

Diss. ETH No. 22632

Occupancy Sensing and Prediction for Automated Energy Savings

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
WILHELM KLEIMINGER
MEng (Hons) in Computing, Imperial College London
born on 8 September 1987
citizen of Germany

accepted on the recommendation of
Prof. Dr. Friedemann Mattern, examiner
Prof. Dr. H.-Jürgen Appelrath, co-examiner
Prof. Dr. Anind Dey, co-examiner
Prof. Dr. Silvia Santini, co-examiner

2015

Abstract

The ability to sense and predict occupancy – *i.e.* to establish when the residents are and will be in a building – represents a basic requirement for the energy-efficient operation of many building automation systems. In residential households, in particular, the absence of all residents allows a heating controller to automatically lower the temperature of the home, thereby saving energy that would have been otherwise wasted on heating an empty building. However, if the home has been thus allowed to cool, a boiler and heat distribution system need a non-negligible time to reheat the home to a comfortable temperature. Therefore, to avoid a loss of comfort, a heating control system also requires a sufficiently accurate prediction of when the occupants are going to return in order to trigger the heating at the right time. Since space heating accounts for a large fraction of residential energy use (*e.g.* 68% in the European Union member states), heating control systems based on occupancy sensing and prediction – often referred to as smart thermostats – play an important role in reducing energy consumption and carbon dioxide emissions, while at the same time ensuring occupant comfort.

The objective of this thesis is thus to investigate how the two main computational components of a smart thermostat – occupancy *sensing*, based on sensors that typically exist in a residential environment, as well as occupancy *prediction* from historical occupancy patterns – can be used to automatically reduce the energy consumption of a heating system while trying to maximise thermal comfort.

Current smart thermostats require the installation of dedicated hardware to sense whether the occupants are at home or away. This increases installation and maintenance costs and thus prevents widespread adoption of such potentially energy-saving solutions. To overcome this hurdle, we investigate the suitability of opportunistically using devices already existing in households to sense occupancy. This opportunistic sensing approach seeks to utilise available devices to replace or augment dedicated infrastructures. An example are smart electricity meters, which are mandated to be installed in many households worldwide. We hypothesise that the information contained in the electrical load of the

household, as measured by the smart electricity meter, can be used to infer its occupancy. To verify this hypothesis, we have performed an extensive data collection campaign over seven months in six Swiss households to collect occupancy ground truth data as well as the aggregated and device-level electrical consumption of the households. Using this data, we employ supervised machine learning algorithms to infer occupancy solely from the households' aggregated electricity consumption. We show that such an approach yields a classification accuracy of up to 94%.

As soon as the occupancy *sensing* infrastructure detects that residents left the house, the temperature can be allowed to drop resulting in energy savings during this setback period. However, a reactive strategy cannot be employed upon the arrival of the occupants as it may take a considerable amount of time to bring the house back to a comfortable temperature. To avoid loss of comfort, occupancy *prediction* algorithms are used to predict the time of arrival of the occupants to determine the right time to start pre-heating the house. To analyse the performance of such prediction approaches we have derived occupancy schedules from a large, publicly available mobile phone location dataset. Using the schedules from 45 participants we show that current state-of-the art occupancy prediction algorithms achieve an accuracy around 85%, which is close to the theoretical optimum given by the *predictability* of the schedules (which in practice always feature some level of irregular behaviour).

The accuracy of the occupancy prediction alone does not necessarily reflect the energy savings and comfort loss that can be achieved or caused by a smart thermostat. The actual savings depend upon the occupancy schedule of the household, the prediction accuracy, the weather conditions and the physical properties of the building. The final part of this thesis thus deals with the simulation of various heating scenarios to investigate the effect of a smart thermostat on the overall energy savings under different environmental conditions. To this end, we assess the overall energy expenditure in several building scenarios. Furthermore, we develop a new methodology to accurately assess the impact of the weather conditions on the energy savings. We show that building parameters result in a range of savings from 6% to 17%, while the savings in the 25% of households with the lowest occupancy are 4-5 times higher than in the quarter with the highest occupancy.

The unifying theme of this thesis is to show how current technology, which already exists in many homes, can help to save energy without sacrificing comfort. For this purpose, we draw upon recent work in the distributed systems domain to access smart electricity meters and machine learning algorithms to derive occupancy data. We show how predictable occupancy schedules are and, by providing a simulation framework to evaluate different occupancy prediction algorithms, we seek to answer the question how much energy a smart thermostat can save.

Kurzfassung

Grundvoraussetzung für eine energieeffiziente Steuerung von Gebäudeprozessen sind Erkennung und Vorhersage der Anwesenheit von Bewohnern in den Gebäuden. In privaten Haushalten erlaubt die Abwesenheit der Bewohner einem intelligenten Heizungsregelungssystem beispielsweise, die Temperatur zu senken und somit Energie einzusparen, die ansonsten für das Heizen des unbewohnten Wohnraums verschwendet werden würde. Allerdings benötigt ein ausgekühltes Haus wieder ausreichend Zeit, um auf eine angenehme Temperatur aufgeheizt zu werden. Um Komforteinschränkungen zu vermeiden, ist für das Heizungsregelungssystem daher eine möglichst genaue Vorhersage über die zukünftige Anwesenheit der Bewohner erforderlich. Die Raumheizung stellt einen signifikanten Teil des Gesamtenergieverbrauchs privater Haushalte dar, in der Europäischen Union liegt er derzeit bei 68%. Daher können Heizungsregelungssysteme, welche auf Anwesenheitserkennung und -vorhersage der Bewohner beruhen, eine wichtige Rolle bei der Reduktion von Energieverbrauch und CO₂-Emission spielen, ohne dass sich Komforteinschränkungen für die Bewohner ergeben.

Das Ziel dieser Arbeit ist die Untersuchung, ob Anwesenheitserkennung – basierend auf Sensoren, die typischerweise in Haushalten bereits vorhanden sind – sowie Anwesenheitsvorhersage anhand historischer Anwesenheitsmuster dazu beitragen können, den Energieverbrauch eines Heizungssystems ohne Komforteinbussen zu senken.

Aktuelle intelligente Heizungsregelungssysteme setzen die Installation von dedizierter Hardware für die Anwesenheitserkennung voraus. Die sich daraus ergebenden Zusatzkosten für deren Einbau und Wartung sorgen dafür, dass solche Lösungen eher Nischenprodukten vorbehalten bleiben. Dieses Problem könnte sich durch eine „Zweckentfremdung“ bereits existierender Haushaltsgeräte für die Anwesenheitserkennung mildern lassen. Ein Beispiel eines solchen Ansatzes sind intelligente Stromzähler, die in vielen Haushalten durch Änderungen in der Gesetzgebung bereits zur Pflicht geworden sind. Wir stellen hierbei die These auf, dass die elektrische Lastkurve eines Haushalts genügend

Informationen beinhaltet, um daraus mit hoher Wahrscheinlichkeit die An- und Abwesenheit seiner Bewohner abzuleiten. Um diese These zu untersuchen, haben wir in sechs Schweizer Haushalten Messtechnik installiert, die den Gesamtstromverbrauch sowie den Verbrauch einzelner Geräte misst. Zusätzlich haben wir die „Ground Truth“ bezüglich der tatsächlichen Anwesenheit in diesen Haushalten aufgenommen. Wir zeigen, dass mit Hilfe von überwachtem maschinellern Lernen, basierend auf den Stromverbrauchsdaten, eine Genauigkeit von bis 94% bei der Erkennung von An- und Abwesenheit möglich ist.

Eine automatisierte Anwesenheitserkennung ermöglicht somit einen Effizienzgewinn beim Heizen. Haben die Bewohner das Haus verlassen, ist eine weitere Beheizung des Wohnraums nicht notwendig, das Heizungsregelungssystem kann die Innentemperatur auf einen tieferen Wert absinken lassen. Allerdings kann solch eine rein reaktive Steuerung nicht verwendet werden, um das Haus erst bei der Ankunft der Anwohner wieder aufzuheizen, da die Aufheizphase eine signifikante Zeit erfordert.

Um den richtigen Zeitpunkt für das Wiederaufheizen des Hauses zu bestimmen, können Algorithmen zur Anwesenheitsvorhersage genutzt werden. Damit kann die Aufheizphase bereits vor Rückkehr der Bewohner gestartet werden, und Komforteinbussen werden vermieden. Um solche Anwesenheitsvorhersagealgorithmen zu analysieren, haben wir Anwesenheitsdaten aus einem öffentlich verfügbaren Datensatz extrahiert und analysiert. Mit Hilfe der Daten von 45 Teilnehmern zeigen wir, dass aktuelle Algorithmen Genauigkeiten aufweisen, welche nur durch die prinzipielle Vorhersagbarkeit der nicht ganz regelmässigen täglichen Routine begrenzt werden.

Natürlich kann die Genauigkeit der Vorhersagealgorithmen nicht direkt den Effizienzgewinn und die Komforteinbussen eines intelligenten Heizungsregelungssystems abbilden. Die tatsächliche Energieersparnis hängt von einigen weiteren Faktoren wie den klimatischen Gegebenheiten der Region sowie den physikalischen Eigenschaften des Gebäudes ab. Daher untersuchen wir im letzten Teil dieser Arbeit den Effekt einer intelligenten Heizungssteuerung auf den Heizenergieverbrauch unter verschiedenen Kontextbedingungen. Für die Analyse entwickeln wir mehrere Gebäudeszenarien sowie eine Methode, um die Effekte von unterschiedlichen Wetterbedingungen auf die Gesamtheizenergie zu untersuchen.

Das Leitmotiv dieser Arbeit besteht darin, beim Heizen unter Ausnutzung von Technologien, die in vielen Haushalten bereits zum Alltag gehören, einen Effizienzgewinn zu erzielen. Um dieses Ziel zu erreichen, nutzen wir technologische und erkenntnisbezogene Fortschritte in den Bereichen verteilter Systeme und maschinellern Lernen, um Anwesenheitsinformationen aus elektrischen Lastkurven zu extrahieren. Ergänzend analysieren wir die prinzipielle Vorhersagbarkeit von Anwesenheit aus historischen Daten und zeigen welche Energieersparnis intelligente Heizungsregelungssysteme ermöglichen können.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my advisor, Professor Friedemann Mattern, who has provided me with his thoughts and guidance throughout the last five years. His support and interest in my research have provided me with great sources of motivation. In the same spirit, I would like to say “grazie mille” to Professor Silvia Santini for her invaluable mentoring and support throughout my PhD.

I would also like to extend my appreciation to my committee members Professor Hans-Jürgen Appelrath and Professor Anind Dey for their advice on my work. I am especially grateful to Professor Dey for being an excellent host during my three months at Carnegie Mellon University in Pittsburgh.

Large parts of this research could not have been done without the help of others. Above all, I would like to thank my colleague and friend Christian Beckel for our collaboration. Without Christian’s support, the collection of the ECO dataset would have been impossible. I would also like to thank to our partners and participants at Energie Thun, who welcomed us to their homes and whose extraordinary commitment made the data collection a success. Over the last couple of years, I was also able to work with a number other researchers. In particular, I would like to thank Paul Baumann and Christian Köhler for our joint work.

I also want to thank my friends and colleagues at the Institute for Pervasive Computing: Gábor Sörös and his wife Zsófia for many nice evenings, barbecues and a memorable trip to Vienna; Simon Mayer for the interesting discussions on all things non-technical – I sincerely enjoyed our disagreements; Hossein Shafagh and Anwar Hithnawi for the unique culinary experiences; Christof Roduner for his refreshing attitude and for first welcoming me in Zurich; Matthias Kovatsch for his support with all things embedded; Leyna Sadamori for many enlightening discussions on Machine Learning; my colleagues Alexander Bernauer, Benedikt Ostermeier, Robert Adelman, Mihai Bâce, Marian George, Vlad Trifa, Dominique Guinard, Christian Flörkemeier, Markus Weiss, Elke Schaper

and Thorsten Staake for many interesting discussions; and finally my students Daniel Pauli, Andreas Dröscher, Sara Kilcher, Andreas Brauchli, Michael Spiegel and Christian Stücklberger.

Last but not least, I would like to extend my deep gratitude to my friends and family: My parents Elke and Jürgen for their continued support and motivation throughout my studies; My sister Lisa for last-minute proofreading and pushing me in the final months; My grandmother Marlies for enduring some comfort loss during various heating experiments; Frank for looking after my physical shape; Malte, without whom I may not have ended up on this path; and Sarah for being very understanding and supportive of my work.

Contents

Acronyms	xv
1 Introduction	1
1.1 Motivation	3
1.2 Research goals and contributions	6
1.2.1 Opportunistic occupancy sensing	7
1.2.2 Classification and analysis of occupancy prediction algorithms . .	8
1.2.3 Analysis of the energy-savings potential of smart heating	8
1.3 Outline of the thesis	8
2 Smart heating	11
2.1 Automatic heating control	12
2.1.1 Occupancy sensing and prediction	12
2.1.2 Optimal control	14
2.1.3 Reactive control	15
2.2 Thermal comfort	16
2.2.1 Static comfort models: Fanger's PMV and PPD	16
2.2.2 Adaptive comfort model	19
2.2.3 Local comfort models	19
2.3 Building constraints	20
2.3.1 Thermal properties	21
3 The ECO dataset	23
3.1 Related research work and datasets	24
3.1.1 Occupancy sensing	24
3.1.2 Datasets containing electricity consumption data	26
3.2 Experimental setup of our occupancy sensing infrastructure	27
3.2.1 Selection of households	27

3.2.2	Overview of the architecture	28
3.2.3	Data collection infrastructure	28
3.2.4	Aggregate electricity consumption	29
3.2.5	Device-level electricity consumption	32
3.2.6	Passive infrared occupancy sensors	34
3.2.7	Occupancy ground truth	34
3.3	Data collection	34
3.3.1	Data formatting	35
3.3.2	Missing data	35
3.4	Description of the dataset	35
3.4.1	A typical day	36
3.4.2	Aggregate electricity consumption	36
3.4.3	Occupancy ground truth	39
3.4.4	Device-level electricity consumption	41
3.4.5	Data cleaning	43
3.5	Conclusions	44
4	Occupancy sensing	45
4.1	Related work	46
4.1.1	Time series analysis	47
4.1.2	Inferring household characteristics from the electric load curve . .	48
4.1.3	Non-intrusive load monitoring	48
4.1.4	Occupancy and the electrical load curve	49
4.2	System design	50
4.2.1	Deriving features from the electrical load curve	53
4.2.2	Cross validation	56
4.2.3	Classifiers	57
4.2.4	Dimensionality reduction	60
4.3	Evaluation	61
4.3.1	Accuracy	62
4.3.2	Matthews Correlation Coefficient	63
4.3.3	False negative and false positive rate	63
4.4	Results	64
4.4.1	Overall occupancy detection performance	64
4.4.2	Performance by classifier	66
4.4.3	Suitability for controlling a thermostat	68
4.4.4	Limits to occupancy sensing using electricity consumption data .	70
4.4.5	Features selected by SFS	71
4.5	Using device-level consumption data	73
4.5.1	Correlation between appliance state and occupancy	73

4.5.2	Detecting activation states	75
4.5.3	Occupancy detection performance	75
4.6	A simple unsupervised approach: Revisiting the mean classifier	76
4.7	Conclusions and lessons learned	78
5	Large-scale occupancy datasets	81
5.1	Related work	82
5.1.1	Significant place sensing	82
5.1.2	Public occupancy and location datasets	83
5.1.3	CDR datasets	84
5.1.4	The Nokia LDCC/MDC dataset	85
5.2	The Homeset Algorithm	85
5.2.1	Occupancy detection using the homeset algorithm	87
5.2.2	Initialisation of the homeset	88
5.3	Evaluation	90
5.4	Conclusion and lessons learned	92
6	Occupancy prediction	93
6.1	A classification of occupancy prediction approaches	94
6.1.1	Schedule-based approaches	95
6.1.2	Context-aware approaches	99
6.1.3	Hybrid approaches	100
6.1.4	Other approaches	100
6.2	Experimental setup	101
6.2.1	Algorithm implementations	102
6.2.2	Preparing the LDCC occupancy schedules	103
6.3	Evaluation	104
6.3.1	Performance measures	104
6.3.2	Cross-validation	105
6.4	Results	105
6.4.1	Prediction accuracy	105
6.4.2	Parameter selection	107
6.4.3	Learning time and prediction accuracy	107
6.4.4	Limits of predictability	108
6.5	Conclusions and lessons learned	109
7	Simulation model	111
7.1	Related work	112
7.2	Lumped capacitance models	113
7.2.1	A simple resistance-capacitance (1R1C) model	114

7.2.2	Limitations of the 1R1C model	116
7.2.3	The ISO 13790 5R1C model	116
7.3	Weather scenarios	118
7.3.1	Annual model	120
7.3.2	Global weather scenarios	121
7.4	Building configurations	122
7.4.1	Transmission losses	124
7.4.2	Design heat load	126
7.4.3	Ventilation losses	127
7.4.4	Internal gains	129
7.4.5	Solar gains	129
7.5	Controller design	133
7.6	Limitations	134
7.7	Conclusions and lessons learned	137
8	Smart Thermostats: How much do they save?	139
8.1	Related work	140
8.1.1	Model predictive control	141
8.1.2	Other control strategies	142
8.2	Discussion	144
8.3	Experimental setup	145
8.3.1	Building model and simulation setup	145
8.3.2	Heating controller	146
8.4	Evaluation	148
8.4.1	Efficiency gain and comfort loss	148
8.5	Results	149
8.5.1	Efficiency gain	150
8.5.2	Heating degree hours	150
8.5.3	Annualised savings	151
8.5.4	Impact of climate conditions	152
8.5.5	Impact of the occupancy schedules	152
8.6	Modelling limitations	154
8.6.1	Building model	154
8.6.2	Baseline metrics	154
8.7	Conclusions and lessons learned	155
9	Conclusions and outlook	157
9.1	Opportunistic occupancy sensing	157
9.2	Occupancy prediction	159
9.3	Energy savings	159

9.4	Future work	160
9.4.1	Sensing	160
9.4.2	Prediction	161
9.4.3	Control	162
9.4.4	Evaluation of energy savings	162
Bibliography		165
Referenced Web Resources		185
A Questionnaires		189
B 1-Resistance 1-Capacitance (1R1C) model		191
B.1	Derivation	191
B.2	Convergence	193
C Occupancy prediction		195
C.1	Dataset overview and prediction results	195
C.2	Probabilistic schedules	197
D Simulation scenarios		203

Acronyms

ADOT average number of daily occupancy transitions.

AIC Akaike information criterion.

ANN artificial neural network.

ANOVA analysis of variance.

AP access point.

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineers.

BfE Swiss Federal Office for Energy.

BMS building management system.

CDR call detail record.

CoAP Constrained Application Protocol.

CTRW continuous-time random-walk.

D4D data for development.

DFT discrete Fourier transform.

DIN Deutsches Institut für Normung.

DPM Dirichlet process mixtures.

DST Dempster–Shafer theory.

DWT discrete wavelet transform.

Acronyms

EC European Council.

ECO Electricity Consumption and Occupancy.

EN European Norm.

EnEV Energieeinsparverordnung.

EU European Union.

FNR false negative rate.

FPR false positive rate.

GMM Gaussian mixture model.

GPS Global Positioning System.

GSM Global System for Mobile Communications.

HMM hidden Markov model.

HS homeset.

HSD honest significant difference.

HTTP Hypertext Transfer Protocol.

HVAC heating, ventilation and cooling.

ICMP Internet Control Message Protocol.

IEC International Electrotechnical Commission.

IECC International Energy Conservation Code.

IoT Internet of Things.

IP Internet Protocol.

IPv6 Internet Protocol version 6.

ISO International Organization for Standardization.

KNN K-nearest neighbour.

LDCC Lausanne data collection campaign.

LED	light emitting diode.
MAC	media access control.
MAT	Mean Arrival Time.
MCC	Matthews correlation coefficient.
MDC	Mobile data challenge.
MDMAT	Minimum Distance Mean Arrival Time.
MPC	model predictive control.
NILM	non-intrusive load monitoring.
NT	Neurothermostat.
NTP	Network Time Protocol.
OBIS	Object Identification System.
OPT	optimal predictive controller.
PCA	principal component analysis.
PDA	personal digital assistant.
PH	Preheat.
PID	proportional-integral-derivative.
PIR	passive infrared.
PMV	Predicted Mean Vote.
PP	Presence Probabilities.
PPD	Predicted Percentage Dissatisfied.
PPS	simplified Presence Probabilities.
RAM	random-access memory.
RC	resistance-capacitance.
REA	reactive controller.

Acronyms

- RF** radio-frequency.
- RFID** radio-frequency identification.
- RMSE** root mean square error.
- ROC** receiver operating characteristic.
- SAX** Symbolic Aggregate ApproXimation.
- SFS** Sequential Forward Selection.
- SI** Système International d’Unités.
- SML** Smart Message Language.
- SMS** Short Message Service.
- ST** Smart Thermostat.
- SVD** singular value decomposition.
- SVM** support vector machine.
- THR** thresholding.
- USB** Universal Serial Bus.
- WiSoC** Wireless System-on-a-Chip.

Introduction and motivation

At the 1933–1934 Chicago World’s Fair, the public caught a first glimpse of what future homes would look like. The “House of Tomorrow” anticipated many modern conveniences now taken for granted¹, including a dishwasher, an electric garage opener and central air conditioning [126]. In terms of home automation, the House of Tomorrow was a sign of things to come.

By that time, heating control had been around for a while. The invention of the first thermostat is commonly credited to Cornelis Drebbel (1572–1633), a dutchman who worked at the courts of King James I and King Charles I. A prolific inventor, Drebbel built an air-conditioning system as well as an incubator for chicken eggs that was able to keep a constant temperature throughout the year [174]. However, it took over 200 years before the concept was commercialised. The first pneumatic thermostat was patented by Warren S. Johnson, in 1883. Two years later, in 1885, the Swiss born inventor Albert M. Butz registered a patent for the first primitive electric thermostat, the *damper flapper* [212]. The two companies that eventually emerged from these days – Johnson Controls and Honeywell – still operate today. In the same year, Hermann Immanuel Rietschel, a German scientist, took “the world’s first chair in ventilation and heating systems at the TH Berlin” [106].

It should take another 60 years to develop what we today generally consider to be a smart home. We define smartness in the context of home automation as the *ability to learn the requirements of the inhabitants in order to actuate the home while increasing comfort and making efficient use of available resources*. One of the first attempts to build such a smart home is the *Neural Network House* introduced by Mozer *et al.* in 1995 [139]. Mozer *et al.* retro-fitted an old schoolhouse with 75 sensors and actuators to control lighting, ventilation and heating. The control strategy is based on artificial neural networks and designed to adapt to the inhabitants’ personal desires. However, containing “nearly five miles of conductor” [2] and requiring a training period of eight months, the

¹It also included a personal aircraft hangar, a prediction that sadly has not come true.

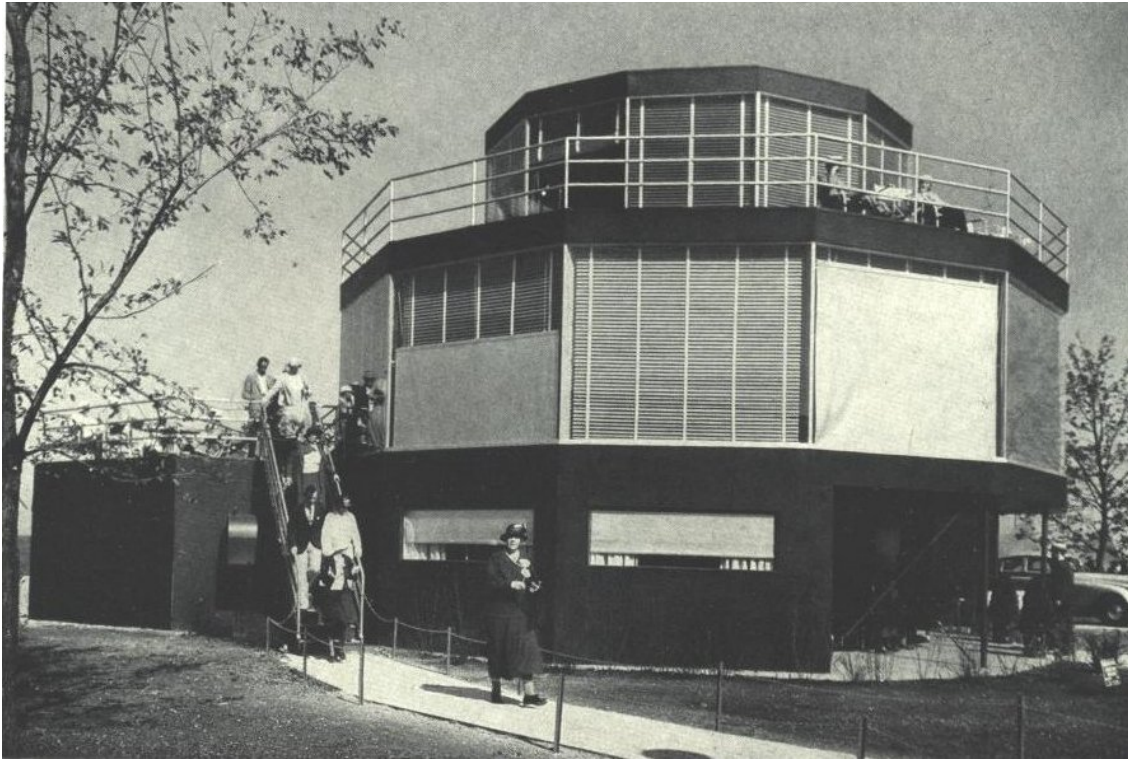


Figure 1.1: House of tomorrow at Chicago World's Fair. Image scanned from original prospectus by Dr. Monica Brooks [201].

technology used in the Neural Network House was still too complex to be applied in the homes of ordinary people.

The trend towards smart homes picked up in earnest with the arrival of the Internet of Things (IoT). The IoT refers to the integration of physical objects from the real world into the virtual world of computers [130]. Thus, while the sales of traditional desktop computers are declining, our environment is becoming more computerised than ever. Advances in microelectronics and wireless communications such as low-power implementations of the IPv6 stack showed that support for the Internet Protocol (IP) could be embedded in small, resource-constrained sensors and actuators [45, 171]. These implementations, low-power Wi-Fi modules and new application layer protocols like the Constrained Application Protocol (CoAP) make connecting physical objects to the Internet feasible at a large scale [78, 111, 153, 170]. Today, an increasing number of appliances such as coffee machines, refrigerators and electricity meters contain microchips and are connected to computer networks [70, 130].

Meanwhile, the abundance of sensors has been identified as a powerful means to reduce the demand for energy [131]. In the home, this has led to *smart thermostats* which automatically control the indoor temperatures based on the actual occupancy of the household. Recently, commercial smart thermostats such as the Nest learning thermostat

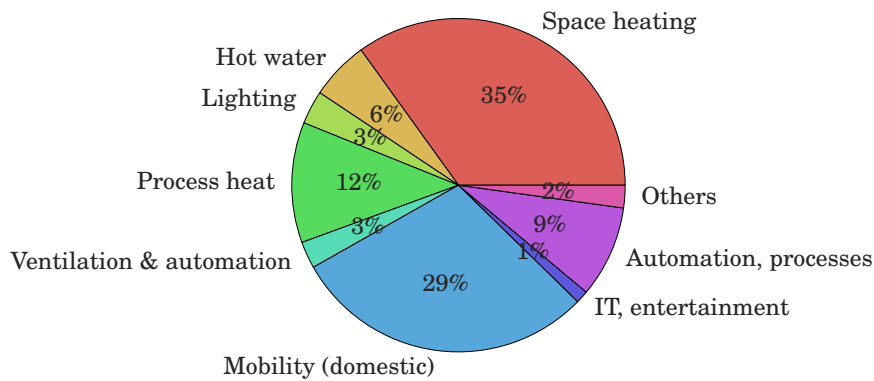


Figure 1.2: Estimated Swiss energy consumption by end use (2013) [93].

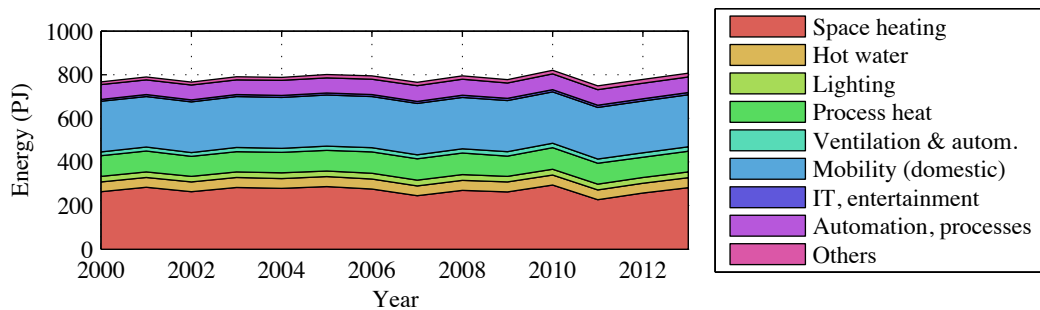


Figure 1.3: Estimated Swiss energy consumption by end use (2000-2013) [20, 93].

and tado° have started to appear [196, 198]. These technologies promise to automatically infer the household's occupancy schedule and enable the occupants to use their mobile phone to track and control the temperature of their home from anywhere in the world.

1.1 Motivation

In Switzerland, central heating in residential buildings is one of the largest contributors to CO₂ emissions and energy bills. According to a study commissioned by the Swiss Federal Office of Energy, 35% of the total domestic energy consumption can be attributed to heating alone [93]. Figure 1.2 shows that heating was the largest single factor influencing the total energy consumption in 2013. Space heating is responsible for five percent more energy consumption than mobility. Furthermore, fluctuations in the energy consumed by heating are much higher than those of the other end uses. Figure 1.3 shows the total domestic energy consumption by end use in Switzerland from 2000 to 2013. Over these 14 years, fluctuations in the energy consumed by space heating (which mainly depends on the outside air temperature in winter) have determined whether the total energy consumption in a given year was higher or lower than previous years.

Table 1.1: Energy consumption in the residential sector by end use. Due to rounding errors, figures may not add up to 100%. Sources: RECS 2009 [222], DECC 2013 [205], BfWE 2012 [202], BFE 2013 [93], EEA 2009 [209], EMSD 2012 [208].

End use	U.S.	U.K.	Germany	Switzerland	EU-27	Hong Kong
Space conditioning	48%	66%	69%	71%	68%	23%
Water heating	18%	17%	15%	13%	12%	19%
Lighting and appliances	35%	18%	16%	16%	19%	58%

Table 1.1 shows that the share of heating of the total energy consumption increases when only residential households are considered. In Switzerland, about 71% of the residential energy consumption can be attributed to space conditioning (*i.e.* heating, ventilation and air-conditioning) [93]. Germany (69%) and the United Kingdom (66%) have similar figures [202, 205]. In these countries, space conditioning mainly requires heating. Across the European Union (EU)², 68% of the residential energy consumption is spent on heating [209]. In the United States, where, due to the different climate zones, space conditioning includes heating, ventilation and air conditioning, the share of the total residential consumption drops to 48% [222]. In Hong Kong, space conditioning accounts only for 23% of the total energy expenditure in households as heating is rarely necessary [208]. The figures show that for moderate climates such as northern Europe, the energy spent on heating is an important factor if the overall energy consumption is to be optimised.

This potential for energy savings in buildings has caused widespread change in legislation regarding building insulation and heating efficiency. In 2010, the EU adopted the 2010/31/EU directive on the Energy Performance of Buildings (EPBD) [156]. The directive requires all member states to establish and enforce “minimum energy performance requirements for new and existing buildings”. Certification of buildings’ energy performance is now mandatory for all new buildings. The certification process consists of measuring or estimating the energy expenditure of a building to be able to draw comparisons to other buildings. In the longer term, the directive requires all new buildings to be “nearly zero-energy buildings”. Similar certification is part of the International Energy Conservation Code (IECC) [213] adopted by many states in the United States.

Energy savings cannot be achieved by focussing on new buildings and renovations alone. In the United States, 60% of homes were built before 1980 [222]. For these homes, programmable thermostats are a reasonable alternative to costly retrofit insulation. Studies have shown that 5% to 15% percent of the energy spent on heating in the United States could be saved by using programmable thermostats [10, 167]. Figures 1.4 and 1.5 show the Honeywell Round *manual thermostat*, which has been produced since the early 1950s, and a modern, *programmable thermostat*. In contrast to the manual thermostat,

²As of the time of writing, the European Union was made up of 28 member states. Croatia joined in 2013 and was thus not included in the statistics.



Figure 1.4: Honeywell Round thermostat [204].

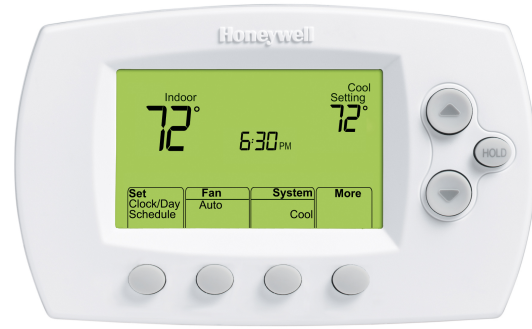


Figure 1.5: Honeywell FocusPRO 6000 programmable thermostat [211].

Table 1.2: Programmable thermostats in the U.S. (N/A: “No Thermostat or Do Not Have or Use Heating Equipment”).

	Yes	No	N/A
Have a programmable thermostat	37%	48%	16%
If yes , reduce temperature during daytime	53%	47%	/
If yes , reduce temperature during nighttime	62%	38%	/

the programmable thermostat allows the occupants to set specific heating schedules. A nighttime setback schedule could for example lower the temperature automatically from 10 p.m. to 6 a.m. by 3 °C in order to save energy. It has been shown that such a nighttime setback saves approximately 1% of heating energy for each degree Fahrenheit forgone [141].

However, a recent survey has shown that only 37% of households in the United States own a programmable thermostat (*cf.* Table 1.2). Moreover, of those who do own such a thermostat, only 53% and 62% use it to reduce the temperature during daytime and nighttime. One reason for this is that programmable thermostats are often too difficult to use. Also, it is often unclear to occupants how much energy could be saved. The potential savings are therefore only realised if the occupants are motivated to save energy in the first place [143, 158]. Thus, when households reduce the temperature, they may not exploit the full savings potential. Only 24% of U.S. households use a setpoint temperature at or below 17 °C during the day when the home is unoccupied [222]. 34% of thermostat owners set the temperature to 21 °C or higher during unoccupied periods. Similarly, only 35% of households have a setpoint below 19 °C during nighttime. Whether the lack of adoption of programmable thermostats is the result of poor interface design [157], the widespread misconception that constant heating is more efficient than a setback schedule [146] or simply the feeling that the potential savings are not worth the extra effort [143], the current state of affairs offers potential for optimisation.

Smart thermostats aim to overcome the drawbacks of manual and programmable ther-

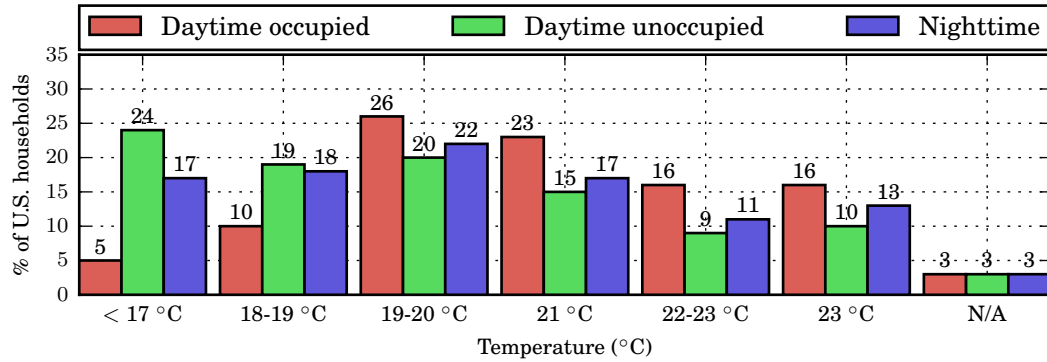


Figure 1.6: Typical setpoint temperatures in U.S. households (N/A: “No Thermostat or Do Not Have or Use Heating Equipment”).

mostats by controlling setpoint temperatures automatically. Instead of requiring the occupants to manually configure a setback schedule, a smart thermostat automatically deduces whether heating is required by *sensing* the occupancy of the household. It reduces the indoor temperature to save energy when nobody is at home and, using occupancy *prediction* algorithms, reheats the house in time for the occupants’ return. This idea to automatically and “intelligently” control heating systems has been investigated for several years. Well-known examples of such *smart heating* approaches include the Neurothermostat [140], the GPS Thermostat [67], the Smart Thermostat [124] and several others [6, 46, 48, 57, 149, 169].

However, the adoption of smart thermostats is impaired by their cost and imprecise savings figures advertised by industry. Existing smart thermostats require additional hardware for sensing the occupancy of the household and controlling the heating and are thus cumbersome and expensive to install. Furthermore, as the actual savings depend on a number of household parameters such as the insulation of the building and its actual occupancy, the projected performance of a smart thermostat may lie well below the figures advertised by vendors. Thus, while it is not clear to the potential buyer how much energy these devices actually save, the cost of additional hardware and uncertainty regarding their amortisation may cause smart thermostats to stay niche products for enthusiasts.

1.2 Research goals and contributions

The goal of this thesis is to provide the technical foundations for the design and evaluation of future smart heating systems. To this end, we investigate the use of opportunistic sensors for detecting occupancy and analyse the energy-savings potential of controlling a heating system using occupancy detection and prediction. We thereby address the following three research questions:

Can existing technology be used opportunistically to sense occupancy? Following recent legislation, smart electricity meters are becoming ubiquitous in many households. Current smart electricity meters report the electricity consumption of a household to a utility company every 15 minutes and often offer the former access to the measured data at 1 Hz for visualisation purposes. This data contains information about the current activity level of the household and could thus be used to detect occupancy. Likewise, many occupants carry mobile phones with Wi-Fi and Global Positioning System (GPS) localisation capabilities from which the occupancy of the household could be derived without requiring to install additional sensors in the household.

How accurately can occupancy be predicted? Accurate occupancy predictions are vital in order to ensure that the home is heated to a comfortable temperature upon the occupants' arrival. Thus, a large number of different approaches for predicting occupancy has recently been proposed in the literature. However, a thorough quantitative analysis of different occupancy prediction algorithms has been missing so far. Furthermore, a lack of analysis of the fundamental limitations of occupancy prediction with respect to the inherent irregularity of human behaviour means that new approaches may only achieve limited improvements over the current state-of-the-art.

How much energy may be saved by a smart heating system using occupancy detection and prediction? Besides the accuracy of the occupancy detection and prediction infrastructure, the energy required for heating a building depends on a number of other factors including the weather, the insulation of the building and the behaviour of the occupants. While previous work has shown the feasibility of using occupancy data, a thorough analysis of the influence of occupancy detection and prediction on the energy savings obtainable under different environmental conditions has thus far been lacking.

Our contributions to the state-of-the-art of automatic heating control systems can be thus summarised as follows:

1.2.1 Opportunistic occupancy sensing

We show how data from smart electricity meters and mobile phones can be used opportunistically to build occupancy schedules. Using concepts from supervised machine learning we first design algorithms to infer occupancy from electric load curves. To this end we collected a dataset in six Swiss households over a period of seven months containing the aggregated electricity consumption, the consumption of selected appliances and ground truth occupancy data [21]. This dataset has been made publicly available [214]. In addition to deriving occupancy from electrical consumption data, we also estimate

long-term occupancy schedules from a large publicly available mobile phone location dataset.

1.2.2 Classification and analysis of occupancy prediction algorithms

Using these schedules, we investigate the performance of occupancy prediction algorithms. For this purpose, we first perform a classification of state-of-the-art occupancy prediction algorithms into *schedule-based*, *context-aware* and *hybrid* approaches. We thereby outline different techniques used in the literature and categorise existing approaches. Using the occupancy schedules of 45 individuals collected over several months, we perform a quantitative comparison of schedule-based occupancy prediction approaches. We show that current schedule-based prediction algorithms can achieve an accuracy around 85%, while further improvements are unlikely due to randomness in human behaviour.

1.2.3 Analysis of the energy-savings potential of smart heating

To investigate the impact of different environmental conditions on the potential savings achievable by occupancy detection and prediction, we derive four different building models based on the ISO 13790 standard and simulate the heating costs for a number of different weather scenarios. We thus show that the energy that a smart thermostat may actually save depends to a large extent on external factors.

1.3 Outline of the thesis

We first give a short introduction to the concepts and paradigms underlying heating control systems in Chapter 2. This chapter serves as an introduction to the terminology used in this thesis and explores the tradeoff between energy savings and thermal comfort.

In the next three chapters, we will deal with the problem of occupancy detection from opportunistic sensors. In Chapter 3 we introduce the infrastructure used to collect the Electricity Consumption and Occupancy (ECO) dataset [21]. Following a description of the dataset, we show how occupancy may be derived from the electrical load curve in Chapter 4. In Chapter 5 we present how we derived long-term occupancy schedules from an unlabelled mobile phone location dataset.

In the following chapters, we use these long-term schedules to analyse the performance of state-of-the-art occupancy prediction approaches. Our classification and quantitative analysis of the prediction accuracy of current occupancy prediction algorithms is documented in Chapter 6. In Chapter 7 we introduce the simulation model to analyse the savings potential of occupancy prediction algorithms under different environmental condi-

tions. Using this simulation model, we investigate the achievable savings and the resulting comfort loss of automatically controlling the heating using occupancy prediction.

We conclude this thesis in Chapter 9 with a summary of our work and a discussion of open challenges in occupancy sensing and prediction.

Smart heating

Heating control systems such as programmable thermostats must address the trade-off between ensuring occupant comfort on the one hand and reducing the energy consumption on the other. Mozer *et al.* succinctly summarise this problem by suggesting:

“If one is merely interested in lowering energy costs, then simply shut off the furnace.” [140]

If the sole goal was to lower the energy consumption, the simplest (albeit not very smart) solution would be to turn heating off at all times. Alas, while forgoing heating altogether clearly minimises the energy consumption, such an approach is obviously infeasible when outside temperatures drop. The main goal of a heating system remains to ensure a comfortable indoor temperature. For many occupants, ensuring this comfort at all times has a higher priority than energy savings [143]. Thus, the thermostat is often constantly left on a comfortable (high) setting regardless of the presence or absence of occupants.

While previous work acknowledges the fact that some energy savings may be gained from persuasive approaches (such as promoting the use of programmable thermostats and/or generally reducing the indoor temperature) [143, 158], recent publications focus on smart heating control systems to achieve savings [48, 55, 67, 124, 128, 150, 169, 181]. The smartness of the system typically lies in its ability to adapt to current *weather conditions*, the *building characteristics* and the *behaviour of the occupants*. The difference between a conventional automatic heating system and a “smart” one is that while the former operates according to a pre-defined and typically deterministic (*e.g.* timer-based) schedule, the latter typically adapts its control strategy to the user context. In both cases, though, the heating is controlled automatically, *i.e.* with the aid of a thermostat that does not require explicit human intervention.

In fact, the energy consumption can be minimised by heating when necessary (*i.e.* ensuring the temperature is always at a comfortable level when the home is occupied) and allowing the temperature to drop otherwise. To this end, a smart heating system uses occupancy detection and prediction strategies to find the correct times for changing the temperature of the home.

This chapter introduces the concepts and terminology behind smart heating based on occupancy sensing and prediction. We will thus first discuss the general operation of an automatic heating control system in Section 2.1. After that, Section 2.2 explores the concept of thermal comfort, before Section 2.3.1 concludes with some general observations on the constraints of a smart heating system posed by the physical characteristics of the building that is to be kept comfortable.

2.1 Automatic heating control

An automatic heating control system can be seen as a regulator that ensures that the (average) air temperature measured within a home is sufficiently close to a given target value. To this end, the system controls the activation and deactivation of the heaters available in the home (*e.g.* radiators and/or electrical heaters). Typically, at least two different target temperatures are defined: the *setback temperature* and the *comfort* (or *setpoint*) *temperature*, indicated as Θ_{setb} and Θ_{comf} respectively. Θ_{comf} is typically set by household occupants depending on their personal preferences and indicates the temperature at which they feel comfortable.

The value of Θ_{comf} will typically be around 21 °C [71]. The setback temperature Θ_{setb} in contrast is defined as the lowest (average) value at which the air temperature of the household is permitted to fall when the occupants are out (or asleep). There are several issues that need to be considered when setting suitable values for the setback temperature. In particular, Θ_{setb} must be sufficiently low to allow for significant energy savings (as the heaters can be – at least temporarily – be deactivated) but still high enough that the time needed to bring the household back up to Θ_{comf} does not exceed a reasonable value.

2.1.1 Occupancy sensing and prediction

Smart heating systems must use adequate procedures to both sensing and predicting the household occupancy state. We define occupancy as follows.

A room or building is said to be *occupied* at a time instant t if at least one of its residents is at home; otherwise, it is said to be *unoccupied*. The *occupancy state* of a house can thus be represented as a binary value (1 for occupied and 0 for unoccupied).

It is usually convenient to represent the historical occupancy states by dividing the hours of the day in N_s equally spaced intervals – called *slots*. An *occupancy vector* Γ is then a

a) Reactive control (sensing only)

The occupancy sensing infrastructure detects when the household becomes unoccupied at 9 a.m. However, since the heating starts only when the household becomes occupied again at 5 p.m., the residents experience a loss of comfort.



b) Predictive control (sensing + prediction)

The occupancy sensing infrastructure detects when the household becomes unoccupied at 9 a.m. Heating starts earlier to ensure a comfortable temperature upon the arrival of the occupants at 5 p.m.

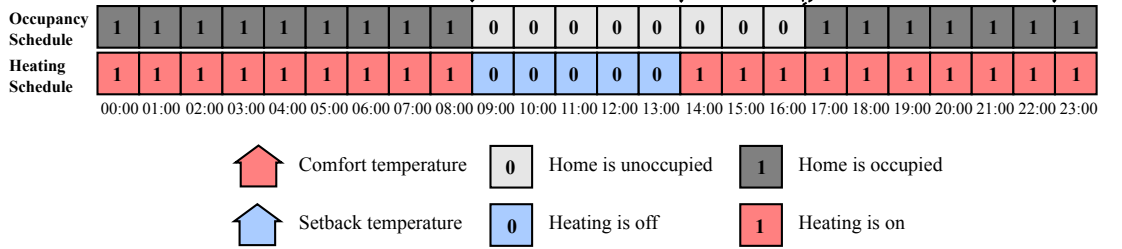


Figure 2.1: Automatic heating control systems based on occupancy sensing and prediction. Figure uses clipart from openclipart.org [206, 207].

$1 \times N_s$ vector of binary values in which the i^{th} element indicates whether the home was occupied or unoccupied during slot i . More specifically, we use $\Gamma_{1..96}$ to denote a 24-hour *ground truth* occupancy vector and $\gamma_{1..96}$ to refer to a 24 hour *predicted* occupancy vector based on 15-minute timeslots. Accordingly, an *occupancy schedule* is a $N_d \times N_s$ matrix containing occupancy data for N_d consecutive days. To accommodate slots for which no data is available, occupancy states can also be represented using three – rather than two – symbols, where one symbol is reserved to represent an *unknown* occupancy state.

An occupancy-based automatic heating control system

Figure 2.1 shows the operation of an occupancy-based automatic heating control system¹, both for *reactive* control (Figure 2.1a) and *predictive* control (Figure 2.1b). In the reactive control system, which is purely based on sensing the current occupancy, the heating schedule (*i.e.* the activation states of the heating system) is equivalent to the sensed

¹In this scenario “Heating is on” is equivalent to setting the thermostat to the comfort temperature, while “Heating is off” corresponds to a setting of the setback temperature. How the heating control system interprets these settings is determined by the available infrastructure and environmental conditions.

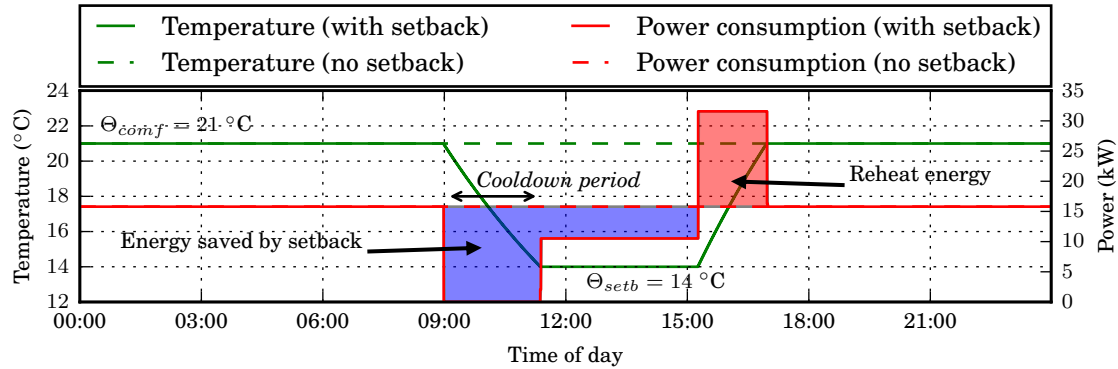


Figure 2.2: Heating automation.

occupancy. As the system only knows the current and past occupancy states, it can only react to a change in occupancy. Thus, as the occupants leave the building at 9 a.m., the heating is duly switched off and the house allowed to cool. However, when the occupants return at 5 p.m., they return to a cool building resulting in a loss of comfort.

This problem is alleviated by occupancy prediction algorithms. Figure 2.1b shows the same scenario with occupancy prediction. Here, the heating is also allowed to switch off at 9 a.m. but it is reactivated prior to the arrival of the occupants at 2 p.m. to reheat the building in time for the occupants' arrival at 5 p.m. Figure 2.1 thus shows that only a predictive heating system can ensure occupant comfort while reducing the energy consumption during the absence of occupants.

2.1.2 Optimal control

To achieve energy savings, an optimal heating system should thus be able to maintain the temperature of a home at Θ_{setb} for as long as possible, so as to reduce the amount of energy spent on heating. At the same time, the system must ensure that the temperature is close to Θ_{comf} whenever at least one occupant is at home (and awake) – so as to avoid any loss of comfort. However, the time needed to bring the home from Θ_{setb} to Θ_{comf} (and vice versa) is typically non-negligible (*e.g.* > 1 hour). An optimal heating system therefore needs to be able to both immediately detect when the home becomes unoccupied – so as to turn off the heating – and also reliably predict when it will be occupied again – in order to restore the temperature to Θ_{comf} by the time the occupants return.

Figure 2.2 shows the operation of such an automatic control system. The figure shows the indoor air temperature for an automatic heating control system based on the same fixed schedule as in the previous section. The house is assumed to be occupied from 12 p.m. to 9 a.m. and from 5 p.m. to 12 p.m. During the day, from 9 a.m. to 5 p.m., the temperature is allowed to drop no further than a setback temperature Θ_{setb} of 14 °C. Following the departure of the occupants at 9 a.m. the temperature quickly drops until the setback is

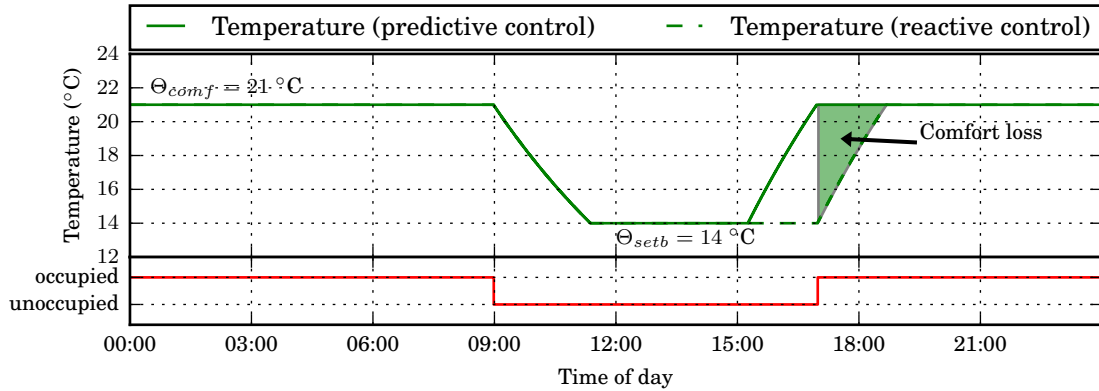


Figure 2.3: Comfort loss.

reached around 11.30 a.m. From this point onwards, the system keeps the temperature equal to Θ_{setb} until approximately 3.15 p.m. At this point, the system has determined that in order to reach Θ_{comf} upon the arrival of the occupants it needs to start heating now.

The shaded areas show the energy required for these two phases. During both the cooldown period and the following setback period, the energy consumption of the heating system is lower than that of a system that keeps Θ_{comf} throughout the day. During the re-heating phase, the energy consumption is naturally higher as heat has to be added to the system. From a visual inspection of the shaded areas one can see that the energy saved by using the setback schedule is clearly larger than the energy required for reheating the building. We will use such optimal control system in our evaluation of occupancy prediction algorithms in Chapter 8.

2.1.3 Reactive control

In contrast to the optimal control system, a purely reactive system shows what may happen if the system incorrectly predicted the occupants. Upon their arrival, the occupants experience *comfort loss* as the current indoor temperature is still at Θ_{setb} . Figure 2.3 shows the temperature for both an optimal predictive controller and a reactive controller. As in the previous examples, the building is unoccupied from 9 a.m. to 5 p.m. The failure to heat prior to the arrival of the occupants at 5 p.m. results in lost comfort for the residents. The heating system requires about one hour to reheat the building to Θ_{comf} . During this time, the temperature is below Θ_{comf} . This comfort loss can be quantified by the shaded area. Right after the arrival of the occupants, the temperature is furthest away from Θ_{comf} . As the heating system starts to reheat the building, this difference becomes smaller. The area between the two curves is often referred to as *heating degree hours*. We will revisit this concept in Chapter 8 when we analyse the performance of smart heating systems

Table 2.1: Fanger’s thermal sensation scale.

Scale	Thermal sensation
+3	hot
+2	warm
+1	slightly warm
0	neutral
−1	slightly cool
−2	cool
−3	cold

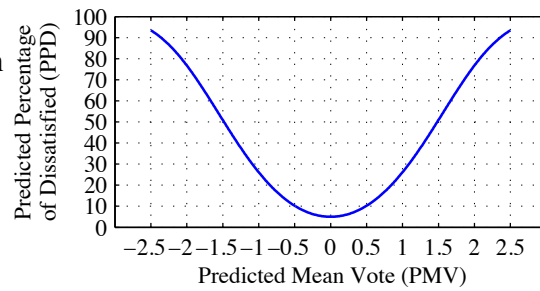


Figure 2.4: Relationship between Fanger’s PPD and PMV.

based on occupancy detection and prediction. However, comfort cannot be captured by the difference between the current and setpoint temperatures, alone.

2.2 Thermal comfort

The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) describes thermal comfort as “that condition of mind that expresses satisfaction with the thermal environment” [13]. As such, it is not solely dependent upon the current air or radiant temperature but determined by a combination of these with other factors such as health, psychology, clothing and activity of the occupants. These relationships were first examined by Ole Fanger’s doctoral thesis on thermal comfort in 1970 [53]. Today, Fanger’s *static comfort model* is one of the foundations of various building standards such as the ASHRAE Standard 55 on the “Thermal Environmental Conditions for Human Occupancy” [13] and the ISO 7730 standard on the “Ergonomics of the Thermal Environment” [83]. In a shared space such as a building, achieving global thermal satisfaction of all occupants is impossible. Fanger’s static model, however, helps to characterise the level of comfort achieved by a particular thermal environment and the number of people dissatisfied with it.

2.2.1 Static comfort models: Fanger’s PMV and PPD

ASHRAE 55 identifies six primary factors influencing thermal comfort (*cf.* Table 2.2). Relating to the occupants it includes their *metabolic rate* and *clothing level*. The metabolic rate of the occupants is determined by their physique and activity. The clothing level depends upon the material and number of layers of clothing worn by the occupants. With respect to the building, ASHRAE 55 identifies the *air* and *radiant temperatures* as well as the *air speed* and *humidity* as relevant factors. A large difference between the air temperature and the mean radiant temperature² is a cause for discomfort. Likewise,

²The mean radiant or masonry temperature is the average temperature of all building parts.

Table 2.2: Sample calculation of Fanger's PPD and PMV (PMV = -0.42, PPD = 9%).

Factor	Typical value
Air temperature	21 °C
Mean radiant temperature	18 °C
Air speed	0.1 m s ⁻¹
Humidity	50%
Metabolic rate	1.2 met (standing, relaxed)
Clothing level	1 clo (typical winter indoor)

draught³ and high or low humidity values cause discomfort to the occupants. These factors were studied in detail by Ole Fanger in 1970 [53]. Fanger came up with a model to use these factors to calculate the thermal sensation felt by the occupants on a seven-point scale (*cf.* Table 2.1). The resulting Predicted Mean Vote (PMV) and Predicted Percentage Dissatisfied (PPD) metrics are discussed below.

Predicted Mean Vote (PMV)

Table 2.1 shows the Fanger's thermal sensation scale⁴ of an individual ranging from +3 (hot) to -3 (cold). The ideal value on the thermal sensation scale is zero, indicating thermal neutrality. Fanger's scale is based on the heat balance of the human body. If the metabolic rate is equal to the heat loss to the environment, thermal balance is obtained. If the heat loss to the environment is greater than the metabolic rate, the person is feeling cold. If the metabolic rate is greater than the heat dissipated to the environment, the person is feeling hot. The human body will try to restore thermal balance by shivering and sweating, respectively. Fanger defined the thermal sensation as "the difference between the internal heat production and the heat loss to the actual environment for a man kept at the comfort values for skin temperature and sweat production at the actual activity level" [53].

To find the relationship between the environmental factors and comfort, Fanger conducted experiments in a climate chamber. He subjected probands to different environmental conditions and recorded their comfort vote. The result are a set of equations leading to an prediction of the participants thermal sensation in a particular environment. Fanger's Predicted Mean Vote (PMV) gives an indication of the mean thermal sensation of a large group of people on the thermal sensation scale. The *comfort zone* is defined for any combinations of the six primary factors such that $-0.5 < \text{PMV} < +0.5$. By varying the parameters, it can be checked which environmental conditions ensure a thermally comfortable environment. ASHRAE 55 and ISO 7730 include a method for calculating the PMV programmatically from the six primary factors highlighted above [83]. Table 2.2 shows an example of an ASHRAE 55 compliant thermal environment. In this example, we assume an indoor air temperature of 21 °C and a mean radiant temperature of 18 °C.

³Currents of cold air through cracks or other openings in the building envelope.

⁴Fanger's thermal sensation scale is also used in ASHRAE 55 and ISO 7730.

The occupants wear typical winter clothing and have a slightly elevated activity level (standing, relaxed). Air speed and humidity are at 0.1 m s^{-1} and 50%. In this environment, the PMV following ASHRAE 55 and ISO 7730 is calculated as -0.43 , which lies within the comfort zone.

Predicted Percentage Dissatisfied (PPD)

The downside of the PMV metric is that it does not reflect the thermal comfort of individual occupants. In order to maximise overall comfort, Fanger's PPD attempts to minimise the number of occupants feeling discomfort in the current thermal environment. Equation 2.1 defines the PPD in terms of the PMV. The formula was derived empirically by Fanger from climate chamber experiments.

$$\text{PPD} = 100 - 95 \times e^{-0.03353 \times \text{PMV}^4 - 0.2179 \times \text{PMV}^2} \quad (2.1)$$

Figure 2.4 shows the relationship between the PMV and PPD [13]. In the comfort zone (*i.e.* $-0.5 < \text{PMV} < +0.5$, as above), 10% of occupants are dissatisfied with the thermal environment. As the comfort zone is left, the number of dissatisfied occupants increases.

Alternatively to the calculation of the PMV-PPD, in an existing building, a heating, ventilation and cooling (HVAC) engineer⁵ may ask occupants to fill in a questionnaire asking for their thermal sensation [13]. The replies can then be used to assess the overall suitability of the current environment to achieve thermal comfort. ASHRAE 55 includes a template for such a questionnaire [13]. A survey on the thermal satisfaction may identify design flaws related to the unintended use of the system. Furthermore, the underlying “occupant psychosocial conditions can impose a strong influence on subjective assessment of the environment” [13]. By conducting such a survey, design protocols can be improved and mitigation strategies for improving comfort can be found.

Limitations

The PMV model deals only with steady-state conditions. It thus for example ignores that occupants may experience different thermal sensations when moving from a cold to a warm place. Furthermore, the static model does not include a variable environment temperature. It implies that the comfortable indoor temperature is not affected by the current season and thus maintains the same indoor temperature year-round. This constant temperature regime is usually not feasible in practice as occupants adapt their clothing to the different seasons. Furthermore, such a static temperature setting usually increases the

⁵Heating, ventilation and cooling (HVAC) is an umbrella term for technology used to ensure indoor thermal comfort and air quality. Heating refers to appliances generating heat in a building. This can either be done locally (*e.g.* by electric room heaters) or centrally (*e.g.* by hydronic heating with a gas-powered boilers). Ventilation is the process of regularly replacing air to ensure air quality (*e.g.* removing carbon dioxide, moisture, odours and smoke). Air-conditioning refers to cooling and humidity control.

difference between the environment and indoor temperature leading to discomfort when moving from the indoors to the outdoors and vice versa.

The calculation of the PPD is based on the simplifying assumption that it is symmetric around a neutral PMV. It does not add any information. As such, the real percentage of dissatisfied occupants may vary considerably from the one predicted by the PPD.

The PMV model is based on mean observations from a climate chamber. As measuring (or even mandating) the metabolic rate of real occupants is infeasible, it is impossible to ensure that all different physiologies are dealt with in the static PMV model. ASHRAE 55 notes this by limiting the scope of the standard: “it is not possible to prescribe the metabolic rate of occupants, and because variations in occupant clothing levels, operating setpoints for buildings cannot practically be mandated by this standard” [13]. Essentially, the PMV can only be used as general guidance for adapting the thermal environment.

The difficulty to measure or estimate the six parameters has led to simplifications. Values corresponding to activity and clothing levels are often obtained by assuming office work and correlating clothing level with outside temperature. As draught is generally accepted as uncomfortable, air speed is also reduced to a minimum.

