Probabilistic Bottom-up Modelling of Occupancy and Activities to Predict Electricity Demand in Residential Buildings

THÈSE N^{O} 5673 (2013)

PRÉSENTÉE LE 21 FÉVRIER 2013

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT
LABORATOIRE D'ÉNERGIE SOLAIRE ET PHYSIQUE DU BÂTIMENT
PROGRAMME DOCTORAL EN ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Urs WILKE

acceptée sur proposition du jury:

Prof. A. Schleiss, président du jury
Prof. J.-L. Scartezzini, Dr F. Haldi, directeurs de thèse
Prof. M. Bierlaire, rapporteur
Dr R. Korsholm Andersen, rapporteur
Prof. R. Madlener, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

To the loving memory of my mother.

Abstract

In Switzerland and most other developed countries, the building sector is one of the most important sources of energy consumption and its related adverse environmental impacts. As the reduction of greenhouse gas emissions and the abandonment of nuclear power have become key policy objectives, considerable effort is undertaken to reduce energy demand, particularly in buildings, and to expand the use of renewable energies. For this reason, it is crucial to develop strategies and solutions, which optimally exhaust the possibilities to improve the efficiency of these systems.

Dynamic simulation models are increasingly used to gain a more precise understanding of the underlying processes of energy flows in buildings. However, many of the fundamental phenomena are still not sufficiently understood, resulting in potentially significant errors. The description of occupants' behaviour still leaves substantial room for improvement in the simulations; in particular, there is a lack of comprehensive and validated stochastic models predicting residential occupancy and activities, as well as their variations between individuals and households. However, the stochastic nature of residential behaviour is central regarding uncertainties in residential buildings' energy demand.

This thesis develops adequate bottom-up models to predict time-dependent residential occupancy and activities, as well as household appliance ownership as a function of individual characteristics, and further proposes an innovative approach to relate the use of electrical appliances to the activities performed. The models are calibrated with detailed survey statistics of individuals' time use, as well as households' appliance ownership and appliances' power consumption.

The approaches to predict presence are based on time-inhomogeneous first- and higher-order Markov processes, where sub-population-specific behaviour is represented by corresponding parameters in the time-dependent transition probabilities and duration distributions. These explanatory variables have been rigorously selected on the basis of statistical significance using backward elimination. The higher-order approach is validated by establishing the relationship to the first-order model, and by comparing the results. The variation of the model predictions for different sub-populations is illustrated and discussed.

Similar approaches have been applied to simulate time-dependent residential activities, based on multinomial logit models predicting starting probabilities for

activities, and modelling their durations by means of survival analysis. An initial model reproducing average population behaviour was refined by including explanatory variables relating to demographic sub-populations, as well as by describing activity transitions according to the Markov property. The explanatory variables have also been selected on the basis of statistical relevance, and cross-validation tests show that the model refinements improve the predictive power.

Approaches to the modelling of the probability of electrical appliance ownership as a function of household characteristics are then presented. For the elimination of insignificant predictors due to correlation effects, two different methodologies are presented, which predictions are validated, and their predictive power compared.

Finally, an approach has been elaborated, providing an activity-dependent prediction of electrical appliance use. Single appliances' power consumption is assigned according to their monitored distributions. The approach has then been applied to predict the residential load profile distribution of simultaneously used appliances, which are in good agreement with the measured data.

It is demonstrated how these models can be integrated into a bottom-up modelling framework, allowing to directly investigate the dependence of the residential electricity load profile distribution on the household's specificities, individual appliances' power demand characteristics, as well as future behavioural changes. This versatile approach results in a valuable methodology to predict electricity demand profiles in various scenarios, which can be used, for instance, to assess the reliability of decentralised power generation infrastructures, being subject to considerable fluctuations.

Keywords

Building simulation, Behavioural modelling, Residential occupancy, Load profile of electricity usage, Markov process, Discrete choice modelling, Survival analysis, Principal components

Résumé

En Suisse, ainsi que dans la plupart des pays industrialisés, le secteur du bâtiment est l'une des sources principales de consommation d'énergie, impliquant des impacts environnementaux. La réduction des émissions de gaz à effet de serre et l'abandon de l'énergie nucléaire étant devenus des objectifs clés sur le plan politique, un effort significatif est nécessaire afin de réduire la demande énergétique, en particulier dans le secteur du bâtiment, et d'accroître l'utilisation des énergies renouvelables. Pour cette raison, il est crucial de développer des stratégies qui exploitent de manière optimale les possibilités d'amélioration de l'efficacité de ces systèmes.

L'utilisation croissante de modèles de simulation dynamique, a pour but d'obtenir une compréhension plus précise des processus sous-jacents aux flux d'énergie et à l'utilisation d'électricité dans les bâtiments. Cependant, beaucoup de phénomènes fondamentaux ne sont pas encore suffisamment compris, ce qui peut mener à des erreurs significatives. La description du comportement des utilisateurs dans les simulations peut être considérablement améliorée ; on constate un manque de modèles stochastiques vérifiés prédisant la présence et les activités résidentielles, ainsi que la variabilité entre individus et ménages de ces dernières. Pourtant, le caractère stochastique du comportement résidentiel est essentiel pour comprendre et prédire les fluctuations de la demande énergétique et électrique des bâtiments résidentiels.

Cette thèse développe des modèles désagrégés adéquats pour modéliser l'utilisation des appareils électroménagers, ainsi que pour simuler la présence et l'activité en fonction du temps et des caractéristiques individuelles des personnes. Elle présente ensuite une approche novatrice associant l'utilisation des appareils électriques aux activités pratiquées. Les modèles sont calibrés sur la base de statistiques détaillées sur l'utilisation du temps, la possession d'appareils électroménagers et la consommation individuelle mesurée de divers appareils.

Les modèles prédisant la présence des individus sont basés sur des processus de Markov d'ordre premier ou supérieur et non-homogènes. Le comportement caractéristique d'une sous-population est capturé par des paramètres spécifiques compris dans la définition des probabilités de transition et des distributions de durées. Ces variables indépendantes sont rigoureusement choisies par élimination descendante. Le modèle d'ordre supérieur est validé en le confrontant avec celui de

premier ordre. La variation des prédictions relatives aux diverses sous-populations est illustrée et discutée.

En outre, une approche similaire a été appliquée, afin de prédire les activités résidentielles en fonction du temps, à l'aide de modèles logit multinomiaux quantifiant les probabilités de commencement des activités. Les durées des activités sont modélisées sur la base d'une analyse de survie. Un premier modèle reproduisant la moyenne comportementale de la population a ensuite été détaillé en incluant des variables indépendantes représentant les sous-populations démographiques, puis en introduisant les transitions entre activités ayant la propriété de Markov. Seules les variables indépendantes significatives ont été sélectionnées, et des tests de validation croisée indiquent que les modèles détaillés ont une capacité de prédiction supérieure.

Par ailleurs, la modélisation des probabilités de possession d'appareils électroménagers en fonction des caractéristiques des ménages est présentée. Afin d'éliminer les variables indépendantes devenues insignifiantes par effet de corrélation, deux méthodologies différentes sont proposées. Leurs prédictions sont validées, incluant une comparaison de leur capacité prédictive.

Finalement, une approche a été élaborée, pour modéliser l'utilisation des appareils électroménagers en fonction de l'activité des personnes. La puissance électrique individuelle des appareils est prédite en fonction de leur distribution empirique. Cette approche a ensuite été appliquée afin d'inférer la distribution du profil de charge total des appareils utilisés simultanément, avec des résultats en accord avec les mesures.

L'intégration de ces modèles dans un système de modélisation désagrégée permet d'étudier directement la dépendance du profil de charge aux spécificités des ménages, aux caractéristiques individuelles de puissance consommée des appareils, ainsi qu'aux possibles changements comportementaux. Cette approche polyvalente mène à une méthodologie concrète permettant de prédire des profils de charge dans divers scénarios, qui peuvent être utilisés, par exemple, pour évaluer la fiabilité des infrastructures énergétiques décentralisées, qui sont sujettes à des fluctuations de puissance considérables.

Mots-Clés

Simulation du bâtiment, Modélisation du comportement, Présence résidentielle, Profil de charge, Utilisation d'appareils électriques, Processus de Markov, Modélisation des choix discrets, Analyse de survie, Composantes principales.

Zusammenfassung

In der Schweiz und anderen Industrieländern ist der Gebäudesektor einer der wesentlichen Verursacher von Energiebedarf, sowie den damit verbundenen Umweltbelastungen. Da die Verringerung der Treibhausgasausstöße und der Ausstieg aus der Nuklearenergie eines der wesentlichen politisches Ziele darstellt, bedarf es erheblicher Anstrengungen, den Energieverbrauch – insbesondere im Gebäudesektor – zu verringern und erneuerbare Energiequellen besser auszuschöpfen. Deshalb ist es von zentraler Bedeutung, Strategien und Lösungswege zu erarbeiten, um möglichst wirksam die Effizienz der relevanten Abläufe und Systeme zu verbessern.

Dynamische Simulationsmodelle werden immer häufiger verwendet, um ein besseres Verständnis der zugrunde liegenden Abläufe von Energieflüssen und Elektrizitätsnutzung in Gebäuden zu erlangen. Die fehlende quantitative Beschreibung vieler relevanter Phänomene kann jedoch erhebliche Fehler in den Berechnungen nach sich ziehen. Hierbei bietet die Beschreibung des Verhaltens von Personen in Gebäuden beträchtliches Verbesserungspotenzial; insbesondere existieren bisher keine fundierten stochastischen Modelle, welche die Anwesenheit und die Aktivitäten einzelner Personen in Gebäuden vorhersagt, sowie Unterschiede in Abhängigkeit der Charakteristika der Personen oder der Haushalte. Die stochastische Natur im Verhalten von Personen ist jedoch von wesentlicher Bedeutung, hinsichtlich der Schwankungen des Energie- und Elektrizitätsbedarfes in Wohngebäuden.

In dieser Doktorarbeit werden zweckentsprechende Bottom-up-Ansätze ausgearbeitet, mit welchen der Besitz von Elektrogeräten in Haushalten, sowie die zeitabhängige Anwesenheit und Aktivitäten in Haushalten vorhergesagt werden können, in Abhängigkeit der Charakteristika der Einzelpersonen und Haushalte. Desweiteren wird ein innovativer Ansatz präsentiert, mit welchem ein Zusammenhang zwischen dem Gebrauch von Elektrogeräten sowie der währenddessen nachgegangenen Aktivitäten hergestellt wird. Diese Modelle stützen sich auf detaillierte Daten aus Erhebungen bezüglich des Zeitbudgets, des Besitzstandes von Elektrogeräten und von Verbrauchsmessungen einzelner Elektrogeräte in Haushalten.

Die Modelle zur Vorhersage der Anwesenheit in Wohngebäuden basieren auf zeitlich inhomogenen Markow-Prozessen erster und höherer Ordnung, bei denen

individuell geprägtes Verhalten von Subpopulationen dargestellt wird durch entsprechende Parameter in den zeitabhängigen Übergangswahrscheinlichkeiten und Dauerverteilungen. Die erklärenden Variablen wurden gemäß dem Kriterium der statistischen Signifikanz in einer Methode der Rückwärtselimination ausgewählt. Der Ansatz höherer Ordnung wurde validiert, indem die das Modell erster Ordnung bestimmenden Größen in jene des ersteren übersetzt wurden. Die geringfügigen Differenzen zwischen den Vorhersagen beider Modelle sind auf Näherungen in der Berechnungsmethode zurückzuführen. Die Abhängigkeit der prognostizierten Verteilungen der Anwesenheitsprofile wird für verschiedene Subpopulationen veranschaulicht und erläutert.

Die Modelle zur Vorhersage von Aktivitäten an Wohnorten basieren auf ähnlichen Ansätzen, wobei Aktivitätsbeginne durch Multinomial-Logit-Modelle und deren Dauern durch Verweildaueranalyse modelliert werden. Ein Ausgangsmodell, welches das durchschnittliche Verhalten reproduziert, wurde verfeinert, indem Erklärungsvariablen eingebunden wurden, welche das spezifische Verhalten von demografischen Subpopulationen und Übergänge zwischen Aktivitäten gemäß der Markow-Eigenschaft widerspiegeln. Auch hier wurden Erklärungsvariablen eliminiert, welche keine statistische Signifikanz aufwiesen. Tests von Kreuzvalidierungsverfahren zeigen, dass die verfeinerten Modelle eine verbesserte Vorhersagekraft aufweisen.

Danach werden Ansätze zur Vorhersage des Besitzstandes verschiedener Elektrogeräte in Abhängigkeit von Haushaltscharakteristika vorgeschlagen. Zwei Methodiken werden präsentiert, in denen statistisch unbedeutende Erklärungsvariablen eliminiert werden, welche daraufhin validiert werden und dessen Vorhersagekraft miteinander verglichen wird.

Schließlich wird eine Methodik ausgearbeitet, in der die Benutzung von Elektrogeräten als bedingte Wahrscheinlichkeit von zu diesem Zeitpunkt nachgegangenen Aktivitäten definiert ist. Der Elektrizitätsverbrauch einzelner Elektrogeräte wird hergeleitet aus den entsprechenden empirisch gemessenen Verteilungen. Es wird dann ein darauf basierender Ansatz präsentiert, mit dem die Verteilung des elektrischen Lastverlaufs von gleichzeitig betriebenen Elektrogeräten hergeleitet werden kann. Die Prognosen des Modells stimmen gut mit den gemessenen Werten überein.

Diese Arbeit zeigt auf, wie die einzelnen Modelle in einem gemeinsamen Rahmen miteinander kombiniert werden können. Dies erlaubt es die unmittelbaren Abhängigkeiten der statistischen Verteilung des Lastprofils von Eigenschaften des Haushaltes und der Verbrauchswerte der einzelnen Elektrogeräte, sowie zukünftigen Verhaltensveränderungen abzuleiten. Die vielseitig anpassbare Methodik stellt einen wissenschaftlich wertvollen Ansatz dar, mit welchem die Verlässlichkeit von Infrastrukturen der dezentralisierten Elektrizitätsversorgung untersucht werden kann, welche beträchtlichen Produktionsschwankungen unterliegen können.

Schlagworte

Gebäudesimulation, Verhaltensmodellierung, Wohnplatzanwesenheit, elektrisches Lastprofil, Markow-Prozess, diskrete Entscheidungsmodelle, Ereigniszeitanalyse, Hauptkomponentenanalyse

Acknowledgements

I am very much grateful to Jean-Louis Scartezzini and Darren Robinson for their trust in me and for offering me the opportunity to begin a doctorate at the Solar Energy and Building Physics Laboratory, where I had the honour and privilege to spend an important part of my life in a very competent and experienced team of wonderful colleagues.

I am highly indebted to my supervisor Jean-Louis Scartezzini for his support and advice, his true leadership qualities, from which all employees and the laboratory benefit, as well as his patient confidence, which assured a high degree of motivation and the strengthening of a pronounced interest in my research topics.

Furthermore, I am very grateful to my former supervisor Darren Robinson, who fostered the development of some of the key subjects in this thesis for which I developed great enthusiasm and interest. His ideas, the careful re-reading and his elaborate rhetoric skills have significantly contributed to the quality of this work.

I especially thank my co-supervisor Frédéric Haldi for his trust in me and for always being extremely supportive – personally, technically, as well as with regard to the general strategic research orientation – which assured my way in this work most substantially. His ample theoretical knowledge often allowed me to rapidly depict a feasible way to put ideas on a sound basis, which accelerated the progress of this work considerably.

I express my great gratitude to the members of the jury, Anton Schleiss (Laboratory of Hydraulic Constructions, EPFL), Michel Bierlaire (Transport and Mobility Laboratory, EPFL), Reinhard Madlener (E.ON Energy Research Center, RWTH Aachen University) and Rune Korsholm Andersen (Department of Civil Engineering, Technical University of Denmark) for their participation in the examination, their careful reading and their constructive comments that significantly improved this thesis.

I would also like to thank Stéphane Ploix (Grenoble Institute of Technology), as well as Kurt Wiederkehr (VSE, Association of Swiss Electricity Providers) for providing data which were essential to the development of some of the models in this thesis. Furthermore, I cordially acknowledge the financial support received from the Swiss National Science Foundation, without which this work would not have been possible.

Special thanks go to Diane Perez, with whom I shared my office for almost four years, as well as Maria Papadopoulou. Both of them were always very helpful and contributed to a pleasant atmosphere, thereby maintaining my enthusiasm.

I also would like to thank my colleagues in the laboratory, who contributed in making it a joyful and charming place to work and study at. It was a privilege to work and share time with David Daum, Friedrich Linhart, Apiparn Borisuit, Nikos Zarkadis, Jérôme Kämpf, Philippe Leroux, Adil Rasheed, Chantal Basurto, Marja Edelmann, Laurent Deschamps, Wanjing Li, Lenka Maierova, Pierre Loesch, Christian Roecker, Mirjam Münch and Nicolas Morel. Many thanks also go to the group of Andreas Schüler, and his collaborators Martin Joly, Antonio Paone, Stefan Mertin, André Kostro, Virginie Le Caër, Luc Burnier, Nicolas Jolissaint and Thomas Gascou.

I would also like to whole-heartedly thank the secretaries of our laboratory Sylvette Renfer, Suzanne l'Eplattenier and Barbara Smith for always being helpful in all kinds of matters, and for always spreading a delightful atmosphere with their friendliness.

I was furthermore very fortunate to have studied with my former fellow students, Johannes Hauk, Henning Gutzmann, Martin Münning, Boris Wolter and Thim Stapelfeldt, as well as the research group of Stefan Heinze with whom I worked on my diploma thesis in Hamburg.

One of the most imporatant things during my doctoral studies was the support of my friends and the uncountable nice moments they spent with me, which helped me not getting lost. I especially thank Jonas Gerking, Jonas Dallinger, Tanja Suworin, Kevin Kosche, Paul Friedrich, Stefan Weiß, Patrizio Araya, Filmon Goitom, Johannes Hauk, Manuel Rupprecht, Martin Münning, Felix Köbisch, Henning Gutzmann, Stefan Sander, Dean Rožić, Michiel Kindt, Thomas Niggli, Burkhard Schiess, Younes Razama, Dirk Gansefort, Jannes Nöldeke, Boris Wolter and Thim Stapelfeldt.

My greatest thanks go to my father, my brothers Wolf and Moritz, my grandmother, as well as Gisela, Christoph, Kim and Kaja.

Lausanne, 15 February 2013

Contents

Chapter 1

Introduction

1.1 General context

As ever since the beginning of industrialisation, energy security has been of central importance in policy. During the last decades, energy demand has attracted increasing concern, its mitigation being a top government priority today. Since the Fukushima Daiichi nuclear disaster in 2011, it was agreed in many countries to rely less extensively on nuclear power generation, which necessitates a more intensive use of alternative energy sources. The need for significant structural changes regarding economic, societal and in particular environmental aspects is now commonly acknowledged, considering predictions of global economic and population growth. In order to support a more sustainable development, one of the most important tasks is research and development that foster a more efficient and parsimonious energy use.

1.1.1 Energy in residential buildings

Residential buildings are responsible for a very important fraction of energy use. According to estimations, this sector is responsible for 16 to 50 % of the global energy demand across countries [1]. In the 27 member states of the European Union, it accounted for about one quarter of the total energy use in the year 2007 [2], showing a steady increase during the last decade [3]. This growth is caused by numerous factors related to improved living standards, such as the increasing use of active cooling and electrical appliances. Regarding the total electricity demand, the residential sector is responsible for more than one quarter, which increased by 1.6 % per year from 708 TWh to 801 TWh in the period from 1999 to 2007 [4].

Therefore, the improvement of the energy performance in buildings, as well as of the development and promotion of decentralised electricity generation from renewable sources is becoming increasingly important. As the latter is already taking place [5], considerably more effort has to be devoted to research assuring

the security of electricity supply. As furthermore, the building envelope efficiency improves, the part of energy used by electrical appliances, as well as their contribution to casual heat gains in buildings increase. It is thus crucial to develop a sound theoretical basis, that accurately describes the comprehensive set of energy-related processes in buildings, as well as in electrical grids, in order to realistically predict the systems' responses to changes of the boundary conditions.

1.1.2 State of the art

Building performance simulation

Much effort has been put into the development of dynamic simulation programs, in order to investigate the dependence of the energy demand of buildings on their characteristics and those of the environment. Pioneering research was provided by Winkelmann and Selkowitz [6], Arumí-Noé and Northrup [7], Clarke [8] and Gough [9], where it was focussed on the dynamical modelling of energy exchanges of buildings with the exterior environment. The predictions of heating energy demand and air temperature issued from multiple simulation tools have already been empirically validated for different cases [10]. The development of dynamic thermal building simulation programs has led to sophisticated tools like ESP-r, which allows for the simultaneous simulation of fluid flow, heating, ventilating and air-conditioning, as well as energy conversion and control systems [11].

Whereas deterministic processes in buildings are relatively well captured in dynamic simulation programs like ESP-r, influences originating from building occupants are insufficiently described. For instance, field surveys showed that in a sample of 28 equally designed houses, there were variations in gas consumption in winter of up to a factor of two [12]. In a more recent study, the average electricity demand of nine identical houses was varying by a factor of up to three and six, respectively on an annual, and a monthly basis [13]. In a study conducted by Iwashita and Akasaka [14], it was found that 87 % of the total air change in the investigated dwellings was caused by occupant behaviour.

As there is a considerable demand of energy-efficient buildings [15], the weight of user behaviour on the buildings' energy balance is increasing. This underlines the need of detailed behavioural models, that allow for a more accurate description of the variation of specific users, in order to optimise building design [16]. Interesting approaches for the prediction of occupants' behaviour towards the building envelope were recently developed, but their use of electrical appliances is much less investigated.

Electricity demand

Since the origin of electrification, electricity loads were studied in the context of planning purposes and electricity pricing [17]. Regarding building energy simu-

lation, the main purposes of electricity demand prediction are:

- The integration of all important energy-related processes.
- The correct prediction of casual heat gains.
- The dimensioning of on-site generation of renewable energy systems.

Today, residential electricity use takes up an important part in total residential energy use (see Section 1.1.1) and, furthermore, electricity generation from renewable sources is steadily increasing [2]. With the prospect of the associated increasingly decentralised electricity supply to the residential sector, there is increasing interest in topics like active load management, as well as micro-generation and local energy storage technologies [18]. This underlines the requirement of a detailed understanding of the involved processes, in order to assure competitive performance of micro-grids. For the latter, the reliable matching of electricity generation with the demand is of significantly higher complexity than on large scales, due to an increased degree of statistical fluctuations. To address this, numerous political, academic and economic efforts are aimed at fostering micro-grid technologies [19]. Regarding the intensified efforts that the electricity supply is to rely more on renewables energies and less on conventional power generation infrastructure, the power supplied becomes less constant, which underpins the increased need to match demand and supply. Here, one of the main challenges consists in an accurate description of the stochastic nature of the latter two [5, 20–25].

1.2 Scope of this work

Figure 1.1 summarises interactions between human beings and building components, as well as electrical appliances. Interactions with window openings enhance natural ventilation. Actions on manually controlled shading devices influence solar heat gains, as well as the use of artificial lighting. The latter, as well as the use of many electrical appliances result in considerable impacts on electricity demand. The use of electrical appliances and individuals present also involve casual heat gains which have, together with natural ventilation and solar gains, important influences on heating demand. The latter also depends considerably on the insulation of the building envelope. As all the above interactions are related to residential presence and activities, and much effort is spent to improve building insulation in the last years, the relative importance of the impact of human behaviour on buildings' energy demand increases.

Although there has been considerable effort to support the dynamic stochastic modelling of individuals' residential occupancy, activities and the use of electrical appliances, as well as actions on building components [e.g., 26–29], there is no existing approach which accurately accounts for significant variations between

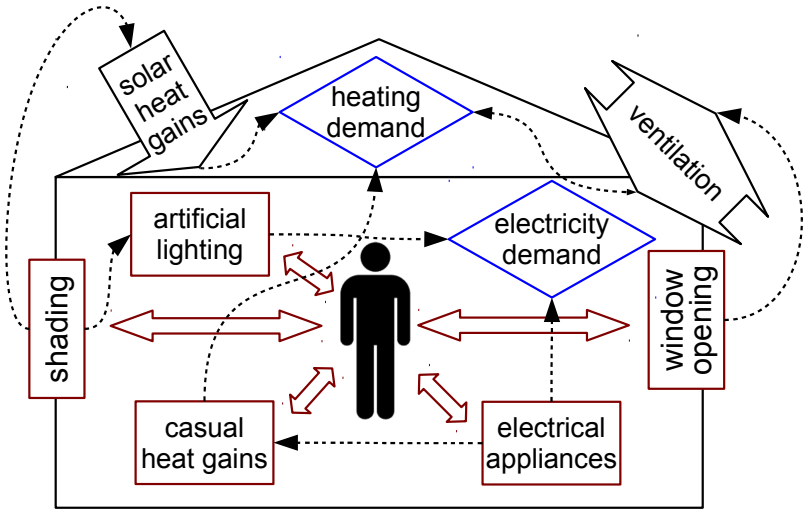


Figure 1.1: Interactions between humans and building components with impacts on heating and electricity demand.

individuals. Either the calibration procedures are designed to match average behaviours, or the level of detail was not tested for statistical significance of parameters. When applying such models, this can often lead to model predictions, which tend to reproduce either average behaviour, or excessively that of individuals of the dataset, with additional statistical noise, resulting in inaccurate distributions.

The scope of this work is the development of models predicting residential presence and activities, as well as the ownership and the use of individual electrical appliances. Much attention will be put on the trade-off between the level of detail and parsimony, simultaneously ensuring model robustness and a high level of accuracy of the predicted time-dependent stochastic variables, in order to develop a generalist approach which is not restricted to the calibration dataset.

1.3 Hypothesis

Current models of residential load prediction and building performance simulation are undermined by an insufficient accounting of variations in the time-dependence of residential occupancy and activities as a function of individual characteristics. However, behavioural patterns are complex regarding their mutual interactions to each other. These interactions complicate the basis for a stochastic treatment, which is required to capture the sources of variability. Therefore, we propose the following hypothesis (also shown in Figure 1.2) to guide our developments:

> **The electricity demand of households can be optimally modelled as a result of the activities performed by**

their occupants. Residential use of electrical appliances is determined by residential activities, which are themselves determined by residential occupancy.

This approach enables a generalisable formulation of the model and a correct prediction of fine resolution power demand accounting for occupants' behavioural diversity and the considered contexts. Furthermore, a joint probabilistic approach considering the stochastic nature of occupancy, activity and appliance use patterns enables an estimation method of the variability of electricity demand.

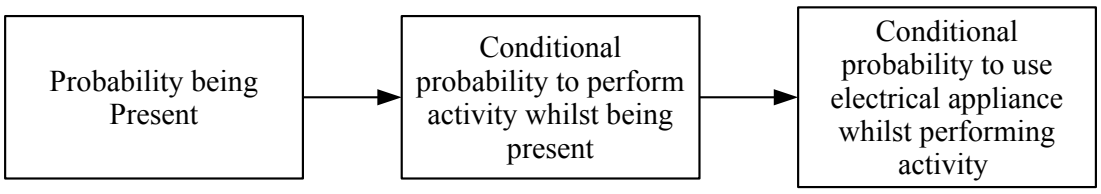


Figure 1.2: Dependences and influences (indicated by arrows) of the different sub-models.

1.4 Structure of this work

In Chapter 2, we present a detailed description of the general characteristics, as well as the time-dependence of the data from the time use survey that was used to calibrate the models of Chapters 3 and 4. We then present in Chapter 3 the model developed to predict the residential occupancy, as a function of time and individual specificities. We go on to describe the model that predicts the time-dependent activities of residential occupants, which also was formulated in dependence of individual characteristics (Chapter 4).

As a next step, in Chapter 5, we present a model to predict the ownership of multiple electrical appliances, depending on the household characteristics. Finally, a detailed methodology resulting from the above stochastic models is developed to predict the activity- and occupancy-dependent use of electrical appliances, allowing to construct the load profile distribution of their simultaneous use. The general bottom-up[1] modelling framework is discussed and synthesised in Chapter 7.

[1]Disaggregated models will be referred to as "bottom-up" models.

Chapter 2

Time use data

This chapter presents the time use data that provided the basis for the calibration of the models to predict residential occupancy and activities. The general included informations, as well as the statistical properties of the included population characteristics in the data are described. Furthermore, the time-dependence of residential presence and activity patterns are described in detail. The relevance of the dataset for our research method is investigated, together with a detailed discussion of statistical artefacts, arising from the measurement methodology and error-prone data.

2.1 Multinational time use study database

The time use data that are considered in this work are available in electronic format in a database that is managed by the Centre for Time Use Research of the University of Oxford [30]. This Multinational Time Use Study (MTUS) database contains the harmonised information of multiple time use surveys that were conducted in the period from 1961 to 2011 in Australia, Austria, Belgium, Bulgaria, Canada, Denmark, Finland, France, Germany, Hungary, Israel, Italy, the Netherlands, Norway, Slovenia, South Africa, Spain, Sweden, the United Kingdom and the United States [31].

The data collected in the surveys comprises information of individuals who completed questionnaires describing the chronological course of activities j_{MTUS} in their diaries in time increments of varying duration (depending on the survey) throughout 24 h, starting at varying times (also depending on the survey). The activities j_{MTUS} are specified according to a list of 41 different activity types (called the "MTUS 41-activity typology", *cf.* Figure 2.7) and, furthermore, the information in which type of place y the respondents were located can take 9 different values (*cf.* Figure 2.6), which will be described in more detail in Section 2.1.1. As a result of the harmonisation of the data of the different TUSs,

not all of the characteristics in the MTUS are available in the datasets of the national surveys.

In Figure 2.1, we show an example of the activity chains $a_{\boldsymbol{x}}(t)$ and occupancy chains $y_{\boldsymbol{x}}(t)$, describing in which types of place they were performed as a function of time for three different individuals $\boldsymbol{x}$ of the database (the legend will be provided in Figures 2.6 and 2.7). As the activity and location codes of the surveys are discretised according to the mentioned typologies, the temporal information of each individual corresponds to a staircase function of different episodes. The marked data points indicate times where a new activity is started.

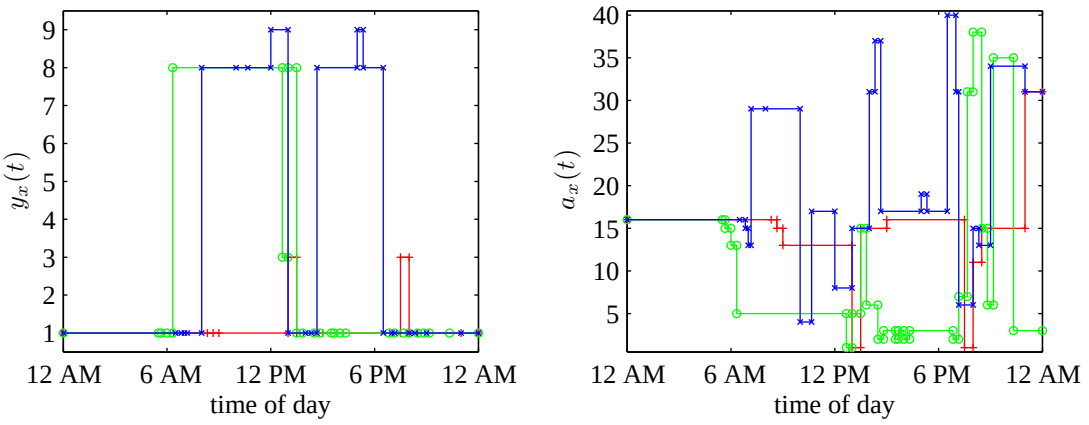


Figure 2.1: Examples of occupancy (left) and activity (right) chains of three individuals of the French TUS.

In addition to the diary plans recorded by the respondents, a detailed list of demographic and socio-economic characteristics is included in the database, containing information such as the timing of the survey, the household composition, the respondent's employment/education status, as well as health classifications. The variable names together with a short description of these characteristics is provided in Table 2.1. Each of these characteristics can take several discrete values. In non-trivial cases, these are shown in Table 2.2, where in some cases, the original value set was aggregated to a coarser resolution.

Table 2.1: List of diary, demographic and socio-economic characteristics in the database.

variable	**description**
countrya	Country or region of study
country	Country where study was conducted
survey	Year the survey began
swave	Longitudinal study wave marker
msamp	Multiple samples using the same diary instrument

hldid	Household identifier
persid	Person/diarist identifier
id	Diary identifier
parntid1	Person identifier of 1st parent of diarist
parntid2	Person identifier of 2nd parent of diarist
partid	Person identifier of spouse or partner
day	Day of week diary kept
month	Month diary kept
year	Year diary kept
diary	Diary order
badcase	Marker of low-quality cases
hhtype	Household type
hhldsize	Number of people in household
nchild	Number of children under 18 in household
agekidx	Age of youngest child in household (categories including adults)
agekid2	Age of youngest child in household
incorig	Original household income
income	Total household income - grouped
ownhome	Whether diarist's household owns or rents home
urban	Urban or rural household
computer	Does household have a computer
vehicle	Does household have a access to a private vehicle
sex	Sex
age	Age
famstat	Individual level family status
cphome	Unmarried child living in parental home
singpar	Whether diarist is a single parent
relrefp	Relation to household reference person
civstat	Civic status
cohab	Respondent is cohabiting
citizen	Whether the diarist is a citizen of the country
empstat	Employment status
emp	In paid work
unemp	Unemployed
student	Whether diarist is a student
retired	Whether diarist has retired
empsp	Employment status of spouse/partner
workhrs	Hours paid work last week including overtime
empinclm	Original monthly income from employment or self-employment
occup	Occupation

sector	Sector of employment
educa	Educational level-original study code
edtry	Harmonised level of education
rushed	Whether diarist generally feels rushed
health	Diarist's general health
carer	Diarist looks after an adult or child with a disability
disab	Diarist has a disability or long-term limiting health condition

Table 2.2: Value sets of the non-trivial variables in Table 2.1.

variable	**value set**	**description**
hhtype	1	One person household
	couple	Couple alone
	couple+	Couple + others
	other	Other household types
agekidx	no	No children in household
	0-4	Youngest child aged between 0-4
	5-12	Youngest child aged between 5-12
	13-17	Youngest child aged between 13-17
	18+	Youngest child aged 18+
incorig	<3.5	Less than 3,500 (in French Francs)
	3.5-10	3,500 to 10,000
	10-21	10,000 to 21,000
	>21	More than 21,000
	n.s.	Doesn't know.
ownhome	own	Own (outright or on mortgage)
	rent	Rent
	other	Other arrangement
urban	urban	Urban/suburban
	rural	Rural/semi-rural
vehicle	N	No
	1	1 car or motorcycle
	2+	2+ 1 cars or motorcycles
famstat	18-39,-	Adult aged 18 to 39 with no co-resident children <18
	18+,<5	Adult 18+ living with 1+ co-resident children aged <5
	18+,5-17	Adult 18+ living with 1+ co-resident children 5-17, none <5
	40+,-	Adult aged 40+ with no co-resident children <18
	<18,parents	Respondent aged <18 and living with parent(s)/guardian(s)

	<18,n.s.	Respondent aged <18, living arrangement other or unknown
cphome	N	Not a child in parental home
	Y	Child in parental home
civstat	Y	Diarist in couple, lives with spouse/partner
	N	Diarist not in a couple
empstat	full	Employed Full Time
	part	Employed Part Time
	n.s.	Employed, unknown status
	no	Not in paid
empsp	n.s.	n.s.
	full	Employed full-time
	part	Employed part-time
	unknown	Employed unknown hours
	no	Not in paid work
occup	n.s.	n.s.
	Manage	Management (senior management, not supervisors)
	Fin	Finance and legal professionals
	Sci	Science and engineering professionals
	Civil	Civil and social service professionals
	Educ	Education and social science professionals
	Other	Other professionals
	Health	Health, education, and social care support,
	Clerical	Clerical and office support
	Secur	Security and armed forces
	Sales	Sales, services, creative support, and cleaning,
	Farming	Farming, forestry, and fishing
	Construct	Construction, assembly & repair, moving goods, transport, extraction
	Self-empl	Self-employed non-professionals
sector	n.s.	n.s.
	Public	Public sector
	Private	Private sector
edtry	1	uncompleted secondary or less Not completed ISCED level 3
	2	completed secondary Completed ISCED level 3 and/or attendance at level 4
	3	above secondary education ISCED level 5 or above
rushed	n.s.	n.s.
	Almost	Almost never
	Sometimes	Sometimes
	Often	Often

In addition to the episodes, the aggregated time spent with each of the activities is also available for the surveyed individuals. All the specified information is available in different versions of the database. For this work, versions 5.53 and 6.0 of the database have been merged together using the unique pair of household and person identifiers, to make use of the full individual chronological diary information (cf. Figure 2.1; only available in version 6.0) and the full set of demographic and socio-economic characteristics (Table 2.1; only available in version 5.53), that were recorded in this TUS. Furthermore, the classification of respondents being minor is stored in separate files which have also been integrated. For a more detailed description of the database and definition of the characteristics and the corresponding value sets, we refer the reader to the description of the database [31].

2.1.1 French time use survey

General characteristics

The modelling approaches presented in this thesis could be calibrated with data from any of the surveys contained in the MTUS as long as the temporal information of the episodes is available. However, in this work, we only use the data from the French time-use survey (TUS) conducted from 16 February 1998 to 14 February 1999 [32], and collected by the French National Institute of Statistics and Economic Studies [33], as the country and the period of this survey corresponds to those of the electricity measurements that are used to calibrate the model in Chapter 6.

This dataset relates to a subset of the French population of $n = 15441$ individuals from 7949 households, whose recorded diary plans are in 10 min time increments throughout 24 h, starting and ending at midnight. In this survey, multiple members of the same household were interviewed by means of two household visits. The response rate was 91.1 % for households, and 88.3 % for individuals, whose age range was from 15 to 80 years. Besides primary activities, secondary activities, as well as information on who else was present was also recorded [34].

The distributions of characteristics of this population regarding a variable selection of the lists in Tables 2.1 and 2.2 are presented in Figures 2.2 to 2.5[1]. The weekday the data was recorded is relatively uniformly distributed, ranging from 10.5 % for Mondays to 16.9 % on Thursdays. The months are also relatively uniformly covered, ranging from 7.6 to 11.0 %, apart from March, August and December which lie between 4.1 and 5.6 %. 47 % of the respondents are male, and 24.6 % are retired.

The distribution of the characteristics in the TUS does not exactly fit to that of the whole population. For instance, with a share of 24.8 % of retired persons in the TUS, these persons are over-represented compared to the real population

[1]The variable "cday" denotes the calendar day which is also part of the episode variables.

share of 16.0 % in the year 2000 [35, 36]. Neither the month nor the weekday are accurately covered. In order to apply the models in scenarios whose population characteristics differ from those of the calibration dataset, a disaggregated model formulation is important.

Time-Dependence

In Figure 2.6, we show the proportions of the population of the French TUS being in a given location as a function of the time of day. In the questionnaires of the French TUS, the variable specifying the type of place y was only recorded for the values 1 and 3 (the other values were deduced from the specified activity types in the diary plans). As in this work, only residential presence and activities are investigated, y is treated as a binary variable, reflecting whether y is equal to 1 (at home) or not (which will be denoted 0).

The populations' activity shares are shown in a stacked presentation in Figure 2.7 as a function of the time of day, regardless of the type of place, which we herein refer to as the activity profile. For every activity in the legend, the first number in the parentheses indicates the daily mean percentage of time that activity is conducted, whereas the second indicates the percentage of time the activity is performed at home (the remainder being carried out elsewhere). The 41st activity type "unclassified or missing" does not occur in the French TUS and was therefore not listed in Figure 2.7. According to these statistics then, respondents are on average engaged in conversation for only 1.3 % of their day (as a primary activity); with a little over two thirds of these conversations taking place within the home.

The activity profile often shows kinks at half and full hours, due to a rounding of the time values by the respondents (respondents appear to prefer to allocate activities to 30 min or 60 min intervals than to 10 min ones or to bias activity starts to these time units; *cf.*, Figure 4.2). Furthermore, the shares of activities conducted at the end of the day often differ significantly from those at the beginning of the day. Again this is due to erroneous allocations of activities by the subjects at the boundaries of the day of the questionnaire. This issue is overcome to a large extent in other surveys, by defining the beginning/end of the monitored time period early in the morning (*e.g.*, 4 am in the US-American TUS [37]).

In this work, only the actions which take place whilst the individuals are at their own homes are investigated. This information may also be error-prone (probably due to errors in the data processing procedures to generate the harmonised activity codes) rather than in the responses of the subjects), as some of the activities (*e.g.*, "visit friends at their homes") have a non-zero share which is performed at the own home, although excluded by definition. However, these errors are only in the order of magnitude of one percent and thus will be ignored.

For reasons of clarity and readability, as well as to simplify the estimation of the multinomial logit models that will be presented in Chapter 4, certain of these

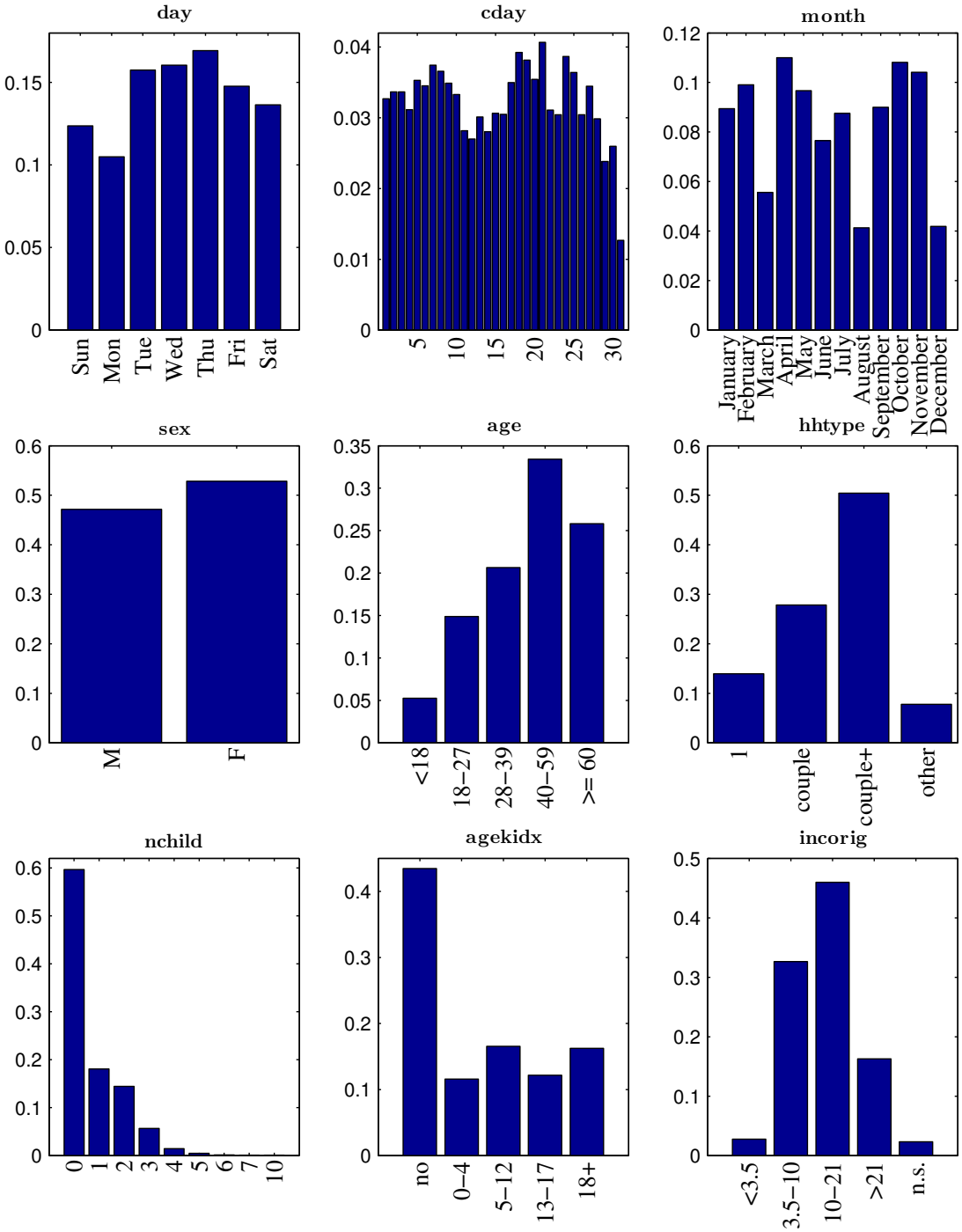


Figure 2.2: Distribution of the characteristics of the French TUS (see Table 2.2).

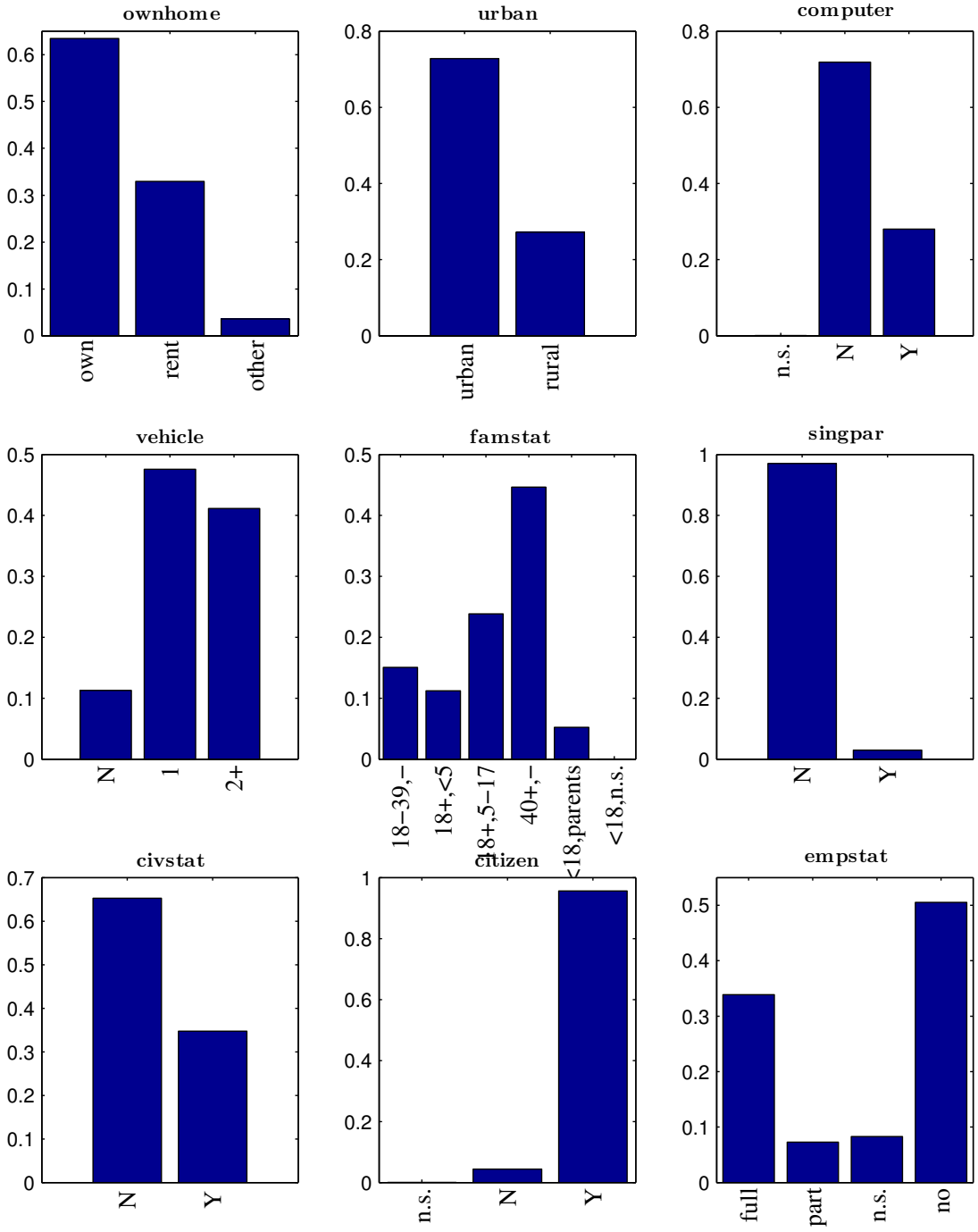


Figure 2.3: Distribution of the characteristics of the French TUS. (see Table 2.2)

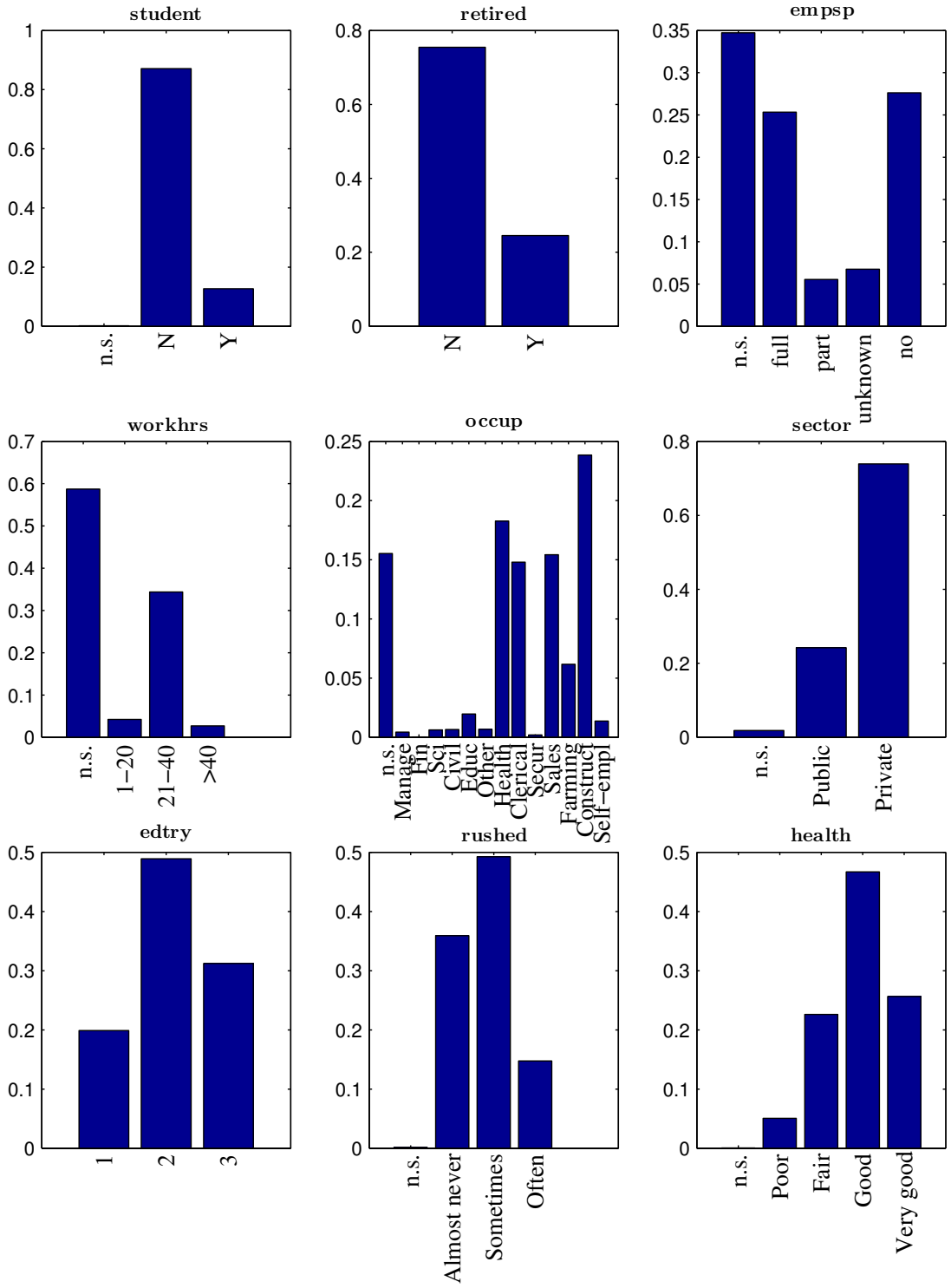


Figure 2.4: Distribution of the characteristics of the French TUS. (see Table 2.2)

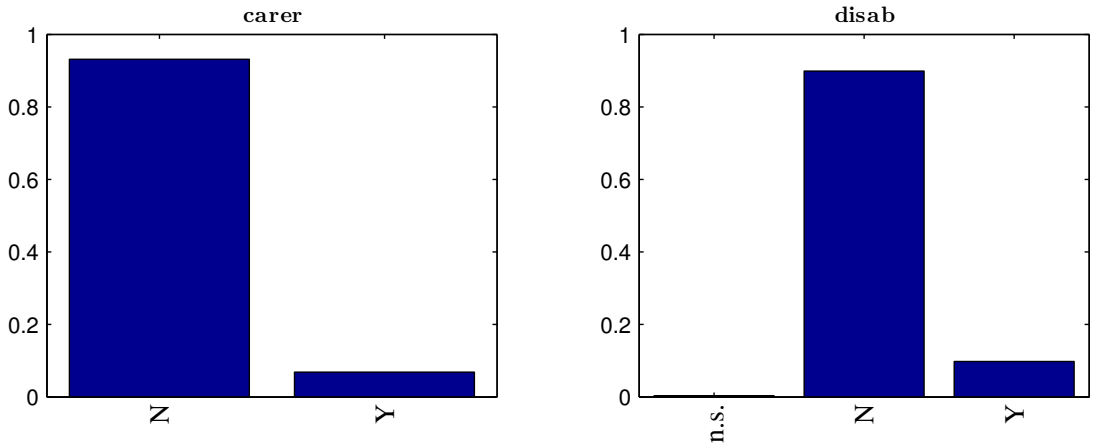


Figure 2.5: Distribution of the characteristics of the French TUS. (see Table 2.2)

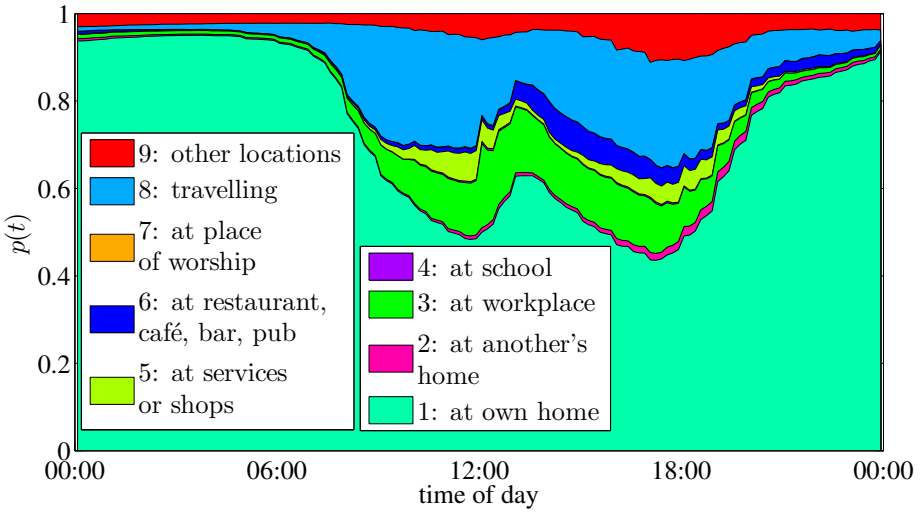


Figure 2.6: Presence profile for the different types of place in the TUS.

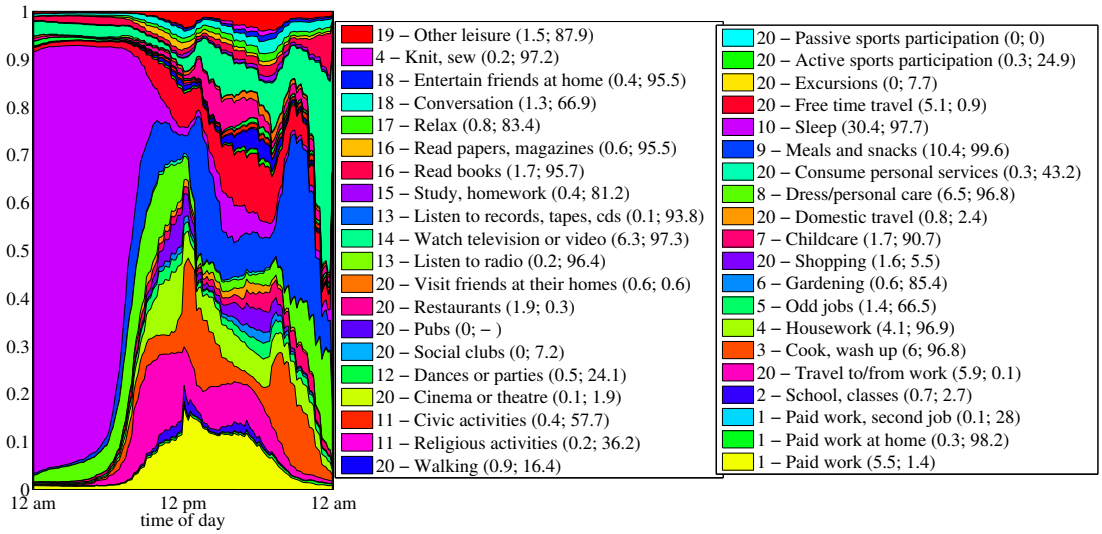


Figure 2.7: Activity profile of j_{MTUS} monitored in the TUS, regardless of the type of place y, where the activities were performed. The first number in the legend shows the index number j of the merged activity category in Figure 2.8. The first number in the parentheses shows the daily mean percentage the activity is performed, the second one which percentage is performed at home.

activities have been merged together in cases where they have a similar impact on a building's energy balance or the use of electrical appliances. These merged activity types are shown in Figure 2.8, which corresponds to the probability distribution of activities that are performed whilst being at home, which will be referred to as residential activity profiles.

