mathematics

Article

Semi-Markov Models for Process Mining in Smart Homes

Sally McClean [1,*,†] and Lingkai Yang [2,†]

[1] School of Computing, Ulster University, Belfast BT15 1AP, Northern Ireland, UK
[2] Research Institute of Mine Big Data, Chinese Institute of Coal Science, Beijing 100013, China; y1163376026@gmail.com
* Correspondence: si.mcclean@ulster.ac.uk
† These authors contributed equally to this work.

Abstract: Generally, these days people live longer but often with increased impairment and disabilities; therefore, they can benefit from assistive technologies. In this paper, we focus on the completion of activities of daily living (ADLs) by such patients, using so-called Smart Homes and Sensor Technology to collect data, and provide a suitable analysis to support the management of these conditions. The activities here are cast as states of a Markov-type process, while changes of state are indicated by sensor activations. This facilitates the extraction of key performance indicators (KPIs) in Smart Homes, e.g., the duration of an important activity, as well as the identification of anomalies in such transitions and durations. The use of semi-Markov models for such a scenario is described, where the state durations are represented by mixed gamma models. This approach is illustrated and evaluated using a publicly available Smart Home dataset comprising an event log of sensor activations, together with an annotated record of the actual activities. Results indicate that the methodology is well-suited to such scenarios.

Keywords: Markov-type model; process mining; Smart Homes; convolution of gamma mixture models

MSC: 60K15

check for updates

Citation: McClean, S.; Yang, L. Semi-Markov Models for Process Mining in Smart Homes. *Mathematics* **2023**, *11*, 5001. https://doi.org/10.3390/math11245001

Academic Editors: Panagiotis-Christos Vassiliou and Andreas C. Georgiou

Received: 21 November 2023
Revised: 14 December 2023
Accepted: 15 December 2023
Published: 18 December 2023

1. Introduction

Worldwide, people are living longer, with increasing numbers of older people forming a larger proportion of the population [1]. It is predicted that, by 2050, the proportion of the world population over 60 years of age will have grown to over 20%. However, the downside of this prediction is that it comes with increased impairment and disability in older people, including cognitive decline, depressive illnesses, and dementia. At the same time, however, there have been major improvements in technology and, in particular, assistive technologies are increasingly being used to facilitate the functioning, independence, and well-being of older citizens.

Older people frequently experience physical and cognitive decline, typically preventing the completion of activities of daily living (ADLs) [2]. Such patients may need family or professional care at home or might require admission to a long-term nursing facility, with a resulting decrease in quality-of-life and an increase in costs. High-tech solutions can help improve such situations, using so-called Smart Homes, which utilise assistive technologies and employ sensors to collect appropriate data and provide suitable analysis to support the diagnosis, monitoring, and treatment of these conditions.

Typically, in Smart Homes, sensors are placed on household objects, such as doors, to monitor ADLs, or may instead utilise imaging techniques to infer (sub-)activities. The activation of sensors can be recorded as a time-stamped log of low-level events. Such data can be thought of as an event history, commonly used in process mining [3], where, in the Smart Home case, the events typically characterise the start or end of an activity, or sub-activity, of daily living. The activity can then be analysed to identify ADL behavioural patterns or abnormalities. This may lead to a reminder to the patient, if required, or an alarm to alert carers or medical services that there may be an issue.

Markov models, and their extensions, offer a well-defined mathematical framework for the movements of individuals through processes, where the individual is known to enter and leave the system through a given state, at a given time. Such models have been widely explored and developed to allow for different application areas and settings, as well as for assessing the legitimacy of assumptions. More specifically, they have already been used for Smart Homes (e.g., [4]) and Process Mining (e.g., [5]). In this paper, we explore several possible directions that exploit the computational advantages of the Process Mining paradigm alongside the mathematical and operational benefits of Markov-type models. Several topics of interest are explored, namely: (i) first passage time distributions for classes of interest, such as specific activities or groups of activities; (ii) aggregation of states to identify levels of detail that are parsimonious and computationally efficient, while providing appropriate inference; (iii) detection of anomalies and concept drift for such activities of interest, which may be indicative of activities taking longer than previously or might signal a new emergent pattern.

This paper is concerned with developing and illustrating some specific mathematical expressions and results for models of human activity. If these are appropriate then computation should be efficient and can provide useful results, e.g., characterising and detecting outliers in performance indicators, such as the length of time taken to perform important (sub-)activities. Appropriate semi-Markov model assumptions can also be built into simulation or digital twin analysis, facilitating the study of different scenarios and their performance in various situations.

The novelty of this paper resides in the development and use of Markov-type models to represent ADL processes and corresponding key performance indicators (KPIs) in Smart Homes, as well as in describing and evaluating strategies to determine anomalies in such transitions and durations. The approach is illustrated and evaluated using a publicly available Smart Home dataset [6], comprising an event log of sensor activations, together with an annotated record of the actual activities, which we use for model development and validation.

2. Literature Review

Processes frequently occur in many different contexts, such as business, telecommunications, and healthcare. Previously, there have been substantial efforts to model and analyse such processes, with the aim of improving understanding and efficiency as well as predicting future outcomes. Also, with recent developments in IT, there exist more sophisticated computer systems to collect, store, process, and exchange data. This has led to the emergence of Process Mining, as a bridge between data mining and process modelling [3], which has already been applied to diverse areas, such as manufacturing (Lorenz et al. [7]), telecommunications (Mahendrawathi et al. [8]), and healthcare (Rojas et al. [9]).

In general, such processes are defined as consisting of a number of activities each with start and end times and corresponding durations. A process instance executes these activities in a sequence, following the logic and rules at work in real-world scenarios. Consequently, Process Mining may involve discovering the activities and trajectories that comprise the process, predicting trajectories, or identifying outliers. Such analysis may use standard data mining approaches such as classification, clustering, association rules, or deep learning. However, in addition, model-based approaches can provide possibilities for incorporating structural process knowledge into the analysis, thereby facilitating improved insight and enhanced forecasting.

A mathematical model can be used to provide a simplified version of a process, where analytic solutions or simulation models can imitate process behaviour without necessarily engaging with real-world scenarios [10]. Such models have frequently been used to address complex problems such as determining correctness, conformance, or performance. Performance analysis typically focuses on the dynamic behaviour of the process, based on key performance indicators (KPIs) such as response time, uptime, or reliability. More specifically, the main KPIs for Markov-type models have been identified by [11] as (i) state occupancy probabilities, i.e., number of visits to a state during an arbitrary time interval; (ii) first passage time probabilities; and (iii) state occupancy duration probabilities [12]. Also, Markov models are widely used probabilistic process models where it is assumed that the Markov property holds, i.e., the current state only depends on the immediately previous state. This assumption facilitates both the prediction of individual transitions [13] and population forecasting [14]. Higher-order Markov models [15] may alternatively be employed if the Markov assumption is not appropriate [16,17]. Also, non-homogeneous Markov models can be used if the model parameters are changing with time [13].

Another type of Markov model is the hierarchical Markov model, where states may be ordered in levels belonging to a hierarchy, where a group of lower-level states may constitute a single state at a higher level ([4,18]). In addition, hierarchical Markov models may only be partially observable, with unobserved states at the lower levels; such situations can be considered as hierarchical hidden Markov models [19]. On the other hand, depending on the level of interest, the lower levels of the hierarchy may be aggregated to combine states, remove kth order effects, and reduce computational complexity [20]. Also, as well as the basic time-stamped events (change of state) data, we may have additional covariates (features) which can be incorporated into the Markov models. In the literature, this has been accomplished by using partitioning approaches [10], conditional Bayesian networks [21], or making the transition parameters functions of the covariates [22].