2.2.2 Adaptive comfort model

While the static comfort model assumes steady-state conditions and thus maintains a constant thermal environment year-round, the *adaptive comfort model* tries to incorporate outdoor climate influences. Previous work has shown that thermal satisfaction is influenced by the context of the occupants [37]. The ability to control the thermal environment as well as past thermal history influence thermal satisfaction especially in naturally ventilated buildings. Compared to sealed and air-conditioned buildings, occupants in naturally ventilated buildings have been found to accept a larger range of comfort temperatures [37]. For buildings where the occupants are free to choose their clothing level, the adaptive comfort model defines the range of acceptable operative temperatures as a function of the *mean monthly outdoor air temperature* [13].

2.2.3 Local comfort models

Both Fanger’s PMV and the adaptive comfort model do not track individual occupants’ thermal comfort levels. To overcome this and in order to assess thermal satisfaction per person, Gao *et al.* introduce the so-called “Predicted Personal Vote” [61], an adaptation of Fanger’s PMV for individual occupants. The authors observe that different micro climates might exist in an office building and therefore suggest that heating and cooling “within the personal work space would be for the benefit of a single worker” [61]. They advocate the use of radiant heaters and fans to “maintain the comfort level of individual workers” [61]. The authors use Microsoft Kinect cameras to monitor the position of occupants and to

Table 2.3: Thermal properties used to describe the energy performance of building materials and components.

Property	Description
Specific Heat J/(kg K)	Heat per unit mass to raise the temperature of a material by one degree Kelvin (<i>e.g.</i> Brick = 0.840 J/(kg K), Wood = 1.7 J/(kg K)).
Heat Capacity (Capacitance) J/K	Thermal mass of a body (<i>i.e.</i> Specific Heat x Mass). A high thermal mass can help to flatten out changes in the outside temperature.
Thermal Conductivity W/(m K)	Measure of a material's ability to conduct heat (<i>e.g.</i> Concrete = 1.7 W/(m K), Wood = 0.04-0.4 W/(m K), Air = 0.025 W/(m K)).
Thermal Resistivity m K/W	Measure of a material's ability to resist a flow of heat.
Absolute Thermal Resistance K/W	The thermal resistance of a body (<i>e.g.</i> a heat sink).
Density kg/m ³	Mass per unit volume.
Thermal Absorbance	Fraction of absorbed long wavelength radiation.
Solar Absorbance	Fraction of absorbed solar radiation.
Visible Absorbance	Fraction of absorbed visible wavelength radiation.

let them use gestures to adjust their thermal preferences. An infrared thermometer is used to measure the clothing surface temperature. The clothing level is estimated using a linear regression model. The PPV is used to reactively control a heater in the office. In a follow-up work [60], the same authors use their system to predictively control the temperature in an office environment.

In contrast to Gao's work, the "Thermovote" approach by Erickson *et al.* [49] is allowing occupants to vote on the current temperature. The voting is facilitated by an iPhone application and uses the ASHRAE thermal satisfaction scale. The approach is simplifying the survey approach of ASHRAE 55 but also requires continuous user interaction.

Lam *et al.* propose a participatory approach based upon a combination of Gao's and Erickson's work [115]. Their system includes a mobile application to vote on the current thermal conditions and a custom comfort model based on the metabolic rate of occupants. The authors thus claim to solve the problems of continuous user interaction and missing model parameters. Unfortunately, the authors made a mistake quoting the comfort zone from ASHRAE 55 to be from -1 to 1 (it is -0.5 to 0.5 as discussed above) [13]. This invalidates the authors' claim that 89% of the votes were within the comfort zone.

2.3 Building constraints

The energy savings obtainable and the potential comfort loss of a smart heating system depend to a great extent on the building itself. Zero energy buildings, which are independent from the energy supply, might not need predictive heating control – although they must be carefully ventilated. For most existing buildings, however, efficient heating control algorithms based on actual occupancy could potentially save a substantial amount of energy as their insulation is insufficient.

2.3.1 Thermal properties

The thermal properties of the building and its environment determine the efficiency of the heating infrastructure. Table 2.3 shows various thermal properties of building materials and components which are currently used to assess the energy performance of buildings.

Storing energy

The *specific heat* is the amount of heat needed to raise the temperature of a material of a certain mass by 1 K. It is therefore a measure of how much energy can be stored inside the material. A body with a high volume and a high specific heat – a high *thermal mass* – can store more energy and is more immune to variations in the surrounding temperature. While this prevents uncomfortable fluctuations in the indoor temperature, it also poses a challenge for a reactive heating control systems as seen in Section 2.1.3. If the house has been left to cool over an extended period of time, comfort is lost due to the slow ramp up time upon the arrival of the occupants.

A lower thermal mass, on the other hand, shortens the ramp up time. As shown in Table 2.3, wood has a high *specific heat* of around $1.7 \text{ J}/(\text{kg K})$. However, allowing for the low density (*i.e.* beech has about $800 \text{ kg}/\text{m}^3$), the volumetric heat capacity of wood ($1360 \text{ kJ}/(\text{m}^3 \text{ K})$) is lower than the one of brick ($1610 \text{ kJ}/(\text{m}^3 \text{ K})$). Brick can therefore store more energy and stabilise indoor temperatures when night and day temperatures vary. This effect is stronger the more temperatures vary.

Retaining energy

The thermal conductivity or its reciprocal thermal resistivity measure a material's ability to conduct or resist a flow of heat. A low thermal conductivity means that the material is a good insulator. As brick has a higher thermal conductivity, additional insulation is needed to prevent the home from losing heat through the brickwork. Wood on the other hand is a better insulator, which does not require additional insulation.

Heating efficiency depends on building properties

This discussion of the thermal properties of buildings goes to show that the efficiency of a heating control system relies to a large extent on how the building is designed and operated. This makes it difficult to analyse the savings of a smart heating system based on occupancy detection and prediction solely from a few specific examples. Moreover, as discussed in the previous section, the indoor air temperature is not the sole factor determining thermal comfort. As homes get better insulated, it becomes increasingly important to regulate the airflow in order to avoid mould or an increase in CO_2 levels.

Towards opportunistic occupancy sensing: The ECO dataset

Occupancy detection is an important component of commercial and residential building automation systems. Systems that typically regulate HVAC are usually based on dedicated sensors to provide occupancy information [124, 169]. Similarly many lighting systems rely on the detection of the presence (or absence) of people to automatically switch lights on (or off) [65]. Occupancy detection is also applied outside the domain of building automation. For example, Dickerson *et al.* showed that sensing a change in occupancy patterns can help to reveal clinical diseases such as depression [40].

In commercial building automation systems, occupancy detection is typically provided by *dedicated infrastructure* such as passive infrared (PIR) sensors, magnetic reed switches and cameras [48, 124]. Despite the large number of application scenarios, such an occupancy detection infrastructure is cumbersome and expensive to install [144]. It necessitates the purchase, installation and calibration of multiple sensors. First commercial products targeting heating in a residential environment [192, 196, 198] are still expensive and non-trivial to install and operate. They thus only cater for a small segment of technologically savvy enthusiasts. In contrast to traditional building management systems (BMSs), which are often used in commercial buildings, in a residential environment it is often also only possible to install a few cheap and possibly imprecise sensors as the overall costs of the infrastructure must be kept low.

Besides the initial cost for the installation of the infrastructure, a building automation system requires continuous maintenance. This poses critical constraints for the adoption of such a system. In a residential scenario, the significant installation and maintenance overhead is typically shouldered by a layman “building administrator” – an (often more or less technically inexperienced) resident of the household. Faulty installations and a lack of maintenance are a frequent consequence.

Taken together, these constraints can cause inconvenience to the users and restrict the

Table 3.1: Overview of dataset.

Sensor / (#)	Description	# Records
Landis+Gyr E750 (6)	Smart Electricity Meters	125,987,285
Plugwise Sting (45)	Smart Power Outlets	686,655,790
Fluksometer (6)	Network monitor (ICMP Echo)	1,585,595
Roving RN-134 (6)	Low-Power Wi-Fi (PIR Sensor)	563,758
Samsung Galaxy Tab P7510 (6)	Occupancy level ground truth data	6,396

acceptance of the system. We will thus investigate if technology that currently already exists in many households can be used to sense occupancy. To this end, we advocate the *opportunistic use* of existing infrastructure such as smart electricity meters to reduce costs and to increase the reliability of occupancy sensing.

To evaluate the feasibility of smart electricity meters as occupancy sensors, we recorded and published the *Electricity Consumption and Occupancy* (ECO) dataset [21]. The ECO dataset contains sensor data from a large set of heterogeneous sensors (smart electricity meters, passive infra-red sensors, smart power outlets and connected network devices) from six Swiss households over a period of seven months. Table 3.1 shows an overview of the data collected.

Our analysis of occupancy sensing using smart electricity meters is split into two separate chapters. In this first chapter, we highlight related datasets and occupancy sensing infrastructure in Section 3.1, before we elaborate on the design and operation of our own data collection infrastructure (*cf.* Sections 3.2 and 3.3). Before we conclude this chapter in Section 3.5, we furthermore describe in detail the ECO dataset in Section 3.4. In the next chapter (Chapter 4) we will then analyse how occupancy information can be derived from the electrical load curve using machine learning algorithms. This chapter is based on the contributions made in [21, 102].

3.1 Related research work and datasets

To evaluate smart electricity meters as occupancy sensors, we first look at related work in occupancy sensing. We then cover the more general topic of significant place sensing, before we cover other datasets containing occupancy, location and electricity consumption data.

3.1.1 Occupancy sensing

Many authors have focused on occupancy detection in residential households. In 1995, Mozer *et al.* introduced the Neural Network House project [139]. The authors retro-fitted an old school house with “nearly five miles of conductor” to sense occupancy and other environmental variables [140]. In a follow-up work, Mozer *et al.* use the

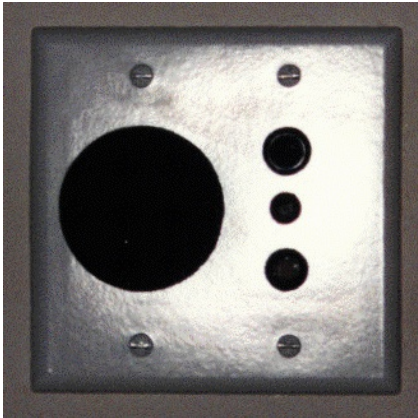


Figure 3.1: Dedicated sensor panel from the Neural network house [139]. Left shows loud-speaker for communication with occupants. On the right are sensors for temperature, ambient light and sound [217].

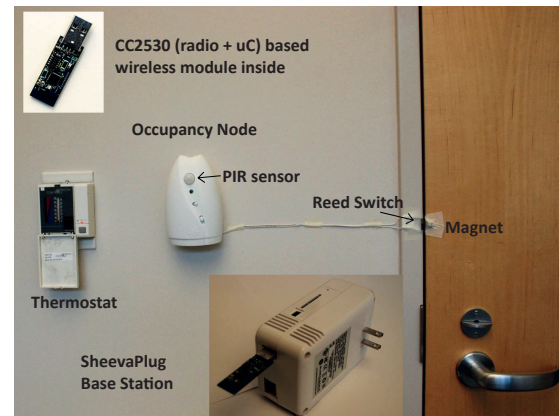


Figure 3.2: Occupancy sensing infrastructure used by Agarwal *et al.* [6]. Occupancy node (containing PIR sensor) is built from an Air Wick air refresher [190].

Neural Network House to investigate how a smart thermostat could be realised in such an environment [140]. Figure 3.1 shows one of the sensor panels used in the Neural Network House. It incorporates a number of sensor and a loudspeaker for communication with the occupants. In their Smart Thermostat paper, Lu *et al.* propose a simpler approach [124]. It relies on cheap off-the-shelf wireless PIR sensors and magnetic reed switches. By using X10 technology the authors aim to keep the cost at around \$25 per household. Other authors rely on recognising occupants using electronic tags. Thus, besides employing traditional PIR sensors to register per-room occupancy, Scott *et al.* use an RFID-based system that requires occupants to deposit their house keys near a receiver at the entrance of the property [169].

Office buildings are also targeted for occupancy detection. Dodier *et al.* use Bayesian network model to fuse data from multi-modal binary sensors to detect occupancy in an office environment [43]. To overcome the inaccuracy of the PIR technology¹, the authors have built a sensor to determine if a telephone conversation is in progress (*i.e.* if the telephone is off-hook). The OBSERVE system by Erickson *et al.* instead relies on a homogeneous wireless camera system to continuously estimate the number of people occupying rooms in public buildings (*e.g.* meeting rooms) [48]. In contrast to other systems, the work by Erickson *et al.* highlights the need to sense the “level” of occupancy in order to adjust ventilation for increased air quality. Beltran *et al.* follow the same objective but address privacy concerns by avoiding the use of cameras. Instead they use an

¹PIR sensors can produce false negatives if occupants do not move around.

array of thermal sensors in conjunction with a PIR sensor to detect occupancy levels in a room [24]. Similarly, Milenkovic *et al.* equipped three offices with PIR sensors and plug-in power meters, which measured the energy consumption of the computer screens [134]. The authors estimated the number of persons in the office as well as their current activity (*e.g.* desk work). Khan *et al.* discuss an occupancy monitoring system based on light, humidity, noise and passive infrared sensors. Like the approach by Erickson *et al.* Khan derive occupancy at various levels of granularity ranging from binary occupancy to the level of occupancy [94]. Agarwal *et al.* presented an infrastructure based on PIR sensors and reed switches for office buildings [6]. Figure 3.2 shows the infrastructure used by Agarwal. A dedicated occupancy node records data from a built-in PIR sensor while an adjacent reed switch monitors the state of the door.

Another strategy for detecting occupancy consists of interrogating sensors carried by the residents, such as dedicated wireless transmitters or GPS modules embedded in mobile phones [67, 113]. Thereby, a mobile phone might share its location with the smart thermostat to determine the optimal time to start re-heating the home. We will discuss mobile phone based occupancy detection in Chapter 5.

Occupancy sensing has a long tradition in building automation. In the context of automatic lighting control, Guo *et al.* provide an overview of occupancy sensing hardware [65].

3.1.2 Datasets containing electricity consumption data

To aid the development of so-called non-intrusive load monitoring (NILM) algorithms that take the aggregated consumption of the household and produce device-level consumption statistics, several authors have collected datasets containing device-level consumption data. The *Reference Energy Disaggregation Dataset (REDD)* collected by Kolter *et al.* contains the aggregate electricity consumption of five homes in the United States along with measurements for individual circuits and appliances over several weeks [110]. Recently, Barker *et al.* published the *UMASS Smart* Home dataset* [16] containing detailed submeter measurements from 21-26 circuit meters in three homes over three months. However, in contrast to our dataset, the three homes in the Smart* dataset are not instrumented to the same level. Only one of the households contains data from PIR sensors. Further NILM-centric datasets include the *GREEND* [137], *BLUED* [11], *AMPds* [125], *UK-Dale* [92], *iAWE* [18] and the *Pecan Street* [75] dataset.

In contrast to the ECO dataset, none of these datasets contains ground truth information on the occupancy patterns of the inhabitants². Furthermore, in addition to including occupancy information, our ECO dataset extends existing datasets in three aspects. First, it covers a long timespan – seven months. From the datasets mentioned above, only the AMPds and the UK-Dale datasets cover a similar timespan. Secondly, the aggregate

²The Smart* dataset contains data from PIR sensors but no user-annotated ground truth.

consumption data contained in the ECO dataset is very detailed and contains measurements of both the real and reactive power for all three phases. This level of detail is only matched by the AMPds, the iAWE and the BLUED datasets. Thirdly, the ECO dataset contains plug-level data at 1 Hz frequency which is only matched by the Smart*, iAWE and GREEND datasets.

3.2 Experimental setup of our occupancy sensing infrastructure

To estimate the occupancy state of a household based on an opportunistic sensing infrastructure, we performed an extensive data collection in collaboration with a utility company in Switzerland. We collected a multi-modal dataset in six households over the course of seven months. In addition to the electricity consumption of a household the dataset contains sensor information collected from PIR sensors and smart power outlets. The households also recorded ground truth occupancy data through a tablet computer. This section describes the selection of households and our measurement infrastructure.

3.2.1 Selection of households

For the data collection we chose the participating households among employees of a utility company in Switzerland. Prospective participants were required to fill in a questionnaire.³ The questionnaire contained 12 questions targeting the number, age and occupation of the occupants, type of property, number of entry doors, typical occupancy, type of heating, pet ownership as well as the level of affinity for technology of the respondent. The affinity for technology was requested through a 7-point Likert scale (1: low, 4: medium, 7: very high) [119]. The purpose of the questionnaire was to ensure households have a reasonable size (*i.e.* 1-4 occupants) and participants are well-disposed to technical equipment. Also, we avoided to include households in which occupants used more than one entrance, because we wanted each participant to log occupancy through a tablet computer located near the main entrance. To each participant we handed a privacy statement that described the data gathered and the household's ability to opt out at any time during the data collection.

Table 3.2 shows an overview of the households ultimately selected to participate in the data collection⁴. Three of the households consist of two occupants, while two of the households are occupied by four persons. Four out of the five respondents live in detached

³Note that we only selected households 1-5 based on the questionnaires. Household 6 did not fill out a questionnaire. It participated in the collection of sensor data, however the participants did not specify their occupancy through the tablet computer. For these reasons, household 6 is omitted from the analysis.

⁴The full table containing all data obtained from the questionnaires can be found in Appendix A.

Table 3.2: Overview of the participants.

Household	No. of occupants	Type of property	Tech. affinity
r1	2 adults, 2 children	House	7/7
r2	2 adults	Flat	7/7
r3	2 adults	House	7/7
r4	2 adults, 2 children	House	4/7
r5	2 adults	House	6/7
r6	2 adults	House	-

houses, only the occupants of household r2 live in a flat. All respondents except for one classified themselves as having a high affinity for technology.

3.2.2 Overview of the architecture

Figure 3.3 shows a schematic view of our opportunistic occupancy sensing architecture for one household. As outlined in by Hnat *et al.* in [74], each type of sensor has its own advantages and drawbacks and can only guarantee limited confidence in estimating the actual occupancy state. For this reason, we have fitted all households with multiple heterogeneous sensors.

The aggregate electricity consumption of the household is measured by a smart electricity meter. It records data such as the real and reactive power at a frequency of 1 Hz. The consumption of selected individual appliances is measured using smart power outlets. In addition to the electricity consumption, our architecture also persists movements recorded by a PIR sensor installed near the entrance of the building. The PIR sensor doubles as a trigger to toggle the display of an Android tablet computer also installed near the entrance. The tablet computer visualises the current electricity consumption and provides buttons for the collection of occupancy ground truth. In some households the system also periodically recorded the media access control (MAC) addresses of devices reachable on the local network.

The data from the system was transmitted to a Web server at ETH Zurich for analysis. In the following sections we will describe the individual parts of the system.

3.2.3 Data collection infrastructure

All the data collected in this deployment is transferred over HTTP to a RESTful Web server at our institution. The server is built around a Java Servlet, which stores the raw values into a database for further processing. Figure 3.6 shows all Web-enabled devices that sent messages (*e.g.* HTTP POST requests) to the Web server. The Web server provides a user interface based on the dojo JavaScript toolkit [191], which allows to assess the current status of the overall system and to add or remove sensors and residences. It also offers a simple visualisation of the data from the smart electricity meters and the smart

3.2 Experimental setup of our occupancy sensing infrastructure

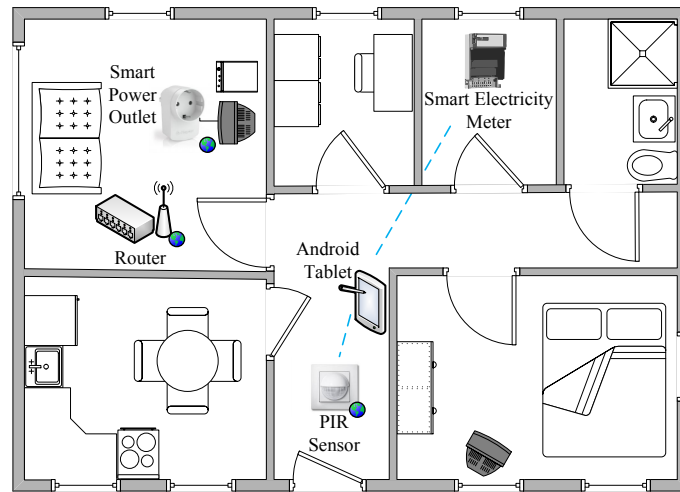


Figure 3.3: Schematic overview of the sensor deployment in one household.

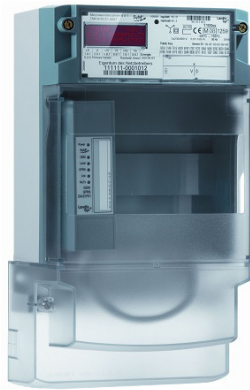


Figure 3.4: Landis+Gyr E750 ZMK400 smart electricity meter.



Figure 3.5: Fluksometer v.2.

plugs. Any unresponsive sensors (*e.g.* sensors which have not communicated new values over a minute) are highlighted to allow for a quicker recovery from failures. The WebUI was not available to the participants during the data collection.

3.2.4 Aggregate electricity consumption

Previous work has shown that electricity meters could provide clues regarding human activity – and thus the presence of residents – within a home [33, 102]. To evaluate the suitability of the electricity consumption as an indicator for occupancy, all six households were fitted with *Landis+Gyr E750 smart electricity meters* (*cf.* Figure 3.4) [116]. The smart meters were connected in series behind the original meter and not used for billing

Table 3.3: Excerpt of data provided by the Landis+Gyr E750 through the SML interface.

Description	OBIS code	Unit
Sum of effective power over all phases	01 00 0F 07 00 FF	Watt
Effective power phase 1	01 00 23 07 00 FF	Watt
Effective power phase 2	01 00 37 07 00 FF	Watt
Effective power phase 3	01 00 4B 07 00 FF	Watt
Effective current neutral	01 00 5B 07 00 FF	Ampere
Effective current phase 1	01 00 1F 07 00 FF	Ampere
Effective current phase 2	01 00 33 07 00 FF	Ampere
Effective current phase 3	01 00 47 07 00 FF	Ampere
Effective voltage phase 1	01 00 20 07 00 FF	Volt
Effective voltage phase 2	01 00 34 07 00 FF	Volt
Effective voltage phase 3	01 00 48 07 00 FF	Volt
Shift between voltage phases 1/2	01 00 51 07 01 FF	Degree
Shift between voltage phases 1/3	01 00 51 07 02 FF	Degree
Shift between current/voltage phase 1	01 00 51 07 04 FF	Degree
Shift between current/voltage phase 2	01 00 51 07 0F FF	Degree
Shift between current/voltage phase 3	01 00 51 07 1A FF	Degree

purposes. The installation was carried out by an employee of our industrial partner. The E750 provides averaged active and reactive power measurements at a frequency of 1 Hz [116]. Table 3.3 shows the variables obtained from the smart meter during the deployment.

Smart Message Language (SML)

For the purpose of remote metering, the E750 implements the Smart Message Language (SML) protocol [183]. SML is a request-response protocol and allows a client to send a request specifying the variables to be read. Once the request is received by the smart meter, a reply is formulated containing the variables requested. Variables such as power, voltage and current used in the request and response are identified using the Object Identification System (OBIS), which is standardised in IEC 62056-61 [79]. However, manufacturers may specify additional codes to communicate vendor-specific variables. Table 3.3 shows the OBIS codes of all the variables used in this deployment.

The E750 uses a monotonic non-decreasing clock (*i.e.* a counter that measures the number of discrete seconds since the meter was taken into operation). Therefore, timestamps have to be dealt with on the client side. In our case, the E750 was read out using a Flukso (*cf.* Section 3.2.4) via its Ethernet interface. The Flukso is associating a timestamp with every measurement after receiving the measurement. The Flukso itself synchronises its clock using the Network Time Protocol (NTP) [135].

3.2 Experimental setup of our occupancy sensing infrastructure

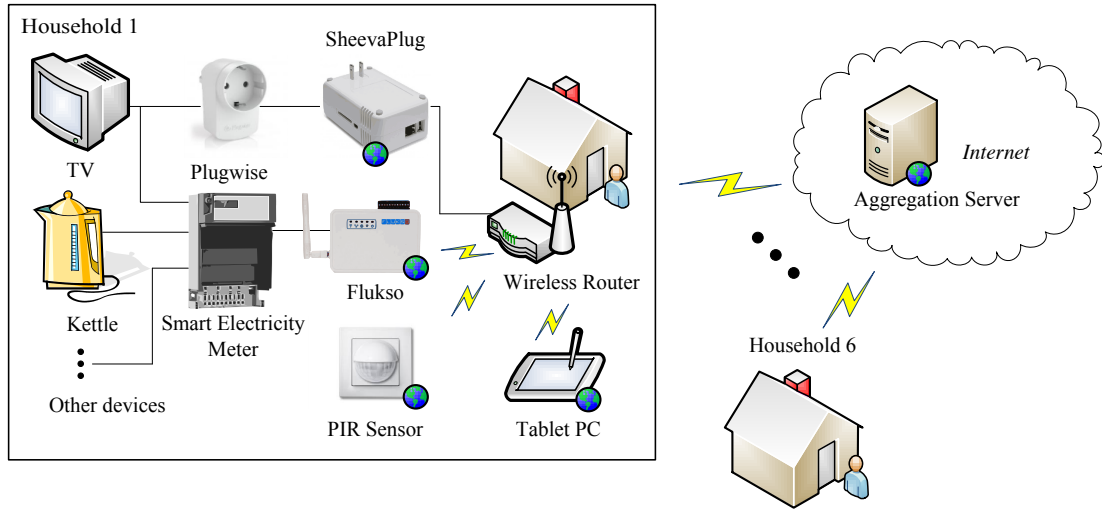


Figure 3.6: Schematic overview of whole deployment.

Communication module

As shown in Figure 3.6 we used a *Fluksometer* (Flukso) to communicate the data from the smart meter to the Web server at ETH Zurich. The Flukso is a community metering device, which allows enthusiasts to measure their electricity consumption online [193]. The Flukso has a built-in sensor board to which up to three current clamps to measure the electricity consumption may be connected. Essentially built on router components, the Flukso consists of an Atheros AR5007AP-G (AR2317) Wireless System-on-a-Chip (WiSoC) with 8 megabyte flash memory and 16 megabyte RAM and runs OpenWrt [52]. The interface to the sensors is realised using an ATmega168 microcontroller. The Atheros offers an Ethernet port and a built-in Wi-Fi module. The system is designed to work with current clamps and to communicate measured values via Ethernet or Wi-Fi to the `flukso.net` platform. However, as the device runs on a version of OpenWrt its software can be easily extended through additional packages [52]. We have chosen the Flukso over a standard router to allow for the integration of households with no smart meter (through the use of current clamps) into our system. However, since OpenWrt is supported by a wide range of devices [194], a different device may have been chosen.

We have extended the Fluksometer to support the SML protocol by porting the current version of libSML, an Open-Source SML library developed by the DAI-Labor at TU-Berlin, to OpenWrt [210]. The libSML library allows us to communicate with the E750. Our client software (pylon), which is built on top of libSML, is available on GitHub [218]. Pylon reads the variables from the smart meter every second and communicates them to the Web server using parallel HTTP POST requests from the curl library. We have parallelised the requests to make sure that the network latency does not impair our ability to record new data. In addition, the Flukso automatically caches data that could not be



Figure 3.7: Smart plug from Plugwise that measures the electricity consumption of a attached appliances.



Figure 3.8: Sheeva plug computer.

send due to network problems. It automatically resends these data when the connection becomes available again. Furthermore, a watchdog process monitors the operation of the Flukso to restart it if needed. A periodic heartbeat is used to further monitor the Flukso.

3.2.5 Device-level electricity consumption

Not all appliances are equally good indicators of occupancy. For instance, the washing machine may or may not indicate occupancy depending on the household. In some households, the residents might program it using a timer ensure the completion of the washing cycle when they return. Others may only do the laundry when they are present. Likewise, electric boilers or heat pumps might be operational during times when the occupants are either away or asleep. The activation state of a television set or electrical stove, on the other hand, usually correlates very well with occupancy. Thus the device-level electricity consumption of the household is a better indicator of occupancy than the aggregated load curve.

Smart plugs

To assess the information gained by measuring individual appliances instead of the aggregated load curve, we have instrumented selected appliances with smart power outlets (*cf.* Figure 3.7). According to [99], the *smart plugs* from Plugwise [195] are currently some of the most accurate and easy to deploy smart power outlets. These *smart plugs* are connected between the measured appliance and the mains. When supplied with power, the smart plugs automatically create a mesh network with their neighbours and communicate their

3.2 Experimental setup of our occupancy sensing infrastructure

Table 3.4: Appliances household r1.

Plug #	Appliance
1	Refrigerator
2	Tumble dryer
3	Router / coffee machine
4	Kettle
5	Washing machine
6	Freezer

Table 3.5: Appliances household r2.

Plug #	Appliance
1	Android tablet
2	Dishwasher
3	Stove exhaust fan
4	Refrigerator
5	Entertainment
6	Freezer
7	Kettle
8	Light
9	Laptops

Table 3.6: Appliances household r3.

Plug #	Appliance
1	Android tablet
2	Freezer
3	Coffee machine
4	PC
5	Refrigerator
6	Kettle
7	Entertainment

Table 3.7: Appliances household r4.

Plug #	Appliance
1	Refrigerator
2	Kitchen appliances
3	Light
4	Stereo and laptop
5	Freezer
6	Android tablet
7	Entertainment
8	Microwave

Table 3.8: Appliances household r5.

Plug #	Appliance
1	Android tablet and telephone
2	Coffee machine
3	Small water fountain
4	Microwave
5	Refrigerator
6	Entertainment
7	PC, router, Sheeva plug, printer

Table 3.9: Appliances household r6.

Plug #	Appliance
1	Lamp
2	Printer and laptop
3	2 Routers and Sheeva Plug
4	Coffee machine
5	Entertainment
6	Refrigerator
7	Freezer
8	Kettle

measurements via Zigbee (802.15.4) to a computer. Tables 3.4 to 3.9 show the appliances instrumented and measured in each household.

Originally, each smart plug stores the total consumption of an appliance and makes it accessible through proprietary software from Plugwise. To access the real-time consumption data at an interval of 1 Hz, we use the open source *python-plugwise* library [199]. The library is running on a *Sheeva* plug mini computer [200] (*cf.* Figure 3.8) which serves as a sink for the Zigbee network. A python script on the Sheeva queries the data from all connected plugs once a second. The data is then transferred to the Web server at ETH Zurich for analysis.

3.2.6 Passive infrared occupancy sensors

To aid the collection of ground truth data, we deployed six Roving RN-134 low-power Wi-Fi modules with passive infrared sensors attached [153]. The sensors implement a coarse occupancy sensing algorithm and transmit binary occupancy values to the Web server via the Sheeva Plug. The RN-134 modules consume very little power when asleep and can sense while the radio is switched off. When a request is to be transmitted, the radio is switched on for a brief period of time. As soon as the request has been completed, the module goes back to sleep. By sending the data to the Sheeva Plug on the same network (which then forwards the data to the Web server) we significantly reduce the time spent in the awake state and thus minimise the module's energy consumption. This allows the module to run for approximately three months on two AA batteries.

3.2.7 Occupancy ground truth

To obtain ground truth data to develop and evaluate our occupancy detection algorithms, we gave each household a *Samsung Galaxy Tab P7510* tablet computer. On the tablet we installed an application that visualised the current electricity consumption, 7-day historical consumption, aggregate consumption and a historical chart with smooth zooming [95]. The application offers an interface for users to record the occupancy status of the residence. For each occupant there is a toggle button that may be pressed to change the status from *present* to *absent* and vice versa. The tablet computer was installed near the main entrance to the property and the occupants were instructed not to move it during the course of the data collection campaign. In order to increase the visibility of the application and to remind the participants to record their occupancy, we automatically switched on the tablet's display whenever the passive infrared sensor sensed a movement.

3.3 Data collection

The collection of sensor data was split into two phases. In the first phase we deployed the system in two households (households r1 and r2), to test the sensors. We then deployed the system in the remaining households. During the initial testing phase we discovered a number of issues that led to several changes in the way we collected the data. We discovered that the Landis+Gyr E750 smart meters had not been calibrated to provide accurate power consumption values. The consumption was rounded to the nearest 10 watts which was not precise enough for our experiment. We therefore replaced the smart meters with re-calibrated ones. In addition, we decided to deploy our own routers whenever possible to minimise the installation time.

3.3.1 Data formatting

The smart electricity meters produce 86,400 measurements per day. In order to be able to directly compare the electricity consumption to the other sensor data, we converted all other data to 86,400 element vectors as well. The smart plugs from Plugwise must be read sequentially [177]. Queries have a round trip time of 80 ms to 120 ms for each plug, depending on the network infrastructure. As there are 6-9 plugs per household it takes about one second to obtain the consumption data of each plug. However, problems that occur for one of the plugs (*e.g.* a slow reply or a timeout due to network interference) can lead to short time periods of 5-10 seconds during which no data from any of the plugs can be obtained. Ultimately, the consumption measurements for each plug are re-sampled to 86,400 measurements a day (*i.e.* 1 Hz).

For each day d , the occupancy states of a household h are captured by $O_{h,d}$. $O_{h,d}$ is a $86,400 \times N_{h,p}$ matrix containing the occupancy state for each member p of the household at every second of the day. $N_{h,p}$ denotes the number of occupants in household h . The element (i,j) of this matrix is set to 1 if – according to the data entered using the tablet – the j^{th} resident is at home at second i . The element is set to 0 otherwise.

Following this notation, we compute the binary occupancy schedule as $B_{h,d}$, a $86,400 \times 1$ vector by computing the bitwise OR among the rows of the matrix. The resulting vector contains 1s to indicate occupancy and 0s to indicate that none of the occupants are present. For the PIR sensors, the matrix contains a sequence of 1s for the next 30 seconds after a sensor event has been triggered.

3.3.2 Missing data

In case of the electricity consumption data from the smart meters, we distinguish between two types of data loss. First, if measurements are missing for up to 10 seconds, the corresponding positions in the vector are filled with the last existing measurement (typically only few seconds are lost each day). Second, in case more than 10 consecutive seconds of data are lost – for example in (rare) cases where the Flukso crashed or was switched off – the values are set to -1 . For the smart plugs, data loss is dealt with similarly. In this case, we chose 100 seconds instead of 10 seconds as a threshold. This is due to the fact that a data loss of 10 seconds is more common for the reasons described above.

3.4 Description of the dataset

In this section we will describe the features of the data collected in the six instrumented households over the 7-month period from June 2012 to January 2013. As we are interested in opportunistic approaches for occupancy detection, we will mainly focus on the data obtained from the smart electricity meter and relate it to the ground truth occupancy

data gathered through the tablet application. In the next chapter, we will then develop algorithms to infer occupancy directly from the electrical load curve.

3.4.1 A typical day

Figure 3.9 shows a representative day of data collected for household r2. Figure 3.9a shows the total electricity consumption of the household augmented with the binary occupancy state as indicated by the occupants on the tablet interface. The electrical load curve shows a small increase in the electricity consumption when the occupants wake up and prepare breakfast. As the occupants leave the household, the PIR near the doorway fires (see Figure 3.9c). As the occupants return again, the PIR sensor fires again and the home entertainment is switched on (see Figure 3.9b). From the total electricity consumption and the consumption of the stove exhaust fan, it can be seen that shortly after 6 p.m. the occupants prepare dinner. Before midnight, the electricity consumption falls to the nighttime mean and the home entertainment system is switched off.

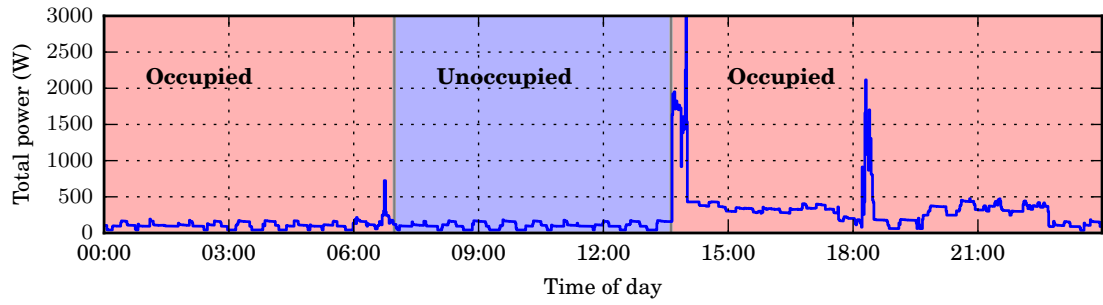
3.4.2 Aggregate electricity consumption

The electricity consumption depends on the current time of day. During the night when occupants are asleep, it is usually at its lowest. During the day, activities like cooking and consuming entertainment are visible. This means that the probability of observing a particular power consumption varies over the course of the day.

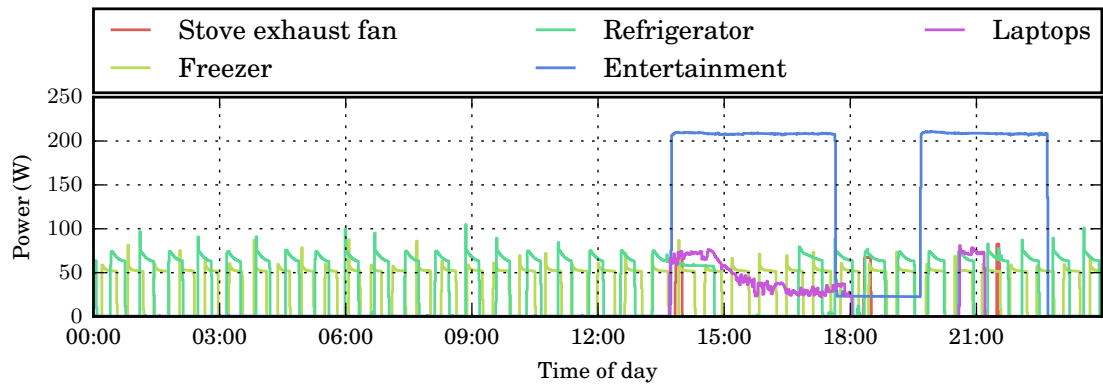
Figure 3.10 shows how this probability distribution varies over for all six instrumented households. The x-axis shows the time of the day in 15-minute intervals while the y-axis shows the distribution of the power consumption during this time. To build the histograms we included all measurements taken in a particular 15-minute interval (*e.g.* measurements taken from 9 a.m. to 9.15 a.m. on any day) and computed the frequency counts for 50 logarithmic bins from 10 watts to 6000 watts. The figure gives an overview of how the electrical load varies during the day. A high probability for a certain power consumption is indicated by a **hot** colour while a low probability is indicated by a **cold** colour.

Household r1 exhibits distinct periods of activity in the morning and in the evening. Figure 3.10a shows a higher activity from 6 a.m. to 9 a.m. and from 6 p.m. to 12 p.m. Both activities can be explained by cooking, while the prolonged activity during the evening indicates the use of entertainment devices. During the night from midnight to 6 a.m., the electricity consumption stays within a narrow range around 100 watts.

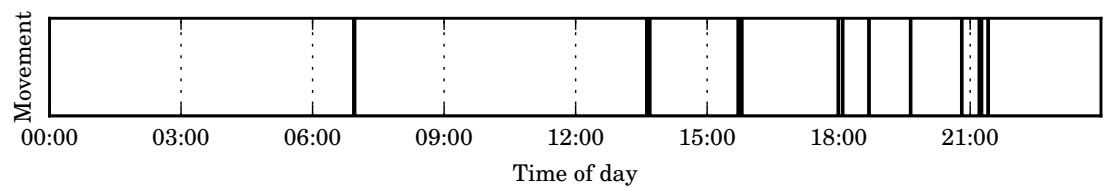
Household r2 In household r2 (*cf.* Figure 3.10b), the electricity consumption during the night is even more stable. There is a faint increase in the probability of a higher electricity consumption from 6 a.m. onwards which may correspond with the occupants getting up



(a) Total power of the household. Blue and red shading indicate periods of occupancy and absence.



(b) Power consumption of individual appliances.



(c) Movements detected by the passive infrared sensor.

Figure 3.9: Sensor measurements for a representative day (15.08.2012) in household r2.

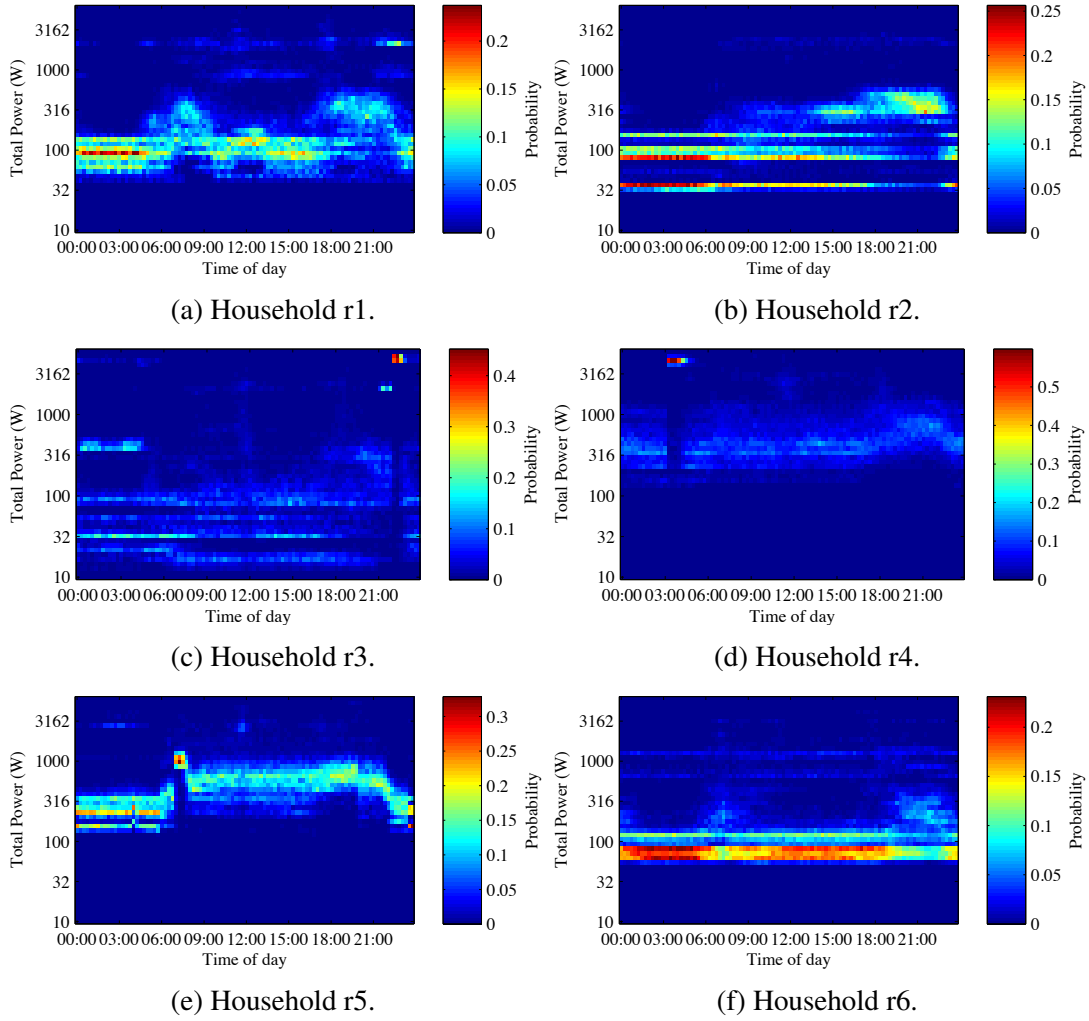


Figure 3.10: Probabilistic electricity consumption over 15-minute intervals.

and starting to use appliances. The increase throughout the morning and early afternoon can be attributed to both an overlapping schedule of the occupants working in shifts and the influence of weekend consumption on the overall probability distribution. After 6 p.m. the probability of a higher consumption noticeably increases. As indicated by Figure 3.9b, this may be explained by the use of the home entertainment system.

Households r3 and r4 Figures 3.10c and 3.10d (households r3 and r4) are more difficult to analyse due to the dominance of the electric boiler during the evening and early morning, respectively. Nevertheless, for both households there is a slightly higher probability for higher electricity consumption during the lunch and dinner hours. In addition, we see the probability of higher consumption figures increasing in the evening.

Household r5 Household r5 shows an interesting behaviour as the electricity consumption is clearly elevated during daytime from 8 a.m. to 8 p.m. As we have learned from the occupants this is due to various timer clocks, one of which is used to operate a pump in the garden during the summer. Again, there is an observable peak in consumption during lunchtime indicating a cooking activity. The plot also shows a high probability for an electricity consumption of 1000 watts around 7 a.m. which might correspond to a coffee machine or another high power kitchen appliance.

Household r6 Figure 3.10f highlights how the variability in consumption may be used to predict occupancy. During the night from midnight to 6 a.m., the electricity consumption is around 100 Watts with a high probability. Around 7 a.m. there is a change with probabilities for higher consumption figures increasing temporarily. During the afternoon, the probability distribution indicates that there is little variance in the consumption. Only after 6 p.m. does the probability for a consumption higher than 100 watts increase again. From this we could infer that the occupants get up at 7 a.m., leave the household and usually do not come back before 6 p.m.

3.4.3 Occupancy ground truth

Due to the importance – and difficulty – to record reliable ground truth occupancy data, we instructed households to particularly pay attention to specify their occupancy during two phases in *summer* (July to September) and *winter* (November to January). During these two collection phases every participant was instructed to click on a button bearing his or her name to indicate presence and absence.

Figure 3.11 shows the occupancy information collected with the tablet computer over the course of the whole deployment in five of the six households. We did not have enough data available to plot the ground truth occupancy for household r6. It is therefore left out from the evaluation. For the plotting we have rounded occupancy figures to binary occupancy values. This means that if one or more persons was present during any one time, the occupancy is assumed to be 1, otherwise the occupancy is set to 0.

Figure 3.11 thus shows a matrix where rows represent days and columns represent 15-minute time slots. White slots indicate that the household was unoccupied (*i.e.* the occupants were away), black periods indicate that the household was occupied (*i.e.* the occupants were home).

Household r1 Figure 3.11a shows how household r1 participated in the summer and winter occupancy collection campaigns. After the summer collection campaign, the occupants went on a holiday and only resumed the ground truth collection around the end of November for the Winter campaign. This means there is no ground truth information in October and November. The data shows that the home is usually occupied over the

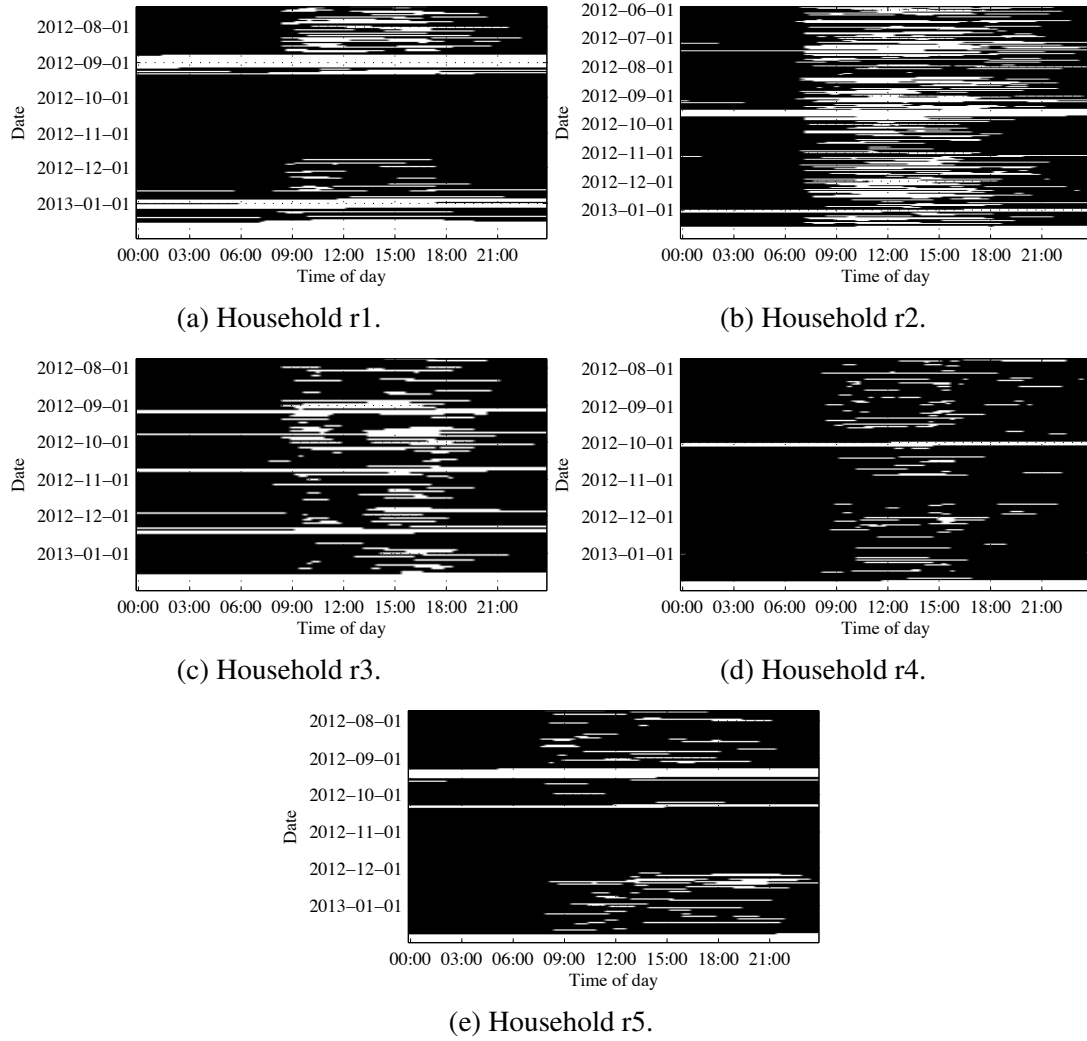


Figure 3.11: Binary occupancy averaged at 15-minute intervals over the whole observation period.

evenings and nights from 6 p.m. to 9 a.m. and that the participants were at home longer and more often during the colder months than during the summer. Overall, the house is occupied more often than not during the daytime which reduces the savings that could be achieved using a smart thermostat.

Household r2 Figure 3.11b shows the ground truth occupancy information for r2. Household r2 is occupied by a couple of young professionals. The occupants exhibit a great regularity, leaving the house around 7 a.m. every weekday. The arrival times in the evening are more distributed due to the different schedules of the inhabitants. However, the occupants are usually at home after 10 p.m. Household r2's holiday absence is also clearly observable from the ground truth occupancy data. The occupancy data for household r2 is

the most complete of the dataset (the occupants collected 7 complete months of ground truth occupancy information). Household r2 also has the lowest day-time occupancy of our participants. This makes it very amenable to the installation of a smart thermostat.

Household r3 The ground truth occupancy for the household r3 is shown in Figure 3.11c. Household r3 also participated over the course of the whole collection campaign. Compared to household r2, household r3 has a much higher occupancy during the day. The building is only left unoccupied for a few hours each day. This means that less of the energy consumed during the day may be saved as occupants are away.

Household r4 Household r4 (Figure 3.11d) has the highest occupancy in the dataset. From 6 a.m. to 10 p.m. the house is occupied over 80% of the time. This is due to the fact that this is a family household with a number of overlapping occupancy routines (*i.e.* people returning before others are coming back, resulting in always at least one person at the property).

Household r5 Figure 3.11e shows the occupancy for another 2-person household of two retirees. Like for household r4, the occupancy in r5 is generally exceeding 80% during daytime.

Quality of the ground truth data

In general, the quality of the ground truth data is quite high. Three of the six households (r2, r3 and r4) participated over the course of the whole experiment. Households r4 and r5, however, exhibit very high occupancy figures and unpredictable schedules. These households may not be suitable for the deployment of a smart thermostat. Furthermore, the high occupancy poses constraints for training the occupancy classifiers in the next chapter. To alleviate some of these problems and to ensure that no erroneous ground truth is used to train the classifiers, we will discuss how we cleaned the dataset in the next section.

3.4.4 Device-level electricity consumption

Besides the aggregated electricity consumption data as measured by the smart meters, we deployed between six and ten smart plugs in each household to measure the consumption of individual appliances. Like the aggregated load curve, this data is available at a frequency of 1 Hz. Figure 3.12 shows the coverage achieved by the smart plugs in all six households. Overall household r2 has the highest coverage with 79% of the total electricity consumption attributable to individual appliances due smart plugs.

In general, apart from household r2, the majority of the electricity consumption is caused by uninstrumented appliances (*e.g.* small electric stoves, boilers and heat-pumps).

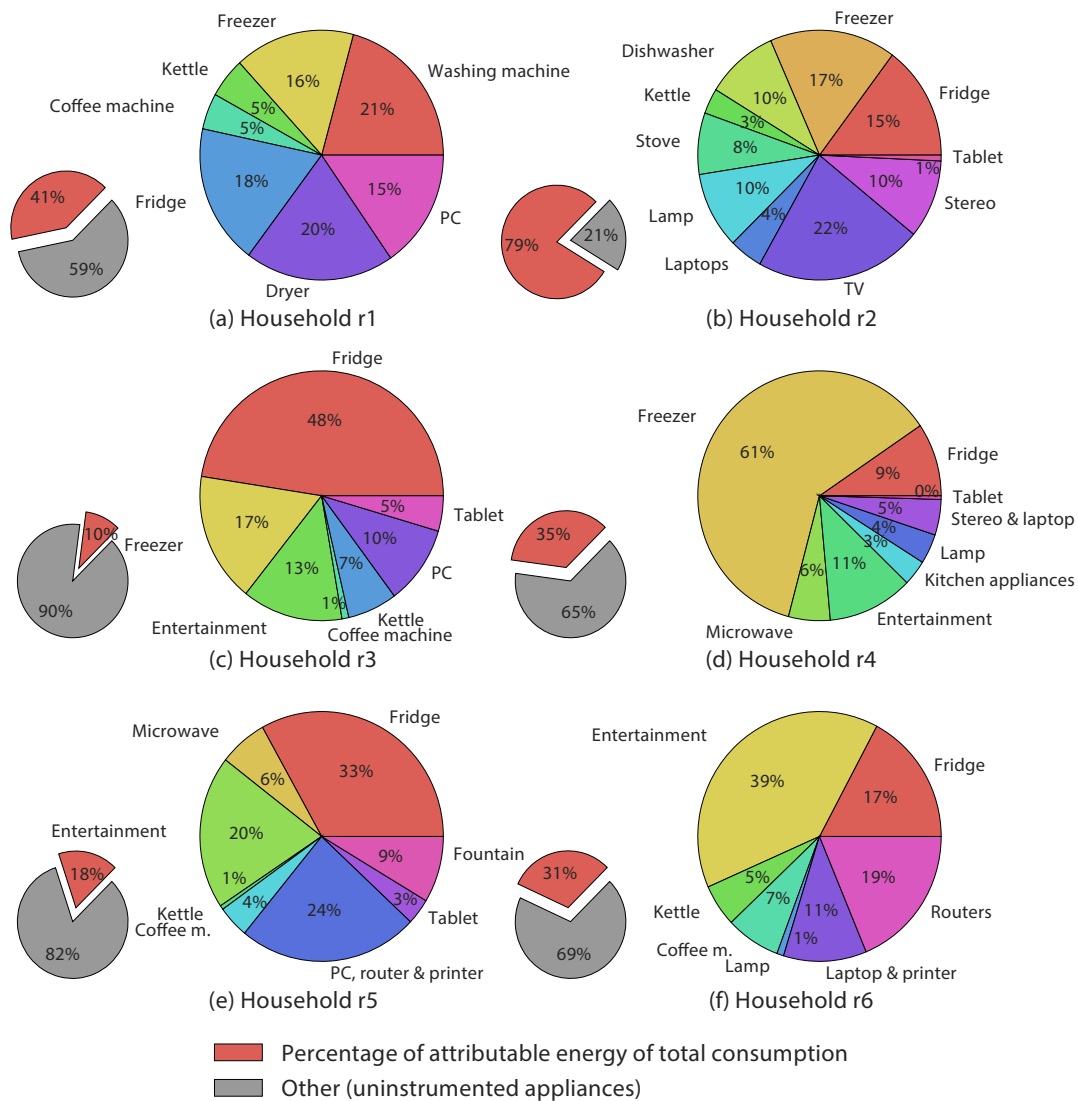


Figure 3.12: Coverage of total energy consumption by smart plugs for all six households.

Table 3.10: Number of days for each household used in the evaluation after data cleaning.

Household	Number of days	
	Summer	Winter
r1	39	46
r2	83	45
r3	57	21
r4	38	48
r5	43	31
r6	<i>no ground truth available</i>	

Due to the limited number of smart plugs, as well as the topology and size of households r1, r3, r4 and r5, it was not possible to fully instrument all appliances in these cases. Household r2, on the other hand, is a small flat (less than 100 m²) that allowed for a comprehensive deployment of smart plugs.

Household r3 has the lowest coverage. Only 10% of the total consumption can be attributed to the instrumented appliances. Household r3 contained a boiler which heated water during the night. More important though, a concrete ceiling in the basement disturbed the radio connection between the gateway and the plugs and thus resulted in measurement outages that incorrectly understate the total consumption of the instrumented appliances. Similarly to the boiler in household r3, household r5 operates a time-triggered pool pump during the summer months which was not measured by a smart plug. This pump consumes around 6 kWh per day, explaining some of the non-attributable consumption.

3.4.5 Data cleaning

In Section 3.3.1 we highlighted how the data was formatted for evaluation. However, days with missing or inconsistent data cannot be used for occupancy classification. Even though the participants noted their occupancy diligently, some mistakes could not be prevented. Occasionally, one or more occupants forgot to record their absence or presence. We have therefore manually removed days where no occupancy information was collected or occupants were supposed to be away but we detected (a) **movement**: a firing of the PIR sensor indicated movement in the household or (b) **appliance usage**: a switch operated device (*e.g.* kettle, TV or oven) was used.

Table 3.10 shows the data gathered by the participants and used in the evaluation. The table shows, for each household, the number of days in both summer and winter phases after erroneous days have been removed from the dataset. This results in an average of 52 days for the summer period and 38 days for the winter period.

3.5 Conclusions

In this chapter we presented an occupancy detection infrastructure and data collection campaign centred around smart electricity meters. To this end, we developed and deployed an infrastructure capable of sampling an off-the-shelf smart electricity meter at 1 Hz granularity in six Swiss households over a period of seven months. We further deployed smart plugs and auxiliary sensors to measure the device-level consumption as well as ground truth indicators. In order to facilitate the development and testing of supervised machine learning algorithms, we further collected ground truth occupancy information using a custom tablet application. The result of our data collection campaign is a comprehensive first-of-its-kind dataset combining detailed electricity consumption with ground truth occupancy data [21].

Building occupancy monitoring using electricity meters

Smart electricity meters are becoming increasingly ubiquitous. In Switzerland, a recent study commissioned by the Swiss Federal Office for Energy (BfE) came to the conclusion that a roll-out of smart meters would be economical and beneficial to consumers [15]. The EU requires¹ Member States to deploy “intelligent metering systems” as part of the Third Energy Package of Directive 2009/72/EC [155]. The directive targets a roll-out of smart electricity meters of 80% by 2020. Today, a total of approximately 45 million smart meters are installed in Finland, Italy and Sweden alone. In Germany, the installation of smart meters was mandated for all new and renovated buildings [1]. According to estimates of the European Commission, commitments of the Member States amount to a roll-out of “close to 200 million smart meters” by 2020 [50].

The trend towards installing smart meters opens new possibilities for opportunistic sensing in buildings. Current building automation systems use a dedicated sensing devices such as PIR sensors and reed switches to provide occupancy monitoring capabilities [7, 124]. Recent experimental systems also show the feasibility of opportunistically using network logins and GPS trackers to monitor occupancy [67, 101, 113, 144]. In such systems, sensors are often combined to increase the overall occupancy detection accuracy. For instance, the system presented in [7] combines door-mounted magnetic reed switches and PIR sensors to compensate for the poor accuracy obtained when only PIR sensors are used. However, for reasons of cost and ease of deployment, current commercial smart thermostats for the residential environment often only include a single PIR sensor [198]. This restricts their ability to accurately monitor the occupancy throughout the building and results in erroneous control decisions. As a result, users of such occupancy-controlled smart thermostats often turn off automatic control in an attempt to regain control of the system [186].

¹The adoption according to 2009/72/EC is subject to a cost-benefit analysis.

In this chapter, we discuss and quantitatively evaluate the suitability of smart electricity meters to be used for occupancy monitoring in residential households. Being already present – or about to be installed – in millions of households worldwide, the installation, use and maintenance of smart electricity meters does not impose additional costs on the residents. The opportunistic use of such existing sensors increases the occupancy monitoring capabilities and thereby the acceptance of building automation systems.

To this end we show that occupancy classification from aggregated electricity consumption of the household using supervised machine learning is feasible. In particular, we show that a detection accuracy of up to 94% can be obtained. For our analysis we utilise the ECO dataset introduced in the previous chapter. The ECO dataset contains the aggregated consumption of six households at a frequency of 1 Hz as well as ground truth annotations and device-level consumption data for selected appliances.

We begin this chapter in Section 4.1 by highlighting related work in the analysis of electricity consumption data². We then go on to describe our proposed system in Section 4.2 and discuss the metrics used for evaluation in Section 4.3. After we present the results of detecting household occupancy from the aggregated electricity consumption in Section 4.4, we discuss how the performance of the system would change if device-level consumption data was available in Section 4.5. Finally, before we conclude in Section 4.7, we present results for a simple, unsupervised classification strategy that works on 15-minute data as recorded by many utility companies today. This chapter is partially based on the contributions made in [102].

