Residential activity pattern modelling through stochastic chains of variable memory length

José Luis Ramírez-Mendiola[a,*], Philipp Grünewald[a], Nick Eyre[a]

[a] *Environmental Change Institute, University of Oxford, UK*

Abstract

Residential activity modelling has attracted considerable attention over the last years. This is particularly due to the fact that residential energy demand loads are highly dependent on the activity patterns of the household. Therefore, activity models are being increasingly used to underpin high-resolution energy demand models. This paper details the implementation of a new methodology for the analysis of empirical activity data that allows for the identification of characteristic behavioural patterns within them. The identified patterns are then used as the basis for the construction of a high-resolution residential user activity model. The model attempts to capture the statistical characteristics of the empirical data in the form of a stochastic process with memory of variable length. The proposed model is compared to a model based on the predominant first-order Markov chain approach. In addition to the modelling approach, a new metric for assessing the quality of activity sequences simulations is proposed. Given the amount of empirical data contained in any of the individual time-use datasets currently available, it would appear that the performance improvement over the predominant first-order Markov chain approach is modest. However, the validation results show that the proposed approach has the potential for broadening our understanding of the scheduling of activities in people's day-to-day lives and how this relates to the observed variability in both activity and energy consumption patterns.

Keywords: Activity modelling, Residential activity patterns, Time-use, Stochastic modelling, Energy demand modelling

1. Introduction

The past couple of decades have seen a substantial increase in the interest in the branch of energy research that looks at the dynamics of energy demand. In particular, considerable efforts have been made to include the associated behavioural aspects into the modelling of residential demand loads through the simulation of people's activity patterns [27, 28].

In the context of a transition towards a more decentralised electricity generation system, in which generation technologies with highly variable outputs are expected to play a major role, it is clear that a complete reworking of the relationship between supply and demand

*Corresponding author:
Email address: joseluis.ramirezmendiola@ouce.ox.ac.uk (José Luis Ramírez-Mendiola)

will be needed. Rather than solely expect the supply to adjust to the variations in demand loads, the active management of the dynamics of such demand loads is increasingly seen as a key component of future low-carbon systems [29]. In addition, it is acknowledged that experimenting with this active demand management in ways beyond conventional approaches is necessary in order to explore the potential for energy efficiency enhancements and to better assess the feasibility and problems of the proposed management strategies [8, 18, 26]. However, in order to effectively explore these issues, more needs to be known about the dynamics and temporalities of the underlying behavioural patterns that give rise to the observed demand loads [30].

Central to the approach that places activities at the base of our understanding of the dynamics of the observed residential demand loads is the idea that the way activities are ordered in time determines the timing of demand [24, 28]. Therefore, a better understanding of the scheduling of activities in people's day-to-day lives and how this relates to the observed variability is of the essence. However, up until now, activity modelling has largely focused on reproducing the overall features of mean activity profiles which are averaged over large numbers of individual daily activity schedules. While attempts have been made to further unravel the complexity of activity scheduling [1, 16, 27, 32], current approaches to the simulation of activity patterns still pose some serious limitations when it comes to learning more about daily activity scheduling.

In particular, the probabilistic methods used in these approaches produce an oversimplified representation of the dynamics observed in empirical activity data. Empirical activity schedules are observed to evolve in numerous ways. Moreover, the evolution of these daily activity schedules is arguably ruled by a self-adaptation process in which the parts of the schedule that are already in place, or a subset of them, play an important role in determining the subsequent steps in the evolution process. These complexities manifest as a wide behavioural diversity, and a probabilistic model aiming to produce a close representation of this diversity would have to attempt to capture such complexity.

However, the stochastic processes represented by current approaches are first-order approximations to the behavioural diversity observed in empirical activity data. Thus, the diversity in the range of behaviours that is observed in the simulated output of one of these first-order models results naturally in a narrow normal distribution around a mean activity pattern. Therefore, the behavioural diversity that results from simulating the observed dynamics by means of such simple processes is equally narrow. The simplicity of the formulation of these first-order models might increase their appeal. However, their simplifying assumptions mask the complexity of the dynamics of the observed activity patterns.

If the ultimate goal is to develop a better understanding of the scheduling of activities in

people's day-to-day lives, an approach to the simulation of daily activity patterns that takes account of the observed complexities would be greatly advantageous and indeed necessary.

In this paper we introduce a novel approach to the analysis and simulation of activity data based on variable memory length stochastic models. The idea behind the development of this approach is addressing the issues associated with the use of the predominant Markov chain technique. Particularly, its inability to take into account the influence of more than one state in the chain of events to determine the subsequent steps in the evolution of daily activity schedules. At the heart of the proposed approach is the analysis of the empirical activity data with a view to identifying characteristic patterns present in the empirical activity sequences so that they can be exploited in the simulation of daily residential activity scheduling. The primary intended purpose of the approach is the simulation of activity schedules to underpin the simulation of residential electricity consumption patterns. Therefore, the activities considered are restricted to those in-home activities that can be reasonably associated with the use (or non use) of electric domestic appliances. The simulated activity sequences are constructed as the result of a series of transitions in which the likelihood of transitioning into the different activity states considered depends on the sequence of activity states observed prior to the transition in question.

The paper is structured as follows. Section 2 presents a brief and general review of current approaches to activity-based electricity demand modelling. In Section 3.2 we introduce the concepts relevant to the modelling approach developed in this paper. The algorithmic process that allows for the identification of activity sub-sequences relevant to the transitions into subsequent states is detailed in Section 3.3, and the implementation for the analysis of time-use data is detailed in Section 3.4. The validation of the proposed model is presented in Section 4 where the activity data simulated by the model are compared with: 1) the data produced by a model based on the commonly employed first-order Markov chain technique, and 2) the empirical activity data contained in the 2015 UK Time-Use Survey. Finally, we conclude by discussing the potential of the approach introduced in this paper, its shortcomings and potential for improvement.

2. Activity-based demand modelling: a brief overview

In the residential electricity demand modelling literature there is a number of examples of models that attempt to simulate activity patterns with a view to incorporating them into the demand modelling process. Some of the first examples of this take a deterministic approach like the one used by Yao & Steemers [34]. Modelling approaches have evolved since and the most recent examples are generally based on the use of empirical activity datasets such as those provided by time-use surveys and make use primarily of the Markov chain

technique. A detailed review of the different models and modelling approaches based on time-use data can be found elsewhere [27].

2.1. Time-use data

Time-use datasets are the best currently available sources of empirical data on the activities carried out by people throughout the day. These datasets come from large-scale surveys that look specifically at how people spend their time. The survey samples are typically based on households and the sample selection procedures aim at gathering the most statistically representative sample of a given population. Central to this kind of studies is a time-diary instrument in which respondents record their activities at regular time intervals. These time-diaries record the activity sequences for prescribed periods, which usually correspond to a single day. These surveys based on the collection of time-diaries are an effective means of gathering comprehensive datasets on how people spend their time, where do they spend it, and who they spend it with. Respondents are typically asked to complete time-diaries for one weekday and one weekend day in order to account for differences in time-use for the different types of day.

2.2. Markov chain technique

A commonly adopted approach to the simulation of stochastic sequences of potential events is based on the representation of these processes as Markov chains. Every step in the evolution of the sequence will result in one of the potential states considered. The set of different potential states is also referred to as the state space of the stochastic process. A Markov chain model consists of this state space and the probabilities associated with the transitions from each of the states into the others.

An abstraction of the way people go about their lives is thinking of them as if they were transitioning between the different elements of a set of potential states of activity. Based on this idea, it is thus possible to simulate the evolution of daily activity sequences as a Markov chain where the state space of the Markov process corresponds to the set of potential states of activity considered. The simplest example of this kind of model consists of a Markov process associated with a binary state space where the only potential states of activity considered are 'active' and 'inactive'. An example of such a model is the occupancy model developed by Richardson et al. [21]. An extension of these model was later introduced by McKenna et al. [16]. The proposed extension included one more activity state, but the technique used for the simulations remained the same.

The stochastic processes that can be adequately described by a first-order Markov chain model are said to satisfy the Markov property. Thus, they are characterised by what is

referred to in the statistical literature as memorylessness. In this context, the concept of 'memory' refers to the number of previously observed states at a given moment in the evolution of the stochastic process. Thus, term memorylessness refers to the fact that previously observed states are not relevant to the evolution of the process. That is, predictions for subsequent events are made based on the current state only. When the stochastic process of interest shows no potential for the presence of complex patterns or particularly distinctive features, a well calibrated first-order Markov chain model can provide a good description of it.

In the case of residential daily activity scheduling, it would be reasonable to assume that the process that describes the observed dynamics is somewhat more complex.

Recognising the limitations of the first-order Markov chain approach, new approaches that aim to capture more of this complexity have been developed. Examples of these are the approaches developed by Wilke et al. [32] and Aerts et al. [1]. The proposed models are variations of the first-order Markov chain models that are known in the statistical literature as semi-Markov models. The key difference between these and simple first-order models is the fact that in a simple first-order model the length of the blocks corresponding to the states in the chain are fixed, whereas in a semi-Markov model these are allowed to vary as a means of altering the 'duration' of the state. Although this class of models do represent an improvement over the conventional first-order Markov models, they still fail to address the memorylessness issue that characterises the simple first-order models. Therefore, the development of approaches that aim to address this issue would prove greatly advantageous, and is indeed necessary if we are to develop a better understanding of residential activity patterns.

2.2.1. Higher-order Markov models

In terms of improving the approach to the simulation of activity patterns, a natural step in this direction would be the development of higher-order Markov models. This kind of models are interesting for the purposes of activity modelling as they attempt to incorporate an element of 'memory' into the simulation of the stochastic process in question.

There are, however, serious issues associated with the use and development of this kind of models. Firstly, the number of free parameters in the model increases exponentially as a function of their order. The higher the order of the model the larger the amount of data required to construct the model. Consequently, as the number of parameters that need to be estimated and the amount of data that needs to be processed increases, so does the the computational intensity associated with both constructing and using the model. Secondly, and more importantly, the collection of all possible full high-order Markov chain models

is limited and completely stratified. That is, there are no models with a number of free parameters in between those defined by the order of the models. This is illustrated in Table 1, where the columns correspond to the levels (strata) represented by the orders of the model.