As a result of such breadth and diversity, Markov-type models are highly applicable to a variety of application areas. A few such examples are manpower planning, e.g., Papadopoulou and Vassiliou [23], Verbeken and Guerry [24], and McClean et al. [25]; hospital planning, e.g., Marshall and McClean [21], Shaw et al. [26]; business process modelling, e.g., Yang et al. [5], Chen et al. [27]. In addition, there have been a number of papers using Markov-like models for Smart Homes, which are our current focus, e.g., Youngblood and Cook [18], Wang et al. [4]. We have also co-authored an introductory paper on this topic which focuses on the use of semi-Markov models for state durations in Smart Homes (Yang et al. [28]). As we will discuss further, this application area is very well suited to a number of the approaches we have mentioned. As such, we believe that Process Mining and Markov-type models can make a very useful contribution to utilising data from Smart Homes, to great advantage.

3. First Passage Time from a Specified Class of States to the Complement of that Class

3.1. The General Case

In what follows, we will consider a notional Smart Home in which we represent human activities of daily living by way of a semi-Markov process, where the states represent changes of (sub-)activity and are typically detected by time-stamped sensor activations, although other mechanisms are possible, e.g., image-processing of a video feed.

We define Class C as a subset of the states of the semi-Markov system and C' its complement in S, the full set of states. Here, S is notionally a set of human activities of daily living carried out in the home and C is a subset of (sub-)activities that are performed to achieve a goal, e.g., making breakfast. Here, S might be the full set of activities of interest in the home, such as "make breakfast", "eat breakfast", "wash dishes", "make dinner", "take shower", and so on, while C is a subset of S, for example, we might have C = "Manage breakfast" = "cook breakfast", "eat breakfast", "wash dishes". On the other hand, $\widetilde{C}$ is an aggregate of all activities in S that are not in C. Initially, we are interested in the distributions of durations in the states of C before the first transition to $\widetilde{C}$.

Then, the sub-transition matrix for class C is $\boldsymbol{A} = \{a_{ij}\}$, where k is the number of states and $i, j = 1, \cdots, k$. Here, we consider $\tilde{C}$ as an aggregate of all states in C' and $\boldsymbol{b}$ is the k-dimensional column vector of transition probabilities from states of C to the aggregated states of $\tilde{C}$. Then, $\boldsymbol{H}(t) = \{h_i(t)\}$ is the $k \times k$ matrix with columns the probability density function's (pdf's) of holding times in each state of C, respectively, and $\boldsymbol{f}(t) = \{f_i(t)\}$ is the column vector of first passage times from each state of C to $\tilde{C}$, respectively. Then,

$$\boldsymbol{f}^*(\boldsymbol{s}) = (\boldsymbol{A}\Box\boldsymbol{H}(\boldsymbol{s}))\boldsymbol{f}^*(\boldsymbol{s}) + \boldsymbol{b}\Box\boldsymbol{h}^*(\boldsymbol{s}), \tag{1}$$

where $\Box$ is the Hadamard product and $\boldsymbol{f}^*(\boldsymbol{s})$ is the moment generating function of $\boldsymbol{f}(t)$, while $\boldsymbol{h}^*(\boldsymbol{s})$ is the moment generating function of $\boldsymbol{h}(t)$. This is a special case of the result in [29]. Regrouping gives

$$\boldsymbol{f}^*(\boldsymbol{s}) = (\boldsymbol{I} - \boldsymbol{A}\Box\boldsymbol{H}^*(\boldsymbol{s}))^{-1}\boldsymbol{b}\Box\boldsymbol{h}^*(\boldsymbol{s}). \tag{2}$$

We note that, in the relation in reference (2), the inverse of the matrix always exists (see [30], P710). We consider an example for $k = 2$, which has been explored previously. Here, states 1 and 2 form the class of interest and we assume that the holding times are exponentially distributed with parameters as in the diagram (Figure 1).

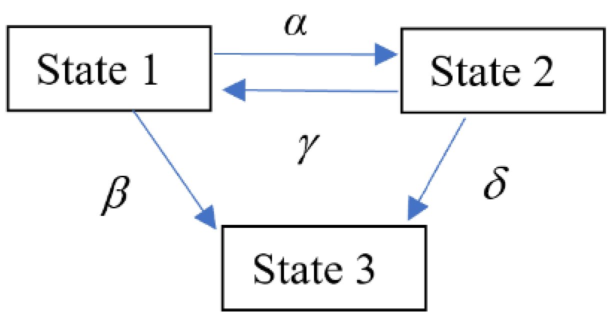

Figure 1. An example where $k = 2$.

In this case, $\boldsymbol{h}(t) = (h_1(t), h_2(t))' = ((\alpha+\beta)e^{-(\alpha+\beta)t}, (\gamma+\delta)e^{-(\gamma+\delta)t})$.

$$\boldsymbol{A} = \begin{pmatrix} 0 & \dfrac{\beta}{\alpha+\beta} \\ \dfrac{\gamma}{\gamma+\delta} & 0 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} \dfrac{\alpha}{\alpha+\beta} \\ \dfrac{\delta}{\gamma+\delta} \end{pmatrix}, \tag{3}$$

$$\boldsymbol{H}(t) = \begin{pmatrix} (\alpha+\beta)e^{-(\alpha+\beta)t} & (\alpha+\beta)e^{-(\alpha+\beta)t} \\ (\gamma+\delta)e^{-(\gamma+\delta)t} & (\gamma+\delta)e^{-(\gamma+\delta)t} \end{pmatrix}, \boldsymbol{h}(t) = \begin{pmatrix} (\alpha+\beta)e^{-(\alpha+\beta)t} \\ (\gamma+\delta)e^{-(\gamma+\delta)t} \end{pmatrix}, \tag{4}$$

and

$$\boldsymbol{H}^*(s) = \begin{pmatrix} (\alpha+\beta)/(\alpha+\beta+s) & (\alpha+\beta)/(\alpha+\beta+s) \\ (\gamma+\delta)/(\gamma+\delta+s) & (\gamma+\delta)/(\gamma+\delta+s) \end{pmatrix}. \tag{5}$$

So,

$$\begin{aligned} \boldsymbol{f}^*(s) &= \begin{pmatrix} 1 & -\beta/(\alpha+\beta+s) \\ -\gamma/(\gamma+\delta+s) & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha/(\alpha+\beta+s) \\ \delta/(\gamma+\delta+s) \end{pmatrix} \\ &= \left(1 - \frac{\beta\gamma}{(\alpha+\beta+s)(\gamma+\delta+s)}\right)^{-1} \begin{pmatrix} 1 & \beta/(\alpha+\beta+s) \\ \gamma/(\gamma+\delta+s) & 1 \end{pmatrix} \begin{pmatrix} \alpha/(\alpha+\beta+s) \\ \delta/(\gamma+\delta+s) \end{pmatrix} \\ &= \left(1 - \frac{\beta\gamma}{(\alpha+\beta+s)(\gamma+\delta+s)}\right)^{-1} \begin{pmatrix} \dfrac{\alpha}{\alpha+\beta+s} + \dfrac{\beta\delta}{(\alpha+\beta+s)(\gamma+\delta+s)} \\ \dfrac{\delta}{\gamma+\delta+s} + \dfrac{\beta\gamma}{(\alpha+\beta+s)(\gamma+\delta+s)} \end{pmatrix}. \end{aligned} \tag{6}$$

Putting $\gamma = 0$ to give a Coxian phase type model with two phases, we get:

$$\boldsymbol{f}^*(\boldsymbol{s}) = \begin{pmatrix} \dfrac{\beta}{\alpha+\beta+s} + \dfrac{\beta\delta}{(\alpha+\beta+s)(\delta+s)} \\ \dfrac{\delta}{\delta+s} \end{pmatrix}, \tag{7}$$

$$\boldsymbol{f}(t) = \begin{pmatrix} \dfrac{\beta\delta}{\alpha+\beta-\delta}e^{-\delta t} + \dfrac{(\alpha+\beta)(\alpha-\delta)}{\alpha+\beta-\delta}e^{-(\alpha+\beta)t} \\ \delta e^{-\delta t} \end{pmatrix}. \tag{8}$$

We note that this equation was previously obtained in McClean [31].

