Energy and Buildings 96 (2015) 30–39

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Four-state domestic building occupancy model for energy demand simulations

Eoghan McKenna*, Michal Krawczynski, Murray Thomson

Centre for Renewable Energy Systems Technology (CREST), School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, UK

ARTICLE INFO

Article history:
Received 8 September 2014
Received in revised form 2 March 2015
Accepted 4 March 2015
Available online 12 March 2015

Keywords:
Domestic occupancy
Markov chain
Energy modelling

ABSTRACT

Stochastic building occupancy models are increasingly used to underpin building energy demand models, especially those providing high-resolution electricity demand profiles. This paper describes the development of an established two-state active-occupancy model into a four-state model in which the absent/present state and the active/inactive state are treated separately. This provides a distinction between sleeping and absence and so offers an improved basis for demand modelling, particularly high-resolution thermal modelling. The model uses a first-order Markov chain technique and the paper illustrates the value of this approach in duly representing the naturally occurring correlation of occupancy states in multiply occupied dwellings. The paper also describes how the model has been enhanced to avoid under-representation of dwellings with 24 h occupancy. The model has been implemented in Excel VBA and made available to download for free. The model is constructed from and verified against UK time-use survey data but could readily be adapted to use similar data from elsewhere.

1. Introduction

The transition to a low-carbon economy may be expected to require high penetrations of low-carbon technologies such as heat pumps, electric vehicles and photovoltaics [1,2]. These large and potentially undiversified loads and generation could present a considerable challenge to the operation of electricity distribution networks, potentially necessitating significant network reinforcement at high cost [3]. Furthermore, the accurate determination of exactly where and when such reinforcement is required is not straightforward – conventional low-voltage network design procedures typically use rather simple representations of the varying demand and rely heavily on experience – experience which is not available for high penetrations of low-carbon technologies [4].

To address this, and for other applications, 'bottom-up' models of domestic electricity demand that use probabilistic methods to provide stochastic high-resolution data for individual dwellings are currently being developed. A common feature of these models is a core representation of the occupancy of individuals within dwellings, which is used as the basis for subsequent modelling of end-use demands. The high-resolution model of domestic electricity demand developed by Loughborough University [5–8] is constructed in this way. It uses a two-state active-occupancy model that feeds into determining stochastic switch-on events for individual lighting and domestic appliances. The published model has been used widely within academia and industry for electricity network modelling [9–11]. It does not, however, include any detailed representation of thermal demands and, therefore, cannot yet be used to properly investigate the effects of the electrification of heating or CHP. Work is now underway at Loughborough to construct an integrated thermal–electrical demand model that can provide a convenient basis for future network studies, including the electrification of heating. This paper describes the first stage of that development.

A requirement for the thermal modelling is to account for casual gains associated with heat produced by lighting, appliances, and occupants. While the first two can be readily derived from the existing lighting and appliance models, the latter requires knowledge of when occupants are present within the dwelling, including when they are sleeping. The existing occupancy model [5], however, does not differentiate between occupants who are asleep and those who are not at home. The model has therefore been extended from a 'two-state' model to a 'four-state' model, where Table 1 describes the various occupancy states. The first aim of this paper is to describe the development and verification of the new four-state occupancy model. The model has been developed as a Microsoft Excel workbook and has been made available to download for free [12].

* Corresponding author. Tel.: +44 07958 531 842.
E-mail address: e.j.mckenna@lboro.ac.uk (E. McKenna).

http://dx.doi.org/10.1016/j.enbuild.2015.03.013
0378-7788/

Table 1
comparison of different states of occupancy for two-state and four-state models.

	Description of state
Two-state occupancy model	
1	Active occupant – at home and active
0	Not an active occupant
Four-state occupancy model	
00	Not at home, and not active
01	Not at home, and active
10	At home, and not active
11	At home, and active

In addition, this paper presents a comprehensive review of the occupancy modelling literature and identifies two modelling assumptions that require further scrutiny. The first is that occupancy can be adequately modelled as a first-order Markov process, and the second is that occupancy patterns within dwellings of more than one occupant can be adequately modelled by assuming the occupants are independent of each other. The second aim of the paper is therefore is to evaluate the discrepancies between real occupancy data and the synthetic occupancy data generated by models with these assumptions.

2. Literature review

2.1. Early 'bottom-up' models of demand

Swan and Ugursal describe the requirements for comprehensive models of residential energy consumption to assess the impacts of technology and behaviour change, and critically assess the strengths and weaknesses of different modelling approaches [13]. A distinction is made between 'top-down' and 'bottom-up' approaches. Top-down models take a whole-system view of demand, and are based on identifying correlations between 'macro' variables (e.g. price, climate, etc.) and aggregated demand data. Bottom-up approaches model individual end-uses at the building level, based on detailed data from samples of the population, which are then aggregated together to produce a wider view of demand. Top-down models tend to be deterministic, while bottom-up models tend to use probabilistic methods to account for demand diversity. Of the two approaches, a bottom-up approach is the more appropriate choice for the requirement of modelling of individual buildings for future network studies as it can represent the diversity of demand at this level.

Early work on developing bottom-up models of domestic electricity demand emphasised the importance of including occupancy as a core variable. Capasso et al. developed a complex residential load model that used a probabilistic 'availability at home' input for each member of the household [14]. The high-resolution domestic lighting model developed by Stokes et al. takes into account the number of occupants but not variations in daily occupancy patterns, with the authors noting that "taking account of these patterns would improve the modelling of diversity" [15]. Jardine used household occupancy derived from measured household demand data as a key input parameter to generate disaggregated high-resolution demand profiles [16]. Yao and Steemers developed a method for generating household load profiles based on five pre-determined occupancy patterns and noted that "the load profile depends very much on the occupancy pattern" [17].

2.2. Markov-chain technique and use of time-use surveys

Following these early models, important advances were made by the Richardson–Thomson occupancy model [5], and the Page occupancy model [18]. These are the earliest published models that use a first-order Markov-chain technique to generate stochastic synthetic occupancy patterns. The concept of a first-order Markov-chain technique is that the probability of being in any state in a given time step depends only on the state in the previous time step (an nth-order Markov-chain would base these probabilities on the previous n time-steps). The probabilities of changing from one state to another ('transition probabilities') are held in 'transition probability matrices' which are derived from observed occupancy states. Both models use 'time-inhomogeneous' Markov-chains, which means the transition probability matrices vary in time.