4.1 Related work

In building automation, the occupancy of a room or building at any given time is commonly determined by interrogating sensors. Firings from a passive infrared sensor or reed switch can indeed give an accurate assessment of the current occupancy. However, when non-binary sensors such as smart electricity meters are considered, algorithms must be designed to translate raw measurements into real information about the occupancy state of the household. We see our work at the intersection of two main areas: (1) the analysis of coarse-grained electricity consumption data to observe and influence users' electricity consumption behaviour; (2) non-intrusive load monitoring (NILM) to sense the activation state of appliances in the household. In this section, we first cover basic concepts of time series analysis. We then discuss how to infer household characteristics from electric consumption data in Section 4.1.2 and NILM in Section 4.1.3. We conclude by highlighting first work in inferring occupancy from the electric load curve in Section 4.1.4.

²For related work on occupancy sensing, in general, the reader is referred to Section 3.1.

4.1.1 Time series analysis

Electricity consumption data typically consists of successive temporal measurements – so called *time series data*. The time interval between these data can vary from a few milliseconds to a number of days. Conventional smart electricity meters transmit the aggregated electricity consumption to the utility company at a period of 15 to 30 minutes. However, while, in general, special hardware is necessary to measure the electricity consumption at multiple kilohertz [68], some current smart meters are able to sample the electricity consumption at 1 Hz or more [116].

Various representations for such time series data exist. Lin *et al.* provide a good overview in [122]. In their paper, the authors differentiate between data adaptive and non data adaptive representations. The former includes techniques like symbolic representations and singular value decomposition (SVD)³. The latter include discrete Fourier transforms (DFTs) and discrete wavelet transforms (DWTs) which transform data into a different domain (*e.g.* from the time to the frequency domain).

Lin *et al.* contribute to the space of adaptive techniques by proposing Symbolic Aggregate ApproXimation (SAX), a method to reduce time series data to a lower-dimensional space by normalising the input data and dividing the resulting data into a number of equal sized frames [121, 122]. Using symbols to denote the frame in which a data point falls, the time series is effectively transformed into a string representation. This representation enables the identification of characteristic patterns (so-called motifs). When applied to electricity consumption data, such motifs could thus be the cooling cycle of a refrigerator or the heating and spinning phases of a washing machine. However, as the z-normalisation step removes all amplitude information from the signal, SAX has difficulties to distinguish characteristic power levels. To address this problem, Reinhardt *et al.* introduce Power-SAX [162] – a method which forgoes normalisation and infers the power levels using clustering on the individual appliance data.

In signal processing, time series data is often transformed from the time to the frequency domain. This transformation enables the identification of specific frequencies in the original data. A Fourier transform could thus identify periodic cooling intervals of fridges and freezers – if the underlying process was behaving linearly. However, from manual inspection of the electric load curves in the ECO dataset, we learned that cooling devices do not always operate in linear fashion. In fact, the length of the interval between successive cooling cycles varies considerably. It is thus not possible to predict the exact time of the next cooling cycle. We assume this is due to environmental changes such as the current load of the refrigerator, the indoor temperature and the opening and closing of the door.

³Also known under the term principal component analysis (PCA).

4.1.2 Inferring household characteristics from the electric load curve

The planned installation of smart electricity meters in millions of households worldwide has sparked interest in the analysis of home electricity consumption data over the last years. Several researchers have investigated how this data can be leveraged to infer knowledge about the behaviour of households' occupants.

Previous work has shown that the use of coarse-grained electricity consumption data (*i.e.* one sample every 30 minutes) is sufficient to infer information about households and their occupants. Energy providers can identify usage patterns in the electricity consumption data to predict future electricity consumption [39] and model daily routines to improve a providers's supply management [5].

Other researchers proposed approaches that can cluster hundreds of households into groups of consumers according to their load profile [166, 180]. Beckel *et al.* show that it is possible to infer socio-economic characteristics of a household from its electricity consumption [22]. To this end, they rely on a dataset of electricity consumption data from more than 4,000 households. Similarly, Albert *et al.* propose a method to infer information on the demographic and appliance stock characteristics of homes [9]. By means of a Hidden Markov Model, the authors first infer the occupancy states of the household from electricity consumption data. In contrast to our work, the authors need to use maximum-likelihood estimations of the emission and transition probabilities as no ground truth occupancy data was available.⁴ They then characterise each household using the magnitude, duration and variability of its occupancy states. Using these techniques, energy providers can for instance identify which households are typically unoccupied during the day. These households represent ideal targets to be offered special tariffs or to be encouraged to adopt a smart heating system [9, 22].

4.1.3 Non-intrusive load monitoring

By analysing the fine-grained⁵ electricity consumption of a household, many researchers have tackled the problem of inferring which appliances are running when. The problem of detecting the activation state of individual appliances and their consumption is usually referred to as non-intrusive load monitoring (NILM). Zoha *et al.* [189] and Zeifman *et al.* [188] provide two good reviews of related work in NILM algorithms. NILM approaches are related to occupancy detection because the activation state of certain home appliances can be used as an indicator of occupancy.

⁴For this reason, this work also does not constitute an evaluation of occupancy detection using electricity consumption data. We will evaluate Hidden Markov Models in Section 4.4.

⁵By fine-grained we mean sampling rates of 1 Hz or more.

One of the first NILM approaches has been proposed by George Hart in 1992 [69]. Hart's method identifies characteristic step changes in the electricity consumption. By comparing these step changes monitored in the electricity consumption with a previously recorded signature database, Hart claims to detect when appliances are being switched on or off. More recent approaches, such as the one from Kim *et al.* [96], pursue unsupervised disaggregation. These unsupervised approaches do not require a training phase, but require only an explicit labelling of those appliances detected in the load curve.

As some devices are (typically) only used when the occupants are at home, NILM would implicitly provide occupancy detection as required by many energy efficiency applications. However, if the electricity consumption is measured at a granularity of at most 1 Hz, only a few appliances (*e.g.* the refrigerator, or the washing machine) can be detected reliably from the data [31]. Also, the activation state of an appliances might or might not correlate well with occupancy. For instance, a refrigerator is typically active irrespective of the presence or absence of the occupants at home. In addition, several home appliances (*e.g.* the dishwasher or the washing machine) can be programmed to start their operation when the occupants are away.

Increasing the accuracy of detecting individual appliances in the electricity consumption data requires a more characteristic signature of each appliance, which can be achieved by increasing the measurement granularity. As Gupta *et al.* show, this approach can identify and classify most consumer electronic and fluorescent lighting devices correctly with a mean accuracy of more than 93% [68]. However, while it requires special hardware to measure the electricity consumption at multiple kilohertz, our approach relies only on 1 Hz consumption data, which can be obtained from an off-the-shelf electricity meter.

4.1.4 Occupancy and the electrical load curve

In [136], Molina-Markham *et al.* suggest that household activities can be inferred from aggregated electricity consumption data. They collect data at one-second granularity from three homes over two months and let household occupants annotate which appliances they have used when. The annotation of the data is performed over “*at least three days*” [136]. The authors also observe that there are differences in the consumption data depending on whether the occupants are present or absent from home. However, their observation is based on visual inspection of the electricity consumption curves. No quantitative analysis of the possibility to use aggregated electricity consumption data to detect occupancy is provided. Boait *et al.* suggest to derive timer settings for a smart thermostat from electricity consumption and hot water use [28]. The authors use a Bayesian model to relate sensor measurements to occupancy. As the model is focussed on heating only, feedback on the current temperature is used to refine the current timer settings and the occupancy detection.

Chen *et al.* have discussed the potential of smart electricity meters to be used for performing non-intrusive occupancy monitoring [33]. In particular, they presented a

threshold-based method to detect occupancy from aggregated electricity consumption data. The authors evaluated their method using data collected in two homes over a summer week. We build upon this work by considering a large set of features including those used by Chen *et al.* and basing our analysis on a dataset collected in five homes and over a period of more than six months.

Other work explores occupancy monitoring using detailed device-level information. Ming *et al.* for example present PresenceSense, a zero-training algorithm based on rough estimates of the participants' working schedules [87]. It uses the average power, standard deviation and absolute maximum power change of individual plug-level loads measured by ACme nodes [86] to estimate occupancy in an office environment. In contrast to Ming *et al.*, however, our work focusses on residential environments and depends only on the installation of a single smart electricity meter instead of multiple smart plugs.

Orthogonal to our work, Yang *et al.* investigate the information leaked when an infrastructure consisting of passive infrared sensors or smart electricity meters is controlled by a third-party. To this end, the authors assess how well the occupancy levels of a university lab can be inferred from its electric consumption data alone [185].

4.2 System design

In residential households, a significant share of the electricity consumption is caused by human interaction with electrical appliances. Therefore, a household's electricity consumption may give an indication of its current occupancy state. As electrical appliances are often used to replace manual labour and to increase comfort, a higher level of consumption often correlates with occupancy⁶. Figure 4.1 revisits the example day from household r2 introduced in the previous chapter. Occupied periods are indicated in red, while blue periods indicate an unoccupied household. The bottom part of the figure shows a very simple occupancy detection strategy. Whenever the current total power is higher than the 24-hour mean of the total power, the house is classified as occupied. The figure shows that even such a rudimentary approach can detect occupancy. In this section we will build upon the correlation between electricity consumption and occupancy and further investigate how one can use features of the electrical load curve to build an occupancy classification engine using machine learning algorithms.

Figure 4.2 shows an overview of our occupancy classification setup. The first step consists of dividing the raw consumption data into 15-minute slots and extracting relevant features from the consumption data. Each of these *examples* is assigned a label (*i.e.* 1 or 0, referring to an *occupied* or *unoccupied* household, respectively) based on the ground truth

⁶A washing machine may for example be used when occupants are present or be programmed to finish upon the arrival of the occupants. If sufficient training data is available, its activation state can be used to infer occupancy in both cases.

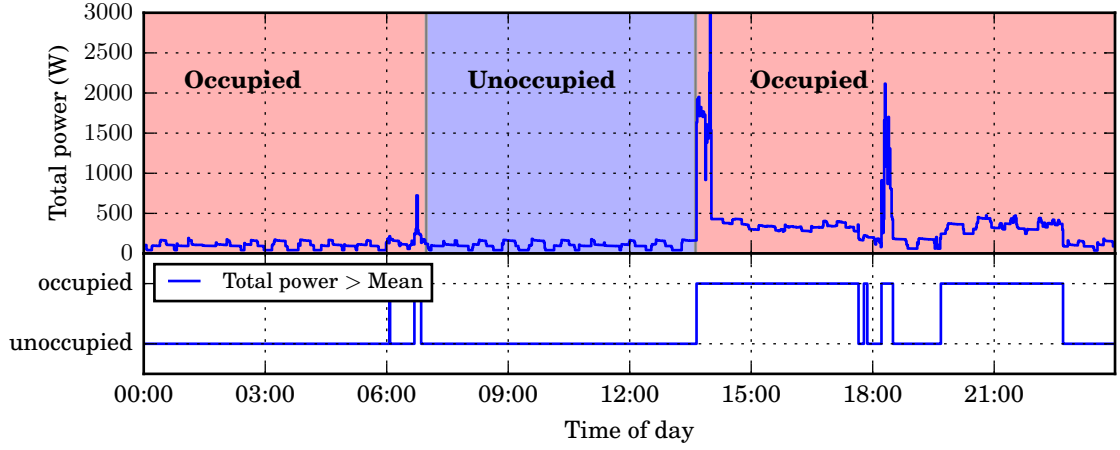


Figure 4.1: Simple occupancy detection algorithm based on thresholding.

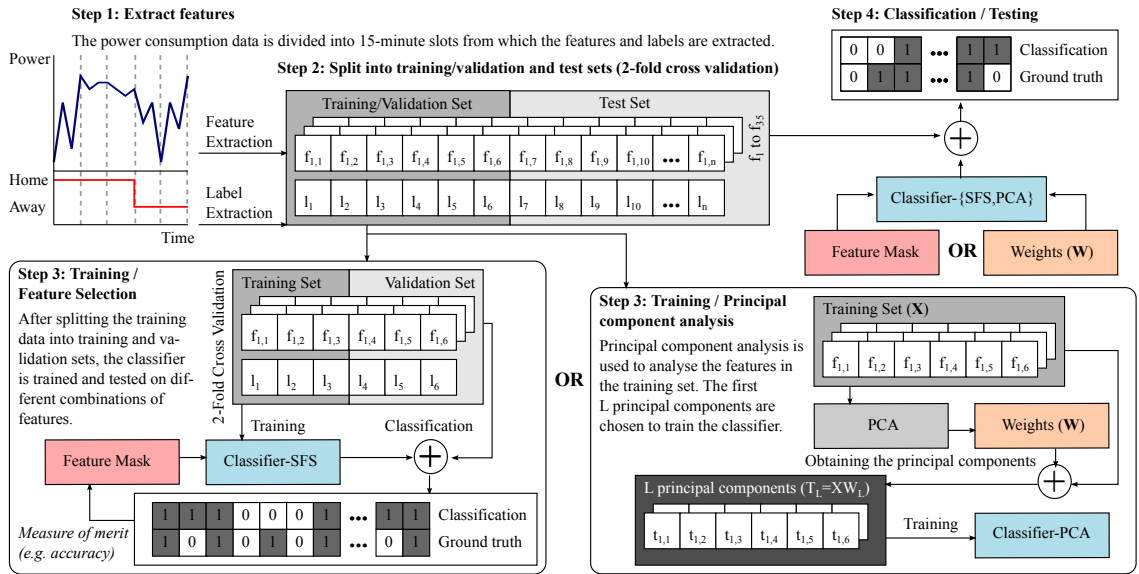


Figure 4.2: Overview of our occupancy classification system.

data gathered by our participants using the tablet application. The examples are divided into training and test sets using cross validation before being assigned to a classifier for training. During training, feature selection selects the most descriptive features, while principal component analysis finds a transformation of the input data and selects the components containing most of the variance. After the training phase is completed, the trained classifier is evaluated on the test data. In the remainder of this section, we will explain each of these steps in detail.

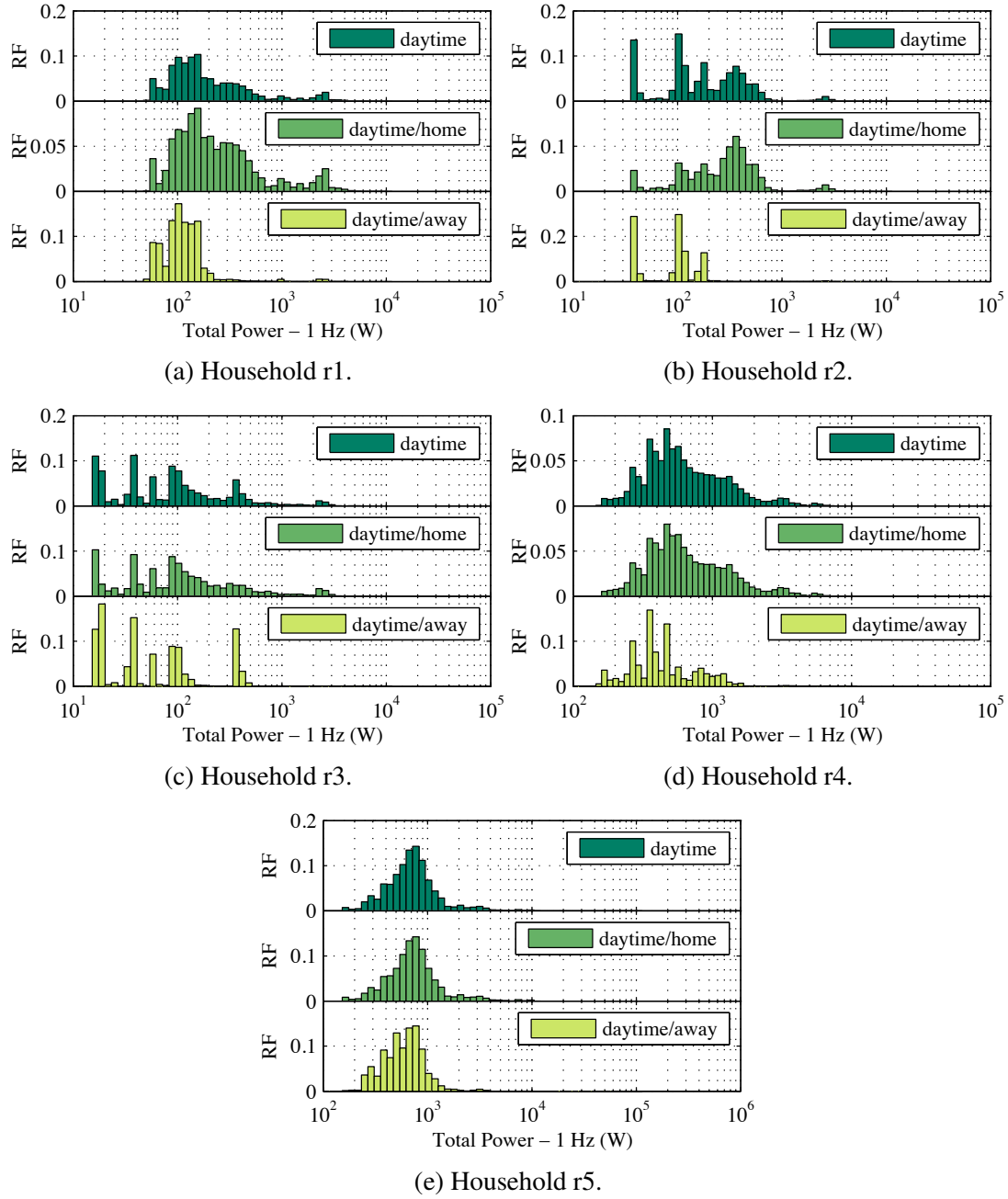


Figure 4.3: Relative frequencies of various total power consumption (sum of all phases at 1 Hz) values during daytime and divided into presence and absence respectively.

4.2.1 Deriving features from the electrical load curve

In order to derive occupancy information from the electrical load curve, it is necessary to identify features that may be indicative of occupants being present in the household. A clear indicator for occupancy are changes in the electric load that require the interaction of an occupant with a device or appliance (*e.g.* the operation of a television, stove or kettle). The electricity consumption induced by appliances such as fridges, freezers or the standby consumption of electric devices (*e.g.* the consumption of the digital video recorder) on the other hand does not give any indication about the occupancy state of the household as no direct interaction is required.

As introduced in Section 4.1.3, a number of authors have looked at NILM approaches to detect the consumption of individual appliances from the electric load curve. However, while these reliably detect high-power devices and cooling appliances, current state-of-the-art NILM approaches require – for each household – detailed training data. This data is typically gathered by recording the aggregated electricity consumption and observing the activation states of installed appliances [21]. This training overhead makes the use of NILM for occupancy detection infeasible. In the following section, we therefore identify a set of appliance-agnostic features of the aggregated electrical load curve that directly relate to occupancy.

Such features can be found by comparing the day-time electricity consumption during periods of occupancy to times when the household is unoccupied. Since the ECO dataset does not contain ground truth data on sleeping patterns, we consider *daytime* slots from 6 a.m. to 10 p.m. in our analysis and leave the detection of sleep to future work.

Figures 4.3a to 4.3e show the relative frequency (empirical probability) of the logarithmically binned total power consumption measurements over summer and winter periods for all five households. Each figure shows from top to bottom the probability distribution of observing a particular power consumption during *daytime* and when occupants are *home* or *away*, for the particular household. We will refer to the latter two probability distributions as the *home* and *away distributions*.

From Figures 4.3a and 4.3b we can see that the power consumption in households r1 and r2 is likely to have a higher *absolute value* and a greater *variability* whenever the respective household is occupied. The away distribution is centred around 100 watts and may be clearly distinguished from the overall daytime distribution. While the household is occupied, the probability of observing a higher consumption increases. However, there is still a significant probability to see lower consumption values even when the household is occupied, resulting in an increased variance of the home distribution. This is due to the fact that occupants may be at home but not using any electrical devices. The two peaks around 30 and 100 watts in household r2 correspond to the operation of cooling appliances and are thus visible in both the home and away distributions. As the home and away distributions in households r3 to r5 are more difficult to distinguish we will focus on

Table 4.1: Features computed on the aggregated electricity consumption traces.

#	Feature names	Description
f_1, f_2, f_3	\min_1, \min_2, \min_3	Minimum of the samples in the slot for phase 1, 2 and 3
f_4	\min_{123}	Minimum of the samples in the slot for the sum of phase 1, 2 and 3
f_5, f_6, f_7	\max_1, \max_2, \max_3	Maximum of the samples in the slot for phase 1, 2 and 3
f_8	\max_{123}	Maximum of the samples in the slot for the sum of phase 1, 2 and 3
f_9, f_{10}, f_{11}	$\text{mean}_1, \text{mean}_2, \text{mean}_3$	Arithmetic average of the samples in the slot for phase 1, 2 and 3
f_{12}	mean_{123}	Arithmetic average of the samples in the slot for the sum of phase 1, 2 and 3
f_{13}, f_{14}, f_{15}	$\text{std}_1, \text{std}_2, \text{std}_3$	Standard deviation of the samples in the slot for phase 1, 2 and 3
f_{16}	std_{123}	Standard deviation of the samples in the slot for the sum of phase 1, 2 and 3
f_{17}, f_{18}, f_{19}	$\text{sad}_1, \text{sad}_2, \text{sad}_3$	Sum of absolute differences of the samples in the slot for phase 1, 2 and 3
f_{20}	sad_{123}	Sum of absolute differences of the samples in the slot for the sum of phase 1, 2 and 3
f_{21}, f_{22}, f_{23}	$\text{cor}_1, \text{cor}_2, \text{cor}_3$	Value of the autocorrelation function at lag 1 computed over the samples in the slot for phase 1, 2 and 3
f_{24}	cor_{123}	Value of the autocorrelation function at lag 1 computed over the samples in the slot for the sum of phase 1, 2 and 3
f_{25}, f_{26}, f_{27}	$\text{onoff}_1, \text{onoff}_2, \text{onoff}_3$	Number of detected on/off events within the slot for phase 1, 2 and 3
f_{28}	onoff_{123}	Number of detected on/off events within the slot for the sum of phase 1, 2 and 3
f_{29}, f_{30}, f_{31}	$\text{range}_1, \text{range}_2, \text{range}_3$	Range of the samples in the slot for phase 1, 2 and 3
f_{32}	range_{123}	Range of the samples in the slot for the sum of phase 1, 2 and 3
f_{33}	p_{prob}	Empirical probability of the slot to be occupied
f_{34}	p_{fixed}	1 (occupied) from 9 a.m. to 5 p.m., 0 (unoccupied) otherwise
f_{35}	p_{time}	Slot number (<i>i.e.</i> 1 – 65)

households r_1 and r_2 to select the features⁷.

Table 4.1 shows the features selected to represent these observations. All features are computed over 15-minute intervals. To this end, we represent a day as a sequence of N_s time slots of length T_s . Given the sampling frequency of 1 Hz and an slot length $N_s = 15$, each feature is thus computed from a 900-element vector (*i.e.* $T_s = 900$). All features – apart from p_{prob} , p_{fixed} and p_{time} – are computed separately for each of the three phases of the smart electricity meter as well as for the sum of the three phases. We use the subscripts $_1, _2$ or $_3$, to indicate that a feature has been computed on the data trace corresponding to phase 1, 2 or 3, respectively. The subscript $_{123}$ indicates that the feature has been computed on the sum of all three phase traces.

⁷We will further discuss the reasons for this behaviour and the resulting implications in Section 4.4.

In all, we consider the features `min`, `max`, `mean`, `std`, `sad`, `cor1`, `onoff`, `range`, `pprob`, `pfixed` and `ptime`. In choosing these features we aim to capture both the absolute value of the consumption as well as its variability.

Absolute value of the power consumption

The features `min`, `max` and `mean` represent the minimum, maximum and arithmetic average of the samples within a 15-minute slot. The boilers in the five households were programmed to operate during the night. Therefore, the absolute level of the consumption is likely to be influenced by presence in the household.

Variability of the power consumption

Features `std`, `sad`, `cor1` and `onoff` serve as indicators of the variability of the power consumption. A high variability indicates that there have been significant changes in the electricity consumption during the observed interval. Such changes may have been caused by human actions (*e.g.* by operating the stove or the kettle) or by appliances with varying consumption patterns (*e.g.* a television set with LED backlight).

Feature `std` denotes the standard deviation of the power consumption. As the standard deviation only measures the distance to the mean of the data we have introduced an additional measure `sad` – the sum of absolute differences. `sad` computes the absolute difference between adjacent power measurements and adds them up, giving another measure of the variability of the data. `cor1` is the value at lag one of the autocorrelation function of the sequence of samples in a slot. Finally, `onoff` is the number of on/off events detected within a slot.⁸

Temporal dependence of occupancy

As the probability of the building being occupied varies with time, we introduced three additional features – `pprob`, `pfixed` and `ptime` to model the temporal aspects of occupancy. `pprob` is the empirical prior probability of a slot to be occupied. `pprob` is computed from the ground truth occupancy data. To this end, only data from the training set is used. `pfixed` is a “dummy” prior probability that assumes the household to be always unoccupied between 9 a.m. and 5 p.m. on weekdays and to be always occupied otherwise. `ptime` introduces a notion of time by adding the slot number as a feature. Slots are numbered from 1 to 65, whereas the first slots corresponds to the period between 6 a.m. and 6:15 a.m. and the last one to the period between 10 p.m. and 10:15 p.m.

⁸An on/off event occurs when a electrical device is switched on or off. We detect on/off events using a simple algorithm: if the difference between a sample and its predecessor is bigger than a threshold ThA and this difference remains higher than ThA for at least ThT seconds, an on/off event is detected. We set $ThA = 30$ W and $ThT = 30$ seconds.

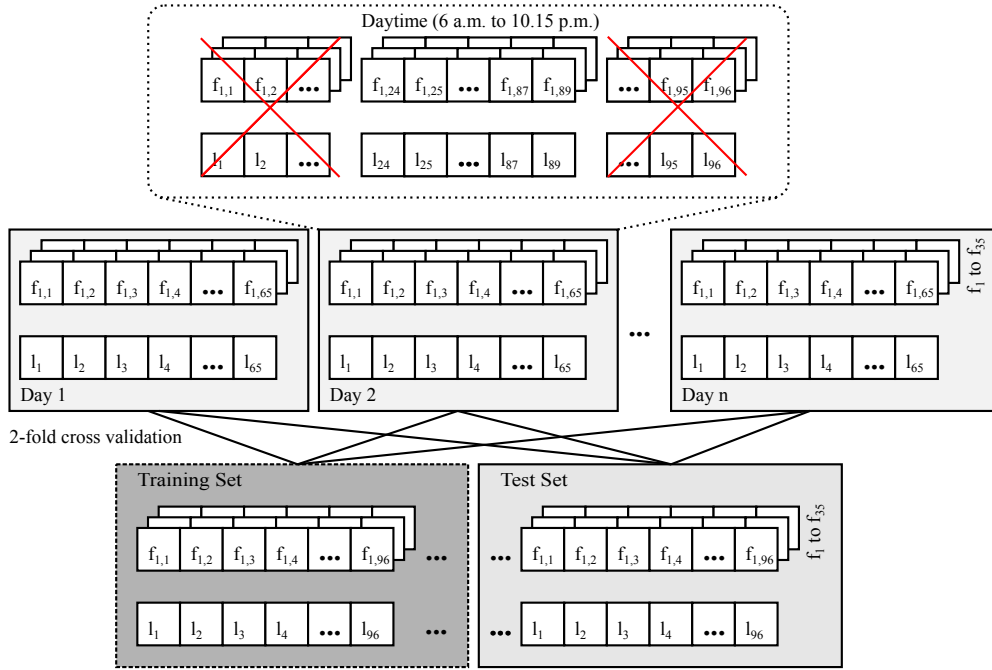


Figure 4.4: Features and labels are computed per day. From the 96 slots produced by computing the features at 15-minute intervals only the 65 slots between 6 a.m. and 10.15 p.m. are used. The days are allocated to the training and test sets using two-fold cross validation.

4.2.2 Cross validation

To avoid a specific choice of the training and test sets to create artefacts in the obtained results, we use two-fold cross validation [184] on ten different, randomly selected pairs of training and test sets. That is, we randomly divide the feature data ten times into different, equi-sized *training* and *test* sets.

For each, the former is used to train a classifier (*cf.* Section 4.2.3) and the latter to evaluate its performance. Then, the role of the two sets is swapped and training and evaluation are repeated. The performance of the classifier is computed as the average of the performance obtained in the two cross-validation runs. The overall performance is computed as the average of the performance obtained in each of the ten runs.

The use of ten runs also allows to analyse the stability of the feature selection (*cf.* Section 4.2.4), that is to ascertain if different test and training sets yield different feature sets. Feature selection is performed using an additional two-fold cross validation on the training data (*cf.* Figure 4.2).

Classification limited to daytime hours Figure 4.4 shows the arrangement of the input data for classification. Table 3.10 in Chapter 3 showed the number of days for each household in the Summer and Winter periods. Each of these days is represented by 96

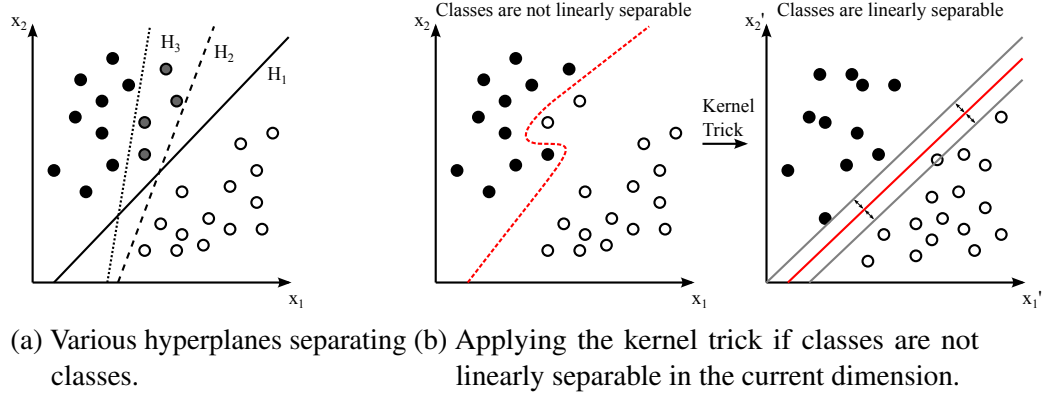


Figure 4.5: Classification using support vector machines.

15-minute slots. For each slot i , the input for the classifier is the feature vector $f_{i,1...35}$. As we do not have ground truth data on the sleeping patterns of our participants, we restrict the estimation of the occupancy states of a household to the time between 6 a.m. and 10.15 p.m. Thus only $f_{24,1...35}$ to $f_{89,1...35}$ are used for training and testing.

4.2.3 Classifiers

To infer occupancy from the electrical load curve, one must infer a mapping function from the feature space (e.g. a mean consumption of 100 watts in the last 15 minutes) to occupancy *classes* such as *home* and *away* (e.g. [feature] \rightarrow class). This mapping function – the *classifier* – can be inferred using models and algorithms from *supervised machine learning*.

Supervised machine learning techniques infer the classifier from labelled training data. During the training phase, the classifier is iteratively refined to correctly assign as many examples (i.e. [[feature], class] tuples) as possible to their respective classes.

To avoid overfitting the classifier – that is, to build a classifier that describes the noise in the data rather than the underlying relationship between features and classes – the data must be split into *training* and *test* sets. The test set is used to provide an unbiased test of how well the trained classifier performs for previously unseen data. We will discuss the details on how we split the data into training and test sets in Section 4.3.

Several learning algorithms for training a classifier have been proposed in the literature [184]. The learning algorithms used in this chapter are support vector machines (SVMs), K-nearest neighbours (KNNs), Gaussian mixture models (GMMs), hidden Markov models (HMMs) and a simple thresholding (THR) approach. The SVM and KNN classifiers have been chosen to evaluate both parametric and non-parametric approaches, while the HMM was chosen to reflect the temporal dependence of occupancy.

Support vector machines (SVMs) are supervised learning models and algorithms to perform linear and non-linear classification. A SVM models the examples of the training set as points in space. It then constructs a hyperplane that separates these examples with the widest possible margin, the *maximum-margin hyperplane*. In the simple case where two classes are linearly separable, it is possible to select two hyperplanes that separate the examples such that there are no data points between them and their distance – the margin – is maximal. The examples closest to these hyperplanes are referred to as support vectors.

There may be an infinite number of different hyperplanes separating the classes. Figure 4.5a shows two classes separated by three different hyperplanes in a simple, two-dimensional feature space. H_1 and H_2 both separate the two classes. H_1 achieves the maximum margin. H_3 does not separate the two classes at all and misclassifies some instances of the first class.

Figure 4.5b shows two classes that are not linearly separable. In such cases, an SVM resorts to a higher-dimensional space using a procedure called the kernel trick [184]. In the case of occupancy detection this may be necessary when occupants regularly run the washing machine while they are away. The operation of the washing machine increases the overall load above a threshold that would in a linear model indicate occupancy but to a lower level than other indicators of occupancy such as the electric stove. A SVM can identify the characteristic power consumption of the washing machine and correctly classify such periods as absence. To implement the SVM classifier we used the LIBSVM library by Chang and Lin [32].

K-nearest neighbour (KNN) classifiers use non-parametric models for classification. This means they do not require an explicit learning phase. Instead during the classification of an instance of test data, the KNN algorithm first finds the k closest examples in the training data according to some distance metric (*e.g.* Euclidean distance, Hamming distance, etc.). Using the known classes of these k neighbours, it then takes a majority vote on the class membership of the unknown class of the test data. For the KNN classifier we used the `ClassificationKNN` classes from the Matlab Statistics Toolbox [215]. We use $k = 1$ and employ the Euclidean distance to find nearest neighbour.

Thresholding (THR) The THR classifier is based on the observation made in Figure 4.1 that a higher electricity consumption positively correlates with occupancy. To this end, the classifier computes the mean of each feature vector [feature] during all *unoccupied* slots. These means are used as thresholds above which a slot is labelled as occupied. The classification of the test data is based on a majority vote of the thresholding applied to all features. The thresholding classifier implicitly assumes that the magnitude of the features positively correlated with occupancy.

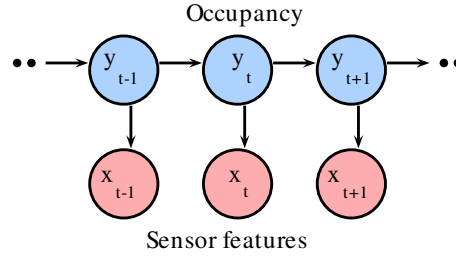


Figure 4.6: Hidden Markov Model.

Gaussian mixture models (GMMs) are parametric probability density functions represented by a weighted sum of individual Gaussian component distributions [163]. Due to the limited size of the ECO dataset and the resulting sparsity in the feature data it is not possible to build empirical multivariate probability density functions as shown in Figure 4.3 for a combination of all features. However, having observed that the process generating the raw power data is approximately following a log-normal distribution, we use multivariate GMMs to approximate the probability density function of the input data.

The GMM is created by iteratively refining the parameters of a combination of K Gaussian distributions (components) to fit the input data. During the training phase we choose a suitable K by minimising the Akaike information criterion (AIC) [8]. The AIC rewards goodness of fit, while including a penalty for the number of components used.

To implement the GMM classifier, we use the `gmdistribution` class from the Matlab statistics toolbox [216]. During training, we build GMMs for both the occupied and unoccupied distributions. During testing, we perform maximum-likelihood classification by evaluating the likelihood of a test example belonging to either distribution. Besides being used on their own as a maximum-likelihood classifier, we also employ GMMs to smooth the emission probabilities for the Hidden Markov Model classifier.

Hidden Markov model (HMM) A HMM (see Figure 4.6) is a statistical state model that relates (hidden) states (*e.g.* occupied, unoccupied) to emissions (*e.g.* the observed electricity consumption) using matrices of emission and transition probabilities. In this case, the observed electricity consumption at time t is then indicated by $x_{t=1\dots T}$. The unknown (occupancy) state is given by $y_{t=1\dots T}$. HMMs improve upon the SVM, KNN, THR and GMM approaches by considering the fact that it is usually not very likely for a household to change its occupancy state during a particular 15-minute interval. Indeed, continuously switching the occupancy state is both unlikely and potentially harmful for a smart heating system.

The training of a HMM requires a matrix of emission probabilities. In the context of occupancy classification from electricity consumption data, this could for example be the probabilities of observing some discrete level of power consumption during both occupied and unoccupied states. Due to the limited observation time, the ECO dataset does not

contain examples of all power levels for both states. To overcome the sparsity in the input data, we build the emission probabilities using 2-dimensional GMMs of the first principal component (*cf.* Section 4.2.4) and the p_{time} feature (*e.g.* the slot number) for both the occupied and unoccupied states, respectively. We obtain the discrete emission probabilities by numerally evaluating the integral of the GMMs.

4.2.4 Dimensionality reduction

Our feature set (*cf.* Table 4.1) contains 35 features. Using this complete set of features allows to capture a variety of characteristics of the electricity consumption curve. However, while some classifiers might take advantage of all the features included in this set, others provide better performance operating only on a subset thereof. Indeed, for each classifier there exists an optimal subset of features that maximises its performance [182]. To limit the set of features to the most descriptive ones, we apply algorithms to reduce this dimensionality as detailed below.

Feature selection

The set of optimal features can be found through an exhaustive search over all possible subsets of the feature set [182]. However, the complexity of performing an exhaustive search grows exponentially with the number of features [85].

To avoid such a computationally expensive operation⁹, we use a so-called *feature selection algorithm* to identify adequate subsets of features. In particular, we adopt the Sequential Forward Selection (SFS) [182] algorithm. SFS is an iterative algorithm that applies a simple heuristic to determine the feature subset that allows to maximise a performance metric J (*e.g.* the accuracy or Matthews correlation coefficient (MCC)).

Listing 4.1: Sequential Feature Selection (SFS)

```

1  $X = [x_0 \dots x_n]$ ; // Set of all features
2  $Y = \{\emptyset\}$ ; // Best feature set of length  $k$ 
3  $m$ ; // Maximum number of features
4
5 while ( $k \leq |X|$  &&  $k \leq m$ ) {
6   // Inclusion of best feature
7    $x^+ = \arg \max_{x \notin Y_k} [J(Y_k + x)]$ ;
8    $Y_k = Y_k + x^+$ ;
9    $k = k + 1$ ;
10 }
11
12 return  $Y_m$ 
```

⁹The number of possible combinations for 35 features is 2^{35} or roughly 34 billion.

Listing 4.1 shows the pseudocode for the SFS feature selection algorithm. During the first iteration, SFS considers one feature at a time and computes the corresponding value of the performance metric J . The feature that allows to achieve the highest value of J is retained and included in the selected feature subset. At each successive iteration, the feature that – added to the already selected ones – allows to obtain the highest improvement of the performance metric J is added to the already selected feature subset. The feature selection procedure may be stopped if all features have been evaluated or a maximum number of features has been reached. In this study, we do not limit the size of the feature subset (*i.e.* we allow it to contain as much as 35 features) and we use the occupancy detection accuracy (*cf.* Section 4.3.1) as the performance metric.

Principal component analysis

Some of the 35 features defined in Table 4.1 are closely related. A change in the max and min features, for example, directly influences the value of the range feature. This makes it difficult for a feature selection algorithm such as SFS to choose the best features. In fact, a combination of different features may be more descriptive of the data. principal component analysis (PCA) solves this problem by transforming the original data into a set of linear combinations of the original features.

Principal component analysis is an orthogonal transformation of a set of observations \mathbf{X} of dimensionality D (*e.g.* $D = 35$ in our case) to a set of linearly uncorrelated variables (*i.e.* *principal components* \mathbf{T}) in a new coordinate system (*cf.* Equation 4.1) such that the variance of the projected data is maximised [77]. Thus, the first component \mathbf{T}_1 explains the largest possible variance. The transformation is then defined recursively, such that the next principal component accounts for both the largest variance in the input data while being orthogonal to the preceding components. This transformation is lossless and does not reduce the dimensionality of the input data.

$$\mathbf{T} = \mathbf{XW}. \quad (4.1)$$

In many cases, the first few components already explain most of the variance of the input data. By restricting the number of components to the first L , we significantly reduce the input data to the classifiers without sacrificing much information from the original feature set. To obtain a suitable L , we use the number of components which accounts for at least 95% of the variance of the input data.

4.3 Evaluation

Binary occupancy classification is a two-class problem. A household can either be occupied or unoccupied during a particular 15-minute slot. Table 4.2 shows the four

Table 4.2: Confusion matrix.

		Actual class (ground truth)		Total
		p (occupied)	n (unoccupied)	
Predicted class	p' (occupied)	True Positive	False Positive	$tp + fp$
	n' (unoccupied)	False Negative	True Negative	$fn + tn$
Total		$tp + fn$	$fp + tn$	N

possible outcomes of classifying such an slot. An instance of correctly assessing the household's occupancy from the electrical consumption data to be *occupied* is called a *true positive* (tp). Likewise, correctly labelling the data to be *unoccupied* is called a *true negative* classification (tn). *False positive* (fp) and *false negative* classifications (fn) denote the instances of incorrectly labelling the household *occupied* or *unoccupied*, respectively. Using this notation, we introduce several different established performance criteria in the remainder of this section.

4.3.1 Accuracy

The *classification accuracy* is the simplest performance measure [184]. The classification accuracy of a classifier c is computed as the number of correct classifications divided by the total number of classifications:

$$\text{Acc}_c = \frac{tp + tn}{tp + tn + fp + fn} \quad (4.2)$$

We use *Prior* – a maximum-likelihood classifier that always assigns an input data to the class of the majority of data points in the training set – as a baseline for our comparisons. Since in our dataset the households are occupied more than 50% of the time, *Prior* always classifies the households as occupied. We use the accuracy of the *Prior* classifier as a baseline for the other methods.

However, the classification accuracy only partially describes the performance of a classifier. It does not take into account the relative costs of making a wrong classification. A false negative classification by a smart thermostat (*e.g.* occupants are assumed to be away although they are, in fact, present) usually results in an automatic reduction of the temperature. On a cold day, this may severely impact the occupants' comfort. Witten *et al.* summarise this problem by saying that an “evaluation by classification accuracy tacitly assumes equal error costs” [184].

In addition, the classification accuracy may be misleading if the distribution of classes is unbalanced. In our case, households r4 and r5 have occupancy figures exceeding 90%. Thus, a high accuracy may be achieved by always predicting these household to be occupied.

Table 4.3: Confusion matrix for biased random classifier.

		Actual class (ground truth)		Total
		<i>occupied</i>	<i>unoccupied</i>	
Predicted class	<i>occupied</i>	$tp = p^2$	$fp = (p-1)p$	$tp + fp$
	<i>unoccupied</i>	$fn = (p-1)p$	$tn = (p-1)^2$	$fn + tn$
Total		$tp + fn$	$fp + tn$	N

4.3.2 Matthews Correlation Coefficient

In the case of occupancy detection, correctly classifying both occupied (true positive) and unoccupied (true negative) states is paramount. For this reason we computed the Matthews correlation coefficient (MCC) over the results of our classifiers [132]. For the MCC, a perfect prediction is represented by a coefficient of +1. A value of -1 indicates that no single instance was classified correctly. The MCC of a classifier c is calculated as:

$$\text{MCC}_c = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \quad (4.3)$$

The MCC provides a balanced measure even if the input data are heavily skewed towards one class. The MCC is undefined for the *Prior* maximum-likelihood classifier as both the numerator and denominator become zero.

4.3.3 False negative and false positive rate

Labelling a household as *unoccupied*, while it is in fact occupied is a *false negative* (fn) classification. A false negative causes a problem for heating control applications: If the household is falsely declared to be unoccupied, the thermostat lowers the temperature while occupants are still present. This results in discomfort for the inhabitants. To be able to quantify the number of such misclassifications we use the false negative rate (FNR). The false negative rate of a classifier c is defined as the number of false negatives divided by all unoccupied slots (true and false negatives).

$$\text{FNR}_c = \frac{fn}{fn + tn} \quad (4.4)$$

Analogously, we call labelling an unoccupied household as *occupied* during a particular slot a *false positive* (fp). A false positive classification means that a smart thermostat will raise the temperature unnecessarily, resulting in a loss of efficiency. We use the false positive rate (FPR) to denote the frequency of such occurrences. The false positive rate of a classifier c is defined as the number of false positives divided by all occupied slots (true and false positives).

$$\text{FPR}_c = \frac{fp}{fp + tp} \quad (4.5)$$

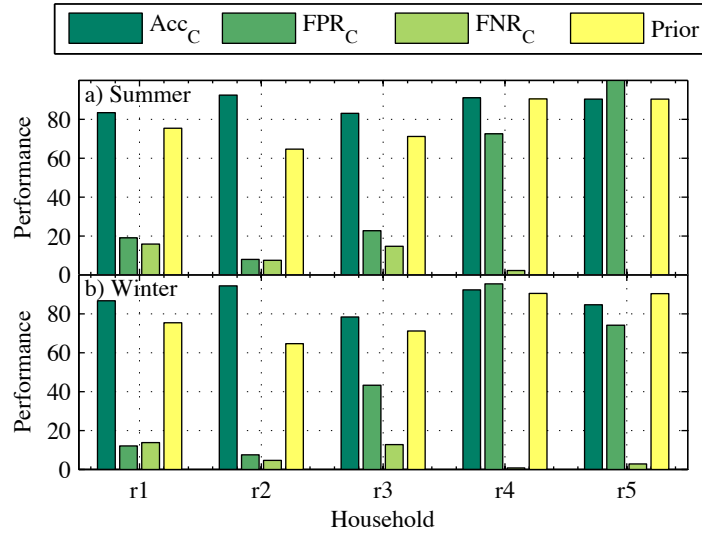


Figure 4.7: Maximum achievable accuracy Acc_C over all classifiers.

4.4 Results

In this section, we quantitatively evaluate the occupancy detection performance achievable using features derived from aggregated electricity consumption data. We consider the set of classifiers introduced in Section 4.2.3. These are SVM, KNN, THR, GMM, HMM and the *Prior* classifier for baseline comparison. The classifiers take as input the set of features computed of aggregated electricity consumption data introduced in Section 4.2.1 and output the estimated state of the household in a specific time slot. For dimensionality reduction, we evaluate the SFS feature selection algorithm and PCA. To indicate the used method, we will append “-SFS” or “-PCA” to the respective classifiers, where applicable (*e.g.* SVM-SFS denotes the usage of the SVM classifier trained using SFS feature selection).

We will first discuss the overall occupancy detection performance before discussing the results of the different classifiers and the results of the feature selection. We will conclude this section by analysing the suitability of these classifiers for occupancy detection in a smart heating scenario.

4.4.1 Overall occupancy detection performance

For each household, we define Acc_C as the *highest accuracy achievable by any classifier*, with C denoting the classifier achieving this accuracy. To put this value into context, we also report the false positive and false negative rates achieved by C . Thus, FPR_C is the false positive rate and FNR_C is the false negative rate of the classifier achieving the highest accuracy. Figure 4.7 shows the values for Acc_C , FPR_C and FNR_C for all five households in the (a) Summer and (b) Winter datasets.

During both summer and winter, for households r1 to r3, Acc_C is higher than the Prior accuracy as determined by the maximum-likelihood classifier. The classification shows the highest performance in household r2 where Acc_C is on average 29% higher than the Prior accuracy. With an accuracy over the whole year of 93%, just over one hour per day is misclassified by the best classifier on average. This is achieved with a low fraction of slots misclassified as unoccupied. The average FNR_C over summer and winter is 6%. Thus the classifier incorrectly assumes the house is unoccupied for an average of 38 minutes per day. The average FPR_C of 8% results in 26 minutes being incorrectly classified as occupied. We will look at the reasons for these misclassifications in Section 4.4.4.

For households r1 and r3 the best classifiers achieve $\text{Acc}_c = 85\%$ and $\text{Acc}_C = 81\%$, respectively. For both households, this is a ten percent improvement over the Prior accuracy. However, in contrast to r2, this comes at the potential cost of significant discomfort in terms of false negative rates. For r1, a false negative rate of 15% means that almost 110 minutes are incorrectly classified as unoccupied while the participants were actually at home. Similarly, in household r3, a false negative rate of 14% results in one hour and 35 minutes being misclassified as unoccupied.

Households r4 and r5 have a high average occupancy around 90%. For these households, the accuracy of none of the classifiers significantly exceeds the accuracy achieved by the Prior classifier. Since we lack detailed data on the behaviour of the occupants we cannot establish the exact reasons for this result. We assume, however, that an explanation may be found in the behaviour of the occupants. As r4 and r5 are almost always occupied, there will inevitably be periods of occupancy during which no electric appliances are used. When training the classifier, these periods look identical to those during which the house is actually unoccupied. Due to the high occupancy, the number of such *inactive* periods is also likely to exceed the unoccupied periods, resulting in the classifier almost always classifying the home as occupied.

In order to alleviate this problem, we changed the behaviour of the classifiers by undersampling the training data to obtain an even split of occupied and unoccupied slots. However, this merely increased the number slots misclassified as unoccupied and reduced the overall accuracy. As the main objective of a smart heating system should be to ensure the comfort of the occupants at all times, this is not feasible. After all, high occupancy households may not be interesting targets for smart thermostats in the first place as the amount of energy that can be saved is reduced as the total occupancy increases¹⁰. While we will list the results for households r4 and r5 in the remainder of this chapter for the sake of completeness, we will not include them in our analysis due to their limited suitability for the smart heating scenario.

¹⁰The connection between energy savings and occupancy will be discussed in Chapter 8.

Table 4.4: **Classification accuracy** (expressed as percentages) obtained for each household and algorithm in the two periods Summer and Winter.

	SFS			PCA				Prior
	SVM	KNN	THR	SVM	KNN	GMM	HMM	
House	Summer							
r1	80	76	77	83	80	78	83	75
r2	91	88	76	92	89	76	90	65
r3	78	76	71	83	79	70	82	71
r4	90	90	85	91	88	70	87	90
r5	90	88	81	90	84	59	79	90
	Winter							
r1	82	78	83	84	81	79	87	73
r2	93	91	77	94	91	88	92	63
r3	70	71	66	78	76	59	71	71
r4	92	92	90	92	90	70	84	93
r5	82	80	77	85	79	63	74	82

4.4.2 Performance by classifier

Table 4.4 shows the classification accuracy for all combinations of households and classifiers for the Summer and Winter datasets. The best classifier(s) are indicated in bold print. The table shows that in terms of classification accuracy, the SVM-PCA classifier outperforms the other classifiers in seven of ten cases. The SVM-PCA classifier achieves an average accuracy of 86% for households r1 to r3. In contrast to the simple occupancy detection algorithm introduced in Figure 4.1 at the beginning of Section 4.2, which classified any power consumption above a certain threshold as occupied, the more complex set of features introduced in this section requires a non-linear classifier¹¹. The SVM classifier is especially well-suited for this sort of classification and thus performs better than the KNN, THR, GMM and HMM classifiers.

The HMM classifier is the second best classifier, outperforming the SVM-PCA classifier for household r1 in winter and performing equally well in summer. The HMM classifier achieves an average accuracy of 84% for households r1 to r3. In contrast to the other classifiers, the HMM classifier does not merely rely on the features, but also takes into account the previous occupancy state. Furthermore, like the SVM classifier, it also allows for non-linear classification. However, as discussed in Section 4.2.3, the HMM classifier operates on a fixed subset of the input data, putting it at a disadvantage to the SVM classifier.

The GMM classifier performs worst, indicating that – while the logarithm of the overall

¹¹The max feature, for example, may assume either (1) a low value, indicating absence, (2) a medium value indicating the operation of a device like the television or (3) unrelated to the actual occupancy, a high value for an electric boiler. Likewise, the p_{time} feature, which assigns slot numbers from 1 to 65 to indicate the current time, has higher associated occupancy probabilities with the first (morning) and last (evening) slots than the ones in-between (lunchtime and afternoon).

Table 4.5: **Matthews correlation coefficient** obtained for each household and algorithm in the two periods Summer and Winter.

	SFS			PCA			
	SVM	KNN	THR	SVM	KNN	GMM	HMM
House	<i>Summer</i>						
r1	0.40	0.35	0.35	0.52	0.46	0.49	0.60
r2	0.81	0.73	0.45	0.84	0.76	0.55	0.79
r3	0.46	0.42	0.32	0.61	0.49	0.44	0.61
r4	0.14	0.15	0.19	0.35	0.35	0.32	0.45
r5	/	0	0.05	/	0.11	0.13	0.19
	<i>Winter</i>						
r1	0.50	0.42	0.55	0.58	0.53	0.55	0.70
r2	0.84	0.81	0.51	0.88	0.82	0.75	0.84
r3	0.18	0.21	0.14	0.46	0.41	0.20	0.32
r4	0.10	0.09	0.19	0.15	0.20	0.22	0.26
r5	0.11	0.24	0.07	0.35	0.32	0.25	0.31

power consumption may follow approximately a normal distribution – the derived features do not. Similarly, due to its reliance on the assumption that a higher absolute value of a feature always indicates presence, the simple THR-SFS classifier only achieves an average accuracy of 75%. While the KNN classifier performs better than the former two, it is outperformed by SVM and HMM on five of six cases for households r1 to r3.

In terms of classification accuracy, using principal component analysis to reduce the number of features prior to classification outperforms feature selection using the SFS algorithm. While SVM-SFS performs similarly to SVM-PCA for r2, it is outperformed by SVM-PCA and the PCA-based HMM in households r1 and r3. In Section 4.4.5 we will look at the selected features and discuss possible explanations of why PCA performs better than SFS on our data.

Table 4.5 confirms that PCA also outperforms SFS feature selection for the MCC (*i.e.* for all households, PCA-based approaches yield a higher MCC). However, if the MCC is taken as the measure of merit, the performance of the HMM and SVM-PCA classifiers converge. Both have an average MCC of 0.64 for households r1 to r3. While HMM now shows the highest performance for household r1 in summer and winter, SVM-PCA performs better or equally well in households r2 and r3. The more clear separation of HMM and SVM-PCA for r1 using the MCC is due to the fact that the MCC rewards a more even split between false positives and false negatives.

To further investigate this issue, Table 4.6 shows the FNR for all combinations of households and classifiers. The FNR is especially important when considering a heating system (*i.e.* a higher FNR produces an uncomfortable environment as the temperature is allowed to drop when occupants are present). Again, bold print indicates the best (lowest) values. The tables show that choosing the classifier according to the classification accuracy (or MCC) may not always be the best strategy for a heating scenario. In the previous

Table 4.6: **False negative rate** (expressed as percentages) obtained for each household and algorithm in the two periods Summer and Winter.

	SFS			PCA			
	SVM	KNN	THR	SVM	KNN	GMM	HMM
House	<i>Summer</i>						
r1	9	15	14	10	14	22	16
r2	8	10	11	7	9	30	9
r3	16	17	21	15	16	36	20
r4	1	2	9	2	7	31	11
r5	0	2	12	0	9	42	17
House	<i>Winter</i>						
r1	8	12	8	9	13	23	14
r2	6	7	15	5	7	15	9
r3	13	13	21	13	16	45	24
r4	1	1	5	1	4	30	14
r5	2	11	9	3	14	39	23

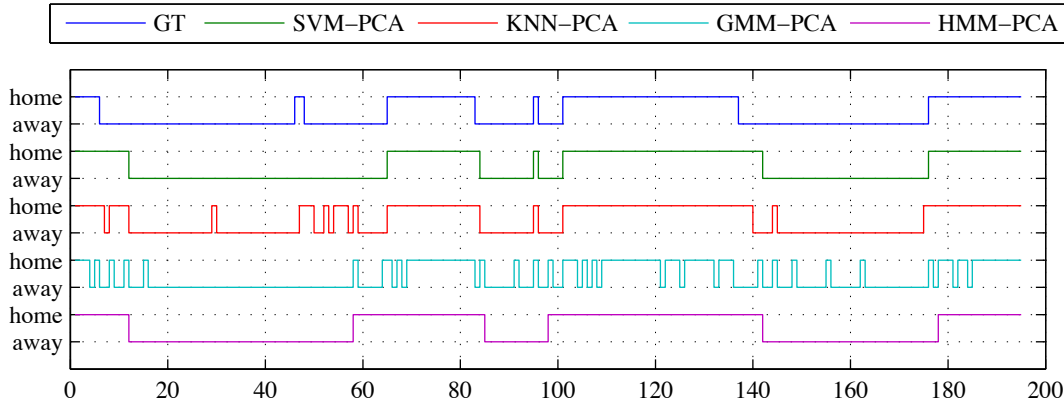


Figure 4.8: Ground truth (GT) and classification results for an exemplary classification of 195 slots (3 days) of household r2.

section, we noted that Acc_C was 85% for household r1. This was achieved by choosing the HMM classifier for both summer and winter periods¹². However, the choice of the HMM classifier results in an average FNR of 15% while choosing the SVM-PCA classifier would have resulted in a FNR of just 10% at the expense of an average 2% reduction in accuracy.

4.4.3 Suitability for controlling a thermostat

Both classification accuracy and MCC assess the performance of a classifier based on the correct classification of individual slots. Any correct or incorrect classification of a slot contributes with the same weight to the metric. The ability to detect *occupancy transitions*

¹²The accuracy for the HMM classifier is slightly higher than that for the SVM-PCA classifier which is not visible in Table 4.4 due to rounding errors.

Table 4.7: **Root mean square error (RMSE)** of the transitions predicted by the classifiers compared with the number of actual occupancy transitions per day and **average number of daily occupancy transitions (ADOT)** for each household and algorithm in the two periods Summer and Winter.

	SFS			PCA				
	SVM	KNN	THR	SVM	KNN	GMM	HMM	ADOT
House	Summer							
r1	11.7	9.3	11.5	7.4	6.2	9.2	8.7	2
r2	3.6	3.9	12.4	3.4	3.8	6.3	3.7	2.5
r3	10.1	7.2	9.9	7.8	6.0	11.1	10.4	2.3
r4	9.2	8.7	11.2	6.5	5.6	20.1	9.1	1.8
r5	12.1	11.1	9.0	12.1	7.4	22.9	11.7	1.3
	Winter							
r1	8.4	8.5	8.5	7.9	9.8	14.0	9.1	1.1
r2	2.6	2.9	6.3	2.5	2.8	3.9	3.0	2.2
r3	5.3	5.9	5.6	4.0	5.2	8.2	3.1	1.9
r4	9.3	9.1	7.1	8.4	5.7	26.9	15.2	1.3
r5	13.8	8.1	9.1	10.3	5.3	12.3	6.9	2.1

– *i.e.* changes in the occupancy state (from occupied to unoccupied and vice versa) – is however crucial to many systems. For instance, when a smart heating system detects that the household has become occupied, it may decide to start heating immediately. As every transition may lead the heating system to thus change its state, correctly identifying the number of transitions is of equal importance to the accuracy of the classification itself. Figure 4.8 shows the classification for the first 195 15-minute slots of household r2 using the SVM-PCA, KNN-PCA GMM-PCA and HMM-PCA classifiers. The ground truth occupancy data shows 10 state transitions. The SVM-PCA classifier misses a short period of occupancy on the first day but otherwise has the same number of transitions as the ground truth. Due to its statefulness, the HMM misses both short occupancy periods but also achieves results close to the ground truth in terms of the overall number of transitions. The KNN-PCA and GMM-PCA however perform significantly worse, inducing a large number of additional transitions in the occupancy state.

To further investigate the ability of the classifiers to detect the correct number of occupancy state transitions, Table 4.7 shows the RMSE between the number of actual occupancy transitions per day and the transitions predicted by the classifiers. For reference, the table includes the average number of daily occupancy transitions (ADOT). We define ADOT for each household and season as follows:

$$\text{ADOT} = \frac{\sum_{d=1}^{\text{Total number of days}} \text{Number of transitions for day } d}{\text{Total number of days}} \quad (4.6)$$

The table shows that household r2 has the highest ADOT of all households. This corresponds to the lowest occupancy (*e.g.* 64%) in the dataset. Again, the SVM-PCA

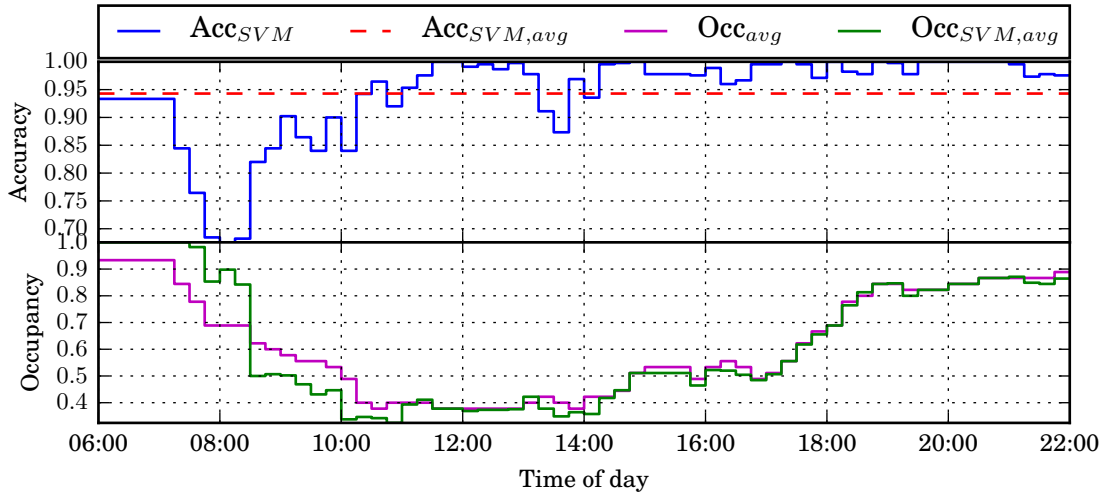


Figure 4.9: Mean accuracy over 24 hours (r2, SVM, winter).

classifier allows for the highest performance in household r2 with an average error of 3 transitions. For the other households, the best classifiers overestimate the number of transitions by 3 to 8. This means that additional smoothing in the heating controller is needed to avoid unnecessary switches of the heating system. The system could for example require to wait an hour before declaring the home to be unoccupied, thereby also reducing the FNR.

4.4.4 Limits to occupancy sensing using electricity consumption data

The accuracy, FNR and RMSE of the SVM-PCA classifier look promising for a heating application. However, Figure 4.9 shows that the classification accuracy still exhibits significant variations throughout the day. The figure shows the average classification accuracy of the SVM-PCA classifier (r2, winter) over the day from 6 a.m. to 10 p.m. The upper graph shows that while the accuracy stays around its average of 94% for most of the day (e.g. from 10.15 a.m. to 10 p.m.), there is a significant drop in the morning.

