Received 19 September 2022, accepted 25 October 2022, date of publication 1 November 2022, date of current version 9 November 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3218785*

RESEARCH ARTICLE

# Long Gaps Missing IoT Sensors Time Series Data Imputation: A Bayesian Gaussian Approach

**HASSAN M. AHMED[1], BESSAM ABDULRAZAK[1], (Member, IEEE), F. GUILLAUME BLANCHET[2,3,4], HAMDI ALOULOU[5], AND MOUNIR MOKHTARI[6]**

[1]Ambient Intelligence Laboratory (AMI-Lab), Département d'informatique, Faculté des sciences, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[2]Département de Biologie, Faculté des sciences, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[3]Département de Mathématiques, Faculté des sciences, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[4]Département des sciences de la santé communautaire, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[5]ReDCAD, ENIS, University of Sfax, Sfax 3029, Tunisia
[6]Institut Mines Télécom, 91120 Paris, France

Corresponding author: Hassan M. Ahmed (hassan.mostafa.ahmed.fahmy@usherbrooke.ca)

**ABSTRACT** Missing sensor data is a common problem associated with Internet of Things ecosystems, which affects the accuracy of associated services such as adequate medical intervention for older adults living at home. This problem is caused by many factors, power down is one of them, communication failure and sensor failure are another two reasons. Multiple missing data imputation methods have been developed to address this issue. However, irregular temporal missing data locations are challenging to handle, due to lack of knowledge of their occurrence probability and their random temporal location. In this paper, we propose a Bayesian Gaussian Process based imputation technique that accounts for temporal forcing to fill in missing sensor data. Our approach; Bayesian Gaussian Process (BGaP); can efficiently impute missing data at any missing rate and for any temporal location using prior knowledge gathered from past observations. We illustrated how our approach performs using real data collected from sensors deployed in the residence of 10 older adults over a two-year period. Using our novel approach, we were able to impute all missing data which allowed us to observe long-term behavior changes that we would not have been able to observe otherwise.

**INDEX TERMS** Missing sensor data, missing data imputation, Bayesian Gaussian process, long-term older adults behavior monitoring.

## I. INTRODUCTION

The Internet of Things (IoT) now makes it possible to deploy a large number of sensory nodes in diverse environments, for example, to monitor older adults' behavior. In retirement homes, sensory data helps in the decision-making process regarding the status of older adults.

In retirement homes, long-term behavior monitoring, and detection changes are important so that physical and cognitive decline can be captured early and, when properly managed, can increase the well-being of older adults for longer periods. In addition, long-term behavioral monitoring provides valuable information for eventual medical intervention. However, long-term behavior monitoring, and change detection need to be carried out continuously to obtain the best results; a constraint that is fulfilled by IoT ecosystems.

Degradation of quality of life for older adults is a consequence of cognitive and physical decline, which, when detected early enough, can result in better intervention and adaptation of medical care [1]. However, monitoring daily changes in older adults' behavior is challenging in practice [2] because it requires medical staff (e.g., nurses) to be constantly available for every patient. Even if this is achievable within small retirement homes with a crew that cycles day and night, it is a difficult, not to say impossible, task for a medium or large retirement home. One way to evaluate changes in the behavior of patients is to monitor the detailed behavior of every patient over time. Typically, medical staff only record a

The associate editor coordinating the review of this manuscript and approving it for publication was Razi Iqbal.

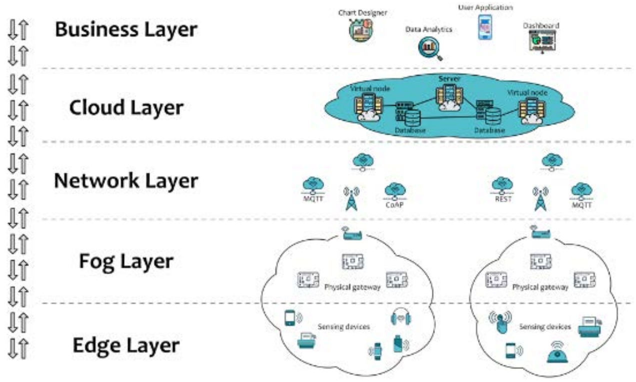

**FIGURE 1.** IoT models architecture, (a) Five-layer architecture IoT model [6].

broad description of a patient's behavior. For example, a nurse would record that a patient goes to the bathroom at night but would not record the time, duration, or frequency of visits. Such information is valuable for assessing patients' health status. Frequent short bathroom visits could indicate urinary tract infection, whereas frequent long bathroom visits could indicate diarrhea [1]. In short, although broad descriptions to assess a patient's physical abilities are valuable, they are incomplete. In this respect, IoT technology can be used to monitor ambient environments unobtrusively and continuously using sensors and sensor nodes [3]. Hence, patient movements can be assessed using IoT sensors and translated into meaningful behavioral data.

IoT ecosystems consist of sensors and actuators that are used to harvest physical data from the environment [4], such as the ambient temperature and/or pressure. Although the three-layer model is considered the basic IoT model, in this study, we assume that a five-layer architecture model is used to realize IoT ecosystems [5], [6], as presented in Figure 1. A three-layer model is composed of a perception layer (the sensors) that resembles the edge layer in the five-layer model, a communication (or transmission) layer that resembles both the fog and network layer in the five-layer model, and an application layer that resembles the cloud and business layer in the five-layer model. The role of the perception layer is to collect data from the environment, while that of the transmission layer is to securely transmit the raw data collected by the sensors. The application layer stores and retrieves collected pre- or post-processed data to and from databases, while also providing special services to be performed on the data, including missing data recovery, anomalous data detection and decision-making.

In the perception layer, there exist several causes that lead to missing sensor data, such as power and hardware failure. However, the loss of data is not limited to the perception layer; it can also occur because of data exchange problems with the communication layer [3]. Regardless of the layer from which the missing data can occur, data can be missing for short as well as long periods. Currently, missing sensor data is one of the important challenges in IoT because incomplete data leads to insufficient information, which in turn results in inaccurate analysis and can ultimately lead to wrong interpretations and decisions that can have costly results both economically and socially [7]. Commonly, missing sensory data shift the statistical parameters estimated from a model using the collected data, resulting in bias in the mean and/or an increase in variance within the collected data that will hinder the possibility to efficiently and accurately monitored patients. Sensory data with missing rates above 50% are unreliable for the decision-making process [3]. Hence, handling sensor data, e.g., through imputation, is essential.

This study presents an approach for imputing long temporal gaps in sensor data that relies on dynamic linear modeling and focuses on the univariate case, where a single variable of interest is imputed. In this study, we will assume that missing data are missing completely at random, which assumes that the mechanisms resulting from the missing data are completely independent from the variables (observed or unobserved) that structure it. In the context of retirement home IoT, it means that the behaviors of the patients being monitored, the sensors and the sensor nodes are completely independent from the events causing missing data, which is a general but fair assumption. Statistically, making such an assumption is practical because it supposes that there is no bias in the available data.

The paper is organized as follows: Section II presents essential information about imputation and the methods used in this subfield of statistics. Section III presents the technical aspects of the proposed imputation approach. Results and discussion are presented in section IV before concluding (section V).

## II. MISSING DATA BACKGROUND

From the sensory data acquisition perspective, what is missing data? It is when no values are obtained from a sensor during the observation process for a specific physical quantity, but they could have been obtained. Aside from malevolent tempering that could affect data acquisition from a sensor, which are situations we do not account for in this study, several causes lead to missing data including: sensor power down, sensor malfunction, and transmission failure issues. As a result, sample size is reduced, which can prevent some analyses from being performed because the statistical power to perform these analyses is too low [8], [9], [10].

There are two stages by which missing data can occur: (1) At the bulk or unit stage and (2) at the data item stage. Missing data at the unit or bulk stage is the result of malfunction, i.e., no data is collected from the sensor in these situations and can result in chunks of missing data. However, missing data at the item stage is sporadic. In this study, we will focus on missing data at the data item stage. To better understand how to handle missing data, the problem should be studied according to the proportion of missing data, the mechanism by which missing data happens and the pattern of missing data [9], [10], [11], [12].

Understanding the mechanisms by which missing data occur is important to properly impute them. In subsection A we explain the different levels at which missing data occur while in subsection B the different methods that have been proposed to impute these missing data are briefly presented.

### A. MISSING DATA LEVELS

In this section we are presenting the different levels by which the missing data can occur. We start with the proportion of the missing data.