The models are differentiated by the type of occupancy data used to calibrate the model. The Page model was based on occupancy data from five single-occupancy office rooms. By contrast, an important feature of the Richardson–Thomson model was that it addressed the issue of the lack of availability of occupancy data with which to calibrate models. This was done by inferring occupancy from the UK time-use survey. Time-use surveys are large nationally representative surveys of how people use their time, which typically contain many thousands of 24-h diary entries. The UK time-use diaries detail participant location and activity at 10-min resolution, and allowed the introduction of the concept of 'active occupancy' – defined as when an occupant is at home and not asleep.

The technique of using a first-order Markov chains combined with national time-use surveys has been widely adopted in the literature. Widén developed a similar model based on the Swedish time-use survey [19]. López-Rodríguez et al. implemented the technique to develop an active occupancy model based on the Spanish time-use survey [20]. Muratori et al. used it to develop an activity model based on the American time-use survey [21].

The first-order Markov chain technique has also been widely adopted to develop models of occupancy in office buildings [22–24]. There are, nonetheless, important ways in which the literature can be distinguished. Broadly, four trends can be identified, as described below.

2.3. Techniques to reduce data input requirements

Given that one of the main drivers for the development of occupancy models was to address the lack of availability of real occupancy data, it is perhaps unsurprising that research has focussed on further techniques to produce synthetic occupancy data given less input data. This has been of particular interest in the field of non-domestic occupancy models.

Page's technique rearranges the Markov-chain formulae such that they depend on the probability of presence and a 'parameter of mobility' [18] – a measure of the likelihood that occupants change state. Wang's technique estimates transition probabilities based on 'expected sojourn times', or the mean time occupants spent in a state [22]. Both techniques are proposed to simplify data input requirements. Page argues that the 'probability of presence is a rather standard input' for simulations which 'should be available to the user'. Wang's technique is proposed to 'further simplify the specifications for [transition probability] matrices'.

Both techniques, however, still require a basis for estimating the model parameters. Page's 'parameter of mobility' is derived from the transition probabilities, while Wang's 'expected sojourn time' should be based on a representative sample of occupancy data. To be accurate therefore, both techniques still have a requirement for actual occupancy data, in which case it is arguably simpler to use this data to calculate and use the transition probabilities directly. In the absence of detailed occupancy data, however, simplifying assumptions about occupancy are required, and these two techniques offer alternative methods for deriving synthetic occupancy profiles.

2.4. Increasing the number of occupancy states

A second trend has been to increase the detail of the output of models in terms of their number of occupancy states. The Page model and Richardson–Thomson models are 'two-state'. Page distinguishes between occupants being 'present' or 'not present', while the Richardson–Thomson model distinguishes between occupants who are 'at home and not asleep' and those who are not. The issue is that a two-state model merges states of occupancy and activity that are important to distinguish for thermal–electrical demand modelling. The Page model, for example, merges occupants who are 'present and asleep' with those 'present and awake', while the Richardson-Thomson model merges occupants who are 'at home and asleep' with those who are 'not at home'.

Recognising this limitation, a number of subsequent models have adopted a three-state approach corresponding to 'absent', 'present and active', 'present and inactive' [19,25]. A three-state model, however, still merges the states 'absent and active' with 'absent and inactive'. To avoid the merging of states altogether, therefore, the new model developed for this paper includes four states of occupancy (see Table 1).

Another way in which models have increased the number of occupancy states is in accounting for the location of the occupant within the building. This has been a development in non-domestic occupancy models, where it is argued that 'multi-zone' models have value [22,26]. As the model presented here is for domestic buildings, however, a single-zone approach would appear to be adequate.

2.5. Model validation and accuracy of state durations

The validation of occupancy models is generally determined by comparing the statistical characteristics of the synthetic occupancy data produced by a model with the original data on which the model is based. A statistical measure that is commonly used for validation is the probability of an occupant being in a given state at a given time ('state probabilities'). While the first-order Markov-chain technique has been validated in terms of state probabilities [5,19,20], the technique has been criticised on its ability to produce accurate distributions of state durations – the duration that an occupant stays in a given state. Wilke et al., for example, state that "activity durations cannot be captured coherently" using the technique [27], and developed instead a 'higher-order Markov process' occupancy model based on the French time-use survey that follows any transition calculation step with an additional step that determines the duration of the state based on the probability distribution of state durations. Aerts et al. adopt a similar technique to develop an occupancy model based on the Belgian time-use survey, stating that compared to the first-order Markov technique "the occurrence of unrealistic occupancy durations is far less probable" with the higher-order Markov technique [25].

Given these criticisms of the first-order Markov chain technique, and the fact that models adopting it have not been validated in terms of their ability to adequately model occupancy state durations, one of the aims of this paper is to quantify the discrepancies in state durations of the model output.

2.6. Accounting for diversity and correlation

The final area in which occupancy models are differentiated is in how they account for diversity and correlation in individual sequences of occupancy. As mentioned previously, one of the benefits of the first-order Markov chain technique is that it produces stochastic output and which therefore introduces the desirable diversity to the resulting occupancy sequence. Beyond this, however, modellers have sought to improve models by accounting for groupings of people that follow similar patterns of occupancy.

Yao and Steemers, for example, propose five distinct occupancy patterns based primarily on the employment status of the occupant [17]. Aerts et al. use 'heirarchical agglomerative clustering' to identify seven distinct patterns of occupancy in the Belgian time-use survey data [25]. The Aerts model is then calibrated separately against the distinct groups, and produces synthetic occupancy data with similarly distinct groups of occupancy patterns.

A further way in which people can be grouped together is in terms of how many people they live with. Differentiating between houses with different numbers of residents has the important consequence of accounting for the correlation in occupancy state transitions in such dwellings. Two people who live together are more likely to have correlated occupancy state transitions than two people who do not. Some models, as a result, treat dwellings with different numbers of residents as distinct, and calibrate separate transition probability matrices for one-person dwellings, two-person dwellings, etc. [5,19,20]. Others, however, do not, and make the assumption that individual occupancy patterns in dwellings with multiple occupants are independent from each other [25,27].