3.2. The Coxian Model

For the general Smart Home case, we consider a Coxian transition matrix structure where transitions to transient states are sequential (Figure 2) and we envisage a phase-type model where there are a number of transient states and exit to a single, or possibly a group of, absorbing state(s). Exit to the absorbing state can occur from any transient case, otherwise, departure from a transient state is to the next transient state in the sequence. A classical phase-type distribution describes a non-negative random variable (usually the duration) generated by a Markov process having a number of transient states (or phases) and a single absorbing state. The duration in each state is therefore exponential. However, in the current case, we envisage a semi-Markov Coxian phase-type model, in particular, a model with mixed gamma durations in each state.

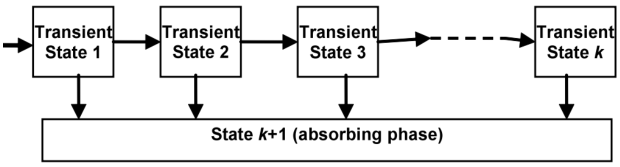

Figure 2. The Coxian phase-type model.

Here, the transient class of states could be all the activities in a visit to the kitchen. Entry to the class is flagged by a door sensor and the class comprises all possible states in the kitchen. For example, a state could be accessing the kettle, which may be accessed for making a cup of coffee (short duration) or for filling a cooking pot (long duration). We therefore assume that the inhabitant enters the kitchen, and then carries out one or more activities in the kitchen class, where an activity duration is described by a mixed gamma duration. In our notional example, these mixture components typically represent latent sub-states, e.g, the short and long kettle durations.

For this semi-Markov Coxian phase-type model, we let a_i be the transition probability from phase S_i to S_{i+1}, for $i = 1, \cdots, k-1$, and let b_i be the probability of transition from S_i to S_{k+1}, for $i = 1, \cdots, k$ where S_{k+1} is the absorbing state. Then, $f_i(t)$ is the pdf of duration in phase S_i, with corresponding generating function $f_i^*(s)$ for $i = 1, \cdots, k$.

In this case:

$$\boldsymbol{A} = \begin{pmatrix} 0 & b_1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & b_2 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & b_{k-1} \\ 0 & 0 & 0 & \cdot & \cdot & 0 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_{k-1} \\ a_k \end{pmatrix}, \tag{9}$$

$$\boldsymbol{H}(t) = \begin{pmatrix} h_1(t) & \cdot & \cdot & \cdot & \cdot & h_1(t) \\ h_2(t) & \cdot & \cdot & \cdot & \cdot & h_2(t) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ h_{k-1}(t) & \cdot & \cdot & \cdot & \cdot & h_{k-1}(t) \\ h_k(t) & \cdot & \cdot & \cdot & \cdot & h_k(t) \end{pmatrix}, \text{and } \boldsymbol{h}(t) = \begin{pmatrix} h_1(t) \\ h_2(t) \\ \cdot \\ \cdot \\ \cdot \\ h_{k-1}(t) \\ h_k(t) \end{pmatrix}. \tag{10}$$

Then, we have:

$$\begin{aligned} \boldsymbol{f}^*(\boldsymbol{s}) &= \begin{pmatrix} 1 & -b_1 h_1^*(s) & 0 & \cdot & \cdot & 0 \\ 0 & 1 & -b_2 f_2^*(s) & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & -b_{k-1} f_{k-1}^*(s) \\ 0 & 0 & 0 & \cdot & \cdot & 1 \end{pmatrix}^{-1} \begin{pmatrix} a_1 h_1(t) \\ a_2 h_2(t) \\ \cdot \\ \cdot \\ \cdot \\ a_{k-1} h_{k-1}(t) \\ a_k h_k(t) \end{pmatrix} \\ &= \begin{pmatrix} 1 & b_1 h_1^*(s) & b_1 b_2 h_1^*(s) h_2^*(s) & \cdot & \cdot & b_1 b_2 \cdots b_{k-1} h_1^*(s) \cdots h_{k-1}^*(s) \\ 0 & 1 & b_2 h_2^*(s) & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & b_{k-1} h_{k-1}^*(s) \\ 0 & 0 & 0 & \cdot & \cdot & 1 \end{pmatrix} \begin{pmatrix} a_1 h_1(t) \\ a_2 h_2(t) \\ \cdot \\ \cdot \\ \cdot \\ a_{k-1} h_{k-1}(t) \\ a_k h_k(t) \end{pmatrix}. \end{aligned} \tag{11}$$

So,

$$f_i^*(s) = \sum_{j=i}^{k} \alpha_j h_j * (s) \prod_{r=j}^{k-1} b_r h_r^*(s), i = 1, \cdots, k. \tag{12}$$

For $k = 2$ and

$$\boldsymbol{H}^*(s) = \begin{pmatrix} (\alpha+\beta)/(\alpha+\beta+s) & (\alpha+\beta)/(\alpha+\beta+s) \\ (\gamma+\delta)/(\gamma+\delta+s) & (\gamma+\delta)/(\gamma+\delta+s) \end{pmatrix}. \tag{13}$$

We therefore obtain the previous result for a Coxian Markov phase-type model. If the initial state entered in C is unknown we use a probability vector $\pi = \{\pi_i\}$, where π_i is the probability of admission to state $S_i, i = 1, \cdots, k$. In this case, the transform of the unconditional pdf of duration in C is:

$$g^*(s) = \boldsymbol{\pi} \boldsymbol{f}^*(s). \tag{14}$$

For mixed gamma holding times and a semi-Markov Coxian phase-type model we have:

$$h_i(t) = \sum_{r=1}^{R_i} \pi_{ir} \frac{1}{\Gamma(\alpha_{ir}) \beta_{ir}^{\alpha_{ir}}} t^{\alpha_{ir}-1} e^{-\frac{t}{\beta_{ir}}}, t \geq 0, \tag{15}$$

and

$$h_i^*(s) = \sum_{r=1}^{R_i} \pi_{ir} (1 + \beta_{ir} s)^{-\alpha_{ir}}, i = 1, \cdots, k. \tag{16}$$

So, we need to compute $f_i^*(s)$, which is a mixture of terms with coefficients $\alpha_j \prod_{r=j}^{k-1} b_r$ of the convolution of exponentials. Thus, the time to exit from class C is prob(exit from 1) $\times$ (mixed gamma 1) + prob(exit from 2) $\times$ (mixed gamma 1 $\bigoplus$ mixed gamma 2) + $\cdots$. We can think of this as a mixture of terms relating to exits from successive states, where the corresponding duration is a convolution of times in all the states involved in this transition. We regard such terms as durations of time to exit from C, following entry to the hidden semi-Markov model, where the gamma mixture components are hidden states. Here, $\bigoplus$ denotes convolution (sum of two independent random variables). We write

$$f_i^*(s) = \sum_{j=i}^{k} a_j (\prod_{r=j}^{k-1} b_r h_j^*(s) h_r^*(s)), i = 1, \cdots, k. \tag{17}$$

