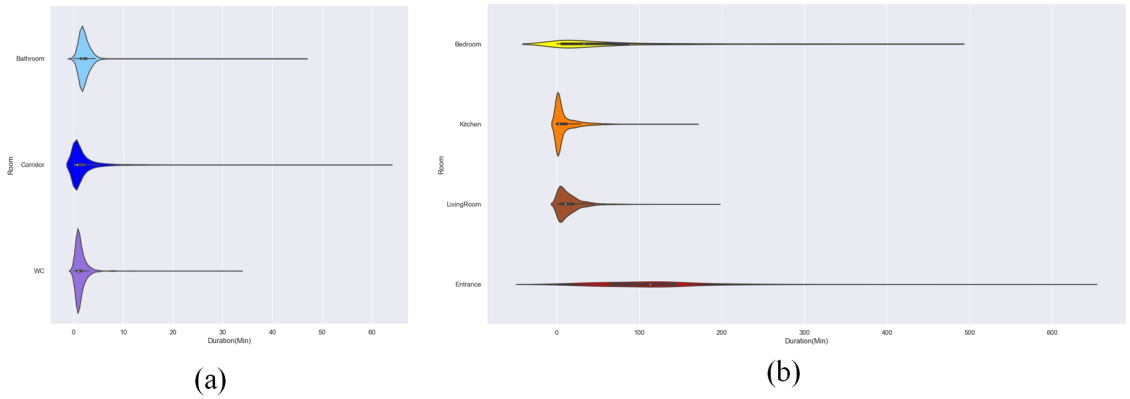
Thus, the FR-dataset is investigated by using different criteria to see which method and values for the thresholds could possibly detect noises better and will produce more precise results. A typical way to determine the thresholds is to use statistical measurements. The average or median and the standard deviation values are normally used within Formula 1 to calculate the minimum and maximum thresholds for each location. Table 1 reports the values measured for each location over the entire duration of observation.

$$Threshold = Average(or\ median)\ \pm X * Standard\,Deviation \quad X = 1, 2\ and\ 3 \tag{1}$$

Table 1: The measured statistical values for each location over the whole log duration.

| Location | Average | Median | Standard Deviation |
|---|---|---|---|
| Bathroom | 00:02:21 | 00:01:48 | 00:02:34 |
| Bedroom | 01:08:12 | 00:32:58 | 01:28:40 |
| Corridor | 00:01:54 | 00:00:44 | 00:03:17 |
| Entrance | 01:55:09 | 01:53:21 | 01:13:36 |
| Kitchen | 00:09:37 | 00:02:42 | 00:15:03 |
| LivingRoom | 00:15:12 | 00:10:27 | 00:16:25 |
| WC | 00:01:25 | 00:00:59 | 00:01:41 |

The number of events that satisfied these conditions was counted. The variability of durations for the events of each place in the FR-dataset is too high, and the positive and negative ranges of standard deviation cover a wide interval, which highlights a significant number of events for inspection. Hence, we concluded that these statistical values are not helpful in calculating our case's min and max values. Figure 5 shows the variability and dispersion of the events' durations for each area in a violin chart. The locations of the house are separated into two groups, one for places that are mostly used for a short period, and the other for areas that can be used for a short or long time.




(a) (b)

Figure 5: The variation of events captured in each area over the whole duration of data collection. (a) the areas which are mostly used for a short period, *i.e.* Bathroom, Corridor and WC, (b) The areas are used for both long and short stays including Bedroom, Kitchen, LivingRoom and Entrance.

Therefore, we decided to interactively define the threshold values in a way that the number of events satisfying the conditions in total does not exceed a pre-determined value of changes to avoid too much alteration of the event log. To this end, up to *five* percent changes in the event log are defined as the error correction method's objective. We decided to define the rules, which led to highlighting at most five percent of the total events for further inspection and considered the following points in the rule-determination step:

1. For the areas which are only used for a single purpose during a short time, including Bathroom, WC and Corridor (which is just used for transition and moving from one area to another), we will consider a top threshold, then inspect the events with higher duration to see whether they are noises or not. The values of "$Mean + (2 * STD\_Dev)$" as the maximum threshold for all these areas lead to less than 5% of the events, which satisfy the selected goal.

2. Bedroom, LivingRoom, Kitchen and Entrance are the areas that can be used both for a short or long duration. Thus, these areas have higher dispersion and noises in irregular short/long-duration events. Therefore, the noises should be inspected by determining both high and low thresholds. Based on the measurement of statistical values and further investigation of the dataset, a 2.5% rule is considered for each place. For this purpose, the 2.5% of events with the lowest duration and the 2.5% of the highest duration events are highlighted for further inspection.

The final applied rules are summarised in Table 2.

Table 2: Determined minimum and maximum threshold for rule-based noise correction.

| Location | Min Threshold | Max Threshold | Applied rule |
|---|---|---|---|
| Bathroom | - | 00:08:29 | Average + 2 ×StdDev |
| Bedroom | 00:00:11 | 05:27:48 | 2.5% rule on both sides |
| Corridor | - | 00:08:29 | Average + 2 ×StdDev |
| Entrance | 00:01:47 | 04:38:10 | 2.5% rule on both sides |
| Kitchen | 00:00:06 | 00:52:10 | 2.5% rule on both sides |
| LivingRoom | 00:00:30 | 00:59:21 | 2.5% rule on both sides |
| WC | - | 00:04:47 | Average + 2 ×StdDev |

### 3.4 The modified PM-based noise correction technique

The proposed hybrid approach for error correction involves two methods. One is the semi-supervised rule-based method, which is designed to detect and correct most of the noisy events. The other is the modified PM-based method, which is used to recover missed transitions and refine the remaining unusual erroneous traces. By addressing most of the noises in the first step, we expect that the traces in the current state mainly represent normal behaviours. Additionally, according to the sensor placements and the identified most problematic areas and error causes, including areas without (blind spots) in the Corridor and LivingRoom, we anticipate there will be several missed transitions in these areas.