Every element of the set of models that represent a given stochastic process has an associated error. The sources of this error are the model bias, the variance of the simulated output, and an irreducible component associated with the noise of the data used to construct de model [11]. The bias is associated with the order of the model and the variance is associated with the number of parameters present in the model. A model with high bias will result in a poor representation of the relevant statistical characteristics of the empirical data; a condition known in the modelling literature as underfitting. A model with high variance will result in simulations overly sensitive to small fluctuations in the empirical data; a condition known as overfitting. Thus, in order to achieve an adequate representation of the stochastic process in question, the model has to strike a good trade-off between bias and variance.

The model bias decreases as the order of the model increases. The variance increases as the dimension (number of parameters) of the model increases. A consequence of such discontinuous and drastic increase in the dimension of the high-order Markov chain models is that, given a real process for which a given set of states is observed, it is not possible to obtain a parsimonious representation of the sate space. That is, the lack of models 'in-between' the exponentially increasing dimensions of the full high-order Markov chain models does not allow for an adequate balancing of the bias-variance trade-off.

The class of full higher-order Markov models offers good alternatives for the representation of theoretically consistent stochastic processes. From the empirical estimation point of view, however, the issues discussed above render this kind of models rather ill suited for the study of many 'real-life' processes, including the daily scheduling of activities. Therefore, alternative approaches to the simulation of activity patterns still need to be sought.

Table 1: Model dimension stratification as a function of the order.

Order (n)	1	2	3	4	5
Dimension	56	448	3,584	28,672	229,376

*The model dimension is given by $Dim = (|S| - 1)\,|S|^n$, where $|S|$ is the size of the state space and n is the order of the model. For the example presented in this paper (see Section 3.4), the state space is $|S| = 8$ (8 different activity states).

2.3. General approach to activity-to-demand conversion

Activity modelling for the purpose of electricity demand modelling has focused on the simulation of home occupancy patterns. Periods of active occupancy are then used as prompters for estimating the engagement in the different activities that may result in or

affect the observed electricity consumption patterns. The reader is referred to reference [27] for reviews of concrete examples of this approach. The implicit assumption underlying this approach is that the engagement in the different types of activities is a consequence of being in a state of active home occupancy (i.e. not asleep). However, arguably in most cases the causal relationship between these two events points in the opposite direction. That is, people are at home in a state of 'active occupancy' because they engage in the activities that result in demand for electricity and not engage in these activities just because they are at home. Particularly in the case of activities highly relevant in terms of demand for electricity such as cooking or laundering.

2.4. Insights from social practice theory

The literature on social practice theory points out that in order to effectively analyse people's daily schedules these should be treated as a whole [25, 28]. Therefore, for the sake of consistency, the study of the sequences of activities corresponding to these schedules should be carried out in the same manner.

The allocation of time to the different activities depends on different factors, but there are certain factors that prevent certain activities from happening at certain times due to the inconvenience they may cause, or the actual impossibility of the action. The scheduling of activities is restricted to some extent by the social constructions of time. For instance, the vast majority of people leave home for work or school at roughly the same time during workdays, lunch takes place at around midday and so on. These restrictions in turn result in a certain degree of synchronisation of the societal activities, which gives rise to the development of characteristic habitual sequences (behavioural patterns) [30].

These characteristic patterns determine to a great extent the way energy demand loads arise. As Torriti [28] points out, energy demand in households is shaped by the fundamental temporal characteristics of social practices, which include time dependence, duration and sequence. Moreover, as the kind of activity or activities being performed by the members of the household dictates the amount of energy required, the rate at which it is consumed (intensity), and the rate at which it varies in time, a more thorough analysis of empirical activity data is becoming increasingly relevant for the power sector. A better understanding of the way demand loads arise and vary would be most helpful for the purpose of future network planning and tackling peak demand issues through effective implementation of Demand-Side Response mechanisms.

2.5. Towards a more robust approach to activity modelling

The simplicity of the Markov chain technique, as well as its ability to provide a reasonable representation of the overall features observed in aggregate daily activity profiles, have led

it to become the current activity modelling paradigm. However, as we discussed in Section 2.2, this approach by definition disregards the influence characteristic patterns might have on the evolution of the stochastic process being modelled. When it is suspected that this kind of structural complexity might exist, as it is arguably the case of daily human activity scheduling, seems appropriate to try and learn more about the structural dependencies that might be present in the empirical data available.

Time-use datasets are, to date, the best source of empirical activity data (see Section 2.1). Therefore, it is reasonable to expect that the kind of characteristic patterns described in the social practice theory literature might be present in the reported activity sequences. In fact, the quality of the data is such that, if the sole purpose were to get a perfect snapshot of the current behaviour patterns, a very large set of activity sequences could essentially work as a model. This might seem like a feasible solution to the complications associated with activity modelling. However, such an approach presents major drawbacks. One of them is that the studies from which the empirical activity datasets stem are very expensive and they are difficult to carry out. But a more important drawback is the fact that in taking such an approach we loose the opportunity to learn more about the behaviours that these activity sequences represent. The development of activity modelling approaches allows us to devise ways in which the characteristics of the behaviours observed in these activity sequences can be quantified. This in turn allows us to develop methods to study how these might change. There is, therefore, great incentive to develop a robust approach to the simulation of activity patterns.

An activity modelling approach that focuses on explicit engagement in activities rather than the representation of periods of home occupancy would give us the opportunity to learn more about the scheduling of daily activity and identify characteristic patterns. Moreover, an approach that seeks taking into account the influence observed characteristic patterns have on the development of daily activity sequences might help: 1) produce better representations of the different behaviours observed, and 2) gain a better understanding of how associated energy demand loads arise.

In the context of activity modelling, the approach to the analysis of the data used as a basis for the model is of great importance. Therefore, if we want to be able to exploit the information provided by characteristic activity patterns we need to integrate this goal into both the approach to data analysis and the development of activity modelling techniques. New approaches to data analysis that focus on the study of the underlying activity patterns observed in empirical data would allow us to build on the strengths and insights provided by previously developed activity modelling approaches and assist in the development of new ones that take into account the influence of such characteristic activity patterns. The

approach presented in this paper attempts to do this.

3. Data and methods

3.1. The 2015 UK Time-Use survey

The latest major time-use dataset for the UK became available in 2017. It comes from the national survey carried out from April 2014 to December 2015 [10]. The survey was designed such that it would follow the guidelines set out by Eurostat on Harmonised European Time Use Studies (HETUS) [9], and the data collected would be compatible with the previous major time-use survey carried out by the Office for National Statistics in 2000 [12].

This study provides information about how individuals aged 8 years and over in the UK use their time on a given weekday and a weekend day. The collected time-diaries provide information about activities, location, co-presence, the use of computers and mobile devices, and even the level of enjoyment of time throughout the day. Questionnaire data provides information about characteristics of individuals and households such as employment and education, and demographic information such as age, gender, marital status, citizenship status and housing.

Respondents were asked to log their activities every 10 minutes. The responses were then re-coded based on a list of over 250 pre-defined activity categories. In addition to the main activity at any given time of day, they are able to specify secondary activities as well as location. Time-diaries are grouped into two broader categories: adult and children diaries. Filtering for adult diaries only we end up with $\sim 1.5 \times 10^4$ individual diary entries. All of this data was used to construct the activity models used for the analysis presented in this paper.

3.2. Variable memory length models

Incorporating an element of 'memory' into the simulation of activity patterns is a non-trivial issue. Taking first-order Markov models as the starting point, a natural way of achieving this would be through the development of higher-order Markov models. However, as we have pointed out, the use of these models poses two main problems: the number of parameters increases exponentially with the order of the models, and the stratification of the dimension of the models does not allow for an adequate balancing of the model fitting to the empirical data (see Section 2.2.1).

The two problems discussed above can be addressed by a conceptually simple idea: allowing the order of the model to vary as the stochastic process evolves.

In the context of activity modelling, this varying order (memory length) proves particularly useful as it allows us to take account of the self-adaptation of the scheduling of

activities throughout the day. Self-adaptation refers to the fact that, as the daily sequence of activities evolves, the probability of observing a given state further down the line might change. Thinking about a particular activity, say laundering, this could be thought of as the effect of having fulfilled the laundering duties for the day. While, on average (fixed-order model), the likelihood of engaging in laundering activities might still be relatively high, if the day's laundering duties have already been 'ticked-off' then the probability of having to do laundry again might not be the same now (variable memory model).

The implementation, however, presents some challenges. The way the order varies is determined as a function of the 'past'. That is, for a given transition in the course of the evolution of the process, the order of the model that would best describe such transition is determined by the the sequence of states observed prior to the transition in question. The number of relevant prior states, that is, the length of the relevant 'past', depends in turn on the transition.

When the order of the model is allowed to vary in this way, we effectively end up with a hierarchy of embedded models of different order with a well-defined structure of variable memory length. In practice, this means that a number of the transitions that would be described by the models embedded in the hierarchy are collapsed into transitions associated with a lower-order model that are deemed equivalent. The result of this is an overall reduction of the size of the state space. Consequently, the state space associated with the variable memory length model can be considerably smaller than the space associated with the full higher-order model corresponding to the highest order considered in the variable memory length model. The extent to which the size of the state space can be reduced depends on the nature of the stochastic process that wants to be modelled.

As Bühlmann et al. [6] point out, if the data for the stochastic process that wants to be modelled allows for the construction of such variable memory length models, there is nothing to lose but only to gain in comparison with the class of ordinary Markov chain models.

Variable memory length models originated in the field of information theory [22]. However, the concepts and methodologies developed for the purposes particular to this field have been expanded and adapted for a variety of purposes across different fields [6, 23, 33].

Such variable memory length models apply generally well to problems involving categorical data. It is therefore sensible to assume that these models can be applied to the study of the categorical time series that correspond to the the observed activity sequences in time-use data. Moreover, a variable memory length model has the potential to be used as an exploratory tool for the dynamics of the simulated processes. The variable memory structure allows for capturing the structural dependencies observed in the sequences of states resulting from the process, which in turn allows for a better representation of them.