In Figure 2.9 we show a summary of median durations of residential activities of the TUS as a function of the hour of the day when they were started, indicated by the height of each single-coloured area. The scarcity of events during the night time amplifies the weight of erroneous recordings in the database leading to a much longer mean duration during this period (we assume that the activity type in the database has sometimes been mistaken for another one when the questionnaire data on paper was copied into electronic format; during the night time this is likely to lead to a replacement of sleeping by another activity, which increases the mean duration of the other activity substantially). Therefore, we do not show the whole range of activities between 12 am and 6 am to focus on the rest of the day which is more reliable (*cf.* Figure 4.3 for an illustration of these truncated activities). Furthermore, the weight of these errors is negligible (cf. Figures 2.8 and 2.10). At the end of the day the means decrease because of the truncation of the questionnaires which stop recording after midnight.

In Figure 2.10 we show a stacked histogram of the numbers the activities that have been started by the n individuals of the TUS sample population as a function of the time interval on the x-axis. Between 1 am and 5 am the total

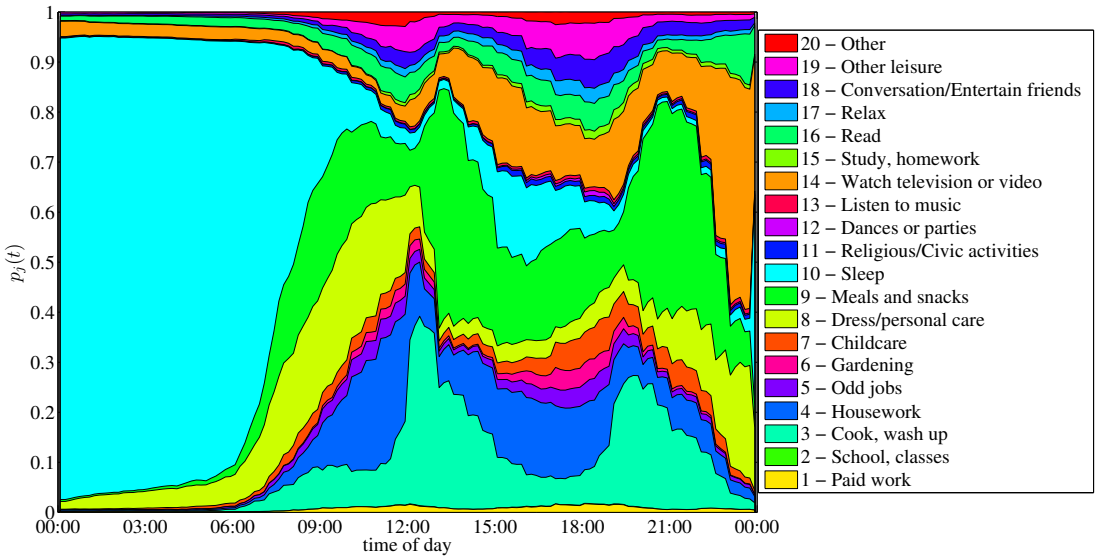


Figure 2.8: Profile of the merged activity types j (see legend) whilst individuals are at home. The merged activity category consists of all the activity types in Figure 2.7, which have the same index j as in the beginning of the line in the legend.

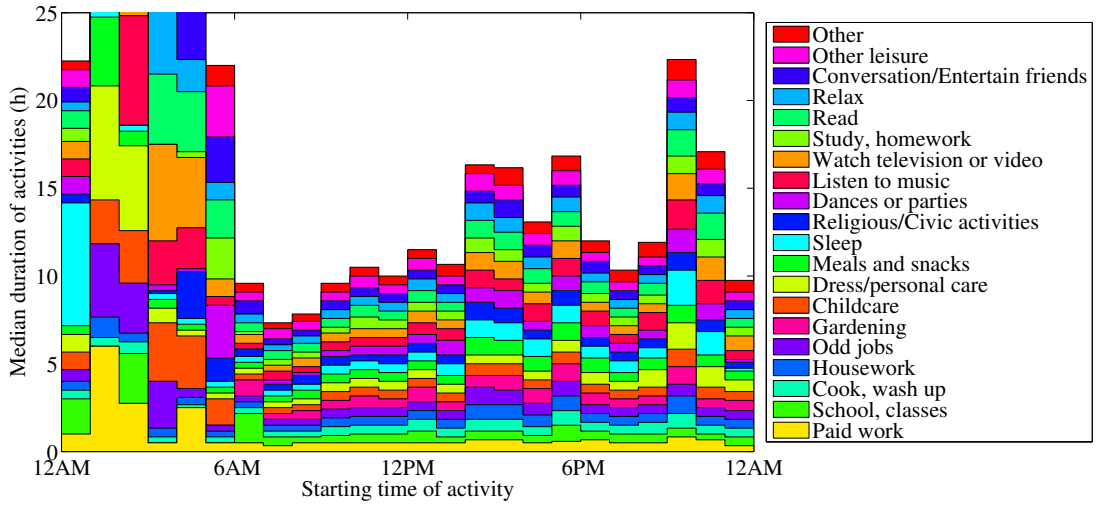


Figure 2.9: Stacked median durations of residential activities j_{MTUS} of the TUS started during the hour interval given on the x-axis. Activities j_{MTUS} which are performed for less than 0.5 % of the time throughout the day are not shown. For readability, the y-axis is bounded to a maximal value of 25 h.

number of started activities is considerably lower than during the rest of the day. This intuitive situation can also be explained by the fact that sleeping is the activity which is most often started in the two intervals before 1 am and which

has a high average duration at that time of day (*cf.* Figure 2.9). As a summary of the Figures 2.9 and 2.10, it can be seen that the probability to start an activity varies more with the time of day than does the average duration of the activity.

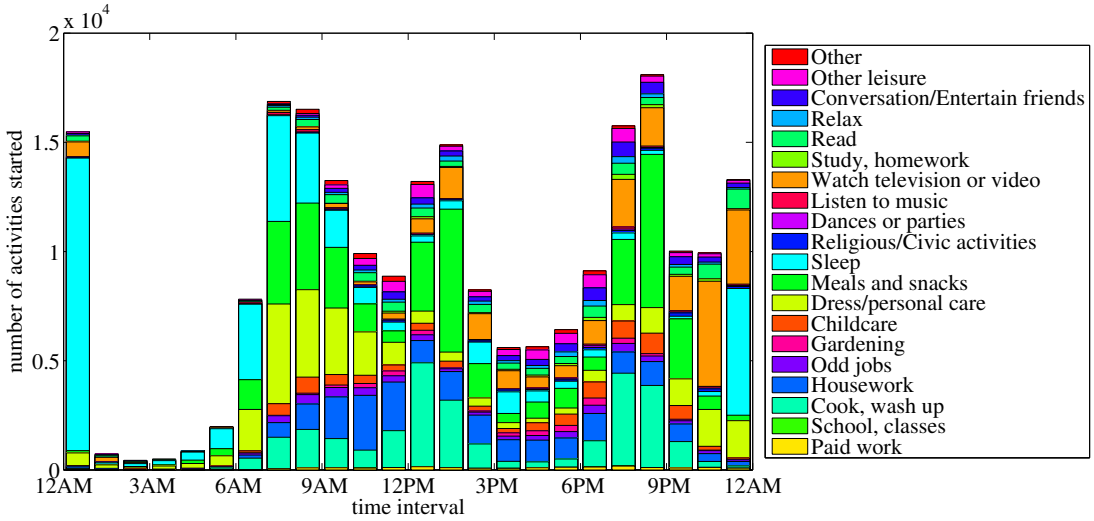


Figure 2.10: Stacked histogram of the total counts of activities started in the corresponding time interval shown on the x-axis.

As mentioned, the models to predict residential presence and activities are calibrated with these data, implying that the TUS data is assumed to reflect reality. Although, the data does obviously not always agree to reality (*e.g.*, the abrupt changes at midnight, incorrect copying of data in the database, the rounding artefacts or other incorrect statements of the questionnaires' respondents), the elucidation of the discrepancies between the monitored data and reality is difficult to provide and is a research topic in itself [*cf., e.g.* 38, 39]. According to the rounding at half and full hour intervals, the standard errors of the time values in the monitored data may be coarsely estimated to respectively 15 min and 30 min. However, the uncertainties might increase with the mean duration of the considered activity type, and also depend on the activity type (for instance, there might be a more important uncertainty in the time specification of sleeping). It is thus not straightforward to estimate uncertainties which appear in the starting/end times (cf. Figures 2.6 and 2.8), as well as in the durations of residential presence and activities (cf. Figure 4.2). As the focus of this work is elsewhere, the data artefacts which are due to erroneous specifications of the survey respondents will not be studied, unless explicitly stated.

2.2 Conclusion

The high time resolution of the TUS, as well as its considerable time-of-day variation provides a powerful source of information for dynamic modelling of human behaviour. Thus, these data will be used to calibrate the models predicting residential presence (Chapter 3) and activities (Chapter 4). The high level of detail of individuals' characteristics in the TUS data allows to investigate in depth their influence on behavioural patterns. The latter is also crucial to adjust resulting predictions for the non-representativeness of the distribution of some of the characteristics in the sample of the TUS.

Chapter 3

Bottom-up stochastic modelling of residential occupants' time-dependent presence

Two novel bottom-up modelling approaches together with a set of calibration methodologies are presented to predict the time-dependent probabilities of residential building occupants' presence by (i) a first- and (ii) a higher-order inhomogeneous Markov process. These models are calibrated depending on individual specificities (of the individual, the household, the household members, the weekday) using French time-use survey data (of 1998/1999), and are based on (i & ii) the transition probabilities and (ii) the presence duration probability distribution. A high convergence speed of both approaches is empirically demonstrated, and the higher-order approach is then validated by establishing the relationship to the one of first-order and comparing their predictions. Furthermore, their predictive power is compared. These models are used to derive the distribution of presence profiles over synthetic populations with different sets of characteristics. Finally, it is demonstrated how the models can be implemented in dynamic simulations to model presence as a dichotomous variable.

3.1 Introduction

As passive design standards of buildings allow to more effectively exploit solar radiation and conserve solar and casual heat gains, buildings' energy and environmental performance becomes more sensitive to the presence [40], activities and activity-related behaviours of their occupants [41], such as the use of shading devices, windows and electrical appliances [see 16, 42]. In particular individuals' behavioural diversity has a great impact on buildings' energy demands [43, 44].

The modelling of occupancy in buildings is also of great importance in electricity demand side management and electricity load modelling, as it determines

the possibility of occupants' interactions with buildings [45–54], as well as for energy demand in general [55]. As decentralised electricity generation from renewable sources becomes more and more important, we need to better match the load profiles of neighbourhoods with (small-scale) power generation. For a better design/sizing of the latter or for structural changes like the "smart grid" concept [*e.g.*, 56–58], we need to model more accurately the stochastic nature of electrical appliance use in single buildings [49, 59]. In particular, one has to account for the variations over time (of day) and of behaviour (use [60], investments, *etc.*) between individuals and households (*cf.* [61] for a comprehensive modelling study of this in office buildings).

With regard to the prediction of presence probabilities, a bottom-up model is needed that is calibrated on the significant characteristics of individuals, to faithfully encapsulate the full range of personal specificities. Such an approach also lends itself well to the modelling of future scenarios to explore responses to changes in occupancy behaviour as well as to the population's demographic characteristics. Last but not least, an important requirement for a sound model is that it should be easy to recalibrate it with other datasets, in order to investigate differences between countries or temporal changes.

3.1.1 Previous research work

In the research field of electricity demand modelling, Capasso et al. developed an approach where active residential occupancy (at home and not sleeping) is modelled stochastically in order to predict daily electricity load profiles [62]. Torriti presents an approach to predict occupancy variances in single-person households, in order to estimate electricity demand related to watching television in different European countries [46]. Widén et al. present a time-inhomogeneous Markov chain approach, calibrated with time use data, to predict residential presence probability profiles in order to model domestic lighting demand [63] and electricity demand in general [20].

Tanimoto et al. present a methodology calibrated with time use data that generates residential occupancy patterns for workdays, Saturdays and Sundays and eight classifications of individuals' attributes [25]. Richardson et al. model the individual's occupancy probability by the proportion of the sample population of a time use data set [64].

Regarding stochastic models of occupancy in offices, Wang et al. propose an approach to predict multiple-zone occupancy based on a first-order time-homogeneous Markov chain with an additional movement process which was calibrated to simulate occupancy in offices [65]. Wang et al. predict occupancy based on an approach, which was calibrated with movement sensor data collected during one year in 35 single person offices in an office building [66]. Dong and Lam used a sensor network [68] to calibrate a model that can predict occupancy in office buildings [67].

One of the most widely used models has been elaborated by Page et al., where occupancy in an office ("zone") is simulated as a time-inhomogeneous Markov chain. However, as the calibration dataset is only based on five offices in the same university building, the model cannot be considered representative for the variety of different behaviours. Furthermore, the calibration methodology is based on a mobility parameter which is ill-defined in case of a zero denominator [69]. Liao et al. model occupancy based on an approach extended to multiple zones, which shows similar results than those of Page et al. [71]. Furthermore, they also use a covariance graph model to reproduce presence proportions in the zones, based on the hypothesis that these proportions are correlated to each other, which implies that individuals do not behave independently of each other [70]. In both works, the models are calibrated with measurements based on movement sensors. Liao et al. furthermore use questionnaire survey data to calibrate the model for multiple person offices. When comparing the two models' predicted occupancy profiles with the observed ones, it is apparent that the applied calibration procedures are vulnerable to overfitting observation artefacts.

In transport research, occupancy and activities in different types of place are of interest in order to predict activity-travel patterns. In this context, much effort was put on the dependence of behaviour on socio-demographic characteristics [e.g., 72–74]. Although also incorporating the modelling of activity durations [cf. 72], the approaches were not validated with observed occupancy patterns.

3.1.2 Limitations of existing models

From the above review we conclude that:

- Existing occupancy models use simulations, in order to derive approximate expectation values of presence profiles. Uncertainties are reduced by increasing the number of simulation replicates, which can be very time-consuming.

- Published research does not provide a common robust and validated approach that models consistently variations of presence profiles and durations based on individual specificities (for instance personal/household characteristics).

- Existing occupancy models are often calibrated with non-representative datasets that are peculiar to specific situations.

- The issue of long absences is mostly neglected.

The objective in this chapter is to establish a bottom-up approach, which allows residential occupancy to be modelled as a function of time and individual

specificities. The time-dependence is crucial for the use in dynamic building simulations and bottom-up modelling of the residential use of electrical appliances, as the probabilities to be present varies significantly with time of day. The bottom-up nature of the approach is also needed to enable its application for scenario testing based on populations with different characteristics to those of the calibration dataset. Both the nature of the model formulation, which can be readily calibrated to other datasets, as well as the application results for populations with different socio-demographic characteristics suggest that this new model is better adapted to building simulation and electrical appliance use modelling than previous variants.

3.2 Methodology

The occupancy status y at a given time of day t represents a dichotomous random variable Y. It will be shown in Section 3.3 that Y can significantly depend on the time of day, as well as on other characteristics (of the individual, the household, the household members, the weekday) $\boldsymbol{x}$. In order to predict $p(\boldsymbol{x}, t)$ (which will be referred to as presence profile, and also denoted by $p(t)$, if the dependence on $\boldsymbol{x}$ is not discussed), the probability of residential presence as a function of time of day and $\boldsymbol{x}$, two approaches will be presented. In the following, we will show how these models are analytically derived and calibrated, in order to meet the requirements that were proposed in Section 3.1.2.

3.2.1 Stochastic models

A straightforward approach to predict $p(t)$, the probability of residential presence ($y = 1$) as a function of time t, is to assign a probability deduced from the observed presence proportion $p_{\text{obs}}(t)$ of the whole sample population $\mathcal{C}$ at each time step t_n [cf. 64].[1] However, this approach leads to an individual-independent model (IIM). Furthermore, this corresponds to a stochastic process of zeroth order, meaning that a state does not depend on the previous one(s). In a dynamic simulation this leads to a strongly fluctuating random variable of the presence state (which will be explained in more detail together with Figure 3.6), whose state has to be determined anew at each time step. As in many applications the occupancy duration is of substantial importance [*e.g.*, 75, 76], this approach will not be investigated in detail in the remainder of this chapter. Instead, two approaches will be presented in the following, where the durations of the occupancy states are modelled more consistently, using a first-order Markov process and a higher-order approach based on survival analysis. In general, the presence profile at the beginning of a given day depends on what has happened at the end of the day before. For reasons of simplicity, it will be assumed that the profiles that

[1] This corresponds to a time-inhomogeneous Bernoulli process.

will be derived in the following approaches have a 24 h periodicity (meaning that when, for instance a Saturday ends the behaviour after midnight corresponds to that of a Saturday and not a Sunday).[2]

First-order Markov process

Regarding time-dependent residential presence, there are 4 different transitions ($0 \mapsto 0,\ 0 \mapsto 1,\ 1 \mapsto 0,$ or $1 \mapsto 1$), denoting "to stay away", "to arrive", "to leave" and "to stay at home", respectively. The corresponding transition probabilities will be denoted by t_{00}, t_{01}, t_{10} and t_{11}, respectively, which satisfy

$$t_{00} + t_{01} = t_{10} + t_{11} = 1. \tag{3.1}$$

These transition probability elements will be regrouped in the transition matrix

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & t_{01} \\ t_{10} & t_{00} \end{pmatrix}. \tag{3.2}$$

Approximating the continuous-time occupancy as a discretised chain, the probability of being present corresponds to a sequence of random variables $(Y_1, Y_2, \ldots)$, and the probability of being at home can be modelled as a first-order (memoryless) Markov process (FOMP).

$$\mathrm{pr}(Y_{n+1} = y_{n+1} | Y_n = y_n, Y_{n-1} = y_{n-1}, \ldots) = \mathrm{pr}(Y_{n+1} = y_{n+1} | Y_n = y_n). \tag{3.3}$$

This implies that the time evolution of the probability of being present is governed by the master equation

$$\frac{d\boldsymbol{P}}{dt} = \boldsymbol{T}\boldsymbol{P}, \tag{3.4}$$

where $\boldsymbol{P} = (p, 1-p)^\intercal$ denotes the vector of the probabilities of being present or absent. In order to generate time-dependent profiles, the transition probability matrix $\boldsymbol{T}(t)$ has to depend on the time of day t. In the discretised approximation of the Markov chain, it follows from Equations (3.1) and (3.4) that the probability of being present at the succeeding time step t_{n+1} in the Markov chain can be calculated from the present one t_n:

$$p(t_{n+1}) = t_{11}(t_n)\, p(t_n) + t_{01}(t_n)\ (1 - p(t_n))\,. \tag{3.5}$$

This means that the probability of being present at t_{n+1} is given by the probability to be present at the time step before times the probability that he/she stayed

[2] A periodicity of one week would be more coherent, as presence is modelled as a function of weekday (cf. Figure 3.1). However, this increases calculation time by a factor of seven. Furthermore, the models are validated (cf. Sections 3.3.2 to 3.3.4) by comparing the predictions to the observations of the TUS, where individuals' diary plans are also recorded during 24 h (cf. Section 2.1.1).

$t_{11}(t_n)\ p(t_n)$ plus the probability of arriving when not having been present at the time step before $t_{01}(t_n)\,(1 - p(t_n))$. By incrementing the time step successively, one can calculate the time evolution of $p(t)$. However, as the initial value is unknown, the solution of Equation (3.4) cannot be immediately calculated. But the structure of the master equation secures that the influence of states to succeeding ones decreases with time to zero (under condition that all elements of $\boldsymbol{T}(t)$ are smaller than one at least for one t). Thus, one can chose an arbitrary initial value and repeat the calculation over several days (periods), so that the calculated $p(t)$ will asymptotically approach its exact solution. As the transition probabilities of a state change are smaller than 1 (in case of a non-deterministic model; cf. Figures 3.2 and 3.3 [69]), this implies that the sequence $p(t_n),\ (t_n \in \{t_1, \ldots, t_{\text{last}}\})$ is asymptotically approaching a continuous function when $(t_n - t_{n-1}) \to 0$ (t_{last} denotes the last time step of the day).

Higher-order Markov process

The modelling of $p(t)$ as a first-order Markov chain, avoids it to fluctuate excessively, like in the example mentioned in the beginning of Section 3.2.1. However, the memorylessness of the Markov process implies that the probability distribution function (PDF) of residential presence durations $f(t)$ is not being modelled consistently, as the transition probability t_{10} does not depend on the beginning time of the residential presence. In other words, the probability of a presence to end in the time interval $[t, t + dt]$ only depends on $p(t)$, but not on the exact history how this evolved before t. It has been shown that for time-homogeneous Markov processes this memorylessness of Equation (3.4) corresponds to exponential $f(t)$ [77, 78].

To model the distribution of presence durations $f(t)$ more realistically, an approach based on the generalized master equation [79–81] will be applied. Here, the basic characterising quantity is given by the survival function, which represents the probability that the presence did not end before time t:

$$S(t) = \text{pr}\,(T_{\text{end}} > t) = 1 - \int_0^t f(t')dt', \tag{3.6}$$

where T_{end} denotes a random variable indicating the time the presence duration ends. As the durations of residential presence depend significantly on the time of day t_{s} they are started (*cf.* Section 3.2.2 and Figure 3.6), the survival functions and the corresponding PDFs do so as well, which will be denoted by $S_{t_{\text{s}}}(t)$ and $f_{t_{\text{s}}}(t)$.

In order to predict the time evolution of $p(t)$, a model will be applied, which will be described in Section 4.2.1. A residential presence is started at t with a probability given by $t_{01}(t)$ (*cf.* Section 3.2.1), and afterwards the presence duration is determined according to $f_{t_{\text{s}}}(t)$, which will be explained in detail in Section 3.4.1.

It was shown that a continuous-time random walk (CTRW) corresponds to the generalized master equation, if the survival functions are not exponential [78, 82]. However, instead of simulating the residential presence by a CTRW to derive an approximative profile of $p(t)$, we will present a solution, which is similar to Equation (3.5). In the discretised approximation, the equation determining the probability of residential presence takes the form

$$p(t_n) = \sum_{i=1}^{\infty} (1 - p(t_{n-i}))\ t_{01}(t_{n-i})\ S_{t_{n-i}}(t_i). \tag{3.7}$$

In this formula, the probability of being present at t_n is given by the sum over the past of all terms that influence this probability. $(1-p(t_{n-i}))\ t_{01}(t_{n-i})$ corresponds to the probability that a presence was started at time t_{n-i} (more precisely in the time interval $[t_{n-1}, t_{n-i+1})$), and the multiplication by $S_{t_{n-i}}(t_i)$ gives the fraction of these started presences, which did not end before t_n. By summing up over the whole past, every contribution to the present $p(t_n)$ is captured in this equation. In this approach only residential presences have higher-order memory, whereas there is a memory effect of first-order for absences, likewise in Equation (3.5).[3] As in this model the impact of states on a later state is also decreasing with time passed, this equation can be approximated by replacing ∞ with $t_{\max}$. This corresponds to a memory effect of $t_{\max}$ time steps and to an approximation, where $S_t(t)$ is set to zero for $t > t_{\max}$. Thus, this represents a higher-order Markov process (HOMP), where the probability to be in a state depends on the $t_{\max}$ preceding states.[4]

The exact solution of Equation (3.7) can be approached recursively, by choosing $t_{\max}$ arbitrary initial values and continuing the stepwise approximation until the desired precision is reached. The results that are shown in this chapter in Section 3.3 are based on a value of $t_{\max} = 24\,h$, corresponding to $24 * 6 = 144$ previous time steps.

3.2.2 Model calibration

The model calibration is a very important and non-trivial part in order to establish a robust bottom-up model that, at the same time, reflects the variety among individual behaviours. For the calibration of the HOMP of Section 3.2.1, the two

[3] This means that the number of preceding states, which influence the probabilities of the following states is not the same for all states, likewise in a variable order Markov model [83]. However, in the approach presented here, the order is pre-defined (one for absences; $t_{\max}$ for presences).

[4] In Equation (6.6) of Section 6.3.1, it will be demonstrated that Equation (3.7) corresponds to the convolution of the survival function of durations with the corresponding starting probability. However, this simplified notation is only valid, in the special case of a system where the starting probabilities are independent of the state itself, and where the survival function is independent of the time of start, unlike Equation (3.7).

time-dependent quantities t_{10} as well as f_{t_s} are needed, whereas for the FOMP of Section 3.2.1, only $\boldsymbol{T}$ is needed. As, in addition to the time of day, these quantities depend on individual specificities $\boldsymbol{x}$, they will be denoted by $\boldsymbol{T}(\boldsymbol{x}, t)$ and $f_{t_s}(\boldsymbol{x}, t)$, when the individual-dependence is discussed. To assure the robustness of the models, the individual-specific calibration of these quantities has to be tested for statistical significance of the influence of the parameters.

Transition probabilities

The transition probabilities are calibrated on an hourly resolution [*cf.* 40] with all the occupancy states that overlap with the corresponding hour interval.[5] To determine the four elements of $\boldsymbol{T}$, there are only two unknowns (*cf.* Equation (3.1)). Thus, at a time t there are two choices an individual with characteristics $\boldsymbol{x} = (x_1, \ldots, x_M)$ in state y can make. One of them is to remain in the current state (present/absent), and the other one is to make a transition. This corresponds to a binary choice, and accordingly, 2x24 logistic regression models were estimated to calibrate $\boldsymbol{T}(\boldsymbol{x}, t)$.

$$t_{01}(\boldsymbol{x}, t) = \frac{1}{1 + \exp(-(\beta_0^{01}(\boldsymbol{x}, t) + \beta_1^{01}(\boldsymbol{x}, t)\, x_1 + \ldots + \beta_M^{01}(\boldsymbol{x}, t)\, x_M))}, \tag{3.8}$$

$$t_{10}(\boldsymbol{x}, t) = \frac{1}{1 + \exp(-(\beta_0^{10}(\boldsymbol{x}, t) + \beta_1^{10}(\boldsymbol{x}, t)\, x_1 + \ldots + \beta_M^{10}(\boldsymbol{x}, t)\, x_M))}. \tag{3.9}$$

In order that the transition probabilities encapsulate a large quantity of significant parameters, a backward elimination technique was applied, which will be explained in detail in Chapter 5. However, in the latter, there is a set of 16 predictors, whereas in the utility functions of Equations (3.8) and (3.9), there are $M = 64$ dummy variables (*cf.* the y-axis of Figure 3.1), which were tested for significant influence. Thus, it would not be possible to estimate an initial model for the whole set of parameters. Therefore, the initial model was restricted to dummy variables, for which there was a significant difference in choice behaviour, when they are considered in isolation from all other dummy variables. This was done by splitting the corresponding sample into two distinct parts (according to the value of the dummy variable) and comparing the resulting two proportions that a state transition is performed (using the two-proportion z test). Parameters were only included in the initial model, if the two proportions were significantly different at the 5 % level of confidence.

As mentioned, the transition probabilities were calibrated for a each individual $\boldsymbol{x}$ to reflect the transitions during a time step of one hour

$$\boldsymbol{P}(t + 1\,\mathrm{h}) = \boldsymbol{T}_\mathrm{h}(t)\boldsymbol{P}(t). \tag{3.10}$$

[5] A time step of one hour was chosen for the calibration, in order to eliminate the rounding artefacts in the temporal information of the TUS respondents (cf. Figures 2.6 and 4.2).

For a homogeneous discrete Markov chain with transition probability $\boldsymbol{T}$, it can be shown using the Chapman-Kolmogorov equation that the transition matrix of n time steps is given by $\boldsymbol{T}^n$ [84]. Therefore, the transition matrix for a time step of 10 min $\boldsymbol{T}_{10\,\mathrm{min}}(t)$ is given by

$$\boldsymbol{T}_{10\,\mathrm{min}}(t) = \boldsymbol{T}_{\mathrm{h}}(t)^{1/6}. \tag{3.11}$$

The transition probabilities $\boldsymbol{T}_{10\,\mathrm{min}}(t)$ were derived in this manner for the 2x24 $\boldsymbol{T}_{\mathrm{h}}(t)$, and afterwards, the values of $\boldsymbol{T}_{10\,\mathrm{min}}(t)$ were derived in the 10 min resolution by interpolating between these hourly values.

The parameter values of $\boldsymbol{T}_{\mathrm{h}}(t)$ are illustrated in Figure 3.1.[6] The values on the left-hand side that are shown for every t (in hourly resolution) correspond to the parameter values of the linear utility function of $t_{01}(t)$ (cf. Equation (3.8)), regarding the binary choice between coming home and staying away. On the right-hand side, they are shown for the utility function of $t_{10}(t)$ (cf. Equation (3.9)), regarding the binary choice between leaving and staying at home. The utility functions of t_{00} and t_{00} are fixed to zero. The parameter values are represented by circles, which are colour-coded according to the scale shown at the top of each graph. The values of the standard errors of these parameters are represented by the colour of the surrounding squares. Here, the value of the standard error was added to or subtracted from the corresponding parameter value for negative and positive values, respectively.

The time-dependent distributions of $t_{01}(\boldsymbol{x}, t)$ and $t_{10}(\boldsymbol{x}, t)$ over the sample population of the TUS are shown in Figures 3.2 and 3.3, respectively. The $24{\cdot}6 = 144$ values (10 min resolution) of the time-dependent density of these distributions were estimated, by the empirical probability distribution for every t using a bin width of $1/144$ times the length of the y-axis. These distributions are classified by the shown selection of their percentiles, as well as their mean values. Furthermore, the density of these distributions is illustrated by the greyscale in the background of the percentile curves.

Distributions of durations

The changing hazard rates $f(t)/S(t)$ of the duration PDFs were incorporated in the HOMP, by describing the PDFs as Weibull distributions. The individual-specific calibration of the duration PDFs[7] will be explained in detail in Section 4.2.1. However, the average duration of residential activities in Section 4.2.1 is considerably shorter than those of presences discussed in this chapter. Therefore, the effect of censored events at midnight (where the time period of the survey starts/ends) is more severe than in the former case. To capture more realistically

[6] The numerical parameter values are available online [85].

[7] The numerical parameter values of the Weibull functions of the individual-dependent duration PDF $f_{t_{\mathrm{s}}}(\boldsymbol{x}, t)$ are available online [85].

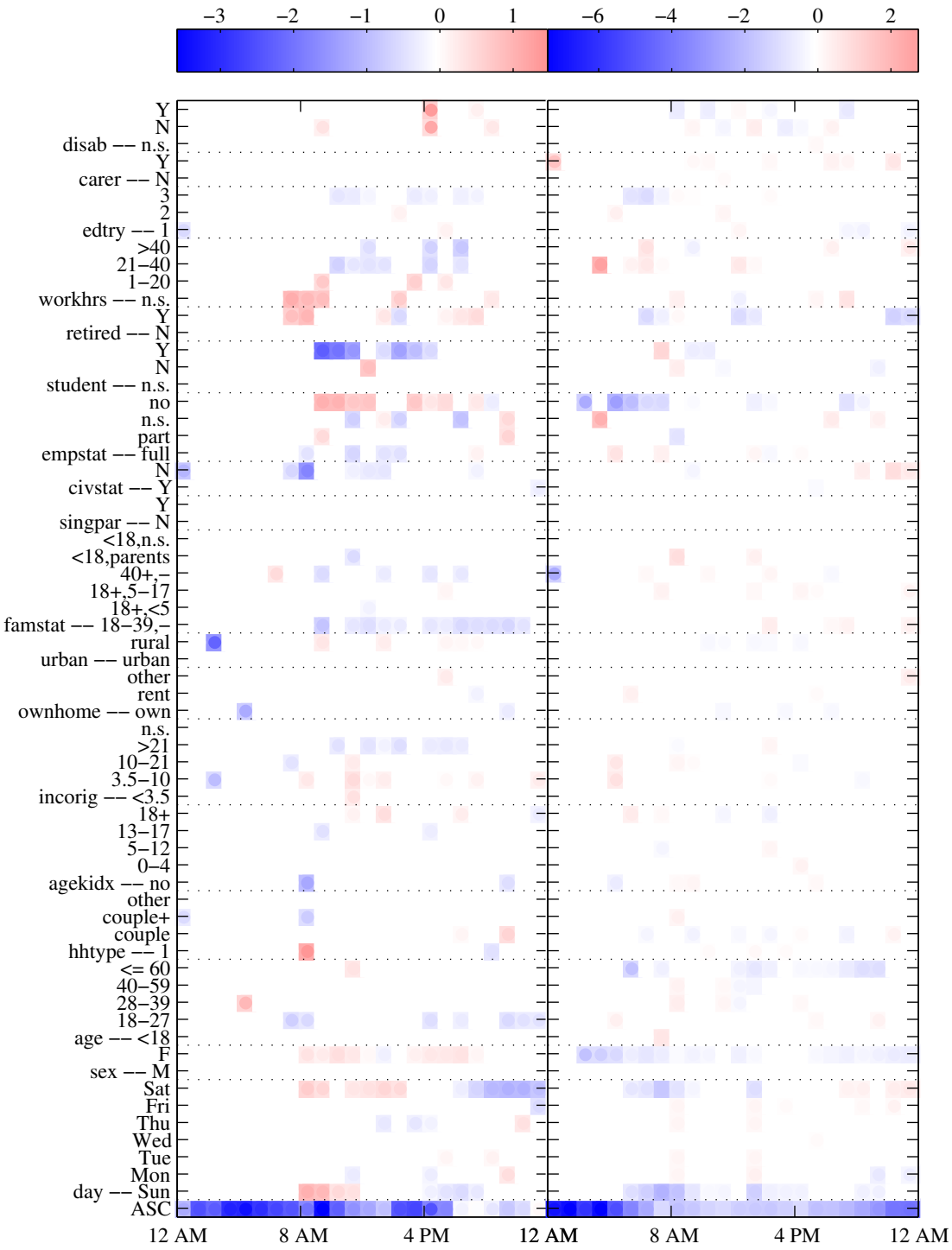


Figure 3.1: Parameter values (cf. Table 2.2) of the binary choice between coming and staying away (left), and staying at home and leaving (right), which are used to calibrate the individual-specific and time-dependent transition probabilities $\boldsymbol{T}(\boldsymbol{x}, t)$.

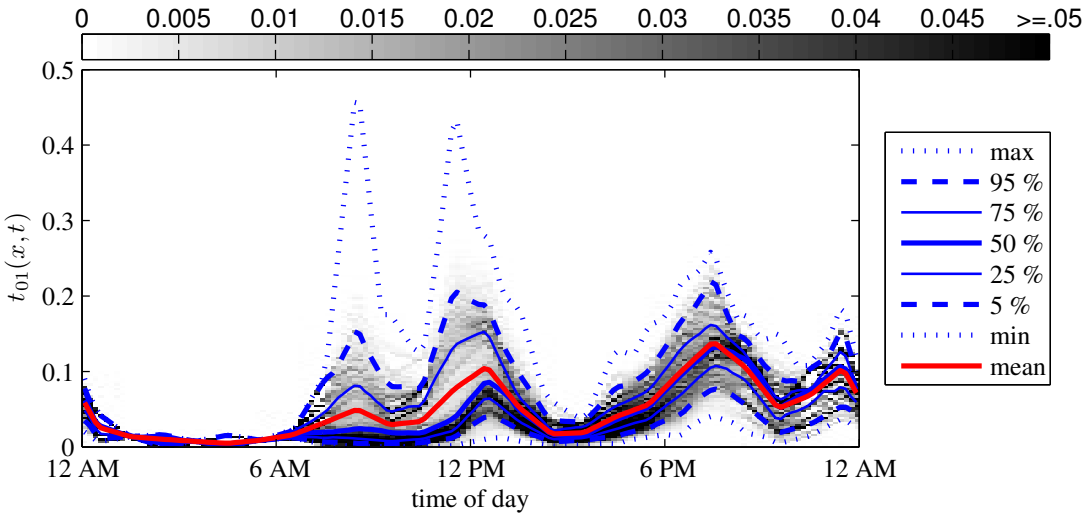


Figure 3.2: Distribution of $t_{01}(t)$ over the sample population.

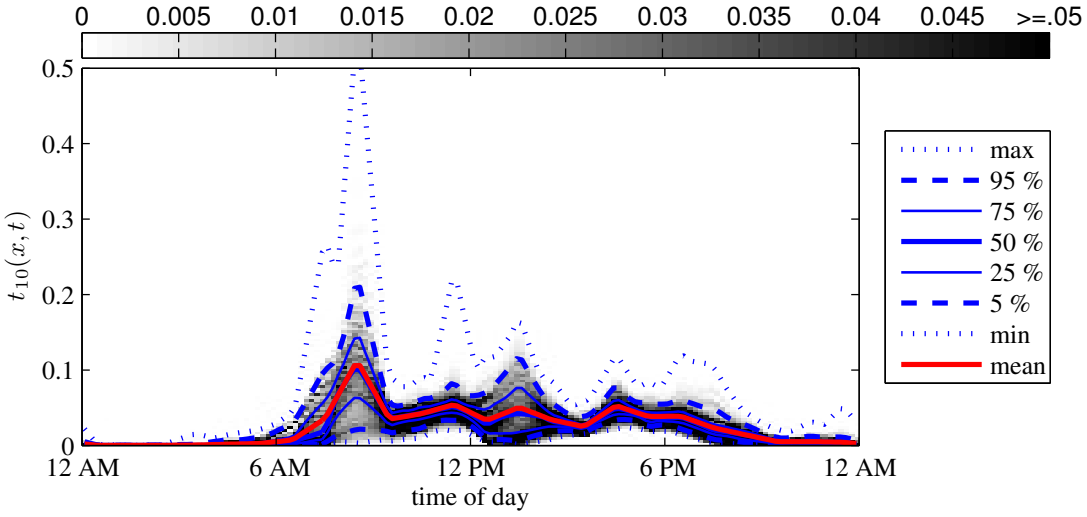


Figure 3.3: Distribution of $t_{10}(t)$ over the sample population.

the durations that outreach midnight, the right-censored events (presences that last until the end of the surveyed period) in all hour intervals were removed. In this way, the distribution of durations $f_{t_s,u}(t)$ which end before midnight was derived. For the censored events, it was assumed that the distribution of the remaining duration after midnight is described by that derived for the first hour interval of the day (between midnight and 1 am) f_1. Thus, the distributions (which are used in the model) are given by

$$f_{t_s}(\boldsymbol{x}, t) = (1 - \rho_{t_s}(\boldsymbol{x}))\ f_{t_s,u}(\boldsymbol{x}, t) + \rho_{t_s}(\boldsymbol{x})\ f_1(\boldsymbol{x}, t - (24\ \text{h} - t_s)), \tag{3.12}$$

where $\rho_{t_s}(\boldsymbol{x})$ denotes the proportion of durations started at t_s that are censored for the sample sub-population with characteristics $\boldsymbol{x}$, and $f_1(\boldsymbol{x}, t)$ is by definition zero for negative durations. The distribution of the values of $\rho_{t_s}(\boldsymbol{x}))$ is shown in Figure 3.4. The mean value of $\rho_{t_s}(\boldsymbol{x}))$ over the sample population is shown in for

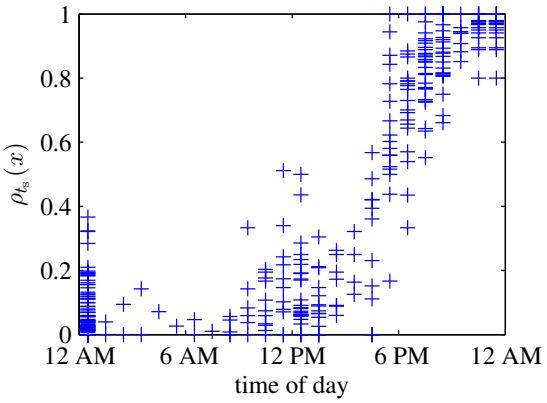


Figure 3.4: Distribution of the fraction of right-censored presence durations at home.

every hour interval in Figure 3.6 in the corresponding sub-graph. [8]

The values of the cumulative distribution function $F_{t_s}(\boldsymbol{x}, t) = 1 - S_{t_s}(\boldsymbol{x}, t)$ for one arbitrarily chosen individual $\boldsymbol{x}$ are shown in Figure 3.5 as a function of the presence duration t and the hour interval when the presence was started $[t_s - 30\ \text{min}, t_s + 30\ \text{min}]$. According to Equation (3.12), the value of $\rho_{t_s}(\boldsymbol{x}))$ determines the value of $F_{t_s}(\boldsymbol{x}, t)$ and the value of t_s determines the value of t, where $F_1(\boldsymbol{x}, t - (24\ \text{h} - t_s))$ starts to be non-zero. This can lead to pronounced bimodal distributions for some of the starting times t_s (for instance, at 4.30 PM or 8.30 PM).

[8] The distribution $f_1(\boldsymbol{x}, t - (24\ \text{h} - t_s))$ in Equation (3.12) corresponds to the distribution of the sum of the two random variables of censored events before midnight (given by the Dirac δ distribution $\delta(t - (24\ \text{h} - t_s))$ and the distribution after midnight (the distribution of the sum of two random variables is given by the convolution of the distributions of the single random variables [84]).

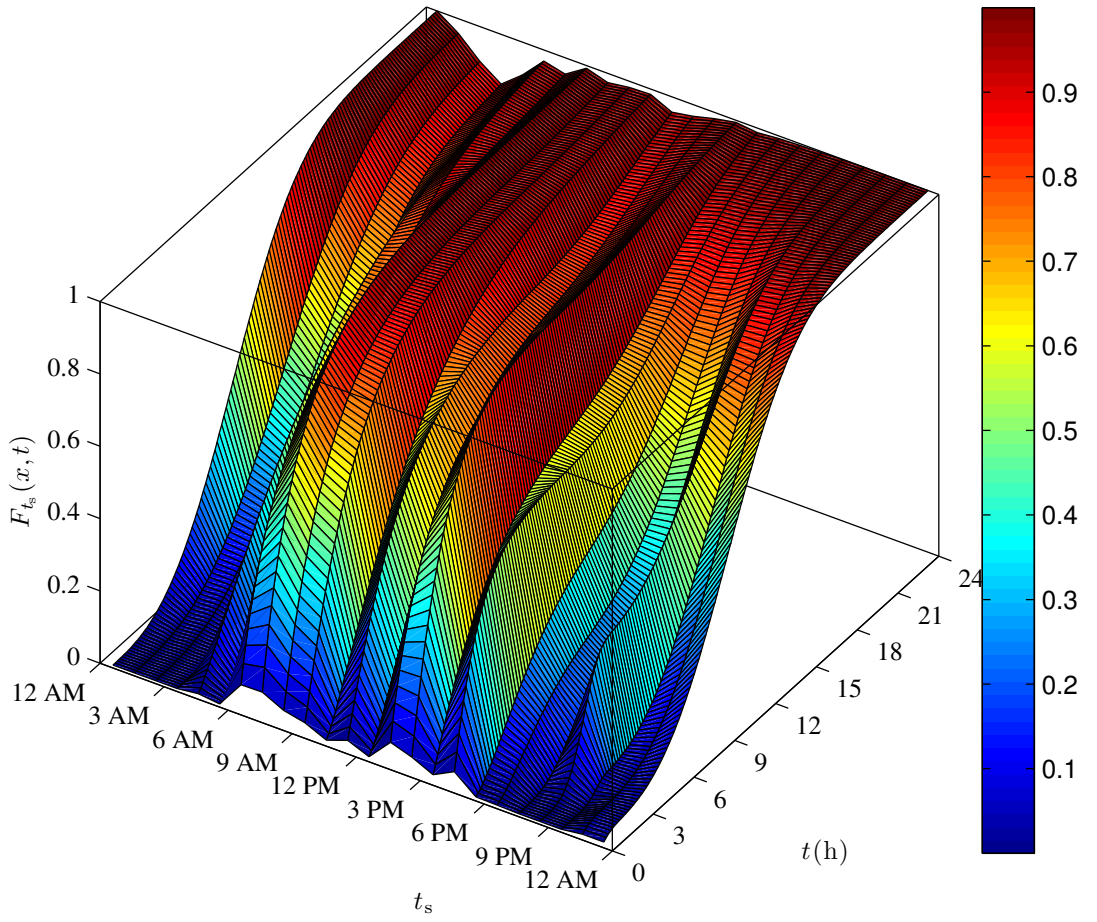


Figure 3.5: Illustration of the value of the Weibull CDF $F_{t_s}(\boldsymbol{x}, t)$ of one individual $\boldsymbol{x}$ as a function of presence duration t and hour interval when the presence was started $[t_s - 30 \text{ min}, t_s + 30 \text{ min}]$.

In order to apply Equation (3.7) on the basis of a 10 min time step, Equation (3.11) had also to be used to derive t_{01} in this resolution. However, in this case, the first column of $\boldsymbol{T}_{\mathrm{h}}(t)$ is given by $t_{10}(t) = S_t(1\ \mathrm{h})$ and Equation (3.1). This can lead to complex-valued presence probabilities, in which case only the real part will be shown.

In order to link the zeroth- and first-order modelling approaches to survival analysis, their PDFs will be derived. That of the first-order Markov process $f_{\mathrm{M},t_{\mathrm{s}}}$ of Section 3.2.1 can be recursively determined in a discretised approximation:

$$f_{\mathrm{M},t_{\mathrm{s}}}(t) = \left(1 - \sum_{1}^{t} f_{\mathrm{M},t_{\mathrm{s}}}(t-1)\right) t_{10}(t). \tag{3.13}$$

As it would not be possible to apply Equation (3.4), if t_{10} was dependent on the presence begin, the latter was not considered in the predictor set for the calibration of t_{10}. The survival function of the IIM is given by

$$S_{0,t_{\mathrm{s}}}(t) = \prod_{t'=t_{\mathrm{s}}+1}^{t} p_{\mathrm{obs}}(t'), \tag{3.14}$$

which also defines the corresponding PDF $f_{0,t_{\mathrm{s}}}(t)$.

In Figure 3.6, a comparison of the empirical PDFs (EPDFs) of observed non-censored presence durations with those of the described models is shown in dependence of t_{s}, the time of day when the presences were started. Here, these distributions were all derived in a 10 min resolution for the whole sample population, i.e. also the FOMP and the HOMP independently of the individual characteristics $\boldsymbol{x}$. The graphs are bounded to the value space of the EPDFs. The strong fluctuations of the IIM appear as a strong weight on short presence durations in $f_{0,t_{\mathrm{s}}}(t)$ for all t_{s}. Hence, there is no agreement with the EPDFs. As the EPDFs are censored for durations that exceed midnight, the values of $f_{\mathrm{M},t_{\mathrm{s}}}$ in the figure were divided by the mean non-censored fraction $(1 - \overline{\rho_{t_{\mathrm{s}}}(\boldsymbol{x})})$, in order to be in accordance to Equation (3.12) and make them comparable to the censored EPDFs. These distributions can capture the bimodal character of the EPDFs in some intervals (for instance, between 12 PM and 1 PM). However, as the FOMP transition probabilities were not calibrated depending on the presence start, there are sometimes large deviations from the EPDFs (for instance, between 1 AM and 4 AM or between 5 AM and 6 AM). The PDFs used in the HOMP were defined as Weibull distributions, which were fitted to the EPDFS. However, the Weibull distributions cannot capture the multimodal character of the EPDFs.

Table 3.1 shows a statistical evaluation of the goodness of fit of the models in Figure 3.6 as a function of the number of the time interval t_{s}. N_{obs} indicates the number of uncensored observations in the corresponding EPDF. k_{df} denotes the number of degrees of freedom of the corresponding model. For the IIM, this corresponds to the number of time intervals of the EPDF (*cf.* Equation (3.14)),

whereas for the FOMP, k_{df} is given by twice the number of hour intervals before midnight (*cf.* Equation (3.13)). As the durations PDFs in the HOMP are represented by Weibull distributions, there are always two degrees of freedom. The values of the Akaike Information Criterion (AIC) [86] and of the more conservative Bayesian Information Criterion (BIC) [87] depend on the log-likelihood $\mathcal{L}$:

$$AIC = -2\mathcal{L} + 2k_{\mathrm{df}}, \qquad BIC = -2\mathcal{L} + k_{\mathrm{df}} \ln N_{\mathrm{obs}}. \tag{3.15}$$

The value of $\mathcal{L}$ of the IIM is inferior to that of the other two models for all time intervals, except for the 23rd time interval, where the that of the FOMP is 2.6 % larger in magnitude. In average the magnitude of $\mathcal{L}$ is approximately twice as high for the IIM than for the two Markov processes. The mentioned exception is not of great significance, as the EPDF in the corresponding interval is based on only 33 observations. Furthermore, a comparison of AIC and BIC between the IIM and the FOMP shows that the latter is preferable, because of the smaller k_{df}. The goodness of fit of the PDFs of the FOMP are superior to those of the HOMP in the first, the twelfth, as well as the 18th to the 21st interval according to the BIC. This also holds for the AIC, in addition to the eleventh and the 22nd interval. During the night, the most significantly inferior goodness of fit of the FOMP may be explained by the small relative appearance of events compared to the first interval after midnight, where N_{obs} is more than two orders of magnitude larger. Therefore, these PDFs have a peak which is dominated by that of $f_{\mathrm{M},1}(t)$ (shifted by the difference of the corresponding starting interval).

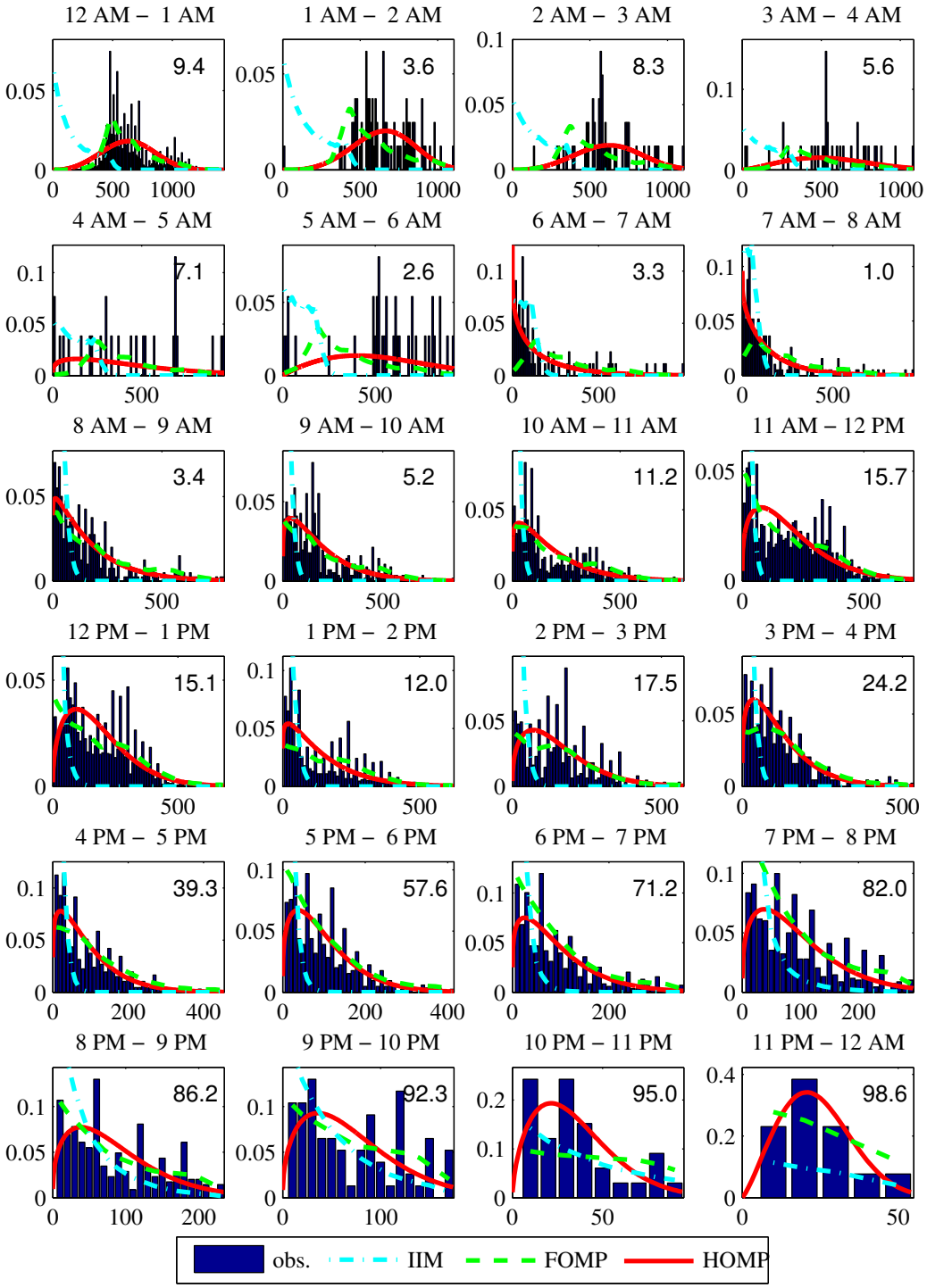


Figure 3.6: Empirical (obs.), zeroth-order (IIM), first-order Markov (FOMP) and fitted Weibull (HOMP) probability distributions of residential presence durations (in min) that were started in the indicated time interval. The number top right of each plot shows the mean right-censored percentage $\overline{\rho_{t_s}(\boldsymbol{x})}$.