Therefore, to improve the preliminary PM-based error correction algorithm derived from [43], we have made adjustments to prevent the incorrect removal of events based solely on their duration. Our approach involves modifying the penalization scores to prioritize the correction of traces with a higher likelihood of adding actions. Moreover, we have explored varying probabilities for deletion action by implementing the algorithm with multiple deletion penalty scores.

To ensure the recovery of missed transitions, the penalization scores for adding Corridor and LivingRoom events have been reduced to give them higher priority. In addition, the Bathroom and WC contact switch sensors are more reliable than the PIR sensors, making captured events less likely to be noisy. As a result, we have assigned deletion penalization scores to these areas in such a way that the algorithm will not delete events from them at all and have set the probability of deletion for Bathroom and WC events to *zero*.

Moreover, the location event log contains events in the Bedroom with long-duration (due to sleeping at night activities) which are followed by short-duration actions in the Kitchen or Bathroom, which represents the sleep interruptions. These particular events can be used to extract valuable features, for instance in behavioural analysis they have great importance. They should not be deleted by the error correction algorithm, hence we decided to also consider an implementation of the algorithm with no deletion option.

We assessed the results of these two implementations of the modified PM-based method to determine how different configurations affect the outcome. Section 4.2 presents a summary of the modified version of PM-based error correction method outcomes.

## 4 Results and Discussion

We evaluated the effectiveness of the two adapted existing methods and the hybrid method, which combines both of them sequentially, for error detection and correction by applying them to the FR-dataset. Our primary objective was to observe how these techniques can detect and fix errors present in the log, which contains data collected by an IoT system in a real-life setting. In addition to the number of errors found, it is also essential to determine the type and source of errors.

This analysis will help us better comprehend the efficiency of the applied techniques in correcting different types of errors.

In this section, we'll discuss the results of using rule-based, modified PM-based error correction techniques and their combination (presented in Section 4.1 and Section 4.2, respectively), by highlighting the types of errors each technique can detect and correct. Additionally, Section 4.3 offers several indicators that can measure the percentage of updates and changes in the actual log. These indicators can help evaluate the performance of error correction techniques on the overall log, and we've used them to compare the impacts of rule-based, modified PM-based, and the combined approach on the data. We've also considered two data quality metrics and analysed the resulting logs from each correction technique to determine which algorithm improves the log's quality the most.

## 4.1 Results of Rule-based error correction

Table 3, lists the results of implementing the rule-based error correction approach, including the number of events before applying the correction method, the number of events after correction, the total number of changes (total altered records, *e.g.* added, removed or fused events) and the percentage of changes in relative to total events in each location. As can be noticed from Table 3, only 3.218% of changes were made to the entire log to correct the traces based on the determined rules and experts' knowledge. This percentage is lower than our predetermined correction target, indicating that the log was only slightly altered after correction.

Additionally, Table 4 provides a summary of the corrections made based on their types and reasons. Based on the results in Table 3 and 4, it appears that the Kitchen and Corridor regions pose the most problems. Specifically, the absence of a sensor in the Corridor results in missed transitions and overly prolonged events. Furthermore, the sensors attached to Kitchen appliances, such as the refrigerator, often detect unnecessary events over the kitchen cabinets when the subject is in the LivingRoom or sitting on the TV-Chair, resulting in anomalous short-duration events. These elements are the primary causes of errors in the log.

Table 3: The outcome of applying the rule-based error correction technique on FR-dataset

| Locations | Number of Events | After Correction Events | Total Corrections (%) |
|---|---|---|---|
| Bathroom | 1655 | 1653 | 10 (0.605%) |
| Bedroom | 1436 | 1369 | 86 (6.282%) |
| Corridor | 2371 | 2375 | 133 (5.600%) |
| Entrance | 242 | 233 | 17 (7.296%) |
| Kitchen | 3279 | 3179 | 158 (4.970%) |
| LivingRoom | 2795 | 2842 | 92 (3.237%) |
| WC | 716 | 716 | 4 (0.559%) |
| Total | 12494 | 12367 | 398 (3.218%) |

Table 4: A summary of the rule-based method's corrections based on the error types events.

| **Error Type** | **Reason** | **Faulty Sensors** | **Correction Action** | **Total** |
|---|---|---|---|---|
| **Missing Data** | Inaccurate Sensor Placement | TV(1), UN(129) | Add missed event | 130 |
| **Noise** | Faulty sensor | C(1), PC(1), S(2), UN(45) | Remove event | 49 |
| | Incorrect events timing and orders | BR(1), E(3), R(2), WC(4), UN(13) | Swap events | 23 |
| | Out of range detection | E(1), F(3), KC(1), M(3), R(42), TC(2), UN(2) | Remove event | 54 |

* The number of captured errors for each error type and by each sensor is mentioned in the parentheses in front of their names.
Abbreviations: TV=TV, C=Closet, PC=PrayingChair, S=Stove, BR=Bathroom, E=Entrance, R=Refrigerator, WC=WC, F=Fridge, KC=KitchenCupboard, M=Microwave, TC=TV Chair, UN=unknown

## 4.2 Results of modified PM-based error correction

In addition to increasing the probability of adding events for correction, in particular for Corridor and LivingRoom, two different configurations with different deletion probabilities have been implemented. The first configuration aims to correct the trace without deleting events from the Bathroom and WC, while the other one solely focuses on adding events without any deletions. The penalization score for delete and add correcting actions for each area of the house is adjusted interactively based on an understanding of the scenario and house floorplan, in order to achieve best configurations for the modified PM-based correction method in this part.

These configurations have been tested on both the original location event log and the rule-based corrected location log (the outcome is the result for the hybrid approach), and the results are presented in Table 5. The added records in fact represent events that the algorithm detected as missed events and added to retrieve the actual events. On the other hand, the delete operation removes events identified as noise. However, as previously mentioned, the PM-based method is sub-optimal, meaning it cannot guarantee the correction of all errors in the best possible way. It tries to find the best option for correcting a trace within a pre-defined time limit.