In the context of daily activity sequences, this would allow for a better representation of the underlying behavioural patterns observed in the empirical data.

The key problem when constructing a variable memory length model resides in finding an effective method for identifying the relevant portions of the 'past' based on their influence on the outcomes of the transitions to subsequent states. This can only be done based on local, pairwise comparisons between the states present in the hierarchy of nested models. To this end, a variety of algorithms have been developed that allow for the comparison of the statistical characteristics observed for the transition from states associated with models based at different levels of the hierarchy. Various examples of these algorithms can be found in the literature where they feature in greatly varied applications ranging from information theory [15, 22, 23, 33] to bioinformatics [3, 14, 17] to statistical physics [2, 13] to language pattern analysis [4, 7].

In this paper we derive a modified version of one of such algorithms and implement it to construct a variable memory length model of daily activity patterns. In the implemented algorithm the criterion for discriminating between relevant past states is based on the the Kullback-Leibler divergence and the log-likelihood ratio test. The process involved is detailed in the following section.

3.3. Algorithm

The aim of the algorithm is to adaptively estimate the lengths of the past sequences of activity states that are relevant to determining the transitions into the next state. As explained in the previous section, these relevant past sequences can be of variable length ranging from 1, which would correspond to the first-order model, to a certain maximum length ℓ_{max}.

A conditional transition probability distribution is associated with each one of these relevant pasts. This distribution determines how likely it is that the next element in the sequence corresponds to each one of the activity states considered. The probabilities are estimated based on the number of times the relevant past sequences are observed followed by each one of the activity states.

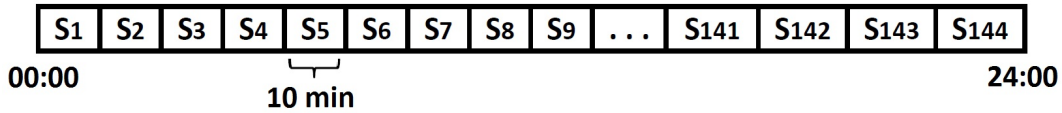

Figure 1: Graphical representation of a daily activity sequence. $S1 \cdots Sn$ =states in the sequence

Let us denote by t the index of the corresponding time interval of the day. The resolution of the time-use diaries is 10 min. This means that the day is partitioned into 144 intervals. Therefore, in this particular case, $t = 1, 2, ..., 144$.

Let $\mathcal{S} = \{s_\alpha, s_\beta, s_\gamma, ...\}$ be the set of activity states that make up the daily activity sequences.

Let $A_d = s_1 s_2 s_3 \cdots s_t \cdots s_{144}$ denote a given daily activity sequence, where $s_a = s_\alpha, s_\beta, s_\gamma, ...$ for any s_a in A_d, and let $s_{t-\ell}^{t-1} = s_{t-\ell} \cdots s_{t-2} s_{t-1}$ denote a given subsequence of A_d preceding the state s_t. This sequence corresponds to a 'past' of length ℓ for the state s_t.

We denote by $N_t\left(s_{t-\ell}^{t-1}\right)$ the number of occurrences of the 'past' denoted by $s_{t-\ell}^{t-1}$.

Assuming that the 'past' corresponding to the sequence $s_{t-\ell}^{t-1}$ exists, that is, it has been observed in the reported activity sequences, we have that $\sum_{s_a \in \mathcal{S}} N_t\left(s_{t-\ell}^{t-1} s_a\right) > 0$, where $s_{t-\ell}^{t-1} s_a$ denotes the activity sequence that results from concatenating the past $s_{t-\ell}^{t-1}$ and the subsequent state s_a.

Therefore, the estimates of the probabilities that make up the transition probability distributions associated with the past in question are given by

$$p_t\left(s_t \mid s_{t-\ell}^{t-1}\right) = \frac{N_t\left(s_{t-\ell}^{t-1} s_t\right)}{\sum_{s_a \in \mathcal{S}} N_t\left(s_{t-\ell}^{t-1} s_a\right)}. \tag{1}$$

The set of transition probabilities obtained for each one of the past sequences considered provide us with probability measures that can then be used to determine how relevant a particular 'past' is for determining the transitions into subsequent states compared to the past states associated with models at lower levels in the hierarchy.

Given two probability distributions defined over the same state space, the Kullback-Leibler divergence provides us with a measure of the difference between them [15]. In the context of information theory, this metric is used to estimate the amount of information lost when one distribution is used to approximate the outcomes resulting from using the other.

If $P_1(x)$ and $P_2(x)$ are the two probability distributions in question, the Kullback-Leibler divergence is given by

$$D\left(P_1 \mid\mid P_2\right) = \sum_x P_1(x) \log\left(\frac{P_1(x)}{P_2(x)}\right). \tag{2}$$

Based on this, we can then estimate the relevance of a given past state by multiplying this measure by the occurrence count of said state [6]. That is, the function for assessing the potential of a model with extended memory $\ell \geq 1$ is defined as

$$\Delta_t\left(s_b s_{t-\ell}^{t-1}\right) = \sum_{s_a \in \mathcal{S}} N_t\left(s_b s_{t-\ell}^{t-1}\right) p_t\left(s_a \mid s_b s_{t-\ell}^{t-1}\right) \log\left(\frac{p_t\left(s_a \mid s_b s_{t-\ell}^{t-1}\right)}{p_t\left(s_a \mid s_{t-\ell}^{t-1}\right)}\right), \tag{3}$$

where $s_b s_{t-\ell}^{t-1} = s_b s_{t-\ell} \cdots s_{t-2} s_{t-1}$, and

$$p_t\left(s_a \mid s_b s_{t-\ell}^{t-1}\right) = \frac{N_t\left(s_b s_{t-\ell}^{t-1} s_a\right)}{\sum\limits_{s_n \in \mathcal{S}} N_t\left(s_b s_{t-\ell}^{t-1} s_n\right)} \tag{4}$$

Given a complete activity sequence $A = s_1 s_2 s_3 \cdots s_t \cdots s_m$, we denote by $\mathcal{P}_\ell(A)$ the likelihood function estimate of that sequence based on the model of fixed memory length ℓ, where

$$\mathcal{P}_\ell(A) = \prod_{t=\ell+1}^{m} \mathcal{P}\left(s_t \mid s_{t-\ell}^{t-1}\right) . \tag{5}$$

Therefore, when individual states associated with models of fixed memory lengths $\ell < \ell'$ are being compared, many terms cancel each other out due to the multiplicative structure of the likelihood function estimate; the only remaining terms are the ones concerning the transition in question. If it were feasible to construct the full models of fixed memory length ℓ and ℓ', and obtain the likelihood function estimates for each one of their associated states, we could then re-state the discrimination criterion as

$$\Delta_t\left(s_b s_{t-\ell}^{t-1}\right) = \log\left(\frac{\mathcal{P}_{\ell'}(A)}{\mathcal{P}_\ell(A)}\right) . \tag{6}$$

What this shows is that the discrimination criterion is essentially a log-likelihood ratio test, only that the acceptance region is extended from [0,1] to $[0,K]$. The upper bound of the acceptance region for the discrimination criterion is chosen by asymptotic considerations. If n denotes the size of the empirical dataset (number of empirical activity sequences), and $|S|$ denotes the size of the set of activity states that make up the observed activity sequences, this upper bound can be estimated by

$$K \sim C \log(n) , \tag{7}$$

where C has to satisfy the condition $C > 2|S| + 4$. The derivation of this bound is provided elsewhere [5].

All the past states for which the value of the discrimination criterion falls within this region represent an improvement over lower-order models. The closer this associated value is to zero, the higher the relevance of the past state considered.

For the purposes of the algorithm, the extended log-likelihood ratio test (Eq. 3) acts as a gain function that will determine whether it is worth looking (further) back in the past in order to determine the probability distributions for the transitions into subsequent activity states. That is, whether it is worth extending the model with initial memory length $\ell = 1$ (first-order model) into a model with memory length $\ell' > \ell$ (higher-order model) for the observed set of transitions in question.

The log-likelihood ratio is a statistic, as it depends on the data. The test based on this statistic consists in comparing the value of this statistic to a critical value, or calculating the corresponding p-value. The null hypothesis of the test is rejected if the p-value of the statistic is too small, or the critical value is exceeded. The threshold is determined by the statistical significance level chosen for the test.

For the log-likelihood ratio test to be meaningful it is required that the models that are being compared are nested. That is, that the more complex model(s) can be transformed into the simpler one by imposing constrains on the parameters. As explained above, the variable length model is essentially an approximation of a full higher-order Markov chain model. The 'order' of the approximation obtained corresponds to the maximum length of the memory sequences ('pasts') deemed relevant. By restricting the length of the memory sequences considered for the generation of the model, we can downgrade the 'higher-order' nested models into the base first-order model.

As Eq. 6 shows, in the log-likelihood ratio the numerator corresponds to the likelihood of the observations being consistent with the extended memory (higher-order) model. This is the null hypothesis for the test. The denominator corresponds to the maximum likelihood of the observations for the different activity states.

It has been proven elsewhere that algorithms based on the log-likelihood assessment criteria converge asymptotically to the most suited variable length memory model [6].

3.4. Implementation

Focusing on the applications to the analysis of time-use data, we then start by re-arranging the data contained in the time-use diaries such that it is consistent with the formalism presented in the previous section. As observed in Section 3.1, the activity data in time-use diaries is coded into a fairly large number of categories. The set of pre-defined codes often correspond to very specific cases of the same kind of activity. Therefore, these can be regrouped into more general categories. We can thus summarise the information in the time-diaries by converting them into sequences of more generally defined activities performed throughout a given day. These sequences are thus composed of a series of 144 random categorical variables[1] with a common activity state space. Namely, we consider the following eight activity states: absence, sleep, generic active occupancy, food preparation, dishwashing, laundering, TV watching and ICT related activities. We take the first letter of each of these categories as the corresponding label. That is, $S = \{a, s, g, f, d, l, t, i\}$ is the

[1] A random categorical variable is a type of statistical variable that can take on one of a limited, and usually fixed, number of possible values, assigning each observation to a particular category on the basis of some qualitative property, i.e. a property that cannot be measured or assigned a numerical value.

state space of the individual transition probability distributions.