The final aim of this paper is therefore to compare the output of the model developed for this paper, which accounts for correlated occupancy in multiply occupied dwellings, with the output of a separate model that assumes occupants are independent of each other.

3. Method

3.1. Model description

The aim of the model is to generate stochastic occupancy data with the same statistical characteristics as the time-use survey on which it is based, notably in terms of state probabilities and state durations. The model distinguishes four states of occupancy as described previously (Table 1). States are described in terms of a combined state variable which consists of a first digit describing the occupancy state (1 = "at home", 0 = "not at home") and a second digit describing the activity state (1 = "active", 0 = "not active"). The model differentiates dwellings by number of residents (up to a maximum of 6 occupants) and weekends are distinguished from weekdays.

The model uses a first-order time-inhomogeneous Markov-chain technique, and is based on the UK time-use survey [28]. This survey data consists of 24-h diaries recoded at 10-min intervals starting at 4 am. The model adopts the 10-min resolution but is configured to run from midnight to midnight. The switch from 4 am to midnight is conceptually straightforward but did reveal various practical issues that are discussed below.

3.2. Inferring location and activity from time-use survey data

The time-use survey diary entries contain location and activity fields which can be used to infer whether a participant is at home and active. For example, an entry in the location field of '2' indicates the participant was at home. Participants were assumed active if not asleep, resting, or sick in bed.

An issue arose, however, as the vast majority (>99.99%) of diary entries where participants specified 'sleep' as their activity have the location field as 'not answered', i.e. the field was not filled in. The time-use survey data does not, as a result, specify where participants are when they are sleeping. The location of sleep had to be inferred therefore from the location of the activity that followed

Table 2
Inferred locations for diary entries where occupants were inactive and the location field was not completed.

Inferred location	Proportion of total number of entries changed
Home	92.1%
Other specified location (not travelling) including second homes, work places, and other people's homes	5.0%
Travelling	1.9%
Location not filled in or unspecified	1.0%

Table 3
illustration of the difference between transition probability matrices used in a two-state model (matrices on the left) and those used in a four-state model (matrices on the right). The top row shows specific examples for one-person dwelling, while the bottom row shows generic matrices for dwelling of n occupants.

$\begin{pmatrix} T_{00} & T_{10} \\ T_{01} & T_{11} \end{pmatrix}$	$\begin{pmatrix} T_{0000} & T_{0100} & T_{1000} & T_{1100} \\ T_{0001} & T_{0101} & T_{1001} & T_{1101} \\ T_{0010} & T_{0110} & T_{1010} & T_{1110} \\ T_{0011} & T_{0111} & T_{1011} & T_{1111} \end{pmatrix}$
$\begin{pmatrix} T_{00} & \cdots & T_{n0} \\ \vdots & \ddots & \vdots \\ T_{0n} & \cdots & T_{nn} \end{pmatrix}$	$\begin{pmatrix} T_{0000} & \cdots & T_{nn00} \\ \vdots & \ddots & \vdots \\ T_{00nn} & \cdots & T_{nnnn} \end{pmatrix}$

the sleep entry or, when sleep was the last activity of the diary, from the location of the preceding activity. The location fields that were inferred are detailed in Table 2.

By changing the location fields in this way, however, another issue arose: it now appeared that a small number of participants were going to sleep in a different location to where they woke up. The amended data showed that people were more likely to be asleep at home at the start of their diary entry (4 am) than at the end, and that this was true for both weekdays and weekends. This resulted in an unrealistic discontinuity of state probabilities that occurred at 4 am – the time when the diary entries start and end.

There are two possible explanations for this. The first explanation would be to do with participants who are away from home and active towards the end of the day, who then return home (but do not specify their location in the diary) and go to sleep within 10 min (the time-resolution of the diary). In this case the method of inferring sleep location counts this as sleep away from home, when in fact the participant was at home. In this case, the location-inferring method will over-estimate the proportion of people going to sleep at the end of the day away from home.

The second explanation would be to do with participants who are asleep at the start of the day and who are away from home (but do not log this) and who then wake up and return home within 10 min. In this case, the location-inferring method would incorrectly count this as sleep at home, and would be over-estimating the proportion of people waking up at home.

The first explanation is, however, the more reasonable, and so the approach was to amend the location-inferring process for the end of diary sleep entries. Whenever a sleep location was determined to be not at home at the end of the diary, then there was a probability that this would be overridden and 'at home' specified instead. A value of 32% for this probability was chosen in order for the distribution of state probabilities to align at 4 am during weekdays for single-occupancy dwellings.

3.3. *Generating the transition probability matrices and synthetic data*

The processes involved in generating synthetic occupancy are the same as for the previous occupancy model, described fully elsewhere [5], with the exception that residents can have four possible occupancy states rather than two. The underlying steps and calculations are the same, there are simply more states involved.

With regard to calculating the transition probability matrices, the main difference is the size of matrices required to hold the transition probabilities between the relevant occupancy states. Table 3 shows the difference between matrices required for a two-state and four-state occupancy model, where T_{XY} stands for the transition probability of transitioning from state X to state Y in the next time step. For the previous model the size of matrices required for each time-step was $(n+1)\times(n+1)=(n+1)^2$ where n is the number of occupants. For the new model, however, the matrices are $[(n+1)\times(n+1)]\times[(n+1)\times(n+1)]=(n+1)^4$ in size and, as a result, considerably bigger. Compared to the previous model, there is more than 30 times the number of data elements.

The starting states for the model are based on the probability distributions of occupancy states observed in the time-use survey data at midnight. Subsequent states are then determined by picking a random number for each time step and using this with the appropriate transition probability matrix to determine the next state.