For example, when $k = 2$:

$$f_1^*(s) = a_1 h_1^*(s) + a_2 b_1 h_1^*(s) h_2^*(s), \tag{18}$$

$$f_2^*(s) = a_2 h_2^*(s). \tag{19}$$

So, for entrance to state 1 and exit from state 1, we have a mixture of durations with an exit from state 1 directly, with probability a_1 and from state 1 to state 2, and then state 2 to exit, with probability $a_2 b_1$. Inverting the transforms, the corresponding durations have pdfs $h_1(t)$ and the convolution of $h_1(t)$ and $h_2(t)$, respectively. Here, $a_1 + b_1 = 1$ and $a_2 = 1$. For entrance to state 2, exit can only occur directly from state 2, and the duration has pdf $h_2(t)$.

In general, of course, different sequences may be formed from the same activities executed in different orders. Here, we have developed the theory for commonly occurring sequences and those potentially of particular interest, e.g., some sequences might indicate that the patient is not managing their activities successfully. In further work, we plan to further consider these aspects, particularly focusing on the aggregation of states, as an aid to understanding, and improved computational efficiency [20].

3.3. Gamma Convolutions

The sum of a number of independently and identically distributed random variables is called their convolution. The exact expressions for the convolutions of n gamma pdfs and cumulative distribution functions (cdfs) are complex and difficult to compute, but are of considerable practical interest [32]. In particular, some authors have proposed that a simple gamma distribution can give a good approximation, especially for convolutions with $n > 2$, (Stewart et al. [33], Covo and Elalouf [34]). On the other hand, approximations and high-performance algorithms have also been developed and are implemented in coga (r-project-org) [35]. In particular, exact solutions have been implemented for $n = 2$ [36] or $n > 2$ [37], while efficient approximate solutions have originated from Barnabani [38].

In our current context, some interesting work has been carried out by Guenther et al. [20] (and in other papers), who describe the modelling and computation of passage-time distributions between groups and aggregates of states.

4. Experiments on Artificially Generated Smart Home Data

In this Section, we generate mixed gamma data representing typical behaviours of each of two activities, in each case consisting of two possible sub-activities, where each sub-activity has two or three possible mixed gamma components. The chosen activities are Toileting and Breakfast, respectively. In each case, we simulate an activity instance by generating two mixed gamma sub-activity instance durations, where we assume these two sub-activities are followed sequentially. For each activity, we then simulate 10,000 representative activity instance durations as the sum of the corresponding mixed gamma sub-activity instance durations.

Using these simulated data, we carry out a series of experiments that, for each of the two activities, compare the empirical activity duration distribution with the convolution model, computed as the convolution of the two sub-activity durations, and the mixed gamma model fitted to the actual (simulated) data for the activity duration. As discussed in the previous section, the latter has previously been shown to give a good approximation to such data [5].

4.1. Datasets

Toileting To generate data that simulates a scenario in which the inhabitant first performs 'use toilet' and then 'wash hands', we suppose the inhabitant has three patterns when using the toilet, requiring 1, 5, and 10 min on average and two duration patterns for washing hands, taking on average 20 and 40 s each. Such patterns are created by gamma distributions with parameters presented in Table 1. The sequential execution of the two activities is referred to as 'Toileting'.

Table 1. Toileting: real gamma distribution parameters and the parameters fitted by the model-based method.

		Real Parameters			Estimated Parameters			
	Activity	**Shape**	**Scale**	**Mean**	**Mixing**	**Shape**	**Scale**	**Mean**
Toileting	Use toilet	3	20	60	0.795	3.076	19.168	58.961
		50	6	300	0.165	43.040	6.946	298.963
		200	3	600	0.040	201.440	2.970	598.284
	Wash hands	20	1	20	0.701	21.257	0.939	19.951
		40	1	40	0.299	46.666	0.861	40.170

Breakfast To simulate a 'Breakfast' scenario including two consecutive activities, 'eat food' (three patterns taking 10, 20, and 40 min on average) and 'wash dishes' (two patterns, requiring 5 and 10 min), gamma distributions are employed with the parameters shown in Table 2.

Table 2. Breakfast: real gamma distribution parameters and the parameters fitted by the model-based method.

		Real Parameters			Mathematical Model			
	Activity	**Shape**	**Scale**	**Mean**	**Mixing**	**Shape**	**Scale**	**Mean**
Breakfast	Eat food	60.00	10.00	600.00	0.69	57.12	10.54	601.86
		60.00	20.00	1200.00	0.24	66.77	18.07	1206.47
		120.00	20.00	2400.00	0.07	121.83	19.69	2398.62
	Wash dishes	10.00	30.00	300.00	0.74	13.35	20.94	279.58
		30.00	20.00	600.00	0.26	29.78	19.08	568.22

4.2. Goodness of Fit of the Model-Based Method for 'Toileting'

Regarding 'Toileting', Table 1 displays real gamma parameters for generating the activities 'use toilet' and 'wash hands', alongside the model-based method estimated from gamma mixture models. Additionally, Table 3 shows the gamma mixture model parameters obtained from the data-driven-based method. Such mixture models are fitted by maximizing the likelihood using the Nelder–Mead simplex algorithm [5].

Table 3. Toileting: parameters fitted by the data-driven-based method.

	Mixing	**Shape**	**Scale**	**Mean**
Toileting	0.796	6.086	13.402	81.561
	0.164	51.745	6.529	337.822
	0.040	189.917	3.376	641.162

Figure 3 demonstrates the model-based method for estimating the duration distribution of the 'toileting' activity (red line). This is achieved by the convolution of estimated gamma mixture models for 'use toilet' (black line) and 'wash hands' (orange line) in Figure 4. The green line represents the distribution that is estimated by the data-driven method. In other words, instead of fitting 'toileting' data directly (the data-driven method), the model-based approach first separately fits the duration distributions of the two sub-activities and then combines them to estimate the 'toileting' duration distribution. The model-based method is therefore suitable for scenarios in which the distribution of sub-activities is known, without the need for additional computational resources.

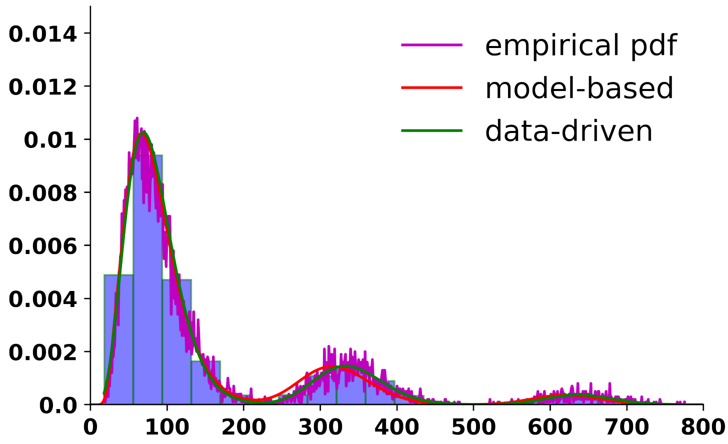

Figure 3. Toileting: pdf of the model- and data-driven-based methods.