The lower graph shows the average ground truth (*i.e.* the actual average occupancy) and the average occupancy predicted by the SVM-PCA classifier. Up until 8.30 a.m., the occupancy is constantly overestimated – probably reflecting the fact that the participants are more likely to forgo their breakfast the earlier they leave the house. After 8.30 a.m., the situation turns and the actual occupancy is higher than the predicted occupancy resulting in a number of false positives. Possible explanations for this behaviour include sleeping in on weekends and a reduced use of electrical appliances in the morning. For the rest of the day, the predicted occupancy closely tracks the ground truth occupancy.

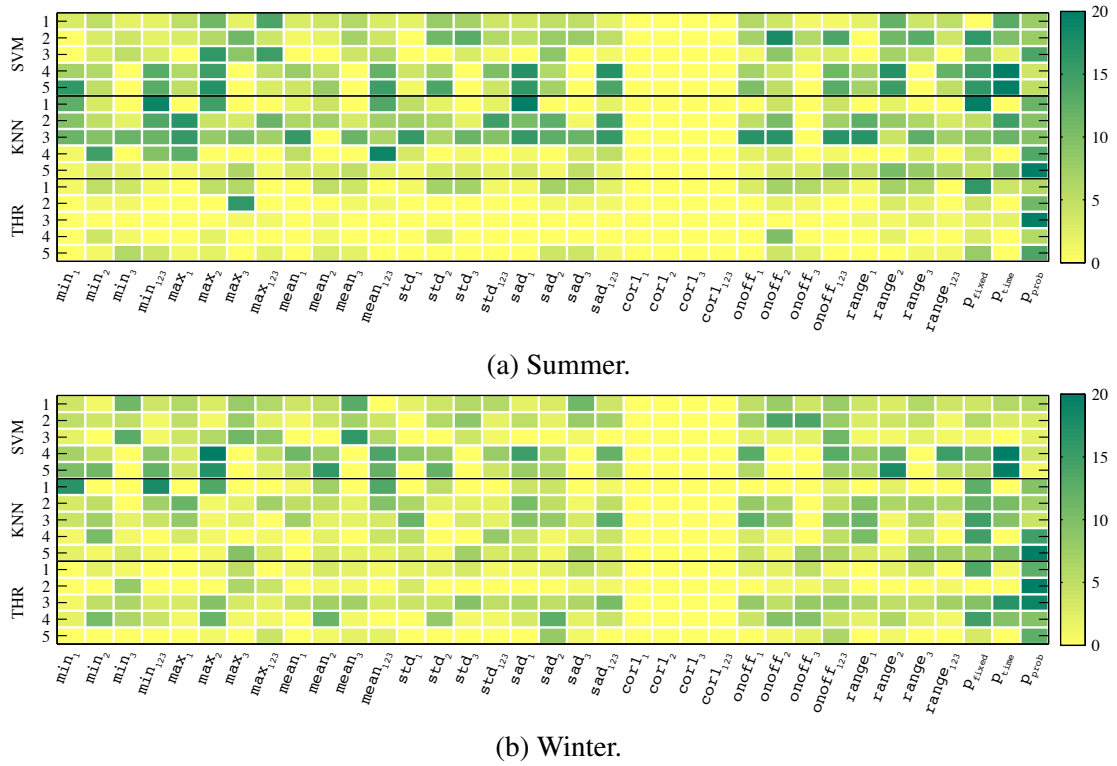


Figure 4.10: Number of times a specific feature has been chosen as part of the feature subset selected by the SFS algorithm for a particular household and classifier. A darker colour indicates a feature was chosen more frequently.

While this drop in accuracy in the morning does not negatively affect the comfort of the inhabitants, it suggests that in order to make sure that a smart heating system does not waste savings potential in the morning, additional instrumentation may be required.

4.4.5 Features selected by SFS

In this section we analyse which features of the electric load curve are most indicative of occupancy. For this purpose, we look at the features chosen by the SFS algorithm¹³. Figures 4.10a and 4.10b display – for the Summer and Winter periods, respectively – the number of times a specific feature has been chosen as part of the feature subset selected by the SFS algorithm. All features are listed on the x-axis. Each row in the plot shows the number of times each feature has been chosen for a specific classifier and household, as indicated on the y-axis.

These plots show that while there is a trend for the SFS feature selection algorithm to select more features during summer than winter, no feature is chosen consistently over all

¹³As the PCA transforms the features into different dimensions, it does not help us with analysing which features are actually well correlated with occupancy.

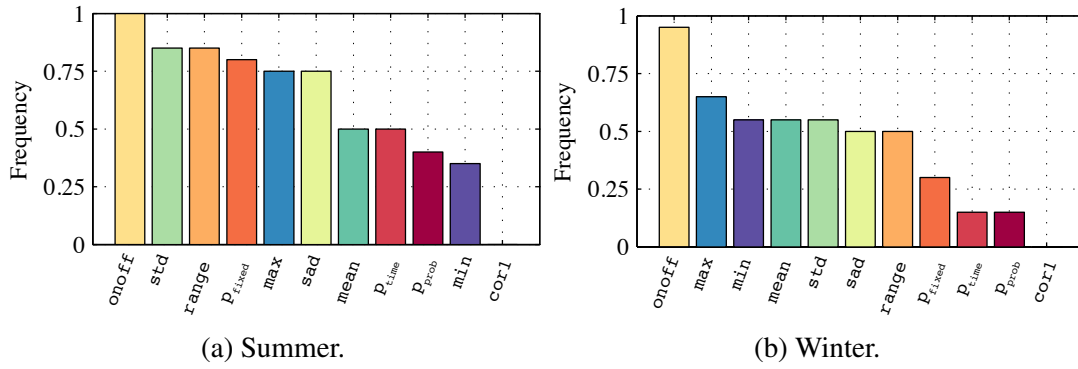


Figure 4.11: Combined features chosen by SFS for SVM classifier (household r2).

households. This is partly due to the fact that in different households, similar appliances may be attached to different phases (*e.g.* the \max_1 feature may detect the operation of a kettle on the first phase in household r1, while the kettle in household r2 is captured by \max_3). A specific feature computed on one of the phases might thus be very valuable for one household but not for the other and vice versa.

Figures 4.10a and 4.10b also show that the feature set obtained by the SFS feature selection is not stable. The feature selection results in different features being chosen on successive runs for the same household. The reason for this is that there is a high correlation between individual features. The range feature for example is composed from the difference of the min and max features. Likewise, the onoff feature is very similar to the sad feature. Whereas the former counts the number of instances of a specific delta (*i.e.* 30 watts over 30 seconds) in the power consumption, the latter aggregates all deltas of the power consumption (*i.e.* it computes the sum of the absolute differences between subsequent measurements). As the features' descriptive power is very similar, the selection by SFS is influenced by small variations in the accuracy that are more to do with the variance of the dataset than the descriptiveness of a particular feature. Incidentally, this is the reason why the PCA dimensionality reduction performs well. By selecting only the n first components comprising 95% of the variance of the training data, the overlap between features is ignored.

Figure 4.11 shows the cumulative probability of a particular feature irrespective of the phase it was computed on¹⁴ being chosen by SFS in the summer (Figure 4.11a) and winter (Figure 4.11b) datasets. As noted previously, the chosen feature set is larger during the summer than during the winter. During summer, the first six features are chosen in more than 75% of runs, while during the winter only the first feature exceeds a 75% probability of being chosen.

Overall, during summer and winter, the onoff feature is used most often. During

¹⁴For features that have been computed on more than one phase (*i.e.* min was computed for 1, 2, 3 and all 3 phases), the figure shows the probability of at least one of these features being used in a particular run.

summer it is used in all runs and in winter it is used in more than 90% of the runs. It is followed by the `max` and `min` features in winter, while in the summer the similar range feature comes in the third position. Furthermore, in winter, the time features – `pprob`, `pfixed` and `ptime` – are less likely to be chosen. This is because during the summer, the time features allow to establish a correlation between occupancy and the current time of day when the electricity consumption itself is not sufficient (*e.g.* apart from a short period of time to prepare breakfast, occupants are using less electricity in the morning). During the winter, the requirement for lighting whenever the house is occupied in the morning or evening makes these features less necessary.

The strong prevalence of the `onoff` feature confirms our initial expectation that by knowing the activation state of individual appliances we can derive the occupancy of the household. To investigate further how the actual activation state of appliances might improve the classification performance we investigate the usage of device-level consumption information for occupancy classification in the next section.

4.5 Using device-level consumption data

In light of the results presented thus far, a natural question to ask would be how much better we could do if we had the disaggregated (*i.e.* device-level) consumption information available. This information could be obtained by NILM algorithms (*cf.* Section 4.1.3) or additional instrumentation of the households.

Knowledge about the activation state of selected household appliances could allow for a more accurate assessment of the occupancy state. In particular, we expect that using information about household appliances that are well-correlated with occupancy and exhibit a small number of false positives (*e.g.* the TV or kettle) should help to increase the performance of the occupancy detection. In this section we will evaluate the occupancy detection accuracy obtainable from the device-level consumption data included in the ECO dataset and compare it to the accuracy obtained from the aggregate load curve.

4.5.1 Correlation between appliance state and occupancy

In order to identify appliances whose on/off states are indicative of occupancy, we will distinguish between three categories of devices. Devices which always exhibit a positive correlation with occupancy we will identify as *class A* devices. Such devices include televisions, kettles and stoves. These devices are usually only switched on when the household is actually occupied. Appliances where the correlation is uncertain – such as the dishwasher or the washing machine – we will identify as *class B* devices. In addition to being operated when the house is occupied, these appliances may be switched on when the occupants are leaving or programmed with a timer in order to run while the occupants

Table 4.8: Appliance-level on/off classification (Household r2, summer). FPR and FNR in percent.

	FPR	FNR	Correlation	Class
Appliance	<i>Summer</i>			
Tablet	1	98	0.00	C
Dishwasher	1	97	0.06	B
Vent	0	97	0.09	A
Fridge	61	31	0.08	C
Entertainment	1	38	0.52	A
Freezer	96	3	-0.01	C
Kettle	0	97	0.09	A
Light	0	96	0.11	A
Laptops	0	58	0.37	A

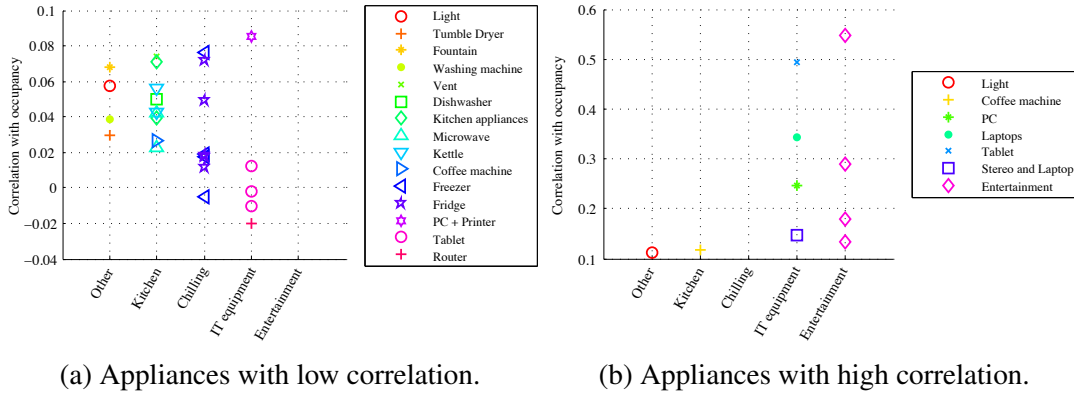


Figure 4.12: Correlation between appliance activation state and occupancy.

are away. Whether a class B device has a positive or negative correlation must be learned by an algorithm as it cannot be known in advance. *Class C* contains all the appliances that have a low correlation with occupancy. Such appliances include the fridge and the freezer, which exhibit relatively constant consumption patterns throughout the day.

Table 4.8 shows the classification of the appliances fitted with smart plugs in household r2. The tablet computer was used as an in-home display to facilitate the data collection and was always plugged into a charger. All devices apart from the freezer exhibit a positive correlation with the occupancy ground truth. Some class A devices such as the ventilation over the stove and the kettle show a weak correlation with occupancy since they are only seldomly operated. The entertainment system and laptops show a strong correlation with occupancy, however. Figure 4.12 shows the correlation between activation state and occupancy for all appliances in the dataset. It can be shown that the activation state of entertainment devices has the highest correlation with occupancy. Kitchen appliances such as the kettle or the coffee machine on the other hand are also good indicators of occupancy. However, they are only activated for a much shorter time.

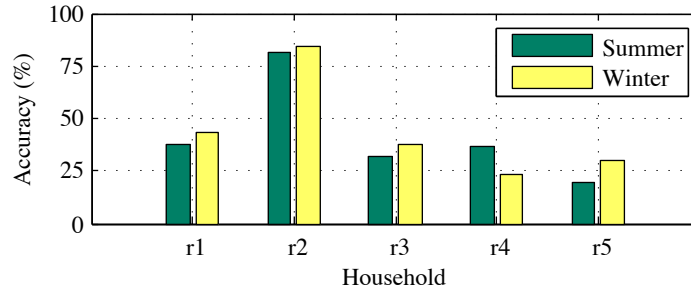


Figure 4.13: Accuracy for on/off only occupancy detection.

4.5.2 Detecting activation states

In order to compute the occupancy classification from the device-level information, we calculate – analogous to the classification of the aggregated data – for each day i and 15-minute slot j , the arithmetic mean of the electricity consumption $mean_{pid,i,j}$ as measured by smart plug pid . If $mean_{pid,i,j}$ is greater than 5 watts, we set the corresponding slot in the classification $C_{pid,i,j}$ to 1. Table 4.8 shows the results of the device-level classification in terms of false positive and false negative rates for r2 during the summer period. The classification of all type A devices, including the home entertainment system and the laptops, exhibits a low false positive rate. These appliances are mostly in use when occupants are present. We use thresholding to exclude all appliances with a false positive rate greater than 5%. While this also yields some class B and C devices, the overall effect of this is negligible as these must also exhibit a low number of false positives to be included. Note, that as the ground truth to establish the effect of including a certain appliance is not available, users may be asked to manually classify their appliances into the three categories.

In order to fuse the occupancy classification from the appliance activation states we take all classifications C_{pid} from appliances pid for which the false positive rate is greater than 5% and take their disjunction C_{onoff} .

4.5.3 Occupancy detection performance

Figure 4.13 shows the occupancy detection accuracy that may be achieved by solely using C_{onoff} , the on/off state of the selected appliances. Due to the low percentage of instrumentation in households r1, r3, r4 and r5, no clear assessment of the performance of device-level occupancy classification can be given, in general. In these cases, the accuracy is below 50% – the performance of a random guess. While this can be remedied by performing occupancy classification analogous to the rest of this chapter, the low correlation of the activation state of the instrumented appliances with occupancy makes this infeasible.

For household r2, however, the classification solely on the activation state of the entertainment and laptop appliances achieves a detection accuracy of 81% in summer and 84% in winter. However, as the lights and appliances such as the stove, microwave, hairdryer and vacuum cleaner are not instrumented, the accuracy does not approach the 92% and 94% obtained when using the aggregated load curve. The analysis of the device-level consumption information shows that the limits identified in Section 4.4.4 such as the tendency of occupants not to use appliances in the morning if they leave the house early cannot be remedied by device-level information.

4.6 A simple unsupervised approach: Revisiting the mean classifier

Up to this point, we have analysed how to design a classifier based on features computed from 1 Hz electricity consumption data and ground truth occupancy. In this section, we revisit the simple mean classifier introduced in Figure 4.1 at the beginning of Section 4.2 to show how well occupancy can be inferred – without having to acquire ground truth data for training – from data that is already available to utility companies today.

To this end we have implemented eight variants of a simple thresholding classifier. All eight variants (DAY-AVG, DAY-MIN, NIGHT-AVG, NIGHT-MIN, LOG-DAY-AVG, LOG-DAY-MIN, LOG-NIGHT-AVG, LOG-NIGHT-MIN) classify a household as occupied if its 15-minute aggregate electricity consumption exceeds a given threshold. To obtain suitable thresholds we computed the *average* and *minimum* of both the *daytime* and *nighttime* electricity consumption and their respective logarithms¹⁵.

Figure 4.14 shows the best accuracy Acc_C achieved by any of the eight classifiers introduced above. For household r2, the accuracy reaches 78%, exceeding the Prior accuracy on average by 14% meaning that the classification is wrong for approximately three and a half hours per day but correct otherwise.

All other households have values for Acc_C close to the Prior accuracy. In these cases, the algorithm almost always classifies the household as occupied. This can be seen most clearly in Table 4.9 for households r1, r3 and r5. In these cases, the best classifications are the ones relying on minimum of the respective consumption value. In contrast, the classifier performing best for household r2 is using the average nighttime consumption as a threshold to determine occupancy.

¹⁵Note, that in contrast to the $mean_{123}$ feature used previously, the 15-minute aggregate electricity consumption is not computed on the logarithm of the 1 Hz data as this data would not be available to the utility company. Instead, as the 15-minute consumption is still approximately log-normally distributed, we implemented LOG-DAY-AVG, LOG-DAY-MIN, LOG-NIGHT-AVG and LOG-NIGHT-MIN on the logarithm of the 15-minute data.

4.6 A simple unsupervised approach: Revisiting the mean classifier

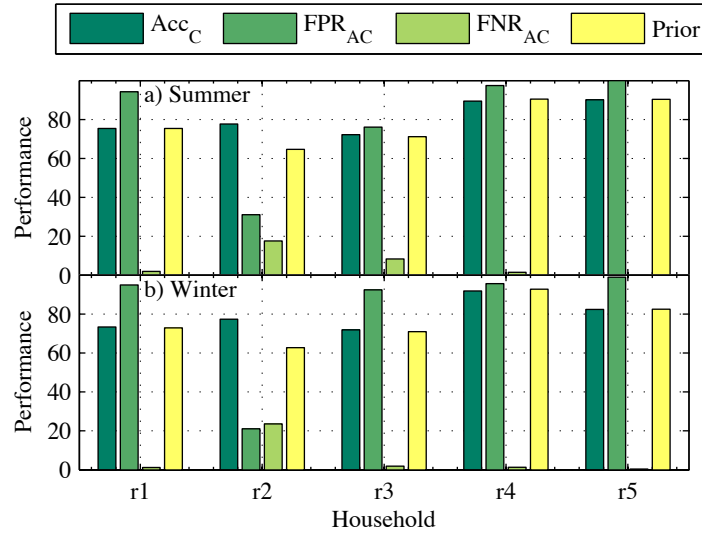


Figure 4.14: Best accuracy Acc_C among all eight simple unsupervised classifiers.

Table 4.9: **Classification accuracy** (expressed as percentages) obtained for each household and simple classifier in the two periods Summer and Winter.

	DAY		NIGHT		DAY (LOG)		NIGHT (LOG)		Prior
	AVG	MIN	AVG	MIN	AVG	MIN	AVG	MIN	
House	Summer								
r1	46	75	55	75	56	75	66	75	75
r2	61	65	77	66	68	65	78	66	65
r3	45	71	36	72	57	71	50	72	71
r4	40	89	16	86	49	89	31	86	90
r5	43	89	84	90	54	89	86	90	90
	Winter								
r1	48	73	61	73	58	73	70	73	73
r2	70	63	77	63	76	63	76	63	63
r3	48	71	35	72	58	71	58	72	71
r4	42	92	40	90	51	92	56	90	93
r5	44	82	25	82	52	82	51	82	82

The classification accuracy of these simple unsupervised approaches can certainly be improved. Such approaches may include clustering of the consumption data using k-Means or inferring the combined probability density function of the occupied and unoccupied states using GMMs.

The conclusion of this experiment is two-fold. In order to use smart electricity meters as opportunistic sensors to control a smart heating system, supervised methods and a sampling rate of 1 Hz are preferable. However, while being unsuitable as input for an automatic control system, the occupancy detection accuracy of these unsupervised approaches is good enough for them to be employed by a utility company to identify households with low occupancy. While, this certainly also warrants privacy concerns (*e.g.* this information may

reveal the employment status of the inhabitants), the information could be used for good by providing these households with information regarding smart heating systems [22].

4.7 Conclusions and lessons learned

In this chapter we addressed the problem of performing automatic home occupancy detection using aggregated electricity consumption data. Our results show that the use of smart electricity meters allows to achieve an average occupancy detection accuracy of up to 94%. We further showed that due to the non-linear relationship between the features derived from the electrical consumption and the occupancy, an algorithm capable of non-linear classification, such as an SVM classifier, is required to achieve the best performance.

Our analysis of the feature space highlights that features that correlate well with occupancy in one household may not be transferred directly to another. Some of the features are derived from the electricity consumption measured on a single phase. While this enhances the descriptiveness of our feature set (*e.g.* the stove can be easily detected as it is usually connected on multiple phases), this also means as different households may have similar appliances attached to different phases, a certain feature that is very descriptive in one household may be of little relevance in another. To select the most descriptive features and to thus reduce the dimensionality of the input data for the classifiers, we have found that using PCA outperforms the SFS feature selection algorithm.

Although our sample is too small to draw any conclusions yet, our results also indicate that occupancy detection from electricity consumption data work best for small households of young professionals with low occupancy. Such households have the ideal occupancy pattern for the intended use case – automating a thermostat. Households with high occupancy levels on the other hand not only do not offer a high potential for energy, but are also inherently difficult to classify as the high imbalance in training information leads to a bias towards occupancy.

Further improvements to the occupancy detection performance may be expected from sensing the device-level consumption information (*i.e.* how much power each appliance draws at any point in time). While a complete instrumentation of the household (*i.e.* sensing the consumption of all appliances) is desirable, in practise, information about long-running appliances such as entertainment systems, lights and computers should be preferred to transient consumers such as kitchen appliances. However, our results show device-level consumption information does not provide significant benefits over sensing occupancy from the aggregated consumption, only. One reason for this might be that there are times when the occupants are at home but the utilisation of electrical appliances is minimal (*e.g.* in the morning before and after the preparation of breakfast).

We began this chapter by noting how smart electricity meters are increasingly being deployed in residential households. In fact, due to privacy concerns and disappointing results (*i.e.* low energy savings between 3% and 7% [31, 54, 154]) from early field trials, their adoption has somewhat stalled. In Europe, the adoption according to 2009/72/EC is subject to a cost-benefit analysis [155]. Subsequently, several countries such as Germany have negatively assessed the impact of smart meters and will not reach the targets of the Commission [50]. One of the goals of the Third Package is to promote the “development of energy services based on data from smart meters” [50] – utilising smart electricity meters as opportunistic occupancy sensors is such a service.

Building large-scale occupancy datasets from unlabelled sensor data

A number of studies have shown how behavioural patterns of both groups and individuals can be discovered by analysing data collected using off-the-shelf mobile devices [34, 38, 59]. In particular, mobile phones have been used to gather mobility traces of individuals [63, 97]. The analysis of these traces enables identification of places of interest in the daily lives of individuals [97] or even the prediction of the places that will be most likely visited next by the mobile phone owners [168].

The use of mobile phones for the collection of mobility traces thus makes it possible to explore, model and predict human behaviour. Retrieving mobility traces at a fine temporal and spatial scale, however, may consume a significant amount of resources. For instance, the continuous operation of a GPS sensor is known to shorten battery lifetime of mobile phones significantly [35]. In practical settings, the use of GPS is thus typically “rationed” and combined with other technologies, in particular cell- or Wi-Fi-based localisation [117]. This, however, also requires reliance on third-party services and might thus raise privacy concerns. To reduce the impact of these issues, collecting data at a much coarser scale might still be sufficient to support a large set of applications and at the same time preserve mobile phone resources and protect users’ privacy. Such scenarios include applications that rely on knowledge about when households’ occupants are likely to return home, like home automation applications (*e.g.* automatic heating control), location-based reminders or notification services to ensure the presence of children at home.

In this chapter, we focus on this specific class of applications and introduce the *homeset* algorithm, a simple approach to estimate occupancy schedules from unlabelled sensor data. The algorithm relies on Wi-Fi scan data (*i.e.* the information that mobile phones gather about visible Wi-Fi access points (APs)) to determine when residents are at home and when not. We validate our approach using the Mobile data challenge (MDC) dataset from the Nokia Lausanne Data Collection Campaign that contains mobile phone traces of 38

participants collected over more than one year. Since the data is unlabelled, we indirectly validate our results by leveraging the information hidden in the anonymised GPS traces contained in the dataset.

While the initialisation phase of the homeset algorithm can be used for a more seamless setup of home automation systems, its main focus lies on the analysis of large-scale, long-term location datasets. The occupancy schedules thus derived can then be used to learn more about the behaviour of occupants and to improve existing occupancy prediction algorithms (*cf.* Chapter 6).

Before presenting the homeset algorithm and discussing its performance in Section 5.2, we summarise related work in Section 5.1. Section 5.4 concludes the chapter. This chapter has been based on contributions made in [100, 101].

5.1 Related work

The idea of using mobile phones to discover human mobility patterns has been explored extensively in the last few years [63]. Several authors have focused on identifying places of interest (*e.g.* the workplace, home or gym) and on predicting transitions between such places [12, 97]. Our work is related to these approaches since we aim to identify – although not locate – the home of a mobile phone user in order to build an occupancy schedule that could be used to control a thermostat.

5.1.1 Significant place sensing

Several authors have extended the problem space from occupancy sensing to the recognition of users’ *relevant places* [12, 73, 97]. Most humans spend a large portion of their day indoors at home, making it the *most relevant place*. Sensing relevant places using a portable sensor such as a mobile phone may thus identify when the users are at home as well as the places visited prior to returning home – information that can be used for predicting occupancy (*cf.* Section 6). A broad categorisation of place sensing algorithms classifies current approaches into *fingerprint-based* and *geometry-based* algorithms [138].

Fingerprint-based algorithms

A fingerprint-based algorithm periodically samples the radio-frequency (RF) signals from stationary sources such as mobile phone towers and Wi-Fi access points [44, 97, 138]. If a certain set of cell towers or Wi-Fi access points is visible over an extended period of time, a so-called “*stable radio environment*” [138] is detected. Such a stable radio

environment implies that the user is currently staying in a place¹. Following the premise that the relevance of a place is correlated with the users' duration of stay, such algorithms identify the relevant places of users.

However, most of these algorithms require sampling rates which are not available in public datasets. The PlaceSense algorithm by Kim *et al.* [97] samples the radio at 0.1 Hz to find semantic places. Such a sampling rate results in short battery life-times that would not be suitable for a long term deployment. As PlaceSense requires a stable scan window (*i.e.* a consistent set of beacons) to detect entrance to a place, it struggles on datasets with relatively low sampling rates such as the MDC dataset used in this chapter. In contrast, our homeset algorithm requires collecting only coarse-grained traces of Wi-Fi scans and can operate locally on the user's phone.

Geometry-based algorithms

As they are agnostic of the actual geographic position of the relevant places, fingerprint-based algorithms cannot directly provide *positioning capabilities* (*i.e.* provisioning of addresses or latitude/longitude coordinates). Moreover, they become inaccurate as the network topology changes and RF transmitters are added or removed. *Geometry-based* approaches, which are based on clustering of latitude/longitude coordinates can provide actual position data for the relevant places [12, 91, 187].

5.1.2 Public occupancy and location datasets

In the domain of ubiquitous computing, several authors have addressed the problem of predicting the next location of a person. In [168], Scellato *et al.* address the problem of estimating the arrival time of a user at specific location as well as “*the interval of time spent in that location*”. In this context, Petzold presented the *Augsburg Indoor Location Tracking Benchmarks* [159]. The dataset covers the rooms of one floor of a university building. The location traces were obtained manually by four participants using a graphical user interface on a personal digital assistant (PDA) computer. The collected traces were collected during two periods – summer and fall and vary in length from one to seven weeks.

In 2005, McNett *et al.* published a dataset containing mobility data of 275 PDA users over 11 weeks [133]. An application on the PDA sampled the available access points as well as the current association of the Wi-Fi module every 20 seconds. The dataset does not contain ground truth on the actual position of the users and therefore relies on coarse localisation using the (known) positions of the access points.

¹Note that an unstable radio environment may not provide evidence of the user's movements. Mobile cells and Wi-Fi networks may “breathe”, resulting in an unstable radio environment even though the receiver is staying in home place.

Existing approaches to discover occupancy schedules often rely on the availability of data from a GPS logger (*e.g.* standalone or embedded into mobile phones) to compute distance from home or ad-hoc sensors (*e.g.* passive infrared sensors) installed in the home [113, 169]. While the installation of ad-hoc sensors poses an additional burden in terms of costs and maintenance effort, the continuous operation of a GPS module is typically avoided due to energy constraints [117]. Thus GPS data is often replaced by, or combined with, information gathered through Wi-Fi- or GSM-based localisation services [97, 117]. Figure 5.3 shows a comparison of GPS based presence detection with our homeset algorithm. Related work [113] has put a user as *home* if she was within a 100 m radius of her home. We therefore argue that being within the coverage area of a Wi-Fi network is sufficient to detect occupancy at a much lower energy cost.

5.1.3 CDR datasets

Mobile phone operators collect data whenever a call has been established or a text message was sent. Such call detail records (CDRs) provide coarse-grained location information for each connection (*e.g.* the location of the cell tower) and can be used to create long-term mobility traces of individual cell phone users. The pervasiveness of mobile phones² enables the use of this data for the analysis of the behaviour of large populations of users. Song *et al.* for example study the predictability of human mobility using a CDR dataset of 50,000 mobile phone users gathered over three months [173]. In a separate paper, Song *et al.* also show that continuous-time random-walk (CTRW) models for human mobility, which are currently employed in a wide range of scenarios from epidemic modelling to traffic prediction, are in conflict with empirical data obtained from CDRs [172].

The increasing appreciation of the value of CDRs for the analysis of human mobility and interaction patterns [4] has led to the release of a number of open and commercial CDR datasets [27, 29]. Released as part of research competition by the mobile phone operator Orange, the data for development (D4D) dataset contains anonymised CDRs of five million mobile phone users in Ivory Coast between December, 2011 and April, 2012 [27]. In [120], Lima *et al.* have extracted mobility data from this dataset and study the spread of diseases and potential counter-measures. Berlingerio *et al.* use the data for developing data-driven ideas to improve public transit systems [25].

The mobile phone operator Telefónica offers a commercial CDRs for the analysis of crowd behaviour [197]. While this dataset is mostly aimed at helping “companies and public sector organizations [to] make informed business decisions”, it has also been used in research. Bogomolov *et al.* use Telefónica’s³ CDRs in conjunction with demographic information to predict crime hotspots in London with high accuracy [29].

²In 2014, the number of mobile phones in the world surpassed the number of people [30].

³The data was gathered by Telefónica’s subsidiary O₂ in the United Kingdom.

5.1.4 The Nokia LDCC/MDC dataset

None of the datasets discussed in the previous section contains fine-grained location data⁴ gathered in a residential environment for a significant amount of time. In order to obtain representative occupancy schedules, however, we need to ensure that the occupancy sensing device is no longer noticed by the occupants and thus does not in itself affect their behaviour.

For this reason we utilise Wi-Fi scans gathered in the context of the Lausanne data collection campaign (LDCC). This campaign was launched in 2009 by the Nokia Research Center in Lausanne, Switzerland and a subset of the collected data was released as part of the Mobile data challenge (MDC) in 2012 [117]. To gather the dataset, almost 200 participants were given Nokia N95 mobile phones with a data collection software installed. The dataset contains more than one year worth of traces of Wi-Fi scans, GPS coordinates, accelerometer readings and several other sensors, as well as demographic information for 38 of the mobile phone users that participated in the data collection campaign⁵.

Since its release, a number of authors have analysed the MDC dataset. Montoliu *et al.* showed that participants spent about two-thirds of their day at home and a quarter at work [138]. Using the PlaceSense algorithm [97], Baumann *et al.* have identified the distribution of relevant places in the MDC dataset [19]. They conclude that the participants spend 56% of their time in the most visited location, on average. Domenico *et al.* show how correlations between the trajectories of friends can be used to improve location prediction [38]. Frank *et al.* present an approach to produce English-language narratives of the events underlying the measured sensor data [59].

5.2 The Homeset Algorithm

The MDC dataset does not contain information about places of interest of the LDCC participants (*e.g.* it does neither contains ground truth occupancy schedules nor location labels). Therefore, we have no explicit ground truth about where the “home” of the participants is. This means spatial proximity to specific Wi-Fi access points or geographic locations cannot easily be associated with the home and recognised as occupancy. To overcome this issues, we propose the *homeset algorithm*.

⁴Here we mean fine-grained both in a spatial and temporal sense. In order to build exact occupancy schedules, we need to know precisely *when* the occupants are *close* (*e.g.* less than 100 m) to their home. This information cannot be obtained from traditional CDR data alone as it merely contains the location of the current cellular tower and is only gathered whenever a call is made or an SMS sent.

⁵The full LDCC dataset was subsequently released by Nokia. In contrast to the MDC subset, the full LDCC dataset does not contain demographic information. In this chapter we therefore evaluate the homeset algorithm on the MDC dataset. In this chapter, the participants are identified by the identifier used in the MDC dataset (*i.e.* participant “007”).

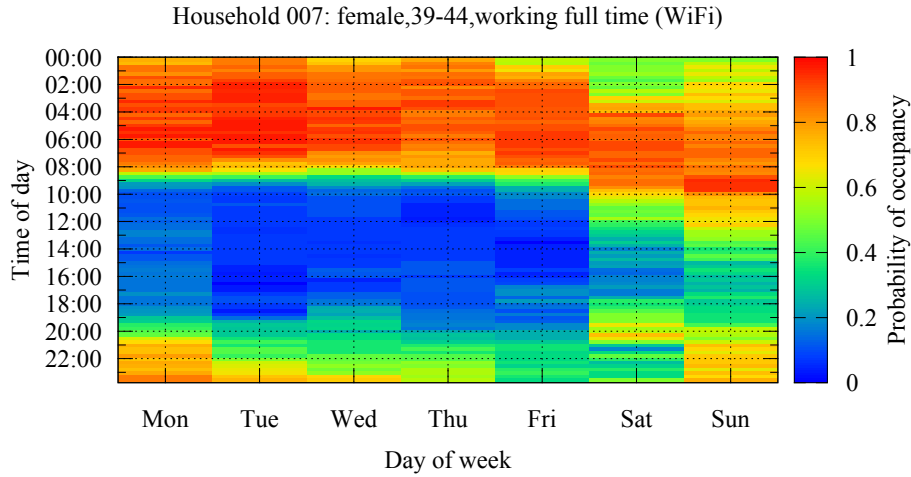


Figure 5.1: Probabilistic Wi-Fi occupancy schedule for participant 007. The participant is most likely to be away from home on weekdays between 8 a.m. and 7 p.m.

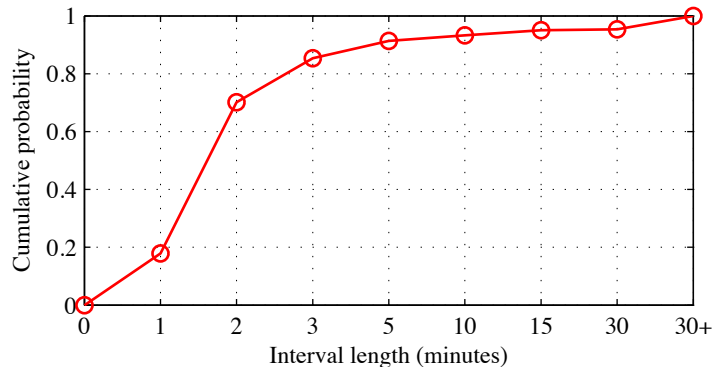


Figure 5.2: Cumulative probabilities for intervals from 1 minute to infinity across all participants.

In contrast to previous approaches such as PlaceSense [97], the homeset algorithm requires scans only to be performed at a coarse temporal scale (*e.g.* every 15 minutes). By performing a time-based clustering of these traces, our algorithm can accurately reconstruct the occupancy schedule of each household’s occupant. Figure 5.1 shows the probabilistic occupancy schedule derived for participant 007 of the MDC dataset. This schedule reveals that participant 007 is usually away from home between 8 a.m. and 7 p.m. during weekdays, while her behaviour is far less regular on the weekends.

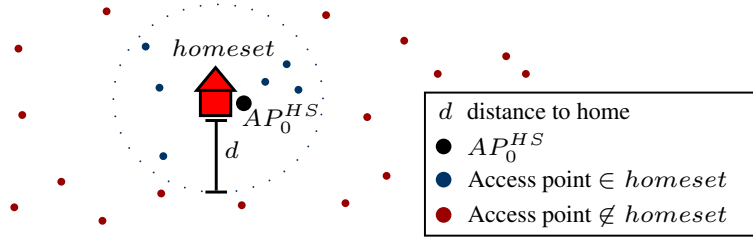


Figure 5.3: The homeset is the set of access points in the vicinity of the home access point AP_0^{HS} .

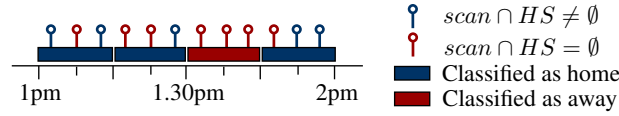


Figure 5.4: Interval classification based on multiple scans and homeset.

5.2.1 Occupancy detection using the homeset algorithm

The goal of the homeset algorithm is to compute the occupancy schedule of the residence of a mobile phone owner. A schedule is represented as a matrix O with N_s columns and N_d rows. N_s is the number of temporal *slots* within a day. N_s can be set to an arbitrary value, depending on the desired time granularity of the schedules. Figure 5.2 shows that in the MDC dataset, the interval between consecutive Wi-Fi scans is less than 15 minutes in 95% of the cases. In the context of this work we thus consider slots of 15 minutes, such that $N_s = 24 \times 60 / 15 = 96$. N_d is the number of days contained in the dataset.

To compute the occupancy schedules, the homeset algorithm relies on logs of Wi-Fi scans. Each time a mobile phone detects the presence of a Wi-Fi access point (AP) it stores several pieces of information. Among these, the homeset algorithm only uses the timestamp of the scan and the MAC addresses of the visible access points. A single Wi-Fi scan is thus a tuple $\langle ts, AP_0, AP_1, \dots, AP_{m-1} \rangle$ where m is the total number of access points seen in a particular scan and AP_i is the MAC address that uniquely identifies a specific access point.

Figure 5.3 shows the homeset algorithm uses these scans to identify a set of access points that are located within, or in the immediate proximity of, the household of a mobile phone owner. We call this set the *homeset* (HS) and assume it contains n access points, so that $HS = \{AP_0^{HS}, AP_1^{HS}, \dots, AP_{n-1}^{HS}\}$. We will for now assume $n > 1$ and discuss the initialisation of the HS below.

Figure 5.4 shows occupancy classification with the homeset algorithm. Given a Wi-Fi scan $\langle ts, AP_0, AP_1, \dots, AP_{m-1} \rangle$ the homeset algorithm tests if one of the sensed access points is contained in the homeset:

$$\{AP_0, AP_1, AP_2, \dots, AP_{m-1}\} \cap HS \neq \emptyset \quad (5.1)$$

If this statement returns true, the algorithm assumes the household to be occupied in the slot i of day j identified by the timestamp of the scan.

5.2.2 Initialisation of the homeset

In order to initialise the homeset in real home automation scenario, one could require the user to manually enter the MAC address of the household's private access point, if one exists or to actively scan for nearby access points while at home. As this information is not available in the MDC dataset and to eliminate this manual effort in initialising the homeset, we base the initialisation of the homeset on the assumption that occupants are most likely to be at home at night.

To find the nightly distribution of access points, we computed the empirical probability ω_x of seeing an access point x at least once between 3 a.m. and 4 a.m. on any particular night from the available MDC data. The access point with the highest value for ω_x is set to be AP_0^{HS} . Once AP_0^{HS} has been identified, the homeset is constructed by including in HS any other access point that appears in a Wi-Fi scan together with AP_0^{HS} . This approach significantly increases the reliability of the homeset algorithm.

Reliability of the occupancy detection

To measure this increase in reliability we define a metric called *stability*. We compute the stability π_x of an access point x over a time interval T_π , which we set to be at night between 3 a.m. and 4 a.m. If AP_0^{HS} is seen at least once within T_π , then it is reasonable to assume that the household must be occupied during the whole period. Indeed, although theoretically possible, it is unlikely that typical household occupants will leave the home between 3 a.m. and 4 a.m. However, in some scans registered in the period T_π AP_0^{HS} does not appear. If the homeset algorithm relied on AP_0^{HS} only, the household would be declared as occupied in given slots within the period T_π and unoccupied in others. This instability would clearly cause false negatives to appear and thus decrease the reliability of our algorithm.

To demonstrate that the homeset approach significantly improves on this aspect, we thus compute the stability π_x as the ratio of two quantities. The numerator is the total number of scans in which the access point x appears in the period T_π . The denominator is the total number of scans in the period T_π , whereby the scans are counted only if the access point x is seen at least once in the period T_π . A value of π_x equal to 1 thus means that if the access point is seen on any given night, it is going to be seen in all other scans between 3 a.m. and 4 a.m. and thus that it is a stable indicator of household occupancy.

Table 5.1: Empirical probability ω and stability π of the primary access point A_0^{HS} only and the extended set of access points, the homeset (HS), for all participants included in the dataset (n.a.: not available, ?: score too low).

ID	$\pi_{A_0^{\text{HS}}}$	$\omega_{A_0^{\text{HS}}}$	π_{HS}	ω_{HS}	Score	In HS?
002	0.555	0.953	0.963	1.0	13	✓
005	0.229	0.424	0.414	0.909	2	?
007	0.859	0.654	0.892	0.946	16	✓
009	0.477	0.78	0.954	0.962	n.a.	n.a.
010	0.75	0.678	0.852	0.956	2	?
017	0.805	0.978	0.981	0.985	8	?
023	0.715	0.487	0.987	1.0	16	✓
026	0.588	0.875	0.963	0.971	3	?
034	0.883	0.481	0.866	0.57	16	✓
042	0.674	0.678	0.964	0.931	10	✓
050	0.668	0.948	0.979	1.0	10	✓
051	0.516	0.982	0.985	1.0	2	?
056	0.943	0.975	0.985	1.0	6	?
060	0.921	0.977	0.983	0.996	12	✓
063	0.851	1.0	0.995	1.0	5	?
068	0.857	0.912	0.998	1.0	5	?
075	0.634	0.481	0.892	0.659	16	✓
077	0.76	0.875	0.961	0.938	n.a.	n.a.
082	0.899	0.968	0.992	0.989	16	✓
083	0.664	0.988	0.998	1.0	16	✓
089	0.512	0.643	0.971	0.857	5	?
094	0.794	0.584	0.832	0.887	n.a.	n.a.
109	0.468	0.884	0.94	0.977	9	?
111	0.676	0.462	0.975	1.0	11	✓
117	0.375	0.825	0.964	0.997	15	✓
120	0.77	0.72	0.96	0.805	13	✓
123	0.813	1.0	0.994	1.0	5	?
126	0.546	0.841	0.957	0.955	2	?
127	0.704	0.689	0.949	0.974	16	✓
139	0.472	0.391	0.916	0.457	n.a.	n.a.
141	0.314	0.839	0.985	0.873	7	?
160	0.538	0.876	0.984	1.0	16	✓
165	0.615	0.968	0.962	0.992	n.a.	n.a.
169	0.692	0.736	0.98	1.0	14	✓
172	0.476	0.915	0.958	0.954	15	✓
179	0.812	0.368	0.971	0.974	16	✓
185	0.448	0.696	0.972	0.983	10	✓
186	0.914	1.0	0.964	1.0	16	✓

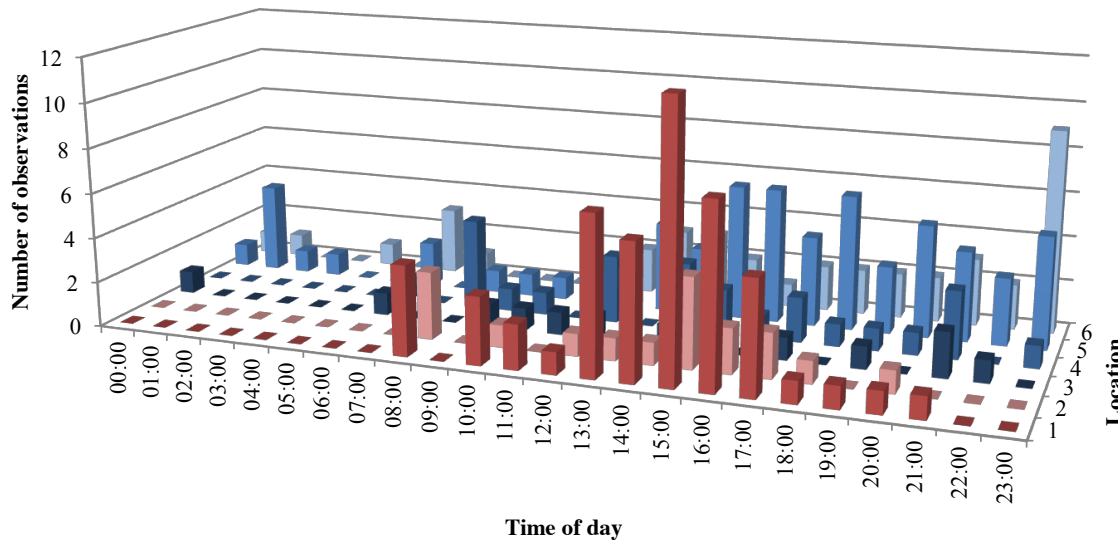


Figure 5.5: Time-frequency analysis of the anonymised locations for participant 002. Locations with less than 10 observations are excluded.

The rationale behind the homeset algorithm is that a set of access points has a higher stability than a single one, even if this one is the private access point of the household. Table 5.1 shows evidence of this observation for selected participants included in the MDC dataset. For instance, for participant 009, using the whole HS instead of the single primary access point only, increased stability from 0.477 to 0.954.

5.3 Evaluation

In order to thoroughly evaluate the homeset algorithm, a precise schedule of the absence from or presence in, the household of the mobile phone owners would be necessary. As this information is not available in the MDC dataset, we set out to evaluate our findings indirectly by verifying whether the access points included in the homeset are plausibly close to the location of the participants' homes. To this end, we used the GPS data available in the MDC dataset and considered the fact this data had been partially modified in order to protect the privacy of the participants. In particular, the latitude and longitude coordinates of sensitive places, like the participants' homes or workplaces, have been occasionally truncated to the third decimal digit. As the coordinates are reported along with a timestamp, we could retrieve statistics about *when* participants were in sensitive places, even though it was not possible to retrieve *where* exactly the participants were at that specific time.

Identifying sensitive places

We thus first extract all the truncated instances of the GPS data from the dataset. We then assign each unique pair of truncated latitude and longitude coordinates to a symbolic location k . For each location, we then create a frequency count vector $\vec{CV}_k = (c_0, c_1, \dots, c_{23})$ with 24 elements, one for each hour of the day. Over the whole dataset, we then count the number of occurrences of a location k in a given hour of the day and store this value in the corresponding element of the vector CV_k . We thus count how many times a specific symbolic location has been “anonymised”.

Figure 5.5 shows the results of this analysis for participant 002, whereby we only display the six most relevant symbolic locations. As visible in this picture, location 1 is anonymised most of the times between 1 p.m. and 5 p.m. and is never anonymised before 8 a.m. or after 9 p.m. We thus conjecture that this location corresponds to the workplace of the participant, as it is likely that between 1 p.m. and 5 p.m. the participant is at work and thus there is a higher need to truncate coordinates that correspond to this sensitive location. On the other side, location 5 is the one that is anonymised most frequently and consistently over the whole course of the day. Therefore, we conjecture that this is the location of the home of the participant.

Automatically validating the homeset

In order to automatically assess if a particular set of coordinates can identify a home location, we compute a score for each location. To make results comparable, we round CV_k to binary values and multiply it with a weighting vector $\vec{w} = (w_0, w_1, \dots, w_{23})$. Times between 9 and 17 (*i.e.* w_9 to w_{17}) are set to $\frac{2}{7}$ while all other times are set to 1. We chose this weighting assuming a normal nine to five schedule with little presence during the day except on weekends. A set of coordinates can score a maximum of 18.3 points under this metric. We have chosen a threshold of 10 for a location to be accepted as a possible home location.

Once we retrieved the (truncated and thus anonymised) location of the home of each participant using the method described above, we compare the symbolic location with the GPS coordinates of the Wi-Fi access points. To this end, we compute the locations of the access points using temporal matching between the Wi-Fi and anonymised GPS data. For 20 out of the 38 participants included in the dataset, a match was found. Of the remaining cases, 13 times the score of the candidate locations was below 10 and in five cases no anonymised coordinates could be found for the homeset access points. By comparing the homesets we could further identify four out of the 13 participants with low scores as couples (*i.e.* intersecting homeset, similar schedule, similar age, male and female). As their candidate anonymised GPS locations are also identical, we could thus lift their combined score over the threshold and validate four additional participants. Thus, for the majority of the participants in the dataset, we could verify that the coordinates of

the symbolic location identified as the home of the participants corresponded with the coordinates of access points included in the homeset, thus establishing the reliability of our homeset algorithm.

5.4 Conclusion and lessons learned

In this chapter we addressed the problem of learning occupancy schedules from unlabelled Wi-Fi scans. We showed that occupancy schedules can be reliably detected by clustering Wi-Fi access points seen during the night. By combining several access points to a *homeset*, we showed that the number of false negatives can be reduced. We evaluated the operation of the homeset algorithm by using artefacts of the anonymisation of GPS locations to serve as implicit labels. We thereby further showed that while anonymising significant locations helps to prevent the disclosure of the actual physical location of the home, the anonymised data itself can be used to infer *when* occupants are at home. The homeset algorithm, evaluated in this chapter on the MDC subset of the LDCC data collection campaign, allows us to derive long-term occupancy schedules at a granularity of 15 minutes. In the next chapter, we will apply the homeset algorithm on the full LDCC dataset and analyse the resulting occupancy schedules with respect to a number of different occupancy prediction algorithms for smart heating applications.

Predicting occupancy schedules

Conventional programmable thermostats operate according to user-defined schedules. Their settings need to be changed manually as the residents' occupancy schedules vary. Smart heating systems seek to overcome this need for manual re-programming by predicting household occupancy and supplying the control schedules to the thermostat without any direct user involvement. So when the occupants leave the building, the heating may be switched off automatically and the temperature allowed to drop to the setback temperature.

This reactive strategy fails when the occupants return, as the thermal properties of the house will result in a certain time lag until the comfortable temperature is reached again. The time lag describes the time taken by the heating system to reach the comfort temperature from the current indoor air temperature. The longer the house has been left unoccupied and the temperature has been allowed to drop, the greater the time lag will be. Therefore, at any given time, if the occupants have left the household, the system needs to know how long it would take to re-heat the property and whether the house is likely to be occupied within this time span. We call the time slots involved in this calculation the prediction horizon. We refer to the process of computing the future occupancy states within the *prediction horizon* as *occupancy prediction*.

In this chapter we will focus on occupancy prediction and perform a classification and review of state-of-the-art approaches. In Section 6.1, we outline different techniques used in the literature and identify three main classes (*schedule-based*, *context-aware* and *hybrid*) into which existing approaches can be categorised. We then perform a quantitative comparison of the prediction accuracy of selected schedule-based occupancy prediction algorithms. Sections 6.2 and 6.3 describe the experimental setup and the evaluation. The latter is based on actual occupancy data for 45 individuals collected over several months. We derived this occupancy data by analysing mobile phone records collected as part of the Lausanne data collection campaign (LDCC) [98]. Lastly, in Section 6.4 we present the results before we conclude this chapter in Section 6.5. This chapter is based on contributions made in [100, 101, 103].

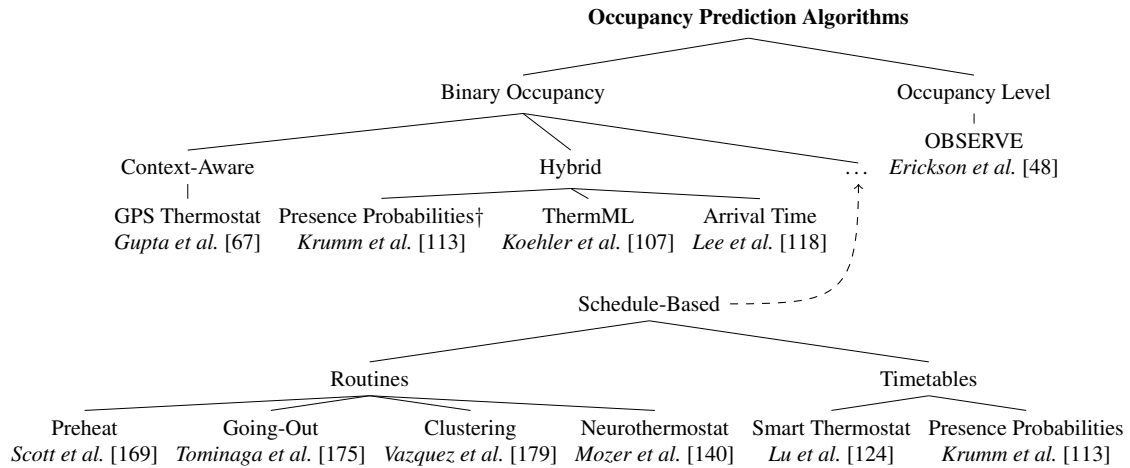


Figure 6.1: Classification of occupancy prediction algorithms.

6.1 A classification of occupancy prediction approaches

A number of occupancy prediction algorithms have been proposed in the literature [48, 67, 113, 124, 169, 175, 179]. Thereby, different mathematical models – including artificial neural networks [140] and Markov chains [48] – have been used.

Figure 6.1 shows an attempt to classify existing algorithms. We first differentiate between algorithms that predict *binary occupancy* (*i.e.* whether the home is occupied or not) and those that predict the *occupancy level* (*i.e.* how many occupants are present). While the occupancy level is relevant for office buildings, where fluctuations in the CO₂ level must be accounted for [48], binary occupancy prediction is sufficient for residential smart heating systems.

Existing binary occupancy prediction algorithms can be broadly categorised into three main classes – so-called *schedule-based*, *context-aware* and *hybrid* approaches. The distinction between these classes depends on the nature of the data being used to predict future occupancy. While schedule-based algorithms work solely on the historical occupancy data of the *building*, context-aware approaches utilise information about the current position, activity and environmental factors (such as the current traffic conditions) – the context – to predict the arrival of *individual occupants*.

Naturally, as outlined in the previous chapter, a body-worn sensor such as a mobile phone can be used to detect entry times to and stay durations in the home and thus provide data for a schedule-based algorithm. However, as it may also be used to sense the current context of the occupants, it may also be used for context-aware prediction. In this case, a hybrid approach can combine advantages of both context-aware and schedule-based occupancy prediction. In the following section we will describe the approaches classified

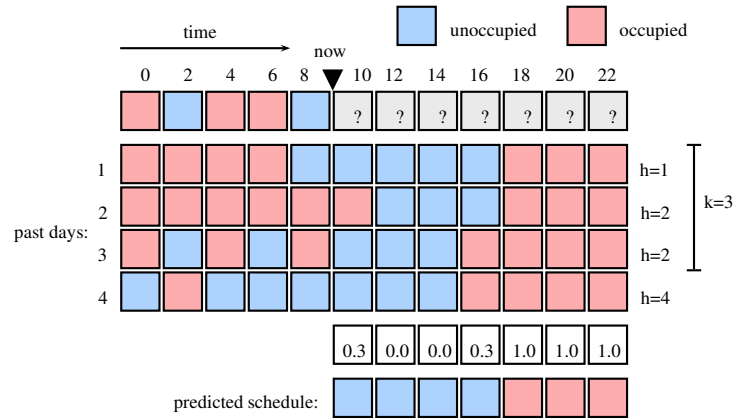


Figure 6.2: Occupancy prediction using the Preheat algorithm from Scott *et al.* [169].

in Figure 6.1 with a specific focus on schedule-based algorithms.

6.1.1 Schedule-based approaches

Several approaches compute occupancy predictions relying on past occupancy schedules only [113, 124, 168, 169]. Such approaches, which we refer to as *schedule-based* algorithms, take as input *historical* data on the household occupancy state. This data is typically collected over an extended period of time (weeks to months). Schedule-based algorithms can be distinguished into those algorithms that try to detect *routines* in the historical occupancy schedules (*e.g.* a late departure time could indicate a late return time) [140, 169, 175, 179] and those that assume that routines can be explained by daily or weekly *timetables* (*i.e.* human routine is assumed to be determined mostly by the current day of the week and the time of the day) [113, 124].

Preheat

The Preheat (PH) algorithm presented by Scott *et al.* [169] is an example of a schedule-based approach. The Preheat algorithm aims to exploit the occupants' routines by analysing the current occupancy and finding the most similar historical patterns. The authors thereby assume that the future occupancy depends on the partial occupancy trace of the current day.

For this purpose, the algorithm maintains a vector for storing the actual occupancy state registered for the current day starting from midnight. Each element of the vector represents the occupancy state of the home in a 15-minute interval. An element is set to 1 or 0 depending on whether the house is occupied or not during the relevant time interval. To compute an occupancy prediction from a given time of day onwards, Preheat first computes the Hamming distance between the occupancy pattern thus far observed for the

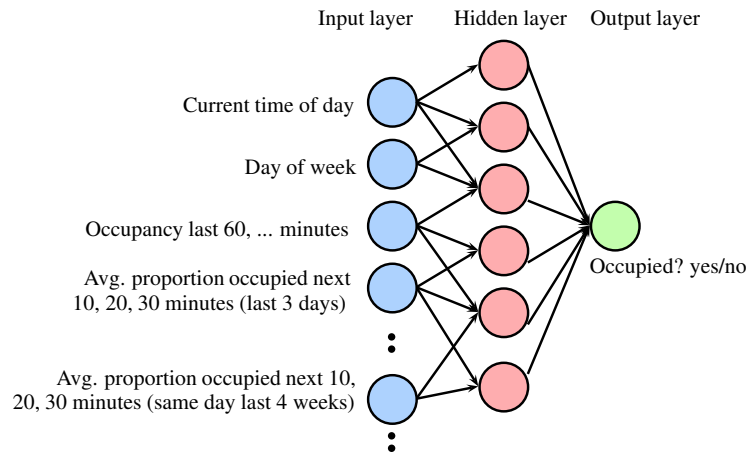


Figure 6.3: Three layer artificial neural network architecture. The number of hidden units is chosen by cross-validation.

current day and the corresponding segments of past occupancy vectors. The k past vectors with the lowest Hamming distances are then selected (k is fixed and equal to 5 in [169]) and averaged element-by-element. These averages approximate to the probability for the home being occupied during the corresponding time interval. The actual prediction is computed assuming that the house will be occupied during a future time interval if the corresponding probability exceeds a given threshold α , or else unoccupied. In [169], the value of α is fixed and equal to 0.5.

Building upon this basic version of the algorithm, Scott *et al.* introduce two additional features. The first consists of differentiating between weekdays and weekends. The second is to pad the current occupancy vector with data for the 4 hours before and after midnight, taken from the previous and following day respectively. This helps the algorithm to predict past midnight. Once the prediction is computed, the algorithm decides whether to start heating. This control decision depends on a number of factors including the current and desired temperatures as well as the rate (in terms of degrees per hour) at which the house can actually be heated.

Figure 6.2 shows a simplified version of the Preheat algorithm using 2-hour slots. The current time is 10 a.m. and thus far five slots have been observed. This vector is compared to previous days and the three most similar days are chosen according to their Hamming distance. The predicted occupancy vector is computed as the average of these three days.

Neurothermostat

The Neurothermostat (NT) by Mozer *et al.* combines a prediction algorithm with fixed-horizon planning to optimise the tradeoff between heating costs and occupant discomfort [140]. Like the Preheat algorithm, NT uses information about recent occupancy to detect routines and predict future occupancy.

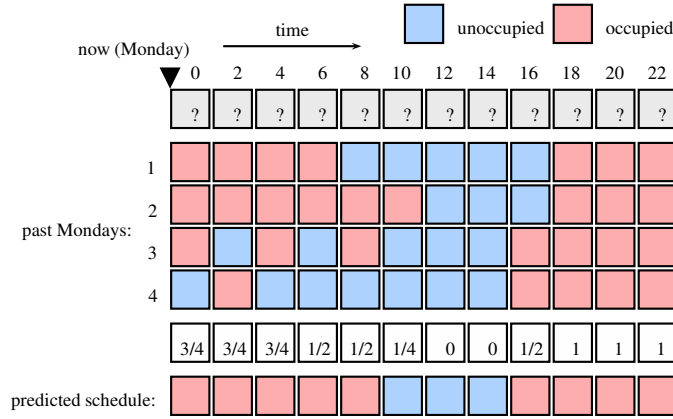


Figure 6.4: Occupancy prediction for Mondays using the Presence Probabilities (PP) algorithm [113].

To find the optimal time to start heating the building, the authors provide a simple first order approximation of the heating cost as well a misery cost function which converts uncomfortable temperatures to a monetary unit. The result is an optimisation problem with the goal to minimise the combined cost of heating energy and discomfort.

Similar to other dynamic programming problems such as the shortest path problem [41], NT considers at every time step all possible decision sequences over the planning horizon to choose the sequence that minimises the expected cost. The occupancy predictor then predicts if the house will be occupied at δ minutes in the future. For the deterministic parts of the schedule (*e.g.* during the night), a lookup table is used. The table is indexed by the time of the day and the current occupancy value. The residual structure is encoded in an artificial neural network. Figure 6.3 shows a three layer artificial neural network like the one used by Mozer *et al.* [140]. The authors used five consecutive months of real occupancy data from the Neural Network House [139] to train the models and tested it on the following month.

Presence Probabilities

The Presence Probabilities (PP) approach presented by Krumm and Brush is another well-known schedule-based approach [113]. In contrast to Preheat, PP does not attempt to find routines in the historical data but builds a fixed 7-day occupancy timetable. The authors thereby assume that any variation in the occupancy is best described by the current day of the week.