Taking into account that 34% of the events in the log contained invalid transitions between areas and that some corrections required modifying multiple records, the number of corrections made using the modified PM-based method and the hybrid method is almost reasonable. Moreover, as previously mentioned, inadequate sensors in the LivingRoom and Corridor and their undiscoverable areas could lead to a high number of errors. The higher number of added events in these areas verify our primary claim regarding the missed events in these places. The modified PM-based method with the deletion option corrected the traces with fewer records and corrections, while the algorithm had to add more events to correct the traces without deleting any records. In addition, the "fuse" operation is not needed, if records were not deleted. Also, the modified PM-based method made fewer corrections to the rule-corrected logs because some errors were already addressed in previous steps.

Table 5: A summary of the corrections for different configurations of the modified PM-based method.

| Log and PM-Configurations | Add | Delete | Fuse | Total Corrections |
|---|---|---|---|---|
| **PM1* on ORL** | BR(41), C(2510), E(27), K(121), LR(1545), WC(3) | BR(119), C(50), E(1), K(558), LR(18) | BR(188), C(17), K(57), LR(10) | 5265 |
| **PM2** on ORL** | BR(7), C(3414), E(263), K(25), LR(2965) | - | - | 6674 |
| **Hybrid1 (PM1 on RCL)** | B(1), BR(36), C(2520), E(32), K(91), LR(1485), WC(3) | BR(115), C(27), E(1), K(516), LR(14) | BR(156), C(9), K(59), LR(8) | 5073 |
| **Hybrid2 (PM2 on RCL)** | BR(6), C(3272), E(172), K(28), LR(2653) | - | - | 6131 |

The number of corrections in each place is mentioned in parentheses.
* PM1= Modified PM-based correction with a higher probability for adding without deleting Bathroom and WC events.
** PM2= Modified PM-based correction with a higher probability for adding without deleting any events.
Abbreviations: ORL= Original Log, RCL= Rule-based Corrected Log.
B=Bathroom, BR=Bedroom, C=Corridor, E=Entrance, K=Kitchen, LR=LivingRoom

## 4.3 Log change metrics and quality indicators

Logs primarily record factual events and errors are uncommon and often caused by environmental factors. Therefore, correction algorithms should aim to preserve the log's meaning without significant alteration. To measure the impact of error correction techniques, the rule-based and modified PM-based methods adjusted to our use case and the proposed interactive hybrid approach, specific metrics are necessary to indicate the extent of changes made. This research outlines the following metrics to clarify the scale of changes resulting from correction algorithms.

- **Number of changes.** The total number of event records that changed in the corrected event log (added, deleted, fused or altered) during error correction.

- **Duration of changes.** The time duration of changed event records that changed in the corrected event log are summed to obtain the total amount of changes applied to the log.

The log change metrics are measured over the entire log and daily with an average value and a maximum value, representing the worst-case situation. In addition, we report the fraction of changes (%) to show the scale of changes relative to the total number of events. Analysing the daily results can help identify problematic days that may need to be removed from the log or treated as outliers. In this way, data with higher quality assurance, for instance, only days with less than a specified threshold of correction, which means less erroneous events were observed within their traces, can be used for analysis purposes by mining techniques, which are sensitive to data variability.

Furthermore, the quality of the resulting logs after corrections will be assessed using the following metrics derived from [25, 70]:

- **Coefficient of Network Complexity (CNC).** The ratio of arcs to nodes in a model, is defined as

$$CNC = \frac{A^2}{N} \quad (A: Nubmer\,of\,arcs,\ N: Number\,of\,nodes) \tag{2}$$

- **Erroneous transition probability (ETP).** The percentage of the transitions that are erroneous. This indicator is obtained by dividing the total number of erroneous transitions by the total number of transitions.

Process models help to understand the underlying process by visualising it and by revealing the connections between the nodes. The complexity measure of a model represents the easiness of visual understanding of that model since understandable models are preferable for analysis. In this way, the CNC measure is an elegant and simple-to-understand index, which reflects the connectedness and the understandability of the discovered model. A higher CNC shows that the model has a higher number of arcs and there are more transitions between nodes, and probably more cycles in the model which increases the complexity (the Spaghetti effect as introduced in Section 2.4). It is worth noting that PM is not the only way to interpret and mine a log. However, in our case study, we used it as a solution for analysis and applied it uniformly to the events log before and after correction to compare the understandability parameter.

Moreover, the ETP indicator shows the number of errors (the transitions which are not valid in the reference model) compared to the total number of transitions. However, the reliability of ETP can be affected if there are many transitions between two adjacent locations (e.g. LivingRoom and Kitchen). Since we are using a single setup and dataset, this issue can affect all the versions of logs and corrections uniformly. Therefore, comparing ETP values together will be useful to assess the quality of logs regarding the fraction of faulty transitions.

## 4.4 Evaluation of error correction techniques

Finally, to evaluate the adapted error correction techniques and the proposed interactive hybrid approach which executes both methods back-to-back, we have measured the log changes metrics and log quality metrics. The results for the log change metrics of the proposed error correction techniques are reported in Table 6, and the quality indicators are reported in Table 7.

Based on the evaluation results, it is evident that the semi-supervised rule-based algorithm is more effective in managing noises. This is because experts possess a better comprehension of typical behaviours, which significantly reduces the probability of inaccurate corrections. The number of small changes in the log made by the rule-based method and the fewer changes in the log made by the hybrid method relative to the modified PM-based corrections on the original log acknowledges this argument. On the other hand, more than the efficiency and quality which are essential, identifying missing events and rectifying them through the addition of relevant subsequent actions can be a laborious and tedious task, which can be automated by applying a PM-based correction method in comparison to doing this job with the help of experts for instance in a fully rule-based approach.