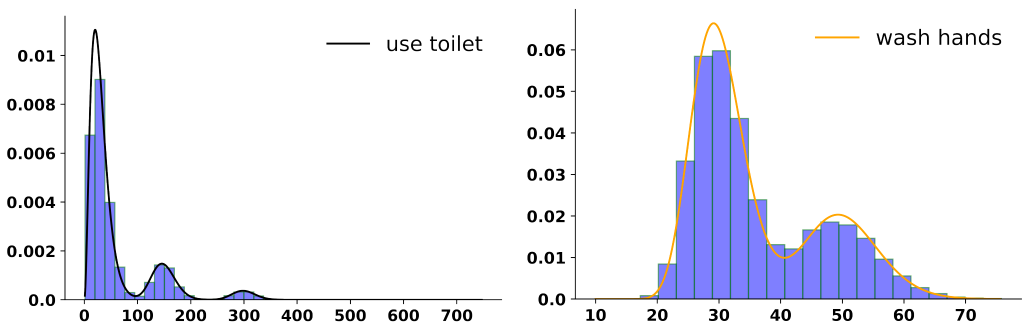

Figure 4. Toileting: pdf of activities 'use toilet' and 'wash hands'.

The KS test [39] is then employed to assess the goodness of fit of the model-based method, with the results presented in Table 4 and Figure 5. We note that the Anderson–Darling test could have been used here as a more powerful alternative to the Kolmogorov–Smirnov test. In summary, both the model-based and data-driven methods effectively represent the original data distribution, as indicated by p-values of 0.219 (model-based) and 0.181 (data-driven), which exceed the predetermined significance level of 0.05. Furthermore, the estimated duration distributions by the two methods is extremely similarly with a p-value of 0.610; the same result can be seen in Figure 5.

Table 4. Toileting: the goodness of data fitting by using the KS test.

Sample 1	Sample 2	Statistic	p-Value
Model-based	Real data	0.047	0.219
Data-driven-based	Real data	0.049	0.181
Model-based	Data-driven-based	0.034	0.610

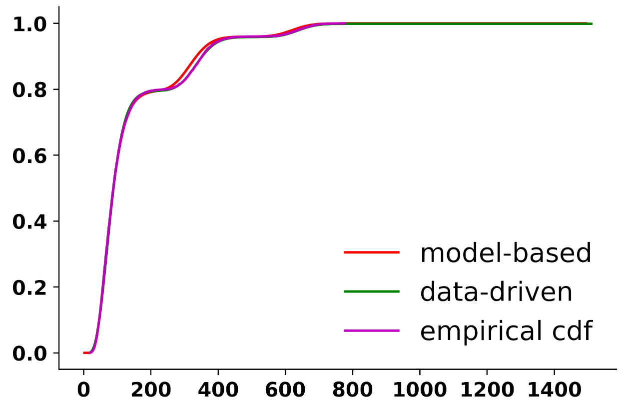

Figure 5. Toileting: cdf of the model- and data-driven-based methods.

4.3. Goodness of Fit of the Model-Based Method for 'Breakfast'

Similarly, parameters to generate the data of 'eating food' and 'washing dishes' and parameters fitted by the model-based method are presented in Table 2. Table 5 displays the data-driven method estimated gamma mixture model parameters. The fitted pdf curves are shown in Figure 6.

Table 5. Breakfast: parameters fitted by the data-driven-based method.

	Mixing	Shape	Scale	Mean
Breakfast	0.698	51.105	17.714	905.274
	0.233	59.564	27.811	1656.523
	0.069	145.588	20.580	2996.230

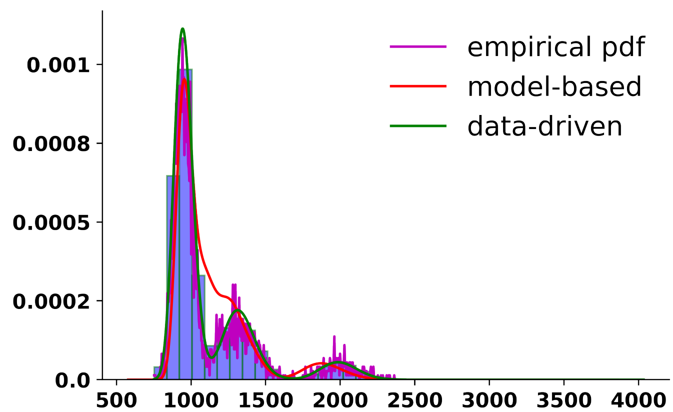

Figure 6. Breakfast: the pdf of model- and data-driven-based methods.

The pdf of activities 'eat food' and 'wash dishes' is given in Figure 7. The KS test results are given in Table 6 and Figure 8. The data-driven method fits the data the best with the highest p-value of 0.523 compared to 0.276 of the model-based method. We highlight that, although the performance of the data-driven method in data fitting outperforms the model-based method, they show less distribution difference with a p-value of 0.821.

Table 6. Breakfast: the goodness of data fitting by using the KS test.

Sample 1	Sample 2	Statistic	p-Value
Model-based	Real data	0.180	0.276
Data-driven-based	Real data	0.147	0.523
Model-based	Data-driven-based	0.114	0.821

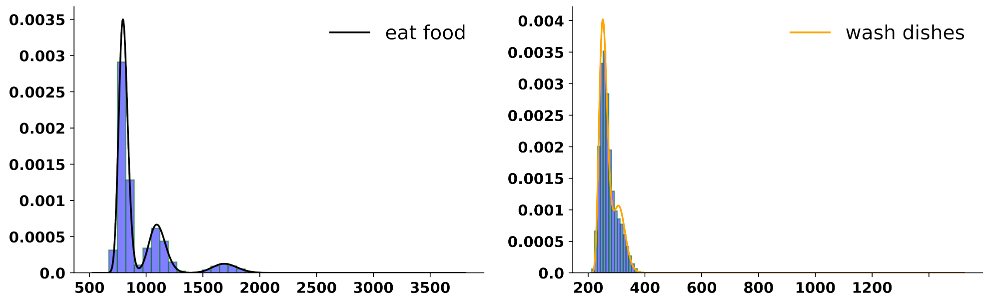

Figure 7. Toileting: pdf of activities 'eat food' and 'wash dishes'.

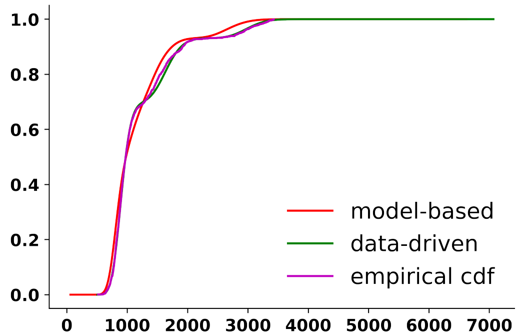

Figure 8. Breakfast: activites of the model- and data-driven-based methods.

5. Example Using Artificially Generated Data for Toileting to Assess Deterioration

Using the artificially generated data described in Section 4.2, we will illustrate the model described in Equations (17) to (19) and discussed in Section 3.3, for gamma convolution. We therefore consider a toileting model as described in the real data part of Table 1. However, we now use a mixed gamma with the given parameters for the two activities 'use toilet' and 'wash hands' but, unlike in Section 4, we assume that 'wash hands' is omitted with probability b. This might be of interest, for example, when assessing deteriorating cognitive or kidney function in an elderly patient.