#### 1) PROPORTION OF MISSING DATA

There are no predefined missing data proportion threshold that will lead to valid (or wrong) statistical inference. Yet, statistical inference quality is directly related to the amount of data available and in turn to the proportion of missing data. However, multiple studies have investigated how different proportions of missing data influence the quality of the statistical inference. For example, Schafer [13] concluded that 5% or less of missing data does not have major influences on the quality of a statistical inference. According to Bennett [14], statistical biasing is more likely to happen when the missing data rate is above 10%. However, based on a simulation study, Madley-Dowd et al. [15] concluded that the proportion of missing data should not be used to guide imputation strategy or inform on their efficiency, e.g., how imputation approaches handle bias in the data. That being said, information on the proportion of missing data is important to guide the decision about the imputation to use. Table 1 presents a commonly used guideline for missing data imputation strategies, which was initially proposed by Hair [9], [10]. However, the mechanisms that lead to missing data and their patterns in the missing data are much more important to account for in missing data analysis than the proportion of missing data [16].

#### 2) STATISTICAL MECHANISMS UNDERLYING MISSING DATA

Rubin [17], stated that there are three mechanisms by which missing data typically occur: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) missing not at random (MNAR). Being able to associate the missing data structure to one of these mechanisms is highly valuable because it guides the users to the best technique to properly handle the particularity of the data [18], [19], i.e., if we are able to correlate the situation at which the missing data happens or the structure of the time series including the missing data segments with one of the mentioned three missing data mechanisms. To understand the different mechanisms underlying missing data, we first need to define a mathematical reference to rely on. As a starting point, let us use a vector of data Y as a reference point. This vector Y is composed of observed and missing values and can thus be partitioned in two parts: the observed values ($Y_{observed}$) and the missing values ($Y_{missing}$). The complete array of the IoT data including the missing data can be defined as:

$$Y = (Y_{observed}, Y_{missing}) \tag{1}$$

Using these two parts of Y, we can calculate the missingness value r for each value of Y, meaning that r has the same dimension as Y. Depending on the statistical mechanism considered, the way the missingness is calculated will change.

##### a: MISSING COMPLETELY AT RANDOM (MCAR)

Data missing completely at random occurs when the missingness is completely unrelated to the mechanisms that structure the collected data. Although there could be variables structuring the missingness, such as punctual meteorological events like a storm that could cause electrical surges preventing a sensor to work, they are unrelated to the observation (movement of a patient or behavior of a sensor). In other words, MCAR occurs when missing data depends neither on the missing data nor the observed data. Mathematically, MCAR is not conditional on any values

$$P(r|Y_{observed}, Y_{missing}) = P(r) \tag{2}$$

Under this assumption, the missing data is considered a random sample of the entire statistical population, which usually means that the standard error of the sample estimates is greater than that of the data. However, MCAR has the advantage of being unbiased. This is because of the reduced sample size [12].

##### b: MISSING AT RANDOM (MAR)

Data missing at random occurs when the probability of data missing depends on the observed data and not on the missing data itself. In more technical terms, for MAR the missingness is conditional on the observed values

$$P(r|Y_{observed}, Y_{missing}) = P(r|Y_{observed}) \tag{3}$$

In MAR, it is assumed that there are variables of important for Y that also define the missingness. As in [20] and [21], it is impossible to test whether MAR assumption is valid for data or not solely with the prior knowledge of observed data. However, it is possible to inspect the tenability of MAR assumption using a t-test that test the difference between the means of the complete dataset and that of the missing dataset [16], [19].

##### c: MISSING NOT AT RANDOM (MNAR)

Data missing not at random occurs when the missingness depends on the missing or the observed data, for example, when a sensor does not make if specific voltage is reached or if a patient tempers with a sensor. Mathematically, MNAR is defined as

$$P(r|Y_{observed}, Y_{missing}) = P(r|Y_{observed}, Y_{missing}) \tag{4}$$

#### 3) PATTERN OF MISSING DATA

There are three patterns missing data can follow [12] univariate, monotone [22] and arbitrary. A univariate pattern of missingness means that missing data can be attributed to a single variable. A monotone pattern of missingness occurs when missing values occur at a regular interval. In addition,

**TABLE 1.** Recommended imputation methods for different percentage of missing data.

| Percentage of Missing Data | Imputation Method | |
|---|---|---|
| <10% | Any imputation method can be applied. | |
| 10%-20% | For MCAR:<br>- All methods are available.<br>- Hot deck case substitution.<br>- Regression. | For MAR:<br>- Model Based methods are preferred. |
| >20% | - Regression method for MCAR. | - Model Based Methods for MAR. |

a monotone pattern of missing data usually means that there is a dependence within the missing data itself. It is important to be aware that this regularity does not have to be temporal (e.g., at a specific time interval), it could also be because a specific voltage threshold is reached preventing the sensor to gather data. Other than the univariate and monotone patterns, in this paper we assumed that missing data patterns are arbitrary. From a computational perspective, univariate and monotone pattern of missingness are straightforward to handle compared to arbitrary missing data [12] hence statistical methods such as univariate regression [23] or mean substitution imputation by class [24] have been designed to approach either of these problems, respectively. A complete scheme representing the missing data problem is presented in Figure 2. In the following subsection we briefly present common imputation methods that can be used for sensory data.

### B. IMPUTATION METHODS

Multiple techniques have been proposed to deal with missing data. Generally, deletion, ignorance and imputation are the three major classes of methods used to handle missing data. The disadvantage of deletion and ignorance is that they create a bias and reduce the amount of data available for analyses which in turn also reduced the quality of results. Conversely, the objective of the imputation is to replace missing data with reasonable values for the problem we are confronted with. It is important to be aware that the way data is imputed may change depending on the goal of our study. When thinking of missing sensor data, there are three classes of imputation methods [3], [25] depending on the type of information used to make the imputation. In the following lines we give a brief explanation of each one of these imputation classes.

#### 1) SPATIAL IMPUTATION

Spatial imputation assumes we have a priori knowledge of the spatial correlation between sensors or sensor nodes that we can use as reference to make an imputation. Specifically, if two sensors are near one another, it is assumed that they capture a similar signal than when they are further apart. Spatial imputation uses this information to handle missing data. In this respect, spatial correlation is calculated using the spatial coordinates of the data. Several studies proposed different approaches to perform spatial imputation. For example, association rule mining techniques such as the Window Association Rule Mining [26] and Freshness Association Rule Mining [25] have been specifically designed for imputing data on networks of sensors. Technically, these methods estimate missing data using association rules among neighbor sensors for which data have been gathered. Although association rule mining was initially designed for spatial imputation, it can also be used for temporal imputation.

#### 2) TEMPORAL IMPUTATION

Temporal imputation requires a priori knowledge of the temporal correlation between the readings collected from a single sensor. Similarly, to spatial imputation, when performing temporal imputation, it is assumed that data gathered at a short temporal interval are more similar than data gathered across longer period. In this respect, temporal correlation is calculated using the time at which each data from a sensor were gathered. Linear interpolation [27], Last Observation Carried Forward (LOCF) [28], autoregressive model [29], and Support Vector Regression (SVR) [30] are commonly used to perform temporal imputation although they can also be used to perform other types of imputations. However, these methods do not handle long temporal gaps efficiently and have a tendency to increase bias.

#### 3) SPATIO-TEMPORAL IMPUTATION

In this type of methods, the imputation is performed based on the a priori joint correlation for both spatial and temporal correlations for sensors or sensor nodes. Among these methods are Spatial and Temporal Imputation [31], Data Estimation using Statistical Model (DESM) [32], k-nearest neighbor estimation (AKE) [33], and Bayesian Gaussian Process (BGP) [34], [35]. The latter method is the closest one to our method used in imputing the missing data in this paper, where the current observation data are considered as Gaussian distributed given the past observation data. In the following section, we present our proposed methodology used in this research. The following section describes the proposed methodology in this research.