3.4. *Inactive arrivals and departures*

The time-use survey data includes a number of diaries where participants start the day (at 4 am) asleep and at home, and end the day asleep away from home. Similarly, there are diaries where the opposite happens and participants finish the day at home having started it away. Due to these diaries, a discontinuity appears in the transition probabilities at 4 am where there is a significant probability of inactive arrival or departure from home (see Fig. 1). While this is an accurate reflection of the original data, it is clearly not appropriate for the model to show residents changing location

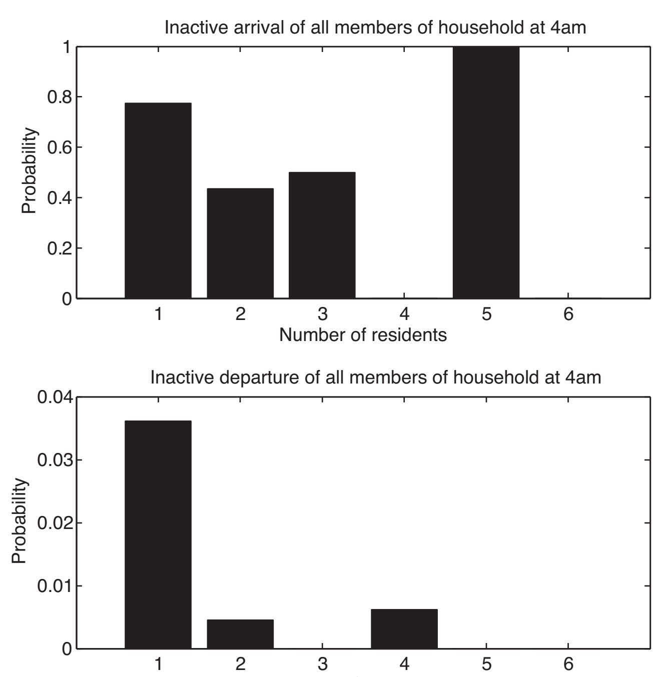

Fig. 1. probability of inactive arrival and departure of residents.

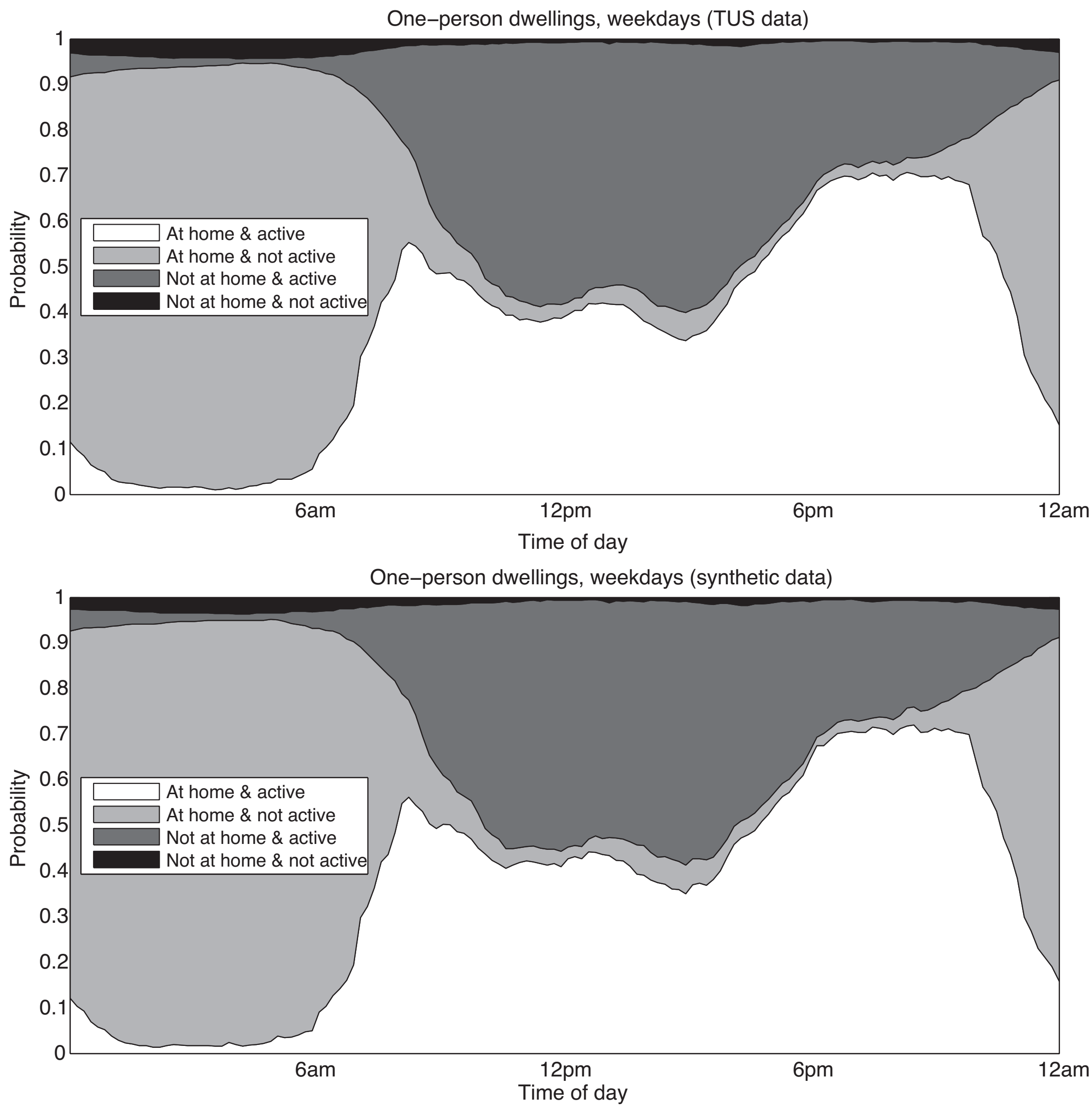

Fig. 2. comparison of state probabilities for one-person dwellings on weekdays between time-use survey data and synthetic occupancy data generated by the model.

while asleep. To deal with this, the transition probabilities for 4 am were simply replaced with those for the previous time-step.

3.5. Estimating state durations

One of the aims of the paper is to quantify any discrepancy in state durations between the original time-use survey data and synthetic data generated using the first-order Markov chain technique. While it is trivial to calculate the durations of occupancy states throughout the day, there is the issue of what to do with the state durations that have been inevitably truncated by the start and end of the diary. For example, a diary might show a participant being asleep and at home at 4 am when the diary starts, and waking up at 8 am. At the end of the day the same diary might show the participant going to sleep at home at 10 pm, and staying asleep until 4 am the following day. Due to the 24 h time limit of the diary, the state durations will be truncated and would be calculated as two separate durations, one of 4 h and the other 6 h.