Through the use of process models extracted from the log and cross-checking with a reference model, PM-based methods can effectively detect invalid transitions and insert necessary actions into the log. These added actions are of sufficient length to cover transition times but not so long that they significantly alter the log's timings. Furthermore, PM-based methods can also eliminate noisy events. However, distinguishing between an actual behaviour and a noise can be difficult, potentially leading to the inadvertent deletion of significant semantics from the log. Additionally,

Table 6: A summary of log change metrics for correction methods

| | Number of Changes | | | Duration of Changes | | |
|---|---|---|---|---|---|---|
| | Daily | | Total | Daily | | Total |
| **Error Corrections** | **Average** | **Max** | | **Average** | **Max** | |
| **Rule-based on ORL** | 2 | 9 | 256 (2.07%) | 00:02:18 (0.160%) | 00:27:35 (1.916%) | 5:35:41 (0.160%) |
| **PM1* on ORL** | 35.82 | 79 | 5265 (33.6%) | 00:32:36 (2.26%) | 02:08:01 (8.89%) | 79:52:30 (2.28%) |
| **PM2** on ORL** | 45.4 | 152 | 6674 (35.27%) | 00:00:08 (0.01%) | 00:00:18 (0.02%) | 00:17:56 (0.01%) |
| **Hybrid1 (PM1 on RCL)** | 34.51 | 70 | 5073 (33.22%) | 00:33:15 (2.31%) | 02:08:01 (8.89%) | 83:10:47 (2.32%) |
| **Hybrid2 (PM2 on RCL)** | 41.71 | 74 | 6131 (33.16%) | 00:00:07 (0.01%) | 00:00:11 (0.01%) | 00:16:03 (0.01%) |

* PM1= Modified PM-based correction with a higher probability for adding without deleting Bathroom and WC events.
** PM2= Modified PM-based correction with a higher probability for adding without deleting any events.
Abbreviations: ORL= Original Log, RCL= Rule-based Corrected Log.

Table 7: Quality indicators for the outcomes of error correction methods

| Error Corrections | CNC | ETP |
|---|---|---|
| **Original Log** | 740.57 | 34.17% |
| **Rule-based** | 622.29 | 32.76% |
| **PM1* on ORL** | 416.51 | 0.66% |
| **PM2** on ORL** | 357.14 | 0.63% |
| **Hybrid1** | 531.57 | 0.65% |
| **Hybrid2** | 357.14 | 0.63% |

* PM1= Modified PM-based correction with a higher probability for adding without deleting Bathroom and WC events.
** PM2= Modified PM-based correction with a higher probability for adding without deleting any events.
Abbreviations: ORL= Original Log, RCL= Rule-based Corrected Log.

the duration of deleted events is crucial, as deleting more events can increase the percentage of changes in the log.

The modified PM-based method is designed to effectively manage deletion actions while prioritizing the addition of relevant actions for correction due to our knowledge about the existence of blind spots within the under-control area. The success of the modified PM-based method can be seen through the decreased ETP value (*i.e.* the percentage of remaining invalid transitions) and the total number of corrections. Both PM1 and PM2 implementations were effective in improving the quality of the event log by reducing errors while avoiding the deletion of specific events. PM1 avoided deleting Bathroom and WC events, while PM2 avoided deleting any events.

Moreover, in terms of the fraction of changes, it should be noted that although modified PM-based implementations approximately altered one-third of the total events, however, the changes made are crucial in recovering missed transitions between main areas of the house that were not adequately covered by sensors. As a result, these changes play a vital role in correcting the log and ultimately contribute to the creation of more comprehensible models. Additionally, the duration of changes is more significant than the quantity of changes. Our results indicate that only 2.5% of the logs were modified using the modified PM-based approach, which is negligible and falls below our predetermined target for error corrections in this study.

From the quality perspective, the correction methods have successfully increased the log quality, resulting in a more accurate log, which can be used in analyses, for instance, to discover more comprehensible models. In particular, when both correction methods are utilised in a hybrid approach, the improvements are even more significant and noticeable according to the log quality indicators.

## 4.5 Discussion

When analysing data obtained from IoT systems, it is crucial to ensure the reliability and applicability of the analytical outcomes derived from the generated data. In this manuscript, we address this challenge by focusing on the problem of error correction. Errors in the sensor and event log can have a negative impact on the analysis and lead to unclear and confusing results. Therefore, an accurate and reliable data log is of utmost importance for any mining technique or algorithm. To tackle this issue, error correction methods must be employed to identify the root cause of errors and rectify any erroneous events. However, it is imperative to approach corrections with caution as they can also affect subsequent analysis. There are several essential aspects related to the efficient correction of errors to utilise IoT technologies for data collection and complex analysis that remain to be considered, and we will discuss three of them in the following.

- **Usability of data with improved quality.**

  From the usability perspective and the expected quality of a log for analysis, when data has fewer errors and is less complex, it can be used in a wider range of mining algorithms, including more complex techniques, and provide greater knowledge and insight. A clear example of this is demonstrated by comparing two models discovered by a single PM discovery technique (PALIA) from the original and corrected logs, as shown in Figure 6. The nodes in the graphs represent locations while the arrows indicate transitions between them. A heat map varying from green (low) to red (high) is also applied to the graph to represent the total events duration of each location and the number of times the transitions were executed.