## III. METHODOLOGY

To efficiently impute long gaps in the data, we need to use a model that can account for long tendencies within the data. Dynamic Linear Models (DLMs) [36] are an appealing option to consider because these models are flexible and can account for short as well as long tendencies in the data. Mathematically, DLM relies on two equations, the observation equation (5) and the state or system equation (6)

$$Y_t = F\theta_t + v_t \tag{5}$$

$$\theta_t = G\theta_{t-1} + w_t \tag{6}$$

where $\theta_t$ defines the model structures (state) at time $t$ that usually depends on different explanatory variables,

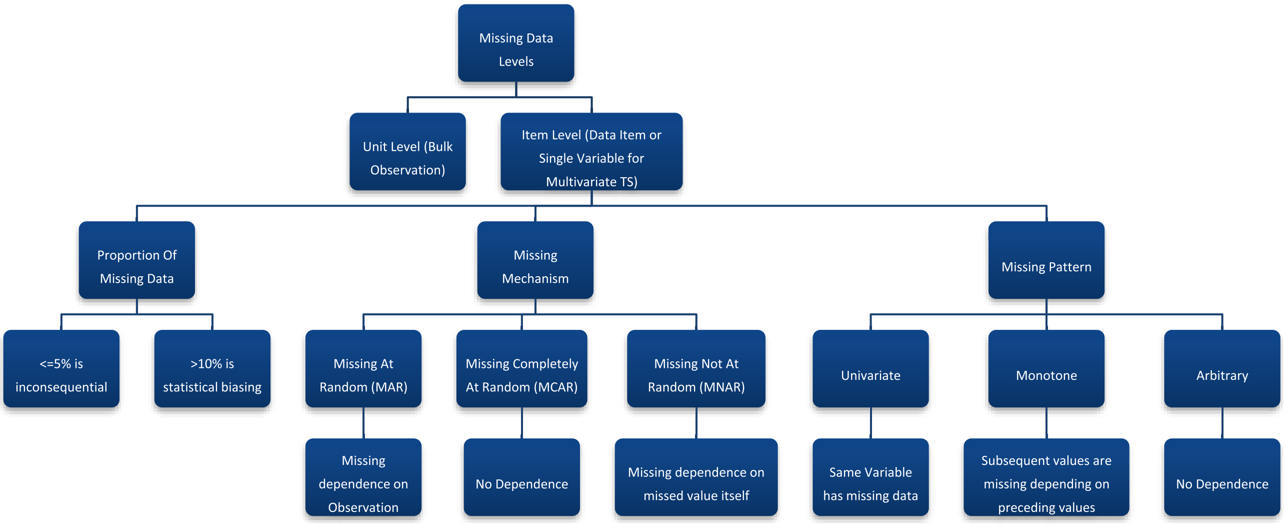
Missing Data Levels
Unit Level (Bulk Observation)
Item Level (Data Item or Single Variable for Multivariate TS)
Proportion Of Missing Data
Missing Mechanism
Missing Pattern
<=5% is inconsequential
>10% is statistical biasing
Missing At Random (MAR)
Missing Completely At Random (MCAR)
Missing Not At Random (MNAR)
Univariate
Monotone
Arbitrary
Missing dependence on Observation
No Dependence
Missing dependence on missed value itself
Same Variable has missing data
Subsequent values are missing depending on preceding values
No Dependence

**FIGURE 2.** Scheme for missing data problem.

$v_t \sim N_m(0, V_t)$ and $w_t \sim N_p(0, W_t)$. In words, $v_t$ follow a multivariate normal distribution of size $m$ with a covariance matrix $V_t$ that needs to be estimated. Similarly, $w_t$ follow a multivariate normal distribution of size $p$ with the covariance matrix $W_t$ that also needs to be estimated. In this general definition of a DLM, $G$ and $F$ are known matrices quantifying the importance of the model structure $\theta_t$ through time.

In the most basic case, a DLM can be simplified to a random walk model with the following observation and state equations

$$Y_t = \theta_t + v_t \tag{7}$$

$$\theta_t = \theta_{t-1} + w_t \tag{8}$$

where, in this simpler case, $v_t \sim N(0, \sigma_v^2)$ and $w_t \sim N(0, \sigma_w^2)$. That is both $v_t$ and $w_t$ follow a univariate normal distribution. Compared to the more general the one presented in (7) and (8), all values associated to $m$, $p$, $G$ and $F$ are equal 1. In more colloquial terms, with the previous model, the time series is modeled as fluctuating around some level $\theta$, which can change through time without any additional structuring constraints.

If we assume that we only have the time series (the observations), we can construct a DLM to fit our data based on the prior information we have.

In the next lines, we briefly present the practical implementation of the DLM and of the data used in our study.

For the practical implementation, ten different older adults were monitored in their residence for periods ranging from one month to several years to identify their long-term behavior over the monitoring period. The monitoring system relies on motion and door sensors spread all over the residence to follow the subjects' activity levels day and night [2]. The parameter used to measure the activity level per day for each subject is the number of movements each subject does in average per day. This is because the motion sensors are triggered by the subject's movement in front of the motion sensor.

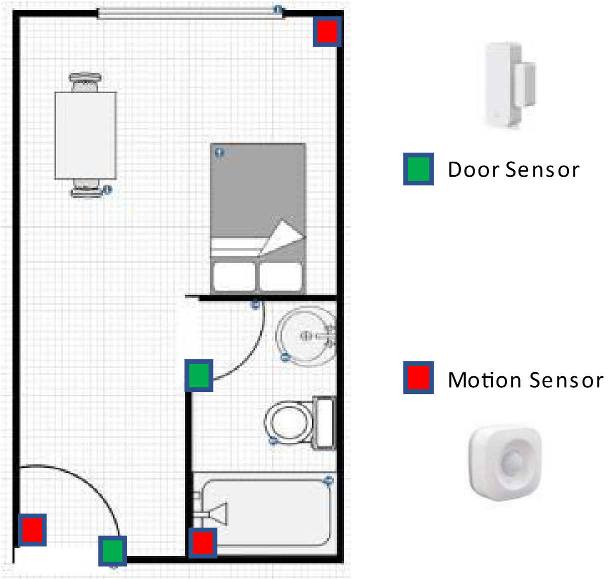

**FIGURE 3.** Room floor plan describing the distribution of both the motion and door sensors along with real photos of deployed sensors.

Figure 3 presents the locations of the sensors in a typical monitored room with pictures of the sensors that were deployed. Motion sensors are used to detect subjects' daily activities in the bedroom and bathroom while door sensors detected outing and visiting activities. Motion sensors were used to detect the subjects' movements within the sensors' line-of-sight, as described by Figure 4.

The overall activity level per day as monitored by the motion sensors for each subject is presented in Figure 5. The missing data are evident along the time series of the captured movements. The proportion of missing data for each subject is presented in Table 2.

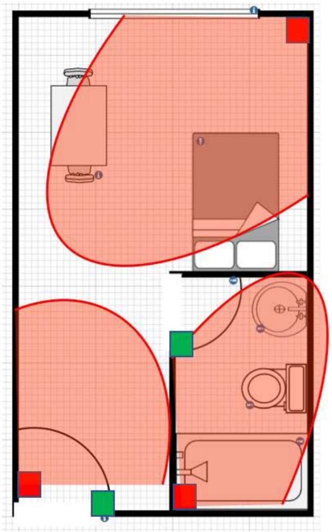

**FIGURE 4.** Room floor plan describing the line-of-sight areas (transparent red areas) for the motion sensors inside the room.

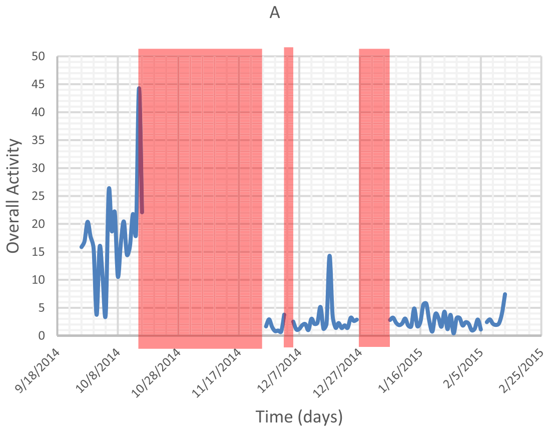

**FIGURE 5.** Activity level per day for subject A where the missing segments are highlighted in red.

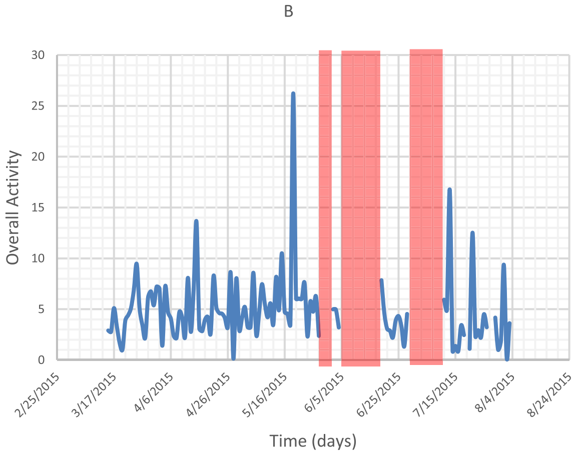

**FIGURE 6.** Activity level per day for subject B where the missing segments are highlighted with red.

After inspecting the data presented in Figures 5-14, we concluded that the missing data was MCAR. That is, the missing data gap location and size are randomly distributed along the monitored period for each subject and as such there are no expected bias in the missing data. Note also that the missing pattern is univariate because only the daily activity of each subject is considered. Hence, state-space models are best suited for this type of missing data. State-space models