As we observed in Section 1, we intend to apply the simulation of activity sequences to the simulation of residential electricity consumption patterns. Therefore, the activity states relevant to the analysis are those which correspond to activities that can be reasonably associated with the use (or non use) of electric domestic appliances. A characterisation of the activity categories and the domestic appliances associated with them can be found in [20].

The starting point is the construction of the first-order model which will be the base of the variable memory length model. Based on the empirical activity sequences contained in the time-use dataset, we extract the transition probability distributions corresponding to this first-order model for each of the state transitions in the activity sequences.

Once the first order estimate is available, the algorithm recursively looks for the observed pasts of lengths larger than one for each of the transitions in the activity sequences. At the same time, estimates for the transition probability distributions associated with the corresponding past are calculated. The maximum length the algorithm is allowed to look back into the past can be estimated based on the characteristics of the empirical dataset used to construct the model. Given the sample size, that is, the number of empirical activity records (n), and the number of activity categories considered ($|S|$), the estimate of the suitable maximum length is given by

$$\ell_{max} = \frac{\log(n)}{\log\left(|S|\right)} \ . \tag{8}$$

A derivation of this estimate can be found elsewhere [31]. Using the values corresponding to the 2015 UK time-use data, $n \sim 10^4$ and $|S| = 8$, we have that this maximum length estimate is $\ell_{max} \sim 5$.

In order to prevent model overfitting we take the following two measures.

Firstly, we impose a restriction over the identified pasts in terms of their occurrence count. That is, we only take into account those pasts for which the condition $N_t\left(s_b s_{t-\ell}^{t-1}\right) \geq N_{min}$ is satisfied. The value of the lower bound for the occurrence count is in fact arbitrary. However, we choose this value to be $N_{min} = |S| = 8$ so that for any given past there is at least one possibility of observing a transition into each one of the eight activity states considered.

Secondly, we implement a probability smoothing step in the algorithm. Generally speaking, no single transition is absolutely impossible. Therefore, if we estimate the transition probability distributions based solely in terms of occurrence counts the risk of overfitting

increases[2]. The transition probability distribution smoothing process occurring during this step helps ensure that no activity state is predicted to have a null transition probability, regardless of the past observed before it.

We start by defining a minimum transition probability which we will denote by p_{min}. If when the transition probabilities are estimated it is determined that the value of the estimate is zero, we replace this null probabilities by p_{min} and then re-normalize the non-null estimates. In analytical terms, if we denote the new transition probability estimates by $p'_t(s \mid \cdot)$, the new set of transition probabilities has to satisfy the condition

$$\sum_{s \in S} p'_t(s \mid \cdot) = 1 \,. \tag{9}$$

That is, the new set of transition probabilities has to be an actual probability distribution.

If we denote the re-normalization constant by α, we can then express the new transition probability estimates in terms of the previous ones as

$$p'_t(s \mid \cdot) = \alpha p_t(s \mid \cdot) + p_{min} \,, \tag{10}$$

so that when $p_t(s \mid \cdot) = 0$, the value of the new estimate is p_{min}. Therefore, when we impose the condition stated above in Eq. 10, we find that the value of this constant is $\alpha = 1 - |S|\, p_{min}$.

Since the value of α has to be greater than zero, this in turn imposes a restriction over the value of the value of p_{min}, which is $p_{min} < \frac{1}{|S|}$. That is, p_{min} can take any value within the range $\left(0, \frac{1}{|S|}\right)$.

Thus far, all model parameters have been estimated based on the characteristics of the dataset used to build the model. For the sake of consistency, we determine the value of this parameter based on these same elements. Therefore, during the smoothing of the transition probability distributions all null probabilities will be replaced by

$$p_{min} = \frac{N_{min}}{n} \sim 5 \times 10^{-4} \,, \tag{11}$$

and the non-null ones scaled accordingly. By estimating the value of p_{min} in this way we ensure that 1) the value is within the appropriate range, and 2) the value does not exceed the probabilities estimated directly from occurrence counts.

When a past has been identified, and its corresponding transition probability distribution

[2]Given a particular empirical dataset, it is possible to find past states for which the transitions to subsequent states are restricted to a subset of the activity states considered.

has been estimated, the discrimination criterion described in Section 3.3 is used to determine whether the model would benefit from including said past in the state space of the transition in question. The process is repeated for all possible pasts of different lengths and for all the time intervals of the activity sequences.

The algorithm was implemented in the Python programming language[3] [19]. The resulting set of parameters for the model was then exported to a file, and arrange in a particular format so they could be reimported later for the purpose of simulating the sets of activity sequences used for the validation analysis.

4. Validation analysis results

The aim of the validation of this kind of models is to verify that the data generated by the stochastic process represented by the model possesses similar statistical characteristics as the empirical data used to construct the model.

Among the examples of occupancy and activity modelling found in the literature, a statistical measure commonly used for the validation of such models is the overall likelihood of finding a given person engaged in one of the occupancy/activity states considered throughout the day. These are commonly called state probabilities, and they vary with respect of time of day.

As we observed in Section 1, the motivation for developing the model presented is to produce a better representation of the characteristics of time-use data than what can be achieved by a first-order Markov model. The primary aim of the proposed approach is to exploit the inherent patterns observed in the empirical activity data.

The validation of this variable memory length model will therefore focus on assessing the improvements over the first-order estimates. This will be assessed in terms of the statistical characteristics of the representations compared to those of the empirical time-use data. In particular, the validation analysis will look at 1) the daily state probability profiles, and 2) the distribution of the state durations. As an additional quality measure, we will also look at how the past states considered by the model are distributed with respect to their relevance.

4.1. Distribution of relevant past states

The relevance of the potential past states considered by the model is assessed by the gain function defined in Eq. 3.

[3]The original implementation makes use of standard modules only. As the number of calculations that need to be performed is large, the implementation makes use of Python's multiprocessing capabilities. This is by no means necessary, but it certainly speeds up the data processing.

The name 'gain function' can be slightly misleading, as one might intuitively expect higher values of this function to indicate higher gain, and therefore higher relevance. However, as the gain function is based on the Kullback-Leibler divergence, lower values are in fact an indicator of higher gain. The lower the divergence the higher the relevance of the past state considered. The relevance of past states is independent of their length. That is, pasts of different lengths can be equally relevant in terms of the transition into subsequent states at a different or the same time of day.

Figure 2 shows the distribution of past states with respect to their associated gain. The gain function value corresponding to each of the past states considered depends on the transition the state is associated with. However, once this value has been estimated, we can simply regroup the past states by length and calculate their distribution with respect to their associated gain.

As we can observe in the different sub-figures in Figure 2, the distributions of the past states considered by the variable memory length model are heavily skewed towards zero. In each of the length groups over a third of the states is concentrated in the vicinity of zero.

As a measure of model quality, what these distributions show is that the benefit of adding the past states in question to the model is substantial. From an information theory point of view, the smaller the gain value of the associated states the larger the amount of additional information gained about the stochastic process. Therefore, the fact that large proportions of the past states considered are concentrated in the vicinity of 0 (Table 2) means that the overall amount of information gained is proportionally large.

Proportions of over 50 % of the states in each length group associated with a gain of less than 1 are clear indicators that the past states considered by the model have significant influence on the outcomes of transitions to subsequent activity states relative to the estimates associated with the first-order model. A summary of the maximum gain and percentage of states with gain values of less than 1 is presented in Table 2.

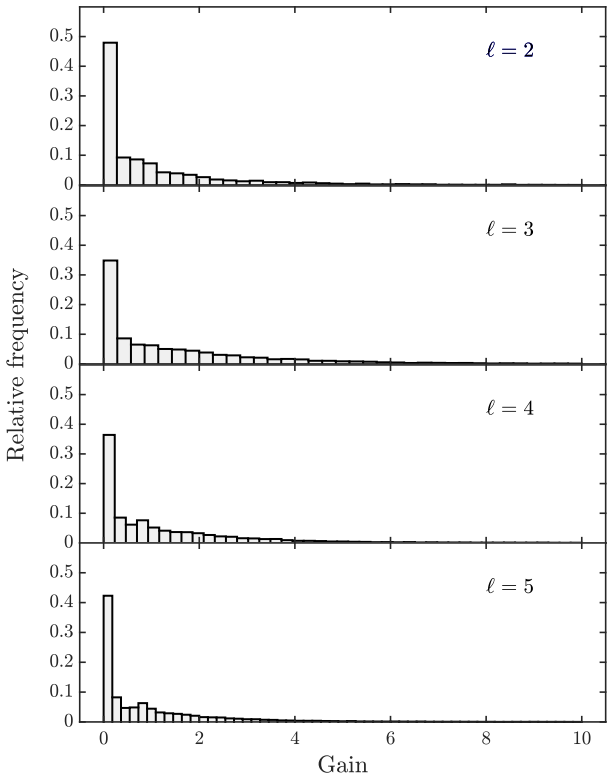

Figure 2: Distribution of past states with respect to their associated gain function value. Sub-figures correspond to sets of states grouped by length (ℓ = length of state).

Table 2: Proportion of states by gain range by length group and maximum gain value.

Length group	Maximum gain	Proportion of states with gain ≤ 1	Proportion of states with gain ≥ 10
2	15.939	71.37 %	0.214 %
3	45.334	57.27 %	1.949 %
4	51.995	61.38 %	1.185 %
5	48.281	69.45 %	0.584 %

In terms of the distribution of the past states considered by the model, another aspect of interest is the way the distribution of states with respect to their length changes throughout the day. The evolution of said distributions can be observed in below in Figure 3.

