Using Smart Meter Data and Probabilistic Models to Augment Energy Performance Assessments of Residential Buildings

Joe Brown

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Contents

List of Tables

List of Acronyms

BEIS . . . . . . Department for business, energy and industrial strategy

CDF . . . . . . Cumulative distribution function

CPLC . . . . . Cooling power loss coefficient

ECO . . . . . . Energy company obligation

EPC . . . . . . Energy performance certificate

GDPR . . . . . General data protection regulation

GHG . . . . . Greenhouse gas

HDD . . . . . . heating degree days

HDH . . . . . . heating degree hours

HLC/HTC . . Heating loss (transfer) coefficient

HMC . . . . . Hamiltonian Monte Carlo

HMM . . . . . Hidden Markov model

HPLC . . . . . Heating power loss coefficient

IEA . . . . . . International energy agency

IPCC . . . . . Intergovernmental panel on climate change

MAPE . . . . Mean absolute percentage error

MCMC . . . . Markov chain Monte Carlo

MNBE . . . . Mean normalised bias error

NAC . . . . . . Normalised annual consumption

NILM . . . . . Non-intrusive load monitoring

NOCT . . . . . Nominal operating cell temperature

NUTS . . . . . No-U-turn sampler

PCC . . . . . . Pearson's correlation coefficient

PDF . . . . . . Probability density function

PNNL . . . . . Pacific Northwest National Laboratory

PTG . . . . . . Power temperature gradient

PV . . . . Photovoltaic

RdSAP . . . . Reduced standard assessment procedure

RMSE . . . . . Root mean square error

SAP . . . . . . Standard assessment procedure

SMETS . . . . Smart metering equipment technical specifications

List of Symbols

A_{sol} . . . . Effective solar aperture of building [m^2]

c_{air} . . . . Specific heat capacity of air [kW/kg]

c_h . . . . PV capacity [kW]

C_p . . . . Constant power due to appliances not used directly for heating [kW]

$CPLC$ . . . . . Cooling power loss coefficient [kW/K]

F_T . . . . Constant used to infer T_{in} from T_{ex}

f_{td} . . . . Solar proxy factor

g_{htd} . . . . Energy generation from PV panel [kW]

h_{on} . . . . Heating on indicator function

H_{pred} . . . . . . Predicted H(P)LC value [kW/K]

H_{real} . . . . . . Real H(P)LC value [kW/K]

$HPLC$ . . . . . Heating power loss coefficient [kW/K]

I . . . . Solar irradiance incident on tangent plane to Earth [kW]

I_{inc} . . . . Solar irradiance incident on PV panel [kW]

I_{sol} . . . . Solar irradiance incident on building [kW/m^2]

i_{sol} . . . . Observed solar irradiance incident on building [kW/m^2]

HLC . . . . . . Heating loss coefficient [kW/K]

k . . . . Scaling factor of Solar irradiance to PV power output

M_T . . . . Thermal mass of building [kW/K]

N_{occ} . . . . Number of occupants in the building

$NOCT$ . . . . . Nominal operating cell temperature [°C]

P_{all} . . . . Total power used by home [kW]

P_c . . . . Power supplied to the cooling system [kW]

p_c . . . . Observed power supplied to the cooling system [kW]

P_b . . . . Power supplied not directly for heating/cooling [kW]

p_b . . . . Observed power supplied not directly for heating/cooling [kW]

P_h . . . . Power supplied to the heating system [kW]

p_h . . . . Observed power supplied to the heating system [kW]

P_s . . . . Power output of PV panel [kW]

Q_b . . . . Heat flow due to appliances not used directly for heating [kW]

Q_{fa} . . . . Heat flow through building fabric [kW]

Q_h . . . . Heat flow due to heating system [kW]

Q_o . . . . Heat flow due to occupants [kW]

Q_{ve} . . . . Heat flow through ventilation [kW]

Q_{we} . . . . Heat flow due to weather [kW]

$Q_{\Delta t}$ . . . . Heat loss/gain change between time-steps [kW]

r_{htd} . . . . Smart-meter reading [kW]

s_{td} . . . . Clearsky irradiance factor

T_0 . . . . Daily average internal temperature given T_{fix} [°C]

T_{ex} . . . . External air temperature [°C]

t_{ex} . . . . Observed external air temperature [°C]

T_{fix} . . . . Fixed external temperature reference point [°C]

T_{in} . . . . Internal air temperature [°C]

t_{in} . . . . Observed internal air temperature [°C]

T_{Panel} . . . . . . Temperature of the PV panel [°C]

V_{air} . . . . Volume of airflow between external and internal environment [m^3]

x_{htd} . . . . Residential energy usage [kW]

X_{htd} . . . . . . Random variable representing residential energy usage [kW]

α_{ht} . . . . Shape parameter for gamma distribution

β_{ht} . . . . Rate parameter for gamma distribution

β . . . . Tilt angle of the PV panel [°]

γ . . . . Azimuth angle of the PV panel [°]

γ_S . . . . Azimuth angle of the sun [°]