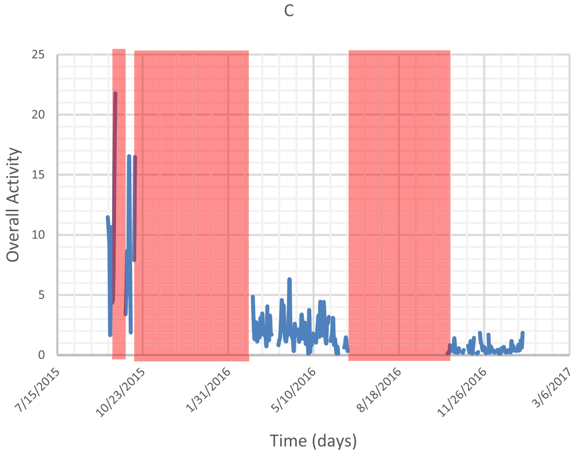

**FIGURE 7.** Activity level per day for subject C where the missing segments are highlighted with red.

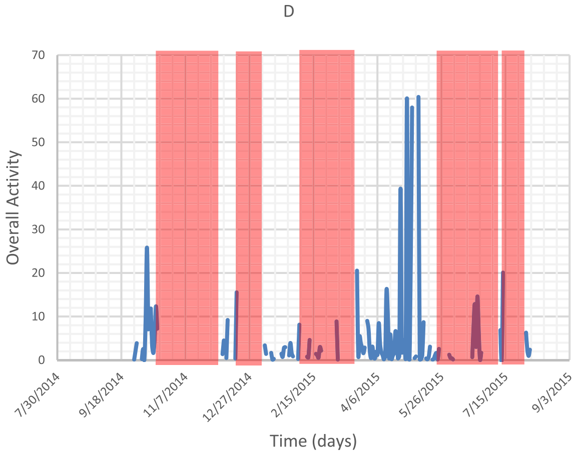

**FIGURE 8.** Activity level per day for subject D where the missing segments are highlighted with red.

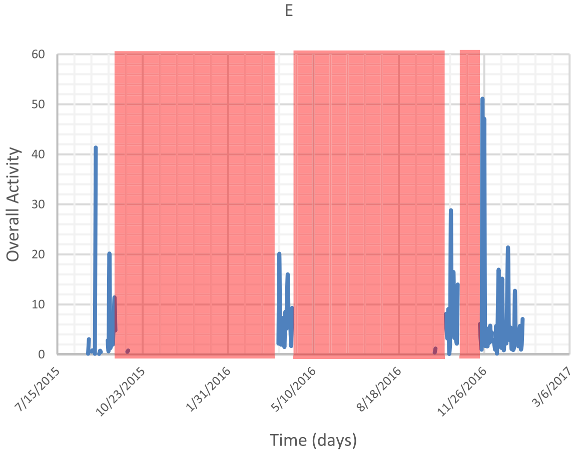

**FIGURE 9.** Activity level per day for subject E where the missing segments are highlighted in red.

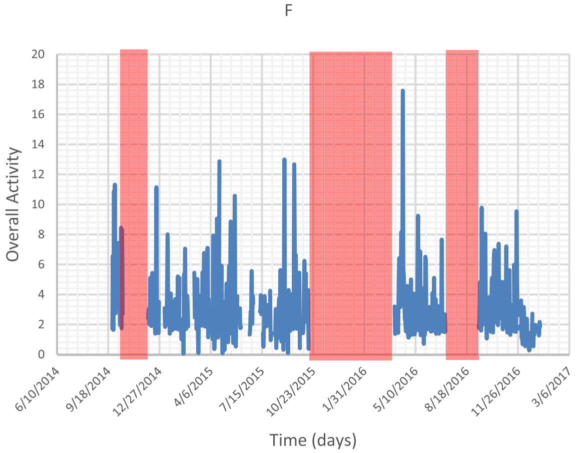

**FIGURE 10.** Activity level per day for subject F where the missing segments are highlighted in red.

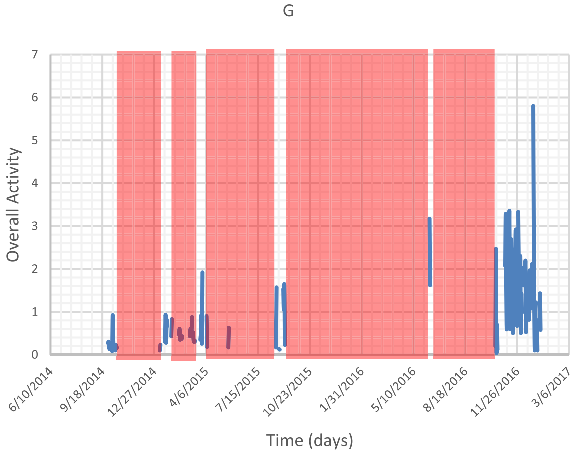

**FIGURE 11.** Activity level per day for subject G where the missing segments are highlighted in red.

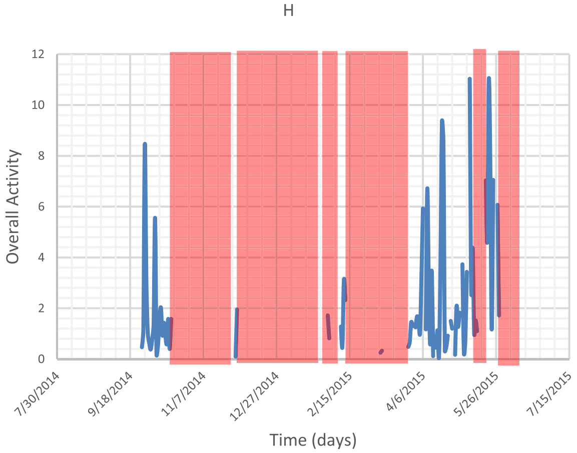

**FIGURE 12.** Activity level per day for subject H where the missing segments are highlighted in red.

consider a time series as the output of a dynamic system perturbed by random noise, which were considered as following a Gaussian distribution around the states of the time-series. Model estimation and forecasting were carried out by recursively computing the conditional distribution of the daily activity, given the available information. In this sense, a natural way to treat this problem is through the Bayesian framework, where missing data can be estimated based on previously collected data.

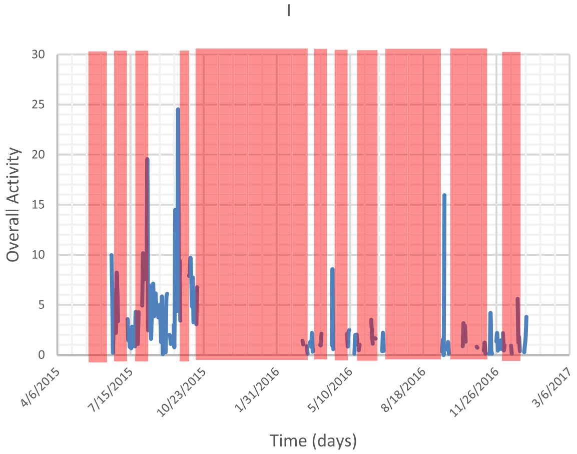

**FIGURE 13.** Activity level per day for subject I where the missing segments are highlighted in red.

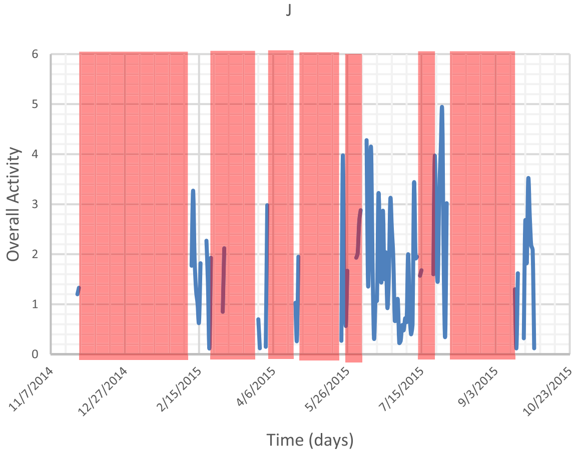

**FIGURE 14.** Activity level per day for subject J where the missing segments are highlighted in red.

**TABLE 2.** Proportion of missing data for each subject organized from least to most missing data.

| Case ID | % Of Missing Data |
|---|---|
| B | 23.20% |
| A | 37.60% |
| F | 44.50% |
| D | 55.40% |
| C | 61.10% |
| H | 64.50% |
| I | 68% |
| J | 69.5% |
| E | 77.40% |
| G | 81.30% |