In [113], household occupancy is detected using a GPS device carried by the residents. The home is assumed to be occupied if the device indicates that a resident is less than 100 meters away from it. Using the GPS data, PP computes the probability for a home being unoccupied – called p_{away} – during any time slot of a day of the week. The values

of p_{away} in slots are computed using the ratios between the number of GPS data points that lie outside the 100-meter radius of the home and the total number of GPS data points available for the slot. The value of p_{away} for each time slot is stored in a vector called p_{week} containing 336 elements (7 days a week, 48 slots a day). The probability within each slot is smoothed using the values of the previous and subsequent slots. To adjust the values of p_{away} for weekdays, a generic vector p_{weekday} that contains the average values of p_{away} for a “generic” weekday is used. Using a regularisation factor λ_{wd} this vector can account for “*greater or lesser variability on weekdays*” [113]. The values of p_{away} in each slot of the final *probabilistic schedule* \tilde{p}_{week} are then computed as the sum of the elements of p_{week} and the relevant elements of p_{weekday} .

Analogous to the example for the PH algorithm, Figure 6.4 shows a simplified example of the PP algorithm for a particular Monday. In order to compute the predicted schedule, all occupancy vectors of previous Mondays are averaged.

Smart Thermostat

The Smart Thermostat (ST) by Lu *et al.* [124] also relies on historical schedules to predict occupancy. In contrast to PP, the authors do not make the distinction between individual weekdays and focus on the use of arrival times to optimise a multi-stage heating system.

The occupancy state of a home is determined using a Hidden Markov Model. The model allows an estimate of whether the home is occupied or not and in the former case also whether the occupants are asleep or active. To compute the estimation, the Hidden Markov Model takes as input both prior information derived from historical schedules and actual data collected by several sensors deployed within the home (*e.g.* PIR sensors).

The model is trained using a set of actual past occupancy schedules and sensor data traces. When the house is classified as unoccupied, ST switches the heating system off and allows the temperature of the household to fall to a “deep” setback temperature. If the occupants were to come back home unexpectedly while the house was at the deep setback temperature they would experience a significant comfort loss. ST thus keeps records of all previously observed arrival times (*i.e.* the time instants at which the house became occupied again after a period of absence).¹ The minimum of such previous arrival times is set as the time by which the household must be preheated to at least a “shallow” setback temperature. This mechanism makes it possible to reduce the risk of comfort loss.

ST also estimates the optimal time instant t^* – called the *preheat time* – at which the heating system must be activated to preheat the house. The preheat time t^* is chosen so as to minimise the average amount of energy wasted to heat the household and maintain it at the comfort temperature when the occupants are out. To identify the preheat time for a given day, ST considers all arrival times $\underline{a} = [a_0, a_1, \dots, a_n]$ observed on previous days.

¹Although this is not specified explicitly in [124], we assume that only one arrival event per day is considered.

Then it considers all time instants $t \in [\max(\underline{a}), \min(\underline{a})]$ for the current day as candidate preheat times. For each $t_i \in [\max(\underline{a}), \min(\underline{a})]$, the system computes the amount of energy waste $w_j(t_i)$ that would occur if t_i were the preheat time and the household were to be occupied again at arrival time a_j . The expected average energy waste that would occur if t_i were the preheat time is then the average: $w(t_i) = \sum_{j=1}^n w_j(t_i)$. The preheat time is chosen as the time instant that minimises the expected average energy waste:

$$t^* = \underset{t_i \in [\max(\underline{a}), \min(\underline{a})]}{\operatorname{argmin}} w(t_i) \quad (6.1)$$

The occupancy prediction mechanism of ST thus requires the identification of arrival times based on past schedules. Both the minimum of these arrival times and their weighted average are used to trigger different stages of the heating system. For the computation of the amount of energy waste, ST assumes a three-stage heating system and the availability of knowledge about the energy consumed by each stage.

Other schedule-based algorithms

Besides the Preheat algorithm, various authors have investigated how to cluster historical occupancy data to obtain characteristic schedules for specific routines. To this end, Tominaga *et al.* [175] present a non-parametric clustering approach based on Dirchlet process mixtures (DPM). In contrast to simple classification approaches such as k-Means, the authors do not require advance knowledge of the number of characteristic occupancy schedules. In [179], Vazquez *et al.* introduce a similar clustering approach based instead on fuzzy c-means [145] and self-organizing maps [108, 109].

6.1.2 Context-aware approaches

Due to the ubiquity of mobile phones with embedded GPS sensors, several authors have proposed techniques that estimate the future occupancy state of a home by observing the current context of its occupants. We refer to these techniques as *context-aware* approaches, since they depend on the current context (*e.g.* location or activity) of the user, rather than the home's historical occupancy schedule. One example of this is the algorithm presented by Gupta *et al.* [67], which estimates the time at which residents will return home based on their current position and driving trajectory. The position is determined using GPS modules embedded either in dedicated devices or in occupants' mobile phones. A web-based mapping service is used to determine the distance from home and the corresponding remaining *drive time*. The thermostat is then instructed to preheat the home if the remaining drive time is less than a given threshold.

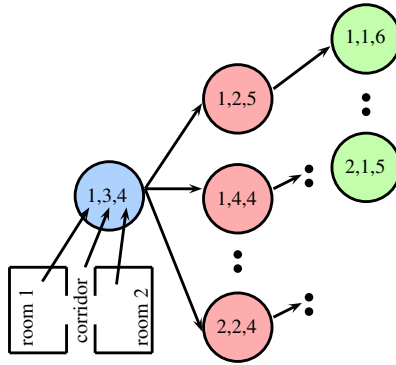


Figure 6.5: Markov Model. Numbers denote occupants in room 1, corridor and room 2 respectively.

6.1.3 Hybrid approaches

As location data from mobile phones can also be used to build historical occupancy schedules, several authors have sought to combine schedule-based and context-aware approaches. In [113], Krumm and Brush show how to combine their PP algorithm with Gupta *et al.*'s drive time prediction approach to build such a *hybrid* prediction algorithm. In contrast to [67], Krumm and Brush allow drive times to be pre-computed, thereby increasing efficiency but reducing accuracy, particularly in areas prone to congestion. In an earlier paper [114], Krumm *et al.* also introduced a method called *Predestination*. This method uses historical data along with information on a user's driving habits to obtain the most likely next destination. A similar system, *TherML*, is presented by Koehler *et al.* [107]. TherML utilises a hybrid prediction algorithm that switches between predicting the next destination and static schedules based on the user's mode of travel (stationary, walking or driving). Other approaches such as [168], [118] and [187] also use context information about the user to predict where he/she is likely to go next.

6.1.4 Other approaches

A number of occupancy detection and prediction approaches focus not only on heating but also on ventilation. Erickson *et al.* argue that in order to control ventilation and heating one needs to sense and predict the *level of occupancy* (*i.e.* how many people are present in each room at any one time) [48].

To detect occupancy, the authors have deployed 16 cameras at transition boundaries (*e.g.* between corridor and offices). A Markov Model was then used to model occupancy levels. The occupancy states are represented as shown in Figure 6.5. In order to predict which rooms to condition, Erickson *et al.* use a Markov Chain Model which given the occupancy distribution at time t , calculates the distribution at time $t + \Delta t$ by multiplying

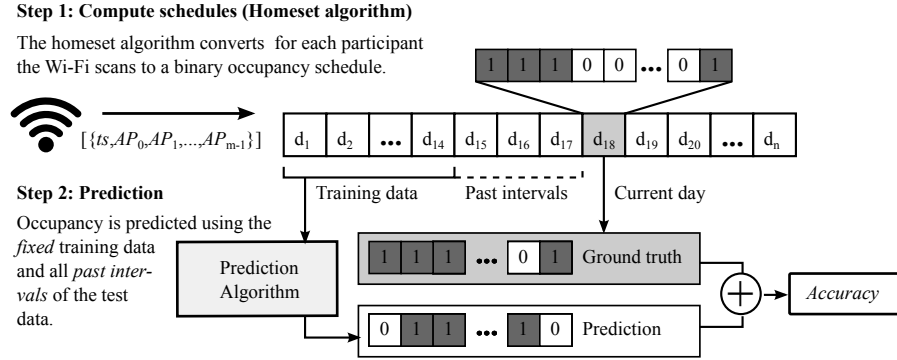


Figure 6.6: Prediction pipeline.

Δt times the transition matrix. The transition matrix gives the probability of moving from one occupancy state to another during a certain time interval.

This naïve approach scales exponentially as more rooms are added. To overcome this problem, the authors only use states observed over a 2-day training period. As the probability of occupancy states is correlated with the time of day, one transition matrix per hour is used. However, as not all states may be present in this matrix, sink states can occur. In addition, due to the partitioning into hourly slots, boundary discontinuities may prevent the system to transition from one to the next transition matrix. The authors overcome this problem by using a Blended Markov Chain in which all hourly transition matrices are blended together with a weighting that favours the current transition matrix over more distant ones.

6.2 Experimental setup

In the remainder of this chapter we report the results of a quantitative analysis of the schedule-based Preheat (PH), Presence Probabilities (PP) and Smart Thermostat (ST) algorithms introduced in Section 6.1. The evaluated algorithms are listed in Table 6.1. Using occupancy schedules derived using the homeset algorithm introduced in the previous chapter we will investigate their prediction accuracy.

Figure 6.6 shows an overview of the occupancy prediction infrastructure. We first compute the occupancy schedules from the raw Wi-Fi scans contained in the LDCC dataset using the homeset algorithm. The resulting occupancy schedule is then split into training and test data using cross validation. The training data is used for the initial setup of the prediction algorithms. The algorithms then learn more information as past days are added to the historical data. In this section, we will first discuss in detail our implementation of the algorithms before reporting on the schedules used.

Table 6.1: Algorithms considered for the comparative performance analysis.

Acronym	Name	Source
PH	Preheat	[169]
PP	Presence Probabilities	[113]
PPS	Presence Probabilities Simplified	[113]
MAT	Mean Arrival Time	Emulating ST [124]
MDMAT	Minimum Distance Mean Arrival Time	Emulating ST [124]

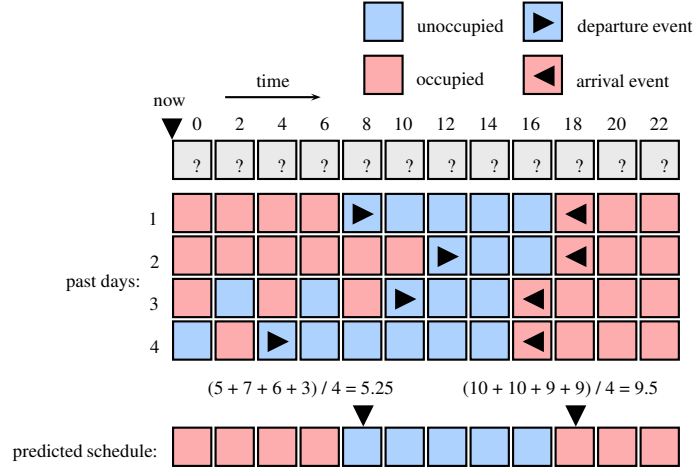


Figure 6.7: Mean Arrival Time (MAT) occupancy prediction algorithm.

6.2.1 Algorithm implementations

Our comparative study focuses on schedule-based approaches and includes both the PP (or PPS) and PH algorithms. We refer to the version of the Presence Probabilities algorithm described above as PP and to a simplified version that does not consider smoothing or the generic weekday schedule as PPS.

In place of ST itself we instead considered two heuristic prediction strategies – called Mean Arrival Time (MAT) and Minimum Distance Mean Arrival Time (MDMAT) – which mimic the occupancy prediction algorithm used by ST. As described in Section 6.1.1, ST uses the minimum of all previously observed arrival times as the time instant at which the household has to change from deep to shallow setback. ST also heats the house to the comfort temperature using a policy that minimises energy waste. To this end, a three-stage heating system with different efficiencies for each stage is assumed to be in place. In this thesis, we analyse performance (*e.g.* efficiency gain) in terms of occupancy prediction separately from a specific heating strategy (*cf.* Chapter 8). Also, we assume a single-stage heating system. Thus, ST would always choose the latest observed arrival time as the preheat time. This is due to the fact that heating reactively guarantees the lowest energy waste when comfort loss is not considered and a single-stage heating system is in place.

We therefore introduce the MAT and MDMAT methods as adaptations of ST’s preheat-

ing strategy. Like ST, the MDMAT algorithm records all n observed arrival times in a vector \underline{a} . For each $a_i \in \underline{a}, i = 1, \dots, n$, MDMAT calculates the distance to all other arrival times $a_j \in \underline{a}, j \neq i$ as:

$$d(a_i) = \sum_{a_j \in \underline{a}, j \neq i} \min(|a_i - a_j|, |a_i - (a'_j + 24)|) \quad (6.2)$$

The *most likely* arrival time for the current day is then chosen as:

$$a^* = \operatorname{argmin}_{a \in \underline{a}} d(a) \quad (6.3)$$

As shown in Figure 6.7, MAT instead computes the expected arrival time for each day as the arithmetic mean of the arrival times recorded on all previous days. To this end, only one arrival time per day is considered. This is selected as the first arrival event after 2 p.m. and before 2 a.m. We impose this restriction to limit the effect of outliers (*e.g.* unusual arrival events in the morning) and to avoid the computation of the arithmetic mean of the arrival times causing misleading results due to the use of a 24-hour interval.² In contrast to ST's original strategy, which targets a reduction in energy consumption, MAT and MDMAT trade off energy efficiency against comfort loss.

6.2.2 Preparing the LDCC occupancy schedules

To compare the performance of different occupancy prediction algorithms in a consistent manner, we evaluate them using a large dataset of actual occupancy schedules. We infer these schedules from sensor data collected as part of the Lausanne data collection campaign (LDCC) [98]. To the best of our knowledge, no publicly available data exists of long-term, high-granularity occupancy schedules, making it necessary to build such schedules in order to conduct our evaluation.

The LDCC dataset contains about 18 months' worth of traces of Wi-Fi scans, GPS coordinates, accelerometer readings and several other sensors, as well as demographic information from mobile phone users [98]. However, the dataset does not contain any information concerning users' relevant places, *i.e.* it is not known where the user's home, office, etc. are located. In the previous chapter we have therefore introduced a technique, called the *homeset algorithm* to infer this information from the available LDCC data.

For the analysis of the prediction algorithms introduced in this chapter, we ran the homeset algorithm on all participants of the LDCC. To exclude participants with too little data for evaluation, we used only occupancy schedules for users who had collected data for at least 100 days (*i.e.* $N_d > 100$) and for whom the occupancy state could be inferred in at

²For example, given two arrival events – one at 1 a.m. and one at 9 p.m. (21:00), their arithmetic mean computed over a 24-hour interval (from 00:00 to 24:00) would return the value 11 a.m., although the desired mean value would be 11 p.m.

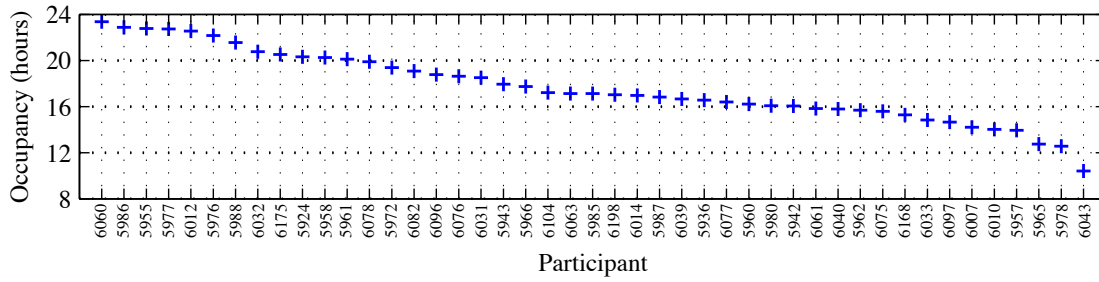


Figure 6.8: Occupancy in hours for all 45 households in the dataset (identified by the unique LDCC participant number).

least 70% of the slots. This was done to ensure sufficiently large training and test sets. We also discarded the schedules of users whose probability of being at home between 3 a.m. and 4 a.m. on weekdays was estimated to be less than 60%. This ensured we considered in the study only users for whom the homeset algorithm could reliably identify the home. This first data cleaning phase enabled us to select 59 occupancy schedules.

The Preheat algorithm by Krumm *et al.* imposes additional constraints on the input data. For instance, daily schedules need to be padded with four hours from the previous day and four hours from the next day [113]. We consequently discarded from the schedules all days for which this information was not available in order to ensure all algorithms were trained and tested on the same data. This left 45 schedules to be used for our evaluation. Figure 6.8 shows the average occupancy in hours per day for all the participants in the dataset. On average, these schedules include 74 days' worth of occupancy data, with the participants staying at home for 17 hours and 40 minutes per day on average. The weekly probabilistic schedules for all 45 participants are included in the Appendix (*cf.* Section C.2).

6.3 Evaluation

In this section we will outline the criteria used in our performance analysis of the MAT, MDMAT, PP, PPS and PH algorithms. We will first describe how we measure the performance before we explain how we use cross-validation to vary the training data.

6.3.1 Performance measures

We say that a *true positive* prediction occurs when an algorithm predicts a house will be occupied during a time slot k and the house is indeed occupied during that time slot. Likewise, correctly predicting the house to be *unoccupied* corresponds to a *true negative* prediction. *False positive* and *false negative* predictions occur when the household is

incorrectly predicted to be *occupied* or *unoccupied*, respectively. If, more formally, tp denotes the number of time slots with a true positive prediction (and likewise for tn , fp and fn), the *prediction accuracy* Acc_p of an algorithm is defined as:

$$\text{Acc}_p = \frac{tp + tn}{tp + tn + fp + fn} \quad (6.4)$$

To compare the considered algorithm against a baseline, we introduced a so-called *naïve predictor*. Given the a priori probability p_{occ} of the home being occupied, the *naïve* algorithm always predicts it to be occupied if $p_{\text{occ}} \geq 50\%$. If $p_{\text{occ}} < 50\%$ the *naïve* predictor always predicts the house to be unoccupied. For our study, we computed p_{occ} from the occupancy schedules as the number of slots containing a 1 in the schedule divided by the total number of slots.³

6.3.2 Cross-validation

All occupancy prediction algorithms evaluated in this chapter require an initial training phase to be able to predict future occupancy. As Figure 6.6 shows, we accommodate for this by means of designated training data. Therefore, we split the data into n 14-day blocks and perform n -fold cross validation on the choice of this initial training data. While for every run those 14 days are only used for training, the other data is progressively learned by the prediction algorithms as it becomes available.

6.4 Results

This section presents the results of our study. We first report on the prediction accuracy achieved by the MAT, MDMAT, PP, PPS and PH algorithms for the occupancy schedules derived from the LDCC dataset. We then show that they achieve a prediction accuracy close to the theoretical upper bound defined by the *predictability* of the input schedules.

6.4.1 Prediction accuracy

Figure 6.9 shows the prediction accuracy of all five algorithms considered in this study along with that of the *naïve* predictor for the LDCC occupancy schedules. For each prediction algorithm, the box plot indicates the median as well as the 25th and 75th percentiles of the accuracy across all 45 households. The interquartile range between the top and the bottom of the box thus represents the accuracy achieved in 50% of the homes. The whiskers represent the extreme data points (within $\pm 2.7\sigma$).

³As noted in [19], the naïve predictor was often quite accurate since typical residents spend a significant amount of their time (60% or more) at home.

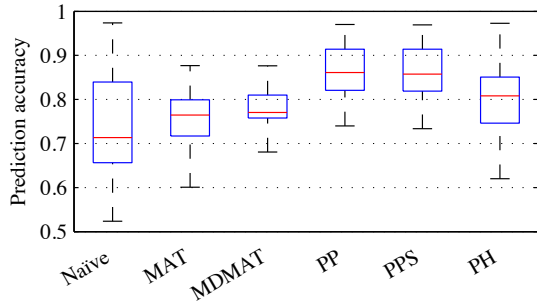


Figure 6.9: Accuracy of prediction algorithms considered in this study.

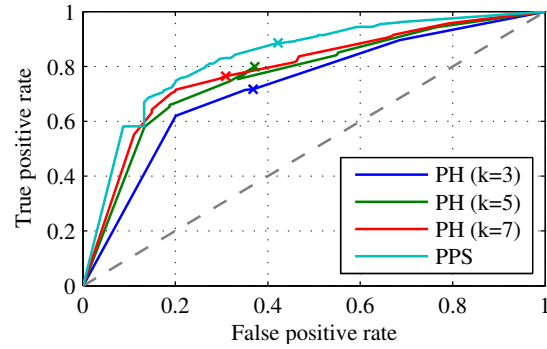


Figure 6.10: ROC curves of PPS and PH. Crosses indicate $\alpha = 0.5$.

With regards to median accuracy, all surveyed algorithms improve upon the baseline provided by the naïve predictor. The PP (or PPS) algorithm achieved the highest prediction accuracy. Its median accuracy lies around 85%, which means that the algorithm achieves at least this accuracy in 50% of the homes in the dataset. It is also the only algorithm for which the accuracy never dropped below 70%, which is the median value of the naïve predictor. We used Tukey’s honest significant difference (HSD) test [161] at the 95% level in conjunction with a one-way balanced analysis of variance (ANOVA) to establish that the mean accuracy of the PP algorithm was significantly different to the accuracy of the other algorithms (except PPS). The ANOVA assumes the distribution of the accuracy for each algorithm to be Gaussian. Confirmation that this assumption holds for the data under analysis was obtained using a two-tailed Shapiro-Wilk test at the 99% confidence level (p-values between 0.23 and 0.75).

The PH algorithm also achieved a median accuracy around 80% although it exhibits larger deviations to both sides of the median. This shows that for selected homes, PH can achieve a higher accuracy. For the average home, however, PP was the algorithm that performed best. In contrast, the prediction performance of MAT and MDMAT, which are considered here as representative of the basic techniques used by the ST algorithm was noticeably worse. The whiskers indicate that MAT and MDMAT are not suitable for schedules resulting in high values for p_{occ} (*i.e.* schedules for users who are almost always or almost never at home). This is due to the fact that for every day, MAT and MDMAT assume a period of absence between the computed mean departure and mean arrival times. A single day containing a 9-hour absence may thus result in a predicted schedule with an implied 63% probability of occupancy. In the case of a house otherwise occupied 90% of the time (*i.e.* $p_{\text{occ}} = 90\%$), this results in a drop in accuracy of 27%.

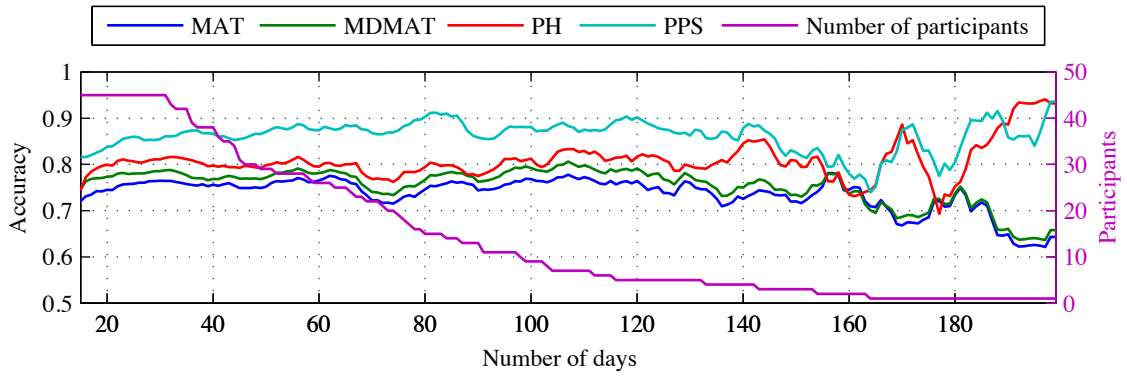


Figure 6.11: Impact of additional training data on prediction accuracy.

6.4.2 Parameter selection

Figure 6.10 shows the receiver operating characteristic (ROC) curves for the PH and PPS algorithms. The curves highlight the tradeoff between the *true positive rate*, defined as $tp/(tp + fn)$ and the *false positive rate*, defined as $fp/(fp + tn)$. The gray dotted line shows the performance of the random predictor (*i.e.* tossing a coin). The curves are obtained by varying the value of the threshold α (*cf.* Section 6.1.1). The cross markers on the curves show the data points corresponding to $\alpha = 0.5$. For both PH and PPS, setting $\alpha = 0.5$ as done in [169] achieved a good balance between true positive and false positive rates. The figure also shows how the performance of the PH algorithm changes for different values of the parameter k (which represents the number of nearest neighbours taken into account when making the prediction). For $\alpha = 0.5$ and $k = 7$, PH achieved a higher true positive rate and a lower false positive rate than with other parameter configurations. As mentioned above, this is the configuration we used for PH in this study as well as the default choice proposed in [169]. For the PH algorithm we used a prediction horizon of 90 minutes.

6.4.3 Learning time and prediction accuracy

Figure 6.11 shows the average accuracy across all households over time. The scale on the x-axis starts at day 15 (*i.e.* after the initial training). The right y-axis shows the number of participants remaining while the left y-axis shows the accuracy of the algorithms. The size of the available data varies between the participants. All participants have at least 30 days of occupancy data. However, less than 10 out of the 45 participants have data exceeding 100 days.

The different curves show the accuracy for the MAT, MDMAT, PH and PPS algorithms. As the prediction performance can vary strongly between subsequent days, their accuracy has been smoothed by a 7-day sliding window. The figure shows that, despite the 14-day

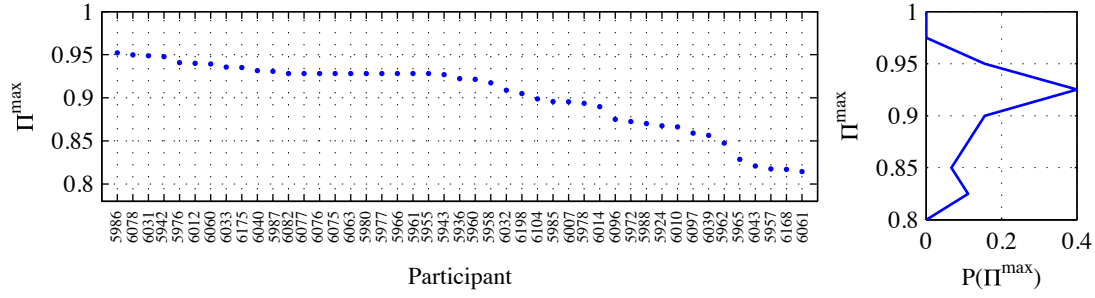


Figure 6.12: Distribution of predictabilities Π^{\max} over all participants.

training phase, there is a trend for the accuracy of all algorithms to improve slightly until day 30. After 40 days, the accuracy does not improve any further. This indicates that a 30-day sliding window approach to train the algorithms would work best. However, due to the limited amount of data, we cannot draw any conclusions about the long term behaviour of the algorithms.

6.4.4 Limits of predictability

The results presented above show that among the algorithms considered in this study, the PP predictor achieved the highest median accuracy of 85%. An obvious question to ask would be: Is it possible to do better? In other words, how close is the performance of PP to that of an “optimal” predictor? To answer this question, we built upon the results presented by Song *et al.* [173]. Their work targets the problem of predicting the next place visited by a person, given that the sequence of places visited thus far – referred to as the *mobility trace* of this person – is known. In this context, they introduce the concept of the *predictability* Π^{\max} of a mobility trace \mathcal{L} and show that it represents the “*upper bound that fundamentally limits any mobility prediction algorithm in predicting the next location based on historical records*” [123].

The predictability Π^{\max} thus corresponds to the upper limit of the prediction accuracy achievable by schedule-based predictors. If the focus is on occupancy prediction, the next place visited by the participant in the LDCC dataset can either be home or “any place but home.” We refer to these two places as L_1 and L_0 respectively. The sequence of places visited by a participant up to a time slot k can then be derived from the schedules. A value of 0 (or 1) in the schedule indicates that the place L_0 (or L_1) has been visited. For instance, assuming 15-minute slots, an excerpt of a schedule indicating a participant is at home for 1 hour and then away from home for 30 minutes corresponds to the sequence $L_1L_1L_1L_1L_0L_0$. In this way, we can derive the mobility trace for each participant and directly apply the method proposed by Song *et al.* to compute predictability values.

Figure 6.12 shows the predictability values of the schedules for the 45 participants con-

sidered in this study (left) along with the corresponding empirical distribution (right). The predictability is computed for each participant over the whole schedule. The participants are sorted in descending order of Π^{\max} from left to right. The maximum value of Π^{\max} is 95% while the minimum is 81%. The average of Π^{\max} over all homes is 90%. This value is thus an upper bound for the average prediction accuracy achievable by any predictor. In Section 6.4.1 (see Figure 6.9) we observed that the median accuracy of the PP algorithm was 85%, which is just 5% below the upper bound of 90%. This indicates that a fairly simple schedule-based approach such as PP can in itself capture most of the predictability intrinsic in typical occupancy schedules. Furthermore, this result indicates that the use of more sophisticated schedule-based algorithms will provide a maximum improvement in accuracy of about 5% only. Note, however, that the use of context-aware algorithms may push the achievable accuracy above the 90% limit, as with such algorithms information other than past occupancy schedules is used to compute predictions.

6.5 Conclusions and lessons learned

In this chapter, we analysed the prediction performance of state-of-the-art schedule-based occupancy prediction algorithms. Among the considered algorithms, the *Presence Probabilities* (PP, PPS) approach by Krumm and Brush [113] provides for the best overall performance in terms of prediction accuracy for the LDCC dataset. The approaches suggested by Lu *et al.* [124] and Scott *et al.* [169] (MAT, MDMAT, PH) perform slightly worse, albeit not by a large margin. All algorithms perform better if additional training information is added. However, after about 30 days no further improvements can be seen.

The reason for this is that the prediction accuracy of existing schedule-based algorithms is close to the achievable *theoretical upper limit*; this limit is expressed by the predictability of the underlying occupancy schedules. Further performance improvements can thus only be achieved by context-aware approaches that consider additional input information rather than occupancy schedules only.

A simulation model to analyse the performance of smart thermostats

The performance of state-of-the-art smart heating systems has been evaluated using a number of different strategies. Several authors show the results of simulations [48, 124, 140], while others report findings from real world deployments [67, 169]. Although there is no lack of models to simulate the energy consumption of a building, few authors have compared their approaches to existing work using the same model and dataset. The use of different evaluation scenarios and parameters, however, makes it difficult to compare results obtained by different authors. Also, a thorough description of the evaluation setup of an approach is often too verbose to be included in a research publication. The lack of such a description, however, makes it infeasible for other authors to reproduce previously achieved results. To ensure the comparability and reproducibility of research results on smart heating control, however, it is crucial to build and improve upon state-of-the-art approaches.

In this chapter, we first present a simple, generic and reproducible methodology based on current building performance standards to evaluate the performance of smart heating control systems. To simulate the building energy consumption we use the 5-resistance 1-capacitance (5R1C) model from the ISO 13790 standard [84]. To this end, we first discuss related work in Section 7.1 before we introduce resistance-capacitance models from first principles in Section 7.2. We then go on to present a framework for the simulation of the energy savings achieved by a smart heating system with a predictive controller (*cf.* Section 7.5). In Section 7.3 we show how using real meteorological data from 20 years, characteristic weather scenarios can be defined that allow for an extrapolation of the annual energy savings of a smart thermostat. In Section 7.4, we outline our parametrisation of the 5R1C model for four fictitious buildings in Lausanne, Switzerland. Section 7.7 concludes this chapter after a discussion of the limitations of the ISO 13790 5R1C model in Section 7.6. This chapter is based on contributions made in [103], [104] and [105].

7.1 Related work

Numerous authors have sought to define thermal models to simulate the energy consumption of buildings. Today, a large number of these approaches is based on the physical principles of electrical circuits [26, 58, 90, 129]. One of the first authors to advocate the use of resistors and capacitors to simulate the energy consumption of a heating system was Clemens Beuken, who introduced the concept in his 1936 PhD thesis on the “Heat loss in periodically powered ovens” [26]. In the Beuken method, the electric voltage corresponds to the temperature and the electric current is equivalent to the heating flux q .

Since Beuken, many authors have presented approaches utilising such resistance-capacitance (RC) models. The number of resistances and capacitances used varies between models. Therefore such models are usually referred to by a shorthand such as N_rRN_cC , whereby N_r and N_c denote the number of resistances and capacitances.

At the macro level, Kämpf *et al.* focus on the energy flows within urban districts and use a simple 1R1C model for modelling an arbitrary number of zones [90]. In the Neurothermostat [140], Mozer *et al.* simulate the energy consumption of the smart heating system utilising a simple 1R1C model. Like Mozer *et al.*, Wooley *et al.* investigate the performance of occupancy-responsive thermostats using a simple 1R1C model. Rogers *et al.* present a practical application of a simple RC model [164]. Their project, MyJoulo, a small USB temperature logger, builds an RC model of the building and automatically infers the operational settings of the heating system in order to predict the effect of intervention strategies such as suggesting a lower setback temperature [164].

To obtain more precise results, other authors increase the number of nodes in the RC model. In [129], Matthews *et al.* present a first-order thermal model to be used in building design using four resistances and one capacitance (4R1C). Similarly, Fraisse *et al.* present a three resistances and four capacities model to simulate a multi-layer wall (3R4C) [58]. Olofsson *et al.* [151] extend the European Standard EN 832 on the “thermal performance of buildings” [82] by incorporating heat loss through the floor and solar gains. Since 2008, EN 832 has been replaced by ISO 13790 which already includes these factors [84].

While the use of RC models is very popular in building performance simulations, several authors also propose alternative approaches. Kalogirou *et al.* for example, use an artificial neural network (ANN) that uses seasonal and building information to predict the energy consumption of passive solar buildings [89]. Similarly, Neto *et al.* analyse the results of modelling a building using an ANN for the energy consumption forecast of an office building in São Paulo [142]. Kramer *et al.* provide a literature review of various simplified building models including neural network models, linear parametric models and lumped capacitance models [112].

In the design process of large building projects, commercial software such as TRN-SYS [219] and EnergyPlus [220] are often used. Thereby, the EnergyPlus software, developed by the United States Department of Energy, has become the de-facto standard

tool for simulating the energy efficiency of buildings. Originally based on the earlier BLAST and DOE-2.1E simulators, version 1.0 of EnergyPlus was published in April 2001. As of the time of writing, the current version stands at 8.2. The software can perform complex multi-zone calculations and output energy and water usage of a building as well as its CO₂ emissions. For this reason, EnergyPlus has been used by a large number of projects including the One World Trade Center in New York City [203]. While several papers have used EnergyPlus models to evaluate the expected energy savings of smart heating systems [48, 124], its modelling approach is focussed at large commercial buildings and its control options are limited [148]. Its complexity results in models too specific to be generalised in the residential environment.

To achieve the right trade-off between complexity and generalisability, we use a 5R1C model for our simulation. This method has been standardised in the EN ISO 13790 energy performance standard [84] and adopted extensively for building simulations in Europe [36, 88]. The standard was mandated by the EU Directive 2002/91/EC on the energy performance of buildings (EPBD) which required a “*common methodology for calculating the integrated energy performance of buildings*” [51]. In the next section we will explain the operation of the resistance-capacitance models in general, before we introduce the specifics of the ISO 13790 5R1C model.

7.2 Lumped capacitance models

In building design, the indoor temperature may be modelled as a transient heat transfer problem. A simple example for transient conduction (*i.e.* heat transfer that is time dependent) is a banana cake taken out of the oven and left to cool down in the kitchen. Energy is transferred from the surface of the cake to its surroundings by convection and radiation. At the same time conduction also occurs between the interior of the cake and its surface. The energy transfer occurs as long as the cake has not reached a steady state temperature distribution.

One of the simplest models for describing such transient conduction is the *lumped capacitance model* [80]. The lumped capacitance model makes the simplifying assumption that there is no temperature gradient within the solid. Thus, the temperature on the surface of the cake is assumed to be the same as the interior temperature. This is clearly impossible as it would imply the existence of infinite thermal conductivity in the cake. However, this is well approximated if the internal conductivity in the cake is higher than the conductivity to the surroundings.

When looking at a building scenario, the rate of heat loss \dot{E}_{out} of the building must be the same as the rate of change of its internal energy \dot{E}_{stored} . This may be written as:

$$-\dot{E}_{\text{out}} = \dot{E}_{\text{stored}} \quad (7.1)$$

Table 7.1: Variables used in calculation of transient heat transfer with the lumped capacitance model.

Variable	Description
\dot{E}_{out}	Heat loss (W)
\dot{E}_{in}	Heat gain (W)
\dot{E}_{stored}	Heat stored (W)
Θ_{in}	Indoor temperature (K)
Θ_{out}	Outside temperature (K)
R	Thermal resistance (K/W)
C	Thermal capacitance (J/K)
h	Heat transfer coefficient (W/(m ² K))
A_s	Surface area (m ²)
ρ	Density (kg/m ³)
V	Volume (m ³)
c	Specific heat (J/(kg K))

7.2.1 A simple resistance-capacitance (1R1C) model

When considering the indoor temperature, we are not merely interested in reaching the steady state temperature distribution of the building with the outside as it cools down. Instead we want to know how much energy must be spent to keep the building at a comfortable temperature level and how long it takes to heat up to this level. Thus we must counteract the rate of change of the internal energy \dot{E}_{stored} by introducing heat gain \dot{E}_{in} into the system. Figure 7.1 shows the simple resistance-capacitance circuit used for this purpose.

$$\dot{E}_{\text{in}} = \dot{E}_{\text{out}} + \dot{E}_{\text{stored}} \quad (7.2)$$

Introducing the temperature difference between the indoor temperature Θ_{in} and outside temperature Θ_{e} , the equation may be re-written as:

$$\dot{E}_{\text{in}} = \frac{\Theta_{\text{in}} - \Theta_{\text{e}}}{R} + C \frac{d\Theta_{\text{in}}}{dt} \quad (7.3)$$

where R stands for the thermal resistance ($R = \frac{1}{hA_s}$) and C for the thermal capacitance ($C = \rho V c$) of the building components facing the outside, respectively. Table 7.1 shows the definition of all parameters.

Now equation (7.3) may be rewritten as:

$$\frac{d\Theta_{\text{in}}}{dt} + \frac{\Theta_{\text{in}}}{RC} = \frac{\dot{E}_{\text{in}}R + \Theta_{\text{e}}}{RC} \quad (7.4)$$

Multiplying by integrating factor $e^{\int \frac{1}{RC} dt} = e^{\frac{t}{RC}}$ and applying the product rule in reverse gives:

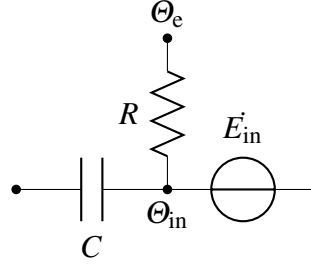


Figure 7.1: Simple 1R1C model.

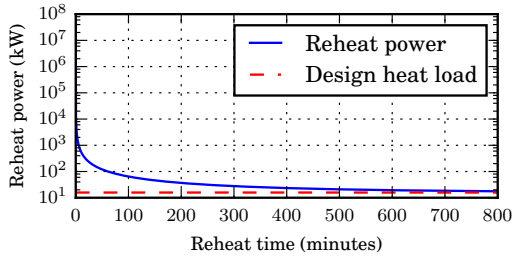


Figure 7.2: Time required to re-heat the building with respect to the available heating power.

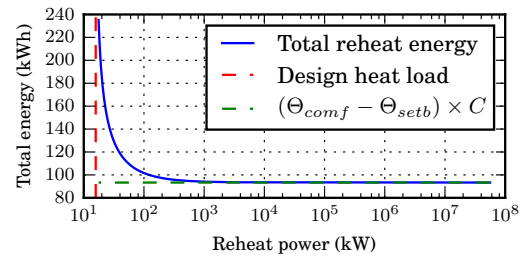


Figure 7.3: Total energy used to re-heat the building with respect to the available heating power.

$$\int \frac{d\Theta_{in}}{dt} e^{\frac{t}{RC}} = \int \frac{E_{in}R + \Theta_e}{RC} e^{\frac{t}{RC}} \quad (7.5)$$

$$\Theta_{in}(t) = (E_{in}R + \Theta_e) + D e^{\frac{-t}{RC}} \frac{E_{in}R + \Theta_e}{RC} \quad (7.6)$$

Fixing $\Theta_{in}(0) = \Theta_{in}(t-1)$ for $t = 0$ we can calculate D:

$$\Theta_{in}(t-1) = (E_{in}R + \Theta_e) + D \frac{E_{in}R + \Theta_e}{RC} \quad (7.7)$$

Substituting D in $\Theta_{in}(t)$ gives the indoor temperature at time t – $\Theta_{in}(t)$ – as a function of the indoor temperature at the previous interval $\Theta_{in}(t-1)$, the outside temperature Θ_e , the resistance and capacitance values (R and C) and the heat added to the system E_{in} . Equation 7.8 thus describes the temporal behaviour of the 1R1C model:

$$\Theta_{in}(t) = \Theta_{in}(t-1) e^{\frac{-t}{RC}} + (E_{in}R + \Theta_e)(1 - e^{\frac{-t}{RC}}) \quad (7.8)$$

7.2.2 Limitations of the 1R1C model

The simple 1R1C model does not explicitly model the losses of the heating system itself. This is important as an over-dimensioned and under-utilised boiler is usually less efficient. Furthermore, the asymptotic behaviour of the system would lead to the conclusion that a more powerful heating system is always preferable.

Figures 7.2 and 7.3 show the asymptotic behaviour of the 1R1C model if the available heating power is increased. As the available heating power tends to infinity, the time needed to heat the building approaches zero. The product of power and time – *i.e.* the energy used to preheat the building, however, approaches a non-zero value. From Equation 7.8 it can be derived that the minimum energy required for preheating E_{\min}^{\cdot} approaches the product of the capacitance and the temperature difference between the comfort and setback temperatures¹:

$$E_{\min}^{\cdot} = (\Theta_{\text{comf}} - \Theta_{\text{setb}}) \times C \quad (7.9)$$

Thus, as the available power approaches infinity, the energy required to reheat the building is not dependent on the resistance R . It is only dependent on the capacitance (*i.e.* the ability of the building to store energy). While it is to be expected that the energy required may not fall below the capacitance – as, after all, the heating process must adhere to the law of the conservation of energy – this result would point to an increase in the design heat load to save energy.

For this reason, we do not use the simple 1R1C model in our evaluation but focus on the ISO 13790 5R1C model. To avoid distorting the results by an over-dimensioned heating system, we further compute the design heat load $\Phi_{H,\max}$ (*i.e.* the maximum heating power available) using the DIN EN 12831 standard [42].

7.2.3 The ISO 13790 5R1C model

The simple 1R1C model has further drawbacks. It does not differentiate between the indoor air temperature and the temperature of walls and other building parts. This produces inaccurate results as the primary goal of the smart heating system is to ensure a comfortable indoor air temperature. Furthermore, the simple model ignores ventilation losses as well as solar and internal gains.

These factors are addressed by the ISO 13790 energy performance standard [84]. To simulate the hourly energy expenditure, ISO 13790 includes 5R1C model. The circuit of the 5R1C model is shown in Figure 7.4. The most significant parameters used by the model are listed in Table 7.2. The RC circuit models the transient conduction between the property and its surroundings and offers a method to calculate the energy needs for heating and cooling while maintaining specified set-point temperatures. In contrast to the simple

¹The interested reader may find the complete derivation in the appendix.

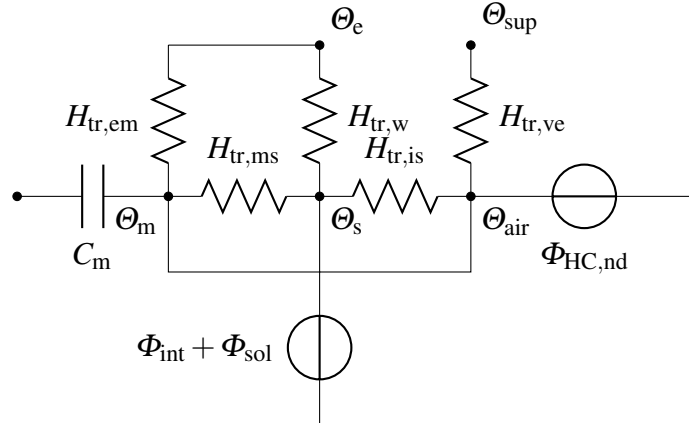


Figure 7.4: ISO 13790 5R1C model.

Table 7.2: Excerpt of ISO 13790 5R1C parameters.

Symbol	Units	Description	Notes
Output			
Θ_m	$^{\circ}\text{C}$	Mean radiant (masonry) temperature	
Θ_{air}	$^{\circ}\text{C}$	Mean air temperature	
Input			
$\Theta_{\text{int,H,set}}$	$^{\circ}\text{C}$	Heating setpoint temperature	$\Theta_{\text{comf}} = 20^{\circ}\text{C}$, $\Theta_{\text{setb}} = 10^{\circ}\text{C}$
$\Theta_{\text{int,C,set}}$	$^{\circ}\text{C}$	Cooling setpoint temperature	
$\Theta_{m,0}$	$^{\circ}\text{C}$	Mean radiant temperature at time $t = 0$	
Θ_e	$^{\circ}\text{C}$	Outside temperature	
Θ_{sup}	$^{\circ}\text{C}$	Temperature of ventilation air	$\Theta_{\text{sup}} = \Theta_e$
$\Phi_{\text{HC,nd}}$	W	Actual heat input	$\Phi_{\text{HC,nd}}$ in ISO 13790 covers heating and cooling
$\Phi_{\text{H,max}}$	W	Maximum heating input	
Φ_{int}	W	Internal heat gains	
Φ_{sol}	W	Solar heat gains	
H_{ve}	K/W	Ventilation heat transmission coefficient	
$H_{\text{tr,w}}$	K/W	Transmission heat transfer coefficient (windows, doors)	
$H_{\text{tr,op}}$	K/W	Transmission heat transfer coefficient (opaque elements)	
C_m	J/K	Thermal capacitance of building mass	
A_f	m^2	Floor area	

1R1C model, the 5R1C model takes into account the heat transfer by transmission and ventilation as well as solar and internal gains. The model also allows for the calculation of the mean transient air temperature Θ_{air} , the mean radiant (masonry) temperature Θ_m and the internal surface temperature Θ_s . In the following sections we will describe how we computed weather scenarios and building configurations for the 5R1C model. We then describe a predictive controller and conclude by discussing limitations of the 5R1C model.

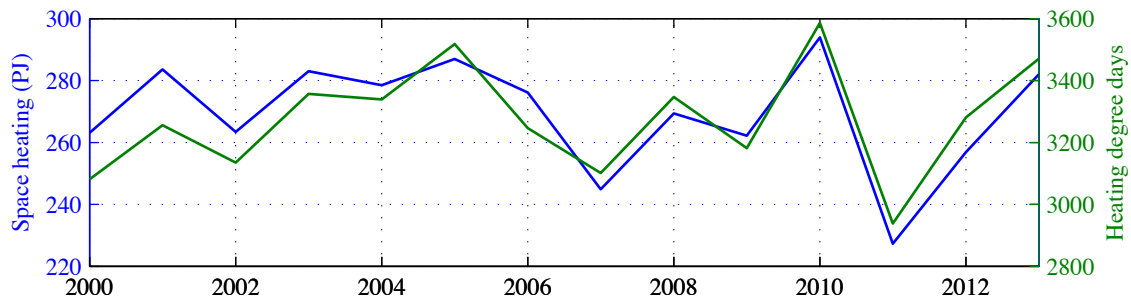


Figure 7.5: In Switzerland, the total energy consumption of space heating is closely following the heating degree hours (2000-2013).

7.3 Weather scenarios

The amount of energy required for heating is closely correlated with the weather. Figure 7.5 shows the average heating degree days and the energy required for space heating in Switzerland from 2000 to 2013. It shows that the energy consumed by space heating closely follows the heating degree days².

To investigate the performance of the smart heating system as close to a real system as possible, we use actual weather data for calibrating the 5R1C model for the design heat load and for simulating different environmental conditions. To set up the system, we use 20 years of weather data from Pully, Switzerland (close to Lausanne) to find Θ_d , the norm outside temperature, and $\hat{\Theta}_e$, the yearly mean of the outside temperature. Both are used in Section 7.4.2 for the calculation of the design heat load Φ_H . The weather data has been provided by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss).³ For details of the design heat load calculation the reader is referred to Section 7.4.2. Table 7.4 shows the weather data used. We use this data to design different environmental scenarios based on the distribution of the outside temperature.

Figure 7.6 shows the distribution of the daily average temperatures over 20 years from January 1994 to January 2014. To generate possible heating scenarios, we only consider days with an average outside temperature $\Theta_{e,d} < 20$, as only these may contribute to the total heating degree days. The left part of the figure shows that there is a peak in the relative frequency around 6°C. The small subfigure on the right shows that the median temperature in the observed sample is 10°C (*i.e.* 50% of the days requiring heating have temperatures below 10°C), while the lower quartile is 5°C (*i.e.* 25% of the days requiring

²The heating degree day is a measure of the energy demand for heating a building. It is derived from the outside temperature and defined relative to a base temperature – the minimum outside temperature at which heating is not required. In theory, the base temperature should vary with the characteristics of the building. A well-insulated building which makes good use of internal and solar gains has a lower base temperature than a poorly insulated building. However, for simplicity reasons, the base temperature is often assumed to be 16 °C or 18 °C.

³MeteoSwiss provides a web interface for researchers at <http://gate.meteoswiss.ch/idaweb>.

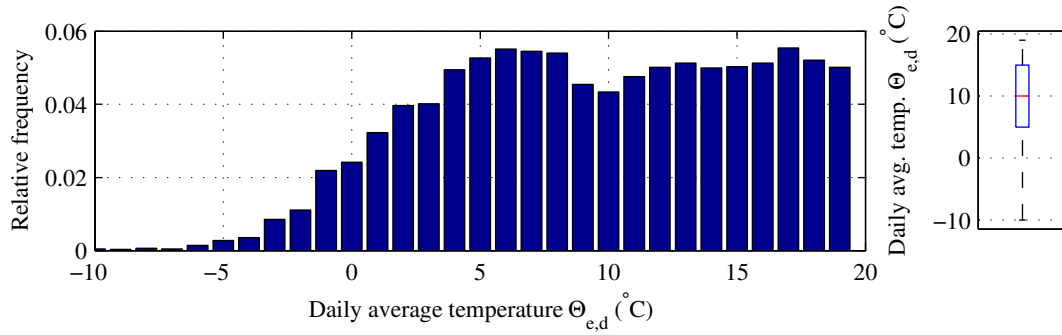


Figure 7.6: Distribution of daily average temperature $\Theta_{e,d}$ in Pully, Switzerland (Jan 1994 to Jan 2014). $\Theta_{e,d} \geq 20^\circ\text{C}$ are excluded. Right subfigure shows median and quartiles.

Table 7.3: Weather scenarios. For each of the eight scenarios, the table shows the daily average temperature $\Theta_{e,d}$ and the daily average of the global radiation I_{avg} for reference.

Scenario	Range	$\Theta_{e,d}$ ($^\circ\text{C}$)		I_{avg} (W/m^2)	
		clear	cloudy	clear	cloudy
Very low temperature	$-6^\circ\text{C} \leq \Theta_e \leq -4^\circ\text{C}$	-5.4	-4.7	142.9	35.5
Freezing temperature	$-1^\circ\text{C} \leq \Theta_e \leq 1^\circ\text{C}$	0.1	0.0	137.5	30.2
Low temperature	$4^\circ\text{C} \leq \Theta_e \leq 6^\circ\text{C}$	5.1	5.1	148.5	26.1
Moderate temperature	$9^\circ\text{C} \leq \Theta_e \leq 11^\circ\text{C}$	10.1	10.0	180.7	29.7

heating have temperatures below 5°C). The minimum value observed is -10°C . Since we have established that the norm outside temperature of our heating system is -6°C , we do not consider the minimum temperature. This led us to the eight test scenarios for our heating simulation shown in Table 7.3.

By choosing the bounds for the *very low* temperature to be slightly above the norm outside temperature, we seek to avoid a scenario where the power of our modelled heating system is insufficient to heat up the property to a comfortable temperature during parts of the day.

The eight weather scenarios are built using data from January 1, 2005 to January 1, 2014. All calculations on the solar radiation as described in Section 7.4.5 were done at a granularity of 10 minutes. The resulting dataset was resampled at 15-minute intervals to coincide with the occupancy schedules derived in Chapter 5. $\Theta_{e,d}$ denotes the daily average of the outside temperature and I_d denotes the daily average of the global radiation. Days with $I_{\text{avg}} \geq 100$ have been considered *clear*. Days with $I_{\text{avg}} \leq 50$ have been considered *cloudy*. The scenarios were built by considering all days in the dataset which fit the criteria. Figures showing the actual values for Θ_e and I_{dir} for all scenarios in conjunction with the response of an optimal controller can be found in the appendix.

Table 7.4: MeteoSwiss dataset: Pully, Switzerland (46°30'44", 06°40'03").

Variable	Description	From	To
I	Global radiation	01/01/2005	01/01/2014
Θ_e	Outside temperature (2 m above surface)	01/01/1994	01/01/2014

Weather data and occupancy Due to the fact that the occupancy traces from different households do not necessarily cover the same timespan (or have the same length), we cannot compare the energy savings obtained in different households if we were to use weather data corresponding to the actual occupancy data. Therefore, we assume that there is no correlation between the occupancy schedule of a household and the weather conditions.

7.3.1 Annual model

In order to estimate how much energy may be saved by a predictive heating system on an annual basis, we must take into account the relative frequency of the weather scenarios and weigh the scenarios accordingly. As Figure 7.6 shows, the very low temperature scenario with its proximity to the norm outside temperature is not very likely to occur on a typical day. The moderate temperature scenario on the other hand is much more likely to occur. Using the relative frequencies for temperatures below 20°C shown in Figure 7.6, we have calculated the probabilities for the different weather scenarios. For this, we extended the ranges shown in Table 7.3 to the ones in Table 7.5. This had no effect on the derivation of the scenarios, but made sure that the performance evaluation of the smart heating system covers the temperatures relevant for heating. Thus, Table 7.5 shows that a temperature between -2.5°C and 2.5°C is half as likely than a temperature between 2.5°C and 7.5°C . We have extended the ranges of the very low and moderate temperature scenarios to cover the whole range of temperatures.

While this does not affect the low temperature scenario (temperatures under -6°C do not occur very often), the probability of the moderate scenario is now overestimated in order to cover the temperatures up to 19°C . Compared to a scenario with equi-sized bounds (*i.e.* 7.5°C to 12.5°C), the probability of the moderate temperature scenario increases from 24% to 63% - an overestimation of 39%.

In order to understand the implications of this one must look at the behaviour of the heating system as the outside temperature increases. For higher temperatures, the absolute available savings drop to zero as Θ_e approaches Θ_{comf} and the solar gains are sufficient to heat the building. Heating becomes unnecessary. At the same time, the relative (percentage) savings increase as shorter preheat times lead to a more reactive heating system (*i.e.* the heating system can stay in the off state for longer). However, the relative savings are bounded above by the occupancy of the household. Heating can only be forgone when the building is unoccupied. The participants in our dataset are absent

Table 7.5: New ranges for weather scenarios with probabilities. Where ranges overlap, each scenario is assigned one half of the relative frequency of that particular temperature.

Range (°C)	Probability (%)	Used weather scenario
$-10 \leq \Theta_e \leq -2.5$	1	Very low temperature
$-2.5 \leq \Theta_e \leq 2.5$	11	Freezing temperature
$2.5 \leq \Theta_e \leq 7.5$	24	Low temperature
$7.5 \leq \Theta_e \leq 19$	63	Moderate temperature

Table 7.6: Average, average lowest and absolute lowest outside temperatures (Θ_e) in °C for selected cities for January to March. Estimated norm outside temperature (cf. Section 7.4.2) for the dimensioning of the heating system. Temperature data obtained from wikipedia.org, if available, otherwise from weatherbase.com.

City	Avg. Θ_e			Avg. lowest Θ_e			Abs. lowest Θ_e			Est. Θ_d
	Jan	Feb	Mar	Jan	Feb	Mar	Jan	Feb	Mar	
Moscow	-8	-7	-2	-11	-11	-5	-36	-33	-27	-20.5
Toronto	-5.8	-5.6	-0.4	-10.1	-10.2	-5.3	-35.2	-25.7	-25.6	-18.7
Beijing	-4	-1	6	-8.4	-5.6	0.4	-17	-15	-8	-8.9
Stockholm	-2.8	-3	0.1	-5	-5.3	-2.7	-27	-27	-20	-14.5
New York	0.5	1.8	5.7	-3	-1.9	1.4	-21.1	-26.1	-16.1	-11.1
Frankfurt	1	2	6	-1	-1	2	-20	-18	-12	-8.3
Lausanne	1.3	2.8	5.5	-0.5	0.5	2.7	-9.7	-13	-9.1	-4.9
Brussels	3.3	3.7	6.8	0.7	0.7	3.1	-17	-13	-7	-5.4
London	4.3	4.5	6.9	1.2	1	2.8	-12	-13	-7	-4.5
Paris	5	5.6	8.8	2.7	2.8	5.3	-14.6	-14.7	-9.1	-4.6
Seattle	5.6	6.3	8.1	2.7	2.7	4.1	-22.8	-27.4	-15	-9.3

from home on average 26% of the day [103]. This means that the average savings are bounded above by 26% as we cannot do better than switching off the temperature during unoccupied periods.

By using the moderate temperature as a model also for temperatures above 12.5°C we thus both overestimate the total energy spent and underestimate the percentage savings, leading to an acceptable error overall.

7.3.2 Global weather scenarios

The potential for energy savings achievable by a predictive heating system varies by region. In fact, some more moderate climates may rarely need any heating at all during the year. Moreover, the performance of a heating system is closely tied to its norm outside temperature and the resulting design heat load. A climate region with a larger variance in the outside temperature requires a more powerful heating system to cope with the lowest temperatures. This “excess capacity” to deal with the norm outside temperature

(cf. Section 7.4.2) then also reduces the ramp-up time during warmer days. Table 7.6 shows the average, average lowest and absolute lowest outside temperatures for the months from January to March for a selection of cities around the world. Since we were missing detailed weather data for the last 20 years for all cities, the norm outside temperature was determined as the mean of the average lowest and the absolute lowest temperatures from January to March.

The table shows that Toronto has the largest differences between the monthly average temperatures and the norm outside temperature Θ_d (varying from 13 °C to 18 °C), while Beijing has the lowest (between 5 °C and 15 °C). A heating system in Toronto is thus designed for a temperature of –19 °C, while a heating system in Beijing is sized to match outside temperatures around –9 °C.

We have used these data to simulate the impact of different climate scenarios by creating, for each city, three weather scenarios with constant temperatures equaling the average temperatures during the months from January to March. This means, for example, that the month of January in Toronto was simulated using a constant temperature of –5.8 °C. All weather scenarios were modelled without solar gains.

7.4 Building configurations

Our goal is to build a simulation framework that allows to show bounds on the potential of occupancy prediction algorithms to save energy in residential heating scenarios. In the simplest case, the possible savings are determined by the insulation of the building (transmission losses), its orientation and number of windows (solar gains), the building's exposure and tightness (ventilation losses) as well as heat gains due to occupants and appliances (internal gains). In the following section, we will describe how we calculated these heat losses and gains for four fictitious building scenarios.

In reality, the extent of the possible energy savings in a heating scenario also depends upon many other factors. Besides the particular solar energy transmittance of the glass used, solar gains are also influenced by the level of shading afforded by blinds, overhangs and other buildings or structures. Furthermore, the potential for solar gains is highly dependent on the location of the building. In mountainous terrain such as Switzerland, the sun may be partly or completely obscured during large parts of the day, resulting in lower solar gains. Similarly, ventilation losses are influenced not only by the characteristics of the building itself but also by its surroundings. An exposed building has higher ventilation losses than a building that is shielded off by adjacent structures. In addition to the tightness of the building, the level of exposure in conjunction with the current wind conditions determines the amount of ventilation losses. For similar reasons, the amount of heat losses due to ventilation depends on the height of the building. For tall buildings, the wind conditions near the top are different to those near ground level. Unless the building

Table 7.7: 5R1C model parameters for different building variants.

Parameter	Building variant				Units
	F-U _{low}	F-U _{high}	H-U _{low}	H-U _{high}	
Thermal transmission coefficient for opaque building elements – $H_{tr,op}$	47.16	184.57	103.57	379.35	W/K
Thermal transmission coefficient for windows and doors – $H_{tr,w}$	12.68	31.50	33.07	102.06	W/K
Thermal transmission coefficient for ventilation – H_{ve}	47.33	47.33	161.57	161.57	W/K
Internal zone capacitance – C_m	8.51	8.51	29.04	29.04	MJ/K
Floor area – A_f	51.56	51.56	176.00	176.00	m ²
Design heat load according to [42] – $\Phi_{H,max}$	2.80	6.86	7.78	16.75	kW

is detached and completely exposed, the energy required to heat might also vary as neighbouring buildings contribute to either transmission gains or losses. In particular, in an apartment complex, a particular party might not need to heat at all if the adjacent parties have heated their apartments to temperatures exceeding the comfort temperature of the first party. In fact, the heating scenario becomes much more complicated when multiple zones with different setpoint temperatures, unconditioned zones and occupancy schedules are considered.

We focus on an idealised scenario in which there is no heat transmission from and to adjacent buildings or zones. The buildings are considered to be a single zone with a single temperature setpoint. This simplification allows us to isolate the effect of the occupancy prediction algorithms on the energy expenditure of the building. We leave the analysis of the savings potential inherent in occupancy schedules of multi-party buildings to future work.

The two building configurations used in this study are a studio flat and a house. Both were simulated with low U-values following recent legislative guidelines [3] (good insulation: F-U_{low} and H-U_{low}) and high U-values (bad insulation: F-U_{high}, H-U_{high}), respectively. We then expose each of these four fictitious properties to a range of environmental conditions. Figures 7.7 and 7.8 show the geometry of the two simulated properties. The studio flat (F-U_{low} and F-U_{high}) has an area of 52 m². The house (H-U_{low} and H-U_{high}) has an area of 176 m². All windows are sized 1.4 m by 1.4 m. The height of the rooms is 2.5 m in both cases. The doors are 1.4 m by 2 m. The flat has one window facing east and three windows facing south. The house has two east-facing windows, four windows on the south side, two to the west and two windows facing north.