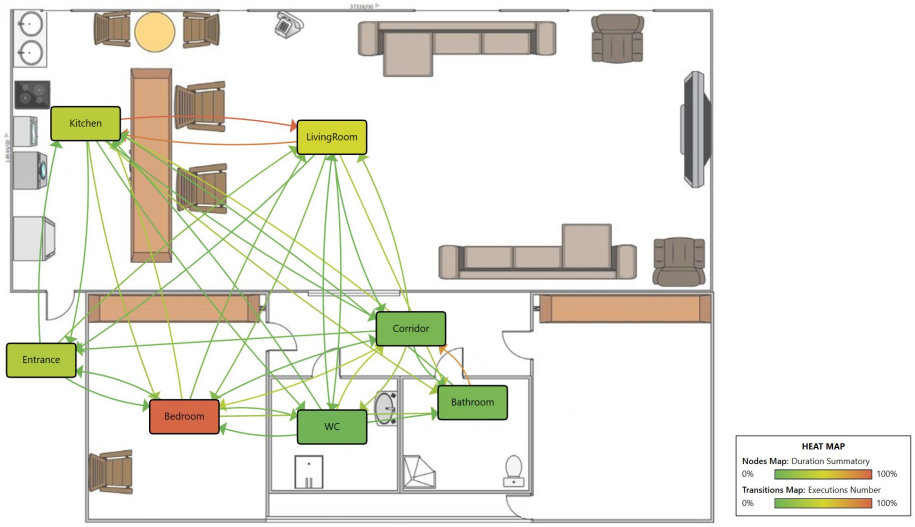
  A mining technique with the purpose of interpreting the model's behaviour will identify different patterns that represent various activities and daily routines with all of their possible variabilities. Assuming a model like (6a), the additional transitions (or actions) due to the capturing of noisy events caused by unwanted external factors, like environmental conditions or the improper design of the data collection systems, increase the complexity of the discovered model. However, the correction algorithm removed a part of these additional and impossible actions in the corrected model (6b). This elimination of unnecessary complexity reduces the load on the mining algorithms, making the resulting model much simpler and easier to comprehend. Improved quality data and simplified models can be used in mining algorithms to identify intricate patterns. Two studies published in [49] and [50], both conducted on corrected event logs, provide examples of how access to logs with improved quality can led to proper outcomes in data analysis. In the absence of efficient error correction techniques, undesired variations in the original data can impose high and improper complexity on the data mining algorithms so that they achieve inaccurate results and models.

- **Selection and development of an appropriate error correcting method.**

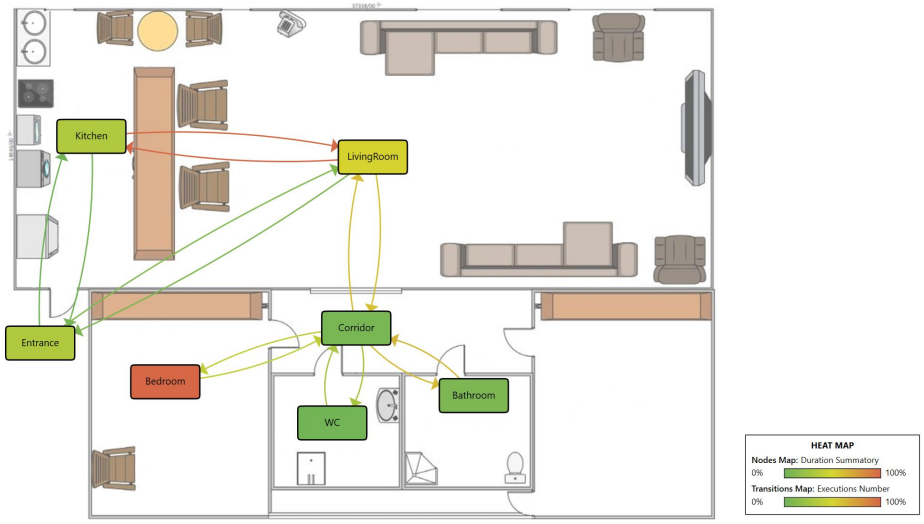
  The next important point is the selection of an error correction method to address the error issues and guarantee the accuracy for further analysis. A precise investigation and selection of error correction algorithms enables effective reviewing and fixing of any errors. It is important to note that no single error correction method can detect every error, as errors can result from various configurations and intrinsic factors. Without a thorough understanding of the actual behaviour, it becomes challenging to accurately identify and comment on all potential errors that may arise during data collection. Additionally, each error correction method comes with its own set of advantages and disadvantages and is primarily designed to address specific types of errors.

  Given these complexities, it is essential to start by carefully examining the event log chosen for the study (or a sample of it) to identify possible error subtypes. Tailoring error correction algorithms based on a comprehensive understanding of the error types and their underlying causes is crucial for ensuring accuracy. For instance, in the selected scenario used as a case study to demonstrate the feasibility of our proposed interactive hybrid approach, ambient binary sensors were employed. We categorised the existing errors in the event log into two subtypes: missed and erroneous events.

  The process of developing an appropriate error correction method involves a multi perspective approach that necessitates the involvement of domain experts in an interactive manner. Due to the diverse sources and reasons for errors, it may be necessary to select and combine different approaches. In the smart home case study presented in this manuscript, we opted for a semi-supervised and heuristic approach to address noises. This decision was driven by concerns about incorrect corrections and the potential for removing valuable patterns in the




(a)




(b)

Figure 6: The extracted TPA models from (a) The original location event log, (b) The corrected log resulted from applying the hybrid1 method.

data. Then, to tackle the high number of missed events disrupting the order of events in the event log's traces, we employed an automated approach based on Process Mining. An existing PM-based approach was modified to suit our specific use case and used for the final round of corrections. Since PM techniques are capable in checking the conformance of a model with a reference model, they proved to be the most suitable options for inspecting and correcting event sequences in a trace.

In the end, evaluating the effectiveness of applied correction methods in improving log quality can be facilitated by quality and log change indicators. Multiple correction methods can be employed in combination to address errors with varying characteristics, forming hybrid approaches that collaboratively tackle different types of errors. Each approach complements the others, leading to enhanced overall performance. Moreover, integrating domain experts in interactive approaches can lead to better results, as these approaches can be modified as needed to meet data quality requirements. However, it is worth noting that directly comparing and evaluating error correction methods from the literature can be challenging due to the non-uniform evaluation process and the prevalent use of non-publicly available datasets. Therefore, comparing different methods on a single dataset (like the conducted evaluations in this study) can offer valuable insights into the strengths and weaknesses of each method in addressing specific error types, thereby informing necessary modifications to achieve better outcomes.

- **Generalisability and scalability.**

  The generalisability of an error correction method is a crucial consideration when evaluating its effectiveness and applicability across different scenarios and settings. While it is impractical to create a one-size-fits-all error correction technique capable of addressing all types of errors with varying characteristics and causes, a flexible and interactive customizable approach can offer significant advantages. We also followed this direction in our proposed error correction method, which outperforms in terms of generalisability compared to traditional rule-based or control-flow-based approaches.