### A. STEPS FOR ESTIMATING DLM PARAMETERS (PROPOSED MATHEMATICAL IMPLEMENTATION)

i. Estimate the rolling mean value for the time series to have a mean (expected) value for each observation in addition of the observation itself.

ii. Calculate the difference between each observation and its expected value to construct a vector of error, which represents the error distribution around the mean (expected) value along the time series.

iii. Fit the error distribution to a Gaussian distribution and estimate its variance parameter. In more technical terms, in this step we estimate $\boldsymbol{\sigma_v^2}$, which is the basis of the error in the observation equation (Equation 7). Note that the mean of the Gaussian distribution is assumed to be 0 because it is accounted for directly by the structure of the model.

iv. Following, we ran an iterative loop to calculate the states of the model ($\boldsymbol{\theta_t}$), since, for each timestamp $\boldsymbol{t}$, we have information on the observation ($\boldsymbol{Y_t}$) and error value ($\boldsymbol{v_t}$). Note that the states of the model can be defined based on the user's preference. In our implementation, we used a polynomial of degree 6 with overall activity as explanatory variable.

v. After estimating the states of the time series, we repeat the previous steps but this time to estimate $\boldsymbol{\sigma_w^2}$ and $\boldsymbol{\theta_{t-1}}$ (Equation 8).

vi. In a nutshell, the key idea is to estimate $\boldsymbol{\sigma_v^2}$ and $\boldsymbol{\sigma_w^2}$ that are the basis of the observation error and the states error, which can then be used to forecast missing data based on the prior knowledge of the observations and states together.

### *B. CONFIDENCE INTERVAL CALCULATIONS*

For the imputation results to be less extreme, it is also possible to constrain $v_t$. A way to do this is by calculating a confidence interval on the resulting values in $v_t$. For example, calculating a confidence interval (CI) at a 95% confidence level can be performed as:

$$CI|_{95\%} = \pm \frac{1.960 \times \sigma_v}{\sqrt{N}}. \tag{9}$$

where ±1.960 are lower and upper quantiles of the Gaussian distribution resulting in the area under the distribution to sum to 95% of the entire distribution, $\sigma_v$ is the observation standard deviation and $N$ the number of missing values estimated.

As an example, if we reconstruct $v_t$ for subject A but relying on the 95% confidence interval instead of the entire data, more extreme values can be seen to have a strong impact on the structure of the missing values to be estimated (Figure 15). The same results hold for the other subject considered here.

With the Gaussian parameters estimated for each subject's time series, we propose a procedure to estimate multiple segments of missing data for multiple subject's time series (Figure 16). The computational algorithm for estimating the missing sensory data based on DLM is presented in Algorithm 1. This algorithm is repeated for each subject independently.

## IV. RESULTS AND DISCUSSION

The Gaussian distribution parameters for each subject with and without considering the 95% confidence interval are presented in Table 3 along with the corresponding missing data rates and available number of observations. The highest missing data rate is associated to subject G while subject B has the lowest missing data rate. The estimated missing data combined with the observed data are presented for each subject in Figures 17-26. For subjects C (Figure 19), D (Figure 20), E (Figure 21), G (Figure 23), H (Figure 24) and I (Figure 25) negative activity values were estimated,

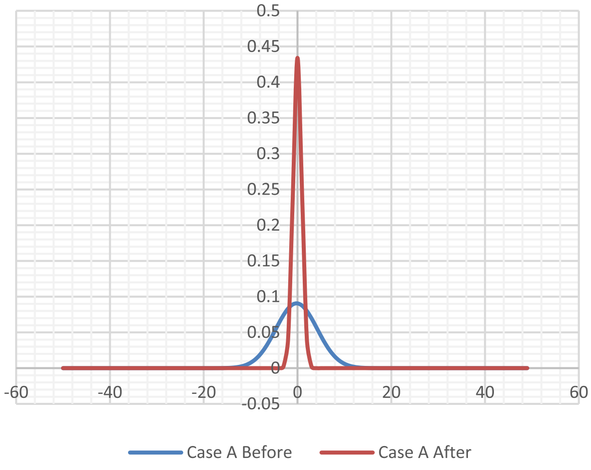

**FIGURE 15.** **$v_t$ distributions for subject A with and without using confidence interval calculation.**

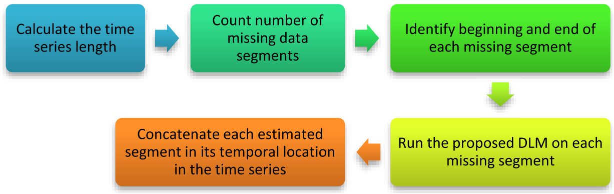
Calculate the time series length
Count number of missing data segments
Identify beginning and end of each missing segment
Run the proposed DLM on each missing segment
Concatenate each estimated segment in its temporal location in the time series

**FIGURE 16.** **Complete estimation process for missing data segments along each subject's overall activity per daytime series.**

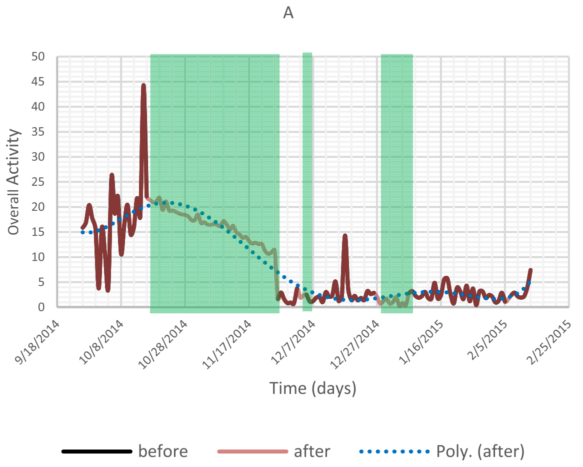

**FIGURE 17.** **Estimated data segments overlapped with existing data segments for subject A, where the estimated segments are highlighted with green.**

**TABLE 3.** Estimated subject's parameters.

| Cases | # Of available samples | Without Confidence Level | | With Confidence Level 95% | | | |
|---|---|---|---|---|---|---|---|
| | | Observation Error | State Error | Observation Error | | State Error | |
| | | Std | Std | Mean | Std | Mean | Std |
| A | 88 | 4.4 | 2.2 | 0 | 0.919 | 0 | 0.459 |
| B | 109 | 2.975 | 1.095 | 0 | 0.558 | 0 | 0.205 |
| C | 190 | 1.920 | 0.787 | 0 | 0.273 | 0 | 0.112 |
| D | 139 | 8.248 | 2.530 | 0 | 1.371 | 0 | 0.420 |
| E | 116 | 7.307 | 2.980 | 0 | 1.329 | 0 | 0.542 |
| F | 465 | 1.777 | 0.709 | 0 | 0.161 | 0 | 0.064 |
| G | 156 | 0.634 | 0.261 | 0 | 0.099 | 0 | 0.041 |
| H | 89 | 2.009 | 0.956 | 0 | 0.417 | 0 | 0.198 |
| I | 188 | 2.427 | 1.022 | 0 | 0.346 | 0 | 0.146 |
| J | 95 | 0.9289 | 0.392 | 0 | 0.186 | 0 | 0.079 |

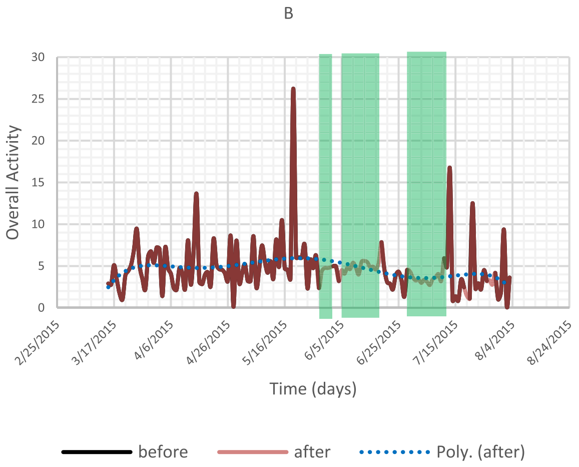

**FIGURE 18.** Estimated data segments overlapped with existing data segments for subject B, where the estimated segments are highlighted with green.