In the remainder of this section, we will show how we calculated the design heat load $\Phi_{H,max}$ as well as the heat gains (Φ_{int} and Φ_{sol}) and heat transfer coefficients for the ISO 13790 5R1C model ($H_{tr,w}$, $H_{tr,op}$ and H_{ve}). Table 7.7 summarises the resulting parameters used in the simulation using the ISO 13790 5R1C model.

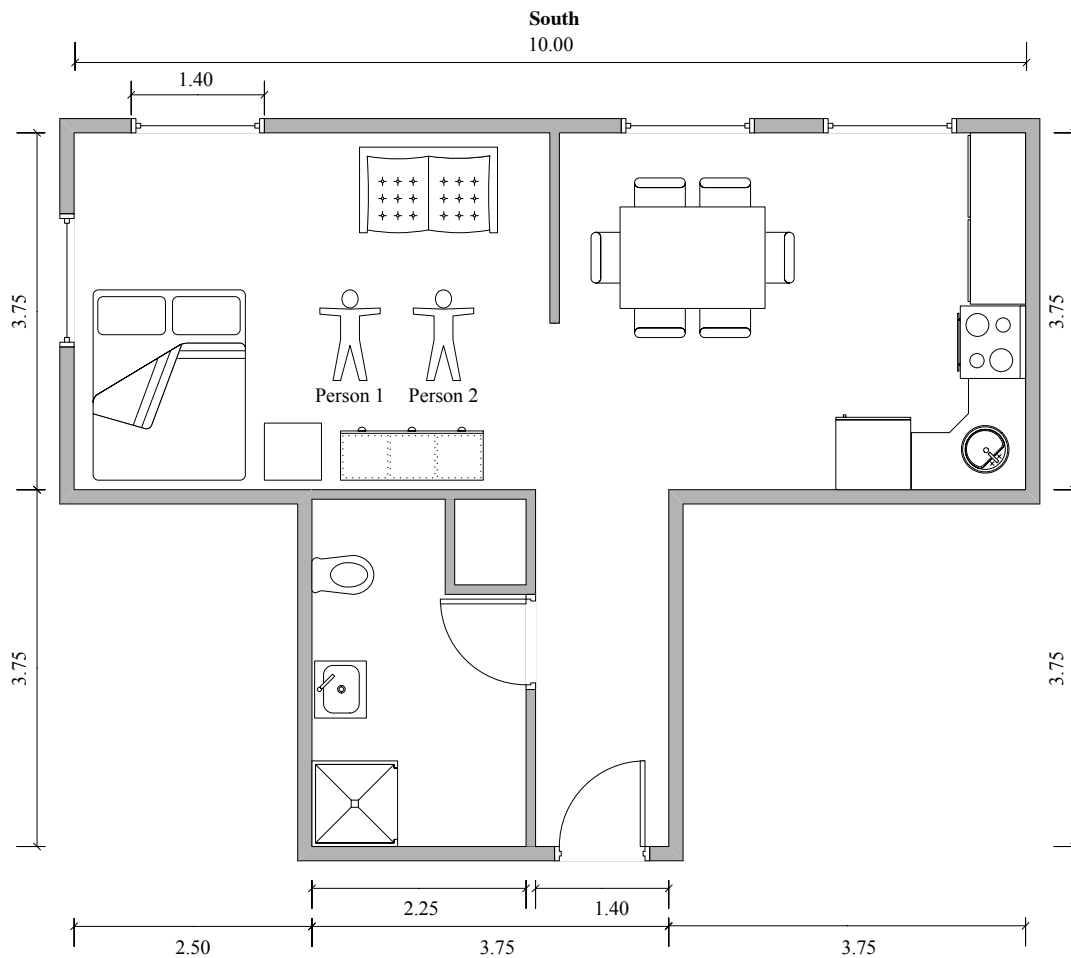


Figure 7.7: Blueprint of studio flat (F-U_{low} and F-U_{high}).

7.4.1 Transmission losses

A large share of the heat lost during cold weather is due to insufficient insulation. The insulation capacity of a material is described using the so-called U-value. The U-value gives the amount of energy (J) which is transmitted across the component at every second for a certain difference in temperature (K). As the difference between the inside and outside temperatures increases, the building loses more energy to its surroundings. The U-value is the inverse of the R-value with SI units of $W/(m^2 K)$. Table 7.8 shows two sets of characteristic U-values. The first set is taken from a low U-value reference building [3]. As there is no standard definition of high U-values (often anything higher than current building regulations allow is considered high), we have taken generic high U-values from wikipedia.org for the low insulation case [223].

The heat transfer coefficients $H_{tr,w}$ and $H_{tr,op}$ are calculated by multiplying the surface area of the building part facing the outside with its U-value.

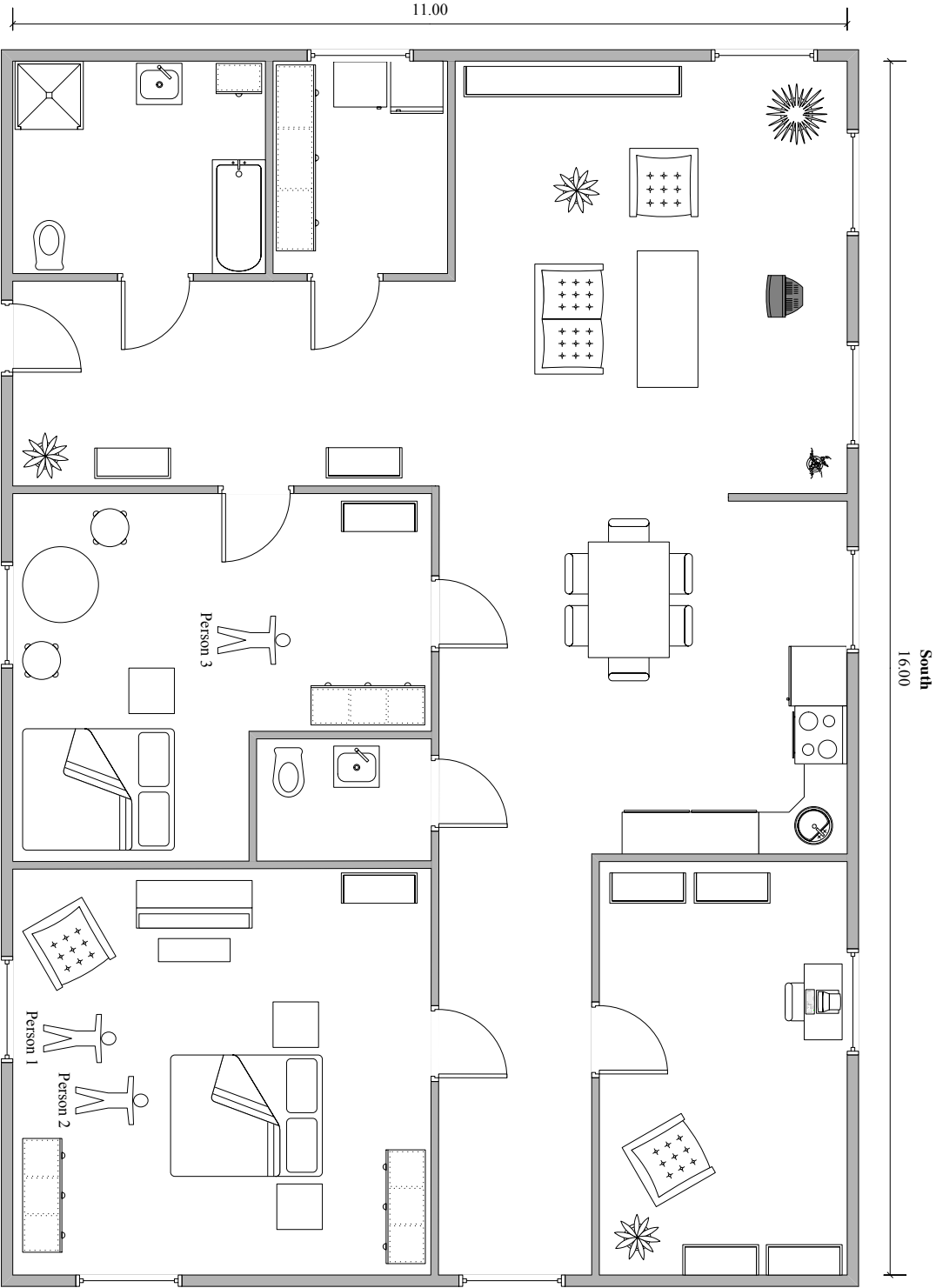


Figure 7.8: Blueprint of house ($H-U_{low}$ and $H-U_{high}$).

Table 7.8: U-values ($W/(m^2K)$).

Component	Low U-Values	High U-Values
Walls, ceiling against outside	0.28	1.5
Ground plate	0.35	1.0
Roof	0.20	1.0
Windows	1.30	4.3
Doors	1.80	1.8

Table 7.9: Design heat load for each building variant.

Term	Building variant				Units
	F·U _{low}	F·U _{high}	H·U _{low}	H·U _{high}	
$\Phi_{H,max}$	2.80	6.86	7.78	16.75	kW
$\Phi_{H,max}$ per m^2	12	30	10	21	W/m^2

$$H_{tr,w} = A_{walls,ceiling} \cdot (U_{walls,ceiling} + \Delta \vec{U}_{WB}) + H_{T,g} \quad (7.10)$$

$$H_{tr,op} = A_{doors,windows} \cdot (U_{doors,windows} + \Delta \vec{U}_{WB}) \quad (7.11)$$

where ΔU_{WB} is a generic thermal bridge correction factor which is set to 0.05 according to [71]. The calculation of the heat transfer coefficient to the ground $H_{T,g}$ is shown in the next section.

7.4.2 Design heat load

In order to appropriately dimension the heating infrastructure (*e.g.* radiators and boilers) in a property, the DIN EN 12831 standard allows for the calculation of the design heat load. The design heat load is the amount of heat that needs to be supplied to a building to keep the target comfort temperature Θ_{int} even when the outside temperature is at its lowest. In the case of the DIN EN 12831 model, Θ_d , the design outside temperature, is the lowest two-day average temperature which was measured at least 10 times over a period of 20 years. We have determined Θ_d using meteorological data from MeteoSwiss. Θ_d was calculated to be -6°C for the period from January 1994 to January 2014 for Pully, Switzerland. Likewise, we determined the yearly mean of the outside temperature $\hat{\Theta}_e$ to be 11.3°C . The temperature variable used for the calculations was the 10-minute average temperature (2 m above ground). The target indoor temperature Θ_{int} was 20°C . The resulting design heat loads are shown in Table 7.9. They are within the range of buildings built from 1978 to 1983 (high U-values) and buildings built after 2001 (low U-values) [56].

Table 7.10: DIN EN 12831 [42] parameters. † DIN EN 12831 refers to the norm outside temperature by Θ_e . To avoid confusion with the current outside temperature Θ_e as defined in ISO 13790, we chose to rename it.

Symbol	Units	Description
$\Phi_{H,\max}$	W	Design heat load: $\Phi_{TL} + \Phi_V$
Φ_{TL}	W	Heat loss due to transmission
Φ_V	W	Heat loss due to ventilation
$H_{T,e}$	K/W	Heat transmission coefficient to the outside (R-value)
$H_{T,g}$	K/W	Heat transmission coefficient to the ground (R-value)
Θ_{int}	K	Indoor temperature
Θ_d^\dagger	K	Norm outside temperature
$\hat{\Theta}_e$	K	Yearly mean of outside temperature
A_f	m ²	Floor area
P	m	Circumference of ground plate in contact with environment

Heat transmission coefficients

The heat loss due to ventilation Φ_V was calculated using the method used in Section 7.4.3 to determine the heat transfer coefficient for the ventilation H_{ve} . The heat *transmission coefficient to the ground* $H_{T,ig}$ was calculated using the method from [71]:

$$H_{T,g} = A_f \times U_{equiv} \times f_{g1} \times f_{g2} \quad (7.12)$$

where $f_{g1} = 1.45$, $f_{g2} = (\Theta_{int,i} - \hat{\Theta}_e) / (\Theta_{int,i} - \Theta_e)$ and U_{equiv} from Table 7.11 using $B' = \frac{A_g}{2P}$. The heat *transmission coefficient to the outside* was calculated as:

$$H_{T,e} = \vec{A} \cdot (\vec{U} + \Delta \vec{U}_{WB}) \quad (7.13)$$

where A and U are vectors containing the areas and U-values of the buildings parts' (*i.e.* walls, windows, doors and ceiling), respectively. ΔU_{WB} is a generic thermal bridge correction factor which is set to 0.05 according to [71]. The U-values used for the calculation of $H_{T,g}$ and $H_{T,e}$ are the same as used in the previous section (*cf.* Table 7.8).

7.4.3 Ventilation losses

Ventilation heat losses occur due to cracks or small openings in the building envelope (natural ventilation) and the need to regularly exchange the air to increase the comfort of the inhabitants (hygienic ventilation). In the following section, we will discuss how we calculated the ventilation heat losses according to the DIN EN 12831 model.

Natural ventilation mainly depends on the tightness of the building envelope, the exposure of the building and the wind speed. Some of the heat losses from natural ventilation may be thus recovered by making windows and doors draught-proof. On the other hand, heat losses due to hygienic ventilation arise from the need for inhabited spaces

Table 7.11: Equivalent transmission coefficient U_{equiv} for buildings without cellar. Table from [71].

$B'[\text{m}]$	$U_{\text{equiv}} (\text{W}/(\text{m}^2 \text{ K}))$				
		$U_{\text{groundplate}} (\text{W}/(\text{m}^2 \text{ K}))$			
	no insulation	2.0	1.0	0.5	0.25
2	1.30	0.77	0.55	0.33	0.17
4	0.88	0.59	0.45	0.30	0.17
6	0.68	0.48	0.38	0.27	0.17
8	0.55	0.41	0.33	0.25	0.16
10	0.47	0.36	0.30	0.23	0.15
12	0.41	0.32	0.27	0.21	0.14
14	0.37	0.29	0.24	0.19	0.14
16	0.33	0.26	0.22	0.18	0.13
18	0.31	0.24	0.21	0.17	0.12
20	0.28	0.22	0.19	0.16	0.12

to be regularly ventilated to reduce the concentration of harmful gases such as carbon dioxide. Ventilation is also necessary to counter the humidity introduced by inhabitants (*e.g.* by breathing, cooking or taking a shower) – which may lead to mold. Therefore these losses cannot be avoided. The *Rechnagel*⁴ [160] minimum (hygienic) ventilation is defined as the minimum amount of air exchange required to maintain 1000 ppm CO₂.

For our model, we have calculated the heat losses due to ventilation H_{ve} using the simplified method from DIN EN 12831. In DIN EN 12831, H_{ve} is defined as the maximum of the hygienic minimum ventilation V_{min} and V_{inf} , the natural ventilation by infiltration. V_{min} is the volume of the room to be heated multiplied by a factor n_{min} . In this case, we chose $n_{\text{min}} = 0.5$ which is given by DIN EN 12831 as the standard value for an inhabited room. The heat loss due to infiltration is given by:

$$V_{\text{inf}} = 2 \times V_{\text{r}} \times n_{50} \times e \times \varepsilon \quad (7.14)$$

Here V_{r} denotes for the volume of the room, n_{50} is a factor for the tightness of the building. We set $n_{50} = 6$, which corresponds to a detached house with tight walls. The shielding factor e we set to 0.09 which corresponds to moderate shielding. We do not employ an elevation correction and thus set $\varepsilon = 1$. Since $V_{\text{inf}} = 2 \times V_{\text{r}} \times 6 \times 0.09 \times 1 = 1.08$, $V_{\text{inf}} > V_{\text{min}}$, which means that the hygienic air flow is already guaranteed by the natural ventilation, $H_{\text{ve}} = V_{\text{inf}}$. The ventilation heat transfer coefficient is thus equivalent to the losses due to natural ventilation.

⁴The *Taschenbuch für Heizung + Klimatechnik* or *Rechnagel* in short is the standard reference for heating and cooling in the German language. First issued in 1897, it is currently in its 76th issue.

7.4.4 Internal gains

Internal heat gains are divided into gains due to *appliances* (e.g. dishwasher, washing machine, dryer and stove), *losses from the heating/cooling system* (e.g. pumps and fans) and the *metabolic heat from occupants*. We do not include internal gains due to appliances since we do not have accurate ground truth data to predict their operation. In addition, we assume that there are no losses from the heating system that may be recovered as part of the internal gains. We do, however, include internal gains due to the metabolic heat rate of the occupants. An average person produces 125 watts of heat [83]. For our simulation we have assumed the flat to be occupied by 2 people and the house by 3 people. Since the algorithms in our sample only consider binary occupancy, we assumed all occupants to be present whenever the property was occupied (*i.e.* $\Phi_{\text{int}} = 250$ watts and $\Phi_{\text{int}} = 375$ watts whenever the flat or house are occupied, respectively).

7.4.5 Solar gains

While solar gains are most important when assessing the need for air-conditioning and ventilation in summer, the sun's radiation must also be taken into account to properly assess the energy needed for heating in winter. Such is the importance of solar gains that the field of passive solar building design focuses on using solar gains in winter and avoiding those gains in summer through the placement of windows, shading and insulation [147].

An increase in temperature through solar gains is the net result from two different processes: (a) long wavelength radiation being trapped inside the building and (b) an increase in the temperature of the building envelope through absorbed sunlight. Here we will focus on the former – solar gain through transparent building parts. When a building material such as glass is more transparent to the shorter wavelengths (visible light) than the longer (infrared radiation), heat is trapped inside the room. This is due to the fact that the room is heated up and any re-emitted (infrared) radiation cannot escape through the windows. This effect is best exhibited in green houses and occurs at a smaller scale in residential buildings.

The incident solar radiation causing the internal gains may be divided into two categories: Direct and diffuse radiation. Direct radiation are direct line-of-sight rays from the sun, while diffuse radiation is light reflected from the surroundings. The direct radiation is the reason why, in the northern hemisphere, buildings are often constructed with windows facing south to maximise exposure to the sun.

The weather data used in the following experiments contains the global (sum of direct and diffuse) radiation on a horizontal surface. To calculate the solar gain through the windows, we must first obtain the position of the sun relative to our location. We then divide the global radiation into direct and diffuse radiation using the Reindl* method [72].

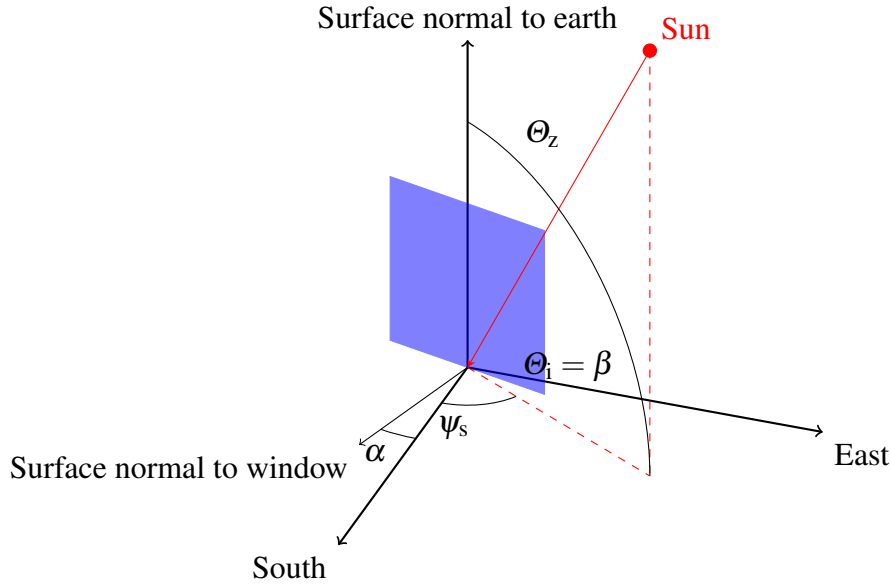


Figure 7.9: Position of the sun. Θ_z is the solar zenith angle. Θ_i is the angle between the normal to the window and the sun. In our case the normal to the surface of the window is assumed to be perpendicular to the normal to the surface of the earth and therefore β , the solar elevation is $\beta = 90 - \Theta_z = \Theta_i$. α is the orientation of the window with zero degrees being south.

Finally, we transform the direct radiation from the horizontal to a vertical plane and thence to the different orientations of the windows.

Position of the sun

In order to correctly compute the incident solar radiation on the windows, we must find the sun's current position in terms of its elevation β and azimuth angle ψ_s . The sun's elevation is at its maximum over noon. Figure 7.9 shows a simplified version of the problem. In this case the normal to the surface of the window is assumed to be perpendicular to the normal to the surface of the earth (*i.e.* there is no tilt or rotation of the window in the vertical plane). β and ψ_s are calculated as follows:

$$\sin \beta = \cos \text{lat} \cos \delta \cos \omega + \sin \text{lat} \sin \delta \quad (7.15)$$

$$\sin \psi_s = \cos \delta \frac{\sin \omega}{\cos \beta} \quad (7.16)$$

The calculation of β and ψ_s needs the local apparent solar time of the current position. This is due to the fact that the Earth follows an elliptical orbit and that its axis is not perpendicular to the plane of that orbit. The result is that the mean solar time (clock time)

Table 7.12: Sun angle parameters.

Variable	Description
lat	latitude of building
δ	$23.45 \sin(284 + n) \frac{360}{365}$, where n = day of the year
ω	$0.25(\text{suntime} - 720)$

Table 7.13: Parameters for the calculation of the local apparent solar time.

Variable	Description
time	local time (CET)
lon	longitude of building
lon _{stored}	standard meridian of building / CET (15)
$4(\text{lat} - \text{lat}_{\text{stored}})$	constant deviation (4 minutes per degree)
E	$9.87 \sin 2B - 7.53 \cos B - 1.5 \sin B$
B	$(n - 81) * 360/364$, where n = day of the year

does not accurately reflect the current position of the sun. The local apparent solar time *suntime* is calculated as follows using the variables defined in Table 7.13:

$$\text{suntime} = \text{time} + 4(\text{lat} - \text{lat}_{\text{stored}}) + E \quad (7.17)$$

Direct and diffuse radiation: Reindl*

With the current elevation β of the sun, we can split the global radiation obtained from the weather data into direct and diffuse components. As the Pully station does not provide detailed information on the cloud cover including the type of clouds, position and number of layers, we must use a decomposition model to determine incident direct and diffuse radiation [72]. The Reindl* is a piecewise regression (*cf.* Equation 7.18) to compute the relationship between the diffuse radiation I_d and the global radiation I with respect to a clearness factor k_t . The clearness factor depends on the extraterrestrial radiation (solar energy) I_0 , the current global radiation I as measured by the weather station as well as Θ_z , the angle between the zenith and the sun. The method is described in detail in [72]. The parameters are listed in Table 7.14.

$$I_d/I = \begin{cases} 1.020 - 0.248k_t & \text{if } 0 \leq k_t \leq 0.3, I_d/I \leq 1.0 \\ 1.400 - 1.749k_t + 0.177 \sin \beta & \text{if } 0.3 < k_t < 0.78, 0.1 \leq I_d/I \leq 0.97 \\ 0.147 & \text{if } k_t \leq 0.78 \end{cases} \quad (7.18)$$

Table 7.14: Reindl* parameters.

Variable	Description
I	Global radiation as measured by the weather station
I_d	Diffuse radiation on the horizontal surface
I_b	Direct radiation on the horizontal surface
I_0	Extraterrestrial radiation / solar energy in W/m^2
I_0	$1356.5 + 48.5 \cos(0.01721 * (n - 15))$
k_t	$\frac{I}{I_0 \cos \Theta_z}$, Clearness factor $0 \leq k_t \leq 1$
Θ_z	$90 - \beta$, angle between zenith and sun

Solar radiation on vertical surfaces

From the incident direct solar radiation on the vertical plane, we can now compute the direct solar radiation on a vertical plane using the angles computed previously as follows:

$$I_{b,vert} = \frac{I_b}{\cos \theta_z} \cos \theta_i \quad (7.19)$$

where

$$\cos \theta_i = -\sin \delta \cos lat \cos \alpha + \cos \delta \cos lat \cos \alpha \cos \psi_s + \cos \delta \sin \alpha \sin \psi_s \quad (7.20)$$

Here α is the clockwise orientation of the window with zero degrees being south (*cf.* Figure 7.9). For our model we assume that the house is positioned directly on the north south axis. This means that that the windows are facing directly to the north, east, south and west directions. In order to calculate the incident solar radiation on the windows we can use the following equation:

$$\Phi_{sol,x} = A_{W,x} \times g(0) \times (1 - \tan^4(\Theta_i/2)) \times \frac{I_b}{\cos \Theta_z} \times \cos \Theta_i \quad (7.21)$$

Here $A_{W,x}$ is the area of the windows ($x \in \{east, south, west\}$) and $g(0)$ is the g-value (solar transmittance) of the window. We set $g(0) = 0.6$, corresponding to double-glazed windows. I_b and Θ_z are the direct solar radiation on the horizontal plane and the angle between the zenith and the sun as previously calculated. Once we have obtained $\Phi_{sol,east}$, $\Phi_{sol,south}$ and $\Phi_{sol,west}$, Φ_{sol} is computed as the sum of the individual solar gains (*i.e.* $\Phi_{sol} = \Phi_{sol,east} + \Phi_{sol,south} + \Phi_{sol,west}$). Like the temperature Θ_e , the solar gains Φ_{sol} have been calculated at 15-minute intervals for all⁵ $\beta > 5$.

⁵The clearness factor k_t is only defined for $\beta > 5$ degrees due to the cosine in $k_t = \frac{I}{I_0 \cos \Theta_z}$.

Table 7.15: Controller parameters.

Symbol	Units	Description
Θ_{comf}	$^{\circ}\text{C}$	Comfort / set-point temperature
Θ_{setb}	$^{\circ}\text{C}$	Setback temperature
t	/	Current 15-minute timeslot
S	$\{1, 0\}$	Actual occupancy schedule
P_t	$\{1, 0\}$	Predicted occupancy schedule at interval t

7.5 Controller design

In order to act upon the predictions made by the occupancy prediction algorithms, we must translate their predicted occupancy schedules into an actual heating schedule containing setpoint temperatures. As the heating system cannot reach the target comfort temperature immediately, these setpoint temperatures have to be chosen so to reach a comfortable temperature upon the occupants' arrival (*e.g.* in order to reach a comfortable temperature upon the arrival of the occupants at 5 p.m. we might have to set the setpoint temperature to the comfort temperature at 3 p.m. already).

Algorithm 1 shows the high-level⁶ controller used to alternate between setpoint Θ_{comf} and setback Θ_{setb} temperatures. The rationale behind our approach is that by simulating the time it takes to heat up the property to a comfortable temperature, we can decide if the predicted schedule gives us enough time to forgo heating for another timestep. For each 15-minute time interval t , the controller looks at the current occupancy S_t of the household given by the occupancy schedule S at time t . If the household is currently occupied, we must keep the setpoint temperature and therefore we set $\Theta_{\text{int,H,set}}$ to Θ_{comf} . If the household is not occupied at time t , we use the predictive policy. The predictive policy first looks at the current prediction from time t onwards P_t and finds the number of intervals until the next occupied interval. It then computes the next indoor temperature $\Theta_{\text{air,noheat}}$, which would result from using the setback temperature for the current interval. Finally, it computes the number of intervals it will take to heat from $\Theta_{\text{air,noheat}}$ to the setpoint temperature Θ_{comf} . If this number is larger than the number of intervals to the next occupied timeslot, we must heat at the current time t .

A reactive policy is obtained if for all times t_1 and all predicted intervals t_2 at these times, the building is predicted to be unoccupied – or more formally: $\forall t_1, t_2 : P_{t_1, t_2} = 0$. Similarly, an always-on policy is obtained if for all times t during the simulation the building's occupancy is set to 1 – formally $\forall t : S_t = 1$. In the appendix, we show the behaviour of the control algorithm for all weather scenarios for a building unoccupied from 9 a.m. to 5 p.m. in terms of the heating setpoint $\Theta_{\text{int,H,set}}$ and indoor air Θ_{air} temperatures as well as the heat input – $\Phi_{\text{HC,nd}}$.

⁶Further to this, there is an internal controller inside the RC model that regulates the heat input to obtain the desired target temperature at each timestep.

Algorithm 1 Control algorithm.

```

1: procedure CONTROLLER
2:    $t \leftarrow$  Current time interval
3:    $S \leftarrow$  Actual occupancy schedule
4:    $P_t \leftarrow$  Predicted occupancy schedule at interval  $t$ 
5:   Reactive policy:
6:   if isOccupied( $S_t$ ) then
7:      $\Theta_{\text{int,H,set}} \leftarrow \Theta_{\text{comf}}$ 
8:   else
9:     Predictive policy:
10:     $n_{\text{horizon}} \leftarrow \text{nextOccupied}(P_t)$ 
11:     $\Theta_{\text{air,noheat}} \leftarrow \text{iso13790}_{5\text{RIC}}(\Theta_{\text{setb}}, \Theta_{m,t-1}, \dots)$ 
12:     $n_{\text{preheat}} \leftarrow$  Preheat time from  $\Theta_{\text{air,noheat}}$  to  $\Theta_{\text{comf}}$ 
13:    if  $n_{\text{horizon}} \geq n_{\text{preheat}}$  then
14:       $\Theta_{\text{int,H,set}} \leftarrow \Theta_{\text{setb}}$ 
15:    else
16:       $\Theta_{\text{int,H,set}} \leftarrow \Theta_{\text{comf}}$ 
17:     $\Theta_{m,t}, \Phi_{\text{HC,nd,t}}, \Theta_{\text{air,t}} \leftarrow \text{iso13790}_{5\text{RIC}}(\Theta_{\text{int,H,set}}, \Theta_{m,t-1}, \dots)$ 
18:     $t \leftarrow t + 1$ 
19:  goto Reactive policy.

```

Obtaining a steady-state When we first start the simulation, we do not know the value for $\Theta_{m,0}$. As Θ_m – the temperature of the building mass – is different from the indoor air temperature Θ_{air} (cf. Figure 7.4), we cannot set $\Theta_{m,0} = \Theta_{\text{setb}}$. We thus first obtain a steady state temperature for $\Theta_{m,0}$ by repeatedly running $\text{iso13790}_{5\text{RIC}}(\Theta_{\text{setb}}, \Theta_{m,t-1}, \dots)$ with the same environmental parameters until $\Theta_{m,t}$ converges. For the initial value of $\Theta_{m,t}$ we use Θ_{setb} .

7.6 Limitations

Mathematical models of complex physical systems are usually simplified renditions of their real-world counterparts. Translating all variables influencing the operation of a heating system is not feasible. Mathews *et al.* summarise this problem as follows:

“Simplified hypotheses, or mathematical models, idealize reality because of the impossible task of accounting for every detail in complicated real life phenomena.” [129]

No current approach to simulate the behaviour of smart heating systems thus accounts for all possible variables involved. Therefore, as discussed in this chapter, we have made a number of simplifying assumptions.

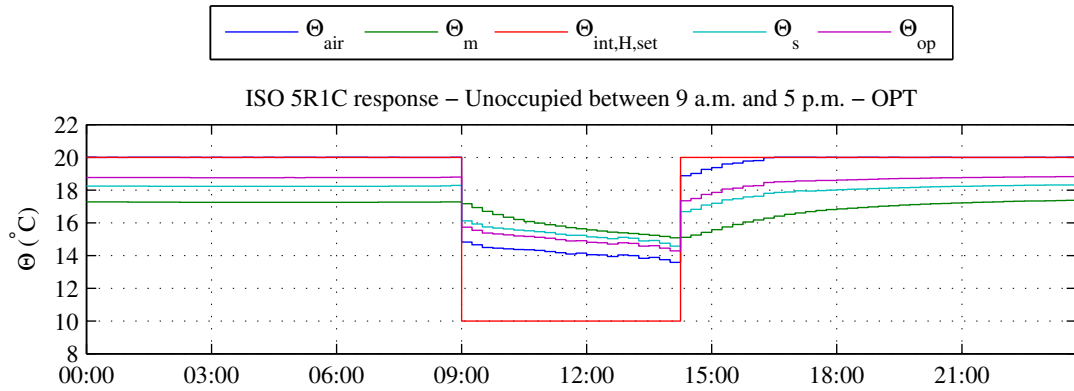


Figure 7.10: Temperature response of ISO 5R1C model.

Weather and occupancy To be able to investigate the effects of occupancy and weather on the overall energy required, we have evaluated both *ceteris paribus*. That is, we have assumed that there is no correlation between weather conditions and household occupancy. Additional work is needed to establish how the current weather conditions might affect occupancy (*e.g.* when the weather is good, occupants may stay outdoors for longer).

Building configurations We have also made a number of simplifying assumptions in the design of our four sample buildings F-U_{low}, H-U_{low}, F-U_{high} and H-U_{high}. For all buildings we assume that they are free-standing, contain a single zone and that there is no heat transfer to and from adjacent buildings. We further used simplified facades and did not take into account the effects of overhangs or other means of shading. We argue that, as there is no single “example” house that can be used, these simplifications are valid to understand the general effect of smart thermostats and to be able to compare their efficacy in small vs. large and well vs. poorly insulated buildings.

ISO 13790 5R1C model Figure 7.10 shows the temperatures⁷ simulated by the ISO 5R1C model for a building occupied from midnight to 9 a.m. and again from 5 p.m. to midnight (well insulated flat). The heating setpoint temperature $\Theta_{\text{int,H,set}}$ is lowered to the setback temperature at 9 a.m. ($\Theta_{\text{setb}} = 10^\circ\text{C}$) in order to save energy. To reach a comfortable temperature (*i.e.* $\Theta_{\text{air}} = \Theta_{\text{comf}}$) upon the arrival of the occupants, $\Theta_{\text{int,H,set}}$ is set to the comfort temperature ($\Theta_{\text{comf}} = 20^\circ\text{C}$) at 2.15 p.m. – this “optimal” control strategy is described in more detail in Section 7.5.

The development of the indoor air temperature Θ_{air} in Figure 7.10 is noteworthy. As the heat input from the heating system is reduced to zero at 9 a.m., Θ_{air} drops by over 5°C . Similarly, there is a significant jump in Θ_{air} when the heat input is increased again at

⁷That is the mean air temperature, Θ_{air} , the mean radiant temperature Θ_{m} , the heating setpoint temperature $\Theta_{\text{int,H,set}}$, the temperature of the surface node Θ_{s} and the mean operating temperature Θ_{op} (*cf.* Table 7.2).

Table 7.16: Example parameters for the calculation of Θ_{air} .

Term	Building variant				Units
	F-U _{low}	F-U _{high}	H-U _{low}	H-U _{high}	
$H_{\text{tr,is}}\Theta_{\text{s}}$ (Heat recovered from surface node)					
$H_{\text{tr,is}}$	1044	1044	3564	3564	W / K
$H_{\text{tr,is}} \times 16^{\circ}\text{C}$	16.70	16.70	57.02	57.02	kW
$H_{\text{tr,is}} \times 18^{\circ}\text{C}$	18.79	18.79	64.15	64.15	kW
$H_{\text{ve}}\Theta_{\text{sup}}$ (Heat loss by ventilation)					
H_{ve}	47.33	47.33	161.57	161.57	W / K
$H_{\text{ve}} \times 0^{\circ}\text{C}$	0	0	0	0	kW
$H_{\text{ve}} \times -5^{\circ}\text{C}$	-0.24	-0.24	-0.81	-0.81	kW
$H_{\text{ve}}\Theta_{\text{sup}}$ (Maximal heat gains when occupied)					
Φ_{ia}	0.13	0.13	0.19	0.19	kW
$\Phi_{\text{HC,nd}}$ ($\Phi_{\text{H,max}}$)	2.80	6.86	7.78	16.75	kW

2.15 p.m. The circuit diagram in Figure 7.4 shows that the heat source $\Phi_{\text{HC,nd}}$ is directly connected to the node for the indoor air temperature Θ_{air} . Θ_{air} loses heat to the outside while heating the supplied air. It further exchanges heat with the building surface. The capacitor C_m , which stores some of the heat is located at the other side of the circuit. While the function for Θ_m is continuous and shows no jumps in the temperature, this arrangement causes the function for Θ_{air} to be discontinuous. This can also be seen from the calculation of Θ_{air} (cf. Equation C11, ISO 13790 [84]). Here, Θ_{air} is derived from Θ_m via Θ_s :

$$\Theta_{\text{air}} = (H_{\text{tr,is}}\Theta_s + H_{\text{ve}}\Theta_{\text{sup}} + \Phi_{\text{ia}} + \Phi_{\text{HC,nd}})/(H_{\text{tr,is}} + H_{\text{ve}}) \quad (7.22)$$

Table 7.16 shows an exemplary calculation of the parameters of the above equation at 9 a.m. The coupling conductance $H_{\text{tr,is}}$ between the air node Θ_{air} and the surface node Θ_s is given by $H_{\text{tr,is}} = h_{\text{is}}A_{\text{tot}}$ where $h_{\text{is}} = 3.45$ (cf. 7.2.2.2 of ISO 13790). A_{tot} is the area of all surfaces facing the building zone and calculated as $A_f \times 4.5$. As $H_{\text{tr,is}}$ is fixed for each building configuration, the first term of the summation only varies with Θ_s . When Θ_s is 18°C (9 a.m.), the heat from the inner surface of the building is 19 kW for the flat and 64 kW for the house. As Θ_s drops to 16°C after the heating is switched off, this heat is reduced to 17 kW and 57 kW for the flat and house, respectively. In both cases, this is a decrease of 11% in the energy supplied by the walls to heat Θ_{air} .

H_{ve} is the ventilation heat transmission coefficient and Θ_{sup} is the temperature of the supplied air – in our case $\Theta_{\text{sup}} = \Theta_e$. Thus, the second term only varies with the outside temperature. If we assume that there is no significant difference in the outside temperature between 9 a.m. and 9.15 a.m., this term does not contribute to the steep drop in the indoor air temperature.

Φ_{ia} is defined as half of the internal gains. $\Phi_{\text{HC,nd}}$ is the heat input, bounded above by the design heat load. The drop in temperature occurs when the building is unoccupied after

9 a.m. as both $\Phi_{\text{HC,nd}}$ and Φ_{ia} become zero⁸. Table 7.16 shows that Φ_{ia} only contributes 0.13 kW to 1.19 kW to the overall gains. In contrast, the impact from switching off the heating system is much larger. Depending on the level of insulation and the size of the building, between 2.8 kW and 16.75 kW are removed entirely from the system at 9.15 a.m. Especially for the poorly insulated buildings, the heat supplied directly by the heating system (rather than through the surface node) has a substantial effect on the indoor air temperature. For the poorly insulated flat, the heating system can supply 7 kW, while the walls supply 19 kW at 18°C and 17 kW at 16°C. Removing this heat from the system entirely, invariably produces the significant drop in temperature after 9 a.m.

To lengthen the ramp-up time and in order for it to more accurately reflect the heating behaviour of hydronic heating systems common in Europe, the 5R1C model should be extended to model the lag caused by boilers and radiators. A hydronic heating system needs to heat up the water in circulation first, before the heat can be transmitted to the indoor air and the structure of the building via radiation. Similarly, the temperature in the radiators does not drop immediately when the setpoint temperature is lowered. Pipes, radiators and the boiler may thus constitute an additional RC circuit, separated from the rest of the building. However, in absence of a better model, the 5R1C model gives us a good first indication of the effects of a smart thermostat based on occupancy prediction.

7.7 Conclusions and lessons learned

In this chapter we have presented a predictive controller and a method for calculating the heating energy consumption of a building based on the ISO 13790 5R1C model. In order to assess the effect of different building characteristics, we have modelled a flat and a house, both with poor and good insulation levels. We have furthermore investigated the effect of weather and climate conditions by introducing characteristic weather scenarios for the Lausanne area. We have shown how these scenarios can be used to determine the annual energy savings of a smart heating system. We will use this modelling framework to analyse a number of smart heating strategies in Chapter 8.

⁸Depending on the setback and outside temperatures, $\Phi_{\text{HC,nd}}$ might stay above zero.

Smart Thermostats: How much do they save?

The main goal of a smart heating system is to save energy while maintaining occupant comfort. The system automatically sets the temperature of a building based on its current occupancy state and a prediction of the future occupancy states of the building. The occupancy prediction algorithms introduced in Chapter 6 thereby aim to reduce the energy consumed by heating, while at the same time avoiding any loss of comfort for the residents. We have seen that state-of-the-art schedule-based occupancy prediction algorithms provide a prediction accuracy around 85% while their performance is limited by the predictability of the participants' schedules. However, although inaccurate predictions may lead to thermal discomfort or a waste of energy, the performance of such a smart heating system is not only dependent on said algorithms. The efficiency of a heating system is also determined by:

- **The actual occupancy.** Occupancy is an important factor in assessing the performance of a smart heating system. Occupancy can vary between different households in absolute terms (*e.g.* one household may have a higher overall occupancy) and in terms of occupancy distribution (*i.e.* variations in the duration of occupied periods). Both the overall occupancy and its distribution influence the savings of a smart heating system.
- **The parameters of the building.** The materials used in and the composition of the building envelope determine the efficacy of the heating system. The ability of the building to store energy as well as its ability to retain it influence the energy required to heat the building. Furthermore, the parameters of the building have a large influence on the preheat times and therefore the prediction horizon. If a

building requires a longer time to heat up, the smart heating system needs to be able to predict further into the future. This increases the potential for wrong predictions, which could lead to smaller energy savings and higher comfort loss.

- **The weather conditions.** Weather arguably plays the most important role as heating is the necessary is only necessary due to changes in the outside temperature. However, the environmental conditions can also substantially lower the energy required by the heating system as solar gains help to increase the indoor temperature.
- **The control strategy.** Especially for large and complex commercial buildings, the controller employed to regulate the heating, ventilation and cooling infrastructure is an important factor for the total energy consumption. Without considering occupancy, a more efficient control strategy can for example save energy by maximising solar gains through intelligent control of the blinds.

In this final chapter we combine the work of Chapters 5 to 7 to investigate the tradeoff between achievable savings and the risk of comfort loss of a smart thermostat for household residents. We evaluate a smart thermostat with a predictive heating controller that uses the MAT, MDMAT, PP(S) and PH algorithms for occupancy prediction. To this end, we utilise the occupancy schedules derived from the LDCC dataset as discussed in Chapter 5, the implementations of the occupancy prediction algorithms from Chapter 6 and the simulation model for smart thermostats from Chapter 7. We will thus analyse the performance of five schedule-based prediction algorithms in 32 different building and weather scenarios and report on the projected annual savings of a smart thermostat based on occupancy prediction.

The structure of this chapter will be as follows. In Section 8.1 we will give an overview of existing approaches that automatically reduce the energy consumption of space heating. We will then describe the setup of our experiment (*cf.* Section 8.3) and discuss the relevant metrics in Section 8.4. We will outline our results in Section 8.5 and conclude in Section 8.7. This chapter is based on contributions made in [103], [104] and [105].

8.1 Related work

Commercial and residential buildings account for a large fraction of the total energy consumption. In the United States, buildings account for over 40% of the primary energy consumption [221]. Their consumption is dominated by HVAC which accounts “for close to half of all energy consumed in the buildings sector” [221]. Various studies have

shown that the operation of HVAC systems can be optimised to reduce the overall energy footprint of buildings [55, 128, 148, 181]. In commercial buildings there is often little to no correlation between the energy consumed by the HVAC system and occupancy [128]. Similarly, many residential households fail to achieve potential savings as programmable thermostats are too difficult to use [143, 158]. For this reason, a number of authors have looked into automatically controlling heating, ventilation and cooling systems to save energy without sacrificing occupant comfort [67, 113, 124, 150, 169].

8.1.1 Model predictive control

Automating the regulation of processes such as space heating by means of control theory without the direct intervention of humans is conventionally referred to as *automatic control*. Automating heating control has been the focus of the research community for some years.

Most notable among current control strategies is the concept of model predictive control (MPC) [62]. In contrast to traditional proportional-integral-derivative (PID) controllers, which react to changes in a control variable, MPC-based controllers are able to incorporate predictions of future events when taking control decisions. These predictions are based on dynamic models of the physical system that are usually obtained by system identification. A smart heating system based on occupancy prediction is an example of such a model predictive system. In order to find the appropriate time to start preheating the building, its future occupancy must be predicted. If it is predicted to be occupied in the near future, a thermal model of the building in conjunction with the projected weather conditions must be used to compute the right setpoint temperatures for the next time intervals.

Many authors have investigated the performance of MPC approaches to reduce the energy consumption of space heating [23, 55, 64, 148, 150, 178, 181]. Oldewurtel *et al.* use stochastic model predictive control in conjunction with uncertain weather forecasts to automatically control blinds and ventilation as well as heating and cooling in order to improve the climate control in commercial buildings [148]. Similarly, Široký *et al.* [181] present an MPC-based control system that uses weather predictions to automatically control the heating and cooling in an office building in Prague, Czech Republic. The authors obtained energy savings between 15% and 28% but note that these savings cannot be generalised as they are dependent upon many factors such as the building parameters and weather conditions. Ferreira address the system identification challenge of MPC (*i.e.* obtaining the model of the building that can be used to predict future behaviour) by the use of artificial neural networks [55] and project energy savings exceeding 50% in a university building.

Model predictive control has so far mainly focussed on large commercial buildings. An exception is the work by Vasak *et al.*, who evaluate the use of an MPC controller to control heating and cooling in a residential household [178]. The authors use an RC model as part of their control strategy and show using the TRNSYS simulation [219] that this choice is a

good fit. The authors do not integrate occupancy prediction and do not evaluate the energy consumption of their approach. Similarly, Rogers *et al.* address the need of residential households for a cheap off-the-shelf heating solution by using MPC to control the radiator valves of a common hydronic heating system [165]. The authors use MPC to improve temperature regulation and focus on an increase in thermal comfort rather than energy savings.

MPC using occupancy data In a follow-up work to [148], Oldewurtel *et al.* compare the use of long-term occupancy predictions to a fixed occupancy schedule as input to a model predictive controller and conclude that a “large part of [the] potential [of occupancy prediction] can already be captured by taking into account instantaneous occupancy information” [150]. However, in contrast to our work, Oldewurtel *et al.*, focus on “long-term vacancies” such as business trips, holidays or illnesses in commercial buildings instead of the daily occupancy patterns in residential households. Goyal *et al.* investigate MPC-based heating control with short prediction horizons and perfect occupancy predictions and find that current ventilation standards¹ prevent significant savings in commercial buildings [64]. Other authors use more complex models for occupancy prediction. Using Erickson *et al.*’s occupancy prediction algorithms (*cf.* Chapter 6), Beltran *et al.* build an MPC-based controller and obtain 9.4% savings for heating in a commercial building [23].

8.1.2 Other control strategies

Traditionally, research in automatic control is produced in civil, mechanical and electrical engineering as well as architecture departments. With the advent of mobile phones and wireless sensor networks, computer science researchers focussing on exploiting advances in information and communication technology to save energy [131] have recently started to discover this field. This has resulted in a number of papers on smart heating systems [6, 48, 67, 67, 76, 118, 124, 140, 169]. In contrast to prior work in the area of automatic control, these approaches focus on fine-grained occupancy sensing and prediction techniques and do not analyse other improvements to the heating control infrastructure.

This occupancy-centric approach has implications on the generalisability of results. Most publications only include a cursory evaluation of energy savings over a small number of non-standard scenarios. Most approaches thereby fail to take into account other environmental factors such as solar and internal gains.

As the respective environmental and baseline data vary significantly, it is difficult to compare the results of individual approaches. Moreover, while some authors based their findings on simulations [48, 67, 76, 124, 140] others used real world deployments to

¹ASHRAE ventilation standard 62.1-2010 prescribes ventilation even during unoccupied times [14].

Table 8.1: Savings achieved in the various projects. S and D denote values gathered in simulation and deployment respectively. †Derived value. ‡Combined savings.

Authors	Environment (S/D)	Savings	Baseline
Mozer <i>et al.</i> [140]	Home (S)	12%†	Always-on
Gupta <i>et al.</i> [67]	Home (S/D)	3.4%	Always-on
Lu <i>et al.</i> [124]	Home (S)	27.9%	Always-on
Erickson <i>et al.</i> [47, 48]	Office (S)	30% and 42%	Scheduled
Agarwal <i>et al.</i> [6]	Office (D)	7.7%†	Scheduled
Scott <i>et al.</i> [169]	Home (D)	8% and 18%	Scheduled
Hong <i>et al.</i> [76]	Office (D)	8% to 28%	Always-on
Beltran <i>et al.</i> [23]	Office (S)	9.4%	Scheduled
Lee <i>et al.</i> [118]	Office (D)	up to 25%	Always-on

measure potential savings [6, 118, 169]. Although actual deployments show the feasibility of the approach, it is difficult to compare different strategies as one cannot replicate the environmental conditions.

The main goal of this chapter is to use the standard simulation techniques introduced in Chapter 7, including a number of representative weather and building scenarios, to evaluate the savings potential of existing state-of-the-art schedule-based occupancy prediction algorithms (*cf.* Section 6.1.1). However, before we go on to describe our own experimental setup in Section 8.3, we will briefly describe how previous work has evaluated the energy savings of smart heating systems using occupancy detection and prediction. Table 8.1 shows an overview of the discussed approaches.

Simulated savings

Lu *et al.* evaluate the energy savings of their ST occupancy prediction algorithm in a residential home (*cf.* Section 6.1.1) in the EnergyPlus simulator [220] over 14 days in winter. The authors show that an optimal prediction yields 35.9% savings on average over maintaining a constant temperature. In contrast, ST achieves 27.9% savings. The results for the ST algorithm are quite specific to the used three-stage boiler. In a later work, Hong *et al.* build upon these results and utilising the same heating infrastructure – with a slightly different prediction algorithm – achieve savings between 8% and 28% [76].

To compute the energy savings of their NT algorithm (*cf.* Section 6.1.1), Mozer *et al.* use a simple 1R1C model (*cf.* Section 7.2) of an old schoolhouse [140]. The mean daily cost of the ST approach in U.S. dollars was then generated from “three training/test splits [] formed by training on five consecutive months and training on the next month”. In contrast to the work by Lu *et al.*, Mozer *et al.* thus do not compute percentage savings. Instead, the authors convert lost comfort to a monetary value and combine it with the heating costs. To be able to compare the savings of the NT algorithm to the other approaches, we have computed the percentage savings shown in Table 8.1 from the monetary values of the NT and constant temperature approaches in Table 2 of [140].

Gupta *et al.* evaluate their GPS thermostat (*cf.* Section 6.1.2) over the course of a 14 day period in a single household. Unlike Lu *et al.* and Mozer *et al.*, the authors focus on cooling rather than heating. Gupta *et al.* simulate the savings of controlling the temperature using their GPS-based thermostat. These simulations resulted in 3.4% savings over the course of the observation period.

Erickson *et al.* use Markov Models to predict the occupancy level in a commercial building (*cf.* Section 6.1.4) and compare the savings of their algorithm to a scheduled baseline strategy “assuming maximum occupancy for ventilation and conditions all rooms from 7:00 – 22:00” [48]. After modelling and simulating the building in EnergyPlus [220], the authors achieve 30% and 42% energy savings on average [47, 48].

Savings in real world deployments

A small number of authors also shows savings estimates from real world deployments. The longest and most comprehensive deployment was done by Scott *et al.* in five homes over an average period of 61 days per house [169]. Scott *et al.* deployed their system in three U.S. households and two households in the United Kingdom. Only the two UK households and one U.S. household yielded energy savings of 8%, 18% and 2%. The other two U.S. households showed increases in the energy consumption by 5% and 1%.

Deployments in commercial buildings also exist. In [118], Lee *et al.* report savings above 25% for a university building. Similarly, Agarwal *et al.* present the implementation of a reactive control system that reduced the energy consumption by an average of 7.7% over the two day deployment in a university building [6].

8.2 Discussion

Table 8.1 shows an overview of recent occupancy-aware smart heating approaches. For homes, the savings achieved by occupancy detection and prediction algorithms range from 3.4% [67] to 28% [124]. For office buildings the approaches yield between 8% and 42%. Due to the limited description of the setup of most experiments is not clear whether the large difference between approaches is caused by superior algorithms, differing baseline results or varying environmental conditions. In the remainder of this thesis we will try to address this issue by simulating the energy savings and comfort loss of the selected state-of-the-art schedule-based algorithms introduced in Chapter 6 on a number of representative scenarios.

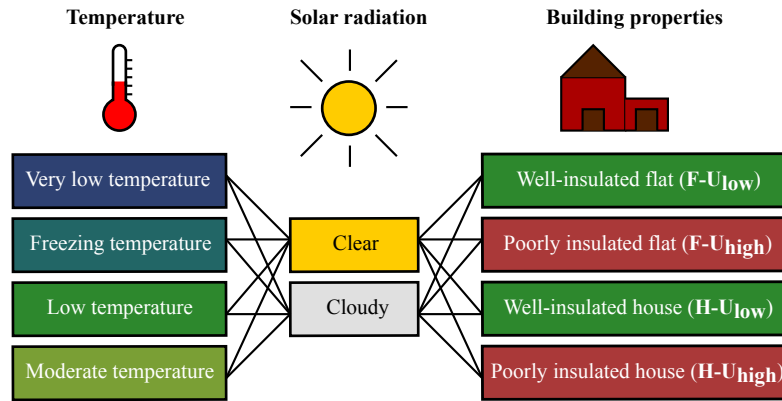


Figure 8.1: Simulated scenarios.

8.3 Experimental setup

The algorithms analysed in this chapter aim to predict occupancy for smart heating control systems. The goal of such systems is to reduce the energy consumed by heating, while at the same time avoiding any loss of comfort for the residents. We therefore assessed the suitability of the prediction algorithms in terms of their ability to save energy and ensure comfortable temperatures when required. To this end, we built a predictive controller to control the temperature of a building based on the current occupancy state and the algorithms' predictions of the future occupancy states of the building. In order to analyse the performance of the controller under different conditions, we ran simulations using the 5R1C thermal building model from the ISO 13790 energy performance standard [84] on 32 different scenarios. In particular, we analysed the influence of different weather conditions, building sizes and insulation levels.

In the following section we will give a high-level overview of our experimental setup. For a detailed description of the occupancy prediction algorithms we refer the reader to Chapter 6. The simulation model is described further in Chapter 7.

8.3.1 Building model and simulation setup

The ISO 5R1C model simulates the transient heat conduction between the property and its surroundings using an analogous electrical resistance-capacitance (RC) circuit and thus offers a method of calculating the energy required for heating and cooling while maintaining specified setpoint temperatures. This modelling principle was first introduced by Beuken in 1936 [26] and has since been widely employed in building design [129]. In contrast to simpler models [140], the ISO 5R1C model takes into account the heat transfer by transmission and ventilation as well as solar and internal gains.

Figure 8.1 shows the simulated scenarios. The response of the heating system was simulated for 32 different weather and building settings. We considered two different building sizes – a 52 m² studio flat (F) and a 176 m² house (H). In order to measure the effect of the building envelope on thermal performance, we also simulated the response of the ISO 5R1C model for low and high U-values². The U-value (W/(m² K)) denotes the overall heat transfer coefficient of a building element. Elements with high U-values conduct more heat per unit temperature difference between the inside and outside. A building with high U-values is considered poorly insulated and thus leaking a significant amount of heat to the outside. For each of the resulting four building configurations (flat F-U_{low}, F-U_{high}; house H-U_{low}, H-U_{high}), the design heat load (maximum heat input) in watts $\Phi_{H,max}$ was determined using the DIN EN 12831 standard [42]. The internal gains Φ_{int} were assumed to be 250 watts and 375 watts, whenever the house was occupied, equivalent to the metabolic heat rate of two and three residents for the flat and house respectively.

The effect of different weather conditions on the heating load was captured by eight representative weather scenarios synthesised from real weather data for the Lausanne (Switzerland) area where also the data used to derive the occupancy schedules was gathered (*cf.* Chapters 5 and 6). Lausanne is situated within a transition zone between a humid oceanic climate zone and a continental temperate zone.

Figure 8.1 also shows the eight weather scenarios used in the evaluation. The scenarios cover four different temperature levels under clear as well as cloudy sky conditions. Each scenario consists of 24-hour vectors of the outside temperature and the direct solar radiation, replicated n times to reflect the number of days in the occupancy data. A detailed description of the methodology used to define the weather scenarios is described in Chapter 7.

8.3.2 Heating controller

We implemented a predictive heating controller to translate the occupancy schedules predicted by the algorithms into actual heating schedules. A heating schedule defines the *target indoor air temperature* $\Theta_{air,set}$ at 15-minute time intervals t . Given the predicted occupancy schedule and the RC model, the heating controller sets $\Theta_{air,set}$ to Θ_{comf} for t if: (1) The house is occupied at time t (reactive policy); (2) The house is expected to become occupied between $t + 1$ and $t + I^*$. The prediction horizon I^* is the time needed to raise the indoor air temperature Θ_{air} to Θ_{comf} (predictive policy), starting from the temperature at time $t + 1$, using the maximum available heating power $\Phi_{H,max}$ (DIN EN 12831 design heat load) and assuming that the target temperature was Θ_{setb} at time t . If neither of these

²The U-values for a well-insulated buildings (F-U_{low} and H-U_{low}) correspond to the maximum allowed U-values for new properties in Germany according to EnEV'14 [3]. For the poorly insulated buildings (F-U_{high} and H-U_{high}), we used a list of high U-values reported in [223].

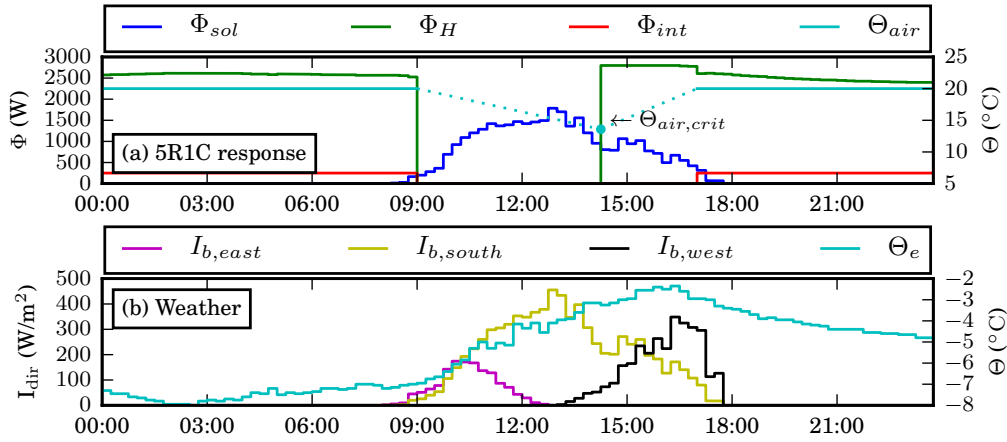


Figure 8.2: Typical behaviour of a heating system according to the ISO 5R1C model ($F-U_{low}$, very low temperature, clear sky) for a scenario where the house is unoccupied between 9 a.m. and 5 p.m. The upper part shows the inputs (solar gain Φ_{sol} , heat input Φ_H and internal gain Φ_{int}), the lower part the direct radiation $I_{b,\{east,south,west\}}$ and outside temperature Θ_e . $\Theta_{air,crit}$ denotes the critical temperature at which the preheating starts to reach Θ_{comf} at 5 p.m.

two conditions is fulfilled, the controller sets the target temperature to Θ_{setb} in order to save energy. The heat input Φ_H at any point in time is directly determined by the current setpoint temperature. In all cases, the controller has perfect knowledge³ of the future weather.

The predictive heating controller is always in one of three different states: the preheat state, the heating state or the cool down state. If the current air temperature is below the setpoint temperature $\Phi_{air,set}$, the controller is in the *preheat* state where the system heats with $\Phi_{H,max}$, the maximum heating power available. If the current air temperature is equal⁴ to the setpoint temperature, the controller is in the *heating* state. Here the heating power is lower than the maximum value and equivalent to the power needed to maintain the setpoint. Otherwise, if the setpoint is lower than the measured air temperature, the system is in *cool down* state and no heat is added to the system (*i.e.* $\Phi_H = 0$).

The upper part of Figure 8.2 shows the behaviour of the controller and the indoor air temperature Θ_{air} for a typical occupancy schedule and the $F-U_{low}$, freezing temperature, clear sky scenario. The lower part of the figure shows the corresponding weather data

³The alternative, predicting the future weather in order to determine when to heat, would prevent us from isolating the performance of the occupancy prediction algorithm.

⁴In practice “equal” is often taken with a grain of salt: To avoid excessive switching and to prevent wear of control equipment, controllers (in particular on-off systems) are typically designed to include hysteresis, effectively substituting the setpoint with a delta interval (the “comfort band”) around the setpoint.

($I_{b,\{\text{east,south,west}\}}$ indicating the direct solar radiation and the outside temperature Θ_e) used in this scenario. When the occupants leave at 9 a.m., the indoor air temperature is allowed to drop until 2.15 p.m. (from 20°C to 13°C), with no heat being added to the system. The controller then preheats the property such that $\Theta_{\text{air}} = \Theta_{\text{comf}} = 20^\circ\text{C}$ when the occupants return home at 5 p.m.

8.4 Evaluation

Having discussed the accuracy of schedule-based occupancy prediction algorithms in Chapter 6, we now investigate the performance of a predictive heating controller that uses the MAT, MDMAT, PP(S) and PH algorithms. For reference purposes we have also included OPT, which uses an oracle to provide a perfect prediction of household occupancy. To measure the energy consumption of the heating system, we built a simulation system in Chapter 7 based on the ISO 5R1C model. We assumed the heating controller behaves as described in Section 8.3.2, irrespective of the algorithm used to predict occupancy. We simulated the response of the controller for the four building variants (F-U_{low}, F-U_{high}, H-U_{low} and H-U_{high}) and eight weather scenarios described in Section 8.3.1, resulting in 32 different configurations.