In this case, the pdf for time spent on the toileting activity becomes a mixture of the pdf of time spent using the toilet, with probability b, and the convolution of time spent on 'use toilet' and 'wash hands', with probability $1 - b$. In Figure 9, we see the pdf of this distribution plotted as a function of b, which shows that, with the increase of b (higher probability to omit washing hands), the length of duration to complete activity 'toileting' is decreasing. Here, b may be regarded as a proxy for impairment, and we can therefore use a statistical test, such as likelihood ratio, to test for drift in b, signalling possible decline. Similarly, Figure 10 demonstrates the pdf that 'wash dishes' is omitted with probability b.

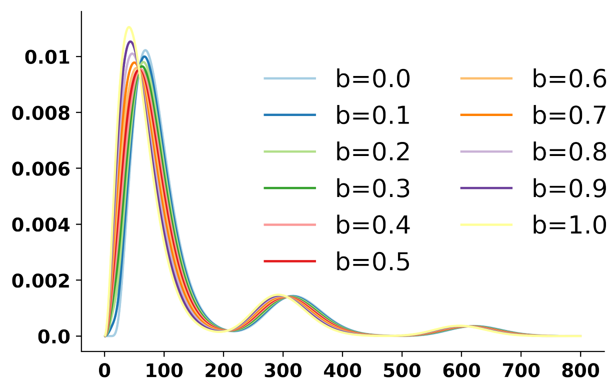

Figure 9. Toileting: pdf of the model-based method given that "wash hands" is omitted with probability b.

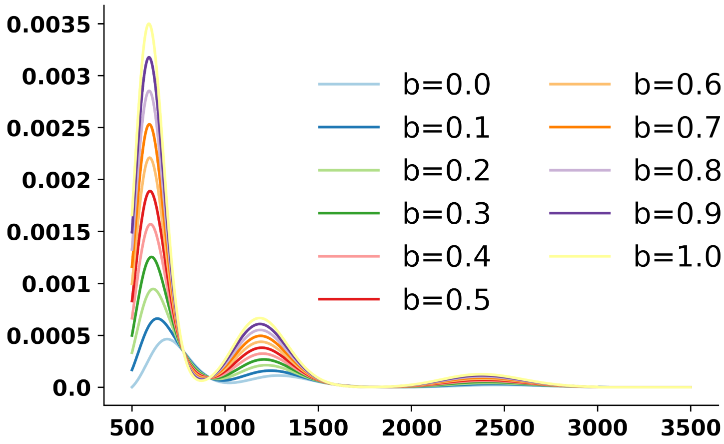

Figure 10. Breakfast: pdf of the model-based method given that "wash dishes" is omitted with probability b.

6. Experiments on Real-World Smart Home Data

Datasets

This section utilizes data pertaining to the daily activities of a 26-year-old individual living in a three-room apartment, as gathered by van Kasteren [6]. The data comprise 14 binary state-change sensors, representing the seven activities shown in Table 7 with the Markov transition probabilities shown in Table 8. These sensors were strategically positioned in various areas such as cupboards, refrigerators, and doors, and were left unattended for a duration of 28 days. Additional insights into the apartment's layout and sensor placement can be found in [6,40].

Table 7. Events in the activity, sensor, and location level [28].

Level	Events	#Events
Activity	'Get Drink', 'Go to Bed', 'Leave House', 'Prepare Breakfast', 'Prepare Dinner', 'Take Shower', 'Use Toilet'	7
Sensor	'Microwave', 'Hall Toilet Door', 'Hall Bathroom Door', 'Cups Cupboard', 'Fridge','Plates Cupboard', 'Front Door', 'Dish Washer', 'Toilet Flush', 'Freezer', 'Pans Cupboard', 'Washing Machine', 'Groceries Cupboard', 'Hall Bedroom Door'	14
Location	'Outdoor', 'Bedroom', 'Kitchen', 'Washroom'	4

Table 8. Transition probabilities between activities [28].

	GD	GTB	LH	PB	PD	TS	UT
Get Drink (GD)	0.15	0	0	0.05	0	0	0.8
Go to Bed (GTB)	0	0	0	0.042	0	0	0.958
Leave House (LH)	0.091	0.03	0.03	0	0.091	0.061	0.697
Prepare Breakfast (PB)	0.05	0	0	0	0	0.55	0.4
Prepare Dinner (PD)	0.6	0	0.1	0	0	0	0.3
Take Shower (TS)	0	0	0.913	0	0.043	0	0.043
Use Toilet (UT)	0.061	0.193	0.096	0.158	0.053	0.088	0.351

Since all seven activities have a high transition probability to activity 'use toilet', these transitions are of interest, with 16, 23, 23, 8, 3, 1, and 40 samples, respectively, starting from 'get drink', 'use toilet', 'go to bed', 'leave house', 'prepare breakfast', 'prepare dinner', 'take shower' and 'use toilet'. Due to the limited sample size and the lack of a representative, several transitions are ignored and, finally, data from 'go to bed' to 'use toilet' and 'use toilet' to 'use toilet' are used as a case study. Their density distributions are shown in Figures 11 and 12. Figures 13 and 14 demonstrate their active curves with the KS test results in Tables 9 and 10.

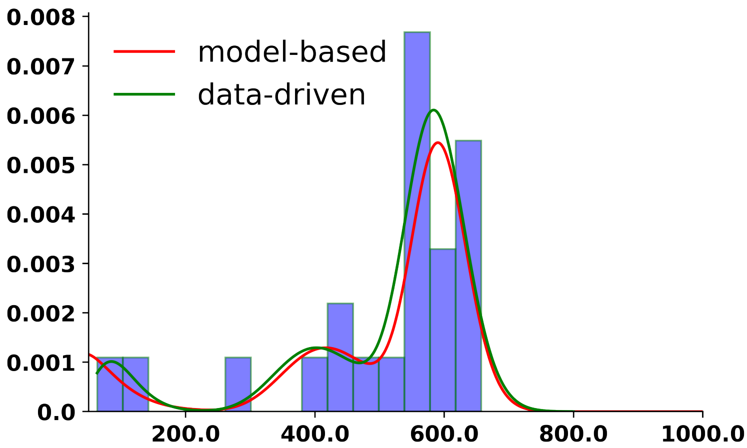

Figure 11. From 'go to bed' to 'use toilet': the pdf of model- and data-driven-based methods.

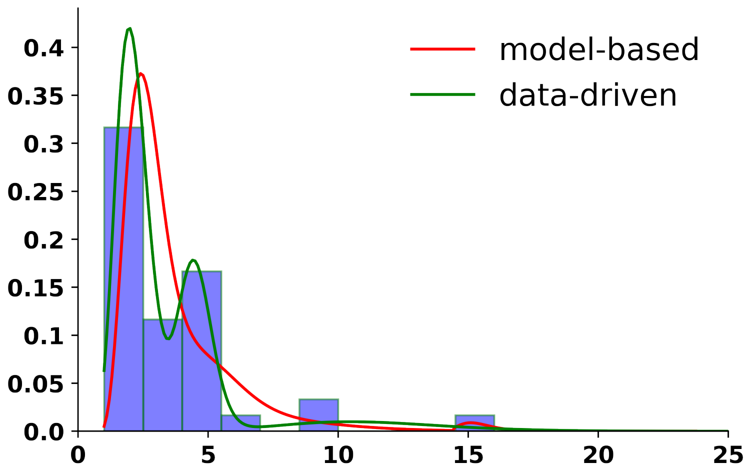

Figure 12. From 'use toilet' to 'use toilet': the pdf of model- and data-driven-based methods.