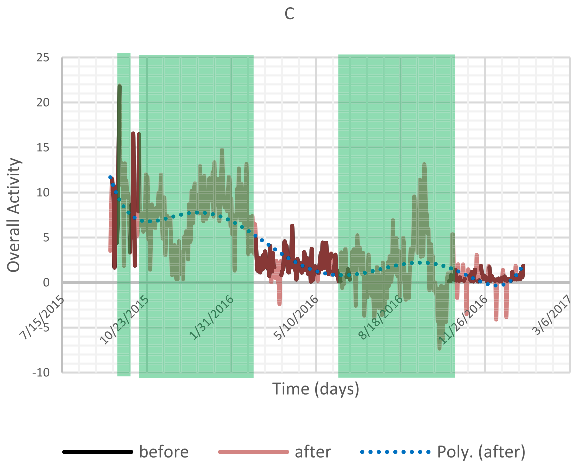

**FIGURE 19.** Estimated data segments overlapped with existing data segments for subject C, where the estimated segments are highlighted with green.

which are non-sensical and should thus not be accounted for in subsequent behavior analysis.

Subjects A (Figure 17), C (Figure 19), F (Figure 22), H (Figure 24), and I (Figure 25) were found to experience a decrease in their overall activity level between October and January.

Subjects were expected to be more frequently indoors during winter and less so during the summer of the same monitoring year. This was confirmed for subjects C (Figure 19), E (Figure 21), F (Figure 22), G (Figure 23), I (Figure 25) and J (Figure 26).

For subject A, there was a significant decrease in activity level between December and January compared to the October-November period, which was due to severe mobility impairments. Similarly, subject C (Figure 19) also experienced severe mobility impairments at the end of November 2015, resulting in very low activity levels, which we were able to detect with our imputation procedure. Note that the decrease in activity level common with the summer period was also clearly observed for subject C in May-August 2016.

There are spikes in the activity levels for all monitored subjects, which were attributed to nurse visiting the residences because with these visits the overall activity level inside the residence increased with the arrival of another person. These nurse visits are observed for subject A in mid-October 2014 and mid-December 2014 (Figure 17). For subject B, they occur in mid-May 2015 and on mid-July and late July 2015 (Figure 17). For subject C, they happen in late September 2015 (Figure 19). For subject D, they take place from late April to early May 2015 (Figure 20).

**Algorithm 1:** Forecasting Missing Sensory Data

```
    Input: resident_(X)_overall_activity_missing.csv
    Output: resident_(X)_overall_activity_estimated.csv
1   Initialization:
    Set df ← resident_(X)_overall_activity_missing.csv
    Set df.index ← df[“Date”]
    Set df.index.sort
    Set df.frequency(“Day”)
2   Calculate Missing Rate:
    Set count_nan ← sum(df.nan())
    Set missing_data_rate ← (count_nan /  length(df)∗100)
3   Cut series into data-containing segments:
    Set events ← df.split(df, where(df.nan()))
4   Calculate the time deltas between data segments:
    Set deltas ← empty
5   for counter in range (length( events) - 1 ):
        Set temp1 ← events[counter]
        Set temp2 ← events[counter] + 1
        Set tempdelta ← temp2.index[0] − temp1.index[-1]
        Set deltas ← deltas.append(tempdelta.days())
    end
6   Obtain mean of the  observation series:
    Set observation ← df
    Set smoothed_observation ← observation.rollingmean()
7   Calculate number of iterations for error estimation process:
    Set length_observation ← length( observation)
    Set length_smoothed_observation ← length( smoothed_observation)
    Set loop_start ← length_observation −  length_smoothed_observation
8   Error estimation process:
    Set estimated_error_vector ← empty
    for  n in range(loop_start,  length_observation ):
        Set temp ← observation[n] − smoothed_observation[n]
        Set estimated_error_vector ←  estimated_error_vector.append(temp)
    end
9   Obtain error vector missing values:
    Set mean ← mean.estimated_error_vector()
    Set std ← std.estimated_error_vector()
    Set missed_distribution_values ← random.normal(mean,  std, size(loop_start))
    Set estimated_error_vector ←
    estimated_error_vector.append(missed_distribution_values)
10  Estimate mu_t vector:
    Set estimated_mu_t ← empty
    for  n in range(length(observation)):
        Set temp ← observation[n] −  estimated_error_vector[n]
        estimated_mu_t.append(temp)
    end
11  Obtain mean of mu_t  series:
    Smoothed_mu_t ← estimated_mu_t.rollingmean()
12  mu_t error estimation process:
    Set smoothed_mu_t ← smoothed_mu_t[0]
    Set estimated_mu_t_error ← empty
    for  n in range(loop_start,  length_observation ):
        Set temp ← estimated_mu_t[n] − smoothed_mu_t[n]
        estimated_mu_t_error ← estimated_mu_t_error.append(temp)
    end
13  Forecasting process:
    Set start_index ← zero
    Set new_series ← empty
    for counter in range(length(events) - 1):
14      Set start_index ← start_index + length(events[counter])
        Set forecasted_observation ← empty
        Set forecasted_mu_t ← [estimated_mu_t[start_index]]
        Set mean ← mean.estimated_mu_t_error ()
        Set std ← std.estimated_mu_t_error()
        for  n in range(deltas[counter] -  1):
            Set temp_mu_t ←forecasted_mu_t[-1] +  random.normal(mean, std, 1)
            Set temp_observation ← temp_mu_t +
            random.normal(mean.estimated_error_vector(),
            std.estimated_error_vector(), 1)
            Set forecasted_mu_t ← forecasted_mu_t.append(temp)
            Set forecasted_observation ←
            forecasted_observation.append(temp_observation)
        end
        Set temp_series ←  events[counter].append(forecasted_observation)
        Set new_series ← newseries.append(temp_series)
    end
    Set resident_(X)_overall_activity_estimated.csv ← new_series
```

For subject E, these spikes are observed in late August 2015, April 2015, October 2015 and November 2015 (Figure 21).

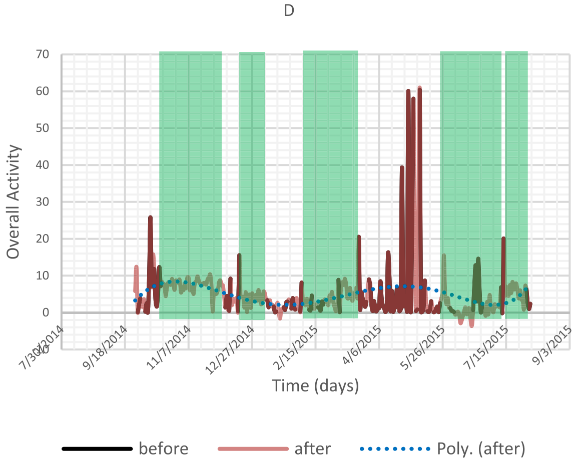

**FIGURE 20.** Estimated data segments overlapped with existing data segments for subject D, where the estimated segments are highlighted with green.

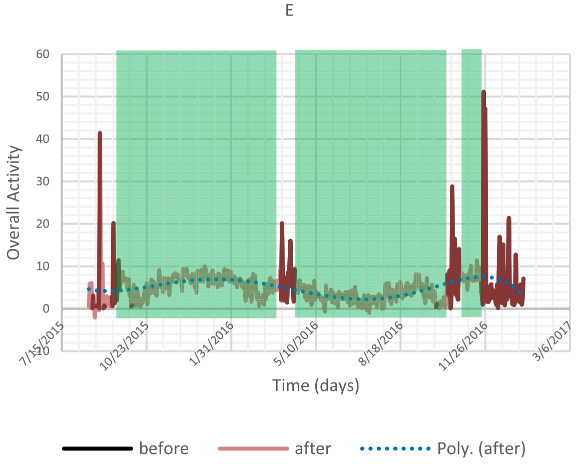

**FIGURE 21.** Estimated data segments overlapped with existing data segments for subject E, where the estimated segments are highlighted with green.

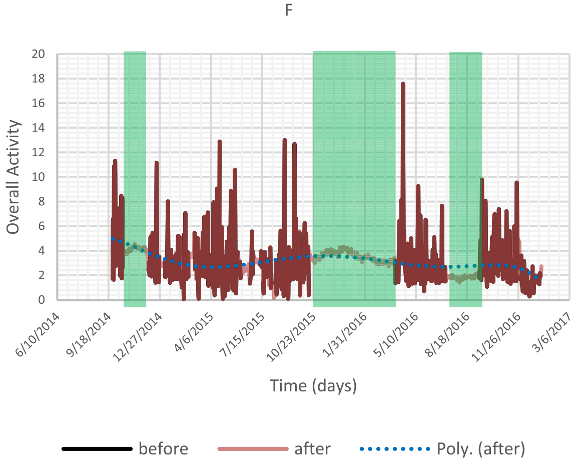

**FIGURE 22.** Estimated data segments overlapped with existing data segments for subject F, where the estimated segments are highlighted with green.