Assuming the participant has the same routine each day, however, it would be reasonable to assume that these two durations could be counted as a single 10 h duration, and that this would be more representative of the participant's actual pattern of behaviour. Following this logic, the approach was taken to 'wrap' the occupancy data; the process was to calculate all the state durations in a diary entry, and if the first and last occupancy states were equal, to add together the first and last state durations. The same technique was applied to calculate state durations for the synthetic data generated from the model.

4. Downloadable model

The model has been developed as an Excel workbook and has been made available for free download [12]. It contains the starting state probabilities, transition probability matrices, and the algorithm to calculate the state transitions written in VBA. The model has been made available under a Creative Commons license and may be adapted for specific applications with due acknowledgement.

5. Results

5.1. Verification of state probabilities

Fig. 2 compares the state probabilities for one-person dwellings for weekdays for the time-use survey and the synthetic occupancy

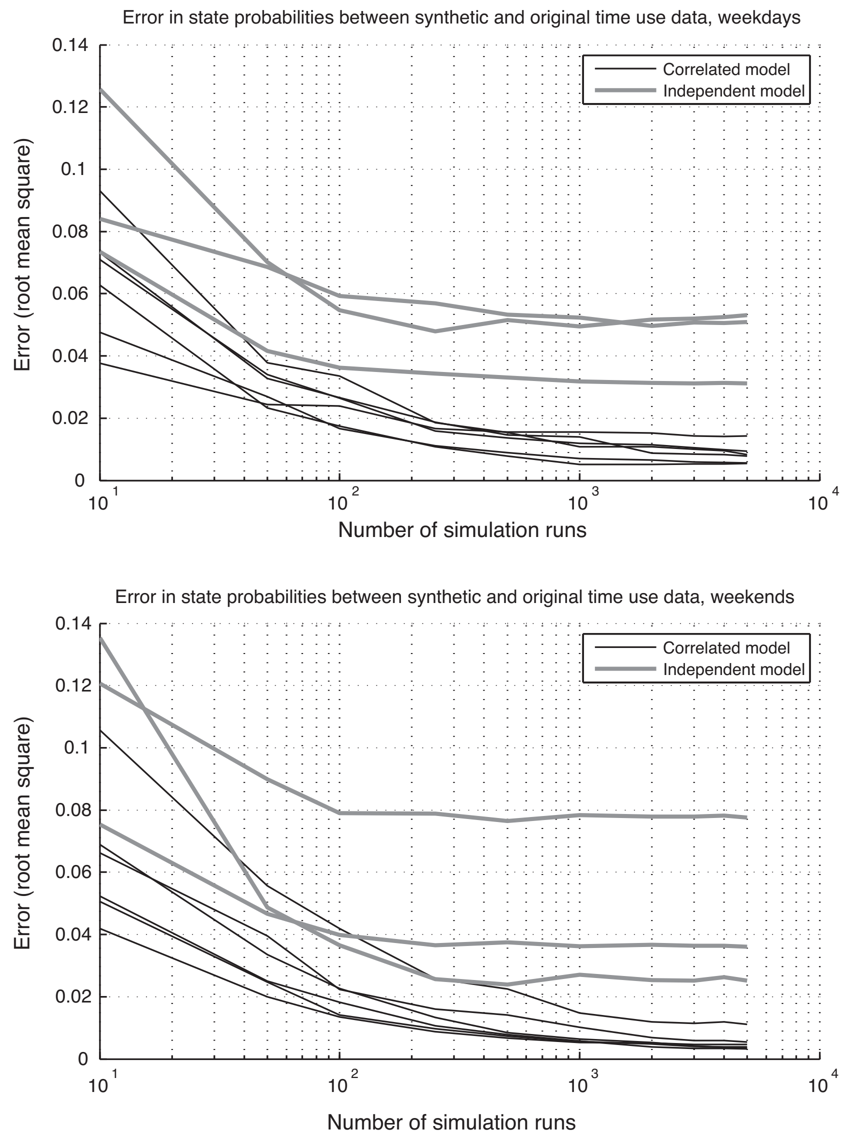

Fig. 3. root-mean-square error in state probabilities between original and synthetic data as a function of the number of simulation runs. The six black lines correspond to the six dwelling size combinations of the correlated model, while grey lines correspond to one-person through to three-person dwellings for the independent model.

data generated by the model. The time-use survey data is based on 1702 diary entries, while the synthetic data is based on an identical number of runs of the model. One-person dwellings have been chosen to facilitate the comparison, as these have the fewest number of possible occupancy states. The profiles show expected features including a low proportion of activity at night, a tendency for people to be out of the dwelling during the day, and peaks in activity around meal times.

The original and synthetic data shown in Fig. 2 are in close agreement – the root-mean-square error between the two sets of state probabilities is less than one percent. The root-mean-square error in state probabilities has been checked for all combinations of number of residents and day types and from 10 simulations runs up to 5000 (see black lines in Fig. 3). All combinations show the same trend: errors are in the range of 4% to 11% for 10 runs with an exponential decay to an asymptote in the range 0.5–1%. For all combinations there is little improvement in accuracy between 1000 and 5000 runs. In general, therefore, it can be said the first-order Markov chain technique accurately reproduces the state probabilities found in the original data.

5.2. *Verification of state durations*

Fig. 4 compares the states duration of occupancy states for the same synthetic and original one-person dwelling data as in the previous section. The four possible states for one-person dwellings are shown. From top sub-figure to bottom, these correspond to “not at home and not active”, “not at home and active”, “at home and not active”, and “at home and active”. Considering all the states, shorter durations are generally more likely to occur, and durations greater than 15 h are unlikely. A notable exception is the peak in probability in the “at home and not active” (“10”) state around 8 h, which corresponds to the duration that people tend to be asleep for. The synthetic data appears to capture the broad variations in durations shown by the original data though there is a slight under-representation of the extremes and over-representation of the middle-of-the-range.