		IIM				FOMP				HOMP			
t_s	N_{obs}	k_{df}	$\mathcal{L}$	AIC	BIC	k_{df}	$\mathcal{L}$	AIC	BIC	k_{df}	$\mathcal{L}$	AIC	BIC
1	13187	143	$-1.48 \cdot 10^{05}$	$2.97 \cdot 10^{05}$	$2.98 \cdot 10^{05}$	48	$-5.55 \cdot 10^{04}$	$1.11 \cdot 10^{05}$	$1.12 \cdot 10^{05}$	2	$-5.82 \cdot 10^{04}$	$1.16 \cdot 10^{05}$	$1.16 \cdot 10^{05}$
2	81	110	$-1.31 \cdot 10^{03}$	$2.84 \cdot 10^{03}$	$3.10 \cdot 10^{03}$	46	$-3.65 \cdot 10^{02}$	$8.22 \cdot 10^{02}$	$9.32 \cdot 10^{02}$	2	$-3.55 \cdot 10^{02}$	$7.14 \cdot 10^{02}$	$7.19 \cdot 10^{02}$
3	55	110	$-7.71 \cdot 10^{02}$	$1.76 \cdot 10^{03}$	$1.98 \cdot 10^{03}$	44	$-2.51 \cdot 10^{02}$	$5.90 \cdot 10^{02}$	$6.78 \cdot 10^{02}$	2	$-2.42 \cdot 10^{02}$	$4.88 \cdot 10^{02}$	$4.92 \cdot 10^{02}$
4	34	108	$-5.28 \cdot 10^{02}$	$1.27 \cdot 10^{03}$	$1.44 \cdot 10^{03}$	42	$-1.67 \cdot 10^{02}$	$4.18 \cdot 10^{02}$	$4.82 \cdot 10^{02}$	2	$-1.60 \cdot 10^{02}$	$3.23 \cdot 10^{02}$	$3.26 \cdot 10^{02}$
5	26	96	$-2.85 \cdot 10^{02}$	$7.62 \cdot 10^{02}$	$8.83 \cdot 10^{02}$	40	$-1.31 \cdot 10^{02}$	$3.42 \cdot 10^{02}$	$3.93 \cdot 10^{02}$	2	$-1.23 \cdot 10^{02}$	$2.50 \cdot 10^{02}$	$2.52 \cdot 10^{02}$
6	37	92	$-5.61 \cdot 10^{02}$	$1.31 \cdot 10^{03}$	$1.45 \cdot 10^{03}$	38	$-1.99 \cdot 10^{02}$	$4.73 \cdot 10^{02}$	$5.35 \cdot 10^{02}$	2	$-1.79 \cdot 10^{02}$	$3.61 \cdot 10^{02}$	$3.64 \cdot 10^{02}$
7	88	99	$-5.21 \cdot 10^{02}$	$1.24 \cdot 10^{03}$	$1.48 \cdot 10^{03}$	36	$-3.95 \cdot 10^{02}$	$8.61 \cdot 10^{02}$	$9.51 \cdot 10^{02}$	2	$-3.44 \cdot 10^{02}$	$6.92 \cdot 10^{02}$	$6.97 \cdot 10^{02}$
8	193	94	$-1.13 \cdot 10^{03}$	$2.45 \cdot 10^{03}$	$2.76 \cdot 10^{03}$	34	$-7.63 \cdot 10^{02}$	$1.59 \cdot 10^{03}$	$1.70 \cdot 10^{03}$	2	$-7.22 \cdot 10^{02}$	$1.45 \cdot 10^{03}$	$1.46 \cdot 10^{03}$
9	399	78	$-3.37 \cdot 10^{03}$	$6.90 \cdot 10^{03}$	$7.21 \cdot 10^{03}$	32	$-1.55 \cdot 10^{03}$	$3.16 \cdot 10^{03}$	$3.28 \cdot 10^{03}$	2	$-1.53 \cdot 10^{03}$	$3.06 \cdot 10^{03}$	$3.07 \cdot 10^{03}$
10	563	85	$-5.73 \cdot 10^{03}$	$1.16 \cdot 10^{04}$	$1.20 \cdot 10^{04}$	30	$-2.20 \cdot 10^{03}$	$4.46 \cdot 10^{03}$	$4.59 \cdot 10^{03}$	2	$-2.18 \cdot 10^{03}$	$4.37 \cdot 10^{03}$	$4.38 \cdot 10^{03}$
11	722	79	$-8.02 \cdot 10^{03}$	$1.62 \cdot 10^{04}$	$1.66 \cdot 10^{04}$	28	$-2.78 \cdot 10^{03}$	$5.61 \cdot 10^{03}$	$5.74 \cdot 10^{03}$	2	$-2.81 \cdot 10^{03}$	$5.63 \cdot 10^{03}$	$5.64 \cdot 10^{03}$
12	1243	70	$-1.58 \cdot 10^{04}$	$3.18 \cdot 10^{04}$	$3.21 \cdot 10^{04}$	26	$-4.82 \cdot 10^{03}$	$9.69 \cdot 10^{03}$	$9.82 \cdot 10^{03}$	2	$-4.91 \cdot 10^{03}$	$9.83 \cdot 10^{03}$	$9.84 \cdot 10^{03}$
13	2710	68	$-3.04 \cdot 10^{04}$	$6.10 \cdot 10^{04}$	$6.14 \cdot 10^{04}$	24	$-1.04 \cdot 10^{04}$	$2.09 \cdot 10^{04}$	$2.10 \cdot 10^{04}$	2	$-1.04 \cdot 10^{04}$	$2.07 \cdot 10^{04}$	$2.07 \cdot 10^{04}$
14	1889	62	$-1.68 \cdot 10^{04}$	$3.37 \cdot 10^{04}$	$3.41 \cdot 10^{04}$	22	$-7.08 \cdot 10^{03}$	$1.42 \cdot 10^{04}$	$1.43 \cdot 10^{04}$	2	$-6.87 \cdot 10^{03}$	$1.37 \cdot 10^{04}$	$1.38 \cdot 10^{04}$
15	611	57	$-6.50 \cdot 10^{03}$	$1.31 \cdot 10^{04}$	$1.34 \cdot 10^{04}$	20	$-2.29 \cdot 10^{03}$	$4.63 \cdot 10^{03}$	$4.71 \cdot 10^{03}$	2	$-2.25 \cdot 10^{03}$	$4.50 \cdot 10^{03}$	$4.51 \cdot 10^{03}$
16	441	53	$-3.56 \cdot 10^{03}$	$7.22 \cdot 10^{03}$	$7.44 \cdot 10^{03}$	18	$-1.55 \cdot 10^{03}$	$3.14 \cdot 10^{03}$	$3.22 \cdot 10^{03}$	2	$-1.50 \cdot 10^{03}$	$3.00 \cdot 10^{03}$	$3.01 \cdot 10^{03}$
17	668	45	$-4.42 \cdot 10^{03}$	$8.93 \cdot 10^{03}$	$9.13 \cdot 10^{03}$	16	$-2.14 \cdot 10^{03}$	$4.31 \cdot 10^{03}$	$4.38 \cdot 10^{03}$	2	$-2.13 \cdot 10^{03}$	$4.27 \cdot 10^{03}$	$4.28 \cdot 10^{03}$
18	846	41	$-5.54 \cdot 10^{03}$	$1.12 \cdot 10^{04}$	$1.14 \cdot 10^{04}$	14	$-2.60 \cdot 10^{03}$	$5.23 \cdot 10^{03}$	$5.29 \cdot 10^{03}$	2	$-2.76 \cdot 10^{03}$	$5.51 \cdot 10^{03}$	$5.52 \cdot 10^{03}$
19	837	35	$-4.05 \cdot 10^{03}$	$8.17 \cdot 10^{03}$	$8.34 \cdot 10^{03}$	12	$-2.44 \cdot 10^{03}$	$4.90 \cdot 10^{03}$	$4.95 \cdot 10^{03}$	2	$-2.67 \cdot 10^{03}$	$5.35 \cdot 10^{03}$	$5.36 \cdot 10^{03}$
20	681	29	$-2.67 \cdot 10^{03}$	$5.41 \cdot 10^{03}$	$5.54 \cdot 10^{03}$	10	$-1.96 \cdot 10^{03}$	$3.94 \cdot 10^{03}$	$3.99 \cdot 10^{03}$	2	$-2.18 \cdot 10^{03}$	$4.37 \cdot 10^{03}$	$4.38 \cdot 10^{03}$
21	346	23	$-1.13 \cdot 10^{03}$	$2.31 \cdot 10^{03}$	$2.40 \cdot 10^{03}$	8	$-1.04 \cdot 10^{03}$	$2.09 \cdot 10^{03}$	$2.12 \cdot 10^{03}$	2	$-1.07 \cdot 10^{03}$	$2.14 \cdot 10^{03}$	$2.15 \cdot 10^{03}$
22	77	17	$-2.30 \cdot 10^{02}$	$4.95 \cdot 10^{02}$	$5.35 \cdot 10^{02}$	6	$-2.14 \cdot 10^{02}$	$4.40 \cdot 10^{02}$	$4.54 \cdot 10^{02}$	2	$-2.22 \cdot 10^{02}$	$4.49 \cdot 10^{02}$	$4.53 \cdot 10^{02}$
23	33	9	$-7.94 \cdot 10^{01}$	$1.77 \cdot 10^{02}$	$1.90 \cdot 10^{02}$	4	$-8.15 \cdot 10^{01}$	$1.71 \cdot 10^{02}$	$1.77 \cdot 10^{02}$	2	$-6.96 \cdot 10^{01}$	$1.43 \cdot 10^{02}$	$1.46 \cdot 10^{02}$
24	13	5	$-3.17 \cdot 10^{01}$	$7.34 \cdot 10^{01}$	$7.62 \cdot 10^{01}$	2	$-1.94 \cdot 10^{01}$	$4.29 \cdot 10^{01}$	$4.40 \cdot 10^{01}$	2	$-1.94 \cdot 10^{01}$	$4.28 \cdot 10^{01}$	$4.39 \cdot 10^{01}$

Table 3.1: Goodness of fit indicators of the presence duration PDFs of the models in Figure 3.6 with the observed PDF.

3.3 Results

In this section, we show the results obtained with the models that were described in the previous section. As it was mentioned in Section 3.2, the two described models yield solutions of the presence profiles for each individual (described by its predictor value set) $\boldsymbol{x}$ that can be asymptotically approached. The convergence during the period of three days of the presence profiles $p(\boldsymbol{x}, t)$ predicted by the two models of Section 3.2 of three individuals is shown in Figures 3.7 and 3.8. The value sets $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ of the three profiles were arbitrarily chosen from three individuals of the sample population. As in the recursive derivation of the presence profiles the initial value(s) (*cf.* Equations (3.5) and (3.7)) is/are not known, and therefore arbitrary value(s) had to be chosen, this can lead to nonphysical presence profiles which are not bounded between zero and one (*cf.* Figure 3.8). However, these errors vanish during the convergence. The discrepancies of the profiles between the second and the third day have already diminished substantially.

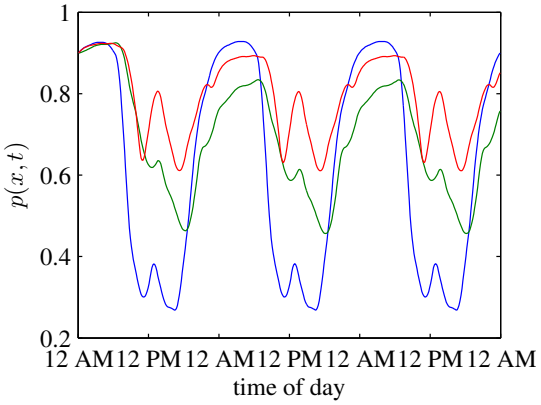


Figure 3.7: Convergence of the presence profiles of 3 arbitrarily chosen individuals derived from the FOMP.

A statistical evaluation of the distribution of the maximum discrepancies Δ_p between the profiles of subsequent daily periods T and $T-1$ is illustrated in Figures 3.9 and 3.10 as a function of T. The value of Δ_p is approximately exponentially decaying for both models. At the fourth period T, the maximum value of Δ_p is $6.0 \cdot 10^{-3}$ for the HOMP and $3.2 \cdot 10^{-5}$ for the FOMP. Thus, the latter represents a more efficient methodology to estimate the expectation value of the presence profile of an inhomogeneous Markov chain than simulation [cf., e.g., 69]. Nevertheless, when estimating the expectation value as the mean of a sample of n_r Monte Carlo simulations (where the convergence speed is proportional to $1/\sqrt{n_\mathrm{r}}$), the convergence speed is inferior to that of both of the models. The errors of the predictions of both of the Markov processes are dominated by the uncertainties in the TUS data that were discussed in Section 2.1.1.

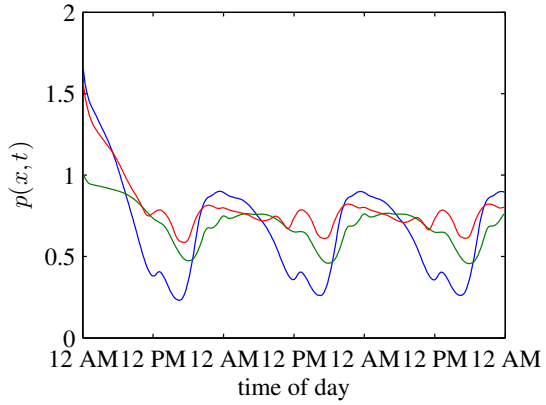


Figure 3.8: Convergence of the presence profiles of 3 arbitrarily chosen individuals derived from the HOMP.

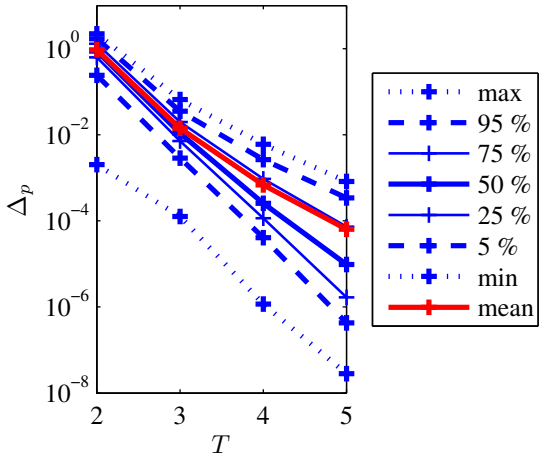


Figure 3.9: Distribution of the maximal difference of the presence profiles of the HOMP between subsequent daily periods as a function of the number of the daily period. The lines are drawn to guide the eye.

To illustrate the resulting predicted presence profiles $p(\boldsymbol{x}, t)$ of the two models, we show the converged curves of 30 arbitrarily chosen individuals $\boldsymbol{x}$ in Figures 3.11 and 3.12.[9] The profiles of the two models vary substantially depending on the time of day as well as on the individual characteristics $\boldsymbol{x}$. A quantitative evaluation of the corresponding time-dependent distributions will be provided in Section 3.3.2, and the influences of the values of the individuals $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ of the population $\mathcal{C}$ on the profiles will be considered in Section 3.3.4.

[9]The resulting presence profiles as a function of $\boldsymbol{x}$ are available online for all individuals of the sample population $\mathcal{C}$ [85].

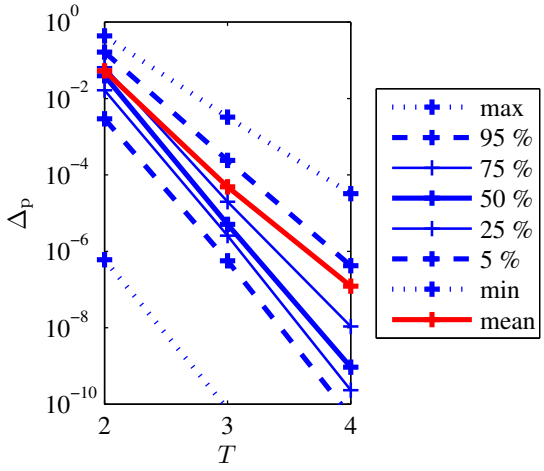


Figure 3.10: Distribution of the maximal difference of the presence profiles of the FOMP between subsequent daily periods as a function of the number of the daily period. The lines are drawn to guide the eye.

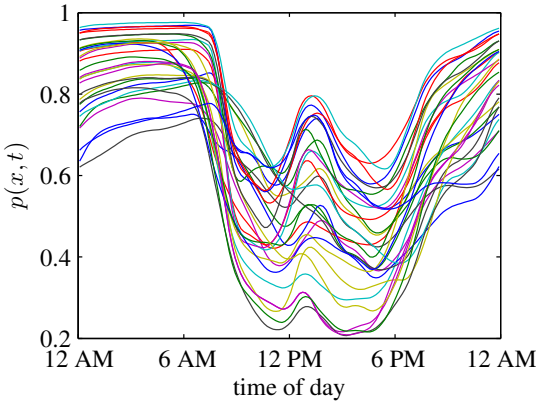


Figure 3.11: Sample of the presence profiles of 30 arbitrarily chosen individuals predicted by the FOMP.

3.3.1 Validation of the survival model

In order to validate the HOMP, it was applied using $f_{\mathrm{M},t_\mathrm{s}}$, the presence duration PDFs of the first-order Markov process of Section 3.2.1 (*cf.* Equation (3.13)). As it was mentioned in Section 3.2.1, the maximal memory was chosen to be $t_\mathrm{max} = 24\ \mathrm{h} = 144 \cdot 10$ min. The 24 h periodicity of $\boldsymbol{T}$ leads to the peculiarity that $S_{t_\mathrm{s}}(24\ \mathrm{h})$ gives a constant value, independent of t_s (as $S_{t_\mathrm{s}}(24\ \mathrm{h}) = \prod_t t_{10}(\boldsymbol{x}, t)$ for all t_s). A comparison of the resulting presence profile of an arbitrarily chosen individual of the sample population with the profile resulting from the first-order Markov property (3.5) and the master equation (3.4) is shown in Figure 3.13. The differences between both curves might arise from the approximation $S_{t_\mathrm{s}}(t_\mathrm{max} +$

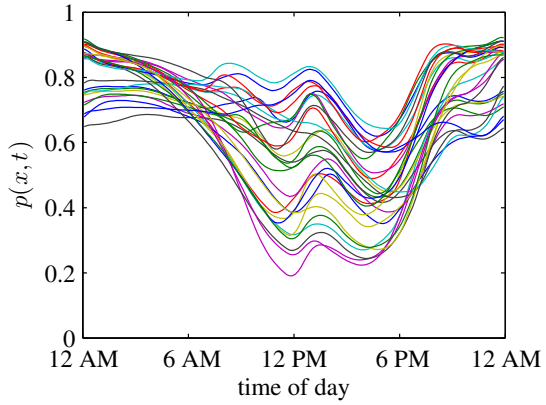


Figure 3.12: Sample of the presence profiles of 30 arbitrarily chosen individuals predicted by the HOMP.

1) = 0, as well as to the interpolation of the transition probability elements that was described in the context of Equation (3.11).

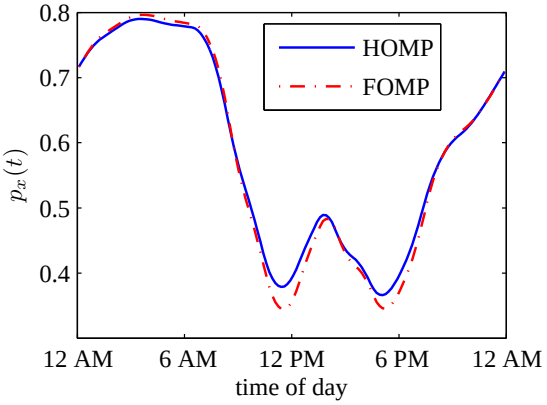


Figure 3.13: Comparison of the resulting presence profiles of the HOMP using Equation (3.13), and the FOMP.

3.3.2 Presence profile distributions

Apart from the influence of the predictor variables $\boldsymbol{x}$ on single individuals' presence profiles predicted by the models of Section 3.2, the distribution of those profiles over a synthetically generated population is of interest, when such a model is applied in a scenario. In order to compare the properties of the synthetic population to the observations (in terms of the distribution of characteristics and correlations among the latter), it was chosen to be equal to the sample population of the TUS, that was used to calibrate the models. When the models are used in applications in order to make aggregate forecasts, the individuals of the synthetic

population have to be weighted in order to bring it in line with the population from which it was drawn or to the scenario population [cf., e.g., 88].

By calculating the presence profiles that are shown in Figures 3.11 and 3.12 for the entire synthetic population, the time-dependent probability distributions $f_p(t)$ can be estimated, which are shown in Figures 3.14 and 3.15. These distributions were derived in the same manner as it was described in conjunction with Figures 3.2 and 3.3.

In Figure 3.14, the presence profile distribution $f_p(t)$ over the population, predicted by the FOMP, has a high level of uncertainty over the whole day (by, for instance, defining the confidence intervals as the 5th and the 95th percentiles). The uncertainty is larger during the day, where the distribution furthermore has a bimodal character. The reason for this will be discussed in Section 3.3.4. The average of the distribution shows good agreement with the observed presence proportion. The maximum overestimation is 11.5 % at 6 PM, and the maximum underestimation of 17.1 % takes place at 9 AM. In general, the predictions during the night are too small, probably due to the censoring of the arrivals at midnight (see difference between the observed presence proportions at the beginning/end of the day), which had to be omitted when estimating the logistic regression models for t_{01}.

In Figure 3.14, we show the distribution over the population $f_p(t)$, that is predicted by the HOMP. During the day, the bimodal character and the level of uncertainty are similar to the distribution predicted by the FOMP. During the night, the level of uncertainty is smaller. This might be due to the fact that then, the variety of the behaviour between individuals is overestimated by the FOMP, as then there are too many fluctuations (cf. Figure 3.6). Another reason might be that the calibration procedure of the individualised $f_{t_s}(\boldsymbol{x}, t)$ for the HOMP does not capture as much the variety of individual behaviour. The average of the distribution shows not as good agreement with the observed presence proportion. The maximum overestimation is 18.3 % at 12 PM, and the maximum underestimation of 21.6 % takes place at 7 AM.

In particular, during the night there are large discrepancies, which can be explained by the representation of $f_{t_s}(\boldsymbol{x}, t)$ by Weibull distributions, where durations that end in the small hours are overestimated in the first hour interval of the day (cf. Figure 3.6). To illustrate this, the average survival function after midnight $f_1(\overline{\boldsymbol{x}}, t)$ is shown in the figure (denoted by "S(t)") for the small hours multiplied by the average value of $f_p(t)$ at midnight as the distribution after midnight has an important weight on the shape of the profiles (cf. Equation (3.12)). The overestimated proportion of shorter presence durations is translated into a too strong decrease of this survival curve.

From 7.30 AM on, there is an increase of the maximum of $f_p(t)$ and in the slope of its average. This is due to the increase of t_{01} for a significant part of the population, which starts to take place then (cf. Figure 3.2), and to the description of absences by a HOMP. In other words, remaining absence durations depend in

reality on their starting times, which is not considered in the FOMP. In this model, likewise the FOMP, the predictions during the night are also too small, due to the mentioned censoring of the arrivals at midnight.

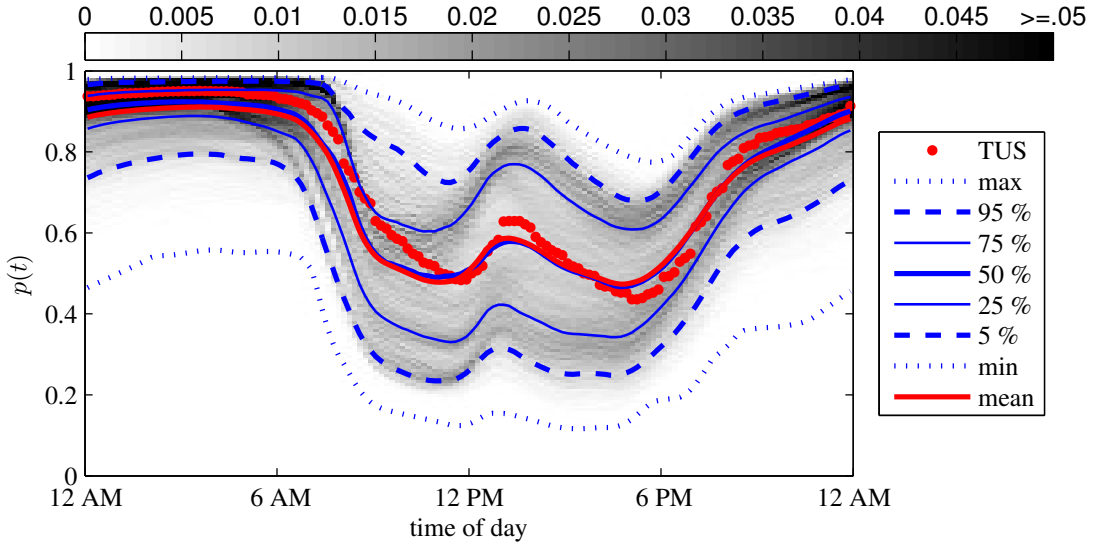


Figure 3.14: Distribution of the presence profiles of the synthetic population predicted by the FOMP.

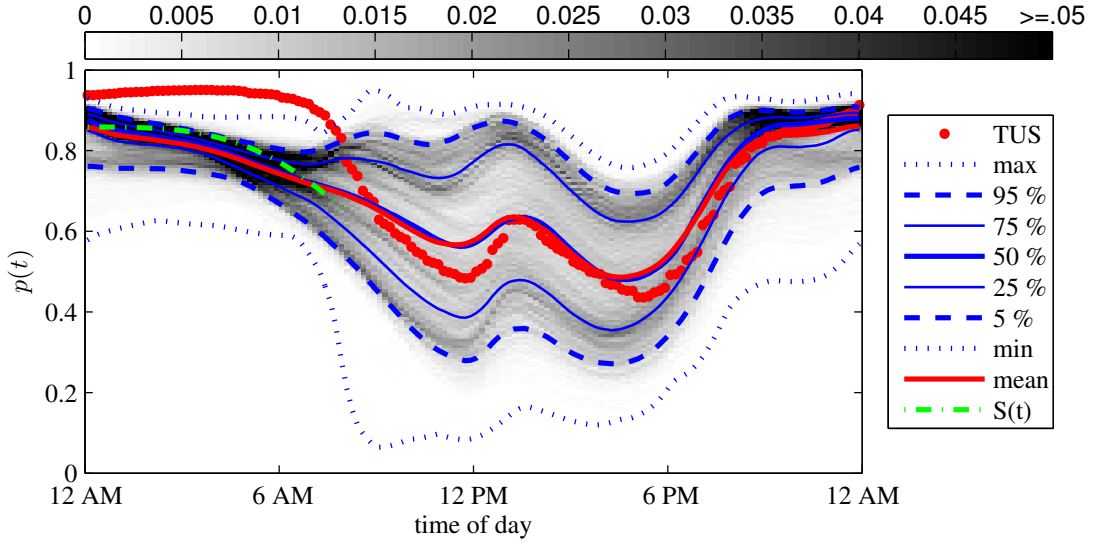


Figure 3.15: Distribution of the presence profiles of the synthetic population predicted by the HOMP.

3.3.3 Model performance comparison

To compare the predictive power of the two models, the log-likelihood $\mathcal{L}$ of the models was calculated. The value of the FOMP is $-9.82 \cdot 10^5$ and the one of the HOMP is $-1.05 \cdot 10^6$. The log-likelihood of the individual-independent model (IIM), (which was introduced at the beginning of Section 3.2.1), amounts to $-1.08 \cdot 10^6$.

The time-dependence of the quality of predictive power can be measured with the time-dependent log-likelihood $\mathcal{L}_\mathrm{t}$, which is evaluated at every time step t, and for which $\mathcal{L} = \sum \mathcal{L}_\mathrm{t}(t)$ holds (it is summed over all time steps of the day). The curves of $\mathcal{L}_\mathrm{t}$ of the three models are shown by the thicker lines in Figure 3.16. The value of the FOMP is above that of the IIM for all time steps, although the former was calibrated in a coarser time resolution. The difference is smaller during the small hours than during the rest of the day, as then there are less significant individual-dependent differences in behaviour than during the day. The value of the HOMP is above that of the FOMP, from 8.30 AM to 11.00 AM, from 12.40 PM to 2 PM, from 3.50 PM to 4.40 PM and at 8.40 PM. Furthermore, it is below that of the IIM from midnight until 8.00 PM, which is due to the inappropriate shape of the used Weibull distributions to fit the empirical PDFs. To illustrate that, the value of $\mathcal{L}_\mathrm{t}$ of the Weibull survival curve (cf. Figure 3.15) is also shown, illustrating that the worse performance is due to the underestimation of the average profile. In addition, the subtotals of $\mathcal{L}_\mathrm{t}$ restricted to presences (1) and absences (0) are also shown for the FOMP and the HOMP. Comparing these, it appears that presences are predicted with a superior predictive power, when they are predicted with the higher order memory from 8.20 AM to 22.50 PM. However, this might be related to the higher mean presence probability that is predicted by the HOMP in this period. In Section 3.3.4, further exemplifications of the time-dependence of the better predictive power of the individual-dependent models will be provided.

In order to capture the distribution of the quality of predictive power over the individuals of the population $\mathcal{C}$, $\mathcal{L}$ was also calculated for every individual $\boldsymbol{x}$, which will be denoted as $\mathcal{L}_{\boldsymbol{x}}$ ($\mathcal{L} = \sum_{x \in \mathcal{C}} \mathcal{L}_{\boldsymbol{x}}$). The distribution of $\mathcal{L}_{\boldsymbol{x}}$ of the 3 models is shown in Figure 3.17. The distribution of the IIM is substantially lower for values of $\mathcal{L}_{\boldsymbol{x}}$ above -50 and larger for values below -190 with respect to the other two models, meaning that there are significantly more very inaccurate predictions and less very good ones, when behaviour is not treated in an individual-dependent manner. The HOMP is performing worse for values above -50, but better for values below -150 compared to that of the FOMP, implying that the proportion of very bad predictions can be reduced when departure transitions are dependent on a longer memory. The worse proportion of very good predictions is probably due to the worse predicted mean presence profile during the night (cf. Figure 3.16).

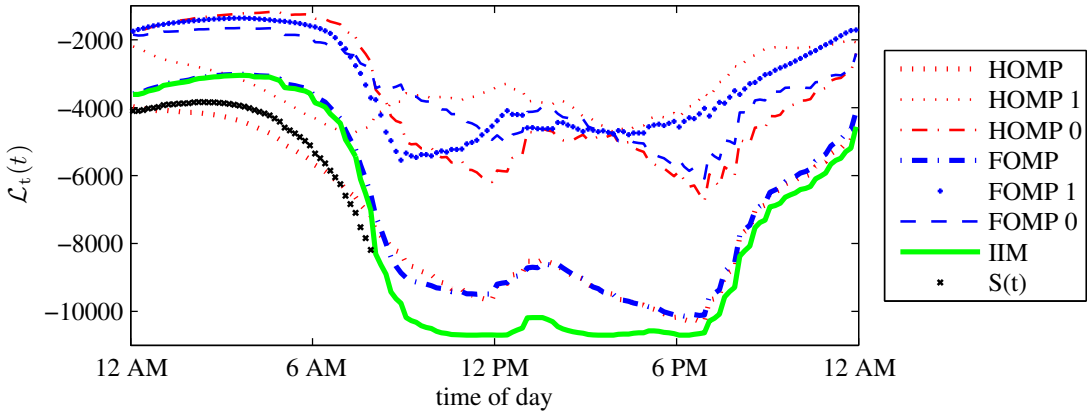


Figure 3.16: The log-likelihood $\mathcal{L}_t$ of the models as a function of time of day, as well as of the individual-independent survival curve after midnight (cf. Figure 3.15) and the sub-totals for presences (1) and absences (0).

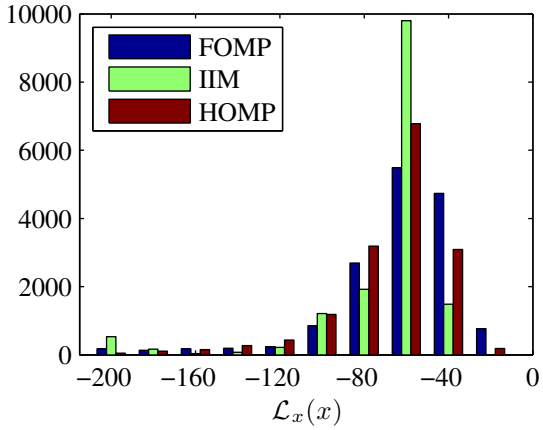


Figure 3.17: Distribution of the log-likelihood $\mathcal{L}_{\boldsymbol{x}}$ of the three models over the synthetic population.

3.3.4 Population characteristics dependence of predicted presence profile distributions

In order to illustrate the dependence of the models' predicted presence profile distributions $f_p(t)$ on the individual characteristics $\boldsymbol{x}$, these distributions will be shown for distinct sub-populations of the synthetic population. The graphs in Figures 3.18 to 3.23 are all constructed in the same manner, which will be demonstrated by the example of Figure 3.18, where the sub-population distributions correspond to a splitting of the synthetic population according to the weekday. In all Figures 3.18 to 3.23, the presence probability during the night is underestimated due to the censoring of arrivals at midnight, which was mentioned

in Section 3.3.2.

FOMP

In Figure 3.18, the part of the population predictions on Saturday and Sunday is specified by (we) and illustrated in magenta, whereas workdays (wd) are illustrated in cyan. The distributions are visually specified by their minima/maxima, as well as their mean values, which are indicated by the curves. Furthermore, the presence profile probability distributions of the sub-populations $f_p(t)$ are shown, which were derived in the same manner as it was described in conjunction with Figures 3.2 and 3.3. The magnitudes of the superposition of the two distributions are colour-coded according to the scale that is shown top-right. The minimum probability of being absent in the forenoon is up to more than twice larger for workdays than for the weekend. The minimum presence probability in the forenoon during the weekend is more than 3 times larger as at workdays.

In the forenoon, the two sub-population distributions are rather unimodal, whereas in the afternoon, both of them are rather bimodal. This behaviour will be explained together with Figure 3.20. At the evening and during the night, the observed presence profile during the weekend is higher than that at workdays, in opposition to the mean predicted profiles. In contrast, the maxima of the predicted presence distributions show the same relationships, whereas the minimum in the small hours is considerably lower at workdays than at the weekend.[10] Accordingly, on a Sunday, the arriving and leaving probabilities that precede the presence probabilities before midnight are over- and underestimated, respectively.

In Figure 3.19, the distributions of the two sub-populations of men and women are illustrated. In this case, the differences between the observed and the predicted mean are smaller than in Figure 3.18, but the sign of these differences is correctly predicted for the averages, the maxima and the minima of the distributions. For men, the distributions are larger than for women, which can be explained by the arriving and leaving probabilities, which are respectively significantly higher and lower, almost during the entire day. The observations are well reproduced by the average model predictions. The largest deviations occur in the morning, whose reason is discussed with Figure 3.20.

In Figure 3.20, the two sub-population distributions of individuals working in a full-time and those not being in paid work are shown (empstat: “full”/“no”; cf. Figure 3.1). In contrast to Figures 3.18 and 3.19, in this case the set union of the

[10] These low night-time presence probabilities at workdays might be related to people working in night shifts, for whom in reality presence probabilities are probably even lower than the predicted ones during this period, implying that these overestimated presence probabilities entail an underestimation of the remaining ones. Another explanation is related to the 24 h periodicity when estimating the profiles. The latter implies that before the small hours on Friday (in the evening), the probability of arriving is underestimated, and the probability of leaving is overestimated (see the corresponding parameter values of Thu/Fri in Figure 3.1).

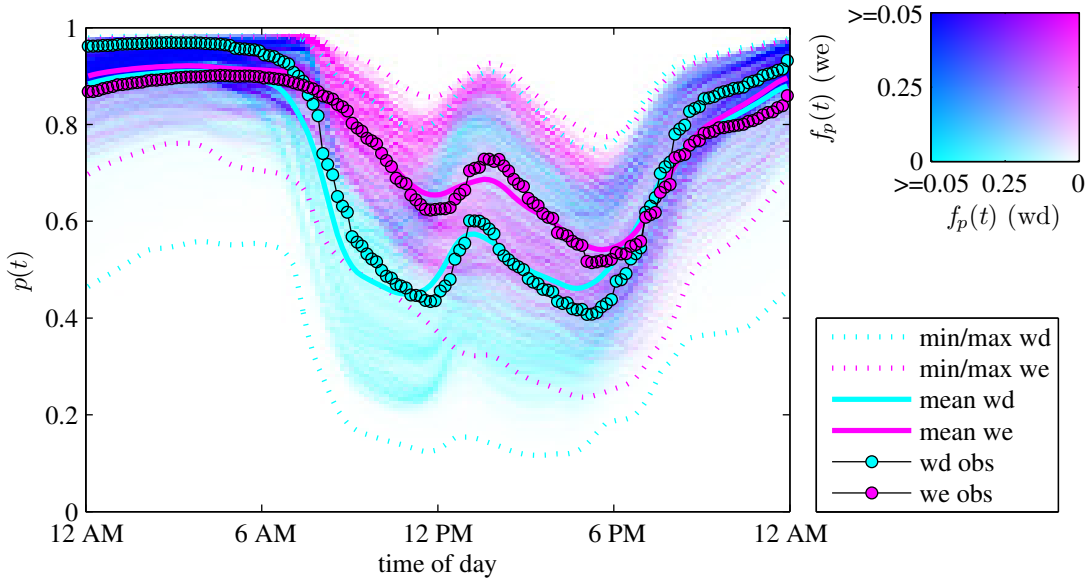


Figure 3.18: Superposed presence profile distributions of the sample subpopulations on weekends (we)/workdays (wd) predicted by the FOMP.

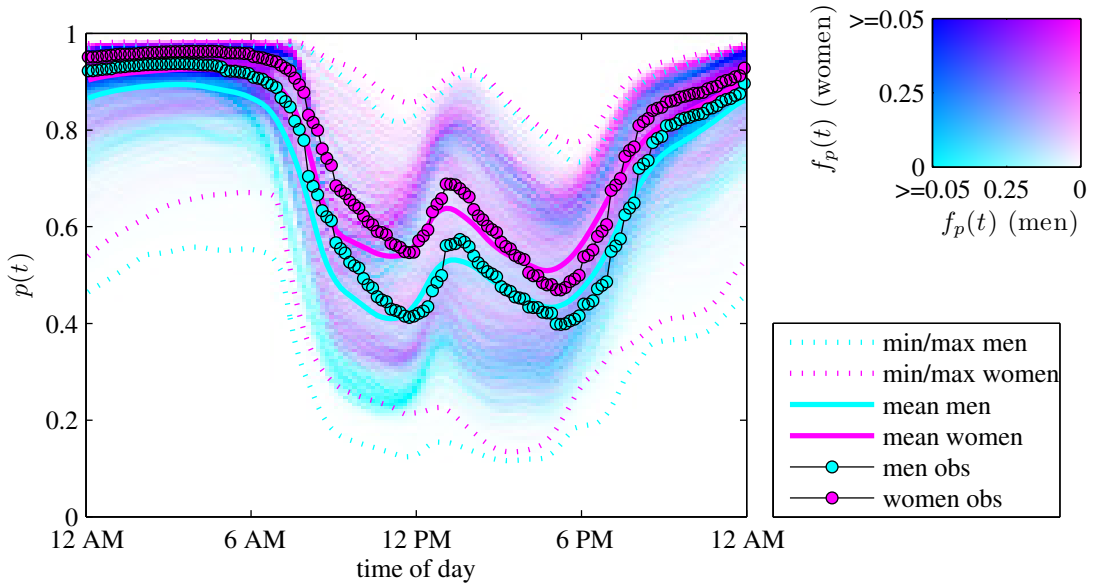


Figure 3.19: Superposed presence profile distributions of the sample subpopulations of men/women predicted by the FOMP.

two sub-populations is not equal to the entire synthetic population. Therefore, the maximum (minimum) of the maxima (minima) of both does not always correspond to the maximum (minimum) in Figure 3.14. Here, the differences between the sub-population presence profiles distributions are most considerable, which is again directly related to the arriving/leaving transition probabilities in Figure 3.1), which are significantly different for the two sub-populations. Here, the intervals of significant different probabilities are fewer but with higher amplitude of the parameters. The minimum predicted absence probability for fully employed is at least 55 % higher during the whole 24 h, and outreaches a factor of four at 7.40 AM. The mean value of the presence probability of individuals without paid work is up to 86 % higher at 2.20 PM, compared to fully employed individuals. The mean predicted probabilities are in very good agreement with the observation for the employed sub-population; for the non-working sub-population the largest deviation of -19.6 % occurs at 9.00 AM. The underestimation of the mean presence profile of non-working individuals might be due to a mimicking of the behaviour of fully employed individuals in the model, related to a removal of parameters in the backward elimination of the transition probabilities, which express these differences. Comparing Figure 3.20 with Figure 3.18, the bimodal character of the two distributions in the latter in the afternoon can be explained by an increase of the presence probability at the weekend, which takes place for individuals with and without paid work.

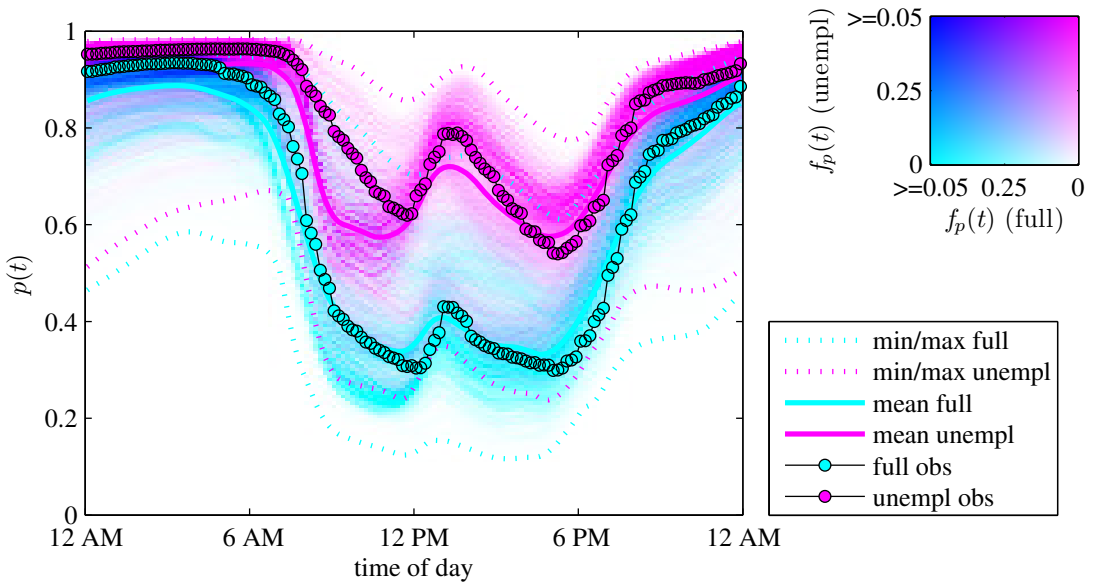


Figure 3.20: Superposed presence profile distributions of the sample sub-populations working in a full-time position (full)/not being in paid work (unempl) predicted by the FOMP.

HOMP

In this section, the distributions of the same sub-populations as in Section 3.3.4 will be illustrated and described. As it was mentioned in Section 3.3.2, the uncertainties of the presence profile distributions are smaller than for the FOMP. Moreover, the extrema of the distributions almost coincide during some periods of the day. This is related to the individualisation procedure of the presence PDFs $f_{t_s}(\boldsymbol{x}, t)$ (cf. Section 3.2.2), which is less diverse than for the logistic regression models of the transition probabilities[11] (in case of coinciding curves, there was no distinction between the PDFs of the considered sub-populations).

In Figure 3.21, the sub-population distributions during the weekend and at workdays is shown. During the night, there are small differences between the averages of the two distributions, and furthermore, the mean of workdays passes under that of the weekend much earlier than the observations. However, the sign of the difference of the two averages better reproduces that of the observations, compared to the predictions of the FOMP. The extrema of the two distributions take a similar course and the two distributions in the daytime also resemble those of the FOMP.

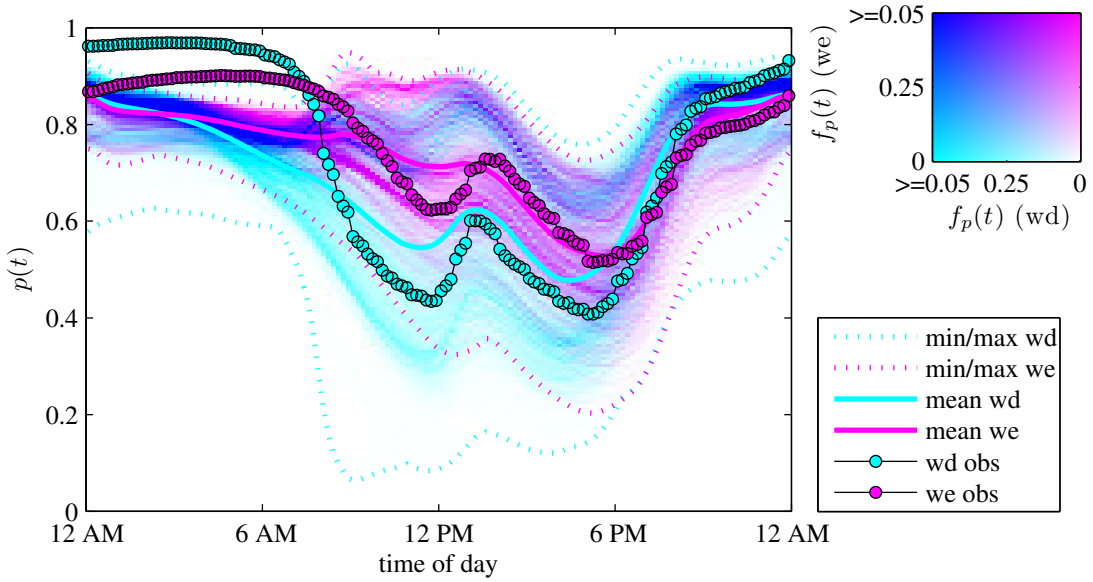


Figure 3.21: Superposed presence profile distributions of the sample sub-populations on weekends (we)/workdays (wd) predicted by the HOMP.

In Figure 3.22, the distributions of the two sub-populations of men and women are illustrated. As in Figure 3.19, the differences between the two sub-populations

[11] There were maximally 91 different $f_{t_s}(\boldsymbol{x}, t)$ (in the first interval after midnight), whereas for the FOMP, the number of different values for the elements in $\boldsymbol{T}(\boldsymbol{x}, t)$ is given by 2^K (K: number of significant parameters in the interval except the ASC; cf. Figure 3.1).

are smaller than those in Figure 3.21. Furthermore, differences between the means and the extrema also corresponds to that of the observations, except for some insignificant exceptions.

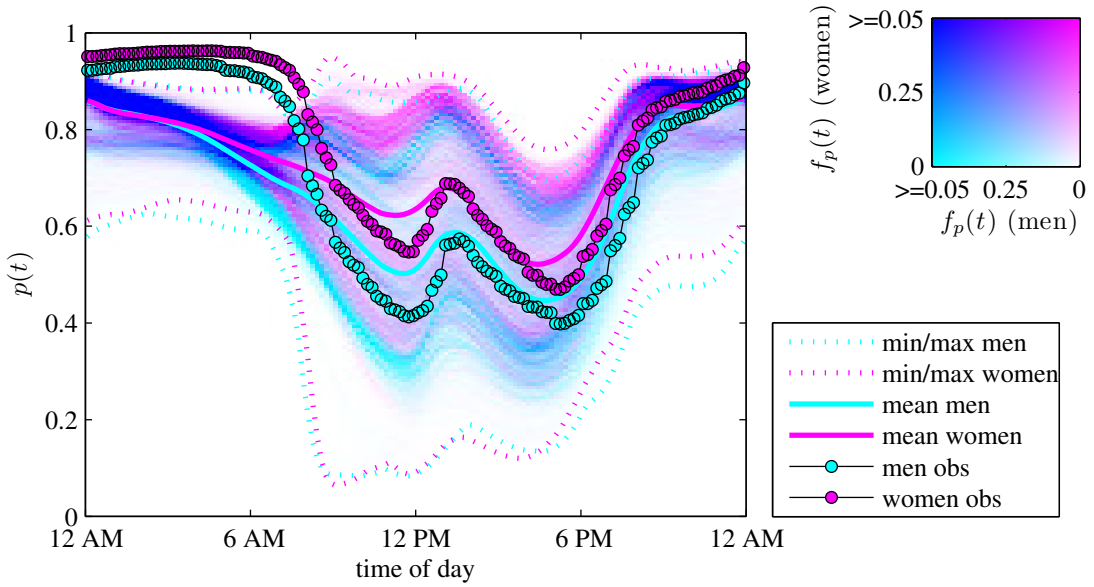


Figure 3.22: Superposed presence profile distributions of the sample sub-populations of men/women predicted by the HOMP.

In Figure 3.23, the two sub-population distributions are shown for fully employed and non-working individuals. Likewise for the FOMP, the differences between means of these two sub-populations are most significant in the daytime. There, the average presence of the working individuals is overestimated by up to 9.6 % at 12.40 PM. Except some exceptions the differences of the extrema have the same sign as the observations.

3.4 Discussion

Two bottom-up models have been formulated to derive asymptotically with fast convergence the time-dependent residential presence profiles of individual members of statistically significant demographic sub-populations. The approach shows a good ability to reproduce the observed profiles of the calibration set. Whereas a straightforward individual-independent approach would exactly reproduce the latter, it yields very inaccurate predictions of individual specificities. Furthermore, it was shown that the two Markov approaches capture more realistically the time-dependent distribution of presence durations. In the first one however, being a time-inhomogeneous first-order Markov process, the observed time-dependent presence distributions cannot be ideally reproduced. This can be done

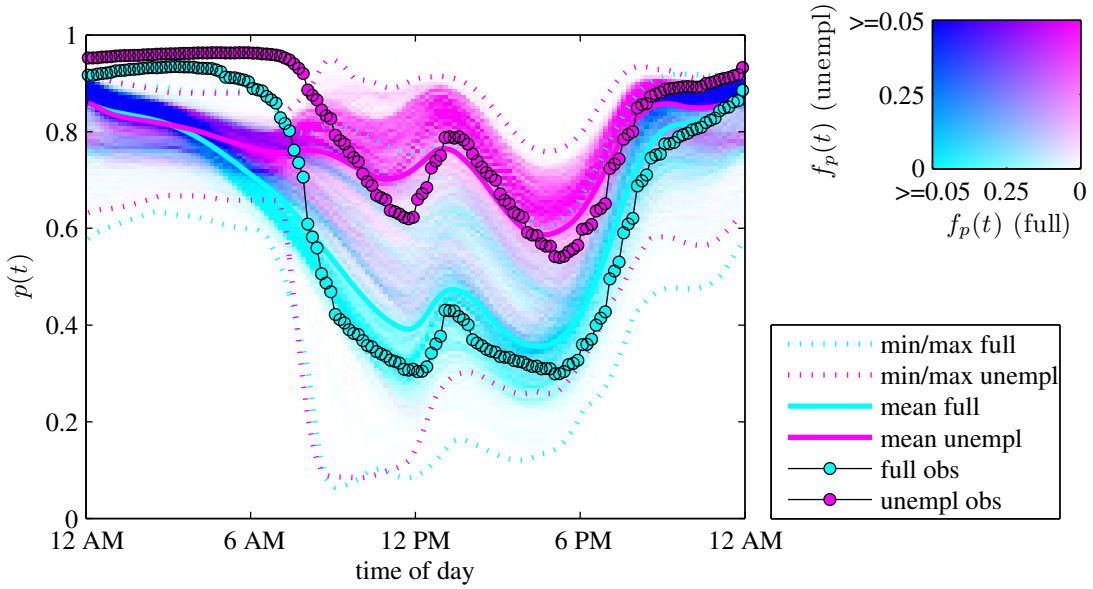


Figure 3.23: Superposed presence profile distributions of the sample subpopulations working in a full-time position (full)/not being in paid work (unempl) predicted by the HOMP.

more consistently on the basis of the second approach, which is based on survival analysis.

The assumptions in this chapter imply that individuals behave individually of each other, which does not reflect reality, for people who live in the same household. The dependence of behaviour between household members could be tackled with a model calibrated with and predicting occupancy patterns of entire households, which was omitted as this would substantially complicate the approach.

The predictive power of the first-order Markov process was shown to better predict individual-specific behaviour than the individual-independent approach at every time step of the day, although the former was calibrated in a coarser time resolution. During the night this superiority is less accentuated than during the day, where there is more variability between individuals and thus the model calibration with significant individual-dependent behaviour is more manifold. The time-dependence of the predictive power of the higher-order Markov process is comparable, except at time periods, where the Weibull distributions used to fit the observed duration distributions were not in agreement with the observations. However, in general there should be a superior predictive power of the survival approach related to a better representation of the presence distributions. In other words, in the first-order Markov approach the time-dependent durations of presences of an individual may end at times when it is not realistic. However, the representation of presence duration distributions with Weibull PDFs is not always

well reflecting reality. This was discussed in conjunction with Figure 3.15, where it was shown that the distribution of presence durations in the first hour interval of the day do not well reproduce the observed ones (*cf.* Figure 3.6). In general, the HOMP does not capture the amplitudes of the oscillations of the observed presence proportion at noon, which might be due to the too coarse calibration resolution, the approximation of the empirical PDFs by Weibull PDFs, as well as the differences due to approximations that were discussed in Section 3.3.1. However, the presence patterns are well reproduced, and from noon on, the predicted mean is in good agreement with the observations.

The worse distribution of the predictive power over the individuals of the synthetic population of the individual-independent compared to the two other approaches shows that individual-specific modelling decreases the proportion of very bad predictions. The worse performance of the survival model of very good predictions compared to that of the Markov model is probably due to the bad representation of the duration distribution after midnight.

The significant dependence of the predictions on individual specificities shows that the latter should not be neglected in a realistic modelling approach and the small deviations between the mean predicted probabilities and the observed means of sub-populations show that the two modelling approaches are appropriate to capture/describe the dependence of the presence profiles on individual characteristics.

3.4.1 Model application in simulations

All the models for the prediction of occupancy $y(\boldsymbol{x}, t)$ of an individual $\boldsymbol{x}$ at time t may be applied in dynamic simulation tools. A general scheme for implementing the HOMP of Section 3.2.1 is provided in Figure 3.24, which consists of the following steps:

1. When an occupant $\boldsymbol{x}$ is absent, a presence is started with a probability of the first-order transition probability to start a presence $t_{01}(\boldsymbol{x}, t)$, by comparing the latter with a uniformly distributed random number $r \in [0, 1[$.

2. (a) If $t_{01}(\boldsymbol{x}, t) < r$:
   The time is incremented by one time step and the occupant stays absent by setting $y(\boldsymbol{x}, t) = 0$.

   (b) If $t_{01}(\boldsymbol{x}, t) \geq r$:
   A presence is started, whose length is determined by drawing a duration Δt from the corresponding PDF at the current time $f_t(\boldsymbol{x}, \Delta t)$.[12]

[12] In a discretised Markov chain the PDF has to be replaced by the corresponding probability mass function, in order to yield durations which are multiple integers of the length of the time step.

The occupancy variable is set to one during the corresponding interval $[t, t + \Delta t]$, and then, the time is incremented by Δt, where a new absence begins by setting $y(\boldsymbol{x}, t) = 0$.

3. The procedure is repeated with step 1 until the desired maximal simulation time is reached.

By replacing step 2b with steps that correspond to the steps 1 and 2a based on the transition probability for leaving $t_{10}(\boldsymbol{x}, t)$, the procedure corresponds to the time-inhomogeneous FOMP of Section 3.2.1. In case of very small probabilities to start a presence, the Markov process is computationally ineffective, as step 2a has to be repeated very often until a presence is started. This can be avoided, by replacing step 2a with a step, corresponding to step 2b based on the distribution of absence durations, which can be derived as it was shown in Equation (3.13) for presences.[13]

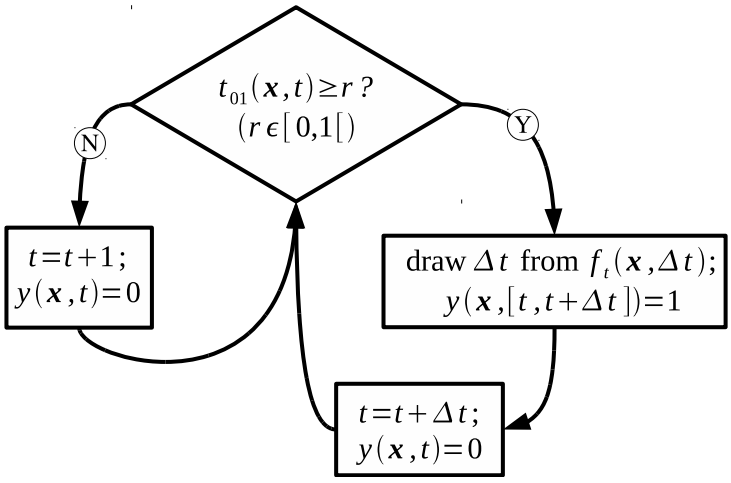


Figure 3.24: Flow chart for the application of the FOMP in a simulation.

3.5 Conclusions

In this chapter, we presented several bottom-up approaches to the modelling of time-dependent residential presence using first- and higher-order Markov processes based on transition probabilities and presence duration distribution functions. The higher-order Markov process was validated by approximately reproducing the solution of the master equation when the according transition probabilities are transformed into the corresponding presence duration distribution

[13] In this case the simulation does not need to be discretised, but can be based on continuous time, only limited to the computational resolution.

functions. The purpose of these models is to support the more accurate representation of occupancy profiles for the use in dynamic building energy simulations and the modelling of the use of electrical appliances. The model parameters were calibrated to reflect the individual behavioural characteristics of observations of a population of respondents to a time use survey questionnaire, fitting the parameters to reflect the individual behavioural characteristics, to support more accurate and tailor-made predictions – for example to accommodate predictions for specific household demographies or to account for the day of the week. In developing these models, considerable effort has been put into the design of the calibration methodologies, to avoid including characteristics which are statistically insignificant. The approaches were then validated by deriving the time-dependent distribution of the presence profiles of sub-populations, showing that the individualised calibration does indeed significantly improve the predictive accuracy.

The first-order Markov process reproduced more accurately observed sub-population presence profiles than the model of higher order. However, it was demonstrated that the worse performance of the latter relates to the shape of the Weibull distributions that were used to model presence durations, and that the latter cannot be consistently reproduced in a first-order Markov process. Thus, we recommend to use the latter in applications, where there is no particular importance of presence durations, and the higher-order Markov process otherwise.

In this chapter, residential absences were modelled based on the first-order Markov property. But as results show, that these should also be treated based on a higher-order memory, a possible future variant of the model might treat presences in the first order and absences in a higher one, or even both states using the generalized master equation. Furthermore, the Weibull distributions used to fit the presence PDFs might be replaced by other candidates, such as Bi-Weibull or kernel density estimations, which are appropriate to reproduce multimodal distributions. The empirical distributions could also be reproduced more accurately with a higher degree of individualisation by means of Equation (3.13), when the corresponding transition probabilities are calibrated in dependence of the presence start. As the model can be readily calibrated to other datasets, it should be applied to study differences between countries or temporal changes, as well as for different types of building use, such as workplaces, restaurants, or to consider travelling periods.