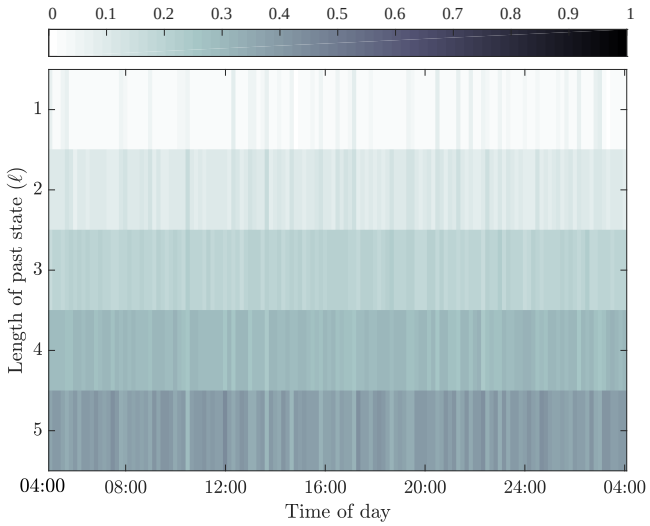

Figure 3: Daily evolution of the distribution of past states with respect to their length (ℓ).

The distribution in Figure 3 provides us with graphical evidence of the effects of the self-adaptation of the daily activity schedules. A more detailed analysis of the changes in these distributions might prove useful. For instance, it could potentially allow for the development of a method for assessing the potential for modifying activity schedules, or quantifying the effects of it. Although this is an interesting line of research, this falls beyond the scope of this paper.

4.2. State probabilities profiles

Based on a set of daily individual activity sequences it is possible to calculate the proportion of people engaging in the different activity states at any given time of day. These proportions correspond to the overall probabilities of finding a given individual engaging in any of the activity states considered and the evolution of these probabilities throughout the day is what is called the state probability profile.

Figure 4 shows the empirical state probability profiles from the UK 2015 time-use data. The eight shaded areas correspond to the different activities considered, as specified in the legend. For a given time of day the shaded area between the delimiting curves correspond to the state probability of the activity in question. These profiles exhibit some expected features such as a considerably low likelihood of finding people engaging in activities such as food preparation or laundering during night-time, or the increase in the likelihood of people engaging in food preparation around mealtimes.

Based on the simulated activity data the corresponding state probability profiles are extracted, so that the estimates obtained from the simulated data can be compared to the empirical values. The discrepancies between these state probability profiles are quantified as root-mean-square deviation.

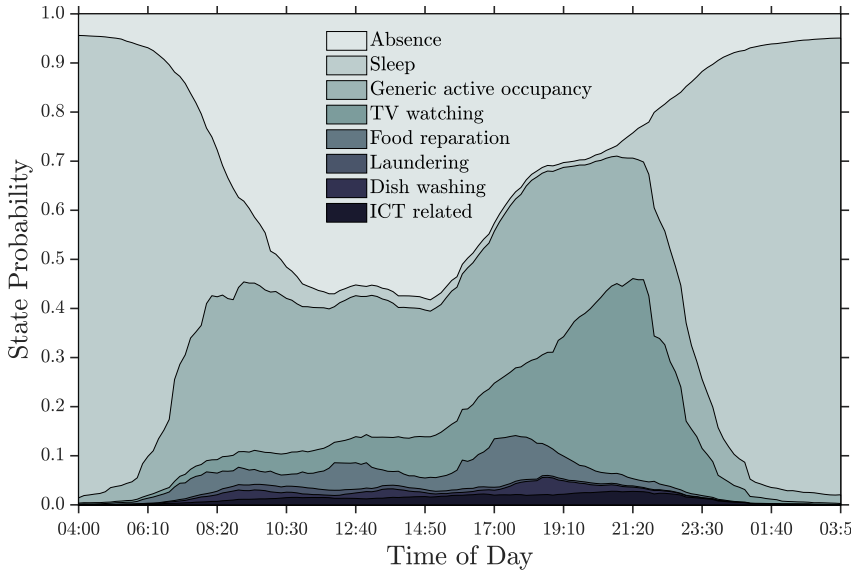

Figure 4: Empirical activity state probability profiles extracted from the UK 2015 time-use data. The order of the elements listed in the legend is the same as the order of the stacked state probability profiles.

Figure 5 shows the comparison between the activity state profiles obtained from empirical activity data and the estimates produced by the first-order Markov model and the variable memory length model. Both the variable memory length model and the first-order Markov model were used to produce sets of simulated activity sequences, with the number of simulated sequences increasing from 10 to 2×10^4. This is done so that the models' performance can be assessed and compared in terms of the rate of convergence of the root-mean-square deviation; as the size of the simulated datasets approaches the size of the empirical dataset, the root-mean-square deviation converges to its minimum value.

As Figure 5 shows, the root-mean-square deviations of the state probability estimates produced by both models converge exponentially to an asymptotic lower bound in the range 0.1 - 0.3 % for all the activity states considered. However, we can also observe that the deviation of the estimates produced by the variable length model are generally lower than those of the first-order estimates, resulting in a faster convergence of the root-mean-square (RMS) deviation.

For the simulated sets containing 10 activity sequences only, the deviations of the first-order estimates range between 12.4 and 2.1%, whereas for the estimates produced by the variable memory length model the deviations range between 9.4 and 1.5%.

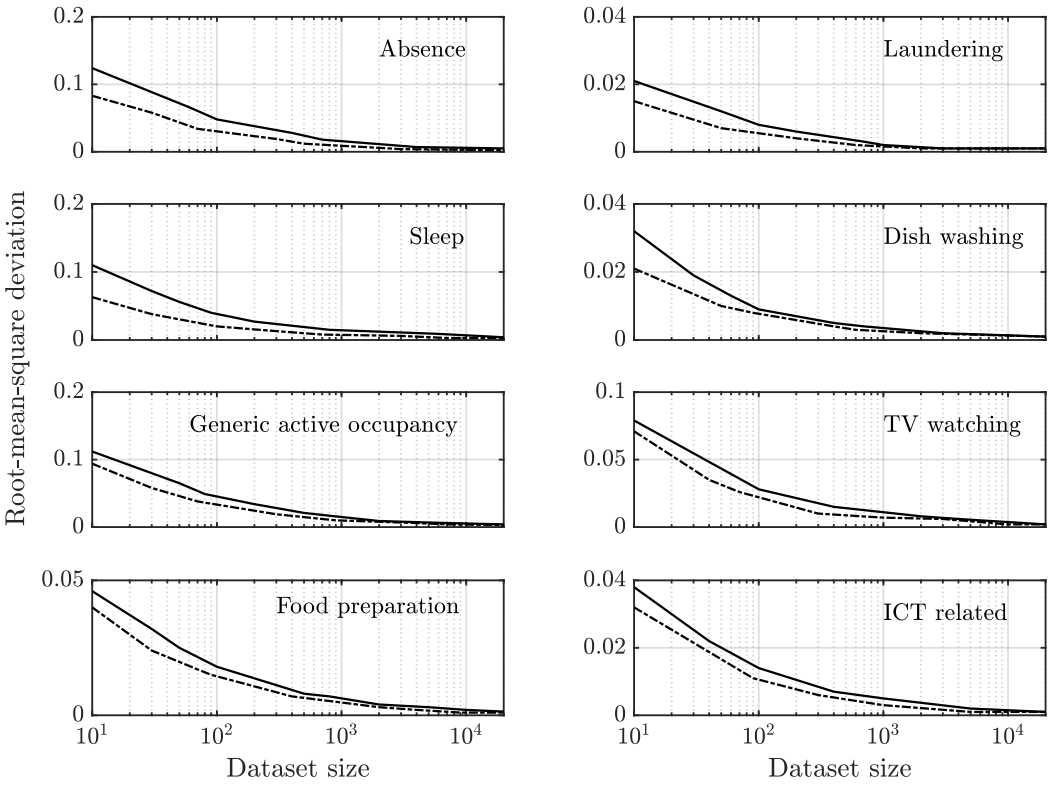

Figure 5: Root-mean-square deviation of the daily state probability estimates produced by the first-order and variable memory length models. Dotted line - Variable memory length model; Continuous line - first-order model.

Table 3 presents a summary comparison of the RMS deviations of the two different models in terms of absolute and relative percentage differences. These differences correspond to the improvements over the estimates produced by the first-order Markov chain model. The overall relative percentage difference averaged over the different dataset sizes is 28%.

Figure 5 and Table 3 give us a good idea of how the estimates produced by the first-order model compare with the estimates produced by the variable memory length model with respect to the size of the simulated dataset. In general, we can observe that the variable length model performs better than the first-order model, and that the level of improvement also appears to increase as the dataset size does. The most clear example is perhaps the case of Food preparation activities, where the improvement over the first-order model starts at a mere 13% and reaches nearly 50% when the size of the simulated dataset is comparable to the size of the empirical dataset.

Table 3: Relative change in RMS deviation by activity by dataset size.

Absence				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.1241	0.0468	0.0161	0.0051
Absolute difference*	0.0412	0.0159	0.0075	0.0023
Relative percentage difference*	33%	34%	46%	45%
Sleep				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.1111	0.0388	0.0159	0.0038
Absolute difference*	0.0472	0.0182	0.0079	0.0007
Relative percentage difference*	43%	46%	49%	19%
Generic active occupancy				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.1122	0.0461	0.0154	0.0039
Absolute difference*	0.0181	0.0121	0.0051	0.0009
Relative percentage difference*	16%	26%	33%	23%
Food preparation				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.0462	0.0178	0.0058	0.0019
Absolute difference*	0.0061	0.0036	0.0015	0.0009
Relative percentage difference*	13%	20%	26%	47%
Laundering				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.0211	0.0079	0.0019	0.0009
Absolute difference*	0.0062	0.0025	0.0005	0.0001
Relative percentage difference*	29%	32%	26%	11%
Dish washing				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.0322	0.0089	0.0034	0.0012
Absolute difference*	0.0111	0.0012	0.0011	0.0003
Relative percentage difference*	34%	13%	32%	25%
TV watching				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.0792	0.0283	0.0114	0.0024
Absolute difference*	0.0081	0.0059	0.0042	0.0001
Relative percentage difference*	10%	21%	37%	4%
ICT related				
Dataset size	10	10^2	10^3	10^4
RMS deviation - 1^{st} order model	0.0383	0.0141	0.0051	0.0014
Absolute difference*	0.0061	0.0034	0.0021	0.0003
Relative percentage difference*	16%	24%	41%	21%
Average percentage difference	24.5%	27%	36.3%	24.4%

* Difference with respect to the RMS deviation of variable memory length model estimates.