Similarly to the state probabilities, the root-mean-square error between original and synthetic data was calculated for a range of simulation runs from 10 to 5000 for one-person to three-person dwellings. A similar trend of exponential decay in error was

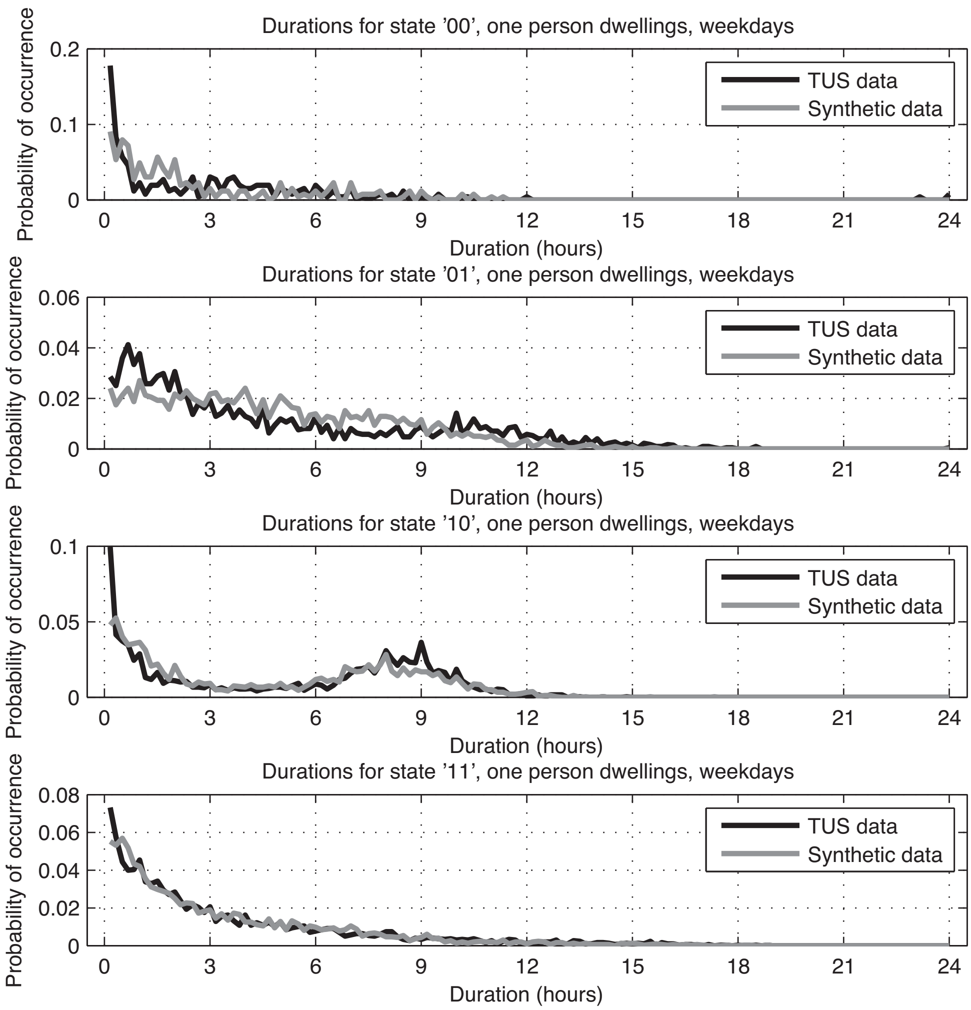

Fig. 4. comparison of state durations for one-person dwellings on weekdays between time-use survey data and synthetic occupancy data produced by the model.

observed, with errors ranging from 2% to 4.5% for 10 runs, reducing to the range 0.5% to 1.5% for 5000 runs. In general, therefore, the first-order Markov chain technique also accurately reproduces the state durations found in the original data.

5.3. Improving 24-h occupancy with an uplift factor

The principle aim of the model is to provide a basis for energy demand simulations. There is a particular need therefore to ensure the model accurately accounts for the proportion of time that dwellings are occupied. Fig. 5 shows the probability distribution of the number of times a dwelling is left unoccupied during the day. The original time-use data is shown alongside the synthetic data from the model and is shown for all household and day types combined. As with the data shown in Fig. 4, the synthetic data under-represents the beginning and end of the distribution and over-represents the middle.

Given the principle aim of the model, there is a particular concern with the model's under-representation of dwellings that are never left unoccupied (that have 24 h occupancy). Table 4 shows the discrepancies in the proportions of dwellings that have 24 h occupancy between the original and synthetic data for all household sizes and day types. As expected, dwellings are generally more likely to have 24 h occupancy on weekends than on weekdays, and if they have larger numbers of residents. The model, however, generally under-represents the proportion of dwellings with 24 h

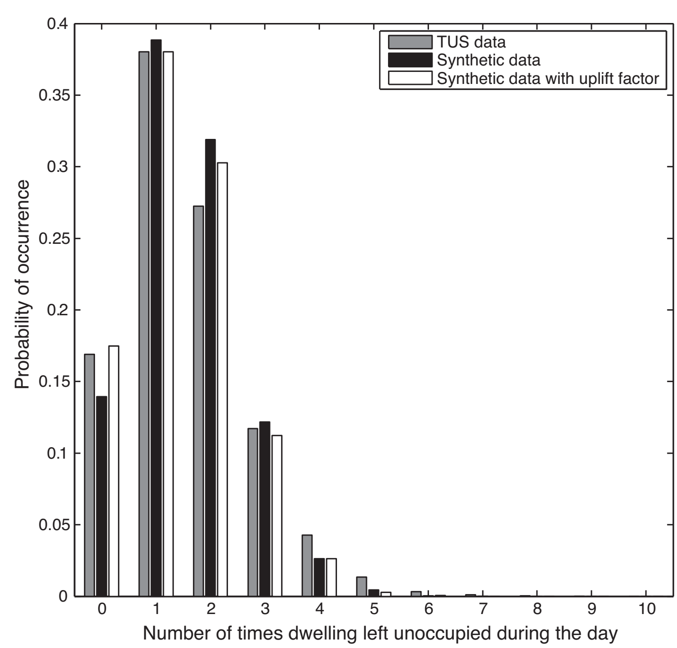

Fig. 5. probability distribution of the number of times a dwelling is left unoccupied during a day (all household and day types combined). Original time-use data is shown alongside the unadjusted synthetic data and synthetic data adjusted using an uplift factor.

Table 4
probabilities of dwellings having 24 h occupancy and uplift factor required for model adjustment.