  In our case study, the rules utilised for the rule-based heuristic part were tailored based on specific noise features to address events of varying durations (too short or too long duration events depending on the areas). This custom-made approach enables the method to be easily adapted to different scenarios by identifying the specific errors that require higher precision and defining appropriate rules with the assistance of domain experts. Subsequently, the automated Process Mining (PM)-based approach, which requires a reference model and multiple scores for correction, can be seamlessly integrated. Even when dealing with larger models in different settings, such as smart buildings with a higher number of rooms or new application scenarios, the algorithm can function effectively, when a reference model is available and specific penalisation scores for corrections are determined. It is also possible and more convenient to use a PM discovery technique to automatically obtain a model from the event log and correct it with the help of experts. In this way, they can also observe the most problematic events or transitions and adapt the scores based on their desired correction in an interactive way.

  In terms of scalability aspect, the scalability of the error correction method is influenced by several key factors, including the size of the dataset, the number of activities and transitions. To manage the computational load and enhance efficiency, our rule-based method incorporates techniques based on statistical parameters or upper and lower boundaries to determine thresholds for highlighting potential errors. This targeted approach ensures that only a fraction of erroneous events require manual inspection by domain experts. In a worst-case situation with a large number of errors, only a percentage of changes in the log will be determined for thorough inspection by experts based on the desired quality and cost considerations.

  Moreover, the automated PM-based approach significantly reduces the need for manual intervention, thereby minimising the time and effort expended by experts. Despite potential increases in computation time due to the number of required corrections, the independence of different traces allows for parallel processing, further enhancing efficiency. Additionally, system designers can balance quality and time spent on error correction by setting limits on the maximum number of corrections or the time budget for each trace.

  To sum up, a method that enables interactive customization and incorporates automated processes can be easily adapted to diverse scenarios and scaled up to handle large datasets

with numerous activities. This flexibility and efficiency make the method a valuable tool for enhancing the reliability and accuracy of data analysis in IoT systems. The proposed method, which utilizes both rule-based and PM-based corrections, provides evidence to support this assertion.

# 5 Conclusions

The quality of data collected by IoT devices can be evaluated by using error detection algorithms, and if data quality is found to be low, error correction algorithms can be employed to rectify the errors. This article highlights that by identifying the main reasons for capturing errors during the data collection, the error correction methods can be adapted in a way to address the errors better. A rule-based method with specifically determined rules to address the IoT system limitations is used to correct noises with higher accuracy. In addition, a modified Process mining-based method is employed to detect any deviations in the sequence of events within traces. This approach compares the model discovered from the data with a reference model to correct the missing event errors. Furthermore, due to the different types of errors caused by varying reasons, a combination of both methods is used to integrate the benefits of them together, and it optimises the error correction method performance.

# References


[1] H. C. Pöhls, V. Angelakis, S. Suppan, K. Fischer, G. Oikonomou, E. Z. Tragos, R. D. Rodriguez, and T. Mouroutis, "RERUM: Building a reliable IoT upon privacy-and security-enabled smart objects," in *Wireless Communications and Networking Conference Workshops (WCNCW'14)*, pp. 122–127, IEEE, apr 2014.

[2] Y. Zhao, H. Haddadi, S. Skillman, S. Enshaeifar, and P. Barnaghi, "Privacy-preserving activity and health monitoring on databox," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pp. 49–54, 2020.

[3] N. Apthorpe, D. Reisman, and N. Feamster, "A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic," *arXiv preprint arXiv:1705.06805*, 2017.

[4] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016.

[5] T. Chun-Wei, L. Chin-Feng, and C. Ming-Chao, "Yang laurence t.. 2013," *Data mining for Internet of Things: A survey. IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 77–97, 2013.

[6] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer networks*, vol. 76, pp. 146–164, 2015.

[7] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: A systematic review," *Journal of Big Data*, vol. 7, no. 1, pp. 1–49, 2020.

[8] Y. Bertrand, R. Van Belle, J. De Weerdt, and E. Serral Asensio, "Data quality patterns in process mining with iot data," *Lecture Notes in Business Information Processing*, 2022.

[9] J. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks. 26th ieee int," in *Conf. Distrib. Comput. Syst. ICDCS06*, pp. 51–51, 2009.

[10] D. Zeng, S. Guo, and Z. Cheng, "The web of things: A survey," *J. Commun.*, vol. 6, no. 6, pp. 424–438, 2011.

[11] Y. Li and L. E. Parker, "Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks," *Information Fusion*, vol. 15, pp. 64–79, 2014.

[12] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

[13] L. Berti-Equille, "Measuring and modelling data quality for quality-awareness in data mining," *Quality measures in data mining*, pp. 101–126, 2007.

[14] D. Hand, H. Mannila, and P. Smyth, "Principles of data mining. 2001," *MIT Press. Sections*, vol. 6, no. 3, pp. 2–6, 2001.

[15] N. Martin, *Data Quality in Process Mining*, pp. 53–79. Cham: Springer International Publishing, 2021.

[16] R. Andrews, S. Suriadi, C. Ouyang, and E. Poppe, "Towards Event Log Querying for Data Quality," in *On the Move to Meaningful Internet Systems. OTM 2018 Conferences*, pp. 116–134, Springer International Publishing, 2018.

[17] A. Whitmore, A. Agarwal, and L. Da Xu, "The internet of things—a survey of topics and trends," *Information systems frontiers*, vol. 17, pp. 261–274, 2015.

[18] J. J. Lull, J. L. Bayo, M. Shirali, M. Ghassemian, and C. Fernandez-Llatas, *Interactive Process Mining in IoT and Human Behaviour Modelling*, pp. 217–231. Cham: Springer International Publishing, 2021.

[19] O. Dogan, A. Martinez-Millana, E. Rojas, M. Sepúlveda, J. Munoz-Gama, V. Traver, and C. Fernandez-Llatas, "Individual behavior modeling with sensors using process mining," *Electronics*, vol. 8, no. 7, p. 766, 2019.