In Figure 6 we can now observe how the estimates for small and large datasets produced by the variable length model compare with one another. For each of the activities considered the estimates for daily state probability profiles produced by the model are compared with the empirical state probabilities. On the left column we have the estimates corresponding to a set of 100 simulated daily activity sequences. On the right column we have the estimates corresponding to a set of simulated activity sequences of the same size as the empirical dataset.

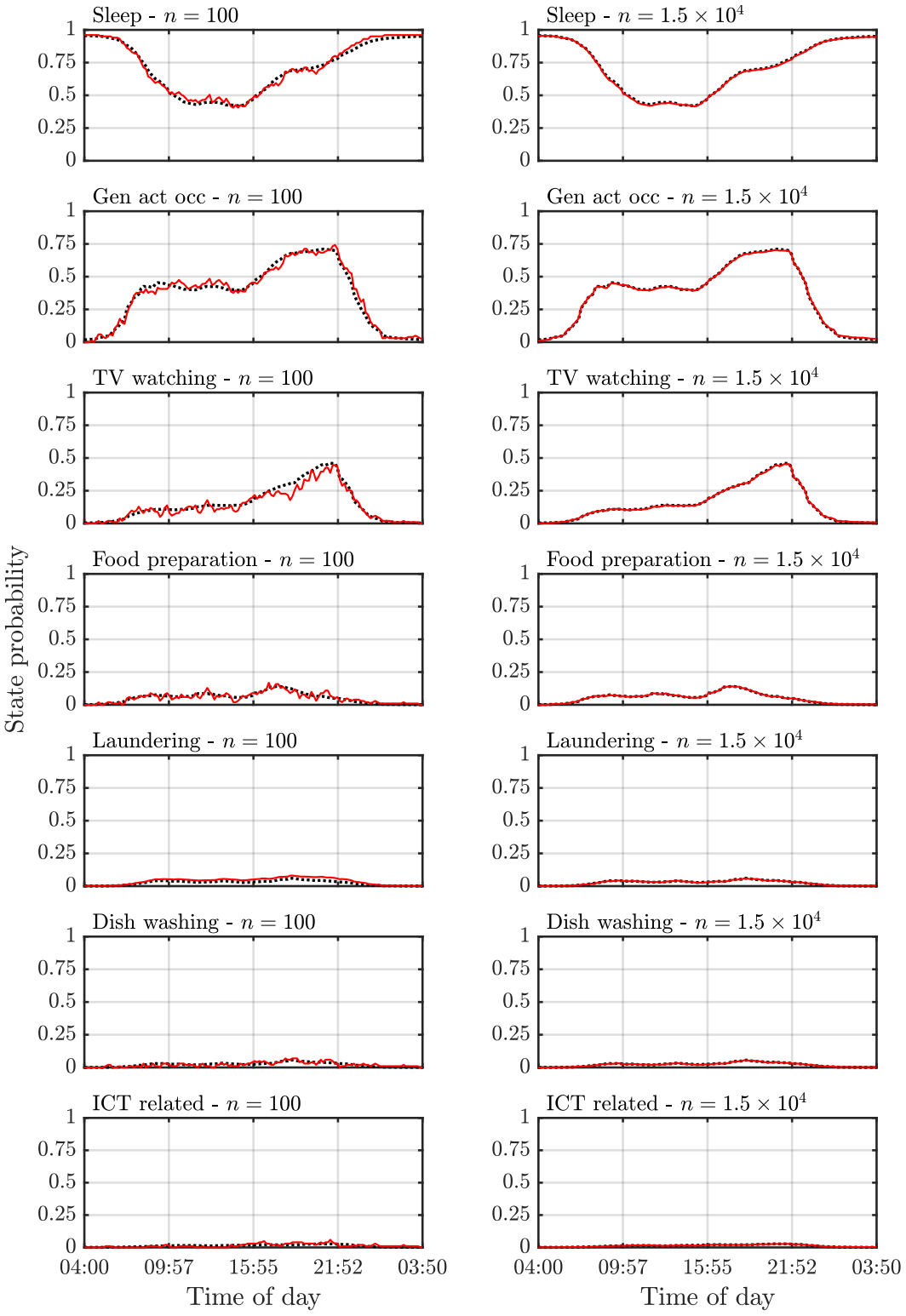

Figure 6: Comparison between state probability profiles as extracted from datasets with 100 and 1.5×10^4 simulated activity sequences. Dotted line - Empirical state probability profiles; Continuous line - Model estimates.

The estimates extracted from the sets of 100 daily activity sequences only are noisier

than those obtained from larger simulated datasets. However, as Figure 5 shows, these state probability estimates provide a closer representation (lower RMS deviation) of what is observed in the empirical data than the estimates produced by the first-order model.

4.3. State durations distributions

Another key aspect to an accurate representation of the observed activity patterns is the duration of the periods of engagement in the different activities considered. The modelling examples found in the literature that make use of the first-order Markov chain approach are mostly validated in terms of the state probability estimates. As in most of those examples, the goal is to develop a model for simulating activity patterns which can in turn be used to estimate residential electricity demand loads. The state probability profiles tell us how likely it is that a given individual would engage in any of the activity states considered. However, as we discussed in section 2, the timing and variability of the observed electricity demand patterns depends just as much on the relative frequency of the activities associated with the loads in question and the time these activities are performed for. Therefore, for the purpose of the activity pattern simulation, once a given person engages in a certain activity, being able to estimate the time said person spends performing such activity is just as relevant and desirable. This is particularly relevant in the case of activities associated with continuous use of electrical appliances such as TV watching or ICT use.

Figure 7 shows the distributions of the durations of the activity states corresponding to the eight activities considered by the model. The distributions observed in the data produced by the first-order and the variable memory length model are compared with the distribution observed in the empirical data.

The plots in Figure 7 would appear to indicate that the distributions obtained from the simulated data are well in agreement both with each other and with the empirical data. However, by comparing these distributions in terms of their RMS deviation (See Table 4), it is observed that in general the distributions obtained from the activity data simulated by the variable memory length model are closer to the empirical distributions than those obtained from the first-order simulations. The relative percentage difference is 30% on average.

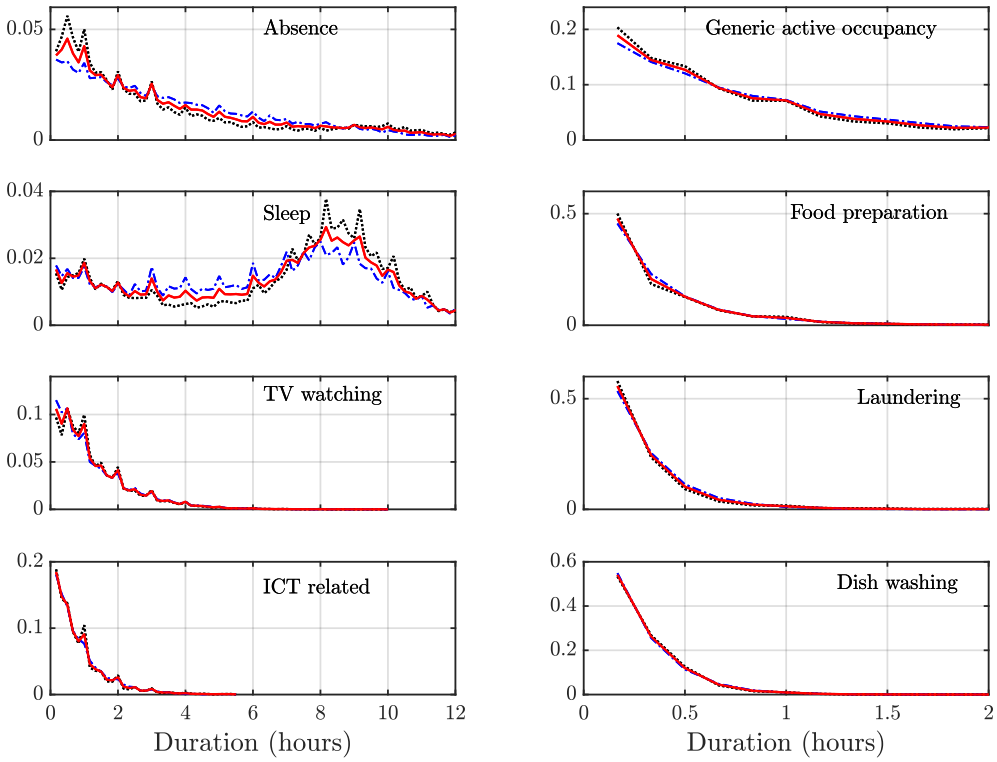

Figure 7: Distribution of state durations. Dotted line - empirical distribution; Red line - Variable length model; Blue line - first-order model.

Table 4: Relative difference in RMS deviation of distributions of state durations by activity.

Activity state	RMS deviation - 1^{st} order model	RMS deviation - Variable Memory Length model	Absolute difference	Relative percentage difference
Absence	0.0038	0.0029	0.0009	24%
Sleep	0.0042	0.0028	0.0015	34%
Generic active occupancy	0.0053	0.0038	0.0014	27%
Food preparation	0.0063	0.0042	0.0021	33%
Laundering	0.0052	0.0037	0.0016	29%
Dish washing	0.0147	0.0098	0.0048	33%
TV watching	0.0163	0.0117	0.0046	28%
ICT related	0.0069	0.0047	0.0022	32%

4.4. Transition Occurrence Frequency error profiles

The primary aim of the development of the activity modelling approach presented in this paper was to improve the simulation of daily activity scheduling. However, the lack of research on this particular area means that there are no standard metrics for quantifying this.

In order to assess the performance of a model in terms of the simulation of activity scheduling, we need to look closer at the behaviour in-between the states that make up the daily activity sequences. To this end, we developed a metric that allows us to estimate and

compare the errors associated with the first-order model and the variable memory length model in terms of the occurrence frequency of pairwise state transitions.