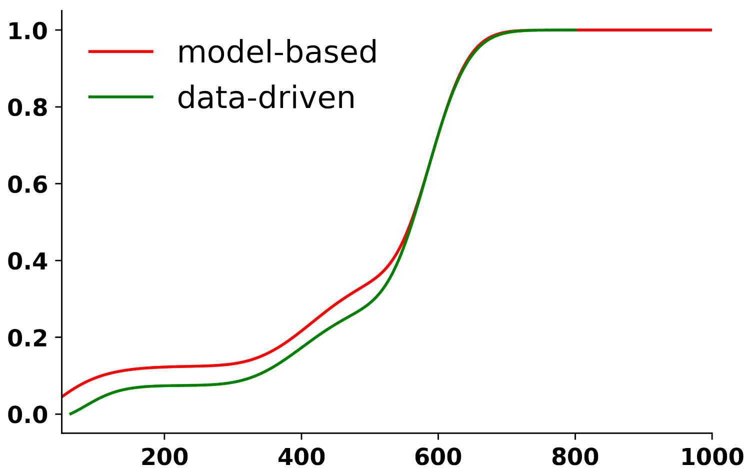

Figure 13. From 'go to bed' to 'use toilet': cdf of the model- and data-driven-based methods.

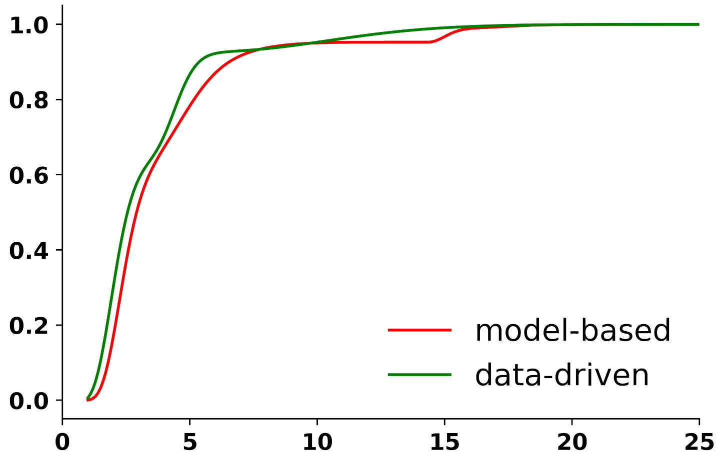

Figure 14. From 'use toilet' to 'use toilet': cdf of the model- and data-driven-based methods.

Table 9. From 'go to bed' to 'use toilet': the goodness of data fitting by using the KS test.

Sample 1	Sample 2	Statistic	*p*-Value
Model-based	Real data	0.215	0.685
Data-driven-based	Real data	0.205	0.705
Model-based	Data-driven-based	0.248	0.532

Table 10. From 'use toilet' to 'use toilet': the goodness of data fitting by using the KS test.

Sample 1	Sample 2	Statistic	*p*-Value
Model-based	Real data	0.582	0.217
Data-driven-based	Real data	0.481	0.384
Model-based	Data-driven-based	0.388	0.632

7. Conclusions and Further Work

Smart Homes and sensorized environments are becoming increasingly prevalent and can provide a useful source of support for disabled inhabitants. In this paper, we have developed suitable equations for semi-Markov models with mixed gamma duration distribution; these have been implemented and evaluated using simulated and real data and have been shown to produce promising results. Such durations can be interpreted as KPIs (Key Performance Indicators) and can be used to assess changes represented by anomalies/outliers or concept drift (changes in duration distribution).

In further work, we plan to extend the research to different Markov-type models, such as: (i) nth order models, where the Markov property no longer hold in all cases, and there may be situations when the next (sub-)activity depends on previous states; (ii) non-homogeneous Markov models [17], where the model parameters vary with time, e.g., day-time and night-time, or week-days and weekends, might have different activity patterns that are reflected in model parameters; and (iii) the aggregation of semi-Markov models with a view to providing key information at a suitable level of detail and to improve computational efficiency.

In general, such models can be used in healthcare situations in several important ways. Firstly, they can provide assistance for the patient, by reminding and prompting when a (sub-)activity has been ongoing for too long or omitted. By providing such services in the home, there are potentially huge financial benefits, in terms of reduced hospital and nursing home admissions, as well as improved quality-of-life for the patient and their carers. Secondly, by providing information to the medical professionals and identifying possible deterioration in the patient's condition, there can be improved monitoring, with the potential for more timely and effective treatment and reductions in hospital (re)admissions.

Finally, by having a suitable deployment of such Smart Home technologies, the care of relevant patients can be better managed, resourced, and planned. However, although there has been a considerable amount of research and some deployment of such solutions, there is still a lot of work that needs to be undertaken before the approach is fully fit-for-purpose and widely adopted ([41,42]).

Author Contributions: Conceptualization, S.M.; Methodology, S.M.; Software, L.Y.; Validation, L.Y.; Formal analysis, L.Y.; Investigation, L.Y.; Resources, S.M. and L.Y.; Data curation, L.Y.; Writing—original draft, S.M. and L.Y.; Writing—review & editing, S.M. and L.Y.; Supervision, S.M.; Project administration, S.M.; Funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The python script to generate the artificial data used in this paper is shared at https://github.com/LingkaiYang/SmatHomeArtificialData/tree/main. The real-world data is publicly available at https://ailab.wsu.edu/mavhome/research.html.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. 2022. Available online: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed on 13 November 2022).
2. Joe, E.; Ringman, J.M. Cognitive symptoms of Alzheimer's disease: Clinical management and prevention. *BMJ* **2019**, *367*, l6217. [CrossRef] [PubMed]
3. Van Der Aalst, W.; Adriansyah, A.; De Medeiros, A.K.A.; Arcieri, F.; Baier, T.; Blickle, T.; Bose, J.C.; Van Den Brand, P.; Brandtjen, R.; Buijs, J.; et al. Process mining manifesto. In Proceedings of the Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, 29 August 2011, Revised Selected Papers, Part I 9. Springer: Berlin/Heidelberg, Germany, 2012; pp. 169–194.
4. Wang, X.; Liu, J.; Moore, S.J.; Nugent, C.D.; Xu, Y. A behavioural hierarchical analysis framework in a smart home: Integrating HMM and probabilistic model checking. *Inf. Fusion* **2023**, *95*, 275–292. [CrossRef]
5. Yang, L.; McClean, S.; Donnelly, M.; Burke, K.; Khan, K. A multi-components approach to monitoring process structure and customer behaviour concept drift. *Expert Syst. Appl.* **2022**, *210*, 118533. [CrossRef]
6. Van Kasteren, T.; Noulas, A.; Englebienne, G.; Kröse, B. Accurate activity recognition in a home setting. In Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Republic of Korea, 21–24 September 2008; pp. 1–9.
7. Lorenz, R.; Senoner, J.; Sihn, W.; Netland, T. Using process mining to improve productivity in make-to-stock manufacturing. *Int. J. Prod. Res.* **2021**, *59*, 4869–4880. [CrossRef]
8. Mahendrawathi, E.; Astuti, H.M.; Nastiti, A. Analysis of customer fulfilment with process mining: A case study in a telecommunication company. *Procedia Comput. Sci.* **2015**, *72*, 588–596. [CrossRef]
9. Rojas, E.; Munoz-Gama, J.; Sepúlveda, M.; Capurro, D. Process mining in healthcare: A literature review. *J. Biomed. Inform.* **2016**, *61*, 224–236. [CrossRef] [PubMed]
10. McClean, S.; Barton, M.; Garg, L.; Fullerton, K. A modeling framework that combines markov models and discrete-event simulation for stroke patient care. *ACM Trans. Model. Comput. Simul. (TOMACS)* **2011**, *21*, 1–26. [CrossRef]
11. Georgiou, A.C.; Papadopoulou, A.; Kolias, P.; Palikrousis, H.; Farmakioti, E. On state occupancies, first passage times and duration in non-homogeneous semi-Markov chains. *Mathematics* **2021**, *9*, 1745. [CrossRef]
12. Vassiliou, P.C.; Tsakiridou, H.L. Asymptotic behavior of first passage probabilities in the perturbed non-homogeneous semi-Markov systems. *Commun. Stat.-Theory Methods* **2004**, *33*, 651–679. [CrossRef]
13. Garg, L.; McClean, S.I.; Barton, M.; Meenan, B.J.; Fullerton, K. Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. *IEEE Trans. Syst. Man Cybern. Part A Syst. Humans* **2012**, *42*, 1332–1345. [CrossRef]
14. Gillespie, J.; McClean, S.; Garg, L.; Barton, M.; Scotney, B.; Fullerton, K. A multi-phase DES modelling framework for patient-centred care. *J. Oper. Res. Soc.* **2016**, *67*, 1239–1249. [CrossRef]
15. Vassiliou, P.C.; Moysiadis, T. Inhomogeneous Markov Systems of High Order. *Methodol. Comput. Appl. Probab.* **2010**, *12*, 271–292. [CrossRef]
16. Moore, S.J.; Nugent, C.D.; Zhang, S.; Cleland, I.; Sani, S.; Healing, A. A markov model to detect sensor failure in IoT environments. In Proceedings of the 2020 IEEE World Congress on Services (SERVICES), Beijing, China, 18–23 October 2020; pp. 13–16.
17. Vassiliou, P.C. *Non-Homogeneous Markov Chains and Systems: Theory and Applications*; CRC Press: Boca Raton, FL, USA, 2022.
18. Youngblood, G.M.; Cook, D.J. Data mining for hierarchical model creation. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2007**, *37*, 561–572. [CrossRef]