Number of residents	Original data (time-use survey)		Unadjusted model output		Uplift factor	
	Weekdays	Weekends	Weekdays	Weekends	Weekdays	Weekends
1	7.4%	10.8%	4.0%	7.1%	3.5%	4.0%
2	14.3%	21.7%	10.5%	19.4%	4.2%	2.9%
3	16.2%	29.5%	9.6%	28.4%	7.3%	1.5%
4	15.0%	33.6%	10.2%	33.7%	5.4%	−0.2%
5	26.4%	52.3%	19.4%	51.2%	8.7%	2.3%
6	20.5%	58.3%	31.5%	55.3%	−16.1%	6.8%

occupancy, with two exceptions: six-person dwellings on weekdays, and four-person dwellings on weekends.

To correct this discrepancy, therefore, the model is amended so that a proportion of dwellings are forced to have 24 h occupancy such that the model output matches the original data. For each simulation run therefore, a random number is compared with an "uplift factor" which will determine whether the dwelling is required to have 24 h occupancy. The uplift factor P_u is calculated as:

$$P_u = \frac{P_o - P_m}{1 - P_m}$$

where P_o is proportion with 24 h occupancy in the original data and P_m the proportion in the unadjusted model output. The calculated values are shown in Table 4 for each combination of household and day type. Note that where the uplift factor is negative the model output is unadjusted.

Fig. 5 shows the resulting model output adjusted by the uplift factor. The proportion of dwellings with 24 h occupancy is in close agreement now. There is also closer agreement in the proportion of dwellings left unoccupied once or twice during the day. The model does still under-represent the right-hand tail of the distribution (dwellings left unoccupied more than 3 times a day), but this is not expected to cause significant errors in the context of energy demand modelling.

5.4. Occupancy correlation in multiply occupied dwellings

The final aim of the paper is to quantify the discrepancies between original time-use data and synthetic output from a model that assumes occupant independence. To do this, a separate set of transition probability matrices was generated based on the time-use data, but treating every individual diary entry as being independent from each other. This resulted in $4 \times 4 \times 144$ matrices for weekdays and weekends. These were then used to construct synthetic data for dwellings with multiple occupants. Occupancy for a 2-person dwelling was created by generating two independent one-person occupancy chains, and then aggregating them together. This will be referred to as the 'independent' model, as opposed to the 'correlated' model described previously, which takes into account correlated occupancy in dwellings with multiple occupants.

As with the correlated model output, the discrepancies between the independent model output and original data were quantified in terms of root-mean-square errors for state probabilities and state durations. And again the errors were seen to reduce exponentially against the number of simulation runs. Overall however, the errors for the independent model were much greater than those the correlated model. After 5000 simulation runs for one-person to three-person dwellings, independent model errors were in the range 3% to 5% for state probabilities, and 0.5% to 2.5% for state durations. The state probability errors are shown for reference by the grey lines in Fig. 3.

6. Discussion

6.1. Verification of the first-order Markov chain technique

The paper has presented a new four-state occupancy model based on the first-order Markov chain technique. One of the aims of the paper was to verify the model's output. The results confirm that the model produces synthetic data with state probabilities that agree well with the original time-use survey data.

In response to criticisms of the technique found in the literature, the paper has also investigated the technique's ability to reproduce realistic state durations. When measured in terms of root-mean-square errors, for a large number of simulation runs, the discrepancies in state durations are comparable to those for the state probabilities. The results therefore indicate that the first-order Markov chain technique also reproduces state durations with reasonable accuracy.

In terms of patterns of occupancy, however, the technique under-represents dwellings with 24 h occupancy and dwellings that are frequently left unoccupied (>3 times per day). The middle of the range is, by contrast, over-represented.

The explanation for this is that the transition probabilities are, by definition, calculated on the mean of the original data. By contrast, the range of the distribution of the synthetic data is not based on the original data, but is rather a function of the random process of selecting transitions, in the same way that tossing a coin and counting the number of consecutive heads will result in a probability distribution about a mean. The first-order Markov chain technique does a relatively good job at recreating the probability distributions of state durations in the original data because the original data happens to have a probability distribution that is similar to that produced by the random outcome of the Markov process. The technique would produce much less satisfactory results if the original data had a bipolar distribution around a mean (e.g. 50% of the population stay at home all day, and 50% are in and out at high frequency), or indeed if the original data had a completely homogeneous population (e.g. 100% of the population left the dwelling unoccupied once per day).

The under-representation of 24 h occupancy is, however, particularly significant given the model's aim of supporting energy demand simulations, and it is for this reason that the authors have implemented the uplift factor to force the model output to better correspond to the original data.

6.2. Accounting for correlated occupancy in multiply occupied dwellings

The discrepancies in models that assume occupancy independence were also investigated and were shown to be greater than those for models that take into account occupancy correlation. The assumption of occupant independence is used in the higher-order Markov process models that have been developed with the aim of correcting the discrepancies in state durations inherent in first-order Markov chain models. While authors of higher-order models have acknowledged that it is "unlikely" that occupants act independently [25], this paper has presented results that quantify the resulting discrepancies. Future occupancy modellers should therefore be aware that in choosing a higher-order Markov technique that assumes occupant independence, they may be correcting one discrepancy, only to introduce another.

6.3. Recommendations for future research

An obvious recommendation to improve occupancy modelling would be to use a higher-order Markov chain technique, but not to assume occupant independence. While such a model would arguably suffer from none of the discrepancies raised here, the problem with developing such a model is to do with the limited amount of data generally available in time-use surveys. An inevitable consequence of accounting for diversity is that it entails taking the original data and dividing it into smaller subsets of data. The model developed here, for example, distinguishes between dwellings with different numbers of occupants, as well as between weekdays and weekends, and calculates separate transition probability matrices for each variation. The 12 smaller data subsets that result will inevitably produce 'noisier' data, with more artefacts. The state duration distributions shown in Fig. 4 provide some examples of the 'spikiness' that is increasingly introduced when sample sizes are reduced – larger sample sizes would produce distributions that were smoother and more 'representative'.

In addition to the 12 sub-divisions mentioned previously, however, a higher-order model that accounted for different dwelling occupancies would need to sub-divide again according to the total number of occupancy states. While for one-person dwellings this only implies a further division of four, larger numbers of occupants will require considerably more (49 in the case for a six-person dwelling). Furthermore, to account for variation in state duration through the day, the current higher-order models sub-divide the day into 24 h periods, requiring a further sub-division again. The technique, therefore, quickly becomes unfeasible due to data limitations and the increasingly unrepresentative subsets of data which are used to calibrate the model.