Given set of (same-length) daily activity schedules, transition occurrence frequency matrices are extracted for each intra-day transition interval. We define transition occurrence frequency as the frequency with which the transition between two given activity states occurs. Thus, the entries of these matrices correspond to the observed frequencies of the pairwise activity state transitions at each transition interval.

The set of empirical activity sequences provides us with the reference values of the transition occurrence frequencies. The transition occurrence frequency matrices extracted from the simulated output of both first-order and variable memory length models are compared to those reference values.

Let us then define the Transition Occurrence Frequency (TOF) Error as the 2-norm of the matrix whose entries are the absolute differences of the transition occurrence frequencies. If we denote this matrix as D^{abs}, we can express this mathematically as:

$$TOF_{err} = \left\| D^{abs} \right\|_2 , \tag{12}$$

The entries of the absolute difference matrix are given by $D^{abs}_{i,j} = \left| TOF^{sim}_{i,j} - TOF^{ref}_{i,j} \right|$, where TOF^{sim} and TOF^{ref} denote the simulated and empirical transition occurrence frequency matrices, respectively.

The 2-norm of a matrix is is also known as the spectral norm and is equivalent to the Euclidean norm of single vectors. For a given matrix A, this norm is defined as:

$$\|A\|_2 = \sqrt{\lambda_{max}\left(A^T A\right)} , \tag{13}$$

where λ_{max} is the maximum eigenvalue and A^T is the transpose of matrix A, respectively.

Both first-order and variable memory length models were, each, used to generate 50 sets of activity sequences with the same number of sequences as the empirical dataset. Average TOF errors were calculated for each of the simulated sets, as well as standard deviation ranges. The resulting average TOF error profiles can be observed in Figure 8.

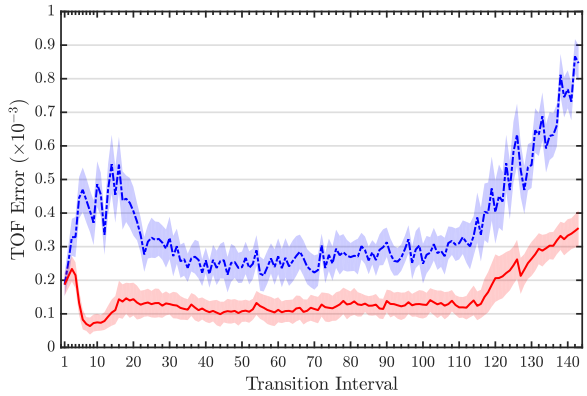

Figure 8: Transition Occurrence Frequency (TOF) Error profiles: Blue line - first-order model error; Red line - Variable memory length error; the shaded regions along the profiles represent the standard deviation ranges.

As we can observe from Figure 8, the TOF error of the variable memory length model is generally lower than that of the first-order Markov model. It is also observed that the TOF error associated with the first-order model would appear to have larger variability than the error associated with the variable memory lenght model. This might be attributed to the fact that the overall performance of the first-order model is not as stable, compared to the variable memory length model. Both of these results would appear to support the evidence for the improvement over the first-order Markov model provided by more standardised metrics such as the RMS error.

5. Discussion

In this paper we have presented a new approach to the stochastic simulation of residential activity patterns. There are a number of elements to the proposed approach which favour it over the current predominant approach based on first-order Markov chain models. In particular, this approach differs from the predominant one in that it focuses on exploiting characteristic behavioural patterns observed in the empirical activity data used to build the model. There are several reasons behind the interest in looking for an alternative approach to the simulation of residential activity patterns. Both the identified issues and the ways in which the proposed approach might help addressing them are outlined below.

- Compatibility with social practice theoretical frameworks

Human behaviour can be very diverse. However, research in the field of social practice theory shows that the restrictions imposed on people's behaviour by the social constructs of time result in a degree of synchronicity and the emergence of characteristic behavioural patterns.

The current predominant approach based on the Markov chain technique appears to perform reasonably well in terms of the representation of the overall activity profiles. Particularly when large simulated datasets are used to assess its performance. However, the inherent memorylessness of the stochastic processes represented by such models entails that any potential information on the presence of characteristic patterns and how they affect the overall evolution of the process is lost.

According to the branch of social practice theory that focuses on the study of daily schedules, the analysis of the temporalities and time dependences of the daily residential activity schedules requires that these are treated as a whole. However, the Markov chain technique by definition focuses exclusively on the transitions between any two adjacent states, paying no attention whatsoever to any potential structural dependencies present in the overall sequence of states.

The process for building the model based on the proposed approach focuses on the identification of characteristic patterns (relevant pasts) within the daily activity sequences observed in empirical datasets. However, in order to identify patterns that are meaningful relative to the entire *daily* empirical activity sequences, the analysis has to be carried out such that the daily activity sequences are processed as a unit. In that sense, the approach is consistent with the postulates of social practice theory outlined above.

- Improved understanding of the scheduling of activities throughout the day

Social practice theory acknowledges the existence of certain characteristic behavioural patterns that shape the overall daily activity sequences. However, research on this particular issue has thus far focused on the potential causes for this, and little attention has been paid to the analysis of empirical activity datasets with a view to identifying such characteristic patterns.

The data analysis carried out in order to identify the relevant pasts used to build the model is a good first step in learning more about the characteristic behavioural patterns observed in the residential sector.

The validation analysis of the model shows that the predictions for the transitions into subsequent activity states throughout the day are generally improved when the influence of these relevant pasts is taken into account. This is supported in particular by the results presented in Sections 4.2 and 4.4

It should also be noted that these results would appear to indicate that the approach to the analysis of empirical activity data presented in this paper has the potential to be used as a tool for a more in-depth analysis of such characteristic behaviours and their influence on the scheduling of activities throughout the day. An analysis of this sort would prove

very informative and would broaden our understanding of the way daily activity schedules are shaped and maybe even provide further insight into the ways in which these can be effectively re-shaped. Such an in-depth analysis, however, falls beyond the scope of this paper.

- Adjustment of cause-effect relationships assumptions in current predominant approach

Most activity-based residential electricity demand models incorporate the behavioural factor through the simulation of dwelling occupancy profiles. These daily active occupancy profiles are then used to produce estimates of the electricity demand loads associated with the different activities considered by the model in question. However, the implicit assumption in this approach is that the states of active dwelling occupancy are the primary events, and the states of engagement in the different activities are a consequence of the former.

In practice, however, the cause-effect relationship between these two kinds of events points in the opposite direction. That is, a state of active occupancy is a consequence of engaging in activities that are carried out at home. Arguably, in order to obtain a better mapping between activity and demand patterns, this should be taken into account. Therefore, an activity model that addresses such conflict would be desirable.

In the approach presented in this paper, the activity sequences are explicitly simulated.

By explicitly simulating the sequences of activities, the the conflict with the cause-effect relationship between active occupancy and engagement in activities is prevented. This also represents an advantage in terms of the applications of the activity model to the simulation of electricity consumption, as it is easier to produce estimates of the demand associated to each of the activities considered. Moreover, should it be required, the information about active occupancy is still readily available.

- Improvement of the simulation of activity-related electricity demand

Electricity demand load modelling has found its most relevant application in the context of network design. Consequently, the requirements for the purpose of conventional network design procedures have been key drivers in the development of particular approaches to demand modelling. Conventional network design relies heavily on experience, and the current model for power systems ensures that the high variability of demand loads at the local level can be largely ignored. In this model, very large centralised generation units supply a very large number of users. Aggregating loads over such large numbers results in a demand profile where most of the variability that would be observed at the local level is smoothed out. Therefore, simplified representations of the changes in demand throughout the day have been enough for the purposes of current network design and planning procedures, and

so, many activity-based demand models have focused on reproducing the characteristics of these largely aggregated profiles.

However, in the context of the transition towards low-carbon energy systems, conventional network design procedures are increasingly becoming inadequate. In order to increase efficiency it is envisioned that generation units will be located closer to where power is needed, i.e. systems will be based on a more decentralised configuration where smaller generation units will be responsible for meeting the demands of smaller numbers of users. In addition, it is expected that generation technologies with highly variable outputs (e.g. solar and wind) will have an increasingly large share in the generation mix. In practice, this will pose serious new challenges to network balancing operations. Smaller numbers of users entail that the variability of the demand observed by the generation centres will be considerably larger. And the intermittent nature of renewables like solar and wind will make it much harder to manage this demand variability.

The way demand varies at the local level and how this might change in the future is less well understood. Therefore, it is necessary to develop approaches to demand modelling that help improve our understanding of this variability and help inform future network operation. In addition, in order to help alleviate the issues associated with the nature of renewables generation, adequate demand-side management strategies are being actively sought. However, our understanding of the way people go about their lives and how that reflects in terms of the observed demand patterns is still limited. Therefore, approaches to demand modelling that better capture the dynamics of the relationship between activity and demand patterns would prove really useful.

The current predominant approach to activity modelling for the purpose of demand modelling is based on the first-order Markov chain technique. The approach seems to perform well in terms of the representation of demand variability at the aggregate level. That is, for large numbers of users. However, as the results presented in Section 4.2 show, the quality of the estimates produced by this approach decreases as the number of simulated users does. This is also the case for the approach presented in this paper. However, the quality decline of the first-order estimates is faster relative to the quality of the estimates produced by the variable memory length model.

Attempts to improve the approach to activity modelling have been made through the development of semi-Markov models [1, 32]. The primary aim of these models is to improve the representation of the duration of the periods of engagement in the different activity states. In this approach, this is done by means of altering a conventional first-order Markov model by incorporating an extra step in-between the state transitions. In this intermediate step the duration of the subsequent states is altered rather artificially.

As the results presented in Section 4.3 show, the approach presented in this paper also provides a better representation of the duration of the periods of engagement in the different activity states. However, this is improved representation of the duration came about as a natural by-product of an improved simulation of the daily activity scheduling.

5.1. Potential for improvement

Given the size of currently available empirical activity datasets, it would appear that the improvement over the first-order model is modest, with an average of 28%. However, current and future activity data might offer the opportunity for a more thorough analysis of the characteristic activity patterns, and the construction of richer, more detailed models.