Chapter 4

Residential activity modelling

A bottom-up modelling approach together with a set of calibration methodologies is presented to predict residential building occupants' time-dependent activities, for use in dynamic building simulations. The stochastic model to predict activity chains is calibrated using French time-use survey data (of 1998/1999), based on three types of time-dependent quantities: (i) the probability to be at home, (ii) the conditional probability to start an activity whilst being at home, and (iii) the probability distribution function for the duration of that activity. The behaviour of the individual agents in the model is first calibrated using a generic approach, where every individual is assumed to behave the same. A refinement is then presented to account for variations in the behaviours of sub-populations, having specific individual characteristics. Furthermore, a statistical approach is introduced for the modelling of transitions between two successive activity types as a Markov process. The models are then validated using a cross-validation technique, and their predictive performance is compared at an individual level, as well as for aggregated (sub-)populations.[1]

4.1 Introduction

As passive design standards of buildings lead to a more efficient use of solar energy and conservation of casual heat gains, so buildings' energy and environmental performance becomes more sensitive to the presence, activities and activity-related behaviours of their occupants [41] such as the use of shading devices, windows and electrical appliances [see 16, 42, 43]. In order to optimise the design of future high performance buildings and better match their energy demand with local generation and storage capacity, it will be necessary to consider the stochastic nature of their occupants. In this regard, the time-dependent activities which

[1] A preliminary version of the models in this chapter was published at the 12$^{\text{th}}$ International Conference of the International Building Performance Simulation Association [89]. A more complete article was published in the journal *Building and Environment* [90].

are performed in buildings and the behaviours that depend on them are crucial [91]. Examples include, the preference of luminosity whilst sleeping, which differs significantly from those whilst other activities are performed; the likelihood of window opening which is highly increased whilst/after cooking to evacuate pollutants, as well as the use of specific electric appliances which is highly correlated with the specific activity that is being performed. In this context, the main benefits of stochastic models are to contribute to estimate quantitatively the uncertainties of the quantities of interest or more generally their distributions.

With regard to the influences of human behaviour on single buildings, a bottom-up approach is needed to faithfully encapsulate the full range and timing of occupants' presence, activities and dependent behaviours on the buildings' energy balance. Such an approach also lends itself well to the modelling of future scenarios to explore responses to changes in the physical composition of buildings or the ownership of appliances as well as to changes to the population's demographic/behavioural characteristics.

4.1.1 Previous research work

In transport research, activity choice modelling is widely applied [92], addressing correlations between members of the same household [93, 94], the occurrence of multiple simultaneous activities [95], time-dependence [96] or the dependence on household characteristics [97].

In the social sciences and economics, models have been developed to forecast temporal changes in behaviour. Helbing, as well as Gershuny and Sullivan give a general overview of the subject [98, 99]. However, these models rather focus on seasonal/annual evolution [*cf.* 100] rather than the dependence on the time of day. Fischer and Sullivan model the latter by applying simulations based on genetic algorithms [101]. Furthermore, besides pairs of subsequent activities, even the likelihood of triples of subsequent activities is taken into account in their model. However, the activity transition probabilities are not time-dependent (for instance, in the evening the probability that sleeping will follow the activity of personal care is typically higher than in the morning). Moreover, the explanatory variables used in the simulations are not dependent on the characteristics of the individuals (although a distinction between weekdays and weekends is made). Numerous studies focus on predicting the use of time depending on personal and socio-economic characteristics [102]. The latter dependence is investigated in many different research fields, where time use survey data is used to describe quantitatively a given activity [*cf.* 103–106]. However, these studies do not account for the dependence of activities on the time of day.

In the context of residential electricity and/or domestic hot water demand prediction, the modelling of activities has already been considered. Torriti presents an assessment of national differences in active occupancy levels in single-person households [107] and a simple time-dependent home-activity model is presented

by Capasso et al. [108]. In the more sophisticated approaches, either the model to predict individual activity chains is based on aggregate proportions and can thus not capture individual-dependent behaviour or the increased/decreased likelihood of the occurrence of a sequence of activities [25, 45] (Yamaguchi et al. present a modified version [48]), or they are based on first-order discrete-time Markov-chains, not being able to model coherently the duration distribution of activities (as the transition probabilities do not depend on the time the activity was initiated) [20, 40, 63]. Richardson et al. follow an approach where domestic activities are predicted as the observed fraction of present non-sleeping occupants of the corresponding weekday type [23]. In this approach, activity durations cannot be captured coherently and all artefacts of the time use survey data are exactly adopted. The probabilistic quantities have not as yet been calibrated depending on individual characteristics. Furthermore, the calibration of the Markov chain transition probabilities with time use data is susceptible to data scarcity issues, so that transition probabilities are set to zero if no such a transition occurred in the survey.

In the context of office building environments a stochastic model was developed that predicts the occurrences of intermediate activities at work [109]. However, the calibration data for the model only contains information relating to 8 persons working in academia and thus cannot be considered representative for all types of working people. Furthermore, the steering quantities of the simulation do not depend on the time of day.

From the above review, we conclude that:

- Residential activities considerably influence building performance, and these should be modelled dynamically to faithfully represent reality.
- There is no existing approach that models the time-dependence of the stochastic activities based on physical quantities.
- Existing activity models do not encapsulate individual specificities (for instance, personal/household characteristics).
- Published research does not suggest a common robust cross-validation procedure, which hinders any meaningful comparison of the quality of alternate models.
- The predictive power of existing models has not been evaluated at an individual level, which is crucial for a robust bottom-up approach.

Looking at the state of the art of activity modelling, it is evident that the detailed validation of models is crucial and, that there is a lack of published research of many aspects influencing human behaviour. This hides model inaccuracies, when they are applied in contexts which differ from the one to which the models are fitted [*cf.*, *e.g.,* 110].

The objective in this chapter is to establish a bottom-up approach, which allows residential activities to be modelled as a function of time and individual specificities. The time-dependence is crucial for the use in dynamic building simulations, as the probabilities to perform activities vary significantly with time of day. The bottom-up nature of the approach is needed to enable its application for scenario testing based on populations with different characteristics to those of the calibration dataset. Both the nature of the model formulation, which can be readily fitted with other datasets, and the results from cross-validation studies suggest that this novel model is better adapted to dynamic building simulation than previous variants.

Summary

A detailed methodology is presented, allowing to predict residential occupants' activities. In Section 4.2, the steering quantities of the simulation models are presented and the time-use data used to calibrate them are described. Then, in Section 4.3, simulation results are presented and the performance of different grades of refinements of the model is compared. Moreover, the accuracy of the different models is verified using a ten-fold cross-validation. In Section 4.4, the approach is discussed and further possible refinements are identified, and finally it is concluded in Section 4.5.

4.2 Method

In this chapter we focus on the modelling of activities based on the information of the database described in Chapter 2.

4.2.1 Model structure and calibration

The physical quantities of interest in this chapter are the probabilities $p_j(t)$ ($j \leq N$) to perform the different activities depending on the time of day. The duration distributions of the different activities are also of importance, *e.g.* when applying the model in dynamic building simulations. In order to model/test scenarios of populations with different demographic characteristics, it is also important that the model be formulated to support examination of individual-specific variations of behaviour, likewise to accurately capture the distribution of outcomes at the scale of one building owing to uncertainties in the household's composition.

At the resolution of the individual, activity chains are characterised by step functions $a_{\boldsymbol{x}}(t)$, of the activity type j of individual $\boldsymbol{x}$ over time t. In the residential context this is only of interest during the time periods whilst $\boldsymbol{x}$ is present in their residential environment. These presence chains are given by time-dependent step functions equal to one when $\boldsymbol{x}$ is at home and zero otherwise.

A modelling approach has been formulated with which the $p_j(t)$ can be approximately determined via Monte Carlo simulation. To simulate residential activities, individuals residential presence/absence chains should be determined by an independent pre-process (by neglecting influences/interactions of the performed activities on the absences/presences, as the latter are modelled in the post-process of the activity model). Although, this occupancy pre-process is to be modelled stochastically in general, in this chapter the observed occupancy data are directly used to assess the performance of the activity model itself, rather than its combination with an occupancy model.

The algorithm is illustrated schematically in Figure 4.1, which shows how the time-dependent activity chain of every individual of the synthetic population in a simulation is generated. After having modelled in a pre-process the occupancy chains for every individual $\boldsymbol{x}$, we need to check whether $\boldsymbol{x}$ is at home at time t. If $\boldsymbol{x}$ is not at home, we forward the simulation time to the next arrival time t_{arrive} of $\boldsymbol{x}$. If $\boldsymbol{x}$ is present, the activity j that is started is determined, according to $p_{\text{s},j}(t)$, the probability to start activity j at time t. Afterwards, the duration Δt of this activity is drawn from the $f_j(t)$, the duration probability distribution function of j at t, and the time is incremented by Δt. The two quantities $p_{\text{s},j}(t)$ and $f_j(t)$ will be explained in more detail in the next paragraph of this section. However, if the departure time t_{depart} of $\boldsymbol{x}$ is exceeded, the end time of j is set to t_{depart} (the residential activity is forced to end when the individual leaves home). As long as the departure time is not exceeded, new activities j and the corresponding time increments are determined repeatedly. Following from an individual's departure, the time is set to the next arrival t_{arrive} and the procedure is repeated.

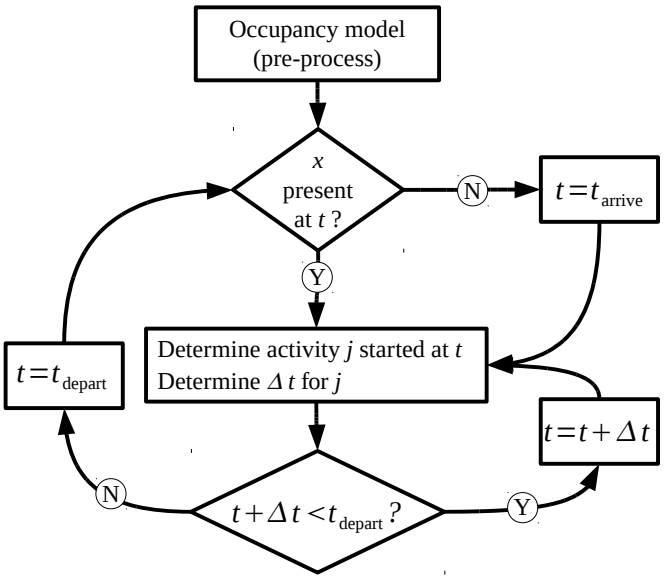


Figure 4.1: Schematisation of the algorithm, which is applied for every individual of the population.

This approach is based on two stochastic quantities whilst individuals are at their residences. First, an activity j is started with the probability $p_{\mathrm{s},j}(t)$ (by choosing $j = \min_j \left(\sum_{j'=1}^{j} p_{\mathrm{s},j'}(t)\right) \geq r$, where r is a uniformly distributed random variable in [0,1[; *cf.*, *e.g.*, [111]; this procedure corresponds to the inverse function method to sample from a discrete state distribution [cf., e.g., 69]). Second, after simulating which activity j is started, its duration Δt is drawn from the corresponding time-dependent probability distribution function (PDF) $f_j(t)$. In order to simplify comparability of the simulation results with the monitored TUS data and to reduce computational overheads, the continuous time values that are drawn are rounded to the 10 min time resolution of the TUS.

Regarding the calibration of $p_{\mathrm{s},j}(t)$ and $f_j(t)$ with the TUS data, there are censored (truncated) events at the beginning/end of the individuals' diaries if no information is given at which time outside of the scope of the questionnaire the first/last activity started/ended. A similar case occurs at the end of a period of residential presence during the day, where it is not known whether the duration of the activity would have lasted longer if the departure of the individual had occurred later. This is modelled in the simulation by forcing an activity to end, as soon as its duration exceeds the time of departure from the residence given by the presence pre-process.

Starting probabilities

Generic approach The conditional probabilities to start an activity j whilst being at home $p_{\mathrm{s},j}(t)$ (which will be referred to as starting probabilities) can be calibrated directly from the corresponding time-dependent ratio of instants the activity j was started over the total number of times any activity was started at home at time t by all the individuals $\boldsymbol{x}$ of the population $\mathcal{C}$ in the TUS data[2]. For ease of notation these conditional starting probabilities will be referred to simply as "starting probabilities".

Multinomial logit model of starting probabilities This more complex approach involves determining starting probabilities accounting for individuals' characteristics. These starting probabilities have been conceived to depend on $M = 41$ dummy variables $x_1, x_2, \ldots, x_M$ describing the characteristics of the individual, assembled in a vector $\boldsymbol{x}$. The corresponding starting probabilities will

[2]There are two problems arising when the probabilities in the model are calibrated in this way. First, for activities which are infrequent at a given time of day, there would be large uncertainties due to data scarcity issues. Second, this calibration method would be sensitive to the apparent rounding of time values to half and full hours mentioned in Section 2 (*cf.*, Figures 2.8 and 4.2). Therefore, the starting probabilities were first calibrated on hourly averages to reduce data scarcity issues as well as to level out a bias due to rounding artefacts. Then, the starting probabilities used in the simulations were deduced by linearly interpolating between hourly averages to achieve a smooth change in those quantities.

be denoted $p_{\mathrm{s},j}(\boldsymbol{x},t)$ and each of the dummy variables x_k ($k \leq M$) correspond to one of the values of the categories in Table 4.1.

A random utility model (RUM) was used to capture the influences of the mentioned characteristics in the calibration of the starting probabilities $p_{\mathrm{s},j}(\boldsymbol{x},t)$ [112, 113]; more precisely, 24 such models have been calibrated for every hour interval of the day [*cf.* 40]. In a RUM, every individual has a finite number of choices N, in the present case corresponding to the number of all activities that can be performed in a residential environment. For a given hour interval $[t, t+1\ \mathrm{h}[$, this approach estimates the N different probabilities $(p_{\mathrm{s},j}(\boldsymbol{x},t))_{j=1,\ldots,N}$ for all the individuals $\boldsymbol{x}$ in the sample. For a given hour, the dimensionless utility functions V_j for every choice j in the sample, were defined to depend linearly on the parameters in $\boldsymbol{x} = (x_1, \ldots, x_M)$

$$\begin{aligned} V_1(\boldsymbol{x}) &= \alpha_1 + \beta_{1,1}x_1 + \beta_{1,2}x_2 + \ldots + \beta_{1,M}x_M, \\ V_2(\boldsymbol{x}) &= \alpha_2 + \beta_{2,1}x_1 + \beta_{2,2}x_2 + \ldots + \beta_{2,M}x_M, \\ &\vdots \\ V_N(\boldsymbol{x}) &= \alpha_N + \beta_{N,1}x_1 + \beta_{N,2}x_2 + \ldots + \beta_{N,M}x_M. \end{aligned} \tag{4.1}$$

Table 4.1: Dummy variables for the starting probabilities. When there is only one value specified, this corresponds to a dichotomous situation, where the complementary value is not listed.

category	**value set**
age	<26, 26-45, 46-62, >62
carer	Diarist looks after an adult or child with a disability
day	Mon-Sun
disab	Diarist has a disability or long-term limiting health condition
edu	Level of education above secondary education
healthy	Diarists self-reported general health status good/very good
hh1	Single household
inc	Household income above lowest 25 %
noveh	Household does not have access to a private vehicle
retired	Diarist retired
owner	Diarist's household rents home
pc	Diarist's household has a computer
sex	Diarist male
stud	Diarist is a student
urb	Diarist lives in urban/suburban area
work	Employed full-time, half-time, not in paid work, unknown

Given these utility functions, the probability of starting an activity j is predicted using a multinomial logit (MNL) model of the form:

$$p_{\mathrm{s},j}(\boldsymbol{x},t) = \frac{e^{V_j(\boldsymbol{x})}}{\sum_{j'=1}^{N} e^{V_{j'}(\boldsymbol{x})}},\ j = 1,\ldots,N. \tag{4.2}$$

The alternative-specific constants (ASCs) α_j and the parameters $\beta_{j,k}$ ($j = 1,\ldots,N;\ k = 1,\ldots,M$) were estimated. This is done by maximising the likelihood function, *i.e.*, the probability that exactly the observed outcome of the choices in the monitored data is predicted. Technically, this is done by maximising the log-likelihood of the model, by adjusting the values of the ASCs and the parameters. However, this represents an optimisation problem and in a RUM only the differences of the utility functions of different alternatives in Equation (4.1) have an influence on the choice probabilities in Equation (4.2). To have a unique optimum of the model in the parameter space, some of the parameters have to be fixed. Namely, all the parameters and the ASC of one of the choices j, as well as one of the parameters in every choice when there is a parameter for a segmentation of the whole "population", *e.g.*, the parameter of one weekday from Monday to Sunday. The optimisation of these discrete choice models has been performed using the open source software Biogeme 1.8 [114–116].[3]

In order to remove the influence of characteristics with insignificant parameters in Equation (4.1), two other models for $p_{\mathrm{s},j}(\boldsymbol{x},t)$ are used in the simulations. The 24 models for the starting probabilities have therefore been determined using the initial model given by Equation (4.1) and eliminating all the parameters where the p value of their t test is not statistically significant (*i.e.* below a given threshold α). However, in the new model which is obtained in this way, new parameters can become insignificant, due to multicollinearity. Therefore, backward elimination was repeated for the newly estimated model, until all parameters of the model are significant with a p value below α. This backward elimination has been done with two values of α of 5 % and 10 %. This led to 24 MNL models for the starting probabilities which incorporate in total 1397 and 1855 significant parameters $\beta_{j,k}$, (corresponding in average to 4.4 and 5.8 parameters per activity and time step) respectively.[4] An MNL model, where there are only ASCs α_j corresponds to the generic approach for $p_{\mathrm{s},j}(t)$ [112].

[3]The complexity of the model was adapted to the data abundance by merging together to the same choice $\tilde{j}$ in the MNL model all activity types j which did not occur more often than 50 times. In the simulation this was retranslated to the original activity types by applying the same generic approach, as in the beginning of Section 4.2.1; meaning that the starting probability in the simulation was defined to be $p_{\mathrm{s},\tilde{j}}(\boldsymbol{x},t)$ times the proportion of j in the total number $\tilde{j}$. This led to 24 MNL models for the starting probabilities which incorporate in total 7950 parameters $\beta_{j,k}$. In average, there were 14.25 choices in the 24 models (including $\tilde{j}$).

[4]The estimated MNL models are available online [85].

Probability distribution of durations

After an individual has started an activity j in the simulation, the corresponding duration is determined by drawing a duration from the corresponding probability distribution function of durations $f_j(t)$. The empirical PDFs of the TUS are unrealistic representations for $f_j(t)$, due to the rounding artefacts in the respondents' questionnaires at integer multiples of half an hour (*cf.* Figure 4.2). Furthermore, using these would entail a substantial data storage overhead for simulation purposes; therefore, fitted Weibull PDFs are used to describe the distribution of durations in the simulation model, whereby

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right), \tag{4.3}$$

in which λ and k are respectively the scale and the shape parameter, which can be readily fitted using standard algorithms. Moreover, the Weibull PDFs are smooth by definition, removing undesirable rounding artefacts. The scale parameter k scales the abscissa of the distribution, whereas the shape parameter λ influences its shape ($\lambda < 1$ indicates that the termination rate decreases over time, i.e. an increased "infant mortality", whereas for $\lambda > 1$ the termination rate increases).

Two examples of the Weibull fits of the empirical duration distributions of sleeping are shown in Figure 4.2. Obviously, the mean durations of activities can vary substantially during the day (*cf.* Figures 4.2 and 4.3); for this reason, the $24 \cdot N$ PDFs $f_j(t)$ ($j = 1, \ldots, N$; $t = 1, \ldots, 24$) have been fitted, based on the empirical duration distributions of all events where j was started in the corresponding time interval. This introduces data scarcity issues – in particular during the night – as many of the probabilities are very small, so that in those time intervals there is an insignificant (or even nil) quantity of times the corresponding activity occurred in the TUS.

By assuming that the PDFs of durations do not change abruptly with time, this is overcome by taking as a basis to fit the PDF the durations of the corresponding activity at multiple adjacent time intervals, if both the corresponding scale and shape parameters of the Weibull fits are not significantly different (according to the two sample z tests). This procedure is repeated by including the Weibull parameters of the subsequent intervals, as long as all parameters in the set are pairwise not significantly different from each other.

The resulting merged intervals are indicated in Figure 4.3. For each activity j, the time intervals for which the durations did not differ significantly to each other are connected by a line. Furthermore, the mean durations of the started activities in the corresponding interval are colour-coded according to a logarithmic scale. The edge length of the squares is proportional to the logarithm of the number of occasions that the activity was started in that interval (see examples on top of Figure 4.3). Due to the logarithmic scale, time intervals cannot be visualised

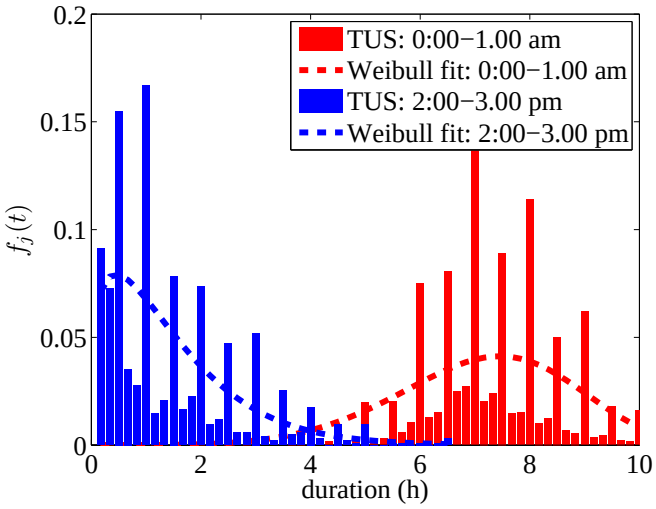


Figure 4.2: Examples of the empirical PDFs of sleeping ($j = 10$) started in two different time intervals, as well as their fitted Weibull distributions $f_j(t)$.

where the corresponding activity was started exactly once. Time intervals are marked with a cross, where the activity was not started at all.

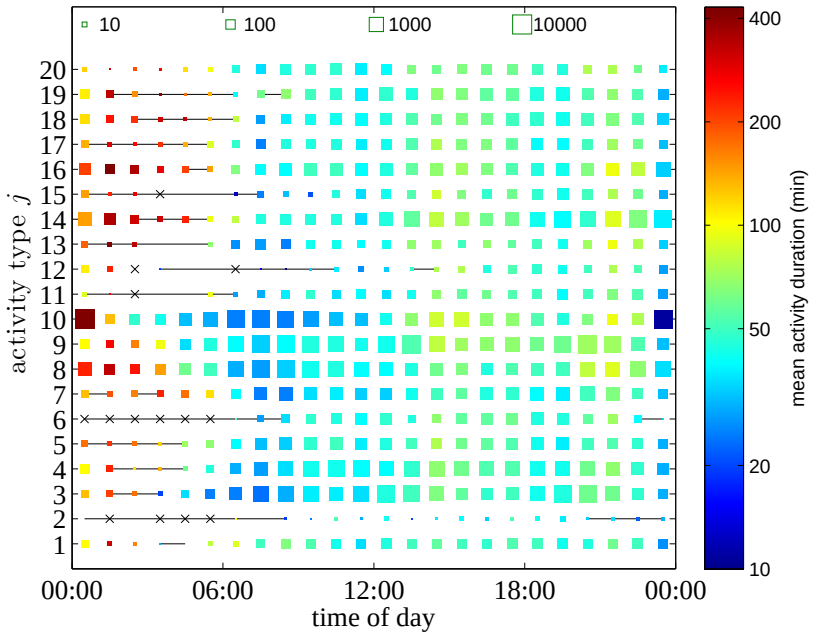


Figure 4.3: Mean durations of the activity types j (*cf.* Figure 2.8) that are started in the time interval (on the x-axis), colour-coded according to a logarithmic scale.

Individual-specific PDFs In contrast to the data scarcity problems which can arise for some activity types when the duration PDFs are fitted on an hourly basis, in some hour intervals certain activities j are started very frequently (see Figure 4.3). This allows us to investigate the statistical dependence of the corresponding PDFs on the characteristics of individuals $\boldsymbol{x} = (x_1, \ldots, x_m)$; *i.e.*, the time-dependent PDFs $f_j(\boldsymbol{x}, t)$ can also depend on $\boldsymbol{x}$.

This was done by subdividing the durations of interest into distinct subsets according to the properties of a criterion (e.g., seven subsets, each one corresponding to one weekday). Then, the mean duration and its standard error was calculated for every subset, and the value was checked for statistical significant difference to all the other subsets of this criterion according to a two sample z test. If the mean duration value is significantly different from all the other values and, the mean magnitude of the corresponding z test is the highest possible of all criteria and all their possible values, the PDF is individualised by fitting it separately for the two subsets, which result of a splitting of the population into one sub-population where the criterion takes the value and its complement. The procedure is repeated recursively with each one of the two resulting subsets, as long as no splitting is possible anymore, according to the specified methodology.

To maintain a meaningful level of statistical explanatory power, splitting into two subsets was only done if all the subsets of the criterion have an element size of $n_{\min} \geq 5$.

In this way one obtains a binary tree structure where each node corresponds to the splitting of the duration data according to a given value of a significant criterion and the branches at the end of the tree correspond to a subset which is characterised by the values that are taken at each ramification of the tree. This implies that any individual $\boldsymbol{x}$ corresponds exactly to one of the distinct subsets at the bottom of the tree, and therefore, the corresponding PDFs $f_j(\boldsymbol{x}, t)$ are also individual-dependent, regarding the characteristics that were chosen from the root of the tree to the end of the branch.

Figure 4.4 shows the subdivision tree for the pairing of activity type and time interval of (sleeping, [12 am, 1 am[) that has yielded the largest number of subsets.[5] The complete set of these durations (at the root of the tree) is subdivided at the nodes according to the value (on the right of the equal sign) of the criterion x_i the individuals have. The set of criteria $(x_i)_{i=1,\ldots,14}$ according to which the population was subdivided and their corresponding possible values are shown in the legend. For the set which branches off to the left the criterion takes the indicated value, for the one which branches off to the right it does not. The nodes at the bottom of the graph cannot be subdivided anymore with this methodology and are used to fit $f_j(\boldsymbol{x}, t)$ (in this particular example this leads to 37 PDFs being fitted). The colour of the nodes indicate the mean duration of the corresponding group.

[5] The full set of estimated model parameters is available online [85].

In Figure 4.4, the initial duration set is divided into two subsets which are distinguished by whether sleeping started on a Sunday or not. If it was started on Sunday, the most significant difference in the duration mean occurs by distinguishing if the individuals are retired or not. If the individuals are retired, no more significant disaggregation (with resulting subsets greater than $n_{\min}$) was possible anymore. The non-retired sub-population was still subsequently split according to the different possible values of employment status of the TUS (full time, unknown, part time and not in paid work).

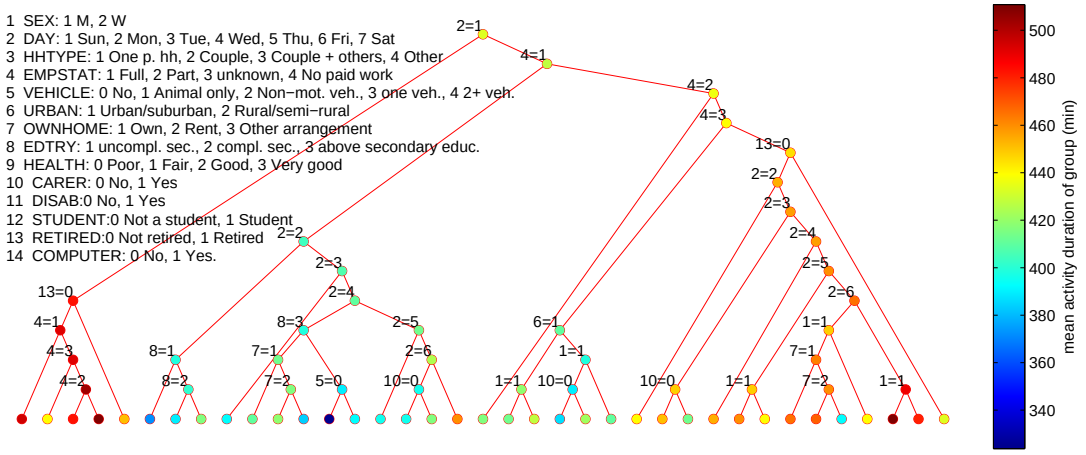


Figure 4.4: Example of the subdivision methodology (described in Section 4.2.1) of the complete duration array of sleeping that is started between midnight and 1 am, which is used to estimate the individual-specific duration PDF $f_j(\boldsymbol{x}, t)$.

Markov model simulation

In reality, the time-dependent probability of transiting from an activity j to a subsequent activity j' can be significantly different of the probability that j' occurs, irrespective of the preceding activity j [*cf.* 20, 101]. This situation corresponds to the first-order Markov property, where the next step depends on the current one only. As our main objective is to model individuals' stochastic activity choices, it is desirable to include the increased/decreased likelihood of subsequent states within the individual-dependent MNL models for the starting probabilities of Section 4.2.1. This can be achieved by adding additional terms $\beta_{j',j}x_j$ for all possible activity pairs (j, j'), where the dummy variable x_j is one when the preceding activity is equal to j and zero otherwise.

In practice however, it would not be feasible to include each one of the transition probability parameters in the MNL models, as for a set of $N = 20$ activity types in the MNL model, $(N-1)^2 = 361$ additional transition parameters would have to be estimated for each time step. This would substantially increase the

computation time and introduce once again the challenge of data scarcity. Therefore, a methodology was followed, whereby an additional transition term $\beta_{j',j}x_j$ was only added to the utility functions in Equation (4.1), if the probability that j' occurs after j in the monitored data is significantly different from the probability that j' occurs regardless of the preceding activity j in that time interval. This was done by checking the two proportion z test of the two shares. Furthermore, as an additional criterion, the term $\beta_{j',j}x_j$ was only included if the number of occurrences of the transition $j \mapsto j'$ in the TUS data is larger than 5. In this way, only 18.5 % of the whole set of possible transition parameters had to be added.[6] To reflect this the simulation algorithm should also be modified, by adding the set of dummy variables $(x_j)_{j=1,\dots,N}$ to the characteristics $\boldsymbol{x}$ describing the individual, to calculate transition-dependent MNL starting probabilities using Equation (4.2).

4.3 Simulation

In this section, the simulations are performed on 100 replicates of the specified model. The simulation period and time resolution were chosen to be the same as in the TUS, *i.e.*, 24 h starting from midnight. This means that the first time step t_1 in the simulation corresponds to the first 10 minutes after midnight, and the simulation ends after the last time step t_{end}. The set of activity types comprises the $N = 20$ partially merged types that were defined in Section 2 and the actual presence information of the survey is directly used as a simulation input, as explained in Section 4.2.1.

4.3.1 Model quality assessment

The performance of the simulations can be evaluated by comparing the residential activity patterns of the simulation to those of the TUS. For this purpose two indicators are used:

1. **Mean relative population share deviation**
   This indicator shows how adequately the model performs regarding global predictions of the population average. The indicator is based on the magnitude of the differences $D_j(t) = p_{j,\text{sim}}(t) - p_{j,\text{obs}}(t)$ between the predicted probabilities to perform a residential activity $p_{j,\text{sim}}(t)$ and the corresponding proportions that were observed in the TUS ($p_{j,\text{obs}}(t)$ as displayed in Figure 2.8).

$$D = \frac{1}{\bar{p}} \frac{1}{N \cdot t_{\text{end}}} \sum_{j=1}^{N} \sum_{t=1}^{t_{\text{end}}} |D_j(t)|. \tag{4.4}$$

[6] The estimated MNL models are available online [85].

Thus, D corresponds to the mean of the $\|\cdot\|_1$ norm of the vectors $D_j(t)$, expressed as the relative error with respect to $\bar{p} = 1/N$, the mean probability to perform one of the activity types. To level out the rounding artefacts in the monitored data, the indicator was based on the hourly averages of the residential activity profiles, *i.e.*, $t_{\text{end}} = 24$. As $\sum_{j=1}^{N} p_j(t) = 1$ for all t, D consists of two equal parts where one of them is restricted to all the positive terms $D_j(t)$ and the other one to all the negative ones (when one of the activities is overestimated by a certain amount, the sum of all the others is underestimated by the same amount; *cf.* Figure 4.6). The possible range of the value set of D is bounded between zero and one. The value of D is a measure for the global performance of the model.

2. **Percentage of correctly predicted activity time steps**
   This indicator A backs up the quality of the model's predictions at an individual level, summarised for the whole population $\mathcal{C}$. It is defined by the ratio of the total number of time steps, where the residential activity chain of the TUS $a_{\boldsymbol{x},\text{obs}}(t)$ has been correctly predicted by the one of the simulation $a_{\boldsymbol{x},\text{sim}}(t)$ whilst the different individuals $\boldsymbol{x}$ were present over the total number of time steps where the individuals were present in their residences. The value space of A is also comprised between 0 and 1, where 1 means that all residential activity chains of the population $\mathcal{C}$ are completely correctly predicted, and 0 means that no activity chains were correctly predicted. The value of this indicator where **n**o **s**leep is considered will be denoted as A_{ns}.

The first criterion is of importance, as research topics which relate to this field often depend on the average behaviour of an aggregate population. However, the distribution of the behaviour over the population may also be of interest, in which case knowledge of the variety of behaviour between the individuals is crucial. In this regard, the second criterion can be used to estimate the discriminating quality of a predictive model.

4.3.2 Generic model

In the following the simulations are based on the generic starting probabilities which were defined at the beginning of Section 4.2.1, as well as with the generic duration PDF that was defined at the beginning of Section 4.2.1, regardless of individuals' specificities. Figure 4.5 presents the probability distribution of activities of the simulated population as a function of time. The model generates a residential activity profile, where the probabilities $p_{\text{s},j}(t)$ are smoothed (with respect to the artefacts noted in the observed data) as a function of time.

The differences on an hourly basis between the simulated and the monitored data profiles $p_{j,\text{sim}} - p_{j,\text{obs}}$ are shown in Figure 4.6. For most of the time intervals

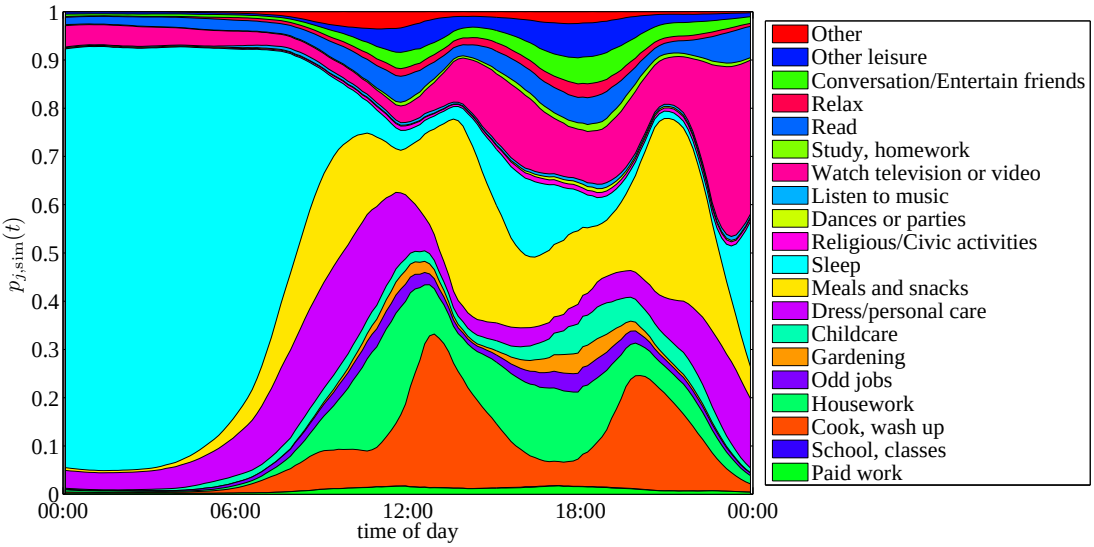


Figure 4.5: Simulated residential activity profile $p_{j,\mathrm{sim}}(t)$, using the generic starting probabilities and duration PDFs.

and activities, the differences are close to zero, meaning that the observed residential activity profiles are in general well reproduced by the simulations. The largest difference of 12.4 % occurs for sleeping in the last time interval, due to strong increases during the last 10 min of the day in the TUS data (cf. Figures 2.8 and 4.3).[7] There are also noticeable underestimations of the share of people having "meals and snacks" between 1 and 2 pm of 9.3 %, preceded by an overestimation of over 2 %. This discrepancy is due to a strong increase of the TUS population's proportion having "meals and snacks" in the hour after noon (*cf.* Figure 2.8). Apart from the other mentioned rounding artefacts, this increase is probably more likely to agree with reality, due to external time constraints. Therefore, the methodology to smooth out abrupt changes in $p_j(t)$ described in Section 4.2.1 might decrease predictive power in this case. Accordingly, the same pattern of over- and succeeding underestimation in the activity to "cook, wash up" occurs in the hour interval before and after noon, as in reality this activity often directly precedes "meals and snacks" and thus is subject to the same time constraints.

Within the last hour interval of the day, the activity "watching TV" is underestimated by the simulations (-8.6 %), due to the overestimation of sleeping, which suppresses the likelihood of performing other activities, as can also be observed in

[7] This increase has a strong weight when the starting probability of this hour interval is derived and thus the model overestimates the percentage of sleeping people. Furthermore, the number of people who are sleeping during the night is underestimated by the simulations. This is due to the fact that the simulation duration is set to 24 h, implying that the individuals who began their sleep before midnight are not accounted for in the simulation (an easily resolved artefact of this particular simulation).

the underestimation of the activities "Dress/Personal care" or "Read". The converse pattern is observable in the first time intervals of the night, where the lack of the individuals who started sleeping before midnight (because the simulation duration was set to 24 h and starting at midnight) amplifies the occurrence of other activities.

The sum of the areas of all curves above zero divided by the number of time intervals corresponds to $D/2$ (defined in point 1 in the beginning of Section 4.3.1). The values of the performance indicators (cf. Section 4.3.1) of this simulation model (Starting Generic Duration Generic, SGDG), are shown in Table 4.2.

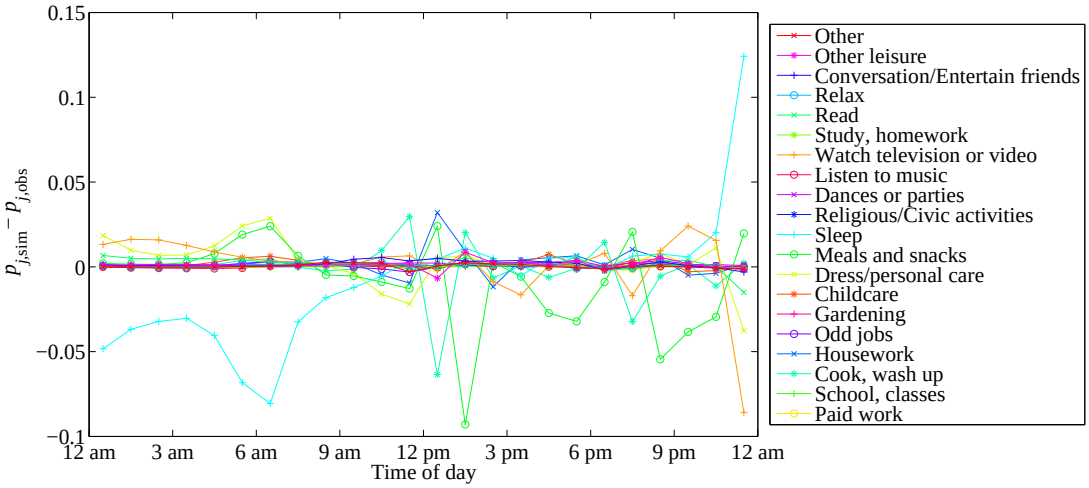


Figure 4.6: Differences between the predicted ($p_{j,\mathrm{sim}}$) and observed ($p_{j,\mathrm{obs}}$) activity profiles on an hourly aggregated basis. The lines are drawn to guide the eyes.

4.3.3 Cross-Validation

To validate the models a ten-fold cross validation was applied, by subdividing the whole TUS data set into ten equally sized distinct subsets, in which the individuals were chosen randomly (following a uniform probability distribution for the individuals to be chosen). These subsets were subsequently used as the validation sets, and the corresponding complementary set (the remaining 90 %) was used as the training set to calibrate the starting probabilities and the duration PDFs.

The final parameter sets that were determined for the training set of the whole TUS population were adopted for all cross-validation training sets, and those parameters were re-estimated for the latter.

The mean values of the indicators introduced in Section 4.3.1 of the ten cross-validation sets are shown in the corresponding columns of Table 4.2. The small relative standard errors of the mean values show that the methodology is robust in

Table 4.2: Performance comparison of the different models introduced in Section 4.2.1 (indicator values in %).†

model	whole population			cross-validation			sub-population	
	D	A	A_{ns}	D	A	A_{ns}	$D(\mathcal{C}_{\text{w}})$	$D(\mathcal{C}_{\text{nw}})$
SGDG	10.10	39.11	14.83	12.72 ± 0.34	39.09 ± 0.15	14.81 ± 0.07	16.17	14.28
S5DG	10.22	40.35	16.48	12.58 ± 0.36	40.33 ± 0.20	16.43 ± 0.09	11.22	11.14
S10DG	10.19	40.44	16.54	12.57 ± 0.35	40.38 ± 0.16	16.48 ± 0.08	11.20	11.04
S100DG	10.29	40.71	16.87	12.59 ± 0.47	40.62 ± 0.19	16.75 ± 0.11	10.54	11.83
SGDI	9.01	40.18	14.92	11.48 ± 0.46	40.86 ± 0.18	14.89 ± 0.07	15.03	13.52
S5DI	8.44	41.47	16.66	11.22 ± 0.41	42.07 ± 0.21	16.54 ± 0.08	9.09	10.08
S10DI	8.39	41.55	16.73	11.18 ± 0.40	42.12 ± 0.20	16.61 ± 0.08	9.12	9.99
S100DI	8.41	41.80	17.03	11.14 ± 0.44	42.32 ± 0.22	16.86 ± 0.08	8.77	9.91
SMDG	9.06	41.48	16.79	12.42 ± 0.38	40.42 ± 0.16	16.54 ± 0.11	9.73	10.71
SMDI	8.15	42.04	17.19	10.79 ± 0.44	41.89 ± 0.29	16.91 ± 0.10	9.54	8.83

† In the beginning of the abbreviations of the different models, the type of starting probabilities are classified. The **G**eneric **S**tarting probabilities are abbreviated by "SG". When there is a number after the "S", this specifies the value of α in percent, which was used during the backward elimination process (*cf.*, Section 4.2.1). "100" represents the model, for which all initial parameters are kept. **S**tarting probabilities modelled as a **M**arkov process (*cf.*, Section 4.2.1) are labelled by "SM" at the beginning. The last two letters indicate whether the **D**uration PDFs are modelled **G**enerically ("DG") (*cf.*, Section 4.2.1) or **I**ndividual-specifically ("DI") (*cf.*, Section 4.2.1).

terms of predictive power when applying the model in a scenario, with properties that differ to those of the calibration set.

Due to the reduced training dataset size, the chosen performance indicators are, as expected, worse in this cross-validation, except for the values of A, where the duration PDFs are individual-dependent (SGDI/S5DI/S10DI/S100DI). This means that the activity sleeping has been better predicted in the simulations of the cross-validation, than for the whole population. The reason for this still has to be investigated in detail.

4.3.4 Model performance comparison

The values of the performance indicators are shown in Table 4.2. The column "whole population" indicates that the corresponding model was calibrated and tested using the whole sample population. The column "cross-validation" shows the results of the cross-validation, which was described in Section 4.3.3. In the last column the indicator D is evaluated for two distinct sub-populations $\mathcal{C}_{\text{w}}$ of individuals who are in paid **w**ork, and $\mathcal{C}_{\text{nw}}$ of individuals who are **n**ot.

The simulations were run for different combinations of the various models for which the starting probabilities and the duration PDFs were calibrated (see Section 4.2.1 and the caption of Table 4.2).

Whole population

In Section 4.3.2 and in connection with Figure 4.6 in particular, the major deviations between the predicted and the observed population shares performing the different activities of the SGDG model were discussed. It was shown that the major contributions to D do not originate from deficiencies of the modelling approach, but rather from peculiarities of the way the TUS data was recorded (*e.g.*, that the recording period starts and ends at midnight), or from time constraints arising from some of the activities (*e.g.*, cooking or eating around noon). It is important to consider this when comparing the performance of different models, as the quality of the model expressed by D of all of the models that are presented suffer from those peculiarities. In other words, a major part of the deviations contributing to D cannot be resolved by the different approaches presented here.

The values of D indicated in Table 4.2 show that the models where the duration PDFs are generic, the starting probabilities have been individualised but the transition probabilities are not included (S5DG/S10DG/S100DG) perform slightly worse than the model where the starting probabilities are also modelled generically (SGDG). This means that an individualised description of the behaviour of starting activities worsens the predicted shares, when the activity durations are generic; but this approach also offers greater flexibility in its application. Comparing the performance of SGDG with SMDG, shows that the population shares are better predicted when transition probabilities are considered in addition to individual characteristics. Given the individual-specific $f_j(\boldsymbol{x}, t)$ and the individual-specific $p_{\mathrm{s},j}(\boldsymbol{x}, t)$, additional consideration of the transition probabilities in the starting probabilities improves the prediction of the correct shares (compare SGDI with S5GI/S10DI/S100DI/SMDI). This suggests that there are synergetic effects regarding the improvement of the prediction of the population shares when both, $p_{\mathrm{s},j}(t)$ and $f_j(t)$ are individualised.

The SGDG model has the worst performance regarding predictivity at an individual level, expressed by A and A_{ns}. The individual-specific modelling of the starting probabilities and of the duration PDFs improves performance compared with the generic model. For both cases, when one individual-specific part is used in combination with the generic counterpart (S5DG/S10DG/S100DG or SGDI), or when both individual-specific quantities are used together (S5GI/ S10DI/S100DI), the inclusion of the modelling of transition probabilities from one activity to another also improves the results of these two performance indicators (compare the results of S10DG with SMDG and S10DI with SMDI).

Comparing the performance of the models with the full set of parameters $\{\beta_{j,k}\}_{j=1,\ldots,N;\, k=1,\ldots,M}$ for the starting probabilities with those where the set was restricted by the backward elimination process (S100DG with S10DG/S5DG or S100DI with S10DI/S5DI), shows that backward elimination does not substantially decrease model predictivity.

Sub-population

For the whole population, the SGDG model has a value of D of 10.10 %. When one considers the corresponding values of 16.17 % and 14.28 % for the two sub-populations $\mathcal{C}_{\mathrm{w}}$ and $\mathcal{C}_{\mathrm{nw}}$, the deterioration of performance is more serious than the comparison of values of D suggests, recalling the discussion in Section 4.3.4.

The comparison of the values of D for the whole population yielded an improvement of about 19 % between the generic model (SGDG) and the best performing one (SMDI). When D is evaluated for different sub-populations with given characteristics (in the case of Table 4.2 distinguishing between people who are in paid work or not), there are more significant improvements of 41 % for $\mathcal{C}_{\mathrm{w}}$ and 38 % for $\mathcal{C}_{\mathrm{nw}}$.

The values of D that are achieved for the two sub-populations for the SMDI model, for instance, are better than those that SGDG yields for the whole population. This suggests that the individual-specific models are more appropriate for scenarios where the characteristics of the target population do not correspond exactly to those of the TUS sample population.

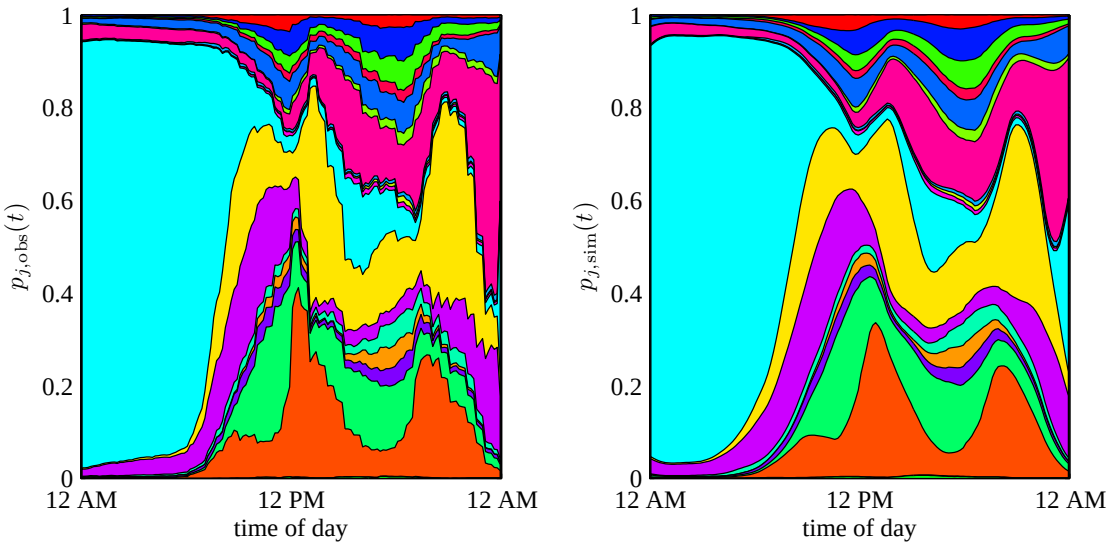
The improvements of D for the whole population are more significant for the individual-dependent description of the duration PDFs $f_j(\boldsymbol{x}, t)$, than for the starting probabilities $p_{\mathrm{s},j}(\boldsymbol{x}, t)$. In contrast, the improvements of D (deviations between the predicted and the observed shares) for the two different sub-populations are more important for the individual-dependent description of $p_{\mathrm{s},j}(\boldsymbol{x}, t)$, than for that of $f_j(\boldsymbol{x}, t)$. However, predictivity also improves for the individual-dependent description of either $f_j(\boldsymbol{x}, t)$ or $p_{\mathrm{s},j}(\boldsymbol{x}, t)$ or of both together.

The inclusion of the Markov property in the starting probabilities in $p_{\mathrm{s},j}(t)$ only improves the performance for the non-working sub-population $\mathcal{C}_{\mathrm{nw}}$, whereas it worsens performance for the working sub-population $\mathcal{C}_{\mathrm{w}}$. However, for the whole population the inclusion of D improves predictivity, as mentioned in Section 4.3.4.

The simulation results of D show that there is considerable additional benefit in modelling behaviour individually, as long as one is not only interested in the aggregate performance of the whole population or in proportions which differ from the TUS sample population. In these cases the assumption that the individuals' behaviours can be approximated by average behaviours is too crude.

To illustrate how individual-specific behaviour is represented, the observed and the simulated (S10DI) activity profiles of the non-working sub-population are shown in Figures 4.7a and 4.7b, respectively (the colour-coding of the different activities is the same as in Figure 4.5). In the SGDG model, the starting probabilities are identical for the whole population: regarding the residential activity profile of a sub-population, the only difference lies in the time-dependent proportion to be at home. Thus, the maximal hourly difference between the residential activity profile of the whole population (*cf.* Figure 4.5) and the sub-population is less than 0.40 %.

In the non-working sub-population, many observed activity probabilities $p_j(t)$ (Figure 4.7a) are similar to those of the whole population (Figure 2.8). However, with respect to the percentage of individuals doing paid work, the observed daily average is 0.12 %, as compared with a predicted percentage using the SGDG model of 0.87 %, which corresponds to an overestimation of a factor of more than 7. The corresponding value of the S10DI model is 0.23 %. This overestimation of 88.5 % can be explained, by the removal of insignificant parameters in the corresponding models of starting probabilities (*cf.* Section 4.2.1), indicating that the specificities of the behaviour of these individuals cannot be completely captured in the model. Furthermore, the mean duration of paid work in the non-working sub-population is shorter than in the whole population. These differences also cannot be fully captured with the algorithm described in Section 4.2.1. However, in comparison to the SGDG model, this also shows that individual-specific behaviour is much better captured by this non-generic model.



(a) Observed in the TUS.

(b) Simulated by the S10DI model.

Figure 4.7: Residential activity profile of the non-working sub-population $\mathcal{C}_{\mathrm{nw}}$.

Integration into Building Simulations

The models presented in this chapter can be integrated into any dynamic building simulation tool. We will describe here the steps that are to be executed, to derive the activity chains whilst an occupant $\boldsymbol{x}$ is at home. We assume that the information of the residential presence/absence chains is provided by a preprocess for the occupants in the simulation.

1. The occupancy status (absence/presence) of the individual $\boldsymbol{x}$ at time t_i as well as the time of the next change t_{i+1} is retrieved.

2. - **Case 1:**
   The occupant is absent. The activity until the end of the occupancy state duration is set to null. The time is incremented to the next occupancy state $t_i \mapsto t_{i+1}$.
   - **Case 2:**
   The occupant is present. An activity j is chosen based on the corresponding starting probability at that time $p_{\mathrm{s},j}(\boldsymbol{x}, t)$. A time increment Δt_j is chosen according to the PDF at that time $f_j(\boldsymbol{x}, t)$ (limited by $t_{i+1} - t_i$ if necessary, corresponding to a censoring of the activity by a departure of the occupant). The time is incremented to $t_i + \Delta t_j$.

3. As long as the time has not reached the maximum simulation time, the simulation returns to step 1.

4.4 Discussion

A bottom-up residential activity model has been formulated to support predictions of individual members of statistically significant demographic sub-populations whilst reproducing well the time-dependent activity probabilities of the entire population. It has been shown that for this latter, a model of the aggregate population is sufficient; but this appears not to be so in the former case as behaviour can vary substantially with individual specificity. To address this, several model refinements have been tested. But experience teaches us that human behaviour can be very diverse. Reflecting this, it is unsurprising that certain behaviours have not been included in this model:

- In Section 4.3.4 it was shown that the model performs worse for sub-population $\mathcal{C}_{\mathrm{w}}$ when modelling transitions to calculate the starting probabilities $p_{\mathrm{s},j}(t)$; whereas the corresponding predictions improve for $\mathcal{C}_{\mathrm{nw}}$ (compare SMDI to S10DI). This is likely due to the fact that transition probabilities are estimated using data for the whole population; whereas in reality these may also depend on the specific characteristics of a sub-population. This drawback could be overcome by estimating sub-population-dependent transition probabilities, as in the work of Fischer and Sullivan [101], where transition probabilities are calculated separately for workdays and weekends. But this would increase the number of transition parameters in proportion to the number of different characteristics to distinguish which may actually degrade the models' predictive power due to data scarcity issues. Therefore, this would require preliminary statistical tests to investigate for which characteristics there are significant differences in the values of the

transition probabilities one is interested in. Another possible weakness of this model is that the duration PDFs of Section 4.2.1 were not deduced depending on the previous activity.

- There is also scope for improving model parsimony with respect to the modelling of sub-populations. For example, it may be appropriate to replace weekdays with the simple distinction between workday and weekend day. However, on a Sunday evening many activity probabilities p_j resemble those of a standard workday, whereas on Friday evening activities resemble better those of a Saturday evening. More parsimony may result from using the models with the full set of characteristics presented in Section 4.2.1 as preliminary estimations, and assembling cases in the same dummy variable when they are not significantly different. In this way one could construct dummy variables containing a group of two or more different values of a given characteristic, in which all elements are pairwise not significantly different from each other.

- Usually, activity modelling is conducted for a much smaller set of activity types N [*cf.* 20, 101]. This would allow for more sophisticated calibration methodologies and testing procedures and improve predictivity, when the set of different activities is smaller. A large set of different N was chosen in the presented models, to prove relevance/applicability for wider interest group of research fields.

- Although this model has been formulated to predict residential activities based on pre-processed residential presence, the approach can be readily applied to predict activities in different types of place or even without restricting it to any type of place by omitting the presence pre-process.

  However, the use of a presence model as a pre-process does in principle increase the versatility of the approach. For instance, one can argue that the last activity before leaving or before going to bed is subject to more stringent constraints than those presented here. One could model this by placing another intermediate process between the presence model and the activity model. This intermediate process would partially assign activities during presence, to be treated like absences by the activity post-process. Sleeping and its preceding and succeeding activities could be modelled in such a way. Another example includes the modelling of activities which are correlated between household members. For instance one could model the increased probability of simultaneous meals, by first running the activity model for the household member who is cooking, and afterwards impose the same time of eating to the remaining household members that are present; in this way removing the inherent limitation of Markov chain models that members behave independently of each other.

4.5 Conclusions

In this chapter, an approach was presented to the modelling of residential activities based on time-dependent probabilities to start activities and their corresponding duration distributions. The purpose of this model is to support a more accurate representation of occupants' energy-related behaviours in building and urban energy simulation programs. An initial model was fitted to data relating to the entire aggregated population of respondents to the time use survey questionnaire. Successive refinements were then implemented by fitting the models' parameters to demographic sub-populations of the survey dataset, to support better more individualistic predictions – for example to accommodate predictions for specific household demographies or to account for the day of the week. Transitions between successive activities were also modelled based on the first-order inhomogeneous Markov property. In developing these refinements, considerable effort has been put into the design of the calibration methodologies, to avoid including characteristics which are statistically insignificant. Results from validation tests show that these refinements do indeed improve model predictability, when influences are included that capture activity transitions and a dependence on individual characteristics in the starting probabilities and the duration distributions.