For subject F, multiple nurse visits occur during between September 2014 and October 2015 and in April 2016 (Figure 22). For subject G, a single nurse visit is observed in

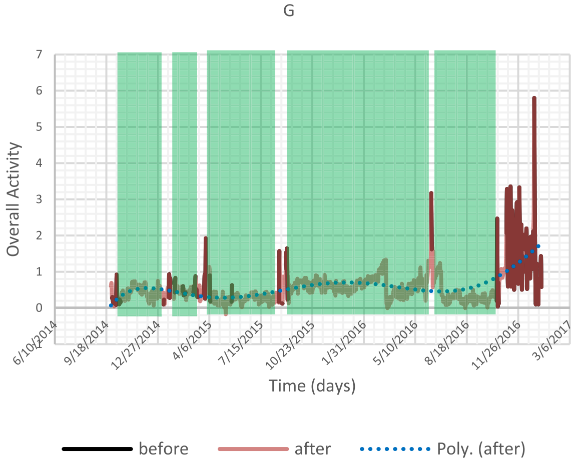

**FIGURE 23.** **Estimated data segments overlapped with existing data segments for subject G, where the estimated segments are highlighted with green.**

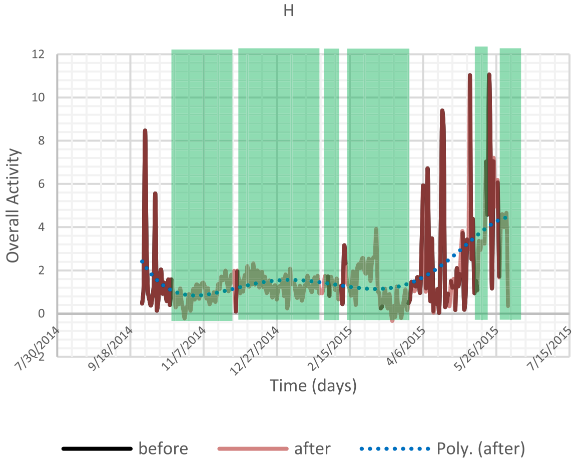

**FIGURE 24.** **Estimated data segments overlapped with existing data segments for subject H, where the estimated segments are highlighted with green.**

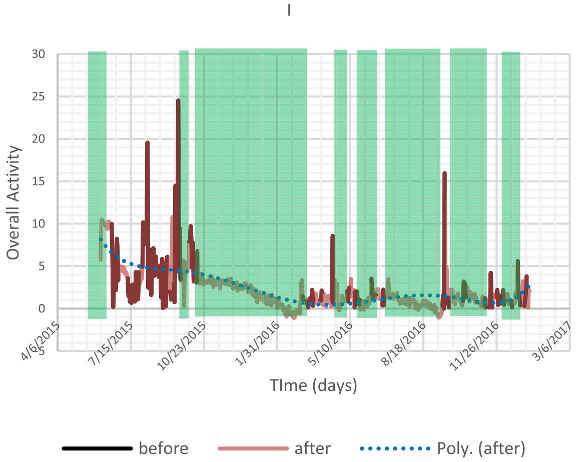

**FIGURE 25.** **Estimated data segments overlapped with existing data segments for subject I, where the estimated segments are highlighted with green.**

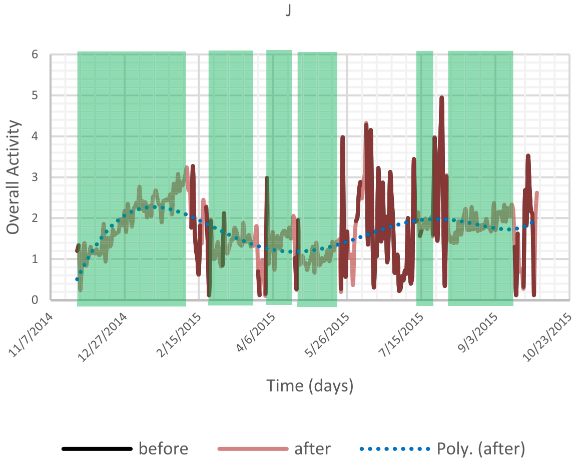

**FIGURE 26.** **Estimated data segments overlapped with existing data segments for subject J, where the estimated segments are highlighted with green.**

late December 2016 (Figure 23). For subject H, nurse visits occur in late September and early October 2014 and also in between late April and late May 2015 (Figure 24). For subject I, nurse visits occur in August and October 2015 as well as in April and September 2016 (Figure 25). Lastly, for subject J, no nurse visits can be observed as the subject's behavior is highly fluctuating (Figure 26).

## V. CONCLUSION

In this paper, we proposed and implemented an approach to estimate the missing sensor data independent of their respective temporal location. This approach is based on Bayesian Gaussian Process and is based on knowledge gain from past observations. We applied this approach to impute the missing sensor data obtained from the IoT overall activity monitoring system for older adults in residences. The imputation process verifies the assumption that indoor activities are higher in winter compared to summer. The imputation approach proposed also enables the identification of long-term behavior changes such as mobility impairment suffered by some subjects. We also verified the assumption that there are behavioral changes seasonally, especially between winter and summer with higher indoor activity during winter compared to summer periods. We were also able to justify nurse visits in the residence, with very high increase in activity level during the visit day compared to the past activity levels for the same subject.

In this study we collected data from real-life deployment and hence, the missing part in the data are not existent in reality. Based on this fact, the dataset collected does not include a no-missing data part, and in consequence the comparison between the missing dataset case and the no-missing dataset case is not feasible. However, we have limited the extreme values that might be obtained due to imputation by calculating the confidence interval for the estimated distribution of the past collected data, and then estimating the missing data based on it. This would limit the error and increase the accuracy of the missing data estimation. Moreover, the mathematical model utilized in our approach, is based on treating the acquired time series as an outcome from a random process, which limits the concept of the evaluation metric itself.

## REFERENCES

[1] F. Kaddachi, H. Aloulou, B. Abdulrazak, P. Fraisse, and M. Mokhtari, ‘‘Technological approach for early and unobtrusive detection of possible health changes toward more effective treatment,’’ in *Proc. Int. Conf. Smart Homes Health Telematics*, 2018, pp. 47–59.

[2] F. Kaddachi, H. Aloulou, B. Abdulrazak, P. Fraisse, and M. Mokhtari, ‘‘Long-term behavior change detection approach through objective technological observations toward better adaptation of services for elderly people,’’ *Health Technol.*, vol. 8, no. 5, pp. 329–340, Nov. 2018.

[3] Y. Liu, T. Dillon, W. Yu, W. Rahayu, and F. Mostafa, ‘‘Missing value imputation for industrial IoT sensor data with large gaps,’’ *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6855–6867, Aug. 2020.

[4] W. Yu, T. Dillon, F. Mostafa, W. Rahayu, and Y. Liu, ‘‘A global manufacturing big data ecosystem for fault detection in predictive maintenance,’’ *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 183–192, Jan. 2019.

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, ‘‘Internet of Things: A survey on enabling technologies, protocols, and applications,’’ *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[6] B. Abdulrazak, S. Paul, S. Maraoui, A. Rezaei, and T. Xiao, ‘‘IoT Architecture with plug and play for fast deployment and system reliability: AMI Platform,’’ in *Proc. Int. Conf. Smart Homes Health Telematics*, 2022, pp. 43–57.

[7] C. C. Aggarwal, N. Ashish, and A. Sheth, ‘‘The Internet of Things: A survey from the data-centric perspective,’’ in *Managing and Mining Sensor Data*. Springer, 2013, pp. 383–428.

[8] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, vol. 5, no. 8. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[9] J. F. Hair, *Multivariate Data Analysis*, 7th ed., J. F. Hair, W. C. B. Barry, J. B. Rolph, and E. Anderson, Eds. 2009.

[10] J. F. Hair, *Multivariate Data Analysis*. 2009.

[11] S. Van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL, USA: CRC Press, 2018.

[12] Y. Dong and C.-Y.-J. Peng, ‘‘Principled missing data methods for researchers,’’ *SpringerPlus*, vol. 2, no. 1, pp. 1–17, Dec. 2013.

[13] J. L. Schafer, ‘‘Multiple imputation: A primer,’’ *Stat. Methods Med. Res.*, vol. 8, no. 1, pp. 3–15, Jan. 1999.