One of the issues with the above is that researchers currently do not have a technique that can be used to determine whether a sub-division has resulted in an unacceptable deterioration of the quality of the model output. It would be useful for example to be able to determine whether there was enough data to justify accounting for diversity in say two-person dwellings, but not enough to justify doing the same for three-person dwellings. Future research is recommended therefore on methods to determine the statistical robustness of sub-dividing time-use data to account for occupancy diversity.

7. Conclusions

This paper presented a new four-state high-resolution stochastic occupancy model developed to provide a basis for an integrated thermal-electrical demand model that is being developed at Loughborough University. The model uses a first-order Markov chain technique and takes into account the correlation in occupancy in multiply occupied dwellings. The outputs from the model are verified by comparing them with the original UK time-use survey data on which the model is based.

The results verify that the first-order Markov chain technique produces output that matches closely the original data, with good agreement in terms of state probabilities and state durations. Compared to the original data, however, the first-order Markov chain technique nonetheless under-represents the beginning and end of the range of state durations and over-represents the middle. The resulting under-representation of 24 h occupancy is significant for the purposes of energy demand modelling, and is corrected in the model using an "uplift factor" that forces an appropriate proportion of dwellings into having 24 h occupancy.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council, UK, within the Top and Tail of Energy Networks project (EP/I031707/1). For the use of the UK time-use survey, the authors gratefully acknowledge Ipsos-RSL, the Office for National Statistics, the survey sponsors, and the UK Data Archive, none of whom bear any responsibility for the further analysis or interpretation contained herein.

References

[1] G.P. Hammond, P.J.G. Pearson, Challenges of the transition to a low carbon, more electric future: from here to 2050, Energy Policy 52 (2013) 1–9.
[2] T.J. Foxon, Transition pathways for a UK low carbon electricity future, Energy Policy 52 (2013) 10–24.
[3] D. Pudjianto, P. Djapic, M. Aunedi, C.K. Gan, G. Strbac, S. Huang, D. Infield, Smart control for minimizing distribution network reinforcement cost due to electrification, Energy Policy 52 (2013) 76–84.
[4] I. Richardson, Integrated High-resolution Modelling of Domestic Electricity Demand and Low Voltage Electricity Distribution Networks, 2010.
[5] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (2008) 1560–1566.
[6] I. Richardson, M. Thomson, D. Infield, A. Delahunty, Domestic lighting: a high-resolution energy demand model, Energy Build. 41 (2009) 781–789.
[7] I. Richardson, M. Thomson, D. Infield, C. Clifford, Domestic electricity use: a high-resolution energy demand model, Energy Build. 42 (2010) 1878–1887.
[8] I. Richardson, M. Thomson, Integrated simulation of photovoltaic micro-generation and domestic electricity demand: a one-minute resolution open-source model, Proc. Inst. Mech. Eng. A: J. Power Energy 227 (2013) 73–81.
[9] A. Navarro, L.F. Ochoa, P. Mancarella, D. Randles, Impacts of photovoltaics on low voltage networks: a case study for the North West of England, in: 22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013), 2013, pp. 1–4.
[10] EA Technology, Assessing the Impact of Low Carbon Technologies on Great Britain's Power Distribution Networks, 2012 (accessed 23.05.14).
[11] A. Collinson, LV Network Management – A Challenge for Network Designers, SP Networks. Presentation given at LCNF Conference Brighton, November 2013, 2014 (accessed 23.05.14).
[12] E. McKenna, M. Krawczynski, M. Thomson, Four-State Domestic Occupancy Model v1.0, 2014, Available at: https://dspace.lboro.ac.uk/2134/15726.1.0
[13] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (2009) 1819–1835.
[14] A. Capasso, W. Grattieri, R. Lamedica, A. Prudenzi, A bottom-up approach to residential load modeling, IEEE Trans. Power Syst. 9 (1994) 957–964.
[15] M. Stokes, M. Rylatt, K. Lomas, A simple model of domestic lighting demand, Energy Build. 36 (2004) 103–116.
[16] C. Jardine, Synthesis of High Resolution Domestic Electricity Load Profiles, 2008, pp. 1–8.
[17] R. Yao, K. Steemers, A method of formulating energy load profile for domestic buildings in the UK, Energy Build. 37 (2005) 663–671.
[18] J. Page, D. Robinson, N. Morel, J. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy Build. 40 (2008) 83–98.
[19] J. Widén, A.M. Nilsson, E. Wäckelgård, A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand, Energy Build. 41 (2009) 1001–1012.
[20] M.A. López-Rodríguez, I. Santiago, D. Trillo-Montero, J. Torriti, A. Moreno-Munoz, Analysis and modeling of active occupancy of the residential sector in Spain: An indicator of residential electricity consumption, Energy Policy 62 (2013) 742–751.
[21] M. Muratori, M.C. Roberts, R. Sioshansi, V. Marano, G. Rizzoni, A highly resolved modeling technique to simulate residential power demand, Appl. Energy 107 (2013) 465–473.
[22] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Build. Simul. 4 (2011) 149–167.
[23] P.D. Andersen, A. Iversen, H. Madsen, C. Rode, Dynamic modeling of presence of occupants using inhomogeneous Markov chains, Energy Build. 69 (2014) 213–223.
[24] C. Liao, Y. Lin, P. Barooah, Agent-based and graphical modelling of building occupancy, J. Build. Perform. Simul. 5 (2012) 5–25.

[25] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, F. Descamps, A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison, Build. Environ. 75 (2014) 67–78.
[26] P. Chenda Liao, Barooah, An integrated approach to occupancy modeling and estimation in commercial buildings, Am. Control Conf. (ACC) 2010 (2010) 3130–3135.
[27] U. Wilke, F. Haldi, J. Scartezzini, D. Robinson, A bottom-up stochastic model to predict building occupants' time-dependent activities, Build. Environ. 60 (2013) 254–264.
[28] Ipsos-RSL and Office for National Statistics, United Kingdom Time Use Survey, [Computer File], 3rd edition, UK Data Archive [distributor], September 2003. SN: 4504, Colchester, Essex, 2003.