The set of variables which potentially influence residential activities was restricted in the presented models (*cf.* Section 4.2.1 and Figure 4.4). But there are some characteristics recorded in the TUS database that may have a considerable influence on residential occupants' activities, such as parental information, the employment status of the spouse/partner or household income. Secondary activities are also recorded in the database. However, this information has so far been neglected to keep the approach as general as possible and to avoid compounding the challenges of data scarcity. Nevertheless, a future variant of this model might usefully consider concurrent activities which have energy implications or simply to improve the predictability of primary activities. Other potential refinements to the modelling approach include:

- sub-population dependent activity transition probabilities,
- the modelling of correlations of activities between members of the same household, and
- an increase in model parsimony by reassembling dummy variables with insignificant differences.

The proposed approach, as well as the calibration and validation methodologies employed are directly applicable to other environments, such as workplaces and all other types of places that are defined in the calibration database. Future work should also concentrate on the use of the predictions of this model as an input

to an algorithm simulating the use of electrical/water appliances, supported by a reliable prediction of appliance ownership.

Chapter 5

Appliance ownership

In this chapter, two approaches are presented to predict the probability of the ownership of M types of electrical appliances, depending on household characteristics, such as income or household size. The models are based on logistic regression and calibrated with data of a Swiss appliance ownership survey of 2005. The specification of the models enables the choice of a set of predictors which assures parsimony together with a high degree of significant parameters using a backward elimination technique. To treat difficulties due to multicollinearity issues, an empirical approach is developed, based on principal components, which allows to bypass the potentially very time-consuming backward elimination procedure. The validation and comparison of the two approaches is done by performing Davidson MacKinnon J tests and by analysing the accuracy of their predictions for sub-populations of the calibration dataset, revealing a similar level of performance.

5.1 Introduction

As we are striving towards a more sustainable energy use, decentralised electricity generation from renewable sources becomes more and more important. To better match the load profiles of neighbourhoods with (small-scale) power generation (and for a better design/sizing of the latter) or for structural changes like those of the smart grid concept [*e.g.*, 56–58], we need to model more accurately the stochastic nature of electrical appliance use in single buildings [49, 59]. In particular, one has to account for the variations over time (of day) and of behaviour (use [60], investments, *etc.*) between individuals and households (*cf.* [61] for a comprehensive modelling study of this in office buildings). It was shown that the ownership of electrical appliances in households is significantly dependent on various drivers, such as personal/dwelling/household characteristics [55, 117, 118]. Obviously, the stock of electrical appliances in single households strongly influences their residential electricity demand [*e.g.* 119, 120]. To model residential

electricity demand in future scenarios, the accurate bottom-up modelling of appliance ownership is a crucial requisite [121].

With regard to the prediction of the probabilities of the ownership of electrical appliances, a bottom-up model is needed that is calibrated on the significant characteristics of households, to faithfully encapsulate the full range of specificities. Such an approach also lends itself well to the modelling of future scenarios to explore responses to changes in the behaviour how appliances are used as well as to changes to the population's demographic characteristics. Last but not least, an important requirement for a useful model is that it should be easy to recalibrate it with other datasets, in order to investigate differences between countries or temporal changes.

5.1.1 Previous research work

A large literature overview over the wider field of consumer behaviour and the use of sustainable energy is provided by Madlener and Harmsen-van Hout [122]. Appliance ownership is also of interest for other fields, such as waste management [123, 124], or even social science [125]. In economics, one of the main interests regarding appliance ownership exists in predicting replacement purchases [126]. However, the explanatory variables in these models usually do not include personal/household characteristics [127].

McNeil and Letschert present an approach based on a modified version of logistic regression, predicting diffusion rates of electrical appliances in different countries in dependence of national characteristics, and calibrated with data of a wide international range [128]. However, due to the need to harmonise the data of the different surveys, these models are only based on a small set of different predictors, which do not include household specificities. Weber and Perrels use discrete choice models to predict ownership of electrical appliances in Western Germany [55]. However, the calibration methodology of the models is not presented in detail, and furthermore, a rigorous validation is missing. Leahy and Lyons use logistic regression models to predict electrical appliance ownership in Ireland, where stepwise deletion of variables was applied to remove variables that are not significant [120]. However, a precise description of the backward elimination methodology is not presented. Matsukawa and Ito use a multinomial logit model to predict the ownership of electric room air conditioners in Japan, depending on income, other household characteristics and the consumption of composite goods [129].

To assure robustness of a model, it is important to reduce the set of predictive parameters to those that are statistically significant. Ndiaye and Gabriel apply principal component analysis, in order to predict electricity needs in residential dwellings as a function of a reduced set of significant predictors [53]. However, the establishment of a theoretical framework that describes the use of principal components in the context of logistic regression is to our knowledge rather in

an early stage. The most comprehensive methodology was discussed by Aguilera et al., where a methodology is established how principal components can be used in logistic regression, in order to manage problems of multicollinearity between predictors in the model [130]. Several strategies are presented how the dimensionality in the principal component space can be reduced to eliminate multicollinearity. However, this methodology is based on the definition and the selection of principal components in the linear case [*cf.*, *e.g.*, 131], and thus this is not applicable to immediately estimate a model, where all insignificant principal components can be eliminated at once, but a stepwise procedure has to be implemented.

Camminatiello and Lucadamo extended the latter methodology to multinomial logit models, but, the model validation is based on a small case study, and a comprehensive quantitative evaluation is not presented, making it very difficult to estimate the viability for the present study [132]. Schaefer use principal components to evolve an alternative estimator to the one using maximum likelihood, when there is multicollinearity in predictor data [133].

Barker and Brown evaluate principal components linear regression besides ridge logistic regression [134]. However, the criterion to decide which principal components to eliminate is adopted from the method of Kaiser [135], which is based on factor analysis. Furthermore, the error propagation between the estimates of the principal components and the original scale is treated by means of bootstrapping and not analytically. Marx presents a methodology for maximum likelihood estimation when the information matrix for the maximum likelihood estimation of a generalized linear model is ill-conditioned [136].

From the above review of the state of the art in predicting the ownership of electrical appliances, we can conclude that a well-documented and detailed approach, which is readily applicable for calibration with different datasets (for instance of different countries or years) and where explanatory variables with insignificant influence are eliminated, is missing. The needed approach furthermore, has to be compatible to encapsulate household-related explanatory variables.

5.1.2 Summary

A detailed methodology is presented to predict the probabilities with which electrical appliances are present in households, depending on their characteristics. In Section 5.2, the calibration data is detailed and two approaches are developed how the models can be calibrated and specified, removing insignificant parameters due to multicollinearity. They are based on a backward elimination procedure of logistic regression models, which are, firstly, based on the original predictors and, secondly, on principal components. In Section 5.3, the models are validated, applied and their predictive power is compared to each other using Davidson MacKinnon J tests. In Section 5.4, the approaches are discussed and

further possible refinements are identified, and finally the chapter is concluded in Section 5.5.

5.2 Methodology

Methodologies are presented in this chapter to predict the ownership of electrical appliances i (*cf.* Table 5.1) in households $\widetilde{\boldsymbol{x}}$, which are characterised by a set of characteristics $\widetilde{\boldsymbol{x}} = (x_1, \ldots, x_M)$ (*cf.* Table 5.2). The choice of the set of predictors $\widetilde{\boldsymbol{x}}$ must be rigorously performed, bearing in mind that the modelling complexity increases with the amount of predictors, but not necessarily the accuracy. In this study, we propose to use only dummy variables, as this facilitates interpretation of the estimated model parameters, as well as the implementation of reliable automatised optimisation algorithms.

5.2.1 Appliance ownership survey

The models are calibrated using data of an electrical appliance ownership survey [137] conducted in 2005 by the Association of Swiss Electricity Companies [138]. To reduce the complexity of the estimated models, the available set of predictor variables was simplified to the dummy variables in Tables 5.1 and 5.2, which are defined to be one if the corresponding statement is true (or respectively if the appliance is present) and zero otherwise[1].

To define the dummy variables in Table 5.2, some of the original variables recorded in the survey have been transformed. In the set of the latter, for every household member the gender and the age class ($1 \mathrel{\widehat{=}} < 21$, $2 \mathrel{\widehat{=}} 21 - 65$, $3 \mathrel{\widehat{=}} > 65$) are recorded. As there are numerous possible combinations of the number of household members, their gender and age class, the inclusion of a dummy variable for each combination would not be feasible within the framework of the modelling approaches presented later. Therefore secondary dummy variables were introduced, which (likely) correspond to given types of household forms, namely:

- Families are indicated by fam, corresponding to the case, where there are at least three household members containing a man and a woman (both not minor) and all the rest are in the first age class.

- Flat shares are characterised by shr, corresponding to the case where there are at least three household members, which are all in the medium age class.

[1] The dummy variable apSf was defined to be 1, if the dwelling is a single family home, and 0 if it is an apartment. The variable rur defines, whether the place of residence of the household is situated in a municipality with less than 10000 inhabitants and not within an urban agglomeration.

- ag2 corresponds to an average household age of at least 2 and less than 3.
- ag3 is one if every household member is in the third age class.
- olPr means that additionally, the household contains one man and one woman to specify elderly couples.
- wom specifies households with at least one woman.

The value of the monthly household income class was also specified in 6 classes (defined by the limits of 3000, 4500, 6000, 9000, 15000 CHF) and the 7th choice of "not specified". The mean rental prices of the dwelling of the corresponding room number [139] was deducted from the mean household income mean value of the corresponding income class (3000 CHF and 16000 CHF for the lowest and highest class, respectively), to derive the average disposable income per household member. The resulting distribution was subdivided in 4 classes using the 5 %, 50 % and 95 % quantiles. The topmost income class was fixed (reference case, together with the non-specified households), resulting in the remaining 3 classes inc1, inc2 and inc3.

Regarding the whole set of predictors, it is evident that there are many variables, which are mutually (anti-)correlated. This needs to be considered, when developing a model to predict the ownership of appliances, which is described in the following section.

5.2.2 Logistic regression

As the main interest in this research topic is to predict the equipment of electrical appliances i in households with characteristics $\widetilde{\boldsymbol{x}}$, a methodology will be presented, which allows to estimate the probabilities $p_i(\widetilde{\boldsymbol{x}})$ that an appliance i is present in the inventory of a household $\widetilde{\boldsymbol{x}}$, using logistic regression:

$$p_i(\widetilde{\boldsymbol{x}}) = \frac{1}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_M\right)} = \frac{1}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^\intercal \widetilde{\boldsymbol{x}}\right)}. \tag{5.1}$$

Here, β_0 is the alternative-specific constant (ASC), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\intercal$ represents the vector of parameter estimates and $\widetilde{\boldsymbol{x}}$ represents the vector of predictors characterising the household $\widetilde{\boldsymbol{x}} = (x_1, \ldots, x_M)^\intercal$. In this study, the dummy variables of Table 5.2 are used in the models[2], which were estimated using the software MATLAB® [140], unless differently specified.

[2]The probabilities of appliance ownership $p_i(\widetilde{\boldsymbol{x}})$ might also depend on appliance attributes, as well as on the ownership of other appliances, which will be neglected.

Table 5.1: Description of the binary choice variables of appliances i that are estimated in the models. The diffusion rates over the total population, as well as sub-populations, defined by some of the predictors (of Table 5.2), are shown in Table 5.7.

abbreviation i	**description**
fr2+	at least 2 fridges
cOv	cooker with oven
cnOv	cooker without oven
ovSp	seperate oven
miwa	microwave oven
dwsh	dishwasher
cafe	coffee machine
tv1+	at least one TV
tvRe	TV receiver
vid	video player
dvd	dvd player
cons	game console
hifi	hifi system
pc1+	at least one PC
pr1+	at least one printer
wlTe	wireless telephone
st1+	at least one electric heater
efBl	at least one energy-efficient bulb
topF	top-opening freezer
frz	freezer
coWa	collectively used washing machine
prWa	private washing machine
tmbl	tumbler
boil	electric water boiler
radH	radiant heater
aqua	aquarium
sol	solarium
pool	swimming pool
wbed	waterbed
flHt	floor heating

Table 5.2: Description of the binary predictors x_j that are considered in the models. The number in parentheses indicates the observed percentage that the corresponding variable is equal to 1.

variable name x_j	**description**
apSf	apartment or single family home (30.7)
fam	family (22.6)
shr	flat share (3.4)
own	owner (42.3)
wom	women in household (78.0)
rur	urban or rural (13.8)
p3+	3 or 4 household members (26.8)
p5+	more than 4 household members (0.8)
r3+	more than 2 rooms (87.1)
ag2	medium average age (70.0)
ag3	high average age (26.5)
kid	minors in household (29.1)
olPr	elderly couple (35.8)
inc1	low household income (12.2)
inc2	medium household income (42.4)
inc3	high household income (40.6)

Preliminary logistic regression models

In general, the set of dummy variables $(x_1, \ldots, x_M)$ is too extensive to find a meaningful solution in the parameter space, due to multicollinearity or even linear dependence between different (sets of) dummy variables (indicated by extreme standard deviations of the involved parameters due to model instability). Another reason for large standard deviations is that the corresponding dummy variable has no significant influence on the binary choice. In this case, the parameter space was reduced by removing one of the involved dummy variables, and afterwards re-estimating the resulting model. This procedure was recursively repeated, until all standard deviations (disregarding the one of the ASC β_0) were smaller than 20.[3] The results of these estimations are shown in Table 5.3. All estimated parameters that are presented in this chapter are printed in boldface if they are significantly different from 0 at the 5 % confidence level, according to their t test.

The parameters that were excluded due to anomalous standard deviations in the logistic regression model of the corresponding appliance are marked with a "/". These preliminary models are not very trustworthy as they depend on many parameters β_j that are not significantly different from zero corresponding to the

[3]The value of 20 has twice the maximum magnitude of the estimated parameters of the constrained optimisation problem, which thus prevents from eliminating significant parameters.

p value of their t test. Therefore, a backward elimination technique was applied to derive two other models, which will be presented below.

5.2.3 Backward elimination models

The models that are shown in Table 5.3 include influences of many parameters which do not result in a significant improvement of the predictive power. However, as in general, there is correlation between different predictors, the estimated parameters β_j are correlated as well. Therefore, one can not immediately obtain a model depending only on significant parameters by excluding all non-significant parameters in the models shown in Table 5.3 at once. In this way, the model would have in general new insignificant parameters, and furthermore, it is likely to exclude parameters which would be significant in a simpler model, not depending on so many insignificant terms.

Table 5.3: Results of the preliminary model logistic regression

		predictors															
i	ASC	apSf	fam	shr	own	wom	rur	p3+	p5+	r3+	ag2	ag3	kid	olPr	inc1	inc2	inc3
fr2+	**-4.85**	**1.22**	-0.59	**1.24**	-0.56	0.83	-0.15	-0.08	0.62	0.74	-0.10	-0.12	0.98	0.12	-0.19	-0.46	-0.44
	0.90	**0.30**	0.39	**0.60**	0.30	0.46	0.29	0.38	0.78	0.55	0.55	0.63	0.50	0.37	0.57	0.54	0.52
cOv	**1.30**	0.02	0.10	0.66	**-1.21**	-0.36	0.05	-0.40	0.31	-0.13	-0.25	-0.10	-0.18	-0.07	0.59	**0.63**	0.31
	0.55	0.18	0.32	0.46	**0.17**	0.23	0.18	0.27	0.76	0.23	0.43	0.46	0.37	0.22	0.35	**0.32**	0.30
cnOv	**-1.58**	-0.02	-0.14	-0.12	**1.49**	0.42	0.11	0.28	-0.90	0.23	0.05	0.04	0.63	0.24	**-0.88**	**-0.92**	-0.45
	0.57	0.18	0.33	0.47	**0.17**	0.26	0.19	0.28	0.87	0.26	0.44	0.47	0.38	0.23	**0.37**	**0.34**	0.32
ovSp	**-1.67**	0.00	-0.20	-0.48	**1.46**	0.45	0.17	0.46	-0.12	0.47	-0.03	0.01	0.54	0.28	**-1.08**	**-1.11**	**-0.67**
	0.58	0.18	0.33	0.47	**0.18**	0.26	0.19	0.28	0.78	0.28	0.44	0.47	0.39	0.24	**0.37**	**0.34**	**0.32**
miwa	**-2.71**	0.06	-0.52	0.40	0.12	**0.62**	0.14	0.27	0.16	**0.70**	**1.02**	0.70	0.36	-0.25	0.48	0.38	0.45
	0.55	0.17	0.32	0.45	0.16	**0.22**	0.18	0.26	0.70	**0.22**	**0.42**	0.45	0.37	0.21	0.34	0.32	0.30
dwsh	**-1.58**	0.03	0.27	**1.34**	**1.16**	0.09	-0.29	-0.28	0.47	**1.13**	0.85	0.29	**1.56**	**0.84**	**-0.90**	**-0.83**	-0.42
	0.61	0.22	0.42	**0.54**	**0.20**	0.23	0.21	0.34	1.12	**0.22**	0.47	0.50	**0.46**	**0.22**	**0.38**	**0.34**	0.33
cafe	**-1.97**	0.19	0.36	0.42	0.21	0.15	0.13	0.06	**-1.60**	**1.30**	0.48	0.28	-0.06	0.40	0.22	0.26	0.39
	0.54	0.19	0.33	0.48	0.17	0.22	0.19	0.28	**0.74**	**0.21**	0.41	0.44	0.37	0.21	0.34	0.31	0.30
tv1+	-0.05	-0.11	-0.43	0.30	0.28	0.58	0.20	0.31	/	0.60	**1.38**	**1.95**	-0.12	0.11	0.83	0.62	0.29
	0.86	0.40	0.65	1.18	0.37	0.39	0.40	0.50		0.34	**0.62**	**0.70**	0.68	0.43	0.64	0.53	0.48
tvRe	**-3.64**	0.31	-0.68	0.37	-0.12	**0.88**	**0.46**	0.04	1.27	0.25	0.21	-0.20	0.35	-0.30	0.21	0.22	0.40
	0.81	0.26	0.38	0.55	0.25	**0.35**	**0.23**	0.36	0.75	0.35	0.57	0.63	0.47	0.30	0.56	0.52	0.50
vid	**-1.45**	0.07	0.04	0.07	0.03	**0.81**	0.23	-0.10	-0.75	**0.68**	0.62	0.05	0.51	-0.13	-0.11	-0.23	0.05
	0.54	0.18	0.35	0.45	0.17	**0.22**	0.19	0.28	0.70	**0.21**	0.42	0.44	0.39	0.21	0.34	0.31	0.29
dvd	-0.05	0.01	0.03	0.17	-0.29	**0.92**	-0.18	0.23	0.16	0.19	-0.11	**-1.62**	**0.79**	-0.17	**-0.75**	**-1.06**	**-0.65**
	0.55	0.19	0.33	0.45	0.18	**0.24**	0.19	0.27	0.74	0.23	0.43	**0.47**	**0.38**	0.22	**0.35**	**0.32**	**0.30**
cons	**-3.28**	-0.41	-0.08	-0.78	0.44	**2.52**	0.27	0.26	0.15	0.61	-0.79	**-2.25**	0.75	**-1.31**	-0.33	-0.39	0.01
	0.88	0.26	0.33	0.62	0.25	**0.57**	0.24	0.30	0.75	0.49	0.46	**0.68**	0.44	**0.35**	0.59	0.55	0.53
hifi	1.22	-0.03	-0.24	0.06	0.12	0.35	-0.33	-0.20	/	0.37	0.39	-0.77	**1.27**	0.04	**-1.16**	**-1.20**	-0.45
	0.74	0.22	0.52	0.53	0.20	0.25	0.21	0.36		0.22	0.59	0.61	**0.55**	0.25	**0.44**	**0.42**	0.41
pc1+	0.11	0.03	-0.44	0.79	0.14	**0.52**	0.16	0.40	/	0.30	1.19	-0.77	**2.47**	0.46	**-1.29**	**-1.56**	-0.42
	0.90	0.25	0.80	0.69	0.22	**0.26**	0.26	0.45		0.24	0.78	0.80	**0.79**	0.26	**0.45**	**0.41**	0.40
pr1+	0.47	0.19	0.49	0.76	**0.40**	**0.63**	0.18	-0.15	0.35	**0.55**	-0.07	**-1.71**	0.90	0.22	**-1.39**	**-1.32**	-0.40
	0.69	0.22	0.46	0.57	**0.20**	**0.24**	0.23	0.36	1.12	**0.23**	0.56	**0.59**	0.50	0.24	**0.40**	**0.37**	0.36
wlTe	1.03	0.05	0.44	-0.03	0.15	0.22	-0.07	0.22	-0.22	0.14	-0.28	**-1.54**	0.02	0.42	-0.07	0.19	0.20
	0.68	0.22	0.43	0.55	0.20	0.24	0.22	0.34	0.86	0.22	0.57	**0.60**	0.47	0.24	0.38	0.34	0.33
st1+	**-3.33**	**1.24**	-0.63	-0.48	-0.08	0.24	0.04	-0.01	-1.34	0.01	0.30	0.92	-0.01	-0.32	0.40	0.29	0.52
	0.63	**0.20**	0.35	0.49	0.19	0.24	0.19	0.30	1.11	0.23	0.48	0.51	0.39	0.23	0.40	0.37	0.35
efBl	**-1.79**	**0.58**	0.03	0.37	0.27	0.12	0.18	0.19	0.53	0.28	-0.37	-0.45	-0.02	0.32	0.10	0.31	0.41
	0.54	**0.17**	0.31	0.44	0.16	0.23	0.18	0.26	0.70	0.22	0.40	0.43	0.36	0.21	0.36	0.33	0.32
topF	**-4.77**	**0.54**	-0.44	-0.14	**0.50**	**0.90**	0.42	0.00	1.07	0.34	0.14	0.35	0.56	0.19	0.50	0.56	0.18
	0.87	**0.23**	0.38	0.61	**0.24**	**0.41**	0.23	0.34	0.75	0.43	0.57	0.62	0.47	0.32	0.59	0.56	0.56
frz	**-3.93**	**0.68**	0.18	**2.01**	**0.65**	-0.22	0.22	-0.15	-0.76	**1.17**	0.67	0.80	**1.20**	**0.77**	0.09	-0.02	0.23
	0.61	**0.18**	0.33	**0.49**	**0.17**	0.26	0.19	0.28	0.73	**0.30**	0.42	0.46	**0.38**	**0.24**	0.39	0.36	0.35
coWa	**4.17**	**-3.09**	-0.77	-0.05	**-1.51**	-0.44	-0.20	0.28	-0.51	-0.48	0.92	1.10	**1.11**	0.18	-0.09	-0.03	-0.33
	0.71	**0.28**	0.48	0.61	**0.18**	0.26	0.24	0.38	1.02	0.25	0.57	0.60	**0.54**	0.25	0.43	0.38	0.36
prWa	**-4.49**	**2.76**	0.51	0.55	**1.47**	0.18	-0.04	-0.39	0.15	**0.74**	-0.44	-0.65	-0.40	0.08	0.13	0.15	0.44
	0.70	**0.26**	0.45	0.59	**0.18**	0.26	0.24	0.36	0.97	**0.26**	0.54	0.57	0.51	0.25	0.43	0.39	0.37
tmbl	**1.53**	**-0.98**	0.17	**1.15**	**0.42**	**-0.49**	-0.28	-0.04	0.53	0.29	0.69	0.52	**0.95**	**0.95**	**-1.37**	**-1.42**	**-1.09**
	0.55	**0.18**	0.32	**0.45**	**0.17**	**0.22**	0.18	0.27	0.75	0.20	0.40	0.43	**0.36**	**0.21**	**0.38**	**0.35**	**0.34**
boil	**-2.56**	**1.21**	0.23	0.25	0.18	**-0.49**	**0.49**	-0.03	1.02	0.16	0.30	-0.04	0.22	-0.07	0.05	0.38	0.27
	0.56	**0.18**	0.33	0.46	0.17	**0.23**	**0.18**	0.27	0.78	0.22	0.42	0.45	0.38	0.23	0.37	0.34	0.32
radH	**-4.52**	**1.15**	-0.54	-1.00	-0.03	-0.62	-0.03	0.65	/	0.71	0.07	0.40	-0.09	0.44	-0.14	-0.09	-0.07
	1.03	**0.32**	0.54	0.90	0.32	0.45	0.32	0.50		0.50	0.75	0.79	0.63	0.43	0.63	0.58	0.56
aqua	**-6.17**	0.23	1.97	**2.39**	0.05	0.82	-1.00	-0.33	0.56	0.51	0.15	-1.04	-0.05	0.84	/	0.30	0.33
	1.81	0.46	1.15	**1.12**	0.46	1.25	0.62	0.59	1.23	1.10	1.14	1.38	1.30	0.87		0.54	0.56
sol	**-6.93**	0.42	/	-1.14	**1.63**	1.41	0.47	0.52	/	-0.07	/	-0.90	-1.37	-0.80	/	/	0.94
	1.52	0.65		1.43	**0.81**	1.18	0.60	0.96		1.16		0.83	1.02	0.79			0.54
pool	**-6.96**	**1.88**	0.33	-1.35	1.67	-2.65	-1.65	2.60	/	/	/	-0.26	-1.50	1.89	0.78	-0.64	-0.04
	1.46	**0.69**	1.25	1.45	0.90	1.40	1.06	1.55				0.54	1.39	1.41	1.14	1.17	1.11
wbed	**-3.90**	-0.52	/	0.07	**1.16**	0.77	-0.01	1.38	1.80	-0.63	/	**-1.56**	-0.37	0.12	0.74	0.10	-0.08
	1.30	0.43		0.95	**0.44**	0.87	0.45	0.78	1.33	0.74		**0.79**	0.83	0.72	1.12	1.10	1.08
flHt	-1.24	**-0.54**	0.60	**1.75**	**0.92**	**-1.28**	**0.51**	0.01	1.49	-0.26	0.22	-0.05	**1.47**	**1.64**	**-1.17**	**-1.28**	**-1.10**
	0.83	**0.27**	0.51	**0.79**	**0.26**	**0.52**	**0.26**	0.43	0.82	0.40	0.68	0.74	**0.67**	**0.52**	**0.51**	**0.45**	**0.43**

To overcome this difficulty a backward elimination technique was used, taking the preliminary models in Table 5.3 as a starting point. Out of the initial model depending on M predictors, M reduced models were estimated by separately eliminating each one of the M parameters. Thus, all the reduced models correspond to the case that the removed parameter has been fixed to zero. Hence, a likelihood ratio test between the initial and the M reduced models could be performed to decide, whether the inclusion of the parameter in the initial model results in a significant increase of predictive power. For this purpose, the likelihood-ratio statistic was compared to a χ^2 distribution with one degree of freedom [112]. If there was at least one reduced model which could not be rejected at a 5% significance level, the one with the lowest likelihood-ratio statistic was chosen as the new "initial" model, and the procedure was repeated until the point where all the reduced models were rejected.

Ordinary logistic regression

The backward elimination of the previous section was applied using the models of Table 5.3 as the initial models. This means that for each appliance the predictors introduced in Table 5.2 were eliminated step-by-step, until a more parsimonious model was connected with a significant loss of predictive power. The final models resulting of this backward elimination procedure of ordinary logistic regression (OLR) are shown in Table 5.4. Disregarding the ASCs, the only two parameters which are not significant are those of the dummy variables rur and own in the model predicting the ownership of a pool. The correlation between those two is less than 0.01. However, multicollinearity between parameters causes that it is not possible to identify directly from the preliminary models (Table 5.3) the parameters not having a significant influence on the prediction. This is apparent, for instance, looking at the influence of women in households on the probability to own a hi-fi system, where the corresponding parameter is insignificant in the preliminary model of Table 5.3, in contrast to the final model in Table 5.4.

Principal component logistic regression

In this section, a methodology is presented to overcome the difficulties in finding a model that is only dependent on significant parameters (related to the correlation of parameters, mentioned in Section 5.2.3). The correlation between different parameters β_j originates from correlations of the different predictors x_j. This can be overcome by using principal components as predictors instead of the original predictors (*cf.*, Table 5.2). The principal components $\boldsymbol{y}$ are defined by

$$\boldsymbol{y} = \boldsymbol{U}^{\mathsf{T}}[\widetilde{\boldsymbol{x}} - \overline{\widetilde{\boldsymbol{x}}}], \tag{5.2}$$

where $\widetilde{\boldsymbol{x}}$ and $\overline{\widetilde{\boldsymbol{x}}}$ represent the set of observations of the original variables (see Table 5.2) and their means. In principal component analysis for linear regression

$\boldsymbol{U}$ represents the covariance matrix of $[\widetilde{\boldsymbol{x}} - \overline{\widetilde{\boldsymbol{x}}}]$. The parameters $(\beta'_0, \ldots, \beta'_M)$ of another logistic regression model can be estimated for each appliance i:

$$p_i(\boldsymbol{y}) = \frac{1}{1 + \exp\left(\beta'_0 + \beta'_1 y_1 + \ldots + \beta'_M y_M\right)} = \frac{1}{1 + \exp\left(\beta'_0 + \boldsymbol{\beta'}\boldsymbol{y}\right)}. \tag{5.3}$$

As the purpose of the transformation in Equation (5.2) is to remove correlation of parameters in the logistic regression model, the matrix $\boldsymbol{U}$ is chosen from the covariance matrix that was estimated in the preliminary logistic regression models of appliance i (*cf.*, Section 5.2.2). However, as in logistic regression, the ASCs β'_0 are also correlated with the rest of the parameters, the corresponding covariance matrices are of one more dimension than the vectors of all predictors. The covariances with β_0 are thus omitted in $\boldsymbol{U}$.

The parameters β' can be retranslated into parameters $\widetilde{\boldsymbol{\beta}}$, which express the dependence of the probabilities on the original predictors $\widetilde{\boldsymbol{x}}$. Combining Equations (5.2) and (5.3), it follows that

$$p_i(\boldsymbol{y}) = \frac{1}{1 + \exp\left(\left[\beta'_0 - \boldsymbol{\beta'}^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}\overline{\widetilde{\boldsymbol{x}}}\right] + \boldsymbol{\beta'}^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}\widetilde{\boldsymbol{x}}\right)} \tag{5.4}$$

Here, the term in the brackets represents the alternative-specific constant $\widetilde{\beta}_0$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{\beta'}$. It follows that the covariance matrix of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\Sigma}^{\widetilde{\boldsymbol{\beta}}} = \boldsymbol{U}\boldsymbol{\Sigma}^{\boldsymbol{\beta'}^*}\boldsymbol{U}^{\mathsf{T}}, \tag{5.5}$$

where $\boldsymbol{\Sigma}^{\boldsymbol{\beta'}^*}$ is derived from the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{\beta'}}$ of $\boldsymbol{\beta'}$, by omitting the covariances to the ASC and by setting to zero all rows/columns corresponding to removed principal components in the backward elimination. This also allows to calculate the standard errors of $\widetilde{\boldsymbol{\beta}}$, given by the square root of the diagonal elements of $\boldsymbol{\Sigma}^{\boldsymbol{\beta'}}$. Furthermore, the variance of $\widetilde{\beta}_0$ (which is by definition not correlated to the components of $\widetilde{\boldsymbol{\beta}}$) is given by

$$\sigma^2_{\widetilde{\beta}_0} = \boldsymbol{a}\boldsymbol{\Sigma}^{\boldsymbol{\beta'}}\boldsymbol{a}^{\mathsf{T}}, \tag{5.6}$$

where $\boldsymbol{a} = (1, -\overline{\widetilde{\boldsymbol{x}}}^{\mathsf{T}}\boldsymbol{U})$.

As the transformation matrix $\boldsymbol{U}$ of the linear transformation of the original predictors, has full rank (*cf.*, Section 5.2.3) and is thus bijective, it follows that $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$, following from the invariance property of maximum likelihood estimators. This means that the logistic regression models that are estimated using the principal components are mathematically fully equivalent to the models that are based on the original predictors.

The major advantage of the PCLR approach is that there is no correlation between the predictors of the initial model (the correlation values between parameters β'_k and β'_l ($k \neq l$ and $k, l \neq 1$) for all appliances i are in the order

of magnitude of the computational precision). Thus, the removal of a principal component does not substantially change the values of the remaining parameters during the backward elimination. However, the more parameters are removed, the more the remaining ones are correlated to each other, as the covariance matrix is an estimation, which represents the real covariance only in the asymptotic mean. The maximal correlation between parameters of all final models amounts to 0.075, which occurred in the model for the ownership of a waterbed (again omitting correlations to the ASC, which could by definition not be treated in the PCLR approach; see above).

Regarding all the initial models, the maximal absolute value of the t test of parameters of principal components that were removed in the corresponding final models, amounts to 1.83; the minimal absolute value of the t test of all parameters that remained was 1.98. As a t test of 1.96 corresponds to a p value threshold of 5 %, this implies that, as a good approximation to the backward elimination procedure, all the principal components can be removed from the initial model at once, when the p value of their t test is not statistically significant. This can strongly reduce computational time compared to the backward elimination in case of complex initial models.

Table 5.4: Results of the original predictor backward elimination

		predictors															
i	ASC	apSf	fam	shr	own	wom	rur	p3+	p5+	r3+	ag2	ag3	kid	olPr	inc1	inc2	inc3
fr2+	-4.41 0.42	0.87 0.20	-	1.08 0.40	-	0.87 0.37	-	-	-	-	-	-	0.48 0.21	-	-	-	-
cOv	1.38 0.16	-	-	0.76 0.36	-1.19 0.13	-0.45 0.17	-	-0.42 0.15	-	-	-	-	-	-	-	0.28 0.13	-
cnOv	-1.78 0.18	-	-	-	1.53 0.13	0.63 0.19	-	-	-	-	-	-	0.54 0.15	-	-0.48 0.22	-0.49 0.15	-
ovSp	-2.27 0.27	-	-	-	1.47 0.14	0.63 0.19	-	-	-	0.52 0.27	-	-	0.50 0.15	-	-0.47 0.22	-0.48 0.15	-
miwa	-1.56 0.22	-	-	-	-	0.49 0.17	-	0.40 0.14	-	0.70 0.21	0.39 0.14	-	-	-	-	-	-
dwsh	-1.68 0.23	-	-	1.03 0.44	1.16 0.15	-	-	-	-	1.13 0.22	0.65 0.16	-	1.50 0.21	0.88 0.17	-0.50 0.23	-0.44 0.17	-
cafe	-1.21 0.20	-	0.36 0.18	-	0.38 0.13	-	-	-	-1.67 0.70	1.34 0.20	0.32 0.14	-	-	0.44 0.15	-	-	-
tv1+	0.27 0.47	-	-	-	-	0.67 0.28	-	-	/	0.70 0.31	1.29 0.42	2.01 0.48	-	-	-	-	-
tvRe	-3.10 0.29	-	-0.78 0.34	-	-	0.79 0.29	0.51 0.23	-	-	-	0.48 0.23	-	0.70 0.31	-	-	-	-
vid	-1.37 0.21	-	-	-	-	0.70 0.16	-	-	-	0.67 0.20	0.65 0.14	-	0.44 0.15	-	-	-	-
dvd	-0.05 0.31	-	-	-	-0.27 0.14	0.89 0.18	-	-	-	-	-	-1.53 0.19	1.08 0.16	-	-0.70 0.35	-1.07 0.31	-0.66 0.30
cons	-3.90 0.52	-	-	-	-	2.27 0.53	-	-	-	-	-	-1.43 0.48	1.04 0.24	-1.06 0.30	-	-	-
hifi	1.40 0.17	-	-	-	-	0.47 0.17	-	-	/	-	-	-1.11 0.16	0.90 0.21	-	-0.75 0.23	-0.80 0.17	-
pc1+	-0.68 0.18	-	-	1.36 0.57	-	0.54 0.24	-	-	/	-	1.86 0.17	-	2.81 0.32	0.55 0.23	-0.86 0.26	-1.18 0.19	-
pr1+	0.20 0.21	-	-	-	0.53 0.16	0.79 0.18	-	-	-	0.62 0.22	-	-1.70 0.17	0.98 0.21	-	-0.98 0.23	-0.90 0.17	-
wlTe	1.27 0.12	-	0.78 0.23	-	-	-	-	-	-	-	-	-1.33 0.16	-	0.59 0.16	-	-	-
st1+	-2.57 0.21	1.16 0.14	-0.46 0.18	-	-	-	-	-	-	-	-	0.60 0.15	-	-	-	-	-
efBl	-1.77 0.19	0.82 0.13	-	-	-	-	-	0.41 0.16	-	-	-	-	-	0.46 0.14	-	-	-
topF	-4.05 0.39	0.60 0.23	-	-	0.56 0.23	1.04 0.32	-	-	-	-	-	-	-	-	-	0.38 0.17	-
frz	-3.14 0.32	0.69 0.18	-	1.62 0.39	0.69 0.17	-	-	-	-	1.16 0.30	-	-	0.86 0.17	0.57 0.17	-	-	-
coWa	4.88 0.36	-3.01 0.27	-	-	-1.50 0.17	-	-	-	-	-0.45 0.23	-	-	-	-	-	-	-0.35 0.16
prWa	-4.75 0.35	2.71 0.26	-	-	1.47 0.17	-	-	-	-	0.79 0.24	-	-	-	-	-	-	0.34 0.16
tmbl	2.10 0.37	-0.97 0.18	-	0.92 0.36	0.43 0.17	-	-	-	-	-	-	-	0.84 0.17	0.71 0.15	-1.33 0.37	-1.40 0.34	-1.05 0.34
boil	-2.35 0.22	1.34 0.14	-	-	-	-0.44 0.17	0.52 0.18	-	-	-	0.34 0.15	-	0.49 0.15	-	-	-	-
radH	-4.26 0.37	1.14 0.22	-	-	-	-	-	-	/	-	-	0.54 0.23	-	-	-	-	-
aqua	-5.02 0.58	-	2.49 0.62	2.80 0.78	-	-	-	-	-	-	-	-	-	1.39 0.65	/	-	-
sol	-5.44 0.58	-	/	-	1.80 0.64	-	-	-	/	-	/	-	-	-	/	/	-
pool	-7.64 1.05	1.55 0.66	-	-	1.70 0.89	-	-1.72 1.03	-	/	/	/	-	-	-	1.12 0.46	-	-
wbed	-4.14 0.34	-	/	-	0.87 0.35	-	-	1.07 0.35	-	-	/	-1.48 0.75	-	-	-	-	-
flHt	-1.82 0.42	-	-	1.47 0.60	0.62 0.21	-0.87 0.38	-	-	-	-	-	-	1.63 0.37	1.23 0.38	-1.18 0.49	-1.28 0.43	-1.14 0.42

Table 5.5: Results of the principal component backward elimination

		predictors															
i	ASC	apSf	fam	shr	own	wom	rur	p3+	p5+	r3+	ag2	ag3	kid	olPr	inc1	inc2	inc3
fr2+	-4.41	1.26	-0.84	0.18	-0.54	0.86	0.03	-0.01	0.35	0.21	-0.19	-0.28	0.82	-0.39	0.07	-0.10	0.10
	0.40	0.30	0.29	0.08	0.29	0.34	0.02	0.16	0.12	0.08	0.11	0.15	0.23	0.19	0.04	0.05	0.05
cOv	1.57	-0.00	0.14	0.82	-1.16	-0.20	-0.07	-0.50	0.17	-0.35	-0.32	0.03	0.06	-0.07	0.26	0.36	0.17
	0.22	0.17	0.13	0.38	0.16	0.07	0.05	0.20	0.07	0.05	0.08	0.07	0.11	0.06	0.13	0.12	0.13
cnOv	-1.47	-0.03	0.41	0.31	1.46	0.30	0.07	0.20	-0.06	0.29	-0.18	-0.25	0.25	0.39	-0.64	-0.78	-0.33
	0.44	0.18	0.12	0.18	0.17	0.10	0.05	0.14	0.02	0.04	0.14	0.12	0.10	0.09	0.32	0.31	0.29
ovSp	-1.47	0.04	0.22	-0.59	1.48	0.38	0.07	0.53	-0.26	0.31	-0.12	-0.12	0.12	0.32	-1.04	-1.03	-0.62
	0.44	0.18	0.17	0.43	0.17	0.10	0.05	0.21	0.08	0.04	0.17	0.14	0.13	0.10	0.33	0.31	0.31
miwa	-2.15	0.30	-0.33	-0.02	0.18	0.59	-0.05	0.34	0.06	0.57	0.37	0.06	-0.14	-0.55	0.47	0.54	0.50
	0.42	0.11	0.12	0.19	0.12	0.18	0.03	0.19	0.03	0.17	0.11	0.09	0.09	0.17	0.29	0.28	0.26
dwsh	-1.61	-0.01	0.29	1.48	1.15	0.20	-0.32	-0.33	0.50	1.18	0.94	0.34	1.41	0.72	-0.61	-0.88	-0.53
	0.52	0.21	0.31	0.44	0.20	0.13	0.21	0.17	0.19	0.21	0.42	0.45	0.37	0.11	0.22	0.23	0.20
cafe	-1.68	0.10	0.30	0.42	0.35	0.06	-0.02	0.05	-1.41	1.27	0.69	0.52	0.16	0.26	0.02	-0.22	0.15
	0.28	0.09	0.25	0.18	0.09	0.12	0.04	0.17	0.66	0.20	0.22	0.23	0.28	0.08	0.03	0.06	0.05
tv1+	1.83	0.06	0.06	0.18	0.26	0.19	-0.10	0.31	/	0.61	0.19	0.82	-0.98	-0.14	-0.20	0.27	-0.19
	0.21	0.07	0.08	0.06	0.08	0.05	0.04	0.15		0.16	0.15	0.21	0.28	0.16	0.07	0.06	0.05
tvRe	-3.14	0.15	-0.45	0.03	0.25	0.98	0.10	-0.34	0.25	0.23	0.13	-0.23	0.38	-0.41	-0.09	0.02	-0.00
	0.28	0.09	0.17	0.01	0.09	0.26	0.02	0.14	0.08	0.05	0.11	0.09	0.12	0.19	0.03	0.07	0.06
vid	-1.15	0.13	0.27	-0.06	0.17	0.93	0.01	-0.06	0.08	0.36	0.28	-0.16	0.24	-0.10	-0.05	-0.08	0.05
	0.18	0.06	0.05	0.03	0.07	0.18	0.02	0.16	0.03	0.07	0.07	0.07	0.06	0.19	0.03	0.06	0.06
dvd	-0.08	-0.05	-0.11	0.63	-0.10	0.54	0.07	0.29	0.03	-0.05	0.26	-1.39	1.02	0.09	-0.70	-1.01	-0.66
	0.35	0.03	0.23	0.17	0.03	0.07	0.01	0.06	0.02	0.04	0.16	0.16	0.19	0.09	0.32	0.29	0.27
cons	-3.01	-0.26	-0.14	-0.46	0.03	2.53	0.05	0.50	0.14	0.62	-0.98	-2.00	0.95	-0.82	-0.57	-0.82	-0.54
	0.61	0.06	0.27	0.23	0.07	0.58	0.02	0.17	0.75	0.14	0.32	0.54	0.34	0.13	0.23	0.23	0.23
hifi	1.52	0.21	-0.23	0.27	0.17	0.27	-0.34	0.15	/	0.02	0.11	-1.09	0.83	0.01	-1.23	-1.10	-0.42
	0.48	0.06	0.29	0.09	0.07	0.07	0.21	0.05		0.04	0.21	0.20	0.24	0.07	0.43	0.40	0.40
pc1+	1.01	0.02	0.06	0.07	0.06	0.72	0.16	0.57	/	0.26	0.27	-1.77	1.54	0.44	-1.28	-1.50	-0.37
	0.50	0.07	0.35	0.15	0.10	0.16	0.04	0.14		0.09	0.28	0.28	0.30	0.15	0.44	0.41	0.40
pr1+	0.18	0.25	0.24	1.33	0.19	0.76	0.09	-0.06	0.12	0.16	0.28	-1.35	1.03	0.40	-1.35	-1.13	-0.35
	0.39	0.07	0.30	0.35	0.08	0.11	0.02	0.13	0.03	0.05	0.18	0.18	0.22	0.09	0.33	0.30	0.28
wlTe	0.12	0.11	0.17	0.03	0.09	0.47	0.06	0.18	0.01	0.15	0.57	-0.56	0.19	0.24	-0.03	0.05	-0.03
	0.15	0.06	0.04	0.00	0.07	0.06	0.01	0.04	0.00	0.02	0.06	0.05	0.05	0.08	0.01	0.06	0.05
st1+	-2.46	1.18	-0.13	0.09	-0.12	0.00	0.00	0.08	-0.00	0.19	-0.33	0.33	-0.13	0.01	0.06	-0.19	0.15
	0.23	0.18	0.03	0.02	0.18	0.05	0.01	0.03	0.00	0.03	0.06	0.06	0.04	0.08	0.01	0.07	0.07
efBl	-1.55	0.41	0.05	0.02	0.45	0.18	0.06	0.08	-0.00	0.12	0.12	-0.09	0.05	0.13	0.03	-0.12	0.10
	0.16	0.05	0.03	0.00	0.06	0.02	0.01	0.03	0.00	0.02	0.02	0.02	0.03	0.03	0.00	0.03	0.03
topF	-3.57	0.49	0.29	-0.06	0.53	0.34	0.29	0.21	-0.02	0.19	-0.26	0.26	0.21	0.37	0.08	0.16	-0.20
	0.27	0.09	0.09	0.03	0.09	0.10	0.07	0.06	0.01	0.04	0.06	0.06	0.07	0.14	0.02	0.05	0.04
frz	-3.52	0.69	0.18	1.87	0.68	-0.00	0.08	-0.37	-1.27	1.15	0.25	0.25	1.11	0.57	0.33	0.03	0.21
	0.34	0.07	0.18	0.45	0.08	0.11	0.02	0.18	0.62	0.27	0.27	0.29	0.26	0.09	0.08	0.10	0.10
coWa	4.79	-3.05	0.26	-0.06	-1.44	-0.23	-0.07	-0.07	-0.01	-0.17	-0.21	0.08	0.02	-0.13	-0.16	0.33	-0.16
	0.34	0.27	0.13	0.13	0.17	0.13	0.05	0.15	0.01	0.08	0.08	0.08	0.14	0.13	0.03	0.07	0.08
prWa	-4.71	2.75	-0.14	0.00	1.37	0.07	-0.01	0.02	0.02	0.71	0.23	-0.16	0.01	0.20	0.17	-0.34	0.14
	0.34	0.26	0.12	0.13	0.17	0.15	0.07	0.16	0.01	0.25	0.08	0.08	0.14	0.12	0.03	0.07	0.07
tmbl	1.46	-0.97	-0.06	0.45	0.45	-0.13	-0.04	-0.11	0.75	0.39	0.65	0.68	0.83	0.65	-1.47	-1.48	-1.32
	0.50	0.17	0.17	0.14	0.16	0.12	0.02	0.06	0.29	0.18	0.34	0.37	0.24	0.11	0.37	0.35	0.34
boil	-2.24	1.09	0.09	-0.08	0.31	-0.37	0.20	-0.11	0.02	0.43	0.17	-0.24	0.23	-0.28	-0.18	0.16	0.02
	0.26	0.14	0.09	0.10	0.15	0.11	0.05	0.14	0.01	0.14	0.08	0.08	0.10	0.10	0.13	0.10	0.10
radH	-4.12	1.13	0.02	-0.04	-0.00	0.07	0.03	-0.06	/	0.07	-0.24	0.24	-0.02	0.11	0.10	-0.09	0.02
	0.39	0.31	0.03	0.02	0.31	0.09	0.01	0.06		0.02	0.09	0.09	0.04	0.13	0.02	0.12	0.11
aqua	-6.78	0.04	0.53	1.16	-0.04	1.32	-1.14	-0.34	0.80	1.21	0.60	-0.53	0.54	0.07	/	0.46	0.19
	1.02	0.07	0.14	0.46	0.03	0.46	0.65	0.28	0.31	0.47	0.39	0.29	0.16	0.14		0.12	0.20
sol	-6.93	0.90	/	-0.04	0.62	0.12	-0.09	-0.04	/	0.22	/	-0.18	0.06	0.37	/	/	1.06
	0.93	0.29		0.01	0.22	0.05	0.09	0.06		0.08		0.11	0.07	0.18			0.50
pool	-6.83	2.04	0.86	-0.31	1.09	-2.88	-0.13	2.79	/	/	/	-0.25	-1.76	1.86	0.25	-0.82	0.28
	0.92	0.41	0.62	0.31	0.25	1.32	0.09	1.50				0.49	0.81	1.29	0.25	0.35	0.16
wbed	-3.82	-0.53	/	-0.15	0.72	0.95	0.09	0.97	-0.33	-0.84	/	-0.97	0.51	0.56	0.07	0.15	-0.23
	0.35	0.32		0.06	0.24	0.28	0.03	0.22	0.11	0.55		0.36	0.12	0.27	0.03	0.08	0.08
flHt	-1.32	-0.50	0.99	0.54	0.95	-1.03	0.17	0.27	2.44	-0.03	-0.19	-0.50	0.30	1.03	-0.73	-0.68	-0.44
	0.41	0.22	0.19	0.24	0.22	0.26	0.03	0.18	0.63	0.05	0.10	0.09	0.10	0.28	0.20	0.16	0.14

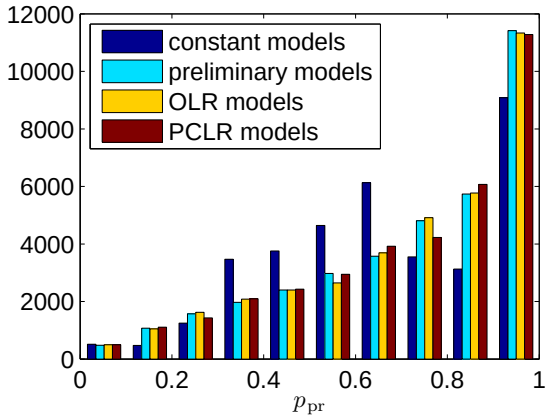


Figure 5.1: Distribution of the probabilities p_{pr} with which the correct outcome is predicted for every appliance and every household in the sample.

5.3 Results

5.3.1 Model comparison

The two models in Sections 5.2.3 and 5.2.3 have been derived from the preliminary model of Section 5.2.2 using likelihood ratio tests, and should thus be favoured over the latter. The probabilities p_{pr}, with which the models predict the correct outcome of the observed appliance ownership over the whole set of modelled appliances and the entire sample population is shown in Figure 5.1. Constant models refer to logistic regression models which are not dependent on any predictor. These models correctly predict the observed overall sample diffusion rate. Preliminary models denote the models that were defined in Section 5.2.2. Except for the constant models, the models show comparable predictive power. Comparing the predictions of the constant models with those of others reveals the need to model appliance ownership in dependence of household specificities. The medians of the distributions are 65.4 %, 78.3 %, 77.9 % and 78.3 % for the constant, the preliminary, the OLR and the PCLR models, respectively. The sum of logarithms of all p_{pr} corresponds to the sum of log-likelihoods of all appliances that are shown in Table 5.6.

In order to estimate which one of the two approaches is preferable over the other, Davidson and MacKinnon J tests [112, 141] were performed using the open source software Biogeme 1.8 [114–116]. This test is used to decide, which one of two competing hypotheses H_1 and H_2 should be preferred:

- $\mathrm{H}_1 : V(\widetilde{\boldsymbol{x}}) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{x}}$, meaning that the deterministic part of the utility function $V(\widetilde{\boldsymbol{x}})$ (given by the exponent in Equation (5.1)) is described by the OLR model.
- $\mathrm{H}_2 : V(\widetilde{\boldsymbol{x}}) = \widetilde{\beta}_0 + \widetilde{\boldsymbol{\beta}}^{\mathsf{T}}\widetilde{\boldsymbol{x}}$, meaning that $V(\widetilde{\boldsymbol{x}})$ is described by the PCLR model.

In order to test H_1 one considers the composite model

$$H_c : V(\widetilde{\boldsymbol{x}}) = (1-\alpha)(\widetilde{\beta}_0 + \widetilde{\boldsymbol{\beta}}^\mathsf{T}\widetilde{\boldsymbol{x}}) + \alpha(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\mathsf{T}\widetilde{\boldsymbol{x}}), \tag{5.7}$$

where $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ denote the parameter values that were estimated in the model of H_1. Hence, $(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\mathsf{T}\widetilde{\boldsymbol{x}})$ is the estimated value of the deterministic part of the utility, and represents a single continuous predictor variable, for which the parameter α is estimated. If the value of α is significantly different from 0, this means that one can not reject the Hypothesis that H_1 is correct. To test H_1 the same procedure is applied, where the roles of H_1 and H_2 in Equation (5.7) have to be interchanged. The results of these tests are shown in Table 5.6 for the OLR, as well as the PCLR models.

This means that for instance for frz (freezer) the value of $\alpha = 0.82 \pm 0.38$ tests the hypothesis, whether the OLR model can be rejected compared to the PCLR model. Thus, the OLR model could be rejected with a 95 % confidence level. The other way round, $\alpha = 0.50 \pm 0.62$ indicates that the PCLR model could not be rejected.

There are four possible outcomes of the J test [112]:

- In two possible outcomes, only one of the two models is rejected, meaning that the other model should be preferred.
- Both models are rejected, indicating that better models should be preferred.
- Both models can not be rejected, indicating that the data is not informative enough to distinguish between the two models.

Omitting the cases where one of both composite models could not be meaningfully estimated, this means that the OLR model was rejected 8 times, whilst the PCLR model could not be rejected. The PCLR model was rejected 10 times, whilst the OLR model could not be rejected. Both models were rejected 5 times, and twice none of them could be rejected, showing in summary that both model specification methodologies are similarly convincing. The PCLR model is more often rejected than the ordinal LR model. However, most often the significance level of the t test of α is not very low, meaning that the models are not rejected with a very high confidence level.

Table 5.6: Results of the Davidson and MacKinnon J tests.

i	#(OLR)	#(PCLR)	$\mathcal{L}$(OLR)	$\mathcal{L}$(PCLR)	α(OLR)	α(PCLR)
fr2+	4	5	-359.76	-359.24	*	**0.70 ± 0.29**
cOv	5	6	-731.10	-730.63	0.52 ± 0.36	0.58 ± 0.34
cnOv	5	7	-675.15	-674.61	0.53 ± 0.36	0.61 ± 0.34
ovSp	6	8	-664.38	-661.43	0.52 ± 0.45	**0.80 ± 0.31**
miwa	4	6	-791.37	-790.45	**0.66 ± 0.24**	**0.62 ± 0.22**
dwsh	8	10	-627.24	-626.20	0.63 ± 0.34	**0.81 ± 0.33**
cafe	6	7	-731.88	-732.73	**0.70 ± 0.32**	0.59 ± 0.33
tv1+	4	2	-269.05	-273.36	**0.83 ± 0.26**	0.45 ± 0.38
tvRe	5	4	-433.76	-434.53	**0.79 ± 0.37**	0.62 ± 0.39
vid	4	5	-757.12	-757.96	**0.98 ± 0.17**	0.69 ± 0.42
dvd	7	6	-687.71	-689.67	**0.71 ± 0.30**	0.51 ± 0.35
cons	4	6	-380.66	-378.22	0.56 ± 0.32	**0.72 ± 0.26**
hifi	5	7	-588.82	-589.10	**0.91 ± 0.41**	0.54 ± 0.34
pc1+	7	8	-486.45	-484.47	0.50 ± 0.33	**0.77 ± 0.31**
pr1+	7	7	-587.08	-589.19	**0.80 ± 0.27**	**0.66 ± 0.32**
wlTe	3	3	-595.75	-595.93	**0.58 ± 0.27**	0.51 ± 0.26
st1+	3	4	-671.97	-672.67	*	0.48 ± 0.39
efBl	3	2	-776.04	-776.06	**0.54 ± 0.25**	**0.56 ± 0.26**
topF	4	3	-440.79	-441.73	**0.64 ± 0.29**	*
frz	6	7	-699.45	-697.26	0.50 ± 0.62	**0.82 ± 0.38**
coWa	4	6	-519.57	-520.16	**0.81 ± 0.35**	0.50 ± 0.30
prWa	4	7	-537.33	-539.38	**0.97 ± 0.28**	0.34 ± 0.30
tmbl	8	6	-781.32	-780.58	**0.87 ± 0.36**	**0.91 ± 0.34**
boil	5	8	-728.19	-728.56	**0.67 ± 0.31**	0.58 ± 0.31
radH	2	4	-311.77	-311.42	0.32 ± 1.36	*
aqua	3	4	-153.72	-152.04	0.64 ± 0.34	**0.73 ± 0.29**
sol	1	2	-79.80	-78.53	0.74 ± 0.54	*
pool	4	4	-95.50	-94.30	0.60 ± 0.18	**0.74 ± 0.28**
wbed	3	3	-153.14	-154.76	**0.85 ± 0.37**	0.42 ± 0.36
flHt	8	4	-355.95	-353.62	**0.58 ± 0.25**	**0.79 ± 0.24**

specifies the number of parameters in the model (apart from the ASC), $\mathcal{L}$ is the value of the log-likelihood and α is the parameter to estimate, whether the other model should be rejected, by testing for statistically significant difference from zero. The values are boldface when α is significantly different from 0 at the 5 % confidence level.

A star indicates that the composite model of the Davison and MacKinnon J test included parameters, which could not be meaningfully estimated (either there were extreme standard errors of parameters, or the parameter value reached the boundaries of the constraint optimisation problem; usually indicating linear dependences between parameters in the composite model).