[20] B. Guo, D. Zhang, Z. Wang, Z. Yu, and X. Zhou, "Opportunistic iot: Exploring the harmonious interaction between human and the internet of things," *Journal of Network and Computer Applications*, vol. 36, no. 6, pp. 1531–1539, 2013.

[21] S. Enshaeifar, P. Barnaghi, S. Skillman, A. Markides, T. Elsaleh, S. T. Acton, R. Nilforooshan, and H. Rostill, "The internet of things for dementia care," *IEEE Internet Computing*, vol. 22, no. 1, pp. 8–17, 2018.

[22] K. Curran, E. Furey, T. Lunney, J. Santos, D. Woods, and A. McCaughey, "An evaluation of indoor location determination technologies," *Journal of Location Based Services*, vol. 5, no. 2, pp. 61–78, 2011.

[23] N. Li and B. Becerik-Gerber, "Performance-based evaluation of rfid-based indoor location sensing solutions for the built environment," *Advanced Engineering Informatics*, vol. 25, no. 3, pp. 535–546, 2011.

[24] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[25] M. Shirali, J.-L. Bayo-Monton, C. Fernandez-Llatas, M. Ghassemian, and V. Traver Salcedo, "Design and evaluation of a solo-resident smart home testbed for mobility pattern monitoring and behavioural assessment," *Sensors*, vol. 20, no. 24, 2020.

[26] J. A. Álvarez-García, P. Barsocchi, S. Chessa, and D. Salvi, "Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 evaal competition," *Journal of Ambient Intelligence and Smart Environments*, vol. 5, no. 1, pp. 119–132, 2013.

[27] T.-H. Tan, M. Gochoo, K.-H. Chen, F.-R. Jean, Y.-F. Chen, F.-J. Shih, and C. F. Ho, "Indoor activity monitoring system for elderly using rfid and fitbit flex wristband," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 41–44, IEEE, 2014.

[28] C. Fernández-Llatas, J.-M. Benedi, J. M. García-Gómez, and V. Traver, "Process mining for individualized behavior modeling using wireless tracking in nursing homes," *Sensors*, vol. 13, no. 11, pp. 15434–15451, 2013.

[29] D. J. Cook and N. Krishnan, "Mining the home environment," *Journal of intelligent information systems*, vol. 43, pp. 503–519, 2014.

[30] S. Palipana, B. Pietropaoli, and D. Pesch, "Recent advances in rf-based passive device-free localisation for indoor applications," *Ad Hoc Networks*, vol. 64, pp. 80–98, 2017.

[31] G. Chen, A. Wang, S. Zhao, L. Liu, and C.-Y. Chang, "Latent feature learning for activity recognition using simple sensors in smart homes," *Multimedia Tools and Applications*, vol. 77, pp. 15201–15219, 2018.

[32] J. Tewell, D. O'Sullivan, N. Maiden, J. Lockerbie, and S. Stumpf, "Monitoring meaningful activities using small low-cost devices in a smart home," *Personal and Ubiquitous Computing*, vol. 23, pp. 339–357, 2019.

[33] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive and mobile computing*, vol. 10, pp. 138–154, 2014.

[34] D. J. Cook, G. Duncan, G. Sprint, and R. L. Fritz, "Using smart city technology to make healthcare smarter," *Proceedings of the IEEE*, vol. 106, no. 4, pp. 708–722, 2018.

[35] F. Viani, F. Robol, A. Polo, P. Rocca, G. Oliveri, and A. Massa, "Wireless architectures for heterogeneous sensing in smart home applications: Concepts and real implementation," *Proceedings of the IEEE*, vol. 101, no. 11, pp. 2381–2396, 2013.

[36] N. Chang, R. Rashidzadeh, and M. Ahmadi, "Robust indoor positioning using differential wi-fi access points," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1860–1867, 2010.

[37] M. Milenkovic and O. Amft, "Recognizing energy-related activities using sensors commonly installed in office buildings," *Procedia Computer Science*, vol. 19, pp. 669–677, 2013.

[38] H. Morsali, S. M. Shekarabi, K. Ardekani, H. Khayami, A. Fereidunian, M. Ghassemian, and H. Lesani, "Smart plugs for building energy management systems," in *Iranian Conference on Smart Grids*, pp. 1–5, IEEE, 2012.

[39] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "Casas: A smart home in a box," *Computer*, vol. 46, no. 7, pp. 62–69, 2012.

[40] A. A. Aramendi, A. Weakley, A. A. Goenaga, M. Schmitter-Edgecombe, and D. J. Cook, "Automatic assessment of functional health decline in older adults based on smart home data," *Journal of biomedical informatics*, vol. 81, pp. 119–130, 2018.

[41] R. J. C. Bose, R. S. Mans, and W. M. Van Der Aalst, "Wanna improve process mining results?," in *2013 IEEE symposium on computational intelligence and data mining (CIDM)*, pp. 127–134, IEEE, 2013.

[42] L. Vanbrabant, N. Martin, K. Ramaekers, and K. Braekers, "Quality of input data in emergency department simulations: Framework and assessment techniques," *Simulation Modelling Practice and Theory*, vol. 91, pp. 83–101, 2019.

[43] C. Fernandez-Llatas, J. M. Benedi, J. M. Gama, M. Sepulveda, E. Rojas, S. Vera, and V. Traver, *Interactive Process Mining in Surgery with Real Time Location Systems: Interactive Trace Correction*, pp. 181–202. Cham: Springer International Publishing, 2021.

[44] C. Fernandez-Llatas, J. Munoz-Gama, N. Martin, O. Johnson, M. Sepulveda, and E. Helm, *Process Mining in Healthcare*, pp. 41–52. Cham: Springer International Publishing, 2021.

[45] W. Van Der Aalst, *Process Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[46] C. Fernandez-Llatas, A. Martinez-Millana, A. Martinez-Romero, J. M. Benedí, and V. Traver, "Diabetes care related process modelling using process mining techniques. lessons learned in the application of interactive pattern recognition: coping with the spaghetti effect," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2127–2130, 2015.