[14] D. A. Bennett, ‘‘How can i deal with missing data in my study?’’ *Austral. New Zealand J. Public Health*, vol. 25, no. 5, pp. 464–469, Oct. 2001.

[15] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, ‘‘The proportion of missing data should not be used to guide decisions on multiple imputation,’’ *J. Clin. Epidemiol.*, vol. 110, pp. 63–73, Jun. 2019.

[16] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using Multivariate Statistics*, vol. 5. Boston, MA, USA: Pearson, 2007.

[17] D. B. Rubin, ‘‘Inference and missing data,’’ *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[18] R. N. Faizin, M. Riasetiawan, and A. Ashari, ‘‘A review of missing sensor data imputation methods,’’ in *Proc. 5th Int. Conf. Sci. Technol. (ICST)*, Jul. 2019, pp. 1–6.

[19] J. R. Carpenter and M. Smuk, ‘‘Missing data: A statistical framework for practice,’’ *Biometrical J.*, vol. 63, no. 5, pp. 915–947, Jun. 2021.

[20] I. R. White, P. Royston, and A. M. Wood, ‘‘Multiple imputation using chained equations: Issues and guidance for practice,’’ *Statist. Med.*, vol. 30, no. 4, pp. 377–399, Feb. 2011.

[21] N. J. Horton and K. P. Kleinman, ‘‘Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models,’’ *Amer. Statistician*, vol. 61, no. 1, pp. 79–90, Feb. 2007.

[22] R. Faria, M. Gomes, D. Epstein, and I. R. White, ‘‘A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials,’’ *PharmacoEconomics*, vol. 32, no. 12, pp. 1157–1170, Dec. 2014.

[23] C. K. Enders, *Applied Missing Data Analysis*, vol. 16. New York, NY, USA: Guilford Press, 2010, pp. 171–180.

[24] G. Kalton and D. Kasprzyk, ‘‘Imputing for missing survey responses,’’ in *Proc. Sect. Surv. Res. Methods, Amer. Stat. Assoc.*, vol. 22, 1982, p. 31.

[25] L. Gruenwald, H. Chok, and M. Aboukhamis, ‘‘Using data mining to estimate missing sensor data,’’ in *Proc. 7th IEEE Int. Conf. Data Mining Workshops (ICDMW )*, Oct. 2007, pp. 207–212.

[26] M. H. Le Gruenwald, ‘‘Estimating missing values in related sensor data streams,’’ in *Proc. COMAD*, 2005, pp. 83–94.

[27] C.-Y. Li, W.-L. Su, T. G. McKenzie, F.-C. Hsu, S.-D. Lin, J. Y.-J. Hsu, and P. B. Gibbons, ‘‘Recommending missing sensor values,’’ in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 381–390.

[28] H. Zhou, K.-M. Yu, M.-G. Lee, and C.-C. Han, ‘‘The application of last observation carried forward method for missing data estimation in the context of industrial wireless sensor networks,’’ in *Proc. IEEE Asia–Pacific Conf. Antennas Propag. (APCAP)*, Aug. 2018, pp. 1–2.

[29] S. Sridevi, S. Rajaram, C. Parthiban, S. SibiArasan, and C. Swadhikar, ‘‘Imputation for the analysis of missing values and prediction of time series data,’’ in *Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT)*, Jun. 2011, pp. 1158–1163.

[30] S. P. Susanti and F. N. Azizah, ‘‘Imputation of missing value using dynamic Bayesian network for multivariate time series data,’’ in *Proc. Int. Conf. Data Softw. Eng. (ICoDSE)*, Nov. 2017, pp. 1–5.

[31] Y. Li and L. E. Parker, ‘‘A spatial–temporal imputation technique for classification with missing data in a wireless sensor network,’’ in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 3272–3279.

[32] Y. Li, C. Ai, W. P. Deshmukh, and Y. Wu, ‘‘Data estimation in sensor networks using physical and statistical methodologies,’’ in *Proc. 28th Int. Conf. Distrib. Comput. Syst.*, Jun. 2008, pp. 538–545.

[33] L. Pan and J. Li, ‘‘K-nearest neighbor based missing data estimation algorithm in wireless sensor networks,’’ *Wireless Sensor Netw.*, vol. 2, no. 2, p. 115, 2010.

[34] M. A. Osborne, S. J. Roberts, A. Rogers, and N. R. Jennings, ‘‘Real-time information processing of environmental sensor network data using Bayesian Gaussian processes,’’ *ACM Trans. Sensor Netw.*, vol. 9, no. 1, pp. 1–32, Nov. 2012, doi: 10.1145/2379799.2379800.

[35] M. A. Osborne, S. J. Roberts, A. Rogers, and N. R. Jennings, ‘‘Real-time information processing of environmental sensor network data using Bayesian Gaussian processes,’’ *ACM Trans. Sensor Netw.*, vol. 9, no. 1, pp. 1–32, Nov. 2012.

[36] G. Petris, S. Petrone, and P. Campagnoli, ‘‘Dynamic linear models,’’ in *Dynamic Linear Models With R*. Springer, 2009, pp. 31–84.

**HASSAN M. AHMED** received the B.Sc. and M.Sc. degrees in biomedical engineering from Helwan University, Cairo, Egypt, in 2012 and 2016, respectively. He is currently pursuing the Ph.D. degree in computer science with the Université de Sherbrooke on behavior change detection. His research interests include signal processing, medical image processing, machine learning, data analysis, mathematical modeling, as well as electromagnetics, elastography, biomechanics, soft tissue shear wave estimation, and dual band RF coils for ultrahigh MRI.

**BESSAM ABDULRAZAK** (Member, IEEE) received the B.Sc. degree in electronics from USTHB, Algeria, the M.Sc. degree in robotics from Paris 6, France, and the Ph.D. degree in computer science from Telecom SudParis, France. He is currently a Professor of computer science with the Université de Sherbrooke and the Director of the AMI Laboratory, Sherbrooke, QC, Canada. He is an active Researcher with the Research Center on Aging and the Interdisciplinary Institute for Technological Innovation. His research interests include the IoT, ubiquitous and pervasive computing, ambient intelligence, smart environments, assistive living technologies, context awareness, and software engineering. He has over 200 peer-reviewed publications, served as the general chair for a number of conferences and workshops, and serves on the editorial board of numerous international journals, as well as program committee of several conferences related to his research interests.

**F. GUILLAUME BLANCHET** received the B.Sc. and M.Sc. degrees in biological sciences from the Université de Montréal, Canada, and the Ph.D. degree in conservation biology from the University of Alberta, Canada. He is currently a Professor with the Department of Biology, Mathematics and Community Heath, Université de Sherbrooke, QC, Canada, and the Director of the Quantitative Biology Laboratory, Sherbrooke. He is an active Researcher with the Research Center on Aging and the Quebec Center for Biodiversity Science. His research interests include the development and the application of statistical methods and mathematical models to approach a variety of problems in biology and health sciences. He has over 40 peer-reviewed publications and serves on the editorial board of *International Journal of Ecology* (Population Ecology and Ecography).

**HAMDI ALOULOU** is currently an Associate Professor with the University of Monastir. He is a Senior Scientist with the Digital Research Centre of Sfax, Tunisia, and an Associate Researcher with the Institut Mines-Telecom, Paris, France. As part of his research work, he was actively involved in different European projects. His research interests include knowledge management and processing for decision-making applied in the domain of ambient intelligence/the Internet of Things (AmI/IoT). The target of his work is to set up intelligent living spaces in order to improve the quality of life and promote public, individual, and collective health.

**MOUNIR MOKHTARI** received the Ph.D. degree in computer science in the field of human–machine interaction and the Research Habilitation degree from the University Pierre & Marie Curie, Paris, in 1997 and 2002, respectively. He is currently a Full Professor classe 1 with the Institut MINES-TELECOM, France, and the Director of IPAL—CNRS (UMI 2955) French-Singaporean joint Laboratory 2014–2018. His background is mainly in human–machine interaction in the domain of ambient assisted living and semantic reasoning. He has successfully supervised 15 Ph.D. students in this topic, examined several Ph.D. candidates, and has over 200 publications in journals, books, and international conferences. He hold the Chair on Quality of Life on Aging people 2011–2016. He is a PI and a Coordinator of several National and European projects and industrial contracts (PSA Peugeot-Citroën, AXA Research Fund, Mondial Assistance, and Mutuelle Generale) on smart living and wellbeing with an overall budget over 40 millions euros over the past decade. He is the Founder in 2003 of annual International Conference on Smart Living and Public Health (ICOST).