19. Asghari, P.; Soleimani, E.; Nazerfard, E. Online human activity recognition employing hierarchical hidden Markov models. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 1141–1152. [CrossRef]
20. Guenther, M.C.; Dingle, N.J.; Bradley, J.T.; Knottenbelt, W.J. Passage-time computation and aggregation strategies for large semi-Markov processes. *Perform. Eval.* **2011**, *68*, 221–236. [CrossRef]
21. Marshall, A.; McClean, S. Conditional phase-type distributions for modelling patient length of stay in hospital. *Int. Trans. Oper. Res.* **2003**, *10*, 565–576. [CrossRef]
22. Jones, B.; McClean, S.; Stanford, D. Modelling mortality and discharge of hospitalized stroke patients using a phase-type recovery model. *Health Care Manag. Sci.* **2019**, *22*, 570–588. [CrossRef]
23. Papadopoulou, A.; Vassiliou, P.C. On the variances and convariances of the duration state sizes of semi-Markov systems. *Commun. Stat. Theory Methods* **2014**, *43*, 1470–1483. [CrossRef]
24. Verbeken, B.; Guerry, M. Discrete Time Hybrid Semi-Markov Models in Manpower Planning. *Mathematics* **2021**, *9*, 1681. [CrossRef]
25. McClean, S.; Montgomery, E.; Ugwuowo, F. Non-homogeneous continuous-time Markov and semi-Markov manpower models. *Appl. Stoch. Model. Bus. Ind.* **1997**, *13*, 191–198. [CrossRef]
26. Shaw, B.; Marshall, A. Modelling the flow of congestive heart failure patients through a hospital system. *J. Oper. Res. Soc.* **2007**, *58*, 212–218. [CrossRef]
27. Chen, Z.; Zhang, S.; McClean, S.; Hart, F.; Milliken, M.; Allan, B.; Kegel, I. Process Mining IPTV Customer Eye Gaze Movement Using Discrete-Time Markov Chains. *Algorithms* **2023**, *16*, 82. [CrossRef]
28. Yang, L.; McClean, S.; Bashar, A.; Moore, S.; Tariq, Z. Using Semi-Markov Models to Identify Long Holding Times of Activities of Daily Living in Smart Homes. Available online: https://faculty.pmu.edu.sa/Admin/ckEditor_Uploads/133318033358462860_Camera_Ready_PROMISE_2023_draft.pdf (accessed on 20 November 2023).
29. Howard, R.A. *Dynamic Probabilistic Systems: Markov Models*; Courier Corporation: North Chelmsford, MA, USA, 2012; Volume 1.
30. Howard, R.A. *Dynamic Probabilistic Systems: Semi-Markov and Decision Processes*; Dover Publications: Mineola, NY, USA, 2007; Volume 2.
31. McClean, S. The two-stage model of personnel behaviour. *J. R. Stat. Soc. Ser. A Gen.* **1976**, *139*, 205–217. [CrossRef]
32. Khaledi, B.E.; Kochar, S. A review on convolutions of gamma random variables. In *Stochastic Orders in Reliability and Risk: In Honor of Professor Moshe Shaked*; Springer: New York, NY, USA, 2013; pp. 199–217.
33. Stewart, T.; Strijbosch, L.; Moors, H.; Batenburg, P.V. A Simple Approximation to the Convolution of Gamma Distributions. 2007. Available online: https://pure.uvt.nl/ws/portalfiles/portal/858139/dp2007-70.pdf (accessed on 20 November 2023).
34. Covo, S.; Elalouf, A. A novel single-gamma approximation to the sum of independent gamma variables, and a generalization to infinitely divisible distributions. *Electron. J. Stat.* **2014**, *8*, 894–926. [CrossRef]
35. Hu, C.; Pozdnyakov, V.; Yan, J. Density and distribution evaluation for convolution of independent gamma variables. *Comput. Stat.* **2020**, *35*, 327–342. [CrossRef]
36. Mathai, A.M. Storage capacity of a dam with gamma type inputs. *Ann. Inst. Stat. Math.* **1982**, *34*, 591–597. [CrossRef]
37. Moschopoulos, P.G. The distribution of the sum of independent gamma random variables. *Ann. Inst. Stat. Math.* **1985**, *37*, 541–544. [CrossRef]
38. Barnabani, M. An approximation to the convolution of gamma distributions. *Commun. Stat. Simul. Comput.* **2017**, *46*, 331–343. [CrossRef]
39. Yang, L.; McClean, S.; Donnelly, M.; Burke, K.; Khan, K. Process Duration Modelling and Concept Drift Detection for Business Process Mining. In Proceedings of the 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), Atlanta, GA, USA, 18–21 October 2021; pp. 653–658.
40. Moore, S.; Nugent, C.D.; Cleland, I.; Zhang, S. Impact analysis of erroneous data on IoT reliability. In Proceedings of the 2019 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI, Leicester, UK, 19–23 August 2019; pp. 1908–1915.
41. Chimamiwa, G.; Giaretta, A.; Alirezaie, M.; Pecora, F.; Loutfi, A. Are Smart Homes Adequate for Older Adults with Dementia? *Sensors* **2022**, *22*, 4254. [CrossRef]
42. Lee, L.N.; Kim, M.J. A critical review of smart residential environments for older adults with a focus on pleasurable experience. *Front. Psychol.* **2020**, *10*, 3080. [CrossRef]