[47] G. Polančič and B. Cegnar, "Complexity metrics for process models–a systematic literature review," *Computer Standards & Interfaces*, vol. 51, pp. 104–117, 2017.

[48] C. Fernandez-Llatas, *Interactive Process Mining Challenges*, pp. 295–304. Cham: Springer International Publishing, 2021.

[49] G. Di Federico, C. Fernández-Llatas, Z. Ahmadi, M. Shirali, and A. Burattin, "Identifying Variation in Personal Daily Routine Through Process Mining: A Case Study," in *6th International Workshop on Process-Oriented Data Science for Healthcare (PODS4H23), International Conference on Process Mining (ICPM 2023)*, 2023.

[50] M. Shirali, Z. Ahmadi, C. Fernández-Llatas, and J.-L. Bayo-Monton, "Synergy of information in multimodal iot systems – discovering the impact of daily behaviour routines on physical activity level," 2024.

[51] G. Di Federico and A. Burattin, "Do you behave always the same? a process mining approach," in *International Conference on Process Mining*, pp. 5–17, Springer, 2022.

[52] M. L. Van Eck, X. Lu, S. J. Leemans, and W. M. Van Der Aalst, "Pm: a process mining project methodology," in *International conference on advanced information systems engineering*, pp. 297–313, Springer, 2015.

[53] N. Martin, A. Martinez-Millana, B. Valdivieso, and C. Fernández-Llatas, "Interactive data cleaning for process mining: a case study of an outpatient clinic's appointment system," in *Business Process Management Workshops: BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers 17*, pp. 532–544, Springer, 2019.

[54] J. Yang, L. Lin, Z. Sun, Y. Chen, and S. Jiang, "Data validation of multifunctional sensors using independent and related variables," *Sensors and Actuators A: Physical*, vol. 263, pp. 76–90, 2017.

[55] Z. Yu, A. Bedig, F. Montalto, and M. Quigley, "Automated detection of unusual soil moisture probe response patterns with association rule learning," *Environmental Modelling & Software*, vol. 105, pp. 257–269, 2018.

[56] O. Bamgboye, X. Liu, and P. Cruickshank, "Towards modelling and reasoning about uncertain data of sensor measurements for decision support in smart spaces," in *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, vol. 2, pp. 744–749, IEEE, 2018.

[57] O. Dogan, J.-L. Bayo-Monton, C. Fernandez-Llatas, and B. Oztaysi, "Analyzing of gender behaviors from paths using process mining: A shopping mall application," *Sensors*, vol. 19, no. 3, p. 557, 2019.

[58] Y. Zhang, O. Martikainen, R. Saikkonen, and E. Soisalon-Soininen, "Extracting service process models from location data," in *Data-Driven Process Discovery and Analysis: 6th IFIP WG 2.6 International Symposium, SIMPDA 2016, Graz, Austria, December 15-16, 2016, Revised Selected Papers 6*, pp. 78–96, Springer, 2018.

[59] W. Kratsch, F. König, and M. Röglinger, "Shedding light on blind spots–developing a reference architecture to leverage video data for process mining," *Decision Support Systems*, vol. 158, p. 113794, 2022.

[60] J. Maeyens, A. Vorstermans, and M. Verbeke, "Process mining on machine event logs for profiling abnormal behaviour and root cause analysis," *Annals of Telecommunications*, vol. 75, no. 9, pp. 563–572, 2020.

[61] R. Andrews, M. T. Wynn, K. Vallmuur, A. H. Ter Hofstede, E. Bosley, M. Elcock, and S. Rashford, "Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in queensland," *International journal of environmental research and public health*, vol. 16, no. 7, p. 1138, 2019.

[62] C. Fernández-Llatas, T. Meneu, V. Traver, and J.-M. Benedi, "Applying evidence-based medicine in telehealth: an interactive pattern recognition approximation," *International journal of environmental research and public health*, vol. 10, no. 11, pp. 5671–5682, 2013.

[63] A. Rogge-Solti, R. S. Mans, W. M. van der Aalst, and M. Weske, "Repairing event logs using timed process models," in *On the Move to Meaningful Internet Systems*, pp. 705–708, Springer, 2013.

[64] G. Chiloiro, M. Savino, A. Romano, S. Di Franco, R. Gatta, N. D. Capocchiano, C. Fernandez-Llatas, V. Valentini, A. Damiani, and M. A. Gambacorta, "Impact of adherence to esmo guidelines on survival outcomes in rectal cancer: A process mining analysis of real-world patient data," *Available at SSRN 4713049*.

[65] K. Tsang and W. L. Chan, "Data validation of intelligent sensor using predictive filters and fuzzy logic," *Sensors and Actuators A: Physical*, vol. 159, no. 2, pp. 149–156, 2010.

[66] M. Falah Rad, M. Shakeri, K. Khoshhal Roudposhti, and I. Shakerinia, "Probabilistic elderly person's mood analysis based on its activities of daily living using smart facilities," *Pattern Analysis and Applications*, pp. 1–14, 2022.

[67] C. R. Wren and E. M. Tapia, "Toward scalable activity recognition for sensor networks," in *Location-and Context-Awareness: Second International Workshop, LoCA 2006, Dublin, Ireland, May 10-11, 2006. Proceedings 2*, pp. 168–185, Springer, 2006.

[68] D. Cook, M. Schmitter-Edgecombe, A. Crandall, C. Sanders, and B. Thomas, "Collecting and disseminating smart home sensor data in the casas project," in *Proceedings of the CHI workshop on developing shared home behavior datasets to advance HCI and ubiquitous computing research*, pp. 1–7, 2009.

[69] G. Di Federico, A. Burattin, and M. Montali, "Human behavior as a process model: Which language to use?," in *ITBPM@ BPM*, pp. 18–25, 2021.

[70] A. M. Latva-Koivisto, "Finding a complexity measure for business process models," 2001.