5.3.2 Application and validation

In this section the models are applied to predict the diffusion of the electrical appliances for sub-populations, to demonstrate the predictive power of both model specification methodologies. This is done by calculating the probabilities (according to Equation (5.1)) that the appliances are present in the 1200 households of the sample population that was described in Section 5.2.1.

The results are shown in Table 5.7. D_{tot} indicates the observed diffusion percentage of the corresponding appliance in the survey dataset, which was described in Section 5.2.1. A model containing no other parameters than the ASC (refered to as "constant model" in Figure 5.1) predicts by definition a mean diffusion probability $\overline{p} = D_{\mathrm{tot}}$, regardless for which sub-population. In the presence of other parameters β_j, the ASCs in the logit models secure that for the whole sample population the mean percentage predicted by the model always reproduces an average diffusion probability equal to D_{tot}. However, regarding sub-populations in the sample the diffusion percentages differ. These percentages D_{sub} are shown in Table 5.7 for the sub-populations, where the corresponding dummy variable apSf, own, rur, kid and inc3 is equal to one. The mean predicted diffusion probabilities over these sub-populations $\overline{p}_{\mathrm{sub}}$ of the OLR model $\overline{p}_{\mathrm{O}}$ and and the PCLR model $\overline{p}_{\mathrm{PC}}$ (according to Equation (5.1)) are also shown. The predicted diffusion rates of the corresponding appliance for the complementary sub-population $\overline{p}^{\mathrm{c}}_{\mathrm{sub}}$ (corresponding dummy variable equal to 0) can be derived via $D_{\mathrm{tot}} = H_{\mathrm{sub}} \cdot \overline{p}_{\mathrm{sub}} + (1 - H_{\mathrm{sub}}) \cdot \overline{p}^{\mathrm{c}}_{\mathrm{sub}}$, where H_{sub} corresponds to the sub-population proportion shown in Table 5.2. The PCLR model prediction of the diffusion rate of dishwashers for the sub-population without minors in the household (kid = 0) is, for instance, $(34.7\,\% - 17.4\,\% \cdot 29.1\,\%)/70.9\,\% = 41.8\,\%$.

Table 5.7 shows that the predictions of both models are convincing. Analysing the predictions $\overline{p}_{\mathrm{sub}}$ for the whole set of sub-populations where one of the predictors is equal to 1 (see Table 5.2, not only the five shown in the table) and all appliances, the maximal deviation of the predictions are found to be -22.3 % for the PCLR model, and -25.1 % for the OLR model. In both cases this maximal deviation occurred for the top-opening freezer (topF) and the sub-population of households with more than 4 household members (p5+). The maximal relative deviation of -70.6 % occurs for the prediction of the PCLR model that more than one TV is in the household for the sub-population of low incomes (inc1). For the OLR model, the maximal relative deviation of +116.7 % occurs for the prediction that more than one TV is in the household for the sub-population of flat shares.

The distribution of the decadic logarithm of the relative error of the predictions $\eta = \log\left((D_{\mathrm{sub}} - \overline{p}_{\mathrm{mod}})\right)/\ D_{\mathrm{sub}})$ is shown in Figure 5.2. The distribution of the OLR model is not shown for x values below -5, as every time the dummy variable is included in the OLS model (142 times in total, *cf.* Table 5.4), the accuracy of the prediction of the diffusion rate $\overline{p}_{\mathrm{O}}$ is given by the computational precision (compare Tables 5.4 and 5.7). In other words, the preliminary models

of Table 5.3 correctly predict the diffusion rates of all sub-populations defined by the dummy variables, which are included in the models. Therefore, the median of the distribution of η of the OLS model of 0.65 % is smaller than the one of the PCLR model of 1.1 %. However, the 95 % quantiles amount to 18.4 % and 16.3 %, respectively, showing that the OLR is less robust against outliers than the PCLR, which is also apparent comparing the maximum of the relative errors (see above).

Table 5.7: Predictions of sub-population shares of the OLR and the PCLR models.

		apSf			own			rur			kid			inc3		
app	D_{tot}	D_{sub}	$\overline{p}_{\text{PC}}$	$\overline{p}_{\text{O}}$	D_{sub}	$\overline{p}_{\text{PC}}$	$\overline{p}_{\text{O}}$	D_{sub}	$\overline{p}_{\text{PC}}$	$\overline{p}_{\text{O}}$	D_{sub}	$\overline{p}_{\text{PC}}$	$\overline{p}_{\text{O}}$	D_{sub}	$\overline{p}_{\text{PC}}$	$\overline{p}_{\text{O}}$
fr2+	90.2	93.5	93.4	93.5	87.4	87.7	86.2	89.1	87.8	87.4	85.4	84.9	85.4	91.6	90.2	91.2
cOv	37.8	31.1	31.1	31.1	54.7	54.5	54.7	42.4	44.8	43.2	45.6	44.4	45.2	39.4	38.7	38.6
cnOv	65.4	73.6	73.5	73.6	45.1	45.3	45.1	57.0	57.7	58.9	55.6	56.3	55.6	63.7	64.0	62.9
ovSp	66.3	74.6	74.7	74.6	46.1	46.0	46.1	57.0	58.4	59.7	56.2	56.8	56.2	65.3	65.3	64.1
miwa	54.1	57.6	58.8	56.4	48.2	47.0	50.7	46.7	49.8	50.1	44.7	45.1	43.7	54.8	55.0	56.0
dwsh	34.7	42.3	42.2	42.3	17.7	17.8	17.7	31.5	32.0	27.4	16.6	17.4	16.6	35.7	36.8	34.9
cafe	37.0	41.7	41.6	41.0	28.1	27.6	28.1	30.3	33.0	32.8	31.5	30.5	30.9	37.0	35.4	38.5
tv1+	6.3	7.0	6.8	6.9	4.5	5.1	5.2	4.8	6.5	6.0	7.4	8.7	7.0	7.4	7.2	6.7
tvRe	87.7	89.3	89.3	88.4	86.0	84.8	86.5	81.2	85.1	81.2	84.5	84.3	84.5	87.1	88.1	88.0
vid	41.9	45.3	46.0	44.6	37.0	35.4	38.1	33.3	36.9	37.5	27.8	27.9	27.8	41.5	43.1	44.2
dvd	56.1	57.7	58.0	57.6	56.9	55.3	56.9	54.5	49.9	52.0	31.2	32.3	31.2	56.5	56.8	56.5
cons	86.2	88.1	87.6	88.3	83.5	85.2	84.6	80.0	82.8	82.6	66.5	66.5	66.5	87.5	88.7	89.4
hifi	24.5	26.0	26.8	25.6	22.6	21.7	23.8	26.1	25.5	21.9	11.2	12.3	11.2	19.5	19.4	18.9
pc1+	27.3	30.3	30.0	29.7	23.8	24.5	25.1	20.0	20.2	22.0	4.6	5.3	4.6	22.4	22.2	21.8
pr1+	32.7	37.3	36.8	36.7	25.6	27.1	25.6	24.2	25.6	26.5	12.0	13.4	12.0	27.7	28.3	27.1
wlTe	23.6	25.6	26.0	24.6	20.7	20.3	22.7	20.6	18.8	20.7	12.6	13.3	13.6	23.4	24.9	23.9
st1+	71.5	78.2	78.3	78.2	63.8	63.8	63.2	69.1	69.7	69.6	75.9	74.4	75.9	68.8	69.3	71.6
efBl	60.2	66.8	66.2	66.8	49.8	49.5	51.9	52.1	54.9	56.2	55.6	55.6	56.4	59.3	58.6	61.2
topF	86.5	90.9	90.5	90.9	79.3	79.6	79.3	78.8	80.4	83.1	81.9	81.9	82.6	89.7	89.9	89.1
frz	58.2	69.0	69.0	69.0	39.6	39.6	39.6	46.7	49.0	50.1	44.7	44.6	44.7	59.8	60.6	61.9
coWa	48.4	27.8	27.9	27.8	82.1	81.6	82.1	63.0	61.6	60.9	54.7	55.3	57.1	49.5	50.1	49.5
prWa	50.6	70.4	70.3	70.4	17.5	18.3	17.5	38.8	38.6	38.3	43.0	42.8	41.7	49.7	49.2	49.7
tmbl	41.7	38.0	38.2	38.0	42.1	41.5	42.1	48.5	43.9	43.2	39.3	40.4	39.3	40.5	42.6	40.5
boil	61.2	71.3	71.0	71.3	48.8	48.3	50.0	45.5	50.7	45.5	49.0	48.7	49.0	62.8	63.1	62.6
radH	92.3	95.0	95.0	95.0	89.0	88.9	89.0	91.5	91.2	91.2	93.1	92.4	92.2	92.2	92.4	92.6
aqua	96.8	97.5	97.2	97.3	96.1	96.5	96.3	98.2	98.5	96.1	94.0	93.8	94.1	97.3	97.7	97.6
sol	98.7	99.3	99.3	99.1	97.4	97.7	97.4	97.6	98.6	98.3	98.6	98.7	98.5	97.9	97.9	98.7
pool	98.0	99.5	99.5	99.5	95.7	95.9	95.7	99.4	97.7	99.4	98.6	98.3	97.7	97.9	97.3	98.5
wbed	96.8	97.4	97.1	97.7	95.1	95.9	95.1	95.8	95.7	95.6	94.6	93.8	93.9	97.7	97.8	97.2
flHt	90.6	91.2	91.3	92.1	87.0	87.1	87.0	85.5	88.3	89.4	86.5	86.6	86.5	91.6	90.7	91.6

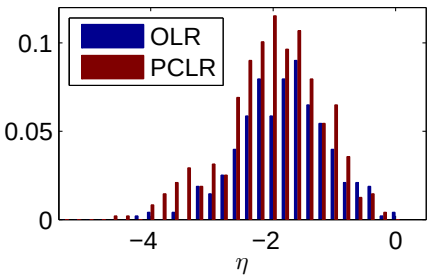


Figure 5.2: Distribution of the decadic logarithm of the relative errors of the predicted sub-population shares of the OLR and the PCLR.

5.4 Discussion

The two approaches to predict appliance ownership in households have been formulated to take into account individual characteristics of the households. The accuracy of both models is comparably satisfying regarding the predictions of the diffusion rate in demographic sub-populations. To address this, two automatised model calibration algorithms have been tested.

As experience teaches us that human behaviour can be very diverse, it is unsurprising that the modelling could still be refined:

- The inclusion/removal of predictor variables could still improve the predictive power of the models. For instance, the number of households with more than 4 persons (p5+) is based on a very small sub-population of 0.8 %, which does not secure statistically significant information. Whilst the backward elimination of the OLR included a parameter for this dummy variable only once, it yielded a significant parameter $\widetilde{\beta}$ in 12 of 24 cases for the PCLR approach. This shows that it is more important for the PCLR to base the approach on a meaningful initial model (*cf.* Table 5.3). However, the validation has shown that the model approaches are nevertheless robust enough, to yield accurate predictions. Further predictors that could be included are, for instance, a dummy variable that men are in the household or a finer resolution of income classes.

  Unfortunately, the household characteristics that were recorded in the survey (*cf.* Section 5.2.1) are not very manifold (fam, shr, olPr are secondary variables derived of gender and age class information; the latter should contain more than 3 different categories; the employment status of household members could also be of interest). Furthermore, a more realistic representation of the disposable income per household member would be based on the equivalence income (according to the OECD modified scale [142]), but this can not be exactly derived as the age class information in the survey is not detailed enough.

- For several appliances, the number of units was also recorded, which could be modelled using a multinomial logit approach. However, the backward elimination of insignificant variables (see Tables 5.4 and 5.5) took up to 10 min using MATLAB. In a multinomial logit model, the significance of parameters is in general choice-specific (choice corresponding to the number of appliances in the household). Therefore, the number of parameters to estimate is proportional to the number of choices, whereas the number of correlation coefficients is proportional to the square of the latter. This causes the elimination procedure for a multinomial logit model to be much more time-consuming. In addition, the modelling of choice-specific parameters in multinomial logit models is not possible in MATLAB, and thus, would also have to be done using Biogeme, necessitating the calibration data reimport for each intermediate model during the backward elimination process, even more increasing the time required.

Looking at the estimated parameters of the two approaches in Tables 5.4 and 5.5, it can be summarized that the type (apartment/single family home), as well as the type of ownership (tenant/owner) of the dwelling, the presence of women and children, the age composition and the revenue of households have an important influence on the ownership of electrical appliances. Looking at the distribution of the predictors (Table 5.2), the demographics of the dataset seem to be representative for Switzerland. For instance, the Swiss average of the year 2000 of the proportion of single or two person households of about 67.6 % is close to the survey value of 72.4 % (of 2005) [143]. However, household sizes of more than 4 are underrepresented in the sample with 0.8 % compared to the national average of 6.3 %. This could be due to a biased sample population/response rate, or to misinformation of the questionnaires' respondents.

5.5 Conclusions

In this chapter, we present approaches to the modelling of the probability of electrical appliance ownership, based on logistic regression models dependent on household characteristics. The purpose of these models is to support more accurate predictions, which electrical appliances are in present households with given specificities. Preliminary binomial logit models were estimated, always using the same set of predictors (occasionally reduced if the standard errors were too large). To eliminate redundant parameters due to correlations, two backward elimination approaches were presented, resulting in models with similar predictive power. In this regard, the approach based on principal components is shown to be a valuable tool for the specification of parsimonious models, by bypassing the backward elimination procedure, which can be very time-consuming in case of complex models.

The set of predictor variables potentially influencing appliance ownership was reduced to the set shown in Table 5.2. However, the formulation of the approach allows to automatically derive a significant model based on any type of predictor set. Other household characteristics, not recorded in the calibration dataset, may have considerable influence on appliance ownership, such as the employment status or a finer age resolution of household members. Besides, the ownership of the number of an electrical appliance could be treated using a multinomial logit approach, which would require more efforts for model specification.

For future research, the principal component logistic regression approach should be tested with more sophisticatedly derived preliminary models (*cf.* Table 5.3), where a wider scope of predictor variables should be included. Moreover, the approaches should also be calibrated using datasets of different countries and years to study cultural and temporal differences.

Chapter 6

Load profile modelling

In this chapter, we present a bottom-up approach to predict residential load profile distributions resulting from the use of individual electrical appliances in dependence of residential activities. The modelling of electrical appliance use is calibrated with measurements of the IRISE campaign, where the electric power consumption of 98 households and individual electrical appliances have been measured as a function of time during approximately one year.

This chapter begins with a brief review of the relevant research in the field of electricity load modelling (Section 6.1), followed by a description of the IRISE dataset (Section 6.2). By comparing the time-dependent probabilities that an appliance is on/switched on with those that relevant activities are done/started (cf. Chapter 4) whilst being at home (cf. Chapter 3), the conditional probability to use/switch on an electrical appliance whilst performing/starting an activity is derived by means of regression analysis (Section 6.3). The distribution of residential load profiles is then treated as the superposition of power consumption distributions of individual appliances (Section 6.4). We go on to discuss further improvements in the modelling and the calibration of residential load profiles, as well as how this can be implemented in simulations. Finally, we will conclude and point out further perspectives for research in load profile modelling.

6.1 Introduction

As many electrical appliances are controlled by occupants, residential electricity load curves depend to a great extent on behavioural aspects. It is thus useful to predict the use of these appliances as a function of the time-dependent residential activities that are performed whilst being at home. We first present a short summary of previous research conducted in this field and outline the requirements for further developments.

6.1.1 State of the art

A detailed statistical evaluation of the patterns of appliance use, as well as the dependence of the use of different electrical appliance types on various factors (such as socio-economic or household characteristics) is provided by Mansouri et al. [144]. This study is based on a questionnaire survey of over 1000 households in the United Kingdom conducted between May and November 1994, where also appliance ownership, attitudinal, and other socio-psychological factors were investigated. Firth et al. [145] present a statistical evaluation of annual consumption values of an electricity consumption monitoring campaign of 72 residential dwellings in the United Kingdom, where they disaggregate the entire household load into "active" and "standby" appliance groups.

A model to predict residential load profiles as a function of occupancy, as well as psychological factors ("proclivity functions") was developed by Walker and Pokoski [50]. Schick et al. [146] present a model to estimate load profiles, based on demographic statistics and climatic data. A simple prediction model for the load profiles of various building types was developed by Jardini et al. [147]. Yao and Steemers [148] present another simple method to predict residential electricity demand profiles based on different pre-defined occupancy and appliance consumption patterns, where random noise is added in a not further defined manner.

A pioneering bottom-up model was developed by Capasso et al. [62], where availability and appliance usage starts are represented by probabilistic functions, which are used together with socio-economic and load data for appliance duty cycles to predict residential load profiles. A similar model to predict residential load profiles was presented by Paatero and Lund [149], calibrated with aggregated statistical data. Apart from the time of day, the model depends on a seasonal and a social random factor, as well as the weather and the week. However, these models were not validated, regarding the distribution of the predictions over households with different characteristics.

Armstrong et al. [150] present a stochastic model to predict electricity load profiles of Canadian households. which is calibrated with average national data about residential electricity use for appliances and lighting [151], as well as with national census data. This model was also used to estimate national residential electricity and domestic hot water loads load using neural network techniques [152]. However, the results of another study revealed that the model of Armstrong et al. does not adequately reproduce either the temporal variability nor the inter-household variations of the distribution of measured electricity load data [52].

Dickert and Schegner [153] model load curves of individual residential appliances as a function of switch-on times, duration distributions and power consumption distributions that are normally distributed around their mean value, if the appliances are in the households according to pre-defined penetration rates of

up to 100 %. Furthermore, they use pre-defined rules, e.g., that a tumble-dryer is used 5 to 60 minutes after the end of the use of the washing machine. The method is validated by comparing the distribution to literature values and the predicted load profiles with measured ones. However, the assumption of normally distributed random variables is not based on measurements. In Figures 6.2, 6.3 and 6.11a, we show that normal distributions are in most cases a too simplistic candidate to accurately represent the observed distributions.

Richardson et al. used time use data to derive profiles of active occupancy (at home and awake), as well as of different activity types. In a simulation each household is then populated with appliances, and residential electricity loads are stochastically modelled based on fixed parameters that were calibrated from measurements [23, 64].

Tanimoto et al. developed a methodology where time-dependent occupant behaviour schedules are stochastically generated, and loads from heating, ventilating and air conditioning systems are simulated as a Markov chain. The load profiles of other electrical appliances are derived from occupancy and behavioural schedules, by means of the respective utility demand schedules [25, 45, 154, 155].

Widén et al. developed approaches where time use patterns (types of place and activities) are predicted deterministically or stochastically based on Markov chains to generate residential electricity loads. The link between occupancy and activities with the power demand of appliances is provided by deterministic conversion functions that are pre-defined for the respective pairs of occupancy, (sets of) activities and the appliances. However, the models were not validated by comparing the distribution of predicted load profiles with the measured one [20, 40, 54].

6.1.2 Perspectives

This review underlines the need of a validated methodology which enables to predict the distribution of residential electricity loads as a function of other influencing parameters, which is also important from an economic point of view [156]. The importance of the uncertainties of the aggregated load profile distribution of multiple households is increasing with a decreasing number of households. Therefore, an accurate prediction of the load distribution is particularly important for the prediction of the aggregated load of households in small-scale neighbourhoods, in order to match the demand with supply from decentralised power generation infrastructures.

In this context, the dependence of the model on significant explanatory variables is crucial, in order to allow for a reliable model application to predict electricity demand in scenarios with boundary conditions that differ from those of the calibration datasets. To enable the accurate application of the model for future scenario predictions, it is furthermore important that the model parameters

be related to observable physical quantities, whose changes can be meaningfully estimated for such scenarios.

6.2 The IRISE survey

6.2.1 General characteristics

The dataset used to calibrate the models originates from the IRISE campaign [157], which has been carried out by Enertech [158] and supported by Électricité de France [159]. In this measurement campaign, the average electricity needs of domestic appliances (in average 9.8 per household) including lighting and cooking during 10 min time steps have been monitored in $N_{\mathrm{hh}} = 98$ households in France during approximately one year from January 1998 to February 1999.

Table 6.1 shows a general summary of the surveyed appliance types l that were monitored in this campaign. The average electric power consumed at every 10 min time step was recorded in integer units of Wh/10 min during the measurement period of about one year. The diffusion rate D denotes the average percentage the appliances were included in the list of measured appliances in the households.[1] The average percentage of time steps the appliances were in use (consuming a non-zero electric power) is denoted by R. $\overline{P}$ indicates the mean non-zero value of consumed power in Watt, as well as its standard error. To account for appliances which consume power in standby mode whilst not operating, the corresponding values of R^* and $\overline{P^*}$ are based on an arithmetic, where every consumption of up to 18 W = 3 Wh/10 min is considered as standby (being switched off). According to these statistics then, the electric ovens were present in 28 % of the households, consumed in average a non-zero power of 374 ± 657 W during 5.8 % of the time steps, and an average power above 18 W of 1040 ± 741 W during 2.0 % of the time steps. The average user and usage characteristics of the IRISE campaign can be found in the REMODECE database [160], for which general characteristics are provided by de Almeida et al. [161].

In addition to the appliance measurements, the commune where the household is situated, as well as its habitable surface and its size are recorded for respectively 84 %, 83 % and 94 % of the households. The distribution of the latter two over the sample is shown in Figure 6.1.

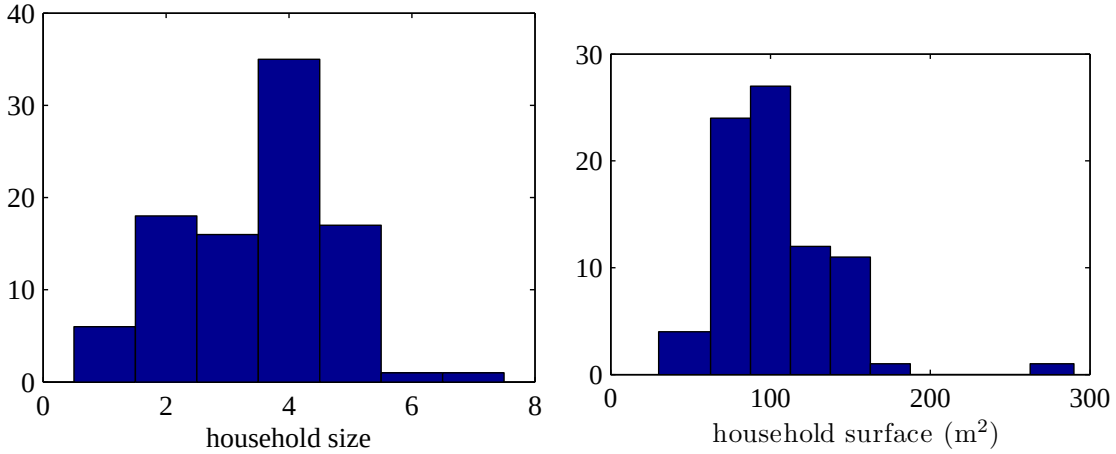
6.2.2 Time-Dependence of appliance use

In the presented modelling approach, the main aspect of electrical appliance use reflecting characteristics of human behaviour lies in the dependence of the probability $p_l(t)$ that the appliance l is in use as a function of time of day, which will be referred to as switched-on profiles in the remainder of this work. Figure 6.2 shows

[1] The list of measured appliances is not exhaustive.

l	appliance type	D	R	$\overline{P}$ (W)	R^*	$\overline{P^*}$ (W)
1	Aquarium	7.3	82.8	69 ± 52	78.6	72 ± 51
2	Chest freezer	30.2	69.6	86 ± 47	63.5	93 ± 43
3	Clothes drier	42.7	4.6	823 ± 834	3.5	1069 ± 804
4	Computer site	2.1	36.9	163 ± 86	36.3	166 ± 85
5	Dish washer	50.0	3.0	1137 ± 977	2.9	1191 ± 968
6	Electric Cooker	30.2	10.5	317 ± 591	4.2	778 ± 729
7	Electric deep fryer	1.0	0.1	1240 ± 721	0.1	1240 ± 721
8	Electric heating	26.0	44.0	902 ± 1079	35.4	1118 ± 1099
9	Electric oven	28.1	5.8	374 ± 657	2.0	1040 ± 741
10	Fridge	97.9	64.4	83 ± 77	55.7	93 ± 77
11	Fridge freezer	61.5	71.0	93 ± 85	61.9	104 ± 85
12	Halogen lamp	116.7	4.8	149 ± 152	4.0	173 ± 154
13	Heat pump	3.1	64.4	581 ± 750	40.3	921 ± 770
14	Heat pump water heater	1.0	51.3	497 ± 168	50.9	501 ± 162
15	Hot plate	10.4	2.4	657 ± 538	2.3	671 ± 535
16	Microwave oven	78.1	4.7	199 ± 244	4.4	213 ± 247
17	Non-Halogen lamp	47.9	4.2	54 ± 40	3.8	58 ± 39
18	TV	133.3	23.8	55 ± 44	18.1	68 ± 44
19	Vertical freezer	28.1	64.6	86 ± 63	58.0	94 ± 62
20	Washing machine	88.5	5.7	452 ± 660	5.5	467 ± 666
21	Washing machine+clothes drier	8.3	3.4	496 ± 633	3.1	538 ± 642
22	Water heater	36.5	14.5	1568 ± 985	14.4	1575 ± 982
23	Water pump	1.0	40.9	306 ± 171	40.3	310 ± 169

Table 6.1: Appliance types l in the IRISE dataset, with the diffusion rate per household D, the mean percentages of being switched on (non-zero electricity use) R (both in percent) and the mean power consumption $\overline{P}$ whilst being on. In the values of R^* and $\overline{P^*}$, values below 18 W are not considered as switched on (standby).



(a) (b)

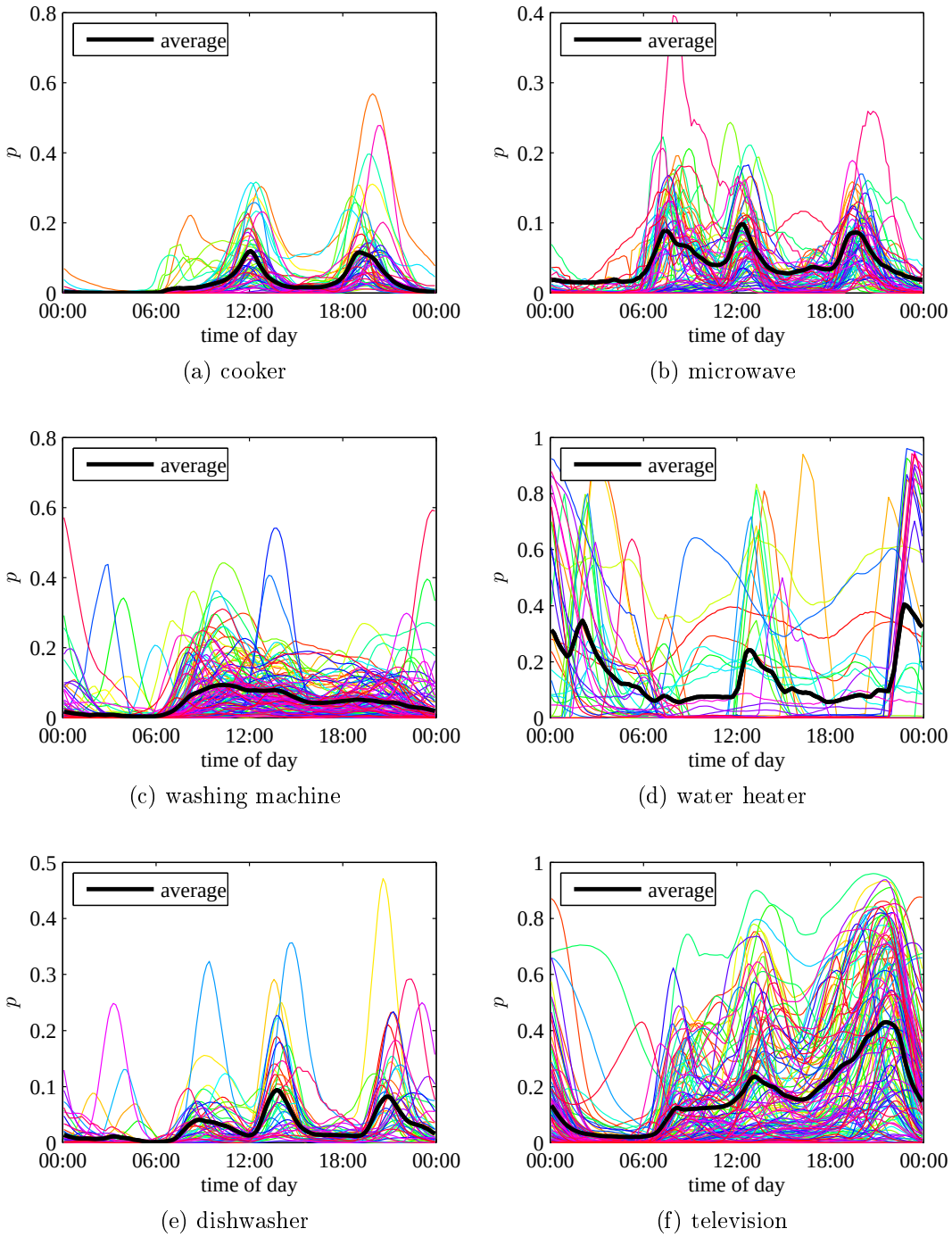
Figure 6.1: Distributions of recorded characteristics of the households in the IRISE dataset.

the switched-on profiles of cookers, microwave ovens, washing machines, water heaters, dishwashers and televisions. Here, the appliance types $l \in \{6, 9, 15\}$ (cf. Table 6.1) were merged together to the generic term of cookers, and $l \in \{3, 20, 21\}$ to washing machines. In these graphs, the average profiles over the measurement period are shown for the individual households (in colour), as well as the average between all households (in black). The illustrated curves were smoothed using a moving average over five time steps[2]. The profiles depend significantly on the time of day, and show a large variety among different households. However, there are common patterns of usage behaviour like, for instance, the peaks around lunch and in the evening for the cooker and microwave oven profiles.

6.2.3 Power demand of appliances

Besides the switched-on profiles, the characteristic electric power demand of electrical appliances is of central importance to predict residential load profiles. As most of the appliances do not have a constant power need, their consumption can be characterised by a distribution. In general this distribution depends on the considered appliance, but as the list of measured appliances (cf. Table 6.1) is too small to be considered as representative for the variety in reality, the approach will be based on uniform distributions for each appliance type.

[2]Likewise, in Chapter 3, the profiles are defined on a 24 h basis, and thus, the moving average was generated in a cyclic manner based on the modulo operation with a divisor given by the number of time steps in one day.



(a) cooker (b) microwave

(c) washing machine (d) water heater

(e) dishwasher (f) television

Figure 6.2: Individual households' (coloured) and average (thick line) switched-on profiles of a selection of appliances.

Variety of power consumption

Figure 6.3 illustrates the empirical probability distribution functions (EPDFs) of power consumption whilst being switched on of the appliance types in Figure 6.2.

The EPDFs were derived in the same manner as R^* and $\overline{P^*}$ in Table 6.1, meaning that power values below 18 W were considered as parasite consumption (standby). This lowers the magnitude of the bins of the smallest power consumption. To illustrate the variety among individual appliances, their respective contribution to the bins of Figure 6.3 is schematised by different colours. This shows that the EPDFs of different cookers or washing machines are relatively similar to each other, whereas those of televisions or water heaters tend to have a very different relative colour gradation. In other words, the energy consumption characteristics of cookers depend less on individual characteristics of the appliance or the household than those of televisions or water heaters. In particular, the latter two types of appliance tend to have a constant electric power consumption in operating mode, which is only dependent on the considered appliance.

Time-dependence of power consumption distributions

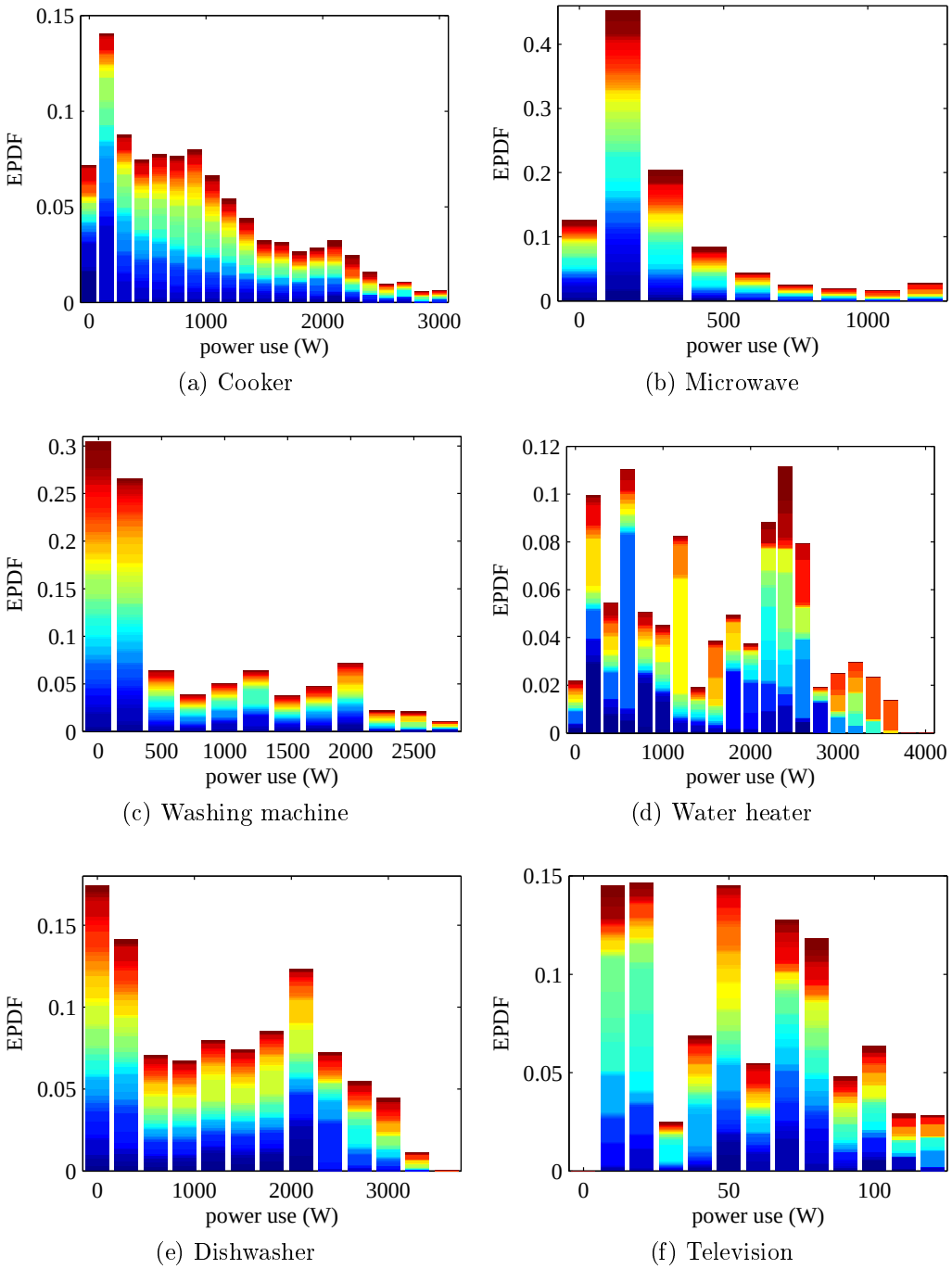
The distribution of the electric load profile $f_{P_{\mathrm{el}}}(t)$ of an appliance l depends on the appliance's switched-on profile $p_l(t)$, as well as on $f^*_{P_{\mathrm{el}}}(t)$, its time-dependent power consumption distribution whilst being in use:

$$f_{P_{\mathrm{el}}}(t) = (1 - p_l(t))\ \delta_0(t) + p_l(t)\ f^*_{P_{\mathrm{el}}}(t). \tag{6.1}$$

In order to study the variations of the distribution of power consumption $f^*_{P_{\mathrm{el}}}(t)$ whilst being in use as a function of time of day, the corresponding EPDFs of measured power demand were derived for several appliance types conditional on being in use. However, the measurements of power demand per 10 min time step contain corrupt records. To illustrate this, a selection of quantiles of $f^*_{P_{\mathrm{el}}}(t)$ of televisions is shown in Figure 6.4. Although the mean power consumption of a common television does not exceed a few hundred Watt (cf. Table 6.1), there are recordings of over 11 kW in the dataset. However, they occur less frequently than 1 ‰ of the time steps. Therefore, the islands of these aberrant recordings were replaced by the mean of the two adjacent values that do not exceed an unrealistic threshold. The resulting distributions of power consumption $f^*_{P_{\mathrm{el}}}(t)$ are shown for the list of appliance types of Figure 6.5. The thresholds over which measurements were considered as corrupt were individually chosen for every appliance, and are shown by the maximum value on the y-axes of the appliances in each graph of Figure 6.5.

As one would expect, $f^*_{P_{\mathrm{el}}}(t)$ does not vary significantly as a function of time of day for dishwashers, washing machines and televisions, because the operating mode of these appliances is not influenced by human behaviour or other daily patterns of any kind.[3] In contrast, the distributions of cookers and microwave ovens are significantly increased around noon and 7 pm. This might be explained

[3]The stronger fluctuations in the small hours are related to statistically insignificant data, due to a less frequent use in this period.



(a) Cooker (b) Microwave

(c) Washing machine (d) Water heater

(e) Dishwasher (f) Television

Figure 6.3: Empirical probability distribution function of power consumption. Every colour represents an individual household, illustrating its contribution to the different bins of the distribution.

by the greater likelihood to prepare food in these periods, which is more likely to involve the use of multiple hot plates of the cooker, or the operation of the microwave during longer periods. The rest of the day the use of these appliances are more dominated by short usage, as for instance, to heat up a small portion of water, and thus, the power demand is distributed around smaller values in these periods [cf. 144].

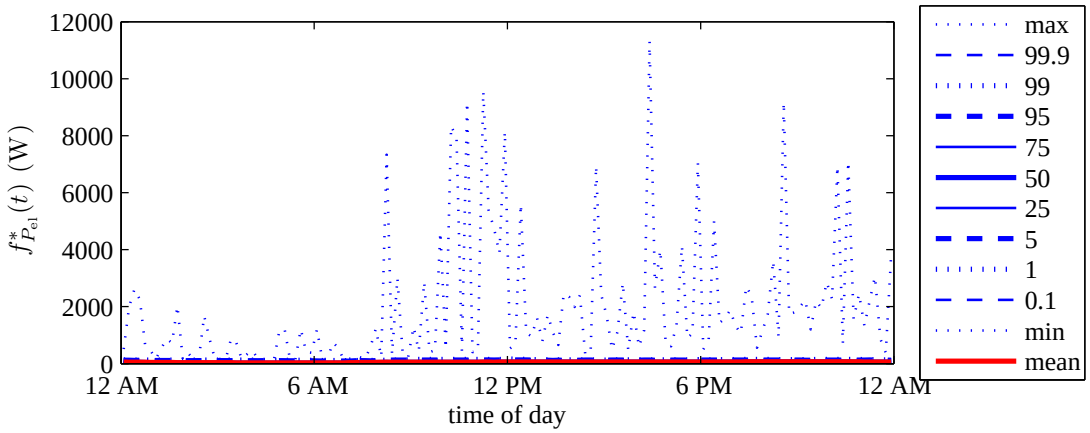


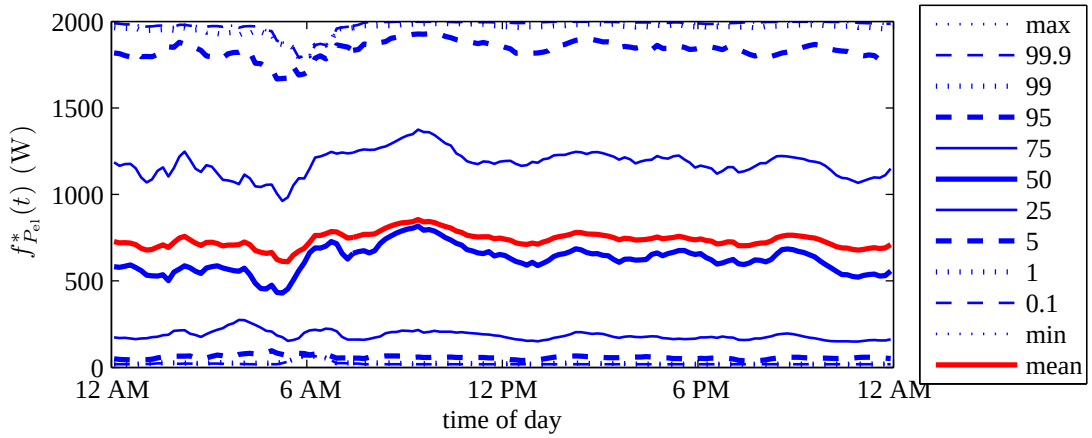
Figure 6.4: Time-Dependent power consumption EPDF of televisions whilst being in use $f^*_{P_{el}}(t)$ without data cleaning.

Total power consumption

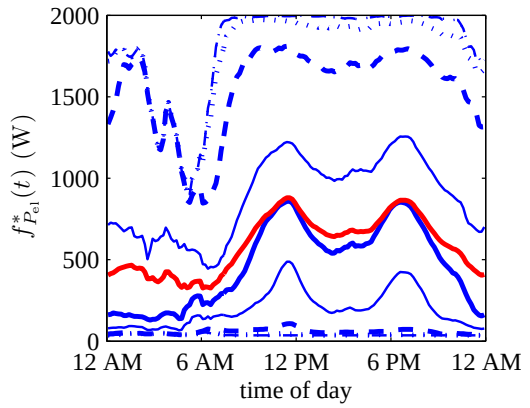
In addition to the measurements of individual appliances, the total power consumption $P^i(t)$ of each household i was also recorded at every 10 min time step. However, the list of individually measured appliances in the households does not include the entire appliance stock (cf. Section 6.2.1), and furthermore, the measured power in an alternating current circuit may not necessarily represent the real consumed power but be distorted by reactive power. Therefore, the value of the total site power consumption does not correspond to the sum of the consumption values of the individual appliances.

Figure 6.6 shows the time-dependent distribution of the total electric power consumption $f^{tot}_{P_{el}}(t)$ of the ensemble of measurements $P^i(t)$ in all households. In Figure 6.6a, the distribution is shown on a logarithmic scale as a function of time of day.[4] The constant minimum of 6 W corresponds to the smallest non-zero

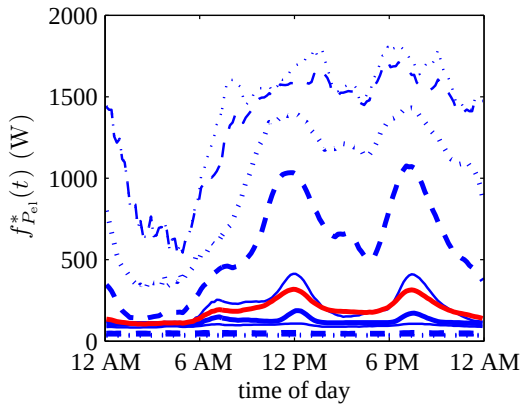
[4] In contrast to the previous section, the displayed distributions are not conditional on "being in use", as the total power consumption is very unlikely equal to zero. However, in the measurements there were occurrences of periods of zero power consumption, which were removed as this is very unrealistic (regarding standby consumption and appliances constantly in use, e.g., fridges) and probably rather due to a malfunction of the metering unit. Thus, the calculation of these EPDFs was identical to that of $f^*_{P_{el}}(t)$ in Figure 6.5.



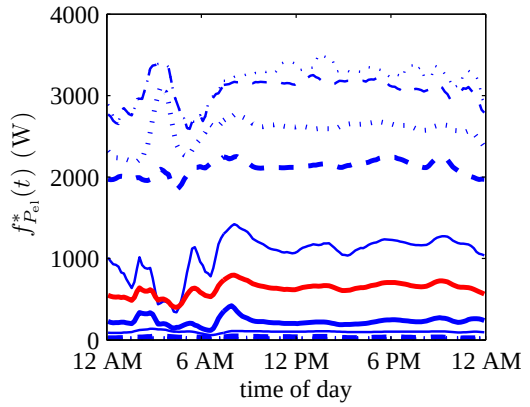
(a) Dishwasher



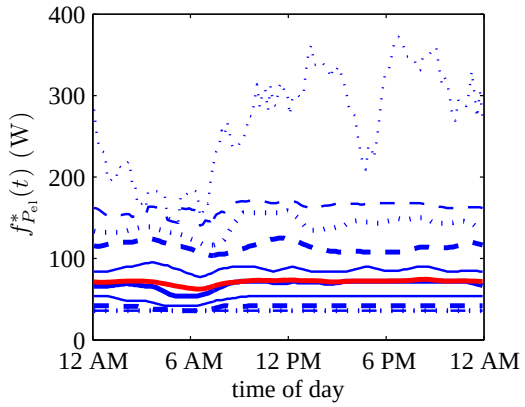
(b) Cooker



(c) Microwave



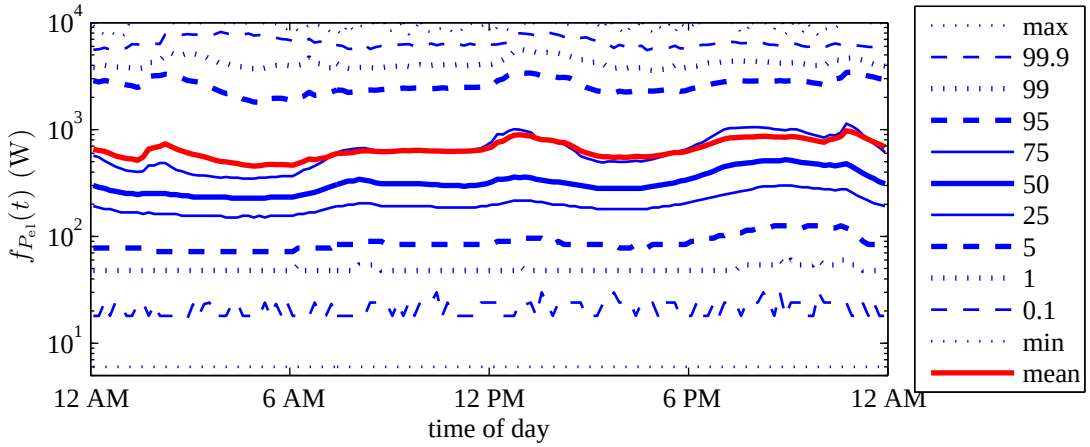
(d) Washing machine



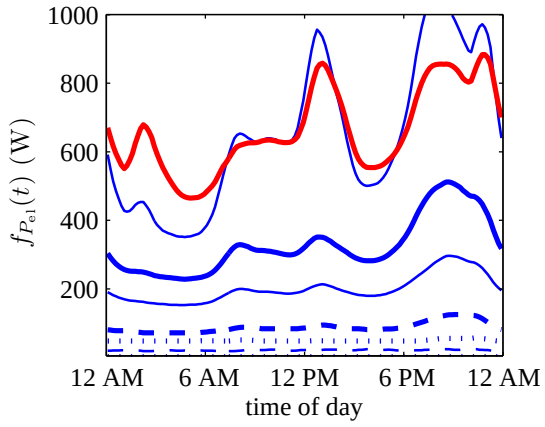
(e) Television

Figure 6.5: Appliances' time-dependent EPDF of power consumption $f^*_{P_{el}}(t)$ of cleaned data, whilst being in use (disregarding standby).

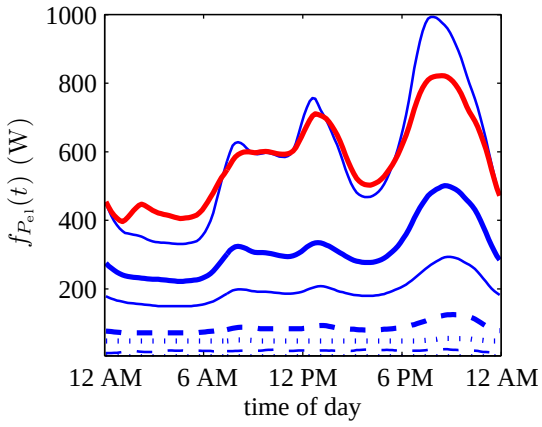
value. However, the position of the first percentile, which is almost constantly at about 50 W, shows that the smaller values can be neglected. The time-dependent distribution is not shown for values above 10 kW, as these are also negligible. In Figure 6.6b, the values of the same distribution are shown below 1 kW on a linear scale. As there is a relatively high share of electric water heaters in the sample (cf. Table 6.1), in Figure 6.6c, the distribution was also derived for the difference of the total power consumption and that of the water heater in case of existence in the household. This distribution is very similar in shape to those of other published research [e.g., 148, 162, 163].



(a) Selected percentiles of the total power consumption.



(b) Total power consumption



(c) Without water heater

Figure 6.6: EPDF of total household power consumption (a,b), and without water heater (c).

Peak and base load

The electric load profiles of households can be separated into peak and base load, which is of particular interest in the field of electricity supply management. It will be shown that, on the demand side, the two different parts can be derived from the load profiles of the different types of appliances. Regarding residential load profiles, a substantial part of the base load is generated by a category which will be referred to as "stuff", which consists of various individual appliances for which it would be very tedious to model them individually [cf. 27]. However, it will be shown that the stuff corresponds to an approximately constant statistical offset. To illustrate this, the average load profiles $\overline{f_{P_{\mathrm{el}}}}(t)$ of the different appliances were derived, which are illustrated in Figure 6.7. These curves correspond to the time-dependent product of the switched-on profiles and the average power consumption whilst being switched on. The power consumption value of stuff was derived by subtracting the sum of consumption profiles of all individually monitored appliances in the household from that recorded for the total site (which should theoretically always be greater or equal than zero, due to the non-measured individual appliances. The figure demonstrates that the average power consumption of the category stuff as a function of time of day is approximately constant.

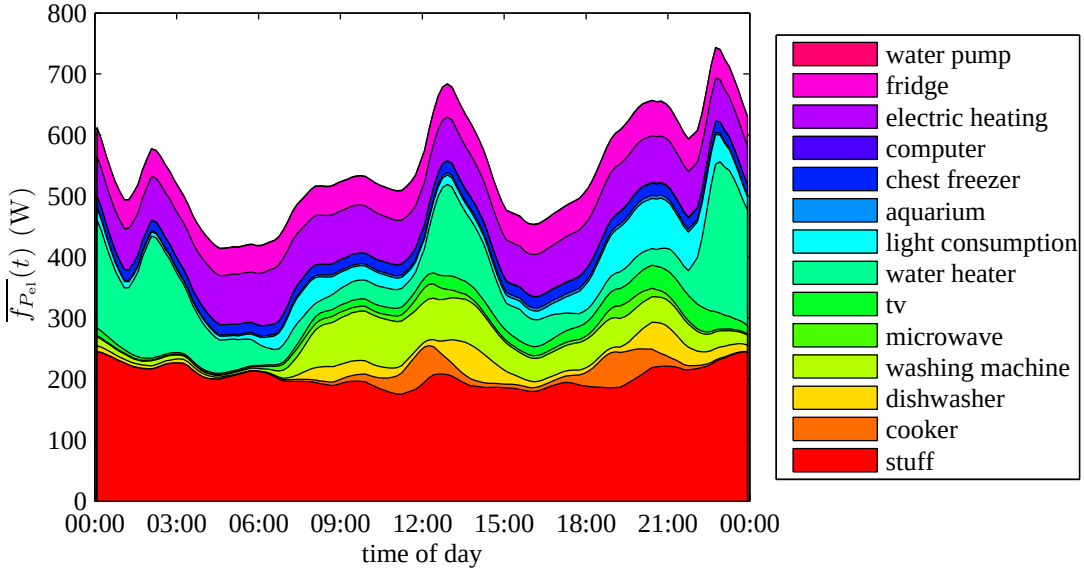


Figure 6.7: Time- and appliance-dependent mean power consumption.

The distribution $f_{P_{\mathrm{el}}}(t)$ of the category stuff is shown in Figure 6.8 for power values of up to 600 W, which are exceeded by 7.0 % of the distribution. The mentioned inconsistencies of the power metering lead to a share of 4.8 % of zero or negative values. Furthermore, the time-dependence of the distribution of stuff $f_{P_{\mathrm{el}}}(t)$ is illustrated in the sub-plot by the corresponding percentiles, where the

deciles and the average are printed with thicker lines (see legend). This shows that the distribution of this statistical offset is also approximately constant throughout the day.

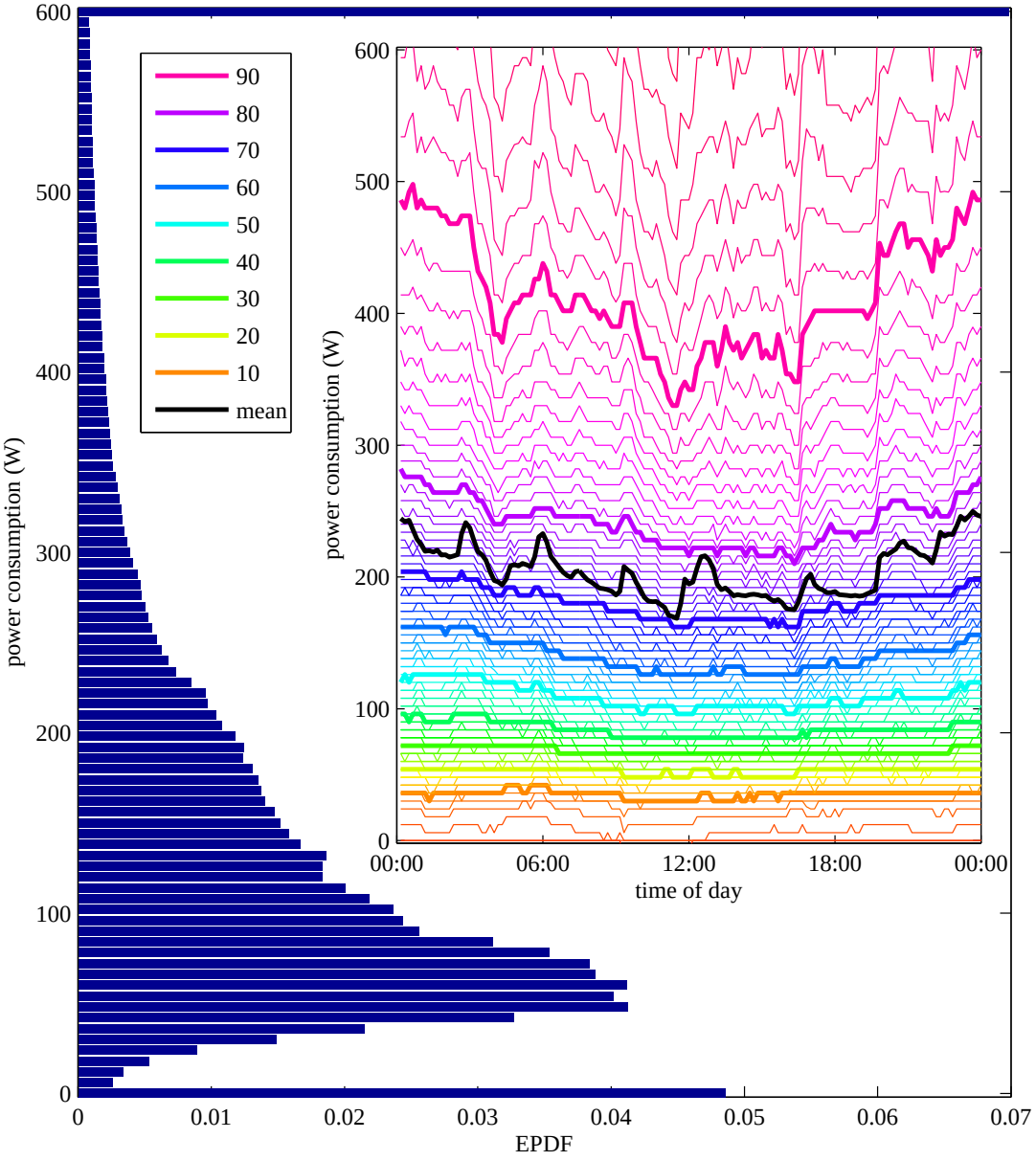


Figure 6.8: Distribution of power consumption of stuff between 0 and 600 W. In the sub-plot, the quantiles of this distributions are shown as a function of time of day, with deciles and the average printed with thick lines (see legend).

6.3 Activity-Dependent electrical appliance use

In order to predict the use of electrical appliances in dependence of specific residential activities, we infer the conditional probabilities that an appliance is in use whilst an activity is performed. Here, we restrict ourselves to appliances where human behaviour bears an immediate temporal impact on their power consumption[5]. Furthermore, the power consumption of water heaters, as well as PCs, deep fryers, heat pump water heaters and water pumps have not been measured in many households (see Table 6.1), and thus do not allow for a meaningful statistical evaluation.

The electric power consumption loads of electric heatings and freezers rather depend on seasonal influences than on residential activities and will thus not be treated in this work. A detailed model to predict the loads of freezers has already been developed by Richardson et al. [23].