Time-use datasets are typically country-specific, and they are approximately the same size in terms of the number of independent activity sequences. However, the interest in time-use data has motivated the creation of standards that have allowed for the compilation of a homogenised time-use database which gathers time-use data from a number of countries.

Each individual dataset is meant to be representative of its country of origin. However, given that the temporal constraints a person is subject to might be the same (or very similar) even for people in different countries, it is reasonable to expect that activity sequences with similar characteristics will be found across many of these country-specific datasets. With the aid of data clustering techniques, this homogenised database offers the opportunity to create larger unified datasets which can then be studied using the data analysis approach proposed in this paper.

The source dataset is of great importance. However, as we have pointed out, the way the data is analysed is just as important. Therefore, further improvement might also come from the technical side. For instance, more detailed analysis of time-use data might help identify particular features of interest. This information can then be used to tailor the algorithm implementation so that these specific features are highlighted during the simulation of activity sequences.

6. Conclusions

In this paper we presented a novel approach to the stochastic modelling of residential activity patterns based on processes with memory of variable length. The construction of the model is based on the analysis of time-use data with a view to identifying characteristic behavioural patterns present in empirical activity data. The primary aim of the proposed approach is to exploit said characteristic patterns for the simulation of residential activity sequences. The model performance is assessed based on the comparison of the statistical characteristics of the simulation outputs with those of the output of a typical first-order

Markov chain model and the empirical activity data. The methods used for the analysis of the empirical activity data are sensitive to the size of the dataset. Therefore, the maximum performance improvement is bound by the restrictions posed by the size of the dataset. Given the amount of data available for the construction of the model, the improvement over the current predominant approach would appear to be modest. The results of the validation analysis show that the simulation output of the proposed model does represent an improvement over the output of the commonly used first-order Markov chain models. A comparison of the RMS deviations reveals that the average improvement across datasets with increasing numbers of simulated users is 28%. In addition to the use of standard error metrics, we also proposed a new metric to assess in more detail the quality of the simulated output in terms of the representation of the intra-day dynamics of the activity sequences. The comparison of the deviations estimated by this new metric support the conclusions drawn by the use of the more standard metrics. Furthermore, the results of the analysis suggest that the proposed approach has the potential to be used as a new tool for a more detailed analysis of the structural complexity of empirical activity data. Such in-depth analysis would help broaden our understanding of the underlying behavioural patterns that shape the daily residential activity schedules. Therefore, an activity model based on this approach might prove really useful both in the context of time-use research and simulation of residential electricity demand loads.

Acknowledgements

Funding: This work was supported by the Consejo Nacional de Ciencia y Tecnología [postgraduate scholarship]; and the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M024652/1].

References

[1] Aerts, D., Minnen, J., Glorieux, I., Wouters, I., Descamps, F., 2014. A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. Build. Environ. 75, 67–78.
URL https://doi.org/10.1016/j.buildenv.2014.01.021

[2] Anderson, D. F., 2007. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. The Journal of chemical physics 127 (21), 214107.
URL https://doi.org/10.1063/1.2799998

[3] Bejerano, G., Yona, G., 2001. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. Bioinformatics 17 (1), 23–43.
URL https://doi.org/10.1093/bioinformatics/17.1.23

[4] Brown, P. F., DeSouza, P. V., Mercer, R. L., Pietra, V. J. D., Lai, J. C., 1992. Class-Based n-gram Models of Natural Language. COMPUTATIONAL LINGUISTICS 18, 467–479.
URL https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.9919

[5] Bühlmann, P., 2000. Model Selection for Variable Length Markov Chains and Tuning the Context Algorithm. Annals of the Institute of Statistical Mathematics 52 (2), 287–315.
URL https://doi.org/10.1023/A:1004165822461

[6] Bühlmann, Peter, Wyner, Abraham J., 1999. Variable length Markov chains. The Annals of Statistics 27 (2), 480–513.
URL https://doi.org/10.1214/aos/1018031204

[7] Charniak, Eugene, 2000. A maximum-entropy-inspired parser.
URL https://dl.acm.org/citation.cfm?id=974323

[8] Darby, S. J., McKenna, E., 2012. Social implications of residential demand response in cool temperate climates. Energy Policy 49, 759–769.
URL https://doi.org/10.1016/J.ENPOL.2012.07.026

[9] EUROSTAT, 2009. Harmonised European time use surveys - 2008 guidelines. EUROSTAT, Luxembourg, Luxembourg.
URL https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-08-014

[10] Gershuny, J., Sullivan, O., 2017. *United Kingdom Time Use Survey, 2014-2015* [data collection]. Centre for Time Use Research, University of Oxford. UK Data Service. SN: 8128,.
URL https://doi.org/10.5255/UKDA-SN-8128-1

[11] Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. No. 2 in Springer Series in Statistics. Springer New York, New York, NY.
URL https://doi.org/10.1007/978-0-387-84858-7

[12] Ipsos-RSL, Office for National Statistics, 2003. United Kingdom *Time Use Survey, 2000* [computer file]. 3rd Edition. Colchester, Essex: UK Data Archive [distributor],

SN: 4504,.
URL https://doi.org/10.5255/UKDA-SN-4504-1

[13] Lampoudi, S., Gillespie, D. T., Petzold, L. R., 2009. The multinomial simulation algorithm for discrete stochastic simulation of reaction-diffusion systems. The Journal of chemical physics 130 (9), 94104.
URL https://doi.org/10.1063/1.3074302

[14] Leonardi, F. G., 2006. A generalization of the PST algorithm: modeling the sparse nature of protein sequences. Bioinformatics 22 (11), 1302–1307.
URL https://doi.org/10.1093/bioinformatics/btl088

[15] MacKay, D. J. C., 2003. Information theory, inference, and learning algorithms. Cambridge University Press.

[16] McKenna, E., Krawczynski, M., Thomson, M., 2015. Four-state domestic building occupancy model for energy demand simulations. Energy and Buildings 96, 30–39.
URL https://doi.org/10.1016/j.enbuild.2015.03.013

[17] Miele, V., Bourguignon, P.-Y., Robelin, D., Nuel, G., Richard, H., 2005. seq++: analyzing biological sequences with a range of Markov-related models. Bioinformatics 21 (11), 2783–2784.
URL https://doi.org/10.1093/bioinformatics/bti389

[18] Peacock, A. D., Owens, E. H., 2014. Assessing the potential of residential demand response systems to assist in the integration of local renewable energy generation. Energy Efficiency 7 (3), 547–558.
URL https://doi.org/10.1007/s12053-013-9236-4

[19] Ramírez-Mendiola, J. L., 2018. Stochastic Chain with Memory of Variable Length Activity Model [Access Permission upon request to the corresponding author].
URL https://github.com/JoseLuisRaMen/scvml-activity-model

[20] Ramírez-Mendiola, J. L., Grünewald, P., Eyre, N., 2018. Linking intra-day variations in residential electricity demand loads to consumers' activities: What's missing? Energy and Buildings 161, 63–71.
URL https://doi.org/10.1016/j.enbuild.2017.12.012

[21] Richardson, I., Thomson, M., Infield, D., 2008. A high-resolution domestic building occupancy model for energy demand simulations. Energy Build. 40 (8), 1560–1566.
URL https://doi.org/10.1016/j.enbuild.2008.02.006

[22] Rissanen, J., 1983. A Universal Data Compression System. IEEE Transactions on Information Theory 29 (5), 656–664.
URL https://doi.org/10.1109/TIT.1983.1056741

[23] Ron, D., Singer, Y., Tishby, N., 1997. The power of amnesia: Learning probabilistic automata with variable memory length. Machine Learning 25 (2-3), 117–149.
URL https://doi.org/10.1007/BF00114008

[24] Shove, E., 2004. Efficiency and Consumption: Technology and Practice. Energy & Environment 15 (6), 1053–1065.
URL https://doi.org/10.1260/0958305043026555

[25] Shove, E., Pantzar, M., Watson, M., 2012. The dynamics of social practice: Everyday life and how it changes. SAGE, London, United Kingdom.
URL https://doi.org/10.4135/9781446250655

[26] Strengers, Y., 2012. Peak electricity demand and social practice theories: Reframing the role of change agents in the energy sector. Energy Policy 44, 226–234.
URL https://doi.org/10.1016/j.enpol.2012.01.046

[27] Torriti, J., 2014. A review of time use models of residential electricity demand. Renewable and Sustainable Energy Reviews 37, 265–272.
URL https://doi.org/10.1016/j.rser.2014.05.034

[28] Torriti, J., 2017. Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. Energy Research & Social Science 25, 37–47.
URL https://doi.org/10.1016/j.erss.2016.12.004

[29] Torriti, J., Hassan, M. G., Leach, M., 2010. Demand response experience in Europe: Policies, programmes and implementation. Energy 35 (4), 1575–1583.
URL https://doi.org/10.1016/j.energy.2009.05.021

[30] Walker, G., 2014. The dynamics of energy demand: Change, rhythm and synchronicity. Energy Research & Social Science 1, 49–55.
URL https://doi.org/10.1016/j.erss.2014.03.012

[31] Weinberger, M., Rissanen, J., Feder, M., 1995. A universal finite memory source. IEEE Transactions on Information Theory 41 (3), 643–652.
URL https://doi.org/10.1109/18.382011

[32] Wilke, U., Haldi, F., Scartezzini, J.-L., Robinson, D., 2013. A bottom-up stochastic model to predict building occupants' time-dependent activities. Build. Environ. 60, 254–264.
URL http://doi.org/10.1016/j.buildenv.2012.10.021

[33] Willems, F., Shtarkov, Y., Tjalkens, T., 1996. Context weighting for general finite-context sources. IEEE Transactions on Information Theory 42 (5), 1514–1520.
URL https://doi.org/10.1109/18.532891

[34] Yao, R., Steemers, K., 2005. A method of formulating energy load profile for domestic buildings in the UK. Energy and Buildings 37 (6), 663–671.
URL https://doi.org/10.1016/j.enbuild.2004.09.007