8.4.1 Efficiency gain and comfort loss

We measured the performance of the controller for each algorithm in terms of *efficiency gain*. Let Q_{pred} be the heat injected by a predictive heating controller into the home and $Q_{\text{no_setback}}$ the corresponding heat injected by a controller that maintains the temperature of the home constantly at Θ_{comf} throughout the day. The *efficiency gain* is then defined as:

$$\text{Efficiency Gain} = (Q_{\text{no_setback}} - Q_{\text{pred}}) / Q_{\text{no_setback}} \quad (8.1)$$

Defining and measuring thermal discomfort in an appropriate way is not easy. In 1970, Gupta proposed using “the ratio of the temperature-time curve area outside the specified comfort zone to that area of the comfort zone” as a “degree of discomfort” [66]. We used a discretised variant of that measure which yields absolute values per day. *Discomfort degree hours* as a measure of comfort loss are defined as the average sum of hourly differences between the actual indoor air temperature Θ_{air} and Θ_{comf} for all occupied intervals, formally:

$$\text{Discomfort degree hours} = 1/4(\Theta_{\text{comf}}\Gamma_{1..96} - \Theta_{\text{air},1..96}) \cdot \Gamma_{1..96} \quad (8.2)$$

Here, $\Gamma_{1..96}$ denotes the ground truth occupancy vector containing 1’s for occupied intervals and 0’s for unoccupied intervals. Thus, if $\Theta_{\text{air}} = 17^\circ\text{C}$ upon the arrival of the

Table 8.2: ISO 13790 average efficiency gain for all experiments with **low U-values (good insulation)**. ☀ and ☁ denote *clear* and *cloudy* scenarios respectively. The rightmost column shows the average total daily energy consumption when no occupancy prediction and setback algorithm is applied.

Weather	Efficiency gain (%)														Σ kWh	
	OPT		MAT		MDMAT		PP		PPS		PH		REA		NO SETB.	
	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁
F-U_{low} (well insulated flat)																
Very low	5	4	4	2	4	2	4	2	4	2	4	3	13	14	51	55
Freezing	8	6	6	5	6	5	6	5	6	5	6	5	10	12	38	44
Low	10	9	8	8	8	8	8	8	8	8	8	8	10	12	27	32
Moderate	11	12	10	11	10	11	10	11	10	11	10	11	11	13	17	20
H-U_{low} (well insulated house)																
Very low	4	3	3	1	3	1	3	1	3	1	3	2	15	16	155	166
Freezing	6	5	4	4	4	3	5	3	4	3	5	4	10	12	119	134
Low	8	7	6	6	6	6	6	6	6	6	7	6	9	10	84	99
Moderate	9	10	8	8	8	8	8	8	8	8	8	8	9	10	53	65

Table 8.3: Same as Table 8.2, but with **high U-values (poor insulation)**.

Weather	Efficiency gain (%)														Σ kWh	
	OPT		MAT		MDMAT		PP		PPS		PH		REA		NO SETB.	
	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁
F-U_{high} (poorly insulated flat)																
Very low	10	9	9	9	9	9	9	9	9	9	9	9	11	11	123	124
Freezing	14	13	14	13	14	13	14	13	14	13	14	13	14	14	95	100
Low	16	17	16	17	16	17	16	17	16	17	16	17	16	17	69	74
Moderate	18	19	18	19	18	19	18	19	18	19	18	19	18	19	45	48
H-U_{high} (poorly insulated house)																
Very low	7	6	6	6	6	5	6	5	6	5	6	5	12	12	328	332
Freezing	11	10	10	9	10	9	10	9	10	9	10	9	13	13	255	269
Low	14	14	13	13	13	13	13	13	13	13	13	13	14	14	186	200
Moderate	15	15	14	15	14	15	14	15	14	15	14	15	15	15	122	133

occupants at 5 p.m. and the heating system requires 1 hour to heat up to $\Theta_{\text{comf}} = 20^\circ\text{C}$ (e.g. $\Theta_{\text{air},17:15} = 18^\circ\text{C}$, $\Theta_{\text{air},17:30} = 19^\circ\text{C}$, $\Theta_{\text{air},17:45} = 19.5^\circ\text{C}$ and $\Theta_{\text{air},18:00} = 20^\circ\text{C}$), then the discomfort degree hours for this day will be 0.75.

8.5 Results

Tables 8.2, 8.3 and 8.4 present the results for all 32 configurations. They show the efficiency gain and discomfort degree hours for all analysed algorithms. It is worth noting that the absolute values for the metrics reported clearly depend on the specific model, data and parameters used in this study. The generalisability of these results is discussed at the end of this section.

8.5.1 Efficiency gain

A predictive heating system is able to achieve the highest efficiency gain in poorly insulated buildings. The potential efficiency gain as determined by OPT is 9% to 19% for the flat F- U_{high} and 6% to 15% for the house H- U_{high} (Table 8.3). For well insulated buildings (low U-values), the efficiency gain under optimal prediction is reduced to a value of 4% to 12% for the flat and 3% to 10% for the house (Table 8.2). Higher U-values mean that the buildings' indoor temperature drops more quickly. At the same time, the prediction horizon I^* is reduced due to a higher design heat load $\Phi_{\text{H,max}}$ (cf. Table 7.7 in Section 8.3.1) and the efficiency gain increases. This happens regardless of the prediction algorithm. As I^* approaches zero, the predictive controller's behaviour approaches that of the *reactive* controller. The reactive controller (REA), which does not predict or preheat (*i.e.* only heats the building when it is occupied), has the highest efficiency gain for all scenarios – 9% to 19%. However, this also comes at the expense of the highest average discomfort degree hours (*i.e.* a large loss of comfort). For this reason, simplified Presence Probabilities (PPS) is clearly not a practical alternative in particular on very cold and freezing days. As the difference between Θ_{comf} and the outside temperature Θ_e becomes smaller, OPT and the reactive strategy converge since it takes less time to heat up the building.

The inability of the analysed algorithms to perfectly predict occupancy has the largest impact on well-insulated buildings (*i.e.* F- U_{low} and H- U_{low}) when solar gains and outdoor temperatures are low (*i.e.* very low temperature, cloudy scenario). In this case, when compared to OPT, the algorithms typically do not achieve much more than 50% of possible savings. This is due to the fact that this scenario requires prediction over a longer prediction horizon I^* .

8.5.2 Heating degree hours

As Table 8.4 shows, none of the prediction algorithms (OPT, MAT, MDMAT, PP, PPS and PH) produced significant comfort loss in terms of discomfort degree hours. Apart from the very low temperature scenario, where the temperature sometimes dropped below -6°C (the design temperature⁵ used for the calibration of $\Phi_{\text{H,max}}$), the average discomfort degree hours are less than one for all scenarios and prediction algorithms. Moreover, even for PPS there was no significant comfort loss for the low and moderate temperature scenarios. We will discuss possible reasons for this behaviour in Section 8.6.1.

One should realise that to achieve significant savings, the response of the “standard” heating controller (cf. Section 8.3.2) to the algorithms' predictions may be too conservative. Especially for lower temperatures and well-insulated buildings, the additional efficiency gain of the reactive over a predictive controller is substantial. This indicates that with

⁵The design temperature is defined as the minimum two-day average temperature that was reached at least 10 times in the last 20 years [42].

Table 8.4: Average discomfort degree hours per day (as a measure for comfort loss) for all experiments. ☀ and ☁ denote *clear* and *cloudy* scenarios.











Discomfort degree hours per day												
Weather	OPT, (MD)MAT, PP(S)		PH		REA		OPT, (MD)MAT, PP(S), PH				REA	
												
F-U _{low} (well insulated flat)						F-U _{high} (poorly insulated flat)						
Very low	0				17	22	0				1	1
Freezing	0				2	7	0					
Low	0				0	1	0					
Moderate	0						0					
H-U _{low} (well insulated house)						H-U _{high} (poorly insulated house)						
Very low	0		1	1	28	35	0				8	8
Freezing	0				5	12	0				1	2
Low	0				0	2	0					
Moderate	0						0					

Table 8.5: ISO 13790 annual efficiency gains.

Efficiency gain (%)														
Building	OPT		MAT		MDMAT		PP		PPS		PH		REA	
	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁	☀	☁
H-U _{low}	8	8					7 / 6						9	11
F-U _{low}	10	10	8	9	8	8	9	8	9	8	9	9	11	12
H-U _{high}	13	14					13						14	15
F-U _{high}	16	17					16 / 17						16	17

some (negligible or at least acceptable) comfort loss or simply by defining a reasonable temperature comfort bound around the setpoint, higher savings should be obtainable by more “courageous” predictive controllers. A modified controller, which not only optimises for zero miss-time (*e.g.* $\Theta_{\text{air}} = \Theta_{\text{comf}} \pm \Delta$) upon the arrival of the occupants) but also assigns a cost to discomfort degree hours and balances this with the actual heating costs, may obtain a higher efficiency gain while incurring only minimal additional discomfort degree hours (and thus comfort loss) per day. This approach has already been suggested by Mozer *et al.* in [140]. We leave the investigation of controllers that trade comfort loss for efficiency gain to future work.

8.5.3 Annualised savings

So far, the results in this section have shown the efficiency gain for selected weather scenarios. The annual efficiency gain is determined by the number of occurrences of each of these scenarios per year. Thus, they can be computed by weighting the efficiency gain of the weather scenarios by their empirical probability as derived from historical weather data. Table 8.5 shows the annualised efficiency gain for all four building scenarios. The weightings for the weather scenarios were determined using the historical weather

Table 8.6: Average outside temperatures for selected cities and simulated efficiency gain for January to March ($F-U_{low}$).

City	Average temperature ($^{\circ}\text{C}$)			Efficiency gain <i>OPT</i> (%)		
	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>
Moscow	-8.0	-7.0	-2.0	6	7	9
Toronto	-5.8	-5.6	-0.4	7	7	9
Beijing	-4.0	-1.0	6.0	5	6	11
Stockholm	-2.8	-3	0.1	7	7	9
New York	0.5	1.8	5.7	8	8	11
<i>Lausanne</i>	1.3	2.8	5.5	6	7	9
Brussels	3.3	3.7	6.8	8	8	10
London	4.3	4.5	6.9	8	8	10
Seattle	5.6	6.3	8.1	10	11	12

distribution of the 20 years from 1994 to 2014. The table shows that all the prediction algorithms (MAT, MDMAT, PP(S) and PH) achieved the same annual efficiency gain, close to OPT, ranging from 6% (well insulated house) to 17% (poorly insulated flat).

8.5.4 Impact of climate conditions

Different climate zones may offer varying potential for energy savings. To indicate how well our findings for Lausanne can be generalised to other locations, Table 8.6 shows the efficiency gain achievable by OPT for the average weather conditions from January to March for selected cities⁶. For these simulations, a simplified model of $F-U_{low}$ with no solar gains and constant outside temperatures was applied. The outside temperature equaled the average temperature for the month in question. Further details are outlined in Chapter 7.

Table 8.6 shows an increase in the efficiency gain of between 5% (Beijing) and 10% (Seattle) in January to a range between 9% (Toronto) and 12% (Seattle) in March. This pegs the efficiency gain closely to the annualised figures obtained for the more detailed Lausanne simulation shown in Table 8.5. Cities with larger differences in the average outside temperature (*e.g.* Beijing has a difference of 10°C between January and March), generally also have a larger variance in efficiency gain. This is due to the fact that the heating system is designed for the lowest temperatures. As the temperatures increase, the additional power of the heating system can be used to heat up the building more quickly.

8.5.5 Impact of the occupancy schedules

As one might expect, the potential for energy savings is highly correlated to a home's occupancy schedule. We analysed the impact of occupancy in the freezing temperature,

⁶Temperature data obtained from wikipedia.org, if available, otherwise from weatherbase.com.

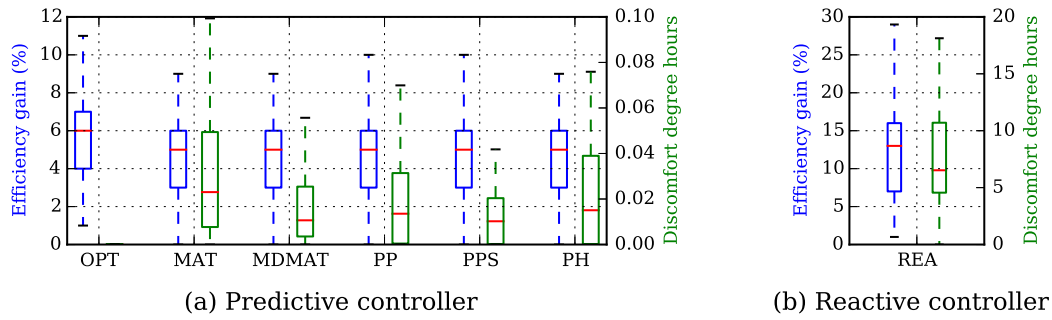


Figure 8.3: Efficiency gain and comfort loss measured in discomfort degree hours per day according to the ISO 5R1C model (F- U_{low} , freezing temperature, cloudy).

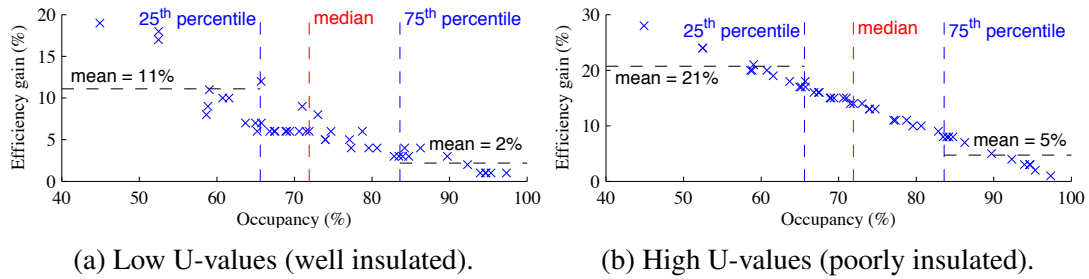


Figure 8.4: Efficiency gain / occupancy correlation: Freezing temperature, cloudy.

cloudy sky scenario weather scenario. Figure 8.3 shows that for the well insulated flat F- U_{low} , efficiency gain and discomfort degree hours vary considerably between the participants. The bar plot shows the median, quartiles and extreme values of metrics for each algorithm (outliers have been removed). The left side of the figure shows the results for the predictive controller in conjunction with the assessed prediction algorithms. The right side shows the results for the reactive controller for comparison. As noted previously, the discomfort degree hours induced by the prediction algorithms are negligible. Overall, there are no significant differences between the algorithms and the distribution of their efficiency gain across the participants.

Figure 8.4 shows the correlation between average occupancy and the efficiency gain that may be obtained by OPT for all 45 participants. Figures 8.4a and 8.4b contrast this relationship between F- U_{low} (good insulation) and F- U_{high} (poor insulation). The figures show that the quarter of homes that are least-occupied (25th percentile) outperformed the most-occupied homes (75th percentile) by a factor of 4-5. Low occupancy houses are clearly much better suited for installing smart heating systems than those with high occupancy.

The figures also show that for the 25% of homes with the lowest occupancy, the efficiency gain almost doubles from 11% to 21% from the well insulated to the poorly insulated flat. Not surprisingly, one can thus conclude that smart heating systems yield the

highest benefits in poorly insulated buildings.

Figure 8.4b shows an almost linear relationship between occupancy and efficiency gain. This relationship is less pronounced in Figure 8.4a. Here, the efficiency gain for the quarter of participants between the 25% quantile and the median is almost constant. As OPT's prediction is perfect, the reason for this effect lies in the structure of the occupancy schedules in conjunction with the increased prediction horizon due to the better insulation. The more arrival and departure events a schedule contains, the more difficult it is for the heating system to lower the temperature to a setback temperature.

8.6 Modelling limitations

Due to their novel nature, performance data from smart heating installations in residential buildings is still sparse. However, to make substantiated claims regarding the impact of different variables such as the building's occupancy and insulation on the efficiency gain and comfort loss of a predictive heating system, one must analyse each variable *ceteris paribus*. Thus, for the time being, in order to analyse the specific impact of different variables, one must resort to simulations. Simulation and modelling naturally involve a trade-off between model complexity and simulation accuracy. In the following, we will briefly discuss some of the shortcomings of the ISO 5R1C model used in this thesis and analyse our choice of baseline strategy for computing efficiency gain.

8.6.1 Building model

To simulate the heating system, we used the 5R1C model from the ISO 13790 standard [84]. In this model, the heat source is connected via the node for the indoor air temperature. As such, even though it has been widely adopted for building design in Europe [88, 152], the ISO 5R1C model more closely resembles a forced-air heating system common in the US, rather than the hydronic systems more typically encountered in Europe. A forced-air heating system typically reduces the preheat time and lowers the penalty for false predictions, thereby resulting in the low comfort loss exhibited by the simulation results (*cf.* Table 8.4). From the variations between different insulation levels (*cf.* Figure 8.4), we have already seen that shorter preheat times induced by more powerful heating systems result in an almost reactive strategy and thus in higher energy savings. As such, our evaluation hints at an upper bound on the savings that can be achieved using predictive heating systems and may lead to an underestimation of comfort loss.

8.6.2 Baseline metrics

In [176], Urban *et al.* describe the problem with choosing a representative baseline scenario for the energy savings of a smart thermostat. They highlight that using *average*

setpoint schedules of a range of users or the *historical energy use* does not yield satisfying results due to the irregularity of the former and differing existing setups in the households influencing the later. They thus argue to use “one fixed temperature setpoint during the entire heating season”.

For this reason, we employed such an *always-on* strategy as the baseline for evaluating the predictive controller and the occupancy prediction algorithms. In practice, however, many households use a (static) *nighttime setback*. Allowing the temperature to drop during the night by 4 °C to 6 °C has been shown to result in savings between 4% and 7% [81, 127]. A baseline strategy using a nighttime setback thus lowers the overall energy consumption, thereby – assuming the predictive setback generally occurs during the day – slightly increasing the efficiency gain of the predictive controller. Using a nighttime setback strategy as the baseline, however, necessitates a clear separation between the efficiency gain achieved by this setback and the predictive strategy.

8.7 Conclusions and lessons learned

The insights gained through our simulation-based performance analysis of occupancy-based approaches for smart heating control, based on real-world weather data and established building standards, can be summarised as follows.

Actual *comfort loss* in terms of *discomfort degree hours* is lower than the values implied by the accuracy of the prediction algorithm. A prediction accuracy of around 80% does not necessarily result in an uncomfortable thermal environment for 20% of the time. This is mainly due to the reactive nature of the heating scenario (*e.g.* heating is not turned off prematurely based on a predicted state if the occupants are still present). Moreover, the comfort loss is bounded by the time it takes to heat from the current temperature to the comfort temperature.

The *efficiency gain* achievable by occupancy prediction depends on the structure of the building, its occupancy and the weather conditions. Annual savings range from 6% to 17% depending on the type of building (*cf.* Table 8.5). Savings are almost doubled for poorly insulated buildings. The 25% of households with the lowest occupancy have a 4-5 times higher potential for efficiency gains than the quarter of homes with the highest occupancy. Lower temperatures and cloudy skies reduce efficiency gain and increase comfort loss as it takes longer to heat the building. Our data confirms similar results by [81] and [127] which showed energy savings of between 6% and 10% for cool and temperate climates using setback thermostats.

The algorithms’ inherent difficulty in correctly predicting the arrival time of the occupants imposes a penalty on the efficiency gain. To save more energy, *additional intelligence* could thus be incorporated into the controller. One example would be to forgo heating if only a short period of occupancy is predicted that would nevertheless result in significant

energy expenditure to heat up the property. A mobile application or simple “override” button on the thermostat to enable the occupants to control the smart thermostat in a simple and easy manner could deal with exceptional cases and increase user acceptance.

Conclusions and outlook

The goal of this thesis is to provide the technical foundations for the design and evaluation of future smart heating systems. Such systems, which reduce the energy consumption automatically by keeping the comfort temperature only when necessary (*i.e.* when occupants are present or will be present in the near future), require occupancy sensing and prediction infrastructure to operate.

In this final chapter, we outline our findings along the three research questions stated in the introduction: (i) Can existing technology be used opportunistically to sense occupancy? (ii) How accurately can occupancy be predicted? (iii) How much energy may be saved by a smart heating system using occupancy detection and prediction? We conclude this thesis with an outlook on future directions for research.

9.1 Opportunistic occupancy sensing

Accurate occupancy detection is an integral factor for the operation of a smart heating system. In fact, for existing commercial solutions, automatic control is often deactivated due to the poor occupancy detection accuracy of the systems [186]. Thus, the first third of this thesis was dedicated to opportunistic occupancy sensing using technology that already exists in many households. While dedicated infrastructure is still costly, the opportunistic use of such technology can reduce the cost of new smart heating installations and improve the accuracy of existing ones. Two examples of such existing technology are smart electricity meters and mobile phones.

Smart electricity meters are increasingly being deployed in households to facilitate billing and encourage a more economical use of resources. Our hypothesis was that current meters, which can measure a range of variables at a sampling rate of 1 Hz, can be used to detect occupancy in households. To investigate this hypothesis, we deployed a large set of sensors in six Swiss households over seven months (*cf.* Chapter 3). The Electricity

Consumption and Occupancy (ECO) dataset contains the total electricity consumption on all three phases, the consumption of selected appliances, events from a PIR sensor and ground truth occupancy data. In total, the dataset, which has been made open to the community, contains over 800 million records.

In Chapter 4 we use the ECO dataset to evaluate supervised machine learning approaches to sense occupancy from the electrical load curve. To this end, we derive 35 features describing the electricity consumption over 15-minute intervals. We then train four different classifiers (*i.e.* support vector machines (SVMs), K-nearest neighbours (KNNs), Gaussian mixture models (GMMs), hidden Markov models (HMMs) and a simple thresholding (THR) approach) in conjunction with SFS feature selection and principal component analysis (PCA) to remove redundant and irrelevant features. Our analysis showed that, using the SVM classifier in conjunction with PCA, an occupancy detection accuracy of up to 94% can be achieved in low-occupancy households.

For high-occupancy households (*e.g.* households with more than 80% occupancy) the accuracy does not significantly exceed the baseline accuracy provided by a simple maximum-likelihood predictor. In these households, the correlation between the energy consumption and occupancy is lower. A possible reason for this is that high occupancy actually increases the probability of observing intervals during which no electrical appliances are used. Furthermore, if the building is only unoccupied for brief periods of time, occupants may be less inclined to switch off appliances to save energy during these absences.

However, high occupancy households also leave little room for optimising the heating schedule. We showed that simple unsupervised approaches – which for example assume that the distribution of the nighttime electricity consumption is similar to the distribution during absence – can identify high and low occupancy households. Thus, our results show that for those households that can benefit from a smart heating system, opportunistic occupancy sensing using smart electricity meters is feasible, while simple unsupervised approaches can reliably infer which households should be targeted.

Mobile phones are a promising proxy for occupancy as current models often include localisation capabilities such as GPS or Wi-Fi. From the past locations of the mobile phone (and thus in most cases its owner), mobility patterns can be inferred that identify whenever a person was at home. In Chapter 5 we introduce the *homeset algorithm* and describe how it can be used to extract occupancy schedules from the LDCC mobile phone location dataset [117]. We show that inferring occupancy is possible using simple heuristics and highlight how a temporal matching of anonymised GPS traces to Wi-Fi scans can be used to infer information about when the participants were at home. By inferring occupancy schedules from a large, unlabelled dataset, we address the problem that currently no large public occupancy dataset exists that could be used to evaluate the performance of smart heating systems.

9.2 Occupancy prediction

Accurate occupancy sensing can be used by the thermostat to let the temperature drift to save energy whenever the building is not occupied. However, if the temperature has been allowed to drop during the absence of the residents, reheating the building requires a non-negligible time. Thus, reactively controlling the temperature can cause significant thermal discomfort for the occupants. In order for the heating system to decide when to start reheating, it thus needs to know the expected arrival time of its occupants.

In Chapter 6 we use 45 occupancy schedules derived from the LDCC mobile phone location dataset [117] using the homeset algorithm (*cf.* Chapter 5) to analyse the performance of a number of state-of-the-art occupancy prediction algorithms. We begin this middle third of the thesis by performing a literature review and classify existing occupancy prediction algorithms into *schedule-based*, *context-aware* and *hybrid* approaches. We then provide five implementations (*i.e.* MAT, MDMAT, PP, PPS and PH) of three different schedule-based occupancy prediction algorithms [113, 124, 169] and evaluate their performance against a maximum-likelihood predictor. Our results show that while all approaches outperform the naïve maximum-likelihood predictor, the Presence Probabilities (PP) approach by Krumm *et al.*, built upon the assumption that occupancy is correlated with the time of the day and the current day of the week, performed best with a median accuracy of 85% over the 45 participants.

To highlight the (lack of) potential for further improvements in schedule-based prediction approaches we further showed that the prediction accuracy of current state-of-the-art approaches is close to the predictability (*i.e.* the fundamental upper limit to the accuracy as posed by the natural irregularity of the participants' occupancy schedules) of the 45 schedules. This shows that major improvements to the prediction accuracy can only be achieved through hybrid approaches which combine schedule-based with context-aware prediction that relies also on the current context (*e.g.* position and activity) of the occupants.

9.3 Energy savings

The potential energy savings and comfort loss of a smart heating system do not only depend on the performance of the occupancy detection and prediction infrastructure, but also on the actual occupancy, the physics of the building and the prevailing weather conditions.

In the final third of this thesis we thus attempt to answer the question how much energy smart thermostats based on occupancy sensing and prediction may save under different environmental conditions. To this end, we develop a set of thermal models based on the ISO 13790 standard [84] (*cf.* Chapter 7) and compute the heating energy consumption for a range of different building and weather conditions (*cf.* Chapter 8). We

then simulate the energy consumption for 45 residents from the LDCC dataset. We evaluate predictive controllers using the five schedule-based occupancy prediction algorithms (*i.e.* MAT, MDMAT, PP, PPS and PH) as well as a reactive controller (REA) and an optimal predictive controller (OPT) with perfect occupancy prediction. We measure energy savings against a thermostat that keeps the same temperature year-round.

We show that the energy savings of a smart heating system are largely dependent on the properties of the building, its occupancy and the weather conditions. From all approaches, the reactive controller saves most energy. However, as expected, such a system also incurs a substantial comfort loss for the residents. All prediction algorithms achieve savings close to the optimum controller, while the potential savings vary substantially with the environmental conditions. Depending on the building parameters, the annual savings of using occupancy detection and prediction range from 6% to 17%. The actual occupancy of the household also has a strong impact on the possible energy savings. The quarter of households with the lowest occupancy can achieve 4-5 times higher savings than those 25% with the highest occupancy.

9.4 Future work

In the following we will list several limitations with our current approaches and suggest ideas for future work. We align these ideas along occupancy *sensing*, *prediction*, heating *control* strategies and the *evaluation of potential energy savings*.

9.4.1 Sensing

Sensor fusion for occupancy detection In Chapter 4 we show how to sense occupancy solely using smart electricity meters. In fact, many households contain additional sensors such as mobile phones (*cf.* Chapter 5) or other even dedicated occupancy sensing infrastructure such as PIR sensors. For future work, we suggest an investigation of how to combine data from multiple heterogeneous sensors with different failure models to increase overall occupancy detection performance. A possible approach would be to use Dempster–Shafer theory (DST) which combines evidence from different, possibly conflicting sources to establish at a degree of belief [17]. Thereby, we can address occupancy detection failures that stem for example from a mobile phone being left behind (or having a depleted battery) or the occupants not using any electric appliances whilst at home.

Sleep detection Previous work has shown that a nighttime setback by 4°C to 6°C can result in savings from 4% to 7% [81, 127]. In order to increase the energy savings, a smart thermostat should thus also sense whenever the occupants have gone to bed to let the temperature drift accordingly.

More complex unsupervised approaches Simple unsupervised occupancy detection algorithms operating at a 15-minute granularity already achieve a considerable detection accuracy (*cf.* Chapter 4). However, the performance of these approaches may be further improved. Possible refinements include k-Means clustering to find distinct power levels and GMMs to infer the joint probability function of the electricity distributions of the occupied and unoccupied states. While we could construct a supervised approach using occupancy data gathered by mobile phones or other sensors, such information is not available when the system is first set up. Therefore unsupervised approaches are important for the system to work seamlessly when it is first taken into operation. Furthermore, unsupervised approaches may be used to establish which households are actually suitable for a smart heating system by establishing with reasonable certainty whether a candidate household has low or high occupancy. A utility company providing gas and electricity, for example, may analyse their data to find customers that are suitable for a subsidised smart heating system.

Room-level sensing, prediction and control In Chapter 4, we highlight that occupancy detection by means of the electricity consumption of the households works best for a smaller living space with fewer inhabitants. One hypothesis is that sensing occupancy at the room level could not only improve the occupancy detection accuracy but also increase the potential for energy savings. By deploying submeters, we could measure the electricity consumption of individual rooms and therefore potentially infer room level occupancy. In addition, we could use Wi-Fi fingerprinting approaches to detect presence using mobile phones. By heating rooms independently, we could then make use of the fact that some rooms such as bedrooms are not used during the day, while others such as the living room are not used at night.

9.4.2 Prediction

Hybrid occupancy prediction In Chapter 6 we show that further improvements to state-of-the-art schedule-based occupancy prediction approaches are limited by the predictability of the schedules. Thus, while it is possible to predict regular behaviour, irregular events such as holidays or business trips cannot be reliably predicted from the past occupancy of the building. In these cases, a hybrid approach that – in addition to the historical occupancy data – makes use of the occupants’ current context (*e.g.* their location or activity) may improve the prediction. If the mobile phone of the occupant is logged into a cellular tower which is located several hours away from the home it is unlikely that the occupant will return shortly. Regardless of the historical occupancy schedule, the heating system can use this information to save energy by letting the temperature drift.

9.4.3 Control

User in the loop First experiences with smart heating systems have shown that automatic heating faces user acceptance problems if the decisions are not made transparent. If the occupants do not feel in control of the system, they are likely to turn it off completely [186]. A smart heating should thus have some means for measuring the comfort of the occupants. If the control modalities remained the same (*e.g.* we use the same valves in a hydronic heating system) comfort could for instance be measured in the number of interactions with the system. Alternatively, the users could be provided with an override button that enables occupants to trigger a reset to the comfortable temperature.

Incorporating comfort models In Chapter 2, we give an overview of existing approaches to quantify comfort. A smart heating system should incorporate models that allow to create a comfortable thermal environment for all occupants. Different occupants may have different thermal comfort criteria depending on their physiology. Furthermore, depending on the current activity, the same person will have different metabolic rates, resulting in temporal variations of the comfort criteria. To address both these issues, body-worn sensors such as activity trackers and smart watches may be used to sense the current activity level and infer metabolic rate of the occupants.

More sophisticated control strategies Currently, our simulation approach uses a very simple predictive controller that has perfect knowledge of the thermal model of the building and the future weather. In reality, a smart heating system must rely on weather forecasts to establish the right time to heat up the building. Furthermore, user comfort models may vary and users may wish to set different comfort temperatures for different rooms and different times. A controller must therefore not only learn the thermal model of the building on-the-fly, but also make sure that the constraints of the user are matched as the environmental conditions remain unpredictable. A number of authors has looked into this problem in the context of model predictive control (MPC). However, little attention has so far been given to enabling MPC in residential environments.

9.4.4 Evaluation of energy savings

Different thermal models While the ISO 13790 model does not simulate a specific heating system, it most closely resembles forced air heating. This type of heating is not very common in Europe, where most households use hydronic heating. Future work to investigate the impact of occupancy detection and prediction must thus also take into account different heat generation and transmission systems. In contrast to a forced air heating system, a hydronic system using wall-mounted radiators has a longer reheat time as the heat is not directly transmitted to the air. At the same time, the home stays warm for

longer as the radiators still store some heat even as the central heating system has already switched off. The time lag increases the prediction horizon and therefore the impact of the prediction accuracy on the overall efficiency gain and comfort loss.

Varying baseline In Chapter 8 we argued for our decision to use an always-on heating schedule as the baseline for our evaluation. This choice is well-founded in the literature – due to a variety of different reasons people do not use setbacks on their programmable thermostats [143, 146, 157, 158]. However, the households that do actually already operate a setback schedule are most likely the ones interested in saving even more energy by utilising a smart thermostat with occupancy detection and prediction. For this reason, a fair comparison should also take into account the effect of various nighttime setback strategies.

Bibliography

- [1] Energiewirtschaftsgesetz vom 7. Juli 2005 (BGBl I S. 1970, 3621), zuletzt geändert am 21. Juli 2014 (BGBl I S. 1066).
- [2] World's "smartest" house created by CU-Boulder team. PEB Exchange, Programme on Educational Building, March 1998.
- [3] Verordnung über energiesparenden Wärmeschutz und energiesparende Anlagentechnik bei Gebäuden in der Fassung der Bekanntmachung vom 24. Juli 2007 (BGBl I S. 1519), zuletzt geändert am 18. November 2013 (BGBl I S. 3951–3990). 2013.
- [4] Ebola and big data: Call for help. *The Economist*, October 25th 2014.
- [5] Joana M. Abreu, Francisco Pereira Câmara, and Paulo Ferrão. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and Buildings*, 49:479–487, June 2012.
- [6] Yuvraj Agarwal, Bharathan Balaji, Seemanta Dutta, R.K. Gupta, and Thomas Weng. Duty-cycling buildings aggressively: The next frontier in HVAC control. In *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN '11)*, pages 246–257, Chicago, IL, USA, 2011. IEEE.
- [7] Yuvraj Agarwal, Bharathan Balaji, Rajesh Gupta, Jacob Lyles, Michael Wei, and Thomas Weng. Occupancy-driven energy management for smart building automation. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '10)*, pages 1–6. ACM, 2010.
- [8] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer, 1998.

- [9] Adrian Albert and Ram Rajagopal. Smart-meter driven segmentation: What your consumption says about you. *IEEE Transactions on Power Systems*, 28(4):4019–4030, November 2013.
- [10] RLW Analytics. Validating the impact of programmable thermostats. Technical report, Middletown, CT, USA, January 2007.
- [11] Kyle Anderson, Adrian Ocneanu, Diego Benitez, Derrick Carlson, Anthony Rowe, and Mario Berges. BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. In *Proceedings of the 2nd workshop on data mining applications in sustainability (SustKDD '12)*, pages 5:1–5:5, Beijing, PRC, August 2012.
- [12] Daniel Ashbrook and Thad Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7(5):275–286, October 2003.
- [13] ASHRAE. Thermal environmental conditions for human occupancy. ANSI/ASHRAE Standard 55-2010, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, USA, 2010.
- [14] ASHRAE. Ventilation for acceptable indoor air quality. ANSI/ASHRAE/ASHE Standard 62.1-2010, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, USA, 2010.
- [15] Michael Baeriswyl, André Müller, Reto Rigassi, Christof Rissi, Simon Solenthaler, Thorsten Staake, and Thomas Weisskopf. Folgeabschätzung einer Einführung von "Smart Metering" im Zusammenhang mit "Smart Grids" in der Schweiz, June 2012.
- [16] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Smart*: An open data set and tools for enabling research in sustainable homes. In *Proceedings of the 2nd Workshop on Data Mining Applications in Sustainability (SustKDD '12)*, pages 1:1–1:6, Beijing, PRC, August 2012. ACM.
- [17] Otman Basir and Xiaohong Yuan. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. *Information Fusion*, 8(4):379–386, October 2007.
- [18] Nipun Batra, Manoj Gulati, Amarjeet Singh, and Mani B. Srivastava. It's different: Insights into home energy consumption in India. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys '13)*, Rome, Italy, November 2013. ACM.

-
- [19] Paul Baumann, Wilhelm Kleiminger, and Silvia Santini. The influence of temporal and spatial features on the performance of next-place prediction algorithms. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*, pages 449–458, Zurich, Switzerland, September 2013. ACM.
- [20] Pia Baumann. Personal communication, 22. October 2014. Estimated Swiss energy consumption by end use for the years 2001 to 2006 (to complement [93]).
- [21] Christian Beckel, Wilhelm Kleiminger, Romano Cicchetti, Thorsten Staake, and Silvia Santini. The ECO data set and the performance of non-intrusive load monitoring algorithms. In *Proceedings of the 1st ACM International Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*, pages 80–89, Memphis, TN, USA, November 2014. ACM.
- [22] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. In *Proceedings of the 4th International Conference on Future Energy Systems (e-Energy '13)*, pages 75–86, Berkeley, CA, USA, May 2013. ACM.
- [23] Alex Beltran and Alberto E. Cerpa. Optimal HVAC building control with occupancy prediction. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*, pages 168–171, Memphis, TN, USA, November 2014. ACM.
- [24] Alex Beltran, Varick L. Erickson, and Alberto E. Cerpa. Thermosense: Occupancy thermal based sensing for HVAC control. In *Proceedings of the 5th ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys '13)*, pages 11:1–11:8, Rome, Italy, November 2013. ACM.
- [25] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 663–666. Springer, 2013.
- [26] Clemens Louis Beuken. *Wärmeverluste bei periodisch betriebenen elektrischen Öfen: Eine neue Methode zur Vorausbestimmung nicht-stationärer Wärmeströmungen*. PhD thesis, Sächsische Bergakademie Freiberg, 1936.
- [27] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data

- for development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.
- [28] Peter J. Boait and R. Mark Rylatt. A method for fully automatic operation of domestic heating. *Energy and Buildings*, 42(1):11–16, 2010.
- [29] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*, pages 427–434, Istanbul, Turkey, November 2014. ACM.
- [30] Zachary Davies Boren. There are officially more mobile devices than people in the world. *The Independent*, October 7th 2014.
- [31] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52(0):213–234, January 2013.
- [32] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:27:1–27:27, April 2011.
- [33] Dong Chen, Sean Barker, Adarsh Subbaswamy, David Irwin, and Prashant Shenoy. Non-intrusive occupancy monitoring using smart meters. In *Proceedings of the 5th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '13)*, pages 9:1–9:8, Rome, Italy, November 2013. ACM.
- [34] Yohan Chon, Nicholas D. Lane, Fan Li, Hojung Cha, and Feng Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proceedings of the 14th ACM International Conference on Ubiquitous Computing (UbiComp '12)*, pages 481–490, Pittsburgh, PA, USA, September 2012. ACM.
- [35] Ionut Constandache, Shravan Gaonkar, Matt Sayler, Romit Roy Choudhury, and Landon Cox. Enloc: Energy-efficient localization for mobile phones. In *Proceedings of INFOCOM '09*, pages 2716–2720, Tel Aviv, Israel, March 2009. IEEE.
- [36] Vincenzo Corrado and Enrico Fabrizio. Assessment of building cooling energy need through a quasi-steady state model: Simplified correlation for gain-loss mismatch. *Energy and Buildings*, 39(5):569–579, May 2007.
- [37] Richard J. de Dear and Gail Schiller Brager. Developing an adaptive model of thermal comfort and preference. *ASHRAE Transactions*, 104(1):145–167, January 1998.

-
- [38] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, December 2013.
- [39] Daswin De Silva, Xinghuo Yu, Damminda Alahakoon, and Grahame Holmes. A data mining framework for electricity consumption analysis from meter data. *IEEE Transactions on Industrial Informatics*, 7:399–407, August 2011.
- [40] Robert F. Dickerson, Eugenia I. Gorlin, and John A. Stankovic. Empath: A continuous remote emotional health monitoring system for depressive illness. In *Proceedings of Wireless Health '11*, San Diego, CA, USA, October 2011. ACM.
- [41] Edsger Wybe Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.
- [42] DIN. Heating systems in buildings – method for calculation of the design heat load. DIN EN 12831-03:2008, Deutsches Institut für Normung, Berlin, Germany, 2008.
- [43] Robert H. Dodier, Gregor P. Henze, Dale K. Tiller, and Xin Guo. Building occupancy detection through sensor belief networks. *Energy and Buildings*, 38(9):1033–1043, September 2006.
- [44] Olivier Dousse, Julien Eberle, and Matthias Mertens. Place learning via direct WiFi fingerprint clustering. In *13th International Conference on Mobile Data Management (MDM'12)*. IEEE, July 2012.
- [45] Adam Dunkels. Full TCP/IP for 8-bit architectures. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MobiSys '03)*, pages 85–98, San Francisco, California, May 2003. ACM.
- [46] Carl Ellis, James Scott, Mike Hazas, and John Krumm. EarlyOff: Using house cooling rates to save energy. In *Proceedings of the 4th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '12)*, pages 39–41, Toronto, ON, Canada, November 2012. ACM.
- [47] Varick L. Erickson, Stefan Achleitner, and Alberto E. Cerpa. POEM: Power-efficient occupancy-based energy management system. In *Proceedings of the 12th International Conference on Information Processing in Sensor Networks (IPSN '13)*, pages 203–216, Berlin, Germany, April 2013. ACM/IEEE.
- [48] Varick L. Erickson, Miguel Á. Carreira-Perpiñán, and Alberto E. Cerpa. OBSERVE: Occupancy-based system for efficient reduction of HVAC energy. In *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN '11)*, pages 258–269, Chicago, IL, USA, April 2011. IEEE.

- [49] Varick L. Erickson and Alberto E. Cerpa. Thermovote: Participatory sensing for efficient building HVAC conditioning. In *Proceedings of the 4th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '12)*, pages 9–16, Toronto, ON, Canada, November 2012. ACM.
- [50] European Commission. Benchmarking smart metering deployment in the EU-27 with a focus on electricity. COM(2014) 356 final, June 2014.
- [51] European Parliament and the Council of 16. Directive 2002/91/EC of 16 December 2002 on the energy performance of buildings. *Official Journal L 001*, pages 65–71, December 2002.
- [52] Florian Fainelli. The OpenWrt embedded development framework. In *Proceedings of the Free and Open Source Software Developers European Meeting*, January 2008.
- [53] Povl Ole Fanger. *Thermal comfort: Analysis and applications in environmental engineering*. PhD thesis, Danmarks Tekniske Højskole, 1970.
- [54] Ahmad Faruqui, Sanem Sergici, and Ahmed Sharif. The Impact of Informational Feedback on Energy Consumption – A Survey of the Experimental Evidence. *Energy*, 35(4):1598–1608, April 2010.
- [55] Pedro M. Ferreira, Antonio E. Ruano, Sergio Silva, and Eusebio Z.E. Conceicao. Neural networks based predictive control for thermal comfort and energy savings in public buildings. *Energy and Buildings*, 55(0):238–251, 2012.
- [56] Forum für Energieeffizienz in der Gebäudetechnik e.V. Leitfaden zum Heizungs-Check nach DIN EN 15378. page 20, May 2010.
- [57] Marc Fountain, Gall Brager, Edward Arens, Fred Bauman, and Charles Benton. Comfort control for short-term occupancy. *Energy and Buildings*, 21:1–13, 1994.
- [58] Gilles Fraisse, Christelle Viardot, Olivier Lafabrie, and Gilbert Achard. Development of a simplified and accurate building model based on electrical analogy. *Energy and Buildings*, 34(10):1017–1031, 2002.
- [59] Jordan Frank, Shie Mannor, and Doina Precup. Generating storylines from sensor data. *Pervasive and Mobile Computing*, 9(6):838–847, December 2013.
- [60] Peter Xiang Gao and Srinivasan Keshav. Optimal personal comfort management using SPOT+. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys '13)*, pages 22:1–22:8, Rome, Italy, November 2013. ACM.

-
- [61] Peter Xiang Gao and Srinivasan Keshav. SPOT: A smart personalized office thermal control system. In *Proceedings of the 4th International Conference on Future Energy Systems (e-Energy '13)*, pages 237–246, Berkeley, CA, USA, May 2013. ACM.
- [62] Carlos E. García, David M. Prett, and Manfred Morari. Model predictive control: Theory and practice – a survey. *Automatica*, 25(3):335–348, May 1989.
- [63] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding Individual Human Mobility Patterns. *Nature*, 453:779–782, June 2008.
- [64] Siddharth Goyal, Herbert A. Ingle, and Prabir Barooah. Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance. *Applied Energy*, 106(0):209–221, June 2013.
- [65] Xin Guo, Dale K. Tiller, Gregor P. Henze, and Clarence E. Waters. The performance of occupancy-based lighting control systems: A review. *Lighting Research and Technology*, 42(4):415–431, August 2010.
- [66] Chamanlal Gupta. A systematic approach to optimum thermal design. *Building Science*, 5(3):165–173, December 1970.
- [67] Manu Gupta, Stephen S. Intille, and Kent Larson. Adding GPS-control to traditional thermostats: An exploration of potential energy savings and design challenges. In *Proceedings of the 7th International Conference on Pervasive Computing (Pervasive '09)*, pages 1–18, Nara, Japan, May 2009. Springer.
- [68] Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel. ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*, pages 139–148, Copenhagen, Denmark, September 2010. ACM.
- [69] George W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, December 1992.
- [70] Peter Hartwell. A sea of sensors. *The Economist*, November 4th 2010.
- [71] Michael Hayner, Jo Ruoff, and Dieter Thiel. *Faustformel Gebäudetechnik*. Deutsche Verlags-Anstalt, München, Germany, 2013.
- [72] Nora Helbig. *Application of the radiosity approach to the radiation balance in complex terrain*. PhD thesis, University of Zurich, Zurich, Switzerland, February 2009.

- [73] Jeffrey Hightower, Sunny Consolvo, Anthony LaMarca, Ian Smith, and Jeff Hughes. Learning and recognizing the places we go. In *Proceedings of the 7th International Conference on Ubiquitous Computing (UbiComp '05)*, Tokyo, Japan, September 2005. Springer.
- [74] Timothy W. Hnat, Vijay Srinivasan, Jiakang Lu, Tamim I. Sookoor, Raymond Dawson, John Stankovic, and Kamin Whitehouse. The hitchhiker's guide to successful residential sensing deployments. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys '11)*, pages 232–245, Seattle, WA, USA, November 2011. ACM.
- [75] Chris Holcomb. Pecan Street Inc.: A test-bed for NILM. In *Proceedings of the 1st International Workshop on Non-Intrusive Load Monitoring*, Pittsburgh, PA, USA, May 2012.
- [76] Dezhi Hong and Kamin Whitehouse. A feasibility study: Mining daily traces for home heating control. April 2013.
- [77] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
- [78] Jonathan W. Hui and David E. Culler. IP is dead, long live IP for wireless sensor networks. In *Proceedings of the 6th ACM Conference on Embedded Networked Sensor Systems (SenSys '08)*, pages 15–28, Raleigh, NC, USA, November 2008. ACM.
- [79] IEC. Electricity metering – data exchange for meter reading, tariff and load control. IEC 62056-61, International Electrotechnical Commission, Geneva, Switzerland, February 2002.
- [80] Frank P. Incropera, Adrienne S. Lavine, and David P. DeWitt. *Fundamentals of heat and mass transfer*. Wiley, 2011.
- [81] John Ingersoll and Joe Huang. Heating energy use management in residential buildings by temperature control. *Energy and Buildings*, 8(1):27–35, February 1985.
- [82] ISO. Thermal performance of buildings – calculation of energy use for heating – residential buildings. ISO 832:2000, European Committee for Standardization, 1999. Superseded by EN ISO 13790:2008.
- [83] ISO. Ergonomics of the thermal environment – analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices

- and local thermal comfort criteria. ISO 7730:2005, International Organization for Standardization, Geneva, Switzerland, 2005.
- [84] ISO. Energy performance of buildings – calculation of energy use for space heating and cooling. ISO 13790-1:2008, International Organization for Standardization, Geneva, Switzerland, 2008.
- [85] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.
- [86] Xiaofan Jiang, Stephen Dawson-Haggerty, Prabal Dutta, and David E. Culler. Design and implementation of a high-fidelity AC metering network. In *Proceedings of the 8th International Conference on Information Processing in Sensor Networks (IPSN '09)*, San Francisco, CA, USA, April 2009. ACM.
- [87] Ming Jin, Ruoxi Jia, Zhaoyi Kang, Ioannis C. Konstantakopoulos, and Costas J. Spanos. Presencesense: Zero-training algorithm for individual presence detection based on power monitoring. *CoRR*, abs/1407.4395, 2014.
- [88] Juha Jokisalo and Jarek Kurnitski. Performance of EN ISO 13790 utilisation factor heat demand calculation method in a cold climate. *Energy and Buildings*, 39(2):236–247, February 2007.
- [89] Soteris A. Kalogirou and Milorad Bojic. Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy*, 25(5):479–491, May 2000.
- [90] Jérôme Henri Kämpf and Darren Robinson. A simplified thermal model to support analysis of urban resource flows. *Energy and Buildings*, 39(4):445–453, April 2007.
- [91] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting Places from Traces of Locations. *Mobile Computing and Communications Review*, 9(3):58, July 2005.
- [92] Jack Kelly and William J. Knottenbelt. UK-DALE: A dataset recording UK domestic appliance-level electricity demand and whole-house demand. *CoRR*, abs/1404.0284, 2014.
- [93] Andreas Kemmler, Alexander Piégisa, Andrea Ley, Philipp Wüthrich, Mario Keller, Martin Jakob, and Giacomo Catenazzi. Analyse des schweizerischen Energieverbrauchs 2000–2013 nach Verwendungszwecken. Technical report, Bundesamt für Energie, Bern, Schweiz, September 2014.

- [94] Aftab Khan, James Nicholson, Sebastian Mellor, Daniel Jackson, Karim Ladha, Cassim Ladha, Jon Hand, Joseph Clarke, Patrick Olivier, and Thomas Plötz. Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*, pages 90–99, Memphis, Tennessee, 2014. ACM.
- [95] Sarah Kilcher and Andreas Dröschner. Smart meters in the field – a sensor framework for a real world deployment. Distributed Systems Laboratory Report, ETH Zurich, May 2012.
- [96] Hyungsul Kim, Manish Marwah, Martin Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised disaggregation of low frequency power measurements. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM '11)*, Mesa, Arizona, USA, April 2011. SIAM.
- [97] Younghun Kim, Thomas Schmid, Zainul M. Charbiwala, and Mani B. Srivastava. ViridiScope: Design and implementation of a fine grained power monitoring system for homes. In *Proceedings of the 11th ACM International Conference on Ubiquitous Computing (UbiComp '09)*, pages 245–254, Orlando, FL, USA, September 2009. ACM.
- [98] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proceedings of the 7th International Conference on Pervasive Services (ICPS '10)*, Berlin, Germany, July 2010. ACM.
- [99] Manuel Klaey. eMeter – Integrating submeters to individually monitor appliances. Distributed Systems Laboratory Report, ETH Zurich, May 2012.
- [100] Wilhelm Kleiminger, Christian Beckel, Anind Dey, and Silvia Santini. Inferring household occupancy patterns from unlabelled sensor data. Technical Report 795, ETH Zurich, September 2013.
- [101] Wilhelm Kleiminger, Christian Beckel, Anind Dey, and Silvia Santini. Poster abstract: Using unlabeled Wi-Fi scan data to discover occupancy patterns of private households. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*, Rome, Italy, 2013. ACM.
- [102] Wilhelm Kleiminger, Christian Beckel, Thorsten Staake, and Silvia Santini. Occupancy detection from electricity consumption data. In *Proceedings of the 5th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '13)*, pages 1–8, Rome, Italy, November 2013. ACM.

-
- [103] Wilhelm Kleiminger, Friedemann Mattern, and Silvia Santini. Predicting household occupancy for smart heating control: A comparative performance analysis of state-of-the-art approaches. *Energy and Buildings*, 85(0):493–505, December 2014.
- [104] Wilhelm Kleiminger, Friedemann Mattern, and Silvia Santini. Simulating the energy savings potential in domestic heating scenarios in Switzerland. Technical report, ETH Zurich, Department of Computer Science, August 2014.
- [105] Wilhelm Kleiminger, Friedemann Mattern, and Silvia Santini. Smart heating control with occupancy prediction: How much can one save? In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*, pages 947–954, Seattle, WA, USA, September 2014.
- [106] Eberhard Knobloch. *The shoulders on which we stand – Wegbereiter der Wissenschaft: 125 Jahre Technische Universität Berlin*. Springer, April 2004.
- [107] Christian Koehler, Brian D. Ziebart, Jennifer Mankoff, and Anind K. Dey. TherML: Occupancy prediction for thermostat control. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*, pages 103–112, Zurich, Switzerland, September 2013. ACM.
- [108] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, January 1982.
- [109] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, September 1990.
- [110] J. Zico Kolter and Matthew J. Johnson. REDD: A public data set for energy disaggregation research. In *Proceedings of the 1st Workshop on Data Mining Applications in Sustainability (SustKDD '11)*, San Diego, CA, USA, August 2011. ACM.
- [111] Matthias Kovatsch, Simon Duquennoy, and Adam Dunkels. A low-power CoAP for Contiki. In *Proceedings of the 8th International Conference on Mobile Adhoc and Sensor Systems (MASS '11)*, pages 855–860, October 2011.
- [112] Rick Kramer, Jos van Schijndel, and Henk Schellen. Simplified thermal and hygric building models: A literature review. *Frontiers of Architectural Research*, 1(4):318–325, December 2012.
- [113] John Krumm and Alice Jane Bernheim Brush. Learning time-based presence probabilities. In *Proceedings of the 9th International Conference on Pervasive Computing (Pervasive '11)*, pages 79–96, San Francisco, CA, USA, June 2011. IEEE.

- [114] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *Proceedings of the 8th ACM International Conference on Ubiquitous Computing (UbiComp '06)*, pages 243–260, Orange County, CA, USA, September 2006. Springer.
- [115] Abraham Hang-yat Lam, Yi Yuan, and Dan Wang. An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings. In *Proceedings of the 5th International Conference on Future Energy Systems (e-Energy '14)*, pages 133–143, Cambridge, United Kingdom, June 2014. ACM.
- [116] Landis+Gyr. ZMK300CE – Basismodul SyM2 E750 Technische Daten, January 2012.
- [117] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. The Mobile Data Challenge: Big Data for Mobile Computing Research. In *Proceedings of the MDC by Nokia Workshop, in conjunction with Pervasive'12*, Newcastle, United Kingdom, June 2012. Springer.
- [118] Seungwoo Lee, Yohan Chon, Yunjong Kim, Rhan Ha, and Hojung Cha. Occupancy prediction algorithms for thermostat control systems using mobile devices. *IEEE Transactions on Smart Grid*, 4(3):1332–1340, September 2013.
- [119] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, June 1932.
- [120] Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. *CoRR*, abs/1306.4534, 2013.
- [121] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '03)*, pages 2–11, San Diego, California, June 2003. ACM.
- [122] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, October 2007.
- [123] Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 14th ACM International Conference on Ubiquitous Computing (UbiComp '12)*, pages 381–390, Pittsburgh, PA, USA, September 2012. ACM.

-
- [124] Jiakang Lu, Tamim Sookoor, Vijay Srinivasan, Ge Gao, Brian Holben, John Stankovic, Eric Field, and Kamin Whitehouse. The Smart Thermostat: Using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys '10)*, pages 211–224, Zurich, Switzerland, November 2010. ACM.
- [125] Stephen Makonin, Fred Popowich, Lyn Bartram, Bob Gill, and Ivan V. Bajic. AMPds: A public dataset for load disaggregation and eco-feedback research. In *Proceedings of the Annual Electrical Power and Energy Conference (EPEC '13)*, Halifax, NS, Canada, August 2013. IEEE.
- [126] William C. Mann. *Smart technology for aging, disability, and independence: The state of the science*. Wiley, July 2005.
- [127] Marianne M. Manning, Mike C. Swinton, Frank Szadkowski, John Gusdorf, and Ken Ruest. The effects of thermostat setting on seasonal energy consumption at the CCHT twin house facility. *ASHRAE Transactions*, 113(1):1–12, 2007.
- [128] Claudio Martani, David Lee, Prudence Robinson, Rex Britter, and Carlo Ratti. ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings*, 47(0):584–591, April 2012.
- [129] Edward H. Mathews, Pieter G. Richards, and Christophe Lombard. A first-order thermal model for building design. *Energy and Buildings*, 21(2):133–145, 1994.
- [130] Friedemann Mattern and Christian Floerkemeier. From the Internet of computers to the Internet of Things. In Kai Sachs, Ilia Petrov, and Pablo Guerrero, editors, *From Active Data Management to Event-Based Systems and More*, volume 6462 of *LNCs*, pages 242–259. Springer, 2010.
- [131] Friedemann Mattern, Thorsten Staake, and Markus Weiss. ICT for green – how computers can help us to conserve energy. In *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking (e-Energy '10)*, pages 1–10, Passau, Germany, April 2010. ACM.
- [132] Brian W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, October 1975.
- [133] Marvin McNett and Geoffrey M. Voelker. Access and mobility of wireless PDA users. *SIGMOBILE Mobile Computing and Communications Review*, 9(2):40–55, April 2005.

- [134] Marija Milenkovic and Oliver Amft. An opportunistic activity-sensing approach to save energy in office buildings. In *Proceedings of the 4rd International Conference on Energy-Efficient Computing and Networking (e-Energy '13)*, Berkeley, CA, USA, May 2013. ACM.
- [135] David L. Mills, Jim Martin, Jack Burbank, and William Kasch. Network Time Protocol Version 4: Protocol and Algorithms Specification. RFC 5905 (Proposed Standard), June 2010.
- [136] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private memoirs of a smart meter. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '10)*, pages 61–66, Zurich, Switzerland, September 2010. ACM.
- [137] Andrea Monacchi, Dominik Egarter, Wilfried Elmenreich, Salvatore D'Alessandro, and Andrea M. Tonello. GREEND: An energy consumption dataset of households in Italy and Austria. In *Proceedings of the 4th International Conference on Smart Grid Communications (SmartGridComm '14)*, pages 511–516, Venice, Italy, November 2014. IEEE.
- [138] Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. Discovering Places of Interest in Everyday Life from Smartphone Data. *Multimedia Tools and Applications*, 62:179–207, January 2013.
- [139] Michael C. Mozer, Robert H. Dodier, Marc Anderson, Lucky Vidmar, Robert F. Cruickshank, and Debra Miller. The neural network house: An overview. In L. Niklasson and Boden M., editors, *Current trends in connectionism*, pages 371–380, Hillsdale, NJ: Erlbaum, 1995.
- [140] Michael C. Mozer, Lucky Vidmar, and Robert H. Dodier. The Neurothermostat: Predictive optimal control of residential heating systems. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 953–959. The MIT Press, 1997.
- [141] Lorne W. Nelson and J. Ward MacArthur. Energy savings through thermostat setback. In *ASHRAE Transactions*, volume 84, pages 319–333, Albuquerque, NM, USA, 1978.
- [142] Alberto Hernandez Neto and Flávio Augusto Sanzovo Fiorelli. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40(12):2169–2176, 2008.

-
- [143] Monica J. Nevius and Scott Pigg. Programmable thermostats that go berserk: Taking a social perspective on space heating in Wisconsin. In *ACEEE Summer Study on Energy Efficiency in Buildings*. ACEEE, August 2000.
- [144] Tuan Anh Nguyen and Marco Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and Buildings*, 56:244–257, January 2013.
- [145] Richard Nock and Frank Nielsen. On weighting clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1223–1235, August 2006.
- [146] Donald A. Norman. *The design of everyday things*. Basic books, September 2002.
- [147] Brian Norton. *Harnessing Solar Heat*, volume 18 of *Lecture Notes in Energy*. Springer, 2014.
- [148] Frauke Oldewurtel, Alessandra Parisio, Colin N. Jones, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann, and Manfred Morari. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*, 45(0):15–27, February 2012.
- [149] Frauke Oldewurtel, Alessandra Parisio, Colin N. Jones, Manfred Morari, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann, and Katharina Wirth. Energy efficient building climate control using stochastic model predictive control and weather predictions. In *Proceedings of the 28th American Control Conference (ACC '10)*, pages 5100–5105, Baltimore, MD, USA, July 2010. IEEE.
- [150] Frauke Oldewurtel, David Sturzenegger, and Manfred Morari. Importance of occupancy information for building climate control. *Applied Energy*, 101(0):521–532, January 2013.
- [151] Thomas Olofsson and T.M. Indra Mahlia. Modeling and simulation of the energy use in an occupied residential building in cold climate. *Applied Energy*, 91(1):432–438, March 2012.
- [152] José A. Orosa and Armando C. Oliveira. Implementation of a method in EN ISO 13790 for calculating the utilisation factor taking into account different permeability levels of internal coverings. *Energy and Buildings*, 42(5):598–604, 2010.
- [153] Benedikt Ostermaier, Matthias Kovatsch, and Silvia Santini. Connecting things to the Web using programmable low-power wifi modules. In *Proceedings of the 2nd International Workshop on the Web of Things (WoT '11)*, San Francisco, CA, USA, June 2011.

- [154] Danny Parker, David Hoak, and Jamie Cummings. Pilot Evaluation of Energy Savings and Persistence from Residential Energy Demand Feedback Devices in a Hot Climate. Technical report, Florida Solar Energy Center, Cocoa, FL, USA, January 2008.
- [155] European Parliament and Council. Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC, July 2009.
- [156] European Parliament and Council. Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast), May 2010.
- [157] Therese Pepper, Marco Pritoni, Alan Meier, Cecilia Aragon, and Daniel Perry. How people use thermostats in homes: A review. *Building and Environment*, 46(12):2529–2541, 2011.
- [158] Daniel Perry, Cecilia Aragon, Alan Meier, Therese Pepper, and Marco Pritoni. Making energy savings easier: Usability metrics for thermostats. *Journal of Usability Studies*, 6(4):226–244, August 2011.
- [159] Jan Petzold. Augsburg indoor location tracking benchmarks. Technical Report 2004-09, Fakultät für Angewandte Informatik der Universität Augsburg, 2005.
- [160] Hermann Recknagel, Otto Ginsberg, Kurt Gehrenbeck, Eberhard Sprenger, Winfried Hönnmann, and Ernst-Rudolf Schramek, editors. *Taschenbuch für Heizung + Klimatechnik*, volume 76. Oldenbourg Industrieverlag, München, Germany, 2013.
- [161] T. Agami Reddy. *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. Springer, 2011.
- [162] Andreas Reinhardt and Sebastian Koessler. PowerSAX: Fast motif matching in distributed power meter data using symbolic representations. In *Proceedings of the 9th IEEE International Workshop on Practical Issues in Building Sensor Network Applications (SenseApp '14)*, pages 531–538, Edmonton, Canada, September 2014.
- [163] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663, 2009.
- [164] Alex Rogers, Siddhartha Ghosh, Reuben Wilcock, and Nicholas R. Jennings. A scalable low-cost solution to provide personalised home heating advice to households. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys '13)*, pages 1:1–1:8, Rome, Italy, November 2013. ACM.