Although the use of electric lighting is only related to human behaviour in residential environments where an automatic control system is usually not installed (even less likely at the time when the measurements of the IRISE campaign were conducted), the description of the corresponding stochastics cannot be treated solely by means of the probabilities to be present and/or to perform activities. Instead, published research shows that, besides (the change of the state of) active occupancy, manually controlled lighting patterns in buildings also depend on irradiation, and thus on seasonal daylight availability, facade orientation and the blind settings. Therefore, the modelling of lighting will not be treated in this work, as in the IRISE dataset, much of this information is not included, and furthermore, published models are relatively detailed and were already validated [63, 164].

6.3.1 Modelling approaches

The probabilities to perform activities[6] $p_j(t)$ which are of relevance for residential appliance use are not related to single individuals but rather to the ensemble of all household members. Therefore, these probabilities were derived from the TUS, depending on whether there is at least one member of the household (defined by the variable "hldid", see Table 2.1) performing the activity. The probability

[5]The water heater is also influenced by human behaviour, where peaks occur after lunch (probably due to dishwashing) and in the evening (probably due to personal hygiene); but also during the night, which might rather be related to technical settings (unfortunately, not known for this dataset) and delayed in time with respect to involved activities (see Figure 6.7). Therefore, this appliance will not be treated in this analysis.

[6]In contrast to the probabilities to perform an activity j conditional on being at home (cf. Chapter 4), the ordinal probabilities to perform an activity have to be considered, given by the product of the probability to be at home (cf. Chapter 3, and Figure 3.14) with the conditional probability to perform an activity whilst being at home (cf. Chapter 4, and Figure 4.5).

profiles that appliances are in use, as well as those to perform activities[7] were smoothed (cf. Section 6.2.2). Regarding the activity profiles, this was done to level out temporal artefacts in the diary plans (cf. Chapter 4). The appliance switched-on profiles were smoothed, due to the occurrence of significant peaks, which – we assume – are related to the behavioural specificities of the non-representative number of 98 households of the IRISE dataset.

Appliance use whilst performing activities

In order to derive $p(l|\{j\})$, the probabilities that an appliance l is used conditional on the set of performed activities $\{j\}$, regression analysis was carried out, based on the two time series of $24 \cdot 6 = 144$ data points of $p_l(t)$, the switched-on profile of the appliance l, and $p_j(t)$. A simple approach to calibrate $p(l|\{j\})$ could be based on multiple linear regression

$$p_l(t) = \beta_{l0} + \sum_{j}^{N} \beta_{lj}\ p_j(t), \tag{6.2}$$

where β_{lj} represents the conditional probability that the appliance l is in use while the activity j is performed, and β_{l0} corresponds to the probability that l is on, regardless of any activity.

In general, it is desirable to identify $p(l|\{j\})$ as a function of all activities $\{j\}$ that may involve the use of the appliance l, as well as of time t. However, the predictions of $p_l(t)$ in the linear regression model of Equation (6.2) are not bounded between 0 and 1.

As there are furthermore only 144 available data points for the regression model of each $p_l(t)$, this leads to an increasing risk of overfitting with increasing number of activities in $\{j\}$. This also has the negative side-effect to obtain significant regression coefficients whose magnitude is much greater than 1, which does not have a physical meaning, and would possibly issue from the correlation between predictions. As there are many activity-appliance pairs which do obviously not have a identifiable relevance to each other, the regression should be restricted to activities with a clear relationship to the appliance profile of l. To eliminate these inconsistencies, another linear regression model was tested:

$$\mathrm{logit}(p_l(t)) = \beta_{l0} + \beta_{lj}\ \mathrm{logit}(p_j(t)), \tag{6.3}$$

where the logit of $p_l(t)$ is fitted against the logit of $p_j(t)$ of only one activity j. It follows that

$$p_l(t) = \left[\mathrm{e}^{-\beta_{l0}} \left(\frac{p_j(t)}{1 - p_j(t)} \right)^{-\beta_{lj}} + 1 \right]^{-1}. \tag{6.4}$$

[7] As the time frame of the diary plans of the French TUS ranges from midnight to midnight, the temporal information during Central European Summer Time (daylight saving time) had to be adapted to align this information with the appliance measurements.

appliance	β_{l0}	β_{lj}
Cooker	-0.77 ± 0.29	0.97 ± 0.07
Dishwasher	-4.13 ± 0.09	0.42 ± 0.02
Washing machine	-1.63 ± 0.11	0.50 ± 0.03
Microwave	-1.76 ± 0.09	0.35 ± 0.02
Television	-2.06 ± 0.08	0.84 ± 0.07

Table 6.2: Results of the linear regression models of Equation (6.3).

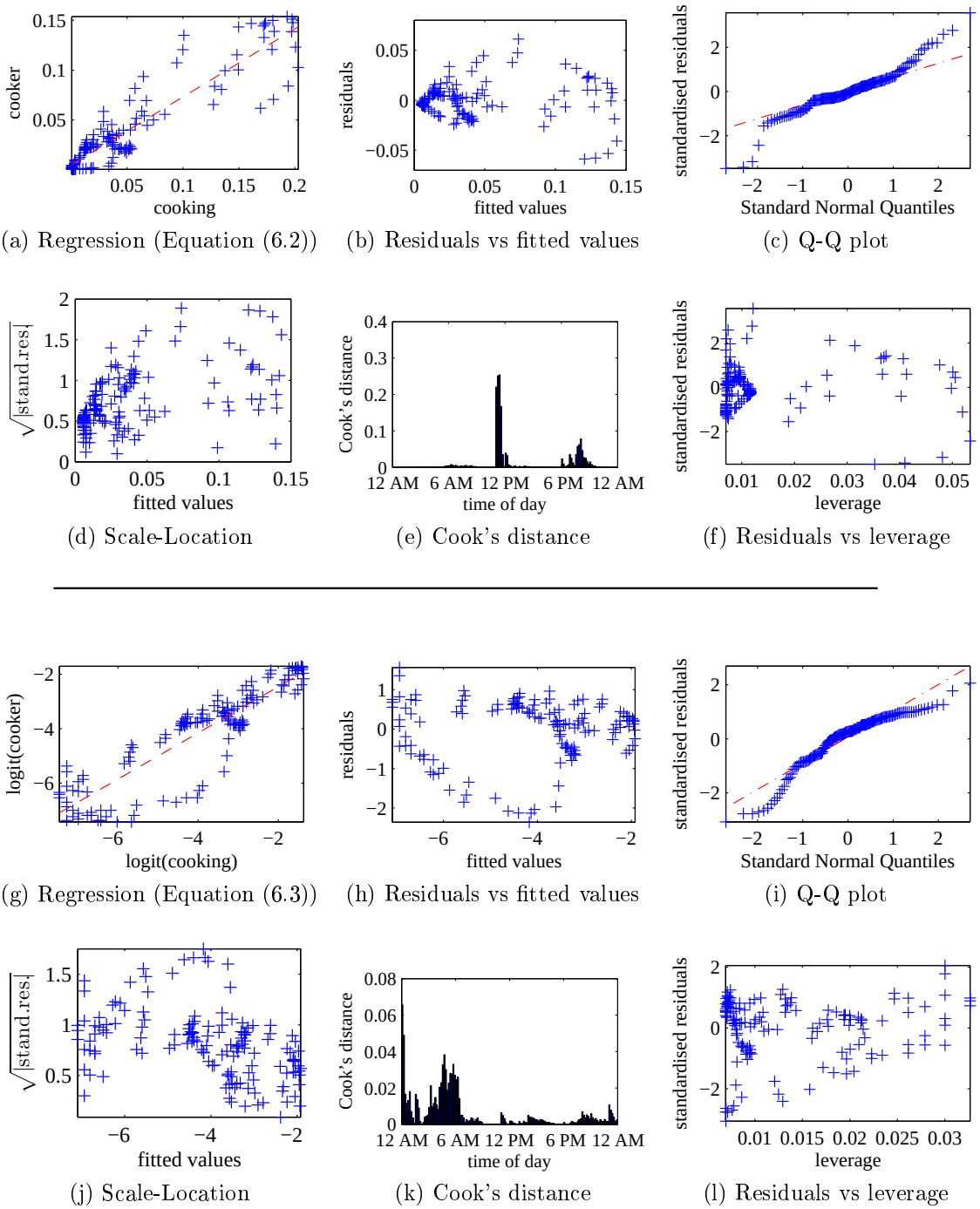
In Figure 6.9, we show regression diagnostics of the two linear regression models of Equations (6.2) and (6.3), of the time-dependent probabilities that the cooker is on.[8] In Figures 6.9a and 6.9g, the results of the regression according to Equations (6.2) and (6.3) are shown. The comparison of the residuals and the square root of the magnitude of the standardised residuals of the regression against the fitted values in Figures 6.9b and 6.9d shows that the assumption of homoskedasticity is violated in Equation (6.2), as the variance of the residuals tends to increase with the fitted value. Furthermore, the Q-Q plot shows evidence that the distribution of the standardised residuals is heavy tailed, which represents another violation regarding the necessary assumptions for linear regression. This logit transformation thus corrects the issues of heteroskedasticity, heavy tails of residuals and influential observations.

The distribution of the standardised residuals of the regression model of Equation (6.3) is left-skewed (see Figure 6.9i) and thus also has a heavy left tail which is however less accentuated. However, the assumption of homoskedasticity is better supported by this model (cf. Figures 6.9h and 6.9j). The Cook's distances of the data points in the regression model, as well as their leverages are smaller in the model of Equation (6.3) than in that of Equation (6.2) (see Figures 6.9e, 6.9f, 6.9k and 6.9l).

As the statistical analysis showed that the linear regression of the logit of the probabilities, defined by Equation (6.3), is preferable, this model will be used in the following. The best results were yielded by defining the probabilities that at least one household member is performing the considered activity, for the reasons that were explained in the beginning of Section 6.3.1. In Figure 6.10, the results of this regression analysis (according to Equation (6.4)) are shown for the pairs of l and j which corresponded most closely to the observations. The corresponding parameter values β_{l0} and β_{lj} are shown in Table 6.2.

In Figure 6.10a, we show the linear regression of the switched-on profile of

[8]Regarding the switched-on profile of the cooker, the most related activity j in the TUS is evidently food preparation. Unfortunately, this activity was merged with washing up and putting away dishes in the "MTUS 41-activity typology" (*cf.* Figure 2.7). Therefore, the code of thee 69-category typology "food preparation, cooking" was used, which is also recorded in the database [31].



(a) Regression (Equation (6.2)) (b) Residuals vs fitted values (c) Q-Q plot

(d) Scale-Location (e) Cook's distance (f) Residuals vs leverage

(g) Regression (Equation (6.3)) (h) Residuals vs fitted values (i) Q-Q plot

(j) Scale-Location (k) Cook's distance (l) Residuals vs leverage

Figure 6.9: Regression diagnostics for (a-f) the linear regression, and (g-l) the regression with the logit transforms of the variables.

cookers with "food preparation/cooking". The peaks of the switched-on profile at lunch and in the evening are underestimated, whereas it is overestimated in the

morning. This might be due to the different character of preparing food at lunch or in the evening, when it is more likely to prepare a warm meal than the rest of the day, whereas in particular early in the morning the use of the cooker is very unlikely when preparing food. This shows that the conditional probability to use a cooker whilst preparing food is time-dependent. Thus the assumption of a constant $p(l|j)$ leads to errors in the predictions, which are however not substantial.

The best-matching regression result to predict the use of the microwave oven was found with the probability profile of starting "meals and snacks" (cf. Figure 2.10), which is shown in Figure 6.10b. This can be interpreted by the short mean duration of the use of a microwave oven, which often occurs immediately before starting a meal. In contrast to the use of a cooker whilst preparing food, the conditional probability of using a microwave oven when starting to eat is higher in the morning than at lunch or in the evening, and thus resulting in the inverse pattern of over- and underestimation than in Figure 6.10a.

The regression predicting the use of washing machines conditional on doing housework is shown in Figure 6.10c.[9] Here, the observed switched-on profile between 7.30 AM and 2 PM is underestimated, whereas it is overestimated in the evening, which might also be due to a time-dependent conditional probability to use the washing machine whilst doing housework. Furthermore, the probability to switch on the washing machine before the night is not well captured and thus underestimated by the model.

In Figure 6.10d, we show the switched-on profile of televisions, as well as the regression resulting from active occupancy (at home and not sleeping). Interestingly, the regression with the activity "watching TV" yields very poor results. This might be related to the fact that watching television is not well represented as a primary activity and is very often performed as a secondary activity. Furthermore, it might be very common that the television is switched on whilst being at home, and often not switched off unless people go to sleep or leave.

Modelling starts and durations of appliance use

In Figure 6.10, evidence is shown that the modelling of the time-dependent use of the appliances l conditional on an activity j that is performed yields relatively accurate results. However, there is no profile of performing or starting an activity j in the TUS with a shape that approximately coincides with the switched-on profile of dishwashers. The influence of human control on dishwashers is comparable with that on washing machines, in the sense that only the starting time

[9] This yielded much better results than the activity "laundry, ironing, clothing repair", which is also specified in the 69-category typology [31], probably due to the fact that the use of a washing machine is usually done in parallel to other activities which can be very manifold and are thus not well represented by the very specific primary activity "laundry, ironing, clothing repair".

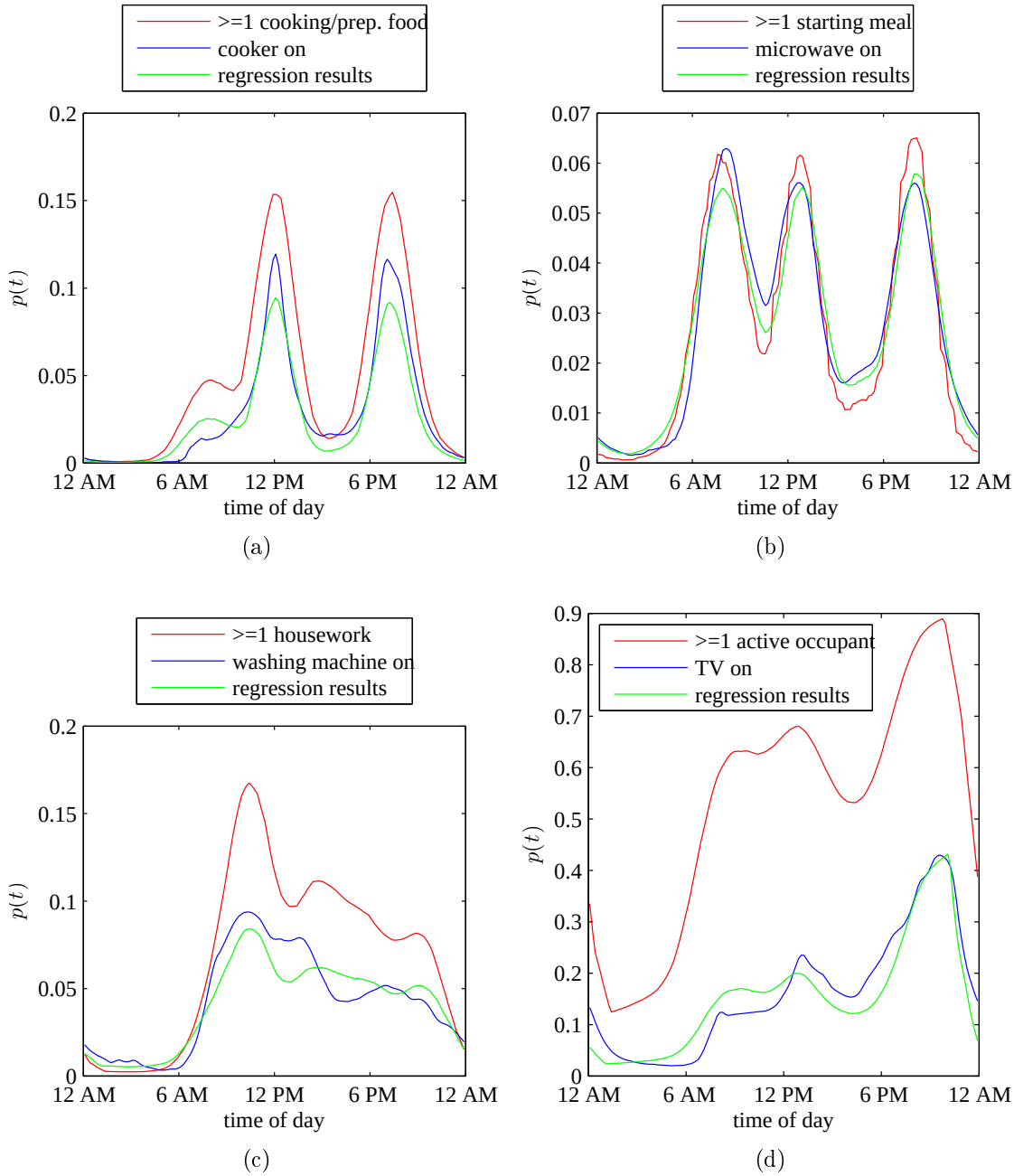


Figure 6.10: Regression results of the activity-dependent use of electrical appliances defined by Equation (6.4).

is influenced and afterwards, the appliance is in use according to a technically pre-set operation mode. However, after the washing cycle the laundry has to be hang out to dry without much delay, in order to avoid malodorous smell, whereas such a constraint is not given with respect to the use of a dishwasher. Therefore,

the probability that a washing machine is in use is more correlated to residential activity profiles, whereas regarding dishwashers, it is rather the probability to start them which is correlated. Thus, the best results of the regression of Equation (6.4) was achieved with the probability "set table, wash/put away dishes"[10] and the time-dependent probability to switch on the dishwasher $p_{l,\mathrm{s}}(t)$, which is shown in Figure 6.11b.

The empirical duration PDF for dishwashers being on $f(t)$, as well as its corresponding survival function $S(t)$ are shown in Figure 6.11a, which do not significantly depend on the time of day. Knowing $p_{l,\mathrm{s}}(t)$ and $S(t)$, the switched-on profile is determined as the superposition of every fraction of usage starts in every time interval $[t, t+\mathrm{d}t]$ which remain switched on at a later time $t+t'$ given by the survival function $S(t')$. Considering the discretised representation of $p_{l,\mathrm{s}}(t)$ and $S(t')$, the switched-on profiles $p_l(t)$ are then given by the circular discrete convolution

$$p_l(t) = S(t) * p_{l,\mathrm{s}}(t) \equiv \sum_{t'=0}^{T-1} \left(\sum_{t''=0}^{\infty} S(t' + t'' \cdot T) \right) p_{l,\mathrm{s}}(t - t'), \tag{6.5}$$

where T is the period, i.e. the number of time steps of one day. As $S(t)$ decreases to zero within about $t_{\max} = 2\ \mathrm{h} < T$ (cf. Figure 6.11a), Equation (6.5) simplifies to

$$p_l(t) = \sum_{t'=0}^{t_{\max}} S(t')\ p_{l,\mathrm{s}}\left((t - t')_{\mathrm{mod}\ T}\right), \tag{6.6}$$

where the argument of $p_{l,\mathrm{s}}$ is taken modulo T.

The result of Equation (6.6), using $S(t)$ of Figure 6.11a and $p_{l,\mathrm{s}}$ derived from Equation (6.3) (cf. Figure 6.11a and Table 6.2) is shown in Figure 6.11c by the green line. The non-smoothed switched-on profile is shown by the blue line, as well as its moving average of 11 time steps. The latter is in good agreement with the observed switched-on profile.

6.4 Modelling of residential load profile distributions

In this section, a methodology will be presented to construct the load profile distribution as a result of the profiles of individual appliances. This bottom-up methodology is based on the switched-on profiles of individual appliances (see Section 6.2.2), as well as their power demand distribution whilst being in use (see Section 6.2.3). As in this work, it is focussed on the use of appliances which are directly operated by household members, the set of investigated appliances will be restricted to the five that were treated in Section 6.3.

[10] Probability that at least one household member performs the activity "set table, wash/put away dishes" of the 69-activity typology [31].

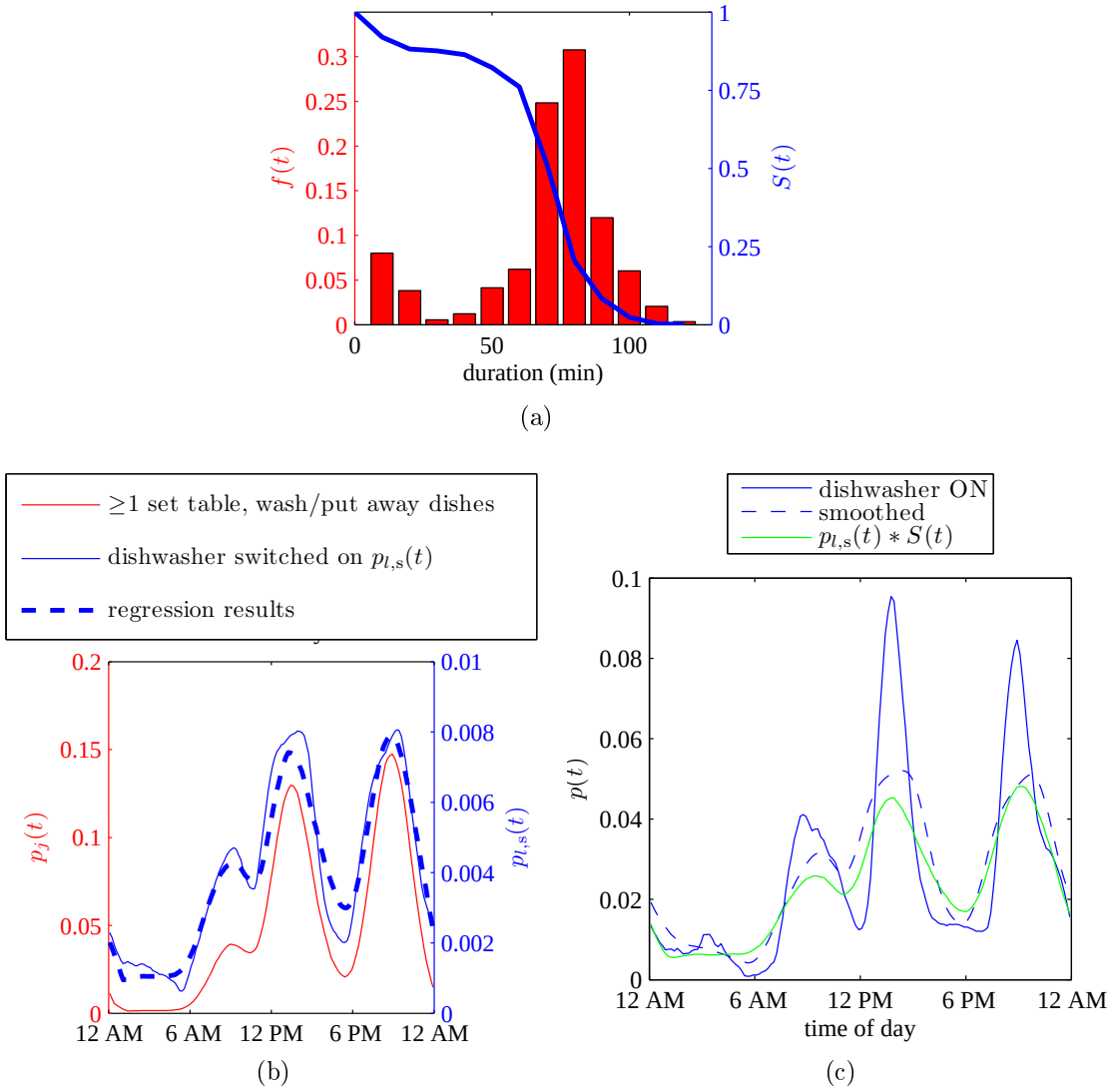


Figure 6.11: (a) Empirical duration PDF $f(t)$ (red) and the corresponding survival function $S(t)$ for dishwashers (blue). (b) Activity profile $p_j(t)$ (red), in blue the probability that the dishwasher is switched on $p_{l,s}(t)$, as well as regression results (dashed; cf. Equation (6.4)). (c) In blue, switched-on profile of the dishwasher and smoothed curve (dashed), as well as the convolution (cf. Equation (6.6)) of the regressed starting probability $p_{l,s}(t)$ with the duration survival function $S(t)$ (green).

6.4.1 Aggregated load profile distribution of multiple appliances

Figure 6.13 shows the monitored mean profiles $\overline{f_{P_{el}}}(t)$ (cf. Figure 6.7) of the subset of these five appliances. As it was mentioned, the measurements of the individual

appliances' power consumption in the IRISE dataset was not exhaustive. This implies that it is unknown, whether the appliance was present in a household, although non-existing in the list of measured appliances. Figure 6.12 illustrates how often the 45 different combinations of measured appliances occurred in the households of the dataset. According to these statistics, in the list of measured appliances of $K(\nu = 5) = 10$ households, there was one television, one microwave oven and one washing machine, and in $K(\nu = 37) = 5$ households, there were two televisions, and one instance of the remaining four appliances.

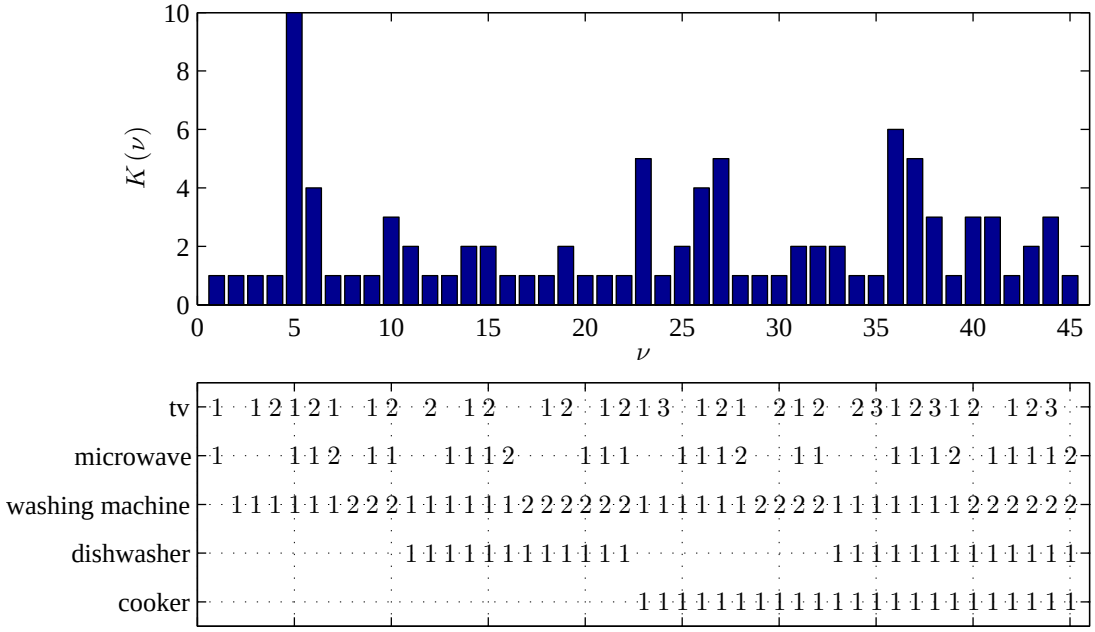


Figure 6.12: Number of households $K(\nu)$, where the corresponding combination ν of individual appliances was monitored.

To construct the aggregated power consumption distribution, when the five appliances are simultaneously used, we assumed that the power consumption distributions, as well as the switched-on profiles of the individual appliances are independent of each other. Then, it follows that the distribution of the sum of the power consumption of multiple appliances $f^{\mathrm{m}}_{P_{\mathrm{el}}}(t)$ can be calculated by means of convolution [84].[11] However, in this case, the distribution of a single appliance l is not represented by the distribution whilst being in use $f^{*}_{P_{\mathrm{el}}}(t)$ (cf. Figure 6.5), but by the distribution regardless whether being in use $f_{P_{\mathrm{el}}}(t)$. The latter (approximated by its corresponding discretised probability mass function) is given by the value of the switched-on profile $p_l(t)$ at the value of 0 W and by $(1 - p_l(t)) \cdot f_{P_{\mathrm{el}}}(t)$ for non-zero values (cf. Equation (6.1))[12].

[11] In case of dependent random variables, the distribution of the sum of them depends on the

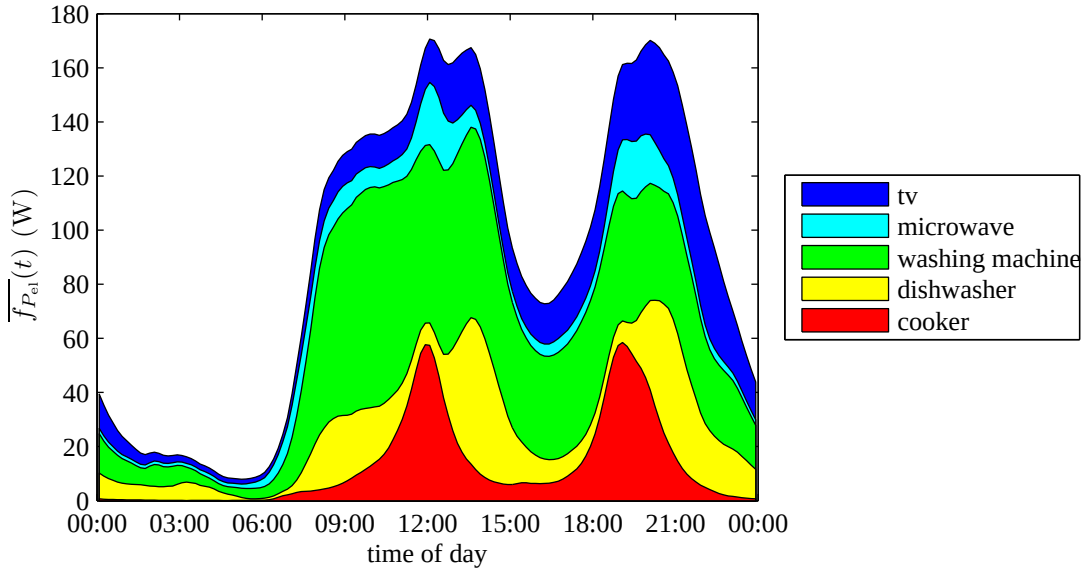


Figure 6.13: Mean power consumption profiles $\overline{f_{P_{\mathrm{el}}}}(t)$.

In this way, the distribution of the overall power consumption can be calculated for each combination of appliances ν. For instance, for $\nu = 4$ it follows that

$$f^{\mathrm{m}}_{P_{\mathrm{el}}}(\nu = 4, t) = f^{\mathrm{TV}}_{P_{\mathrm{el}}}(t) * f^{\mathrm{TV}}_{P_{\mathrm{el}}}(t) * f^{\mathrm{WM}}_{P_{\mathrm{el}}}(t), \tag{6.7}$$

where $f^{\mathrm{TV}}_{P_{\mathrm{el}}}(t)$ and $f^{\mathrm{WM}}_{P_{\mathrm{el}}}(t)$ respectively denote the load profile distribution of a television and a washing machine, regardless whether being in use. To calculate the distributions of the different combinations with Equation (6.7), it was assumed that only the distribution of cookers and microwaves are time-dependent (cf. Figure 6.5). The distribution $f^{\mathrm{tot}}_{P_{\mathrm{el}}}(t)$ over the entire sample of households is then given by

$$f^{\mathrm{tot}}_{P_{\mathrm{el}}}(t) = \sum_{\nu=1}^{45} K(\nu)/N_{\mathrm{hh}} \cdot f^{\mathrm{m}}_{P_{\mathrm{el}}}(\nu, t) \tag{6.8}$$

To validate this model, its results were compared with the monitored distribution $f^{\mathrm{tot,m}}_{P_{\mathrm{el}}}(t)$. For this purpose, the sum of the power consumption of the five appliances was calculated for each household as a function of the measured time, from which the measured load profile distribution of the five appliances was deduced. The time-dependent value of $(1 - f^{\mathrm{tot,m}}_{P_{\mathrm{el}}}(t) = 0)$ corresponds to the mean observed probability profile that at least one of the five appliances is present in the household and switched on. In Figure 6.14, this value is compared with

joint probability distribution [cf., e.g., 165].

[12] The standby power that was omitted when deriving the switched-on profiles will be neglected.

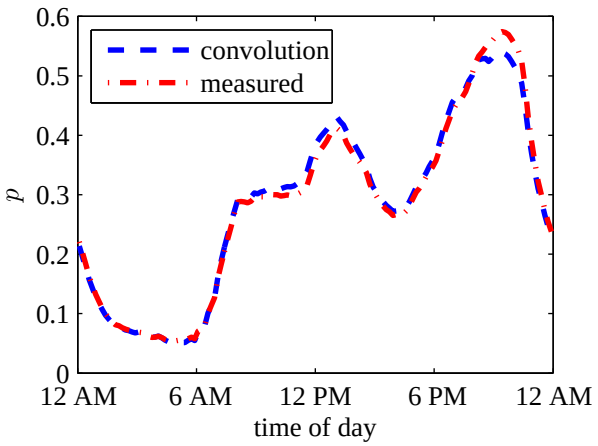


Figure 6.14: Mean measured probability that any of the appliances is switched on (red), and probability profile resulting of the convolution of Equation (6.7) using the monitored switched-on profiles (blue, cf. Figure 6.2).

the theoretical probability profile predicted by Equation (6.8).[13] The predictions are in very good agreement with the measurements. Around noon the measured values are overestimated, and in the evening they are underestimated, respectively corresponding to anti-correlation and correlation of individual switched-on profiles.

A set of quantiles, as well as the mean of the normalised non-zero part (corresponding to the load profile distribution whilst being "switched on") of $f_{P_{\mathrm{el}}}^{\mathrm{tot,m}}(t) \neq 0$ and $f_{P_{\mathrm{el}}}^{\mathrm{tot},*}(t) := \left(f_{P_{\mathrm{el}}}^{\mathrm{tot}}(t) \neq 0\right)$ is shown in Figure 6.15. Between 8 AM and 10 PM, the predicted distribution is in good agreement with the measured one. Only the maximum is overestimated by a factor of more than 2. In contrast the 99.9 % quantile is already underestimated. The minima of both distributions are constantly 6 W, which corresponds to the smallest non-zero measured value. During the night, the measured distribution is significantly overestimated by the predictions.

The reason for the overestimated maximum might be an anticorrelation of very high consumption values in the individual appliances' distributions. However, these predicted values are not completely reliable, due to numerical imprecisions when calculating the convolutions of Equation (6.7)[14]. Most importantly, the overestimation might result from the assumption of time-independent power consumption probability distributions of some of the appliances (the distributions of dishwashers and washing machines are overestimated during the night). The

[13] This corresponds to the probability that $f_{P_{\mathrm{el}}}^{\mathrm{tot}}(t)$ is non-zero as a function of time of day.

[14] The maximal value of the corresponding cumulative distribution function of $f_{P_{\mathrm{el}}}^{\mathrm{tot}}(t)$ was in average 0.98 instead of 1. Furthermore, the theoretical maximum non-zero value of $f_{P_{\mathrm{el}}}^{\mathrm{tot}}(t)$ is given by the sum of the maxima of the individual appliances in Figure 6.5, which is also not exactly reproduced.

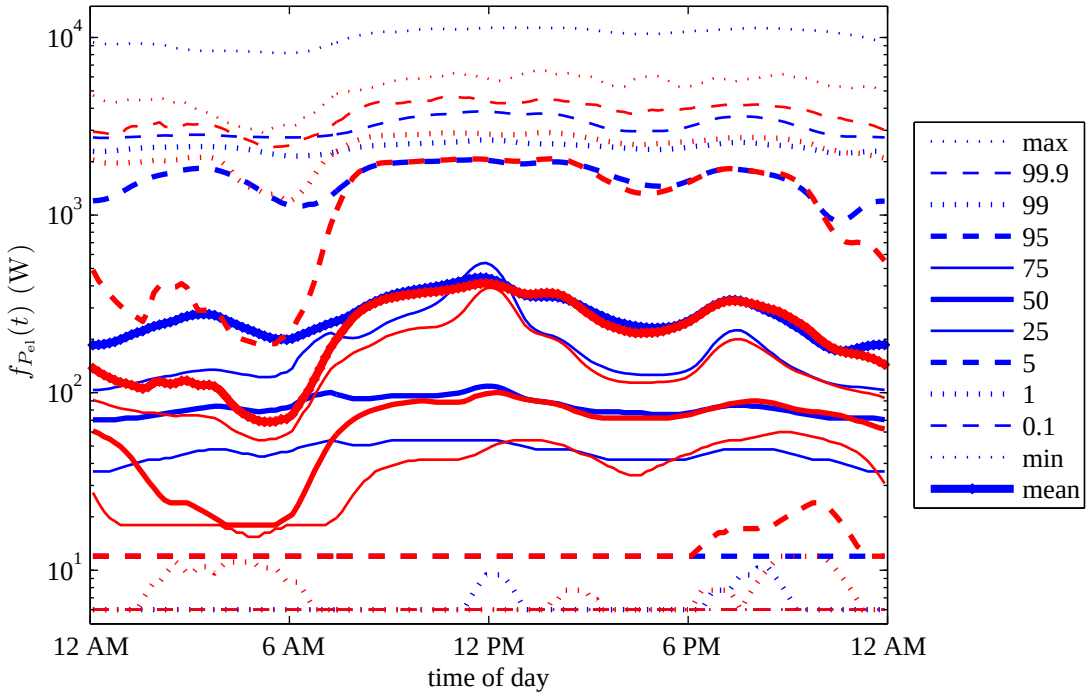


Figure 6.15: Comparison of the measured total power consumption distribution whilst being non-zero $f_{P_{el}}^{tot,m}(t) \neq 0$ (red), as well as the calculated one $f_{P_{el}}^{tot,*}(t) = (f_{P_{el}}^{tot}(t) \neq 0)$ (blue).

constant overestimation during the night might also be related to anticorrelation of the power consumptions[15], as for instance, during the night it might be less probable that the cooker and the microwave are simultaneously used in a very large extent. However, this might also be most related to the time-independent description of the power consumption distributions of some appliances (during the night the monitored data is less statistically significant; cf. Figure 6.5).

6.4.2 Activity-Dependent prediction of load profile distribution

In this section, we will investigate the influences of the activity-dependent switched-on profiles (Section 6.3) on the predicted aggregated load profile distribution. For this purpose, the aggregated distributions of the individual households $f^{\mathrm{m}}_{P_{\mathrm{el}}}(t)$ were calculated with the activity-dependent switched-on profiles (see Figures 6.10 and 6.11).

Figure 6.16 shows a comparison of the predicted probability profiles that at least one of the five appliances is switched on, based on the measured individual switched-on profiles (blue, cf. Figure 6.15), as well as the activity-dependent ones (green). The differences between the two curves can be attributed to the deviations of the activity-dependent switched-on profiles and the measured ones, which are both shown in Figures 6.10 and 6.11.

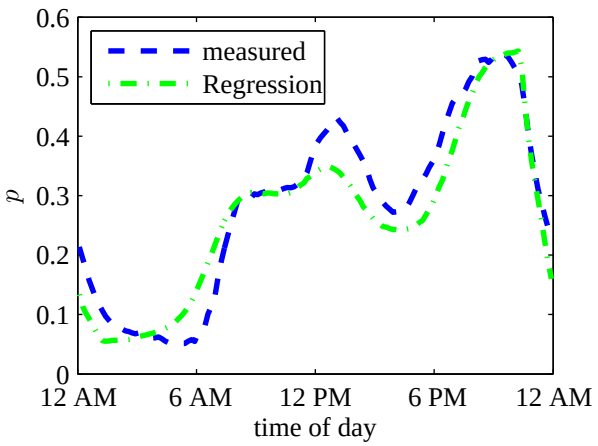


Figure 6.16: Probability that any of the appliances is switched on, according to Equation (6.8), resulting of the measured (blue) and of the activity-dependent switched-on profiles (green).

Figure 6.17 shows a set of quantiles and the mean of the predicted normalised non-zero part of $f^{\mathrm{tot}}_{P_{\mathrm{el}}}(t)$, based on the measured (blue) and of the activity-

[15] Correlations are neglected, due to the assumption that different distributions are independent of each other.

dependent switched-on profiles (green). The comparison shows that the description of the switched-on profiles by the activity-dependent model of Section 6.3 does not result in substantial errors in the predictions of the load profile distributions.

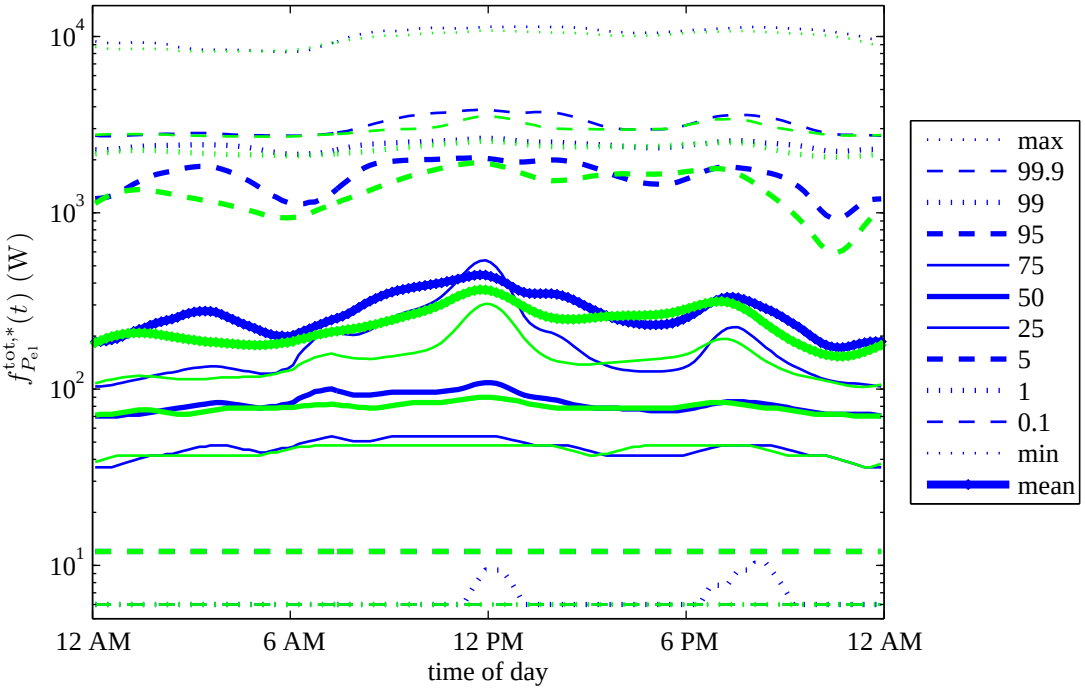


Figure 6.17: Comparison of the predictions of the load profile distribution whilst being in use $f_{P_{el}}^{tot,*}(t)$ according to Equation (6.8) of the appliances of Figure 6.13, based on the measured switched-on profiles (blue, also shown in Figure 6.15), and those predicted conditional on residential activities (green, cf. Figures 6.10 and 6.11).

6.5 Discussion

A modelling approach has been formulated to support predictions of the distributions of residential load profiles of individually and simultaneously used electrical appliances. These distributions depend on the load profile distributions whilst being in use, as well as on the time-dependent probabilities that the appliances are in use. In order to model the influences of human behaviour on appliance use, a methodology was developed to relate the probabilities to use human-controlled electrical appliances to residential activities performed. It has been shown that the time-dependence of appliance use and of the load profile distributions are accurately predicted by these models, respectively based on linear regression of the

logits of appliance and activity profiles and on convolution of the time-dependent load profile distributions. Furthermore, an approach was presented, which also reproduces the measured distribution of usage durations.

In the model to predict the aggregated load distribution of the simultaneous use of individual appliances, correlations were neglected. The resulting profile that at least one appliances is in use did not show strong deviations from the measured curve. In contrast, the predictions of the non-zero load profile distribution during the night show significant deviations from the measured one, which are probably related to an overestimation of the measured values by time-independent appliance load profile distributions.

It would be desirable to develop a model that predicts the conditional probability $p(l|j)$ of the use of an electrical appliance l whilst performing an activity j, as a function of time of day, as well other characteristics (such as the characteristics of the household/its members, or the weekday), similarly as in Chapters 3 and 4. Furthermore, the household ownership of the appliance l (cf. Chapter 5) might also have significant influences (for instance, the conditional probability to use the cooker whilst preparing food might depend on, whether there is a microwave oven in the household). Whereas the latter would have been relatively easy to additionally record in the IRISE campaign, the acquisition of a representative dataset with simultaneous recordings of occupancy and activities, as well as appliance use is complicated in reality, due to the nature of the two different kinds of survey. Appliance monitoring campaigns are usually longitudinal studies over a long-time period of a smaller set of households, whereas transversal time use surveys tend to cover a large sample population, where diaries are usually not longer recorded than a day. However, if the characteristics of the households and their members were recorded in the same level of detail than those of the TUS, the dependence of $p(l|j)$ on the characteristics and on time could be well determined via the approach shown in Section 6.3 applied on sub-population profiles.

The two modelling approaches of Section 6.3.1 can be implemented in simulations as a post-process of residential occupancy (cf. Chapter 3) and residential activities (see Chapter 4). In the approach where appliance use is modelled as a conditional probability whilst performing an activity (according to Equation (6.3)), the dichotomous variable whether the appliance is in use can be modelled as a time-inhomogeneous Bernoulli process. When furthermore, appliance use durations are explicitly modelled, this can then be implemented in the same way as it was schematised in Figure 4.1. As mentioned, the activity model then corresponds to the pre-process of the appliance model. In bottom-up simulations of individual appliance use, the methodology of Section 6.4 is not needed, but instead, the power consumption of the appliances can be directly generated from the individual distributions (cf. Figure 6.5). However, the presented approach can be used to model individual residential load profiles by directly generating values from the derived distributions, which saves a substantial amount of computational time.

In Equation (6.7), it was explained how the load profile distribution of simularily used appliances can be obtained by means of convolution. Figure 6.18 shows schematically, how this methodology can be extended to derive $f_{P_{\mathrm{el}}}^{\mathrm{NH}}(t)$, the distribution of aggregated power consumption values of a neighbourhood (NH) of N households $\sum_{i=1}^{N} P^i(t)$. The load profile distribution $f_{P_{\mathrm{el}}}^{i}(t)$ of a household i represents the 24 h periodic time average of the random variable of the total household demand $P^i(t)$. According to the logic of Equation (6.7), the load profile distribution of the neighbourhood as a function of time is given by

$$f_{P_{\mathrm{el}}}^{\mathrm{NH}}(t) = *_{i=1}^{N} f_{P_{\mathrm{el}}}^{i}(t), \tag{6.9}$$

where $*_{i=1}^{N}$ denotes the convolution of all individual distributions $f_{P_{\mathrm{el}}}^{i}(t)$ for $i = 1, \ldots, N$.[16] This makes the approach highly versatile in application, as on the one hand the individual appliances' load profile distributions can be treated in a high level of detail, either dependent on human behaviour or according to other models. On the other hand, the unknown appliances can be regrouped into the statistical offset of the category stuff.

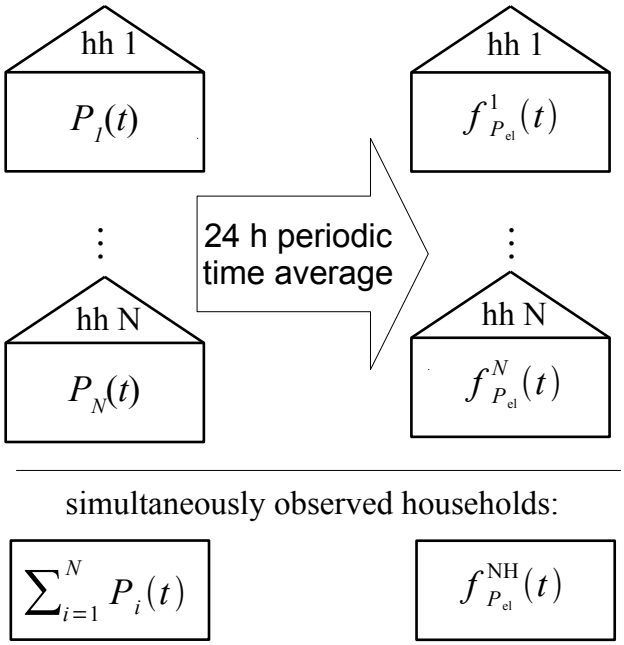


Figure 6.18: Methodology to predict the load profile distribution $f_{P_{\mathrm{el}}}^{\mathrm{NH}}(t)$ of the total power demand $\sum_{i=1}^{N} P^i(t)$ of a neighbourhood of N households.

This methodology can be readily implemented to predict residential load profiles for future scenarios with changed boundary conditions. For instance, changes

[16] The identity of $f_{P_{\mathrm{el}}}^{\mathrm{NH}}(t)$ with the 24 h periodic distribution of $\sum_{i=1}^{N} P^i(t)$ follows from the reasonable assumption that all $P^i(t)$ are pairwise independent from each other for $i = 1, \ldots, N$.

in residential behaviour only have impacts on the residential activity profiles $p_j(t)$ of the individual households. Changes in appliance ownership affect only $K(\nu)$ in Figure 6.12. Changes in population characteristics would change the frequency of the $1, \ldots, N$ household characteristics in Equation (6.9). In order to model changes in individual appliances' power consumption (for instance an increased diffusion of appliances with an A++ energy label), only the corresponding $f^*_{P_{\mathrm{el}}}(t)$ has to be modified.

6.6 Conclusion

In this chapter, a novel approach was developed which predicts the use of individual electrical appliances, as well as the resulting aggregated load profile distribution of simultaneously used appliances, based on the residential activities performed. The approach was limited to electrical appliances which are directly controlled by occupants, but the methodology can be readily extended to predict the distribution of the entire residential appliance stock. As it was furthermore shown that the non-measured appliances correspond to a statistical offset with a distribution that is approximately constant with time, the methodology is a valuable approach for application in the field of demand side-modelling or the smart grid.

The purpose of this model is to support the more accurate estimation of residential electricity demand, in dependence of household characteristics. The methodology was designed to provide a theoretical basis to calculate analytically/ numerically the load distributions, but it can also be implemented in simulation tools to generate concrete outcomes of the distributions.

The methodology provides a robust, adaptable and generalist approach to predict electricity load profiles of households or residential neighbourhoods in future scenarios, in order to investigate for the impact of changes in residential behaviour, appliance ownership, population characteristics or individual appliance power consumption distributions. It was discussed that discrepancies between observations and predictions could be diminished, using refined calibration procedures that are based on more comprehensive measurement sets. For instance, the load profiles may also depend on many other factors, such as the household size or the habitable surface [49], national differences [46] and many others [53], which was neglected as the sample size of the IRISE dataset does not allow to calibrate a meaningful bottom-up model that can be considered as representative. Thus, the characteristics of the two different datasets which were used to calibrate the models could not be fully aligned. Apart from that, it is of central importance to investigate the different variables for correlations. Furthermore, it would be of high interest to apply this methodology to other datasets, in order to study for temporal or national changes, as well as other characteristics.

Chapter 7

Conclusion

In this final chapter, we review and discuss the key contributions of this research work, where we developed a bottom-up modelling framework which addresses the prediction of residential presence and activities, as well as of the distribution of load profiles related to the presence and the use of individual electrical appliances in households.

Summary

A modelling framework was established which provides a robust, adaptable and generalist approach to predict electricity load profiles of households or residential neighbourhoods in future scenarios, in order to take into account the impact of changes in residential behaviour, appliance ownership, population characteristics or individual appliance power consumption distributions. Based on survey data of time use information, as well as of households' ownership of appliances, coupled with data of a measurement campaign of individual appliance use in households, careful statistical analysis was carried out, which results in the following advances:

- A validated model for the prediction of the time-dependent probability of residential presence, as well as of its durations, calibrated with a set of rigorously selected and significant explanatory variables. The developed approaches allow the determination of the temporal evolution of time-inhomogeneous first- and higher-order Markov processes with fast convergence.
- A validated model to predict the time- and individual-dependent conditional probabilities to perform residential activities whilst being present at home, reproducing observed duration distributions, as well as activity transitions, verified by a classical cross-validation procedure, which is used to select the optimal model formulation, as well as relevant input parameters.
- A methodology to predict the dependence of the probabilities of households' ownership of a large range of electrical appliances on the household

characteristics, which predictions were validated. The approach facilitates the fast elimination of non-significant parameters in principal component logistic regression by re-translating the values of the principal components into the original predictors.

- A model to predict the time-dependent probabilities of the use of electrical appliances, conditional on the ongoing residential activities, where measured distributions of usage durations, as well as of load profiles can be accurately reproduced.

Contribution of the developed stochastic models

The proposed stochastic models to describe residential occupant behaviour open new perspectives for the use in dynamic simulation programs for the prediction of energy flows. We will present a list of topics of particular interest:

- The impact of individual variations in residential occupancy on buildings' energy balance can be more accurately investigated.
- The detailed description of the variation in residential activity profiles also allows for a more accurate treatment of numerous other processes in buildings, such as the interaction of occupants with building components or the influences of activities on environmental comfort in dependence of the individual characteristics.
- The rigorous and generalist formulation of the individual models provides a robust basis to investigate temporal and cultural changes of the described processes, which only requires general input data of the temporally resolved disaggregated data of time use and appliance measurements, as well as of the appliance ownership in households.
- The dependence of the models on individual characteristics and the bottom-up formulation lends itself well to the modelling of future scenarios to explore responses to changes to the population's demographic/behavioural characteristics, the diffusion rate of appliances, of appliances power consumption or in the electrical appliance usage behaviour.

Figure 7.1 summarises the modelling framework to predict the distribution of the load profile of a household $\widetilde{\boldsymbol{x}}$ consisting of the individuals $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. The appliance stock in the household is modelled according to the probabilities of appliance ownership as a function of household characteristics $\widetilde{\boldsymbol{x}}$ (Chapter 5). The models of Chapters 3 and 4 allow to determine the probabilities to perform activities j whilst being at home $p(\boldsymbol{x}_i, t) \cdot p_j(\boldsymbol{x}_i, t)$ for each household member $\boldsymbol{x}_i$. These enable the modelling of $p_l(t)$, the time-dependent probabilities that a (user-controlled) appliance is in use (Section 6.3). The latter together with the power

consumption characteristics of the individual appliances in the household are then used to calculate the load profile consumption of the entire household by means of convolution (Section 6.4). In Section 6.5 it was discussed how this methodology can be extended to model the electricity load of a residential neighbourhood.

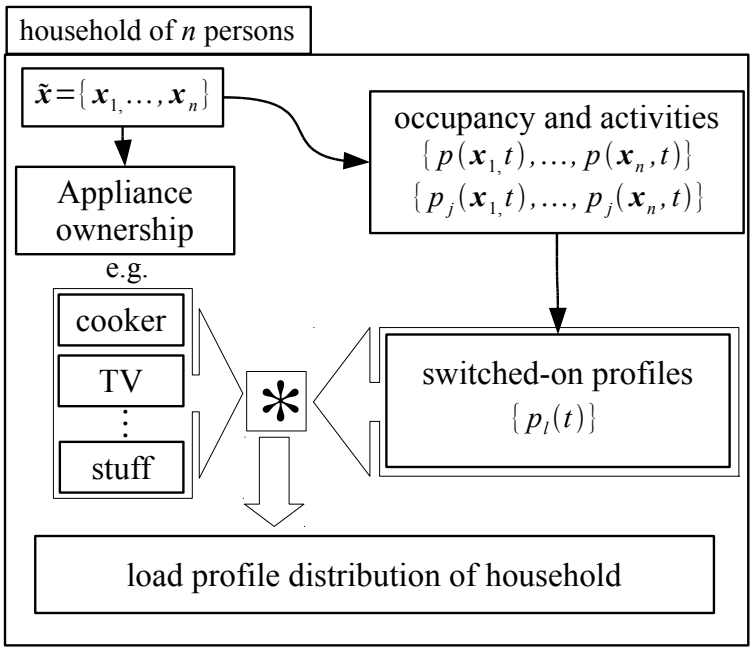


Figure 7.1: Methodology to predict the load profile distribution of a household $\widetilde{\boldsymbol{x}}$ accommodating the individuals $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$.

Longer term perspectives

In spite of the mentioned advances, the examples of the following issues show that there is still considerable scope for further improvements in the modelling of the studied processes:

- The correlations between occupancy and activities of different members of the same household, as well as between appliance ownership and appliance use needs to be investigated in dependence of the household characteristics, which would require in-depth designed measurements.

- By fitting the models with data of other surveys and monitoring campaigns, the methodology can be applied to quantitatively estimate temporal and cultural changes in human behaviour that affect energy demand.[1] It is to be

[1] An investigation of cultural/temporal specificities is also important to quantify model inaccuracies of the current study, which is based on Swiss household appliance ownership data of 2005, and French time use and appliance power consumption data of 1998/1999.

expected that not all aspects of the findings on human behaviour are valid in a different temporal/cultural context. However, the robustness of the calibration methodologies should ensure that the methodologies are readily applicable to other contexts.

- The integrated use of the models, as presented in Figure 7.1, would allow to investigate in more detail the dependence of residential electricity load profile distributions on human behaviour and household characteristics.

- The calibration of the electricity use model with a dataset containing more detailed individual characteristics of a larger sample of households would allow to investigate the influences of explanatory variables on the usage behaviour, and thus to describe the involved conditional probabilities more accurately.

- When realising smart grid concepts, remote control [cf. 166] might gain importance for specific types of electrical appliances, which may lead to a different relevance of residential presence regarding residential electricity demand. However, the methodologies of the presented modelling approaches could then be further developed to quantitatively estimate the impacts of such behavioural changes on electricity loads.

  The methodology of Section 6.3.1 may be applied to predict load profiles of individual appliances in the context of a more flexible use to match electricity demand with supply from renewable energy sources [167].

Appendix A

Dynamics of electric power consumption distribution

Apart from the specific appliance, the distribution of the electric power demand of electric appliances $f_{P_{\mathrm{el}}}$ as defined in Chapter 6 might be influenced by many other factors. Regarding cookers, the distribution might depend most significantly on the characteristics of the household and its individuals. As these characteristics are unfortunately not available for the IRISE dataset, and furthermore, the sample size does not allow for a meaningful investigation of the dependence on the household size or surface, this is not considered. In this chapter, we will focus on dependence of $f_{P_{\mathrm{el}}}$ on other temporal characteristics than the time of day.

A.1 Duration of being in use

In Figure A.1, we show the dependence of $f_{P_{\mathrm{el}}}$ on the duration for which the cooker was already in use (in 10 min time steps; indicated on top of each sub-graph). In other words, in the upper-left sub-graph $f_{P_{\mathrm{el}}}$ is shown for time steps where the cooker was not switched on in the time step before. In addition, the duration data were fitted with a Weibull (cf. Section 4.2.1) [cf. 18], as well as an exponential probability distribution function

$$f_{P_{\mathrm{el}}} = \begin{cases} 1/\mu \, \mathrm{e}^{-1/\mu \cdot P_{\mathrm{el}}}, & \text{if } P_{\mathrm{el}} \geq 0, \\ 0, & \text{if } P_{\mathrm{el}} < 0. \end{cases} \tag{A.1}$$

For the latter their confidence intervals are shown by dotted lines. There is a considerable variation of the fitted mean values μ of the exponential distributions, ranging from 259 ± 10 W to 551 ± 5 W. In all time intervals after being switched on, the estimated value of k of the Weibull distribution is not greater than one, which implies that smaller power consumption values are more likely to occur than in an exponential distribution. With increasing cooking duration, the value of k tends to decrease. This might be explained by the fact that in every of these

intervals there is a considerable proportion of cooker uses that end during the considered interval, and therefore the distribution is more pronounced for small power values (as then, power consumption is only taking place in a fraction of the time interval). Another reason for the increased proportions of small power consumption in the distributions is the use of the cooker on low flame. The latter might be more likely to occur with increasing cooking duration, whereas the use might more likely involve strong heating in the beginning.

Regarding appliances other than cookers, the consideration of the duration already being on might also be of importance. For instance, the power demand of dishwashers or washing machines may also depend on the duration due to different phases in the operating mode. In Figure 6.7, for instance, it was shown that water heating takes an important part of energy consumption, which can take a considerable part of energy consumed by those two appliances [cf. 153]. Although for dishwashers and washing machines the power consumption is not influenced by the user but purely determined by the machine program, it would be too tedious to calibrate a model with all the necessary information of different devices and cycles. Thus, this methodology would allow to detect common probabilistic patterns in the ensemble of the latter.

This implies that in simulations, the power demand of appliances would have to be modelled according to a distribution that depends on the time of day (see Figure 6.5), as well as on the duration already being in use. The latter cannot be treated with the approach of Equation (6.4), as it does not explicitly model durations of appliance use. Instead, an approach would have to be used, where appliance usage starts and their durations are modelled (cf. Equation (6.6)).

A.2 Transition of power demand

The phenomenon of a decreasing power consumption of cookers with increasing duration can also be seen in Figure A.2. Here, we show the transition probability matrix of power consumption intervals I_P of successive time steps t and $t+1$ that was observed in the measurements of the IRISE dataset. The measured transition probabilities that a power consumption transits from a value in the interval I_P to one in $I_P(t+1)$ is represented by circles whose radii are proportional to the transition probabilities' magnitude (see examples on the right). The time steps where the cooker was and stays switched off at both time steps ($I_P(t) = I_P(t+1) = 0$) were omitted as they are several orders of magnitude higher than the other transition probabilities in this row (being a transition matrix, the sum of all transition probabilities of the same row is equal to one). An appliance type where all individual instances have a constant power consumption would be characterised by a transition matrix where all non-diagonal elements are equal to zero. As it was discussed together with Figure A.1, this example also shows

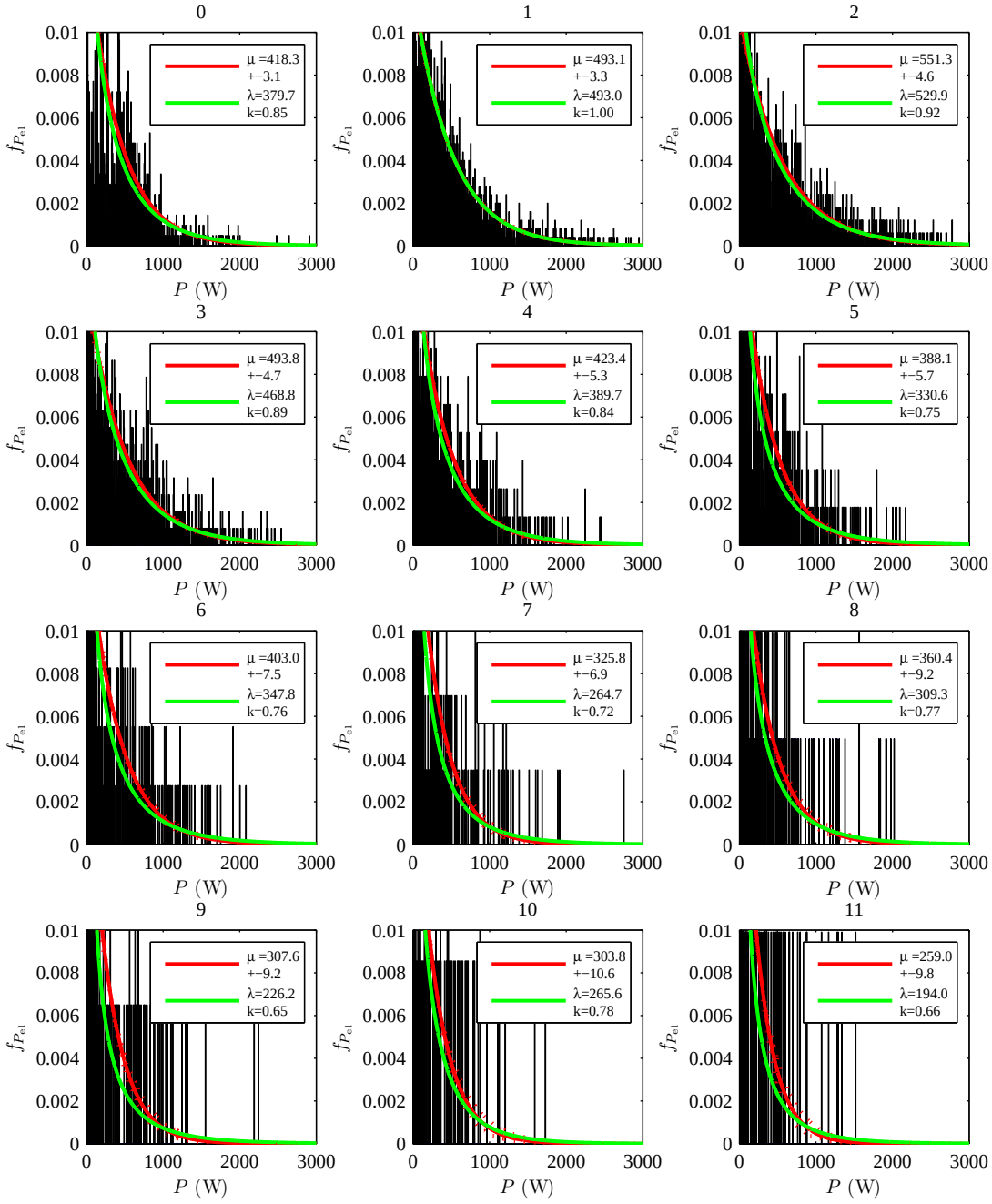


Figure A.1: Dependence of cookers' EPDF of power consumption on the number of time steps already being in use, with exponential (red) and Weibull fits (green).

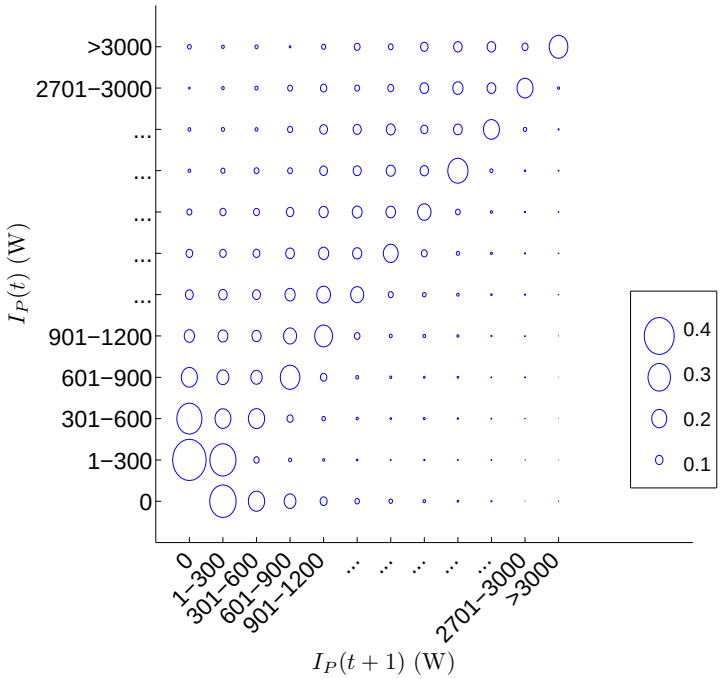


Figure A.2: Transition probability of the power consumption interval I_P between adjacent time steps t and $t+1$.

that the power consumption of cookers is more likely to decrease than to increase with increasing duration.

By calibrating these transition matrices for all considered types of appliance, one could more accurately simulate the appliances' electric power consumption whilst being in use, by means of a first-order Markov process (cf. Section 3.2.1). However, the transition matrix might be much closer to identity for other appliance types with constant power demand as it was mentioned. In these cases, the modelling of the distribution of the appliances' load profiles as a Markov process would not lead to an improvement of the model's predictive power.

A.3 Discussion and conclusion

The shown dependences of the cookers' power demand distributions can be used to model the load profile distribution of electrical appliances more accurately, as there are first- and higher-order memory effects in the evolution of the electric power demand of appliances that are switched on. However, this also leads to increases in computational effort, which therefore have to be carefully evaluated against the improvements in predictive power.

List of Figures

List of Tables

Nomenclature

α_j	ASC in the utility function of activity j in the MNL model of $p_{\mathrm{s},j}(\boldsymbol{x}, t)$.
i	Index for appliance type (in Chapter 5), and to specify a household (in Chapter 6).
$\beta_{j',j}$	Parameter in the utility functions of the MNL models accounting for the utility for transiting from an activity j to a subsequent activity j'.
$\boldsymbol{\Sigma}^{\boldsymbol{\beta}'}$	Covariance matrix of estimated parameters $\boldsymbol{\beta}'$.
η	Decadic logarithm of the relative error of the predictions of appliance ownership.
$f_{P_{\mathrm{el}}}(t)$	Time-Dependent distribution of power consumption.
$\widetilde{\boldsymbol{x}}$	Variable indicating the characteristics of a household.
$\boldsymbol{y}$	Vector of principal components.
T	Transpose of a vector/matrix.
j_{MTUS}	Activity type according to the "MTUS 41-activity typology".
λ	Scale parameter in the Weibull distribution function.
$\boldsymbol{U}$	Covariance matrix of $[\widetilde{\boldsymbol{x}} - \overline{\widetilde{\boldsymbol{x}}}]$.
$\mathcal{C}$	Entire sample population of the TUS.
$\mathcal{C}_{\mathrm{nw}}$	Sub-population of individuals not being in paid work.
$\mathcal{C}_{\mathrm{w}}$	Sub-population of individuals in paid work.
$\mathcal{L}$	Log-likelihood.
$\mathcal{L}_{\mathrm{t}}$	Time-Dependent log-likelihood.
$\mathcal{L}_{\boldsymbol{x}}$	Log-likelihood dependent on individual $\boldsymbol{x}$.

H_1	Hypothesis that system is represented by model 1.
H_2	Hypothesis that system is represented by model 2.
H_c	Hypothesis that system is represented by the composite model.
pr	Probability.
$\overline{P^*}$	Mean consumed power above 18 W.
$\overline{P}$	Mean non-zero value of consumed power in Watt.
$\overline{p}_\mathrm{O}$	Mean predicted diffusion probabilities over sub-populations by the OLR model.
$\overline{p}_\mathrm{PC}$	Mean predicted diffusion probabilities over sub-populations by the PCLR model.
$\overline{p}_\mathrm{sub}$	Mean predicted diffusion probabilities over sub-populations.
β_k	Parameter capturing the influence of the dummy variable x_k on the utility function.
$\boldsymbol{x}$	Variable describing (the characteristics of) an individual.
f	Probability distribution function of residential presence durations.
$f_{0,t_\mathrm{s}}(t)$	Duration PDF of presences started at t_s according to the IIM.
$f_{\mathrm{M},t_\mathrm{s}}$	Duration PDF of presences started at t_s according to the FOMP (cf. Section 3.2.1).
$f_{t_\mathrm{s},\mathrm{u}}(t)$	Probability distribution function of durations which end before midnight.
$f_{t_\mathrm{s}}(t)$	Probability distribution function of durations t of presences that were started at t_s.
$\boldsymbol{P}$	Row vector of the probabilities of being present or absent.
y	Variable specifying the type of place.
$y_{\boldsymbol{x}}(t)$	Occupancy chain of an individual $\boldsymbol{x}$ as a function of time t.
$p_i(\widetilde{\boldsymbol{x}})$	Probability that an appliance i is present in the household $\widetilde{\boldsymbol{x}}$.
$\rho_{t_\mathrm{s}}(\boldsymbol{x})$	Proportion of durations started at t_s that are censored for the sample sub-population with characteristics $\boldsymbol{x}$.
Y	Random variable of occupancy status y.

T_{end}	Random variable indicating the time the presence duration ends.
$S(t)$	Survival function of presence after t.
$S_{0,t_s}(t)$	Survival function of presences started at t_s according to the IIM.
$S_{t_s}(t)$	Survival function of presence that was started at t_s after t.
D_{tot}	Observed diffusion percentage of the corresponding appliance in the survey dataset (cf. Section 5.2.1).
t_{00}	Transition probability to stay away.
t_{01}	Transition probability to come home.
t_{10}	Transition probability to leave home.
t_{11}	Transition probability to stay at home.
$\boldsymbol{T}$	Transition probability matrix.
$\boldsymbol{T}_{\text{h}}(t)$	Time-dependent transition probability matrix, calibrated for transitions during a time step of one hour.
$\boldsymbol{T}_{10\,\text{min}}(t)$	Time-dependent transition probability matrix for transitions during a time step of 10 min.
A	Average ratio of the total number of time steps, where the residential activity chain of the TUS has been correctly predicted by the simulation.
A_{ns}	Average ratio of the total number of time steps, where the residential activity chain of the TUS has been correctly predicted by the simulation, disregarding sleeping.
$a_{\boldsymbol{x}}(t)$	Activity chain of an individual $\boldsymbol{x}$ as a function of time t.
AIC	Akaike Information Criterion.
BIC	Bayesian Information Criterion.
$D_j(t)$	Indicator for average difference between the predicted probabilities to perform a residential activity $p_{j,\text{sim}}(t)$ and the corresponding proportions that were observed in the TUS ($p_{j,\text{obs}}(t)$.
$f_j(t)$	Duration probability distribution function of activity j started at time t.

$f_p(t)$	Distribution of mean presence over the population $\mathcal{C}$ as a function of time of day t.
$F_{t_\mathrm{s}}(\boldsymbol{x}, t)$	Cumulative distribution function of durations t started at t_s by the individual $\boldsymbol{x}$.
j	Index for merged activity types (cf. Figure 2.8).
k	Shape parameter in the Weibull distribution function.
l	Electrical appliance index in Chapter 6.
M	Number of dummy variables for the estimation of the MNL/binomial logit models in Chapters 4 and 5.
N	Number of merged activity types.
n	Number of individuals in the TUS.
N_hh	Number of households in the IRISE dataset.
n_min	Minimum size of the disaggregated sub-populations when calibrating the individual-dependent $f_j(\boldsymbol{x}, t)$ (cf. Section 4.2.1).
n_r	Number of replicates in a sample of simulations.
$p(\boldsymbol{x}, t)$	Probability of individual $\boldsymbol{x}$ to be at home at time t.
$p(t)$	Probability to be at home at time t.
p_pr	Probability with which the correct outcome is predicted by the models in Chapter 5.
$p_j(t)$	Conditional probability to perform activity j whilst being at home.
$p_\mathrm{obs}(t)$	Probability to be at home at time t observed in the TUS.
$p_{\mathrm{s},j}(t)$	Probability to start activity j at time t.
$p_{j,\mathrm{obs}}$	Mean probability to perform activity j whilst being at home observed in the TUS.
$p_{j,\mathrm{sim}}$	Simulated mean probability to perform activity j whilst being at home.
R	Average percentage of time steps the appliances were in use (consuming a non-zero electric power).
R^*	Average percentage of time steps the appliances were consuming an electric power above 18 W.

T	Period of 24 h.
t	Time.
$t_{\max}$	Integer denoting the maximum memory of the higher-order Markov process in Section 3.2.1.
t_n	Discretised time step.
t_{end}	Last time step of the simulation.
t_{last}	Last time step of the day in the discretised Markov chain.
t_{s}	Time of day when presence is started.
V_j	Utility function of activity j in the MNL model of $p_{\text{s},j}(\boldsymbol{x}, t)$.
x_k	Dummy variable in the utility function of the MNL/binomial logit models.
ASC	Alternative-Specific constant.
CTRW	Continuous-Time random walk.
EPDF	Empirical PDF.
FOMP	First-Order Markov process.
HOMP	Higher-Order Markov process.
IIM	Individual-Independent model (time-inhomogeneous Bernoulli process; cf. Section 3.2.1).
MNL	Multinomial logit.
MTUS	Multinational Time Use Study [30].
PDF	Probability distribution function.
RUM	Random utility model.
TUS	French time use survey [32].

Bibliography

[1] R. Saidur, H. Masjuki, and M. Jamaluddin. An application of energy and exergy analysis in residential sector of Malaysia. *Energy Policy*, 35(2):1050 – 1063, 2007. ISSN 0301-4215. doi: 10.1016/j.enpol.2006.02.006.

[2] European Commission and others. EU Energy and Transport in Figures 2010. *Luxembourg: Publications Office of the European Union*, 2010.

[3] L. Pérez-Lombard, J. Ortiz, and C. Pout. A review on buildings energy consumption information. *Energy and Buildings*, 40(3):394 – 398, 2008. ISSN 0378-7788. doi: 10.1016/j.enbuild.2007.03.007.

[4] P. Bertoldi and B. Atanasiu. Electricity consumption and efficiency trends in the enlarged European Union. Status report 2006-EUR 22753 EN. Joint Research Centre. *Institute for Environment and Sustainability. Ispra, Italy*, 2007.

[5] S. Conti and S. Raiti. Probabilistic load flow using Monte Carlo techniques for distribution networks with photovoltaic generators. *Solar Energy*, 81 (12):1473 – 1481, 2007. ISSN 0038-092X. doi: 10.1016/j.solener.2007.02.007.

[6] F. C. Winkelmann and S. Selkowitz. Daylighting simulation in the DOE-2 building energy analysis program. *Energy and Buildings*, 8(4):271 – 286, 1985. ISSN 0378-7788. doi: 10.1016/0378-7788(85)90033-7.

[7] F. Arumí-Noé and D. O. Northrup. A field validation of the thermal performance of a passively heated building as simulated by the DEROB system. *Energy and Buildings*, 2(1):65 – 75, 1979. ISSN 0378-7788. doi: 10.1016/0378-7788(79)90021-5.

[8] J. Clarke. *Environmental systems performance.* PhD thesis, University of Strathclyde, 1977.

[9] M. Gough. *Modelling heat flow in buildings: An eigenfunction approach.* PhD thesis, University of Cambridge, 1982.

[10] K. Lomas, H. Eppel, C. Martin, and D. Bloomfield. Empirical validation of building energy simulation programs. *Energy and Buildings*, 26(3):253 – 275, 1997. ISSN 0378-7788. doi: 10.1016/S0378-7788(97)00007-8.

[11] J. Clarke. *Energy Simulation in Building Design (Second Edition)*. Butterworth-Heinemann, Oxford, United Kingdom, 2001.

[12] R. H. Socolow. The twin rivers program on energy conservation in housing: Highlights and conclusions. *Energy and Buildings*, 1(3):207 – 242, 1978. ISSN 0378-7788. doi: 10.1016/0378-7788(78)90003-8.

[13] A. Bahaj and P. James. Urban energy generation: The added value of photovoltaics in social housing. *Renewable and Sustainable Energy Reviews*, 11(9):2121 – 2136, 2007. ISSN 1364-0321. doi: 10.1016/j.rser.2006.03.007.

[14] G. Iwashita and H. Akasaka. The effects of human behavior on natural ventilation rate and indoor air environment in summer — a field study in southern Japan. *Energy and Buildings*, 25(3):195 – 205, 1997. ISSN 0378-7788. doi: 10.1016/S0378-7788(96)00994-2.

[15] S. Banfi, M. Farsi, M. Filippini, and M. Jakob. Willingness to pay for energy-saving measures in residential buildings. *Energy Economics*, 30(2): 503 – 516, 2008. ISSN 0140-9883. doi: 10.1016/j.eneco.2006.06.001.

[16] P. Hoes, J. Hensen, M. Loomans, B. de Vries, and D. Bourgeois. User behavior in whole building simulation. *Energy and Buildings*, 41(3):295 – 302, 2009. ISSN 0378-7788. doi: 10.1016/j.enbuild.2008.09.008.

[17] W. Lackie. The influence of load and diversity factors on methods of charging for electrical energy. *Journal of the Institution of Electrical Engineers*, 42(193):100 –114, February 1909.

[18] S. Borg and N. Kelly. The effect of appliance energy efficiency improvements on domestic electric loads in European households. *Energy and Buildings*, 43(9):2240 – 2250, 2011. ISSN 0378-7788. doi: 10.1016/j.enbuild.2011.05.001.

[19] A. Hawkes and M. Leach. Impacts of temporal precision in optimisation modelling of micro-combined heat and power. *Energy*, 30(10):1759 – 1779, 2005. ISSN 0360-5442. doi: 10.1016/j.energy.2004.11.012.

[20] J. Widén and E. Wäckelgård. A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy*, 87(6): 1880–1892, 2010.

[21] L. Kelly, A. Rowe, and P. Wild. Analyzing the impacts of plug-in electric vehicles on distribution networks in British Columbia. In *Electrical Power Energy Conference (EPEC), 2009 IEEE*, pages 1 –6, Oct. 2009. doi: 10.1109/EPEC.2009.5420904.

[22] J. Widén. *System studies and simulations of distributed photovoltaics in Sweden*. PhD thesis, Umeå University, 2010.

[23] I. Richardson, M. Thomson, D. Infield, and C. Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10): 1878–1887, 2010.

[24] G. Tina, S. Gagliano, and S. Raiti. Hybrid solar/wind power system probabilistic modelling for long-term performance assessment. *Solar Energy*, 80 (5):578 – 588, 2006. ISSN 0038-092X. doi: 10.1016/j.solener.2005.03.013.

[25] J. Tanimoto, A. Hagishima, and H. Sagara. A methodology for peak energy requirement considering actual variation of occupants' behavior schedules. *Building and Environment*, 43(4):610–619, 2008.

[26] F. Haldi. *Towards a Unified Model of Occupants' Behaviour and Comfort for Building Energy Simulation*. PhD thesis, EPFL, Lausanne, 2010.

[27] J. Page. *Simulating occupant presence and behaviour in buildings*. PhD thesis, EPFL, Lausanne, 2007.

[28] R. Fritsch, A. Kohler, M. Nygård-Ferguson, and J.-L. Scartezzini. A stochastic model of user behaviour regarding ventilation. *Building and Environment*, 25(2):173 – 181, 1990. ISSN 0360-1323. doi: 10.1016/0360-1323(90)90030-U.

[29] C.-A. Roulet, P. Cretton, R. Fritsch, and J.-L. Scartezzini. Stochastic model of inhabitant behavior in regard to ventilation. Technical report, 1991.

[30] K. Fisher, M. Bennett, J. Tucker, E. Altintas, A. Jahandar, J. Jun, and other members of the Time Use Team. Multinational Time Use Study, Versions World 5.5.3, 5.80 and 6.0 (released October 2011). Created by Jonathan Gershuny and Kimberly Fisher, with Evrim Altintas, Alyssa Borkosky, Anita Bortnik, Donna Dosman, Cara Fedick, Tyler Frederick, Anne H. Gauthier, Sally Jones, Jiweon Jun, Aaron Lai, Qianhan Lin, Tingting Lu, Fiona Lui, Leslie MacRae, Berenice Monna, José Ignacio Giménez Nadal, Monica Pauls, Cori Pawlak, Andrew Shipley, Cecilia Tinonin, Nuno Torres, Charlemaigne Victorino, and Oiching Yeung. Centre for Time Use Research, University of Oxford, United Kingdom. `http://www.timeuse.org/mtus/`.

[31] K. Fisher, M. Bennett, J. Tucker, E. Altintas, A. Jahandar, J. Jun, and other members of the Time Use Team. Multinational Time Use Studies. Version 5, November 2012. Centre for Time Use Research, University of Oxford, United Kingdom. http://www.timeuse.org/files/cckpub/858/mtus-user-guide-r5.pdf, 2012.

[32] Readme File for France. http://www.timeuse.org/files/cckpub/mtus/study/2926/readme_fra1998.doc, 2009. [Online; accessed 12 February 2013].

[33] INSEE - National Institute of Statistics and Economic Studies - France. http://www.insee.fr/en/default.asp, March 2012. [Online; accessed 14 March 2012].

[34] CTUR Table of Time Use Studies Entry. http://www-2009.timeuse.org/information/studies/data/france-1998-99.php.

[35] National Institute of Statistics and Economic Studies, France. Évolution de la population jusqu'en 2012 (Population evolution until 2012). http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATnon02145, . [Online; accessed 12 February 2013].

[36] National Institute of Statistics and Economic Studies, France. Public pension plan contributors, retired persons and demographic ratio, 2010. http://www.insee.fr/en/themes/tableau.asp?reg_id=0&ref_id=nattef04560, .

[37] Bureau of Labor Statistics. American time use survey website. http://www.bls.gov/tus/. [Online; accessed 2 April 2012].

[38] J. Moore, L. Stinson, and E. Welniak. Income measurement error in surveys: A review. *Journal of Official Statistics, Stockholm*, 16(4):331–362, 2000.

[39] W. Rodgers, C. Brown, and G. Duncan. Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association*, pages 1208–1218, 1993.

[40] J. Widén, A. Molin, and K. Ellegård. Models of domestic occupancy, activities and energy use based on time-use data: Deterministic and stochastic approaches with application to various building-related simulations. *Journal of Building Performance Simulation*, 5(1):27–44, 2012.

[41] Y. Shimoda, T. Fujii, T. Morikawa, and M. Mizuno. Residential end-use energy simulation at city scale. *Building and Environment*, 39(8):959 – 967, 2004. ISSN 0360-1323. doi: 10.1016/j.buildenv.2004.01.020.

[42] R. Andersen, B. Olesen, and J. Toftum. *Occupant behaviour with regard to control of the indoor environment*. PhD thesis, Technical University of Denmark, Department of Civil Engineering, Section for Indoor Environment, 2009.

[43] F. Haldi and D. Robinson. The impact of occupants' behaviour on building energy demand. *Journal of Building Performance Simulation*, 4(4):323–338, 2011.

[44] M. A. R. Lopes, C. H. Antunes, and N. Martins. Energy behaviours as promoters of energy efficiency: A 21st century review. *Renewable & Sustainable Energy Reviews*, 16(6):4095–4104, Aug 2012. ISSN 1364-0321. doi: 10.1016/j.rser.2012.03.034.

[45] J. Tanimoto, A. Hagishima, T. Iwai, and N. Ikegaya. Total utility demand prediction for multi-dwelling sites by a bottom-up approach considering variations of inhabitants' behaviour schedules. *Journal of Building Performance Simulation*, 6(1):53–64, 2013. doi: 10.1080/19401493.2012.680498. URL `http://www.tandfonline.com/doi/abs/10.1080/19401493.2012.680498`.

[46] J. Torriti. Demand Side Management for the European Supergrid: Occupancy variances of European single-person households. *Energy Policy*, 44: 199–206, May 2012. ISSN 0301-4215. doi: 10.1016/j.enpol.2012.01.039.

[47] Y. Shimoda, Y. Yamaguchi, T. Okamura, A. Taniguchi, and Y. Yamaguchi. Prediction of greenhouse gas reduction potential in Japanese residential sector by residential energy end-use model. *Applied Energy*, 87(6):1944–1952, Jun 2010. ISSN 0306-2619. doi: 10.1016/j.apenergy.2009.10.021.

[48] Y. Yamaguchi, T. Fujimoto, and Y. Shimoda. Occupant behavior model for households to estimate high-temporal resolution residential electricity demand profile. *Proceedings of Building Simulation*, 2011.

[49] Y. G. Yohanis, J. D. Mondol, A. Wright, and B. Norton. Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings*, 40(6):1053–1059, 2008. ISSN 0378-7788. doi: 10.1016/j.enbuild.2007.09.001.

[50] C. Walker and J. Pokoski. Residential load shape modeling based on customer behavior. *IEEE Transactions on Power Apparatus and Systems*, 104 (7):1703–1711, 1985. ISSN 0018-9510. doi: 10.1109/TPAS.1985.319202.

[51] R. Baetens, R. De Coninck, J. Van Roy, B. Verbruggen, J. Driesen, L. Helsen, and D. Saelens. Assessing electrical bottlenecks at feeder

level for residential net zero-energy buildings by integrated system simulation. *Applied Energy*, 96:74–83, Aug 2012. ISSN 0306-2619. doi: 10.1016/j.apenergy.2011.12.098.

[52] N. Saldanha and I. Beausoleil-Morrison. Measured end-use electric load profiles for 12 Canadian houses at high temporal resolution. *Energy and Buildings*, 49:519–530, Jun 2012. ISSN 0378-7788. doi: 10.1016/j.enbuild.2012.02.050.

[53] D. Ndiaye and K. Gabriel. Principal component analysis of the electricity consumption in residential dwellings. *Energy and Buildings*, 43(2-3):446–453, 2011. ISSN 0378-7788. doi: 10.1016/j.enbuild.2010.10.008.

[54] J. Widen, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegard, and E. Wackelgard. Constructing load profiles for household electricity and hot water from time-use data - Modelling approach and validation. *Energy and Buildings*, 41(7):753–768, Jul 2009. ISSN 0378-7788. doi: 10.1016/j.enbuild.2009.02.013.

[55] C. Weber and A. Perrels. Modelling lifestyle effects on energy demand and related emissions. *Energy Policy*, 28(8):549–566, Jul 2000. ISSN 0301-4215. doi: 10.1016/S0301-4215(00)00040-9.

[56] H. Brown, S. Suryanarayanan, and G. Heydt. Some characteristics of emerging distribution systems considering the smart grid initiative. *The Electricity Journal*, 23(5):64–75, 2010.

[57] R. Brown. Impact of smart grid on distribution system design. In *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, pages 1 –4, July 2008. doi: 10.1109/PES.2008.4596843.

[58] A. Faruqui, R. Hledik, and S. Sergici. Piloting the smart grid. *The Electricity Journal*, 22(7):55–69, 2009.

[59] E. Ghisi, S. Gosch, and R. Lamberts. Electricity end-uses in the residential sector of Brazil. *Energy Policy*, 35(8):4107–4120, Aug 2007. ISSN 0301-4215. doi: 10.1016/j.enpol.2007.02.020.

[60] Y. G. Yohanis. Domestic energy use and householders' energy behaviour. *Energy Policy*, 41:654–665, Feb 2012. ISSN 0301-4215. doi: 10.1016/j.enpol.2011.11.028.

[61] T. Zhang, P.-O. Siebers, and U. Aickelin. Modelling electricity consumption in office buildings: An agent based approach. *Energy and Buildings*, 43(10):2882–2892, Oct 2011. ISSN 0378-7788. doi: 10.1016/j.enbuild.2011.07.007.

[62] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi. A bottom-up approach to residential load modeling. *IEEE Transactions on Power Systems*, 9(2):957–964, 1994.

[63] J. Widén, A. M. Nilsson, and E. Wäckelgård. A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. *Energy and Buildings*, 41(10):1001 – 1012, 2009. ISSN 0378-7788. doi: 10.1016/j.enbuild.2009.05.002.

[64] I. Richardson, M. Thomson, and D. Infield. A high-resolution domestic building occupancy model for energy demand simulations. *Energy and Buildings*, 40(8):1560–1566, 2008.

[65] C. Wang, D. Yan, and Y. Jiang. A novel approach for building occupancy simulation. *Building Simulation*, 4(2):149–167, Jun 2011. ISSN 1996-3599. doi: 10.1007/s12273-011-0044-5.

[66] D. Wang, C. Federspiel, and F. Rubinstein. Modeling occupancy in single person offices. *Energy and Buildings*, 37(2):121–126, Feb 2005. ISSN 0378-7788. doi: 10.1016/j.enbuild.2004.06.015.

[67] B. Dong and K. Lam. Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network. *Journal of Building Performance Simulation*, 4(4):359–369, 2011.

[68] B. Dong, B. Andrews, K. P. Lam, M. Hoeynck, R. Zhang, Y.-S. Chiou, and D. Benitez. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. *Energy and Buildings*, 42(7):1038–1046, Jul 2010. ISSN 0378-7788. doi: 10.1016/j.enbuild.2010.01.016.

[69] J. Page, D. Robinson, N. Morel, and J. L. Scartezzini. A generalised stochastic model for the simulation of occupant presence. *Energy and Buildings*, 40(2):83–98, 2008. ISSN 0378-7788. doi: 10.1016/j.enbuild.2007.01.018.

[70] C. Liao, Y. Lin, and P. Barooah. Agent-based and graphical modelling of building occupancy. *Journal of Building Performance Simulation*, 5(1, Part 2, SI):5–25, 2012. ISSN 1940-1493. doi: 10.1080/19401493.2010.531143.

[71] C. Liao and P. Barooah. An integrated approach to occupancy modeling and estimation in commercial buildings. In *American Control Conference (ACC), 2010*, pages 3130–3135. IEEE, 2010.

[72] D. Ettema, F. Bastin, J. Polak, and O. Ashiru. Modelling the joint choice of activity timing and duration. *Transportation Research Part A: Policy and Practice*, 41(9, SI):827–841, Nov 2007. ISSN 0965-8564. doi: 10.1016/j.tra.2007.03.001. 45th Congress of the European-Regional-Science-Association, Amsterdam, Netherlands, Aug 23-27, 2005.

[73] D. Ettema, T. Schwanen, and H. Timmermans. The effect of location, mobility and socio-demographic factors on task and time allocation of households. *Transportation*, 34(1):89–105, Jan 2007. ISSN 0049-4488. doi: 10.1007/s11116-006-0007-3.

[74] B. Alexander, M. Dijst, and D. Ettema. Working from 9 to 6? An analysis of in-home and out-of-home working schedules. *Transportation*, 37(3):505–523, May 2010. ISSN 0049-4488. doi: 10.1007/s11116-009-9257-1.

[75] F. Haldi and D. Robinson. Interactions with window openings by office occupants. *Building and Environment*, 44(12):2378–2395, 2009.

[76] F. Haldi and D. Robinson. Adaptive actions on shading devices in response to local visual stimuli. *Journal of Building Performance Simulation*, 3(2):135–153, 2010. ISSN 1940-1493. doi: 10.1080/19401490903580759.

[77] D. Bedeaux, K. Lakatos-Lindenberg, and K. Shuler. On the relation between master equations and random walks and their solutions. *Journal of Mathematical Physics*, 12:2116, 1971.

[78] V. Kenkre, E. Montroll, and M. Shlesinger. Generalized master equations for continuous-time random walks. *Journal of Statistical Physics*, 9(1):45–50, 1973.

[79] L. Van Hove. The approach to equilibrium in quantum statistics: A perturbation treatment to general order. *Physica*, 23(1):441–480, 1957.

[80] R. Zwanzig. On the identity of three generalized master equations. *Physica*, 30(6):1109–1123, 1964.

[81] E. Montroll. Fundamental problems in statistical mechanics. *Compiled by EGD Cohen North-Holland Publishing Company, Amsterdam*, page 230, 1962.

[82] M. Shlesinger. Asymptotic solutions of continuous-time random walks. *Journal of Statistical Physics*, 10(5):421–434, 1974.

[83] R. Begleiter, R. El-Yaniv, and G. Yona. On prediction using variable order Markov models. *J. Artif. Intell. Res. (JAIR)*, 22:385–421, 2004.

[84] A. DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer, 2011.

[85] U. Wilke. Database containing results and parameters of the models developed. `\\icare-srv03\Public & FTP\Everyone\Wilke`, 2013.

[86] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, dec 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

[87] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[88] F. Koppelman. *Travel prediction with models of individual choice behavior.* PhD thesis, Massachusetts Institute of Technology, 1975.

[89] U. Wilke, F. Haldi, and D. Robinson. A model of occupants' activities based on time use survey data. In *Proceedings of Building Simulation*, 2011.

[90] U. Wilke, F. Haldi, J.-L. Scartezzini, and D. Robinson. A bottom-up stochastic model to predict building occupants' time-dependent activities. *Building and Environment*, 60(0):254 – 264, 2013. ISSN 0360-1323. doi: 10.1016/j.buildenv.2012.10.021.

[91] H. Allcott and S. Mullainathan. Behavior and energy policy. *Science*, 327 (5970):1204–1205, 2010.

[92] C. Bhat and F. Koppelman. A retrospective and prospective survey of time-use research. *Transportation*, 26(2):119–139, 1999.

[93] J. Gliebe and F. Koppelman. A model of joint activity participation between household members. *Transportation*, 29(1):49–72, 2002.

[94] J. Zhang, H. Timmermans, and A. Borgers. A model of household task allocation and time use. *Transportation Research Part B: Methodological*, 39(1):81–95, 2005.

[95] C. Bhat. A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological*, 39(8):679–707, 2005.

[96] H. Huang, Z. Li, W. Lam, and S. Wong. A time-dependent activity and travel choice model with multiple parking options. In *Transportation and Traffic Theory. Flow, Dynamics and Human Interaction. 16th International Symposium on Transportation and Traffic Theory*, 2005.

[97] T. Golob. A simultaneous model of household activity participation and trip chain generation. *Transportation Research Part B: Methodological*, 34 (5):355–376, 2000.

[98] D. Helbing. *Quantitative sociodynamics: stochastic methods and models of social interaction processes*. Springer, 2011.

[99] J. Gershuny and O. Sullivan. The sociological uses of time-use diary analysis. *European Sociological Review*, 14(1):69–85, Mar 1998.

[100] R. McCleary, R. Hay, E. Meidinger, D. McDowall, and K. Land. *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA, 1980.

[101] I. Fischer and O. Sullivan. Evolutionary modeling of time-use vectors. *Journal of Economic Behavior & Organization*, 62(1):120–143, 2007.

[102] L. Feldman and J. Hornik. The use of time: An integrated conceptual model. *Journal of Consumer Research*, pages 407–419, 1981.

[103] M. Basner, K. Fomberstein, F. Razavi, S. Banks, J. William, R. Rosa, and D. Dinges. American time use survey: Sleep time and its relationship to waking activities. *Sleep*, 30(9):1085, 2007.

[104] Y. Lee and L. Waite. Husbands' and wives' time spent on housework: A comparison of measures. *Journal of Marriage and Family*, 67(2):328–336, 2005.

[105] L. Craig. Does father care mean fathers share? *Gender & Society*, 20(2): 259–281, 2006.

[106] S. Daniels, I. Glorieux, J. Minnen, and T. Tienoven. More than preparing a meal? Concerning the meanings of home cooking. *Appetite*, 2012.

[107] J. Torriti. Demand Side Management for the European Supergrid: Occupancy variances of European single-person households. *Energy Policy*, 44: 199–206, May 2012. ISSN 0301-4215. doi: 10.1016/j.enpol.2012.01.039.

[108] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi. A bottom-up approach to residential load modeling. *IEEE Transactions on Power Systems*, 9(2):957–964, 1994.

[109] V. Tabak and B. de Vries. Methods for the prediction of intermediate activities by office occupants. *Building and Environment*, 45(6):1366–1372, 2010.

[110] M. Schweiker, F. Haldi, M. Shukuya, and D. Robinson. Verification of stochastic models of window opening behaviour for residential buildings. *Journal of Building Performance Simulation*, 5(1):55–74, 2012.

[111] A. Bortz, M. Kalos, and J. Lebowitz. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics*, 17(1): 10–18, 1975.

[112] M. Ben-Akiva, M. Bierlaire, D. Bolduc, and J. Walker. *Discrete Choice Analysis, (draft version)*. 2010.

[113] K. Train. *Discrete choice methods with simulation*. Cambridge University Press, 2003.

[114] M. Bierlaire. Biogeme: a free package for the estimation of discrete choice models. *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.*, 2003.

[115] M. Bierlaire. Estimation of discrete choice models with BIOGEME 1.8, User's manual. `biogeme.epfl.ch`, 2008.

[116] M. Bierlaire. Biogeme website. `biogeme.epfl.ch`, March 2012. [Online; accessed 14 March 2012].

[117] P. Rickwood. Residential Operational Energy Use. *Urban Policy and Research*, 27(2):137–155, 2009. ISSN 0811-1146. doi: 10.1080/08111140902950495.

[118] B. Ang, T. Goh, and X. Liu. Residential electricity demand in Singapore. *Energy*, 17(1):37–46, Jan 1992. ISSN 0360-5442. doi: 10.1016/0360-5442(92)90031-T.

[119] J. O'Doherty, S. Lyons, and R. S. J. Tol. Energy-using appliances and energy-saving features: Determinants of ownership in Ireland. *Applied Energy*, 85(7):650–662, Jul 2008. ISSN 0306-2619. doi: 10.1016/j.apenergy.2008.01.001.

[120] E. Leahy and S. Lyons. Energy use and appliance ownership in Ireland. *Energy Policy*, 38(8):4265–4279, Aug 2010. ISSN 0301-4215. doi: 10.1016/j.enpol.2010.03.056.

[121] S. Roberts. Demographics, energy and our homes. *Energy Policy*, 36(12): 4630–4632, Dec 2008. ISSN 0301-4215. doi: 10.1016/j.enpol.2008.09.064.

[122] R. Madlener and M. Harmsen-van Hout. 10 Consumer behavior and the use of sustainable energy. *Handbook of Sustainable Energy*, page 181, 2011.

[123] C.-h. Lin. A model using home appliance ownership data to evaluate recycling policy performance. *Resources, Conservation and Recycling*, 52(11): 1322–1328, Sep 2008. ISSN 0921-3449. doi: 10.1016/j.resconrec.2008.07.015.

[124] A. Curran, I. D. Williams, and S. Heaven. Management of household bulky waste in England. *Resources, Conservation and Recycling*, 51(1):78–92, Jul 2007. ISSN 0921-3449. doi: 10.1016/j.resconrec.2006.08.003.

[125] M. Bittman, J. Rice, and J. Wajcman. Appliances and their impact: the ownership of domestic technology and time spent on household work. *British Journal of Sociology*, 55(3):401–423, Sep 2004. ISSN 0007-1315. doi: 10.1111/j.1468-4446.2004.00026.x.

[126] V. Fernandez. Decisions to replace consumer durables goods: An econometric application of wiener and renewal processes. *Review of Economics and Statistics*, 82(3):452–461, Aug 2000. ISSN 0034-6535. doi: 10.1162/003465300558948.

[127] A. Prinzie and D. Van den Poel. Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems*, 44(1):28–45, Nov 2007. ISSN 0167-9236. doi: 10.1016/j.dss.2007.02.008.

[128] M. A. McNeil and V. E. Letschert. Modeling diffusion of electrical appliances in the residential sector. *Energy and Buildings*, 42(6):783–790, Jun 2010. ISSN 0378-7788. doi: 10.1016/j.enbuild.2009.11.015.

[129] I. Matsukawa and N. Ito. Household ownership of electric room air conditioners. *Energy Economics*, 20(4):375–387, Sep 1998. ISSN 0140-9883. doi: 10.1016/S0140-9883(97)00011-X.

[130] A. Aguilera, M. Escabias, and M. Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8):1905–1924, 2006.

[131] J. Jackson and J. Wiley. *A user's guide to principal components*, volume 19. Wiley Online Library, 1991.

[132] I. Camminatiello and A. Lucadamo. Estimating multinomial logit model with multicollinear data. *MTISD 2008. Methods, Models and Information Technologies for Decision Support Systems*, 1(1):51–54, 2008.

[133] R. Schaefer. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1-2): 75–91, 1986.

[134] L. Barker and C. Brown. Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine*, 20(9-10):1431–1442, May 2001. ISSN 0277-6715. doi: 10.1002/sim.680.

[135] H. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958. ISSN 0033-3123. doi: 10.1007/BF02289233.

[136] B. Marx. A continuum of principal component generalized linear regressions. *Computational Statistics & Data Analysis*, 13(4):385–393, 1992.

[137] J. Nipkow, S. Gasser, and E. Bush. Der typische Haushalt-Stromverbrauch [The typical power consumption of households]. *Bulletin - Association pour l'électrotechnique les technologies de l'énergie et de l'information*, 98(19): 24, 2007.

[138] Verband Schweizerischer Elektrizitätsunternehmen. `www.strom.ch`, August 2012. [Online; accessed 6 August 2012].

[139] Investigation of rental price statistics of the Swiss Federal Statistical Office. `http://www.bfs.admin.ch/bfs/portal/de/index/themen/05/06/blank/key/einfuehrung.html`, January 2012. [Online; accessed 9 August 2012]; (in German).

[140] MATLAB. *version 7.8.0.347 (R2009a)*. The MathWorks Inc., Natick, Massachusetts, 2009.

[141] R. Davidson and J. MacKinnon. Several tests for model-specification in the presence of alternative hypotheses. *Econometrica*, 49(3):781–793, 1981. ISSN 0012-9682. doi: 10.2307/1911522.

[142] K. Vos and M. Zaidi. Equivalence scale sensitivity of poverty statistics for the member states of the European Community. *Review of Income and Wealth*, 43(3):319–333, 1997.

[143] Household size statistics of the Swiss Federal Statistical Office. `http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/04/blank/key/haushaltsgroesse.html`, July 2010. [Online; accessed 23 August 2012];.

[144] I. Mansouri, M. Newborough, and D. Probert. Energy consumption in UK households: Impact of domestic electrical appliances. *Applied Energy*, 54 (3):211 – 285, 1996. ISSN 0306-2619. doi: 10.1016/0306-2619(96)00001-3.

[145] S. Firth, K. Lomas, A. Wright, and R. Wall. Identifying trends in the use of domestic appliances from household electricity consumption measurements.

Energy and Buildings, 40(5):926 – 936, 2008. ISSN 0378-7788. doi: 10.1016/j.enbuild.2007.07.005.

[146] I. Schick, P. Usoro, M. Ruane, and J. Hausman. Residential end-use load shape estimation from whole-house metered data. *Power Systems, IEEE Transactions on*, 3(3):986 –991, Aug 1988. ISSN 0885-8950. doi: 10.1109/59.14551.

[147] J. Jardini, C. Tahan, M. Gouvea, S. Ahn, and F. Figueiredo. Daily load profiles for residential, commercial and industrial low voltage consumers. *Power Delivery, IEEE Transactions on*, 15(1):375 –380, Jan 2000. ISSN 0885-8977. doi: 10.1109/61.847276.

[148] R. Yao and K. Steemers. A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings*, 37(6):663 – 671, 2005. ISSN 0378-7788. doi: 10.1016/j.enbuild.2004.09.007.

[149] J. V. Paatero and P. D. Lund. A model for generating household electricity load profiles. *International Journal of Energy Research*, 30(5):273–290, 2006. ISSN 1099-114X. doi: 10.1002/er.1136.

[150] M. M. Armstrong, M. C. Swinton, H. Ribberink, I. Beausoleil-Morrison, and J. Millette. Synthetically derived profiles for representing occupant-driven electric loads in Canadian housing. *Journal of Building Performance Simulation*, 2(1):15–30, 2009. doi: 10.1080/19401490802706653. URL `http://www.tandfonline.com/doi/abs/10.1080/19401490802706653`.

[151] Natural Resources Canada. `http://www.nrcan.gc.ca/home`. [Online; accessed 12 February 2013].

[152] L. G. Swan, V. I. Ugursal, and I. Beausoleil-Morrison. Occupant related household energy consumption in Canada: Estimation using a bottom-up neural-network technique. *Energy and Buildings*, 43(2–3):326 – 337, 2011. ISSN 0378-7788. doi: 10.1016/j.enbuild.2010.09.021.

[153] J. Dickert and P. Schegner. A time series probabilistic synthetic load curve model for residential customers. In *PowerTech, 2011 IEEE Trondheim*, pages 1 –6, june 2011. doi: 10.1109/PTC.2011.6019365.

[154] J. Tanimoto and A. Hagishima. State transition stochastic model for predicting off to on cooling schedule in dwellings as implemented using a multilayered artificial neural network. *Journal of Building Performance Simulation*, (1):1–9, 2011. doi: 10.1080/19401493.2010.533388.

[155] J. Tanimoto, A. Hagishima, and H. Sagara. Validation of probabilistic methodology for generating actual inhabitants' behavior schedules for accurate prediction of maximum energy requirements. *Energy and Buildings*, 40(3):316 – 322, 2008. ISSN 0378-7788. doi: 10.1016/j.enbuild.2007.02.032.

[156] A. Wright and S. Firth. The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations. *Applied Energy*, 84(4):389 – 403, 2007. ISSN 0306-2619. doi: 10.1016/j.apenergy.2006.09.008.

[157] Informations about the IRISE campaign in the REMODECE website. `http://www2.isr.uc.pt/~remodece/database/Campaign_Irise.htm`.

[158] SARL Enertech & Cabinet Sidler, 26160 Félines sur Rimandoule, France. `http://www.enertech.fr/`.

[159] Électricité de France S.A., France. `www.edf.com/`.

[160] Residential Monitoring to Decrease Energy Use and Carbon Emissions in Europe. `http://remodece.isr.uc.pt/`.

[161] A. de Almeida, P. Fonseca, B. Schlomann, and N. Feilberg. Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations. *Energy and Buildings*, 43(8): 1884 – 1894, 2011. ISSN 0378-7788. doi: 10.1016/j.enbuild.2011.03.027.

[162] A. Ihbal, H. Rajamani, R. Abd-Alhameed, and M. Jalboub. Statistical predictions of electric load profiles in the UK domestic buildings. In *Energy, Power and Control (EPC-IQ), 2010 1st International Conference on*, pages 345 –350, 2010.

[163] J. Barton, S. Huang, D. Infield, M. Leach, D. Ogunkunle, J. Torriti, and M. Thomson. The evolution of electricity demand and the role for demand side participation, in buildings and transport. *Energy Policy*, 52(0):85 – 102, 2013. ISSN 0301-4215. doi: 10.1016/j.enpol.2012.08.040.

[164] C. F. Reinhart. Lightswitch-2002: A model for manual and automated control of electric lighting and blinds. *Solar Energy*, 77(1):15 – 28, 2004. ISSN 0038-092X. doi: 10.1016/j.solener.2004.04.003.

[165] J. Kharoufeh. *Density estimation for functions of correlated random variables*. PhD thesis, Ohio University, 1997.

[166] P. Parikh, M. Kanabar, and T. Sidhu. Opportunities and challenges of wireless communication technologies for smart grid applications. In *Power and Energy Society General Meeting, 2010 IEEE*, pages 1 –7, Jul 2010. doi: 10.1109/PES.2010.5589988.

[167] R. Stamminger, G. Broil, C. Pakula, H. Jungbecker, M. Braun, I. Rüdenauer, and C. Wendker. Synergy potential of smart appliances. *Report of the Smart-A project*, 2008.

Urs Wilke

Avenue d'Yverdon 5
1004 Lausanne
Switzerland

German
April 21st, 1981
0041 78 928 76 49
urs.wilke@gmail.com

Education

2009 - 2013	**PhD** at Solar Energy and Building Physics Laboratory, EPFL, Switzerland. *Probabilistic Bottom-Up Modelling of Occupancy and Activities to Predict Electricity Demand in Residential Buildings.*
2001 - 2008	**Physics Diploma**, University of Hamburg, Germany. Subsidiary subjects: Chemistry, Electronics, Mathematics (representation theory, functional analysis and algebra). Diploma thesis in Spintronics group at Institute of Applied Physics, *Magnetization Dynamics of One- Dimensional Chains Studied by Kinetic Monte Carlo Simulations.*
2000 - 2001	**Studies** in information engineering, French linguistics, media sciences and history at University of Konstanz, Germany.
1991 - 2000	**Highschool certificate** (Abitur) at the Geschwister-Scholl-Schule in Konstanz, Germany

Professional Experience

2009 - 2013 **Research and teaching assistant**, EPFL, Solar Energy & Building Physics Laboratory (Lausanne, Switzerland)

- Development of dynamic stochastic models to predict residential occupancy & activities, as well as the ownership and the use of individual electrical appliances.
- Work for the project *A Bottom-Up Model of Energy Flows to Investigate Strategies Leading to a 2000 W City* of the Swiss National Science Foundation.
- Teaching assistant for the course *Building Physics* (2009 - 2012).
- Direction of the teaching assistants of the course *Building Physics* (2011 - 2012).

2008 - 2009 **Research associate** in the Spintronics theory group of Prof Dr Stefan Heinze, University of Hamburg

- Investigation of the spin dynamics of magnetic systems based on statistical mechanics approaches.
- Publication writing of the expansion of the Glauber model due to the inclusion of temperature and finite-size effects and the comparison to experimental results.

2007 **Assistant** of the exercises of the course *Solid State Physics*, University of Hamburg.

Languages

German	Mother tongue
English	Fluent
French	Fluent

Further qualifications

Statistical modelling	Stochastic processes, behavioural modelling, discrete choice analysis, survival analysis, kinetic Monte Carlo simulations, principal component analysis.
General physics	Solid state physics, surface physics, statistical physics.
Computer science	- **Operating systems**: Linux (Ubuntu, Suse), Unix, Windows. - **Office**: Word, Excel, Power Point, LaTeX. - **Other**: Matlab, Python, Biogeme, SQL, C, C++, Java.

Publications

Refereed journal articles

U. Wilke, F. Haldi, J.-L. Scartezzini, and D. Robinson. A bottom-up stochastic model to predict building occupants' time-dependent activities. *Building and Environment*, 60(0):254 – 264, 2013.

D. Robinson, U. Wilke, and F. Haldi. Multi-agent stochastic simulation of occupants' presence, activities and behaviours. *Building Research and Information*, (in press) 2012.

Conference articles

U. Wilke, F. Haldi, and D. Robinson. A model of occupants' activities based on time use survey data. In *Proceedings of Building Simulation*, 2011.

U. Wilke, F. Haldi, and D. Robinson. Stochastic activity modelling in residential buildings. In *Proceedings of Cisbat 2011, Lausanne, Switzerland*, 2011.

U. Wilke, M. Papadopoulou, and D. Robinson. Towards a 2kw city, the case of zurich. In *Proceedings of of World Renewable Energy Congress 2011, Linköping, Sweden*, 2011.

D. Robinson, U. Wilke, and F. Haldi. Multi agent simulation of occupants' presence and behaviour. In *Proceedings of Building Simulation*, 2011.

N. Filchakova, U. Wilke, and D. Robinson. Energy modelling of city housing stock and its temporal evolution. In *Proceedings of Eleventh International IBPSA Conference, Glasgow, Scotland*, 2009.

D. Robinson, F. Haldi, J. Kämpf, P. Leroux, D. Perez, A. Rasheed, and U. Wilke. Citysim: Comprehensive micro-simulation of resource flows for sustainable urban planning. In *Proceedings of Building Simulation*, 2009.

Monographs

U. Wilke. *Probabilistic Bottom-Up Modelling of Occupancy and Activities to Predict Electricity Demand in Residential Buildings.* PhD thesis, EPFL, Lausanne, 2013.

U. Wilke. Magnetization Dynamics of One-Dimensional Chains Studied by Kinetic Monte Carlo Simulations. Diploma thesis, Institute of Applied Physics, University of Hamburg, 2008.