-
- [165] Dan Rogers, Martin Foster, and Chris Bingham. Experimental investigation of a recursive modelling MPC system for space heating within an occupied domestic dwelling. *Building and Environment*, 72(0):356–367, February 2014.
- [166] Ignacio Benítez Sánchez, Ignacio Delgado Espinós, Laura Moreno Sarrion, Alfredo Quijano López, and Isabel Navalón Burgos. Clients segmentation according to their domestic energy consumption by the use of self-organizing maps. In *Proceedings of the 6th International Conference on the European Energy Market (EEM '09)*, pages 1–6, Leuven, Belgium, May 2009. IEEE.
- [167] Marla Sanchez, Richard Brown, Greg Homan, and Carrie Webber. Savings estimates for the United States Environmental Protection Agency’s ENERGY STAR voluntary product labeling program. Technical report, June 2008.
- [168] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T. Campbell. NextPlace: A spatio-temporal prediction framework for pervasive systems. In *Proceedings of the 9th International Conference on Pervasive Computing (Pervasive '11)*, pages 152–169, San Francisco, USA, June 2011. IEEE.
- [169] James Scott, Alice Jane Bernheim Brush, John Krumm, Brian Meyers, Mike Hazas, Stephen Hodges, and Nicolas Villar. Preheat: Controlling home heating using occupancy prediction. In *Proceedings of the 13th ACM International Conference on Ubiquitous Computing (UbiComp '11)*, pages 281–290, Beijing, PRC, September 2011. ACM.
- [170] Zach Shelby, Klaus Hartke, and Carsten Bormann. The Constrained Application Protocol (CoAP). RFC 7252 (Proposed Standard), June 2014.
- [171] Zach Shelby, P. Mahonen, J. Riihijarvi, O. Raivio, and P. Huuskonen. NanoIP: The zen of embedded networking. In *Proceedings of the 38th IEEE International Conference on Communications (ICC '03)*, volume 2, pages 1218–1222, Anchorage, Alaska, USA, May 2003. IEEE.
- [172] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, October 2010.
- [173] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, February 2010.
- [174] Gerrit Tierie. *Cornelis Drebbel*. PhD thesis, Rijksuniversiteit Leiden, 1932.
- [175] Shoji Tominaga, Masamichi Shimosaka, Rui Fukui, and Tomomasa Sato. A unified framework for modeling and predicting going-out behavior. In *Proceedings of*

- the 10th International Conference on Pervasive Computing (Pervasive '12)*, pages 73–90, Newcastle, UK, June 2012. Springer.
- [176] Bryan Urban and Kurt Roth. A data-driven framework for comparing residential thermostat energy performance. Technical Report 2004-09, Fraunhofer Center for Sustainable Energy Systems, July 2014.
- [177] Bert Vande Meerssche, Geert Van Ham, Geert Deconinck, Jeroen Reynders, Mathias Spelier, and Nathalie Maes. Practical use of energy management systems. In *Proceedings of the 10th International Symposium on Ambient Intelligence and Embedded Systems (AmiEs '11)*, Chania, Crete, Greece, September 2011.
- [178] Mario Vašak, A. Starčić, and Anita Martinčević. Model predictive control of heating and cooling in a family house. In *Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO '11)*, pages 739–743, Opatija, Croatia, May 2011. IEEE.
- [179] Félix Iglesias Vázquez and Wolfgang Kastner. Clustering methods for occupancy prediction in smart home control. In *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE '11)*, pages 1321–1328, Gdansk, Poland, June 2011. IEEE.
- [180] Sergio Valero Verdú, Mario Ortiz Garcia, Carolina Senabre, A Gabaldón Marín, and Francisco J. García Franco. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21(4):1672–1682, November 2006.
- [181] Jan Široký, Frauke Oldewurtel, Jiří Cigler, and Samuel Prívara. Experimental analysis of model predictive control for an energy efficient building heating system. *Applied Energy*, 88(9):3079–3087, September 2011.
- [182] A. Wayne Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 100(9):1100–1103, September 1971.
- [183] Martin Wisy. Smart Message Language version 1.03, November 2008.
- [184] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
- [185] Longqi Yang, Kevin Ting, and Mani B. Srivastava. Inferring occupancy from opportunistically available sensor data. In *Proceedings of the 12th International Conference on Pervasive Computing and Communications (PerCom '14)*, pages 60–68, Budapest, Hungary, March 2014.

- [186] Rayoung Yang and Mark W. Newman. Learning from a learning thermostat: Lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*, pages 93–102, Zurich, Switzerland, September 2013. ACM.
- [187] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. Mining individual life pattern based on location history. In *Proceedings of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware (MDM '09)*, pages 1–10, Taipei, Taiwan, May 2009. IEEE.
- [188] Michael Zeifman and Kurt Roth. Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*, 57(1):76–84, February 2011.
- [189] Ahmed Zoha, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors*, 12(12):16838–16866, December 2012.

Referenced Web Resources

All links were accessed on 22nd February 2015 and were correct at the time of publication.

- [190] Air Wick air fresheners. <http://www.airwick.com>.
- [191] dojo JavaScript toolkit. <http://dojotoolkit.org>.
- [192] ecobee thermostat. <https://www.ecobee.com>.
- [193] Flukso community metering. <http://www.flukso.net>.
- [194] OpenWrt – Table of hardware. <http://wiki.openwrt.org/toh/start>.
- [195] Plugwise smart plugs. <http://www.plugwise.com>.
- [196] tado° smart thermostat. <http://www.tado.com/en>.
- [197] Telefónica Smart Steps.
<http://dynamicinsights.telefonica.com/488/smart-steps-2>.
- [198] The Nest Learning Thermostat. <http://www.nest.com>.
- [199] python-plugwise.
<https://bitbucket.org/hadara/python-plugwise/wiki/Home>, March 2011.
- [200] SheevaPlug mini computer.
<http://www.globalscaletechnologies.com/t-sheevaplugs.aspx>, 2012.
- [201] Monica Brooks. A Century of Progress 1933-34 Chicago World's Fair.
http://users.marshall.edu/~brooks/1933_Chicago_World_Fair.htm, 2013.

Referenced Web Resources

- [202] Bundesministerium für Wirtschaft und Energie. Zahlen und Fakten Energiedaten [spreadsheet]. <http://www.bmwi.de/BMWi/Redaktion/Binaer/energie-daten-gesamt,property=blob,bereich=bmwi2012,sprache=de,rwb=true.xls>, October 2014.
- [203] Drury B. Crawley. EnergyPlus: DOE's next generation simulation program [presentation]. http://www1.eere.energy.gov/buildings/pdfs/eplu_webinar_02-16-10.pdf, February 2010.
- [204] Vincent de Groot. Honeywell round thermostat [image]. http://commons.wikimedia.org/wiki/File:Honeywell_thermostat.jpg, June 2006.
- [205] Department of Energy and Climate Change. Energy Consumption in the UK – Overall data tables [spreadsheet]. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/358232/overall.xls, 2014.
- [206] dominiquechappard. Al sleeping [clipart]. <https://openclipart.org/detail/91651/al-sleeping-by-cybergedeon>, October 2010.
- [207] dominiquechappard. Al walking [clipart]. <https://openclipart.org/detail/91549/al-walking-by-cybergedeon>, October 2010.
- [208] Electrical & Mechanical Services Department. Hong Kong energy end-use data 2014. http://www.emsd.gov.hk/emsd/e_download/pee/HKEEUD2014.pdf, 2014.
- [209] European Environment Agency. Household energy consumption by end-use in the EU-27 [spreadsheet]. http://www.eea.europa.eu/data-and-maps/figures/households-energy-consumption-by-end-uses-4/ener22_fig6_excel/at_download/file, November 2012.
- [210] Juri Glass, Mathias Runge, and Nadim El Sayed. libSML. <https://github.com/dailab/libsm1>.
- [211] Honeywell. Honeywell FocusPRO 6000 programmable thermostat [image]. https://www.forwardthinking.honeywell.com/products/thermostats/thermostat_products.html, 2014.

- [212] Honeywell International Inc. Honeywell history.
<http://honeywell.com/About/Pages/our-history.aspx>.
- [213] International Code Council. International Energy Conservation Code.
<http://publicecodes.cyberregs.com/icod/iecc/>, 2011.
- [214] Wilhelm Kleiminger and Christian Beckel. ECO (Electricity Consumption and Occupancy) data set.
<http://vs.inf.ethz.ch/res/show.html?what=eco-data>.
- [215] MathWorks. ClassificationKNN class.
<http://ch.mathworks.com/help/stats/classificationknn-class.html>.
- [216] MathWorks. gmdistribution class.
<http://ch.mathworks.com/help/stats/gmdistribution-class.html>.
- [217] Michael C. Mozer. Neural network house – Sensor panel [image].
<http://www.cs.colorado.edu/~mozer/Research/Projects/Adaptive%20house/photos/sensors/tls.gif>.
- [218] Daniel Pauli and Wilhelm Kleiminger. libSML.
<https://github.com/wkleiminger/pylon>.
- [219] LLC Thermal Energy System Specialists. TRNSYS: Transient System Simulation Tool. <http://www.trnsys.com/>, January 2015.
- [220] U.S. Department of Energy. EnergyPlus energy simulation software.
<http://apps1.eere.energy.gov/buildings/energyplus/>.
- [221] U.S. Department of Energy. Buildings Energy Data Book.
<http://buildingsdatabook.eren.doe.gov/ChapterIntro1.aspx>, March 2012.
- [222] U.S. Energy Information Administration. 13th Residential Energy Consumption Survey (RECS) [spreadsheet].
http://www.eia.gov/consumption/residential/data/2009/xls/HC6%20RECS%20Household%20Characteristics_Final%20Tables.ZIP, May 2013.
- [223] Wikipedia. Thermal transmittance.
http://en.wikipedia.org/wiki/Thermal_transmittance, February 2015.

Appendix A

Questionnaires

This appendix contains the replies given by the participants in the questionnaires prior to the deployment. We have omitted data that was necessary for the deployment but could identify individual participants. The responses of participants who were not selected for a deployment are not listed.

Table A.1: Overview of the participants (detailed). FT: *full-time employment*, PT: *part-time employment*, HM: *house-maker*, S: *student*.

Using electricity for:

#	Tech. affinity	Type of property	Occupants (employment, age)	Pets	# Entrances	Heating	Hot water	Vacancy / h per day
r1	7/7	House	(FT, 33), (HM, 33), (-, 3), (-, 1)	-	1	yes, separate meter	no	3
r2	7/7	Flat	(FT 34), (PT (50%), 32)	-	1	no	no	5
r3	7/7	House	(FT, 40), (other, 40)	-	1	yes, separate meter	yes	3
r4	4/7	House	(FT, 55), (HM, 49), (S, 17), (S, 15)	1	2	no	yes	2
r5	6/7	House	(FT, 62), (HM, 64)	1	3	no	yes	1
r6	6/7	House	2	n/a	n/a	n/a	n/a	n/a

Appendix B

1-Resistance 1-Capacitance (1R1C) model

This appendix contains the complete derivation of the transient heat transfer equation for the 1-resistance 1-capacitance model introduced in Chapter 7. In addition, we show that the energy required to heat up the system from Θ_{comf} to Θ_{setb} asymptotically approaches $(\Theta_{\text{comf}} - \Theta_{\text{setb}}) \times C$ as t tends to zero.

B.1 Derivation

Figure B.1 shows the RC circuit of a simple 1-resistance 1-capacitance model. In this model, the energy input to the system must be equal to the energy lost and the energy stored by the building.

$$\dot{E}_{\text{in}} = \dot{E}_{\text{out}} + \dot{E}_{\text{stored}} \quad (\text{B.1})$$

Introducing the temperature difference between the indoor Θ_{in} and outside temperatures Θ_e , the Equation B.1 may be re-written as:

$$\dot{E}_{\text{in}} = \frac{\Theta_{\text{in}} - \Theta_e}{R} + C \frac{d\Theta_{\text{in}}}{dt} \quad (\text{B.2})$$

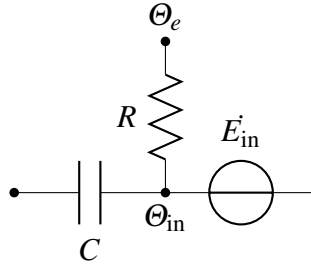


Figure B.1: 1-resistance 1-capacitance (1R1C) model.

where R stands for the thermal resistance ($R = \frac{1}{hA_s}$) and C for the thermal capacitance ($C = \rho Vc$) of the building components facing the outside, respectively.

Now equation (B.2) may be rewritten as:

$$\dot{E}_{in} = \frac{\Theta_{in}}{R} - \frac{\Theta_e}{R} + C \frac{d\Theta_{in}}{dt} \quad (B.3)$$

$$\frac{d\Theta_{in}}{dt} + \frac{\Theta_{in}}{RC} = \frac{\dot{E}_{in}}{C} + \frac{\Theta_e}{R} \quad (B.4)$$

$$\frac{d\Theta_{in}}{dt} + \frac{\Theta_{in}}{RC} = \frac{\dot{E}_{in}R + \Theta_e}{RC} \quad (B.5)$$

Multiplying by integrating factor $e^{\int \frac{1}{RC} dt} = e^{\frac{t}{RC}}$ and applying the product rule in reverse gives:

$$\int \frac{d\Theta_{in}}{dt} e^{\frac{t}{RC}} = \int \frac{\dot{E}_{in}R + \Theta_e}{RC} e^{\frac{t}{RC}} \quad (B.6)$$

$$\Theta_{in}(t) e^{\frac{t}{RC}} = \frac{\dot{E}_{in}R + \Theta_e}{RC} (RC e^{\frac{t}{RC}} + D) \quad (B.7)$$

$$\Theta_{in}(t) = \frac{\dot{E}_{in}R + \Theta_e}{RC} e^{-\frac{t}{RC}} (RC e^{\frac{t}{RC}} + D) \quad (B.8)$$

$$\Theta_{in}(t) = (\dot{E}_{in}R + \Theta_e) + D e^{-\frac{t}{RC}} \frac{\dot{E}_{in}R + \Theta_e}{RC} \quad (B.9)$$

Fixing $\Theta_{in}(0) = \Theta_{in}(t-1)$ for $t=0$ we can calculate D :

$$\Theta_{in}(t-1) = (\dot{E}_{in}R + \Theta_e) + D \frac{\dot{E}_{in}R + \Theta_e}{RC} \quad (B.10)$$

$$D = [\Theta_{in}(t-1) - (\dot{E}_{in}R + \Theta_e)] \frac{RC}{\dot{E}_{in}R + \Theta_e} \quad (B.11)$$

Substituting D in $\Theta_{in}(t)$ gives the indoor temperature at time $t - \Theta_{in}(t)$ – as a function of the indoor temperature at the previous interval $\Theta_{in}(t-1)$, the outside temperature Θ_e , the resistance and capacitance values (R and C) and the heat added to the system \dot{E}_{in} :

$$\Theta_{in}(t) = (\dot{E}_{in}R + \Theta_e) + [\Theta_{in}(t-1) - (\dot{E}_{in}R + \Theta_e)] \frac{RC}{\dot{E}_{in}R + \Theta_e} e^{-\frac{t}{RC}} \frac{\dot{E}_{in}R + \Theta_e}{RC} \quad (B.12)$$

$$\Theta_{in}(t) = (\dot{E}_{in}R + \Theta_e) + [\Theta_{in}(t-1) - (\dot{E}_{in}R + \Theta_e)] e^{-\frac{t}{RC}} \quad (B.13)$$

$$\Theta_{\text{in}}(t) = \Theta_{\text{in}}(t-1)e^{\frac{-t}{RC}} + (E_{\text{in}}R + \Theta_e)(1 - e^{\frac{-t}{RC}}) \quad (\text{B.14})$$

B.2 Convergence

$$\Theta_{\text{comf}} = \Theta_{\text{setb}}e^{\frac{-t}{RC}} + (E_{\text{in}}R + \Theta_e)(1 - e^{\frac{-t}{RC}}) \quad (\text{B.15})$$

$$\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}} = (E_{\text{in}}R + \Theta_e)(1 - e^{\frac{-t}{RC}}) \quad (\text{B.16})$$

$$\frac{\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}}}{1 - e^{\frac{-t}{RC}}} = E_{\text{in}}R + \Theta_e \quad (\text{B.17})$$

$$\frac{\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}}}{1 - e^{\frac{-t}{RC}}} - \Theta_e = E_{\text{in}}R \quad (\text{B.18})$$

$$E_{\text{in}} = \frac{\frac{\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}}}{1 - e^{\frac{-t}{RC}}} - \Theta_e}{R} \quad (\text{B.19})$$

$$E_{\text{min}} = \lim_{t \rightarrow 0} t \times \frac{\frac{\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}}}{1 - e^{\frac{-t}{RC}}} - \Theta_e}{R} \quad (\text{B.20})$$

$$E_{\text{min}} = \lim_{t \rightarrow 0} t \times \frac{\Theta_{\text{comf}} - \Theta_{\text{setb}}e^{\frac{-t}{RC}}}{R - Re^{\frac{-t}{RC}}} - \frac{\Theta_e}{R} \quad (\text{B.21})$$

$$E_{\text{min}} = \lim_{t \rightarrow 0} \frac{t\Theta_{\text{comf}} - t\Theta_{\text{setb}}e^{\frac{-t}{RC}}}{R - Re^{\frac{-t}{RC}}} - \frac{t\Theta_e}{R} \quad (\text{B.22})$$

Since $\frac{t\Theta_e}{R}$ goes to zero we can simplify as follows:

$$E_{\text{min}} = \lim_{t \rightarrow 0} \frac{t\Theta_{\text{comf}} - t\Theta_{\text{setb}}e^{\frac{-t}{RC}}}{R - Re^{\frac{-t}{RC}}} \quad (\text{B.23})$$

Since as $\lim_{t \rightarrow 0} t\Theta_{\text{comf}} - t\Theta_{\text{setb}}e^{\frac{-t}{RC}} = \lim_{t \rightarrow 0} R - Re^{\frac{-t}{RC}} = 0$ we can employ l'Hôpital's rule:

$$\lim_{t \rightarrow c} \frac{f(t)}{g(t)} = \lim_{t \rightarrow c} \frac{f'(t)}{g'(t)} \quad (\text{B.24})$$

where

$$f(t) = t\Theta_{\text{comf}} - t\Theta_{\text{setb}}e^{\frac{-t}{RC}} \quad (\text{B.25})$$

Applying the product rule twice

$$f'(t) = \Theta_{\text{comf}} - \Theta_{\text{setb}}(e^{\frac{-t}{RC}} + \frac{-t}{RC}e^{\frac{-t}{RC}}) \quad (\text{B.26})$$

$$\lim_{t \rightarrow 0} f'(x) = \lim_{t \rightarrow 0} \Theta_{\text{comf}} - \Theta_{\text{setb}}(e^{\frac{-t}{RC}} + \frac{-t}{RC}e^{\frac{-t}{RC}}) = \Theta_{\text{comf}} - \Theta_{\text{setb}} \quad (\text{B.27})$$

and

$$g(t) = R - Re^{\frac{-t}{RC}} \quad (\text{B.28})$$

applying the chain rule once:

$$g'(t) = \frac{R}{RC}e^{\frac{-t}{RC}} \quad (\text{B.29})$$

Thus

$$\lim_{t \rightarrow 0} g'(x) = \lim_{t \rightarrow 0} \frac{R}{RC}e^{\frac{-t}{RC}} = \frac{1}{C} \quad (\text{B.30})$$

Therefore as $\lim_{t \rightarrow c} \frac{f(t)}{g(t)} = \lim_{t \rightarrow c} \frac{f'(t)}{g'(t)}$:

$$E_{\text{min}} = (\Theta_{\text{comf}} - \Theta_{\text{setb}}) \times C \quad (\text{B.31})$$

Occupancy prediction

This appendix contains a description of the LDCC data used to evaluate the occupancy prediction algorithms presented in Chapter 6. It further includes the prediction accuracy of the MAT, MDMAT, PP(S) and PH algorithms for all 45 participants. A detailed overview of the results is shown in Section C.1. For reference, Section C.2 includes the probabilistic schedules of all 45 participants.

C.1 Dataset overview and prediction results

Table C.1: Complete results in percent, sorted by occupancy. Π^{max} is the predictability of the schedules.

LDCC #	Occupancy	Π^{max}	# Days	Prediction accuracy				
				MAT	MDMAT	PP	PPS	PH
6060	97	94	83	67	69	97	97	97
5986	95	95	153	73	82	95	95	95
5955	95	93	35	74	88	93	93	94
5977	95	93	36	67	87	92	92	94
6012	94	94	72	63	79	93	93	94
5976	92	94	115	72	81	92	92	92
5988	90	87	44	65	80	87	87	90
6032	87	91	97	70	78	86	86	84
6175	85	94	31	53	77	88	88	86
5924	85	87	45	73	76	84	84	84
5958	84	92	58	60	73	82	82	84
5961	84	93	51	78	78	92	91	84
6078	83	95	199	72	74	85	85	83
5972	81	87	163	69	75	81	81	77
6082	79	93	65	76	77	84	84	76
6096	78	87	79	61	64	77	76	65

Continued on next page

Table C.1 – continued from previous page

Prediction accuracy								
LDDC #	Occupancy	Π^{max}	# Days	MAT	MDMAT	PP	PPS	PH
6076	78	93	44	77	79	83	83	75
6031	77	95	142	80	80	89	89	78
5943	75	93	76	80	81	86	86	81
5966	74	93	77	78	76	91	91	79
6104	72	90	102	77	76	87	87	75
6063	71	93	35	77	79	86	86	73
5985	71	90	57	76	76	81	81	75
6198	71	90	41	78	77	80	80	74
6014	71	89	43	72	73	74	73	71
5987	70	93	35	83	77	83	83	84
6039	69	86	98	82	82	92	92	81
5936	69	92	40	76	77	85	85	71
6077	68	93	44	84	82	87	87	81
5960	68	92	90	85	79	94	94	88
5980	67	93	74	81	81	91	91	80
5942	67	95	111	86	85	90	90	85
6061	66	81	71	71	72	76	76	62
6040	66	93	103	85	82	92	92	83
5962	65	85	66	88	85	91	91	87
6075	65	93	40	88	87	91	91	88
6168	64	82	86	79	76	89	89	77
6033	62	94	31	78	74	81	81	70
6097	61	86	90	76	76	81	81	75
6007	59	90	32	83	82	84	84	82
6010	58	87	62	80	77	85	85	78
5957	58	82	48	79	77	86	86	79
5965	53	83	75	73	71	77	77	67
5978	52	89	132	72	68	75	75	57
6043	43	82	68	76	76	78	78	74
Average:	74	90	74	75	78	86	86	80

C.2 Probabilistic schedules

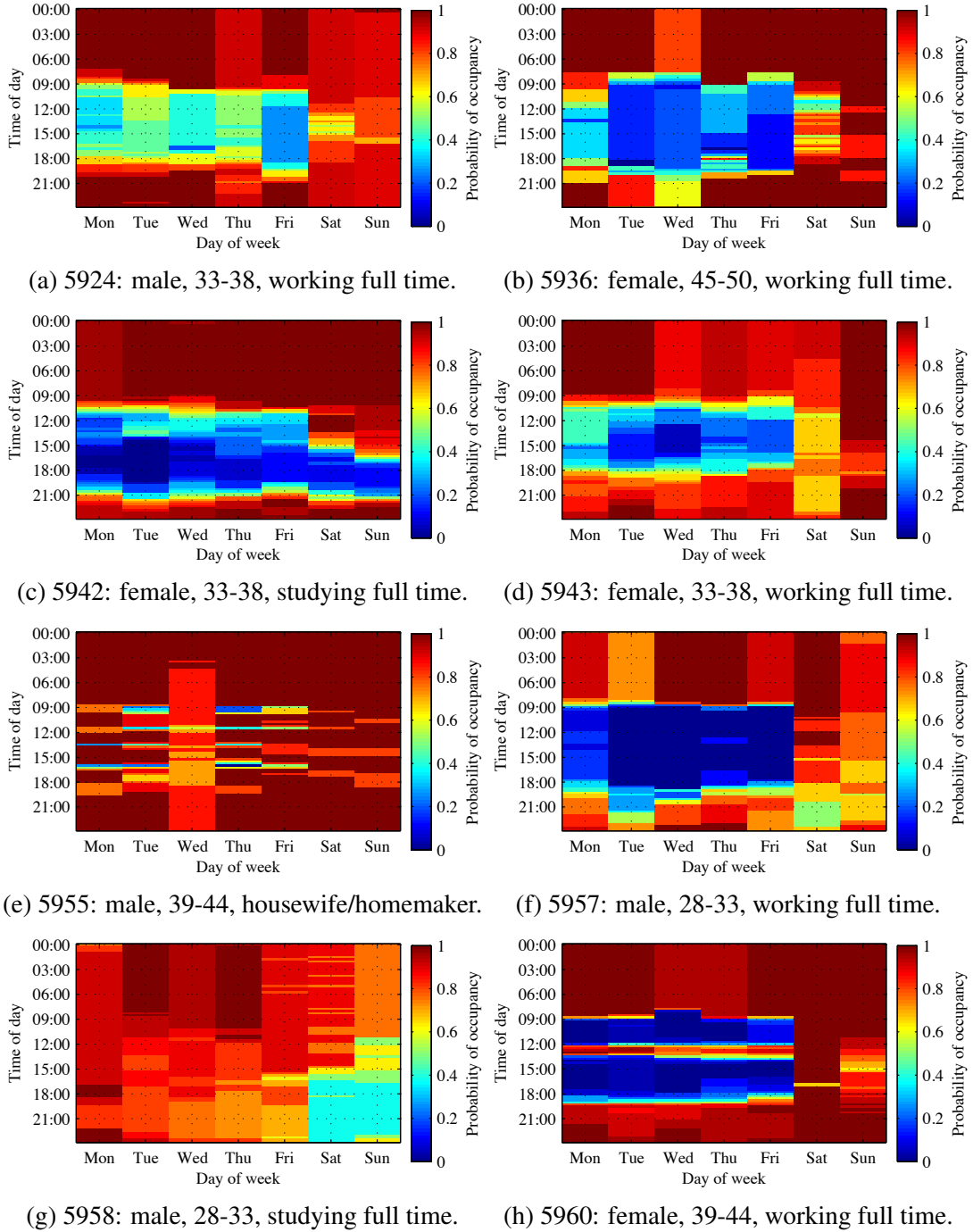


Figure C.1: Probabilistic occupancy schedules (households 5924 to 5960).

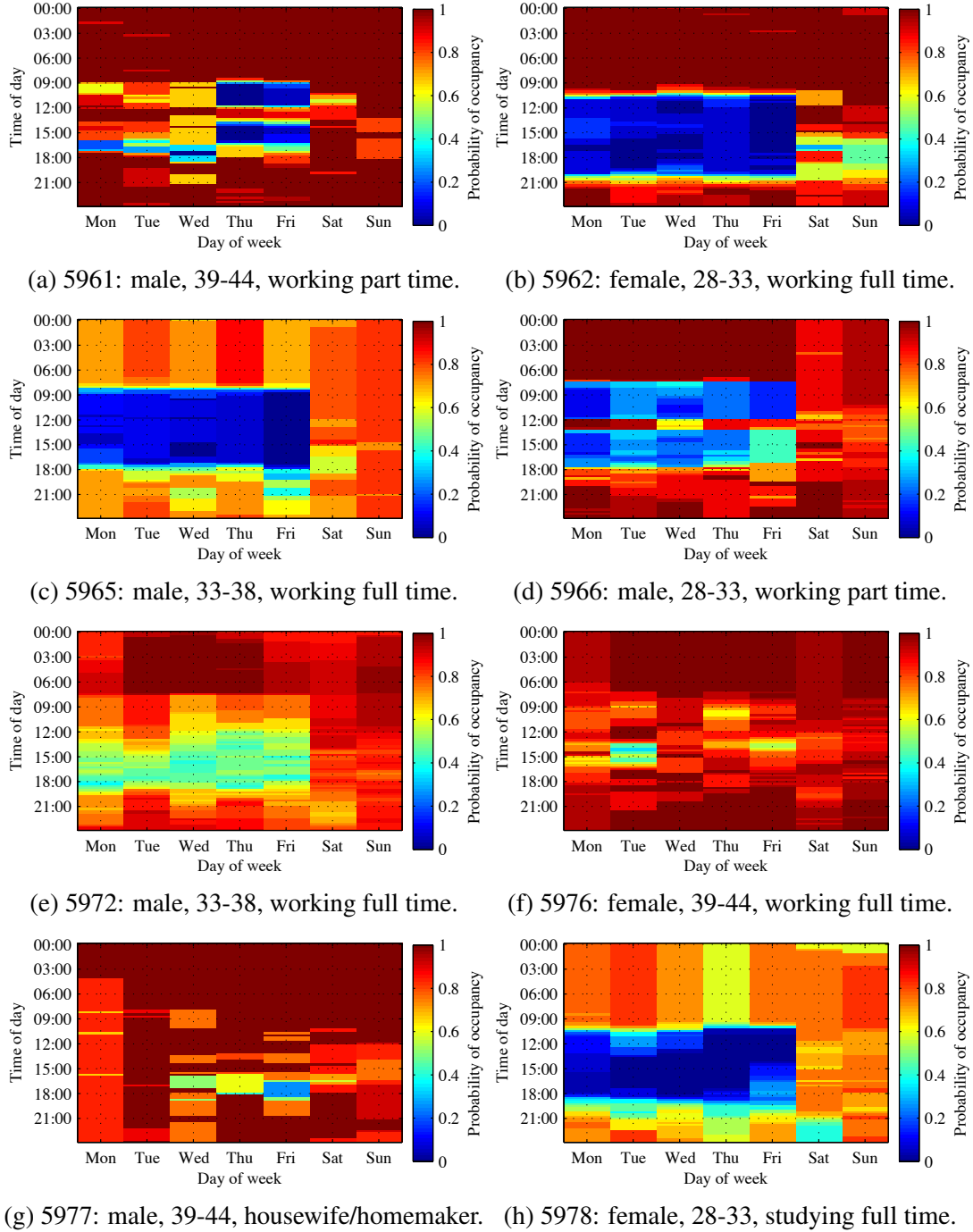


Figure C.2: Probabilistic occupancy schedules (households 5961 to 5978).

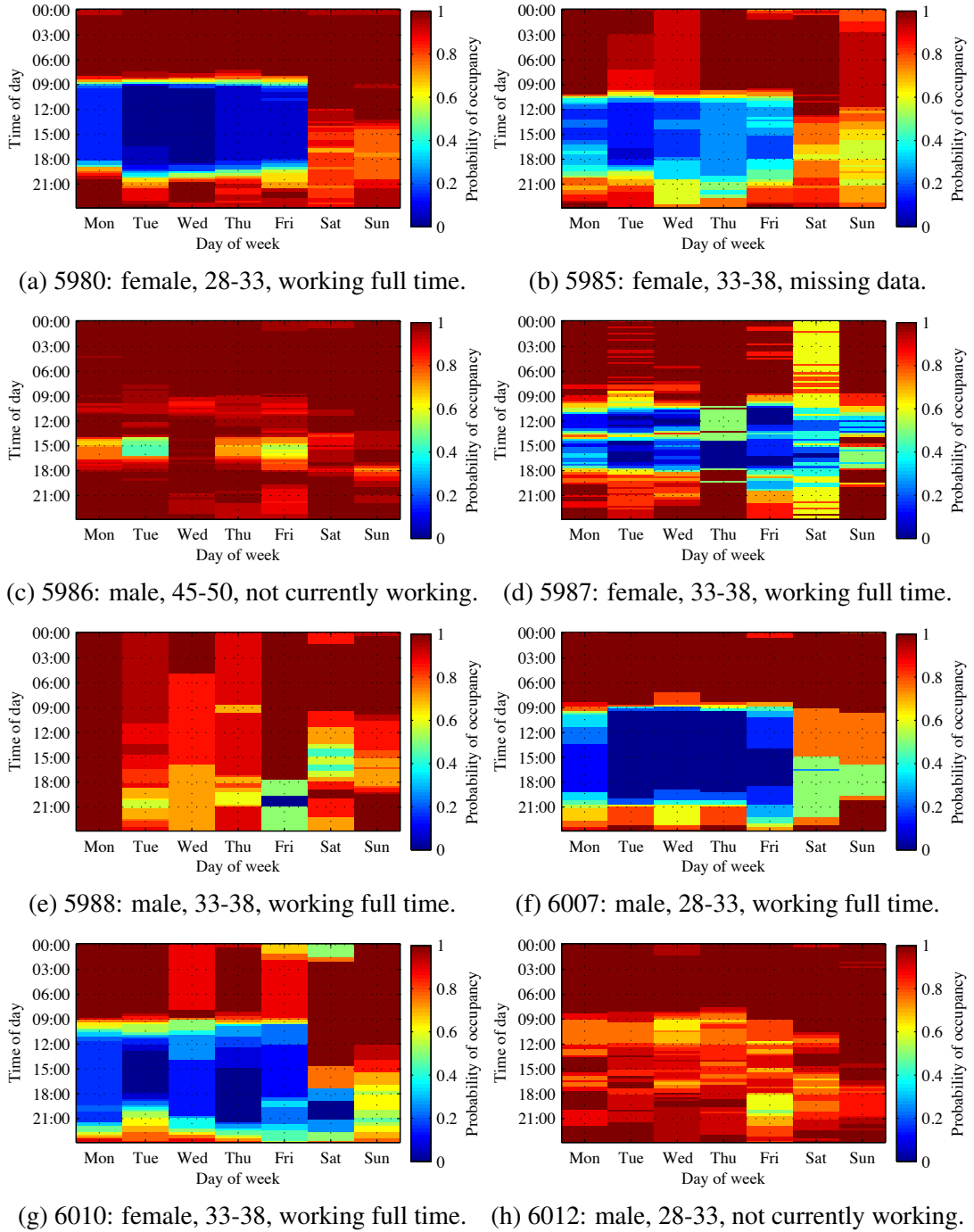


Figure C.3: Probabilistic occupancy schedules (households 5980 to 6012).

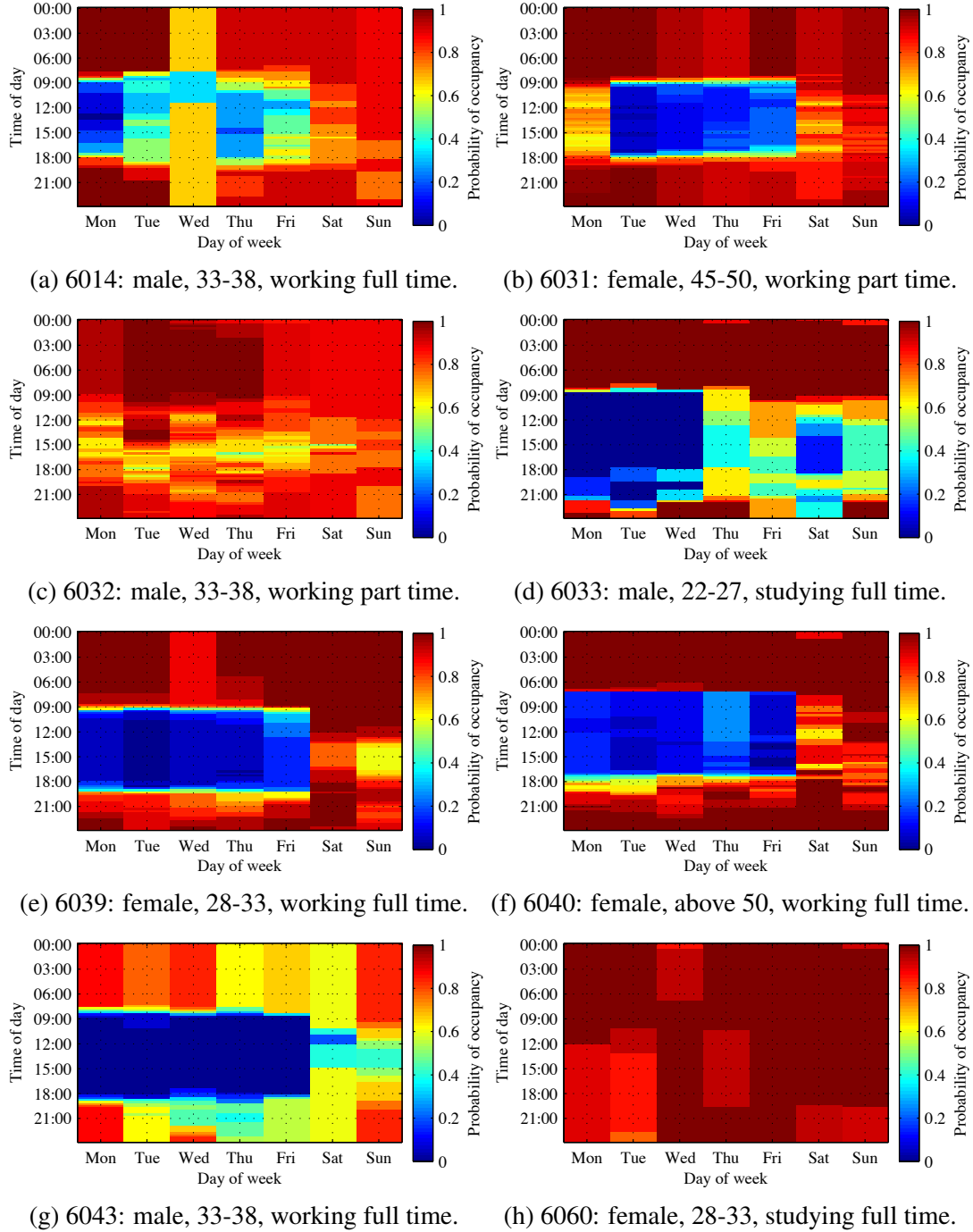


Figure C.4: Probabilistic occupancy schedules (households 6014 to 6060).

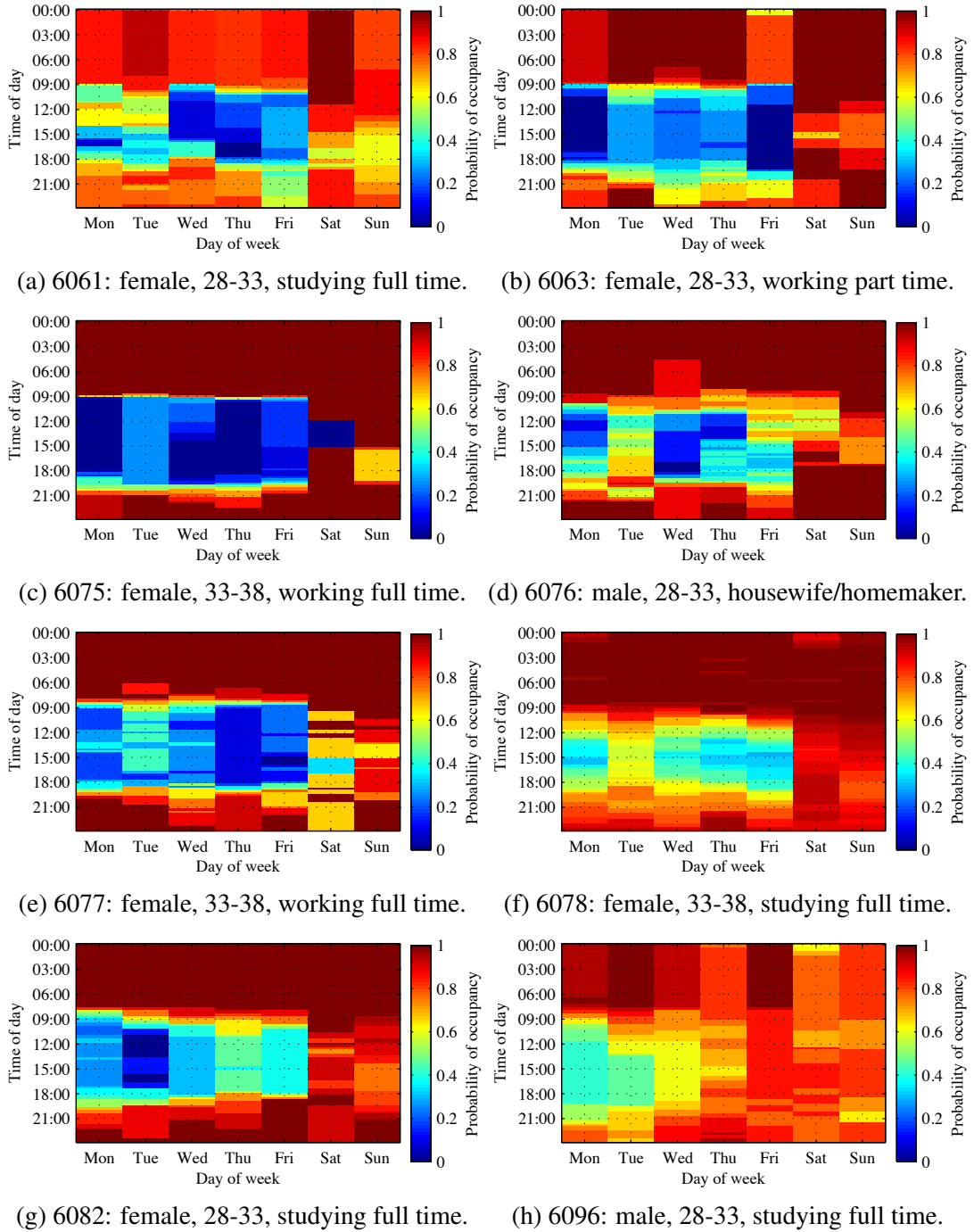


Figure C.5: Probabilistic occupancy schedules (households 6061 to 6096).

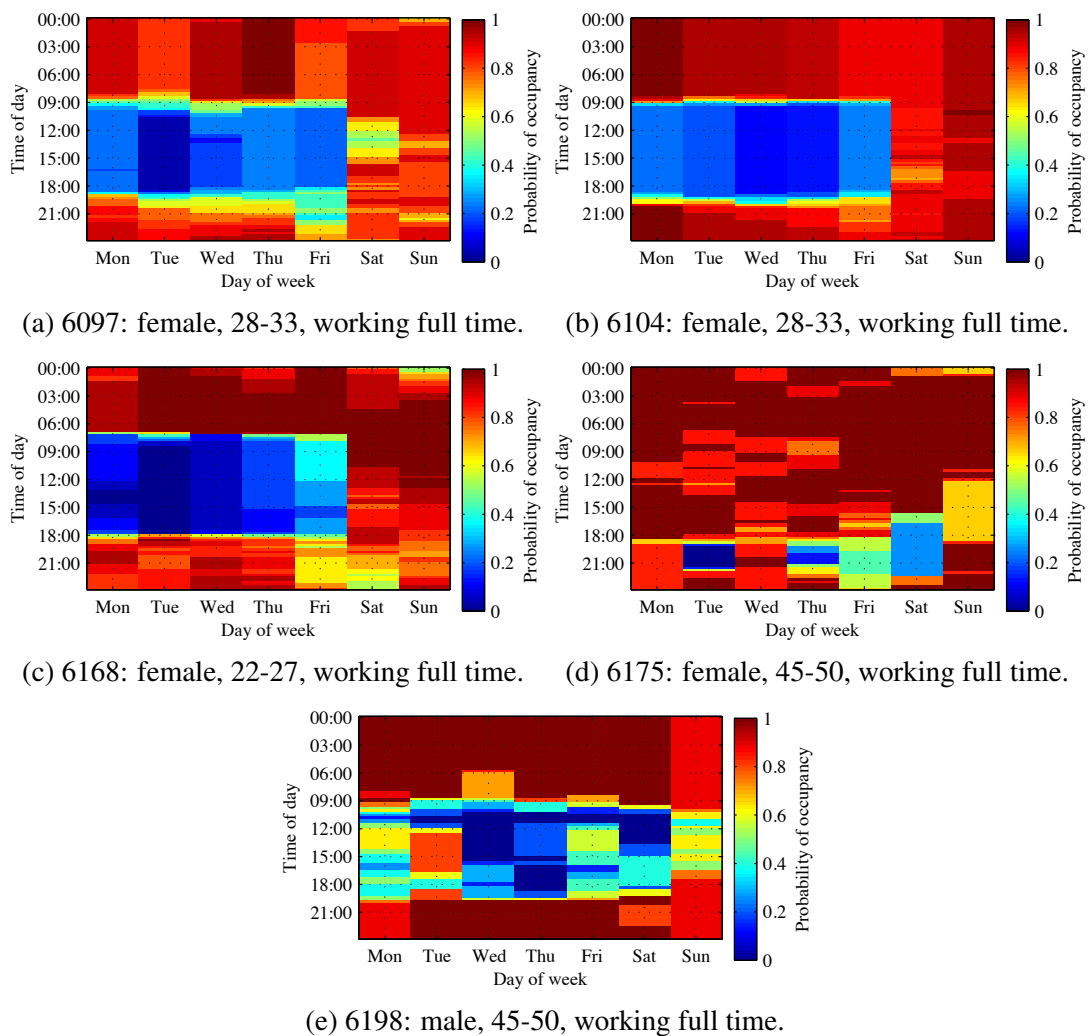


Figure C.6: Probabilistic occupancy schedules (households 6097 to 6198).

Appendix D

Simulation scenarios

This appendix shows the details of all 32 building and weather scenarios derived in Chapter 7 and used in the evaluation of the energy savings of a smart heating system in Chapter 8. Figures D.1 to D.4 show the typical behaviour of a heating system according to the ISO 5R1C model for a scenario where the building ($F-U_{\text{low}}$, $F-U_{\text{high}}$, $H-U_{\text{low}}$ and $H-U_{\text{high}}$) is unoccupied between 9 a.m. and 5 p.m.

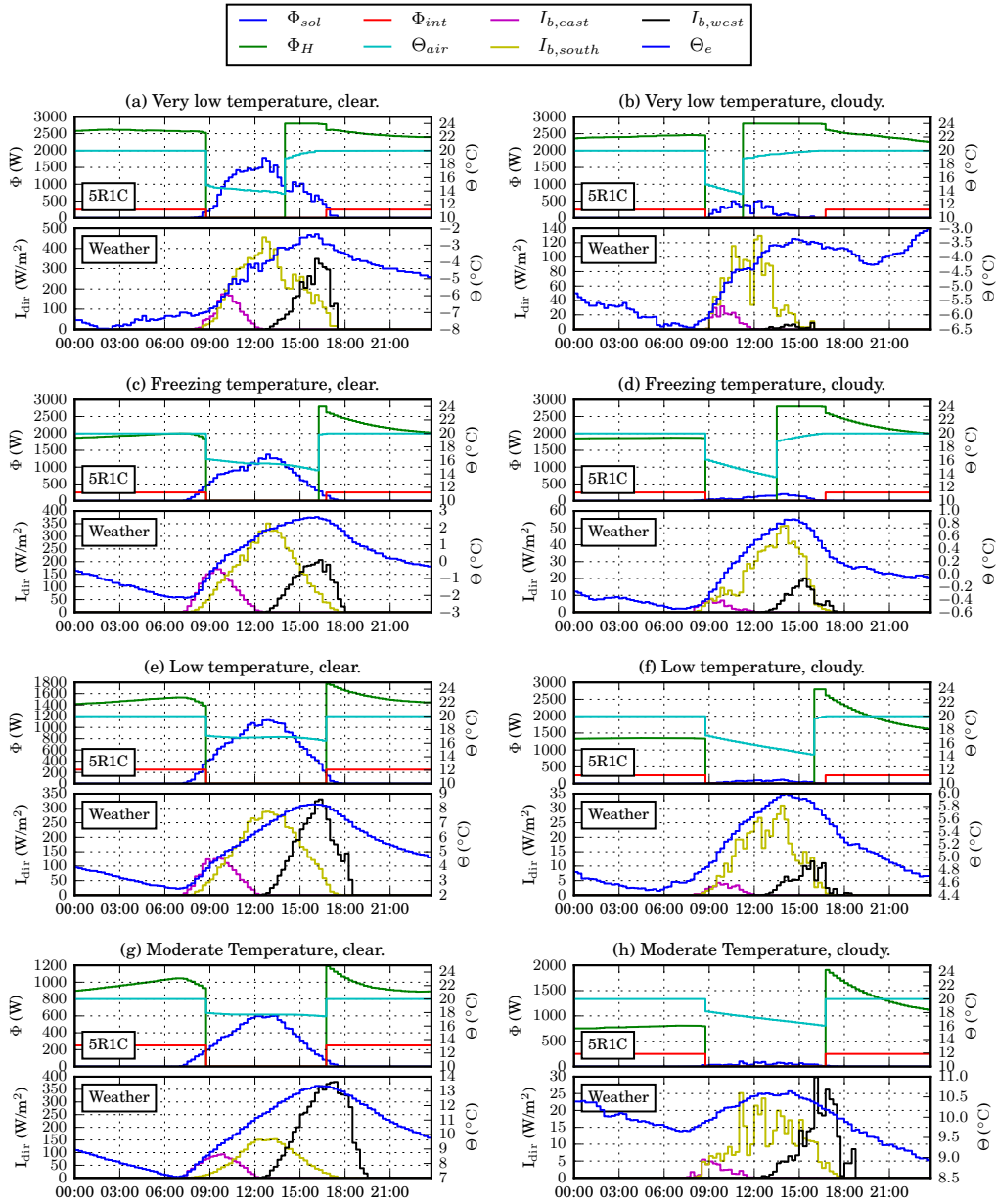


Figure D.1: Typical behaviour of a heating system according to the ISO 5R1C model for a scenario where the **well insulated flat** ($F-U_{low}$) is unoccupied between 9 a.m. and 5 p.m. For each, (a) to (h), the upper part shows the heat inputs of the 5R1C model (solar gain Φ_{sol} , heat input Φ_H and internal gain Φ_{int}) and the resulting indoor air temperature Θ_{air} , while the lower part shows the direct radiation $I_{b,\{east,south,west\}}$ and outside temperature Θ_e of the weather scenario.

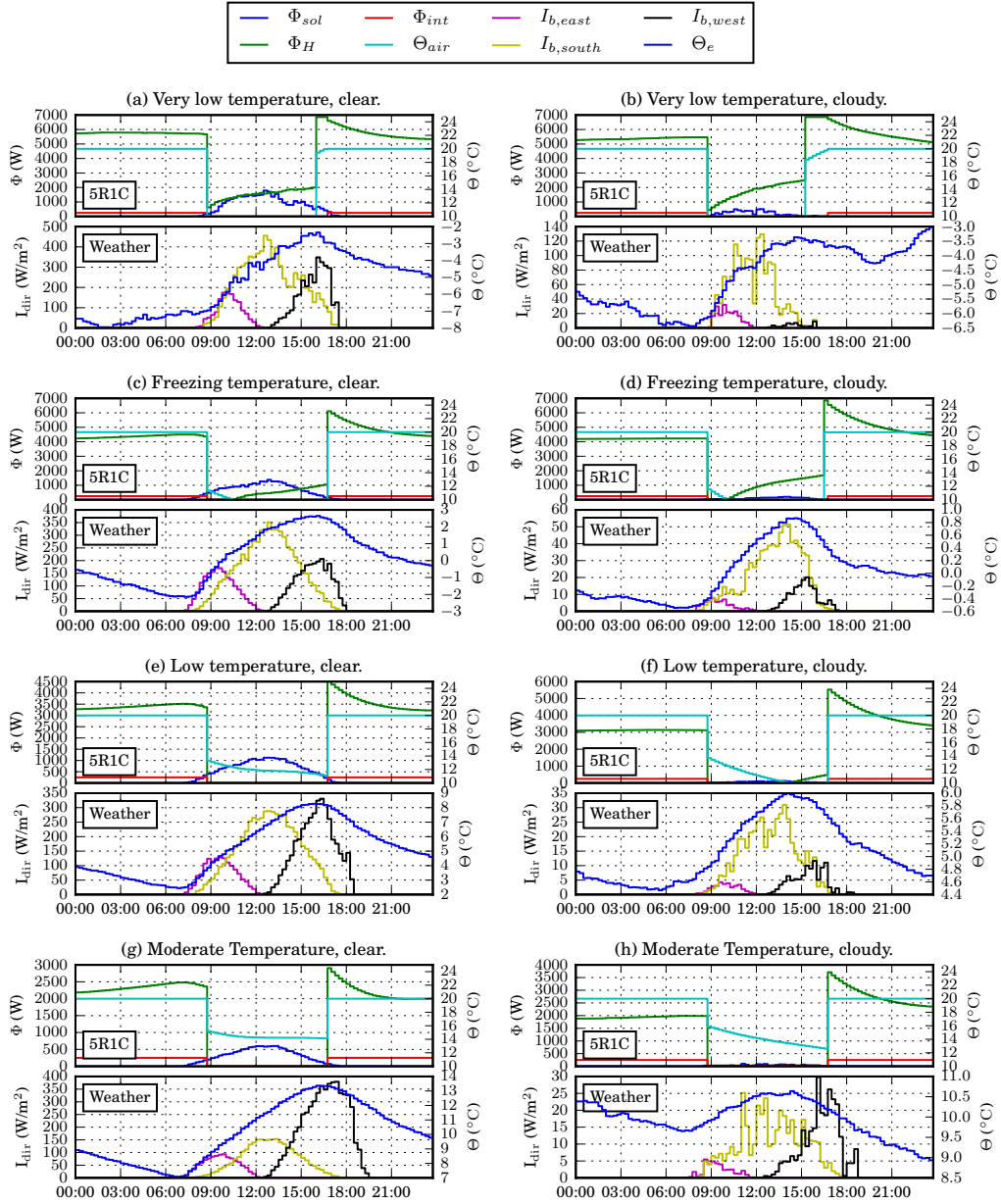


Figure D.2: Typical behaviour of a heating system according to the ISO 5R1C model for a scenario where the **poorly insulated flat** ($F-U_{\text{high}}$) is unoccupied between 9 a.m. and 5 p.m. For each, (a) to (h), the upper part shows the heat inputs of the 5R1C model (solar gain Φ_{sol} , heat input Φ_H and internal gain Φ_{int}) and the resulting indoor air temperature Θ_{air} , while the lower part shows the direct radiation $I_{b,\{\text{east},\text{south},\text{west}\}}$ and outside temperature Θ_e of the weather scenario.

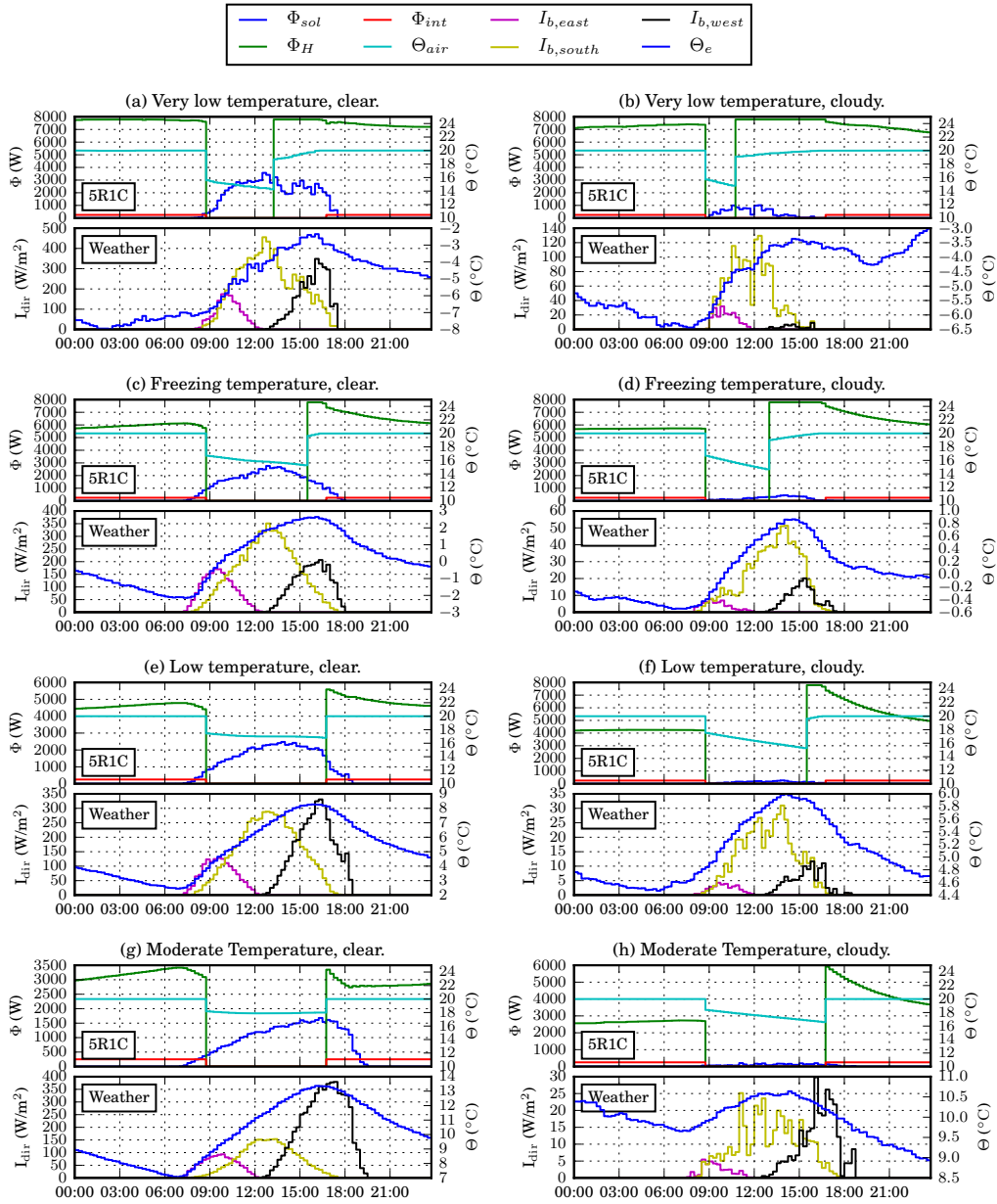


Figure D.3: Typical behaviour of a heating system according to the ISO 5R1C model for a scenario where the **well insulated house** ($H-U_{low}$) is unoccupied between 9 a.m. and 5 p.m. For each, (a) to (h), the upper part shows the heat inputs of the 5R1C model (solar gain Φ_{sol} , heat input Φ_H and internal gain Φ_{int}) and the resulting indoor air temperature Θ_{air} , while the lower part shows the direct radiation $I_{b,\{east,south,west\}}$ and outside temperature Θ_e of the weather scenario.

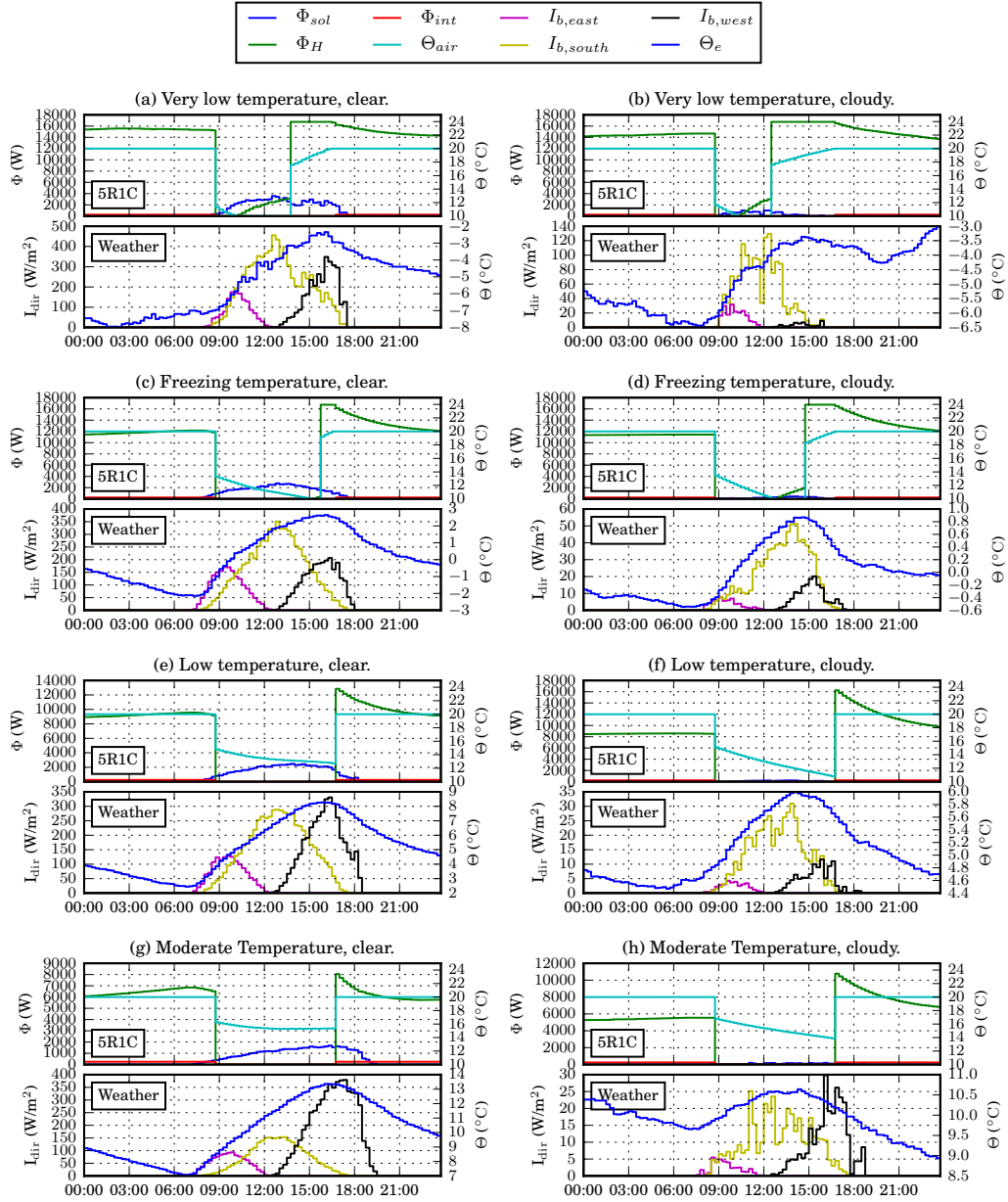


Figure D.4: Typical behaviour of a heating system according to the ISO 5R1C model for a scenario where the **poorly insulated house** ($H-U_{\text{high}}$) is unoccupied between 9 a.m. and 5 p.m. For each, (a) to (h), the upper part shows the heat inputs of the 5R1C model (solar gain Φ_{sol} , heat input Φ_H and internal gain Φ_{int}) and the resulting indoor air temperature Θ_{air} , while the lower part shows the direct radiation $I_{b,\{\text{east},\text{south},\text{west}\}}$ and outside temperature Θ_e of the weather scenario.