ΔT . . . . Difference between external and internal air temperature [°C]

Θ_{ht} . . . . Set of probability distribution parameters

η_b . . . . Proportion of heat generation from non-heating appliances

η_c . . . . Efficiency of cooling system

η_h . . . . Efficiency of heating system

θ_z . . . . Zenith angle of the sun [°]

ρ_{air} . . . . Density of air [kg/m^3]

σ_i . . . . Standard deviation of the subscript term

τ . . . . Time-step size [hours]

ξ_i . . . . Noise term corresponding to subscript [kW]

$\boldsymbol{C}$ . . . . Set of PV capacity values for all houses

$\boldsymbol{F}_t$ . . . . Set of solar proxy values for all days

$\mathbf{H}_{pred}$ . . . . . . Set of predicted H(P)LC values

$\mathbf{H}_{real}$ . . . . . . Set of real H(P)LC values

$\boldsymbol{R}_{ht}$ . . . . Set of smart-meter readings for all days

$\boldsymbol{R}_{td}$ . . . . Set of smart-meter readings for all houses

$\boldsymbol{X}_{ht}$ . . . . Set of energy usage values for all days

$\boldsymbol{X}_{td}$ . . . . Set of energy usage values for all houses

$\boldsymbol{\alpha}_t$ . . . . Set of gamma distribution shape parameters for all houses

$\boldsymbol{\beta}_t$ . . . . Set of gamma distribution rate parameters for all houses

$\mathbb{I}_{htd}$ . . . . Indicator function for smart-meter censoring

$\mathbb{I}_{ht}$ . . . . Set of indicator functions for all days

$\mathbb{I}_{td}$ . . . . Set of indicator functions for all houses

1
Introduction

1.1 Motivation

To keep global warming to 1.5°C action must be taken to reduce greenhouse gas (GHG) emissions before 2030 according to the 2018 IPCC report [65]. If nothing is done climate change is predicted to cause rising sea levels, ocean acidification, and more extreme weather events [138, 55]. Such outcomes will have a catastrophic effect on human health, wildlife, and ecosystems across the globe [97, 35]. Action needs to be taken to mitigate global warming and climate change by reducing and replacing large sources of GHG emissions [65]. In order to achieve this a better understanding of energy usage and generation is required to identify where GHG reducing interventions can be taken.

Residential energy usage is one of the largest sources of GHG emissions, accounting for 29 % of energy usage, and 17 % of the total GHG emissions from the UK according to the 2020 carbon budget report [95]. This means that reducing and decarbonising residential energy usage can play a significant role in reducing global GHG emissions. Additionally, measures need to be taken to protect vulnerable people living in fuel poverty; this is defined as those who cannot keep their home warm at a reasonable cost, as per the UK Warm Homes and Energy Conservation Act [1]. Currently, the energy company obligation (ECO[1]) means that large-scale energy suppliers in the UK are obliged to help households take energy-saving measures, particularly those in fuel poverty.

To identify where energy-saving interventions can be taken, it is important that accurate evaluations of the energy performance of residential buildings are conducted. Across the EU and the UK an energy performance certificate (EPC) is produced as the result of an assessment which evaluates how energy efficient a building is and identifies where energy saving interventions can be taken. To produce an EPC for a new building the standard assessment procedure (SAP) is followed, typically before the residents move in, which evaluates the energy performance using the building specification and plans [23]. For existing buildings, if 10 years have passed since the EPC was last updated the reduced standard assessment procedure (RdSAP) is conducted which consists of an in-person survey of the house to update the EPC and identify opportunities for energy saving interventions [11]. Unfortunately, both of these processes are time-consuming, expensive, and often inaccurate [73, 85]. Hence, there is potential to augment the EPC with data-driven approaches which capture the real-world energy performance of buildings and are more accurate, scalable and low cost.

To understand the real world energy performance of residential buildings it is important that accurate measurements of energy usage are taken. The installation of

[1] `https://www.ofgem.gov.uk/environmental-programmes/eco`

smart meters provides insight into the energy usage of houses by providing data on a half-hourly basis if the homeowner chooses to opt-in, and daily if they do not. The UK smart meter roll-out is currently at 45 % of all gas and electric meters, according to the Q1 2021 update from the UK government, and all customers will be offered a smart meter by 2025[2] [12]. Smart meters provide insight into energy usage at a granularity and scale not previously available.

The access to more granular data at scale creates an opportunity to augment the EPC and provide more accurate, lower cost, and scalable methods to evaluate the energy performance of residential buildings. A more accurate EPC will enable individuals, energy providers and surveyors to better identify opportunities to improve the energy performance of residential buildings, and reduce the GHG emissions they are responsible for. Furthermore, it may be performed more frequently if it can be conducted remotely and cheaply with readily available data. By leveraging the smart-meter data, it provides a data source to build new tools to identify where the most effective energy-reducing interventions and shifts to renewable sources can be made.

To reduce energy usage in residential buildings it is important to address the areas of energy usage where the largest reductions can be made. For example, in cooler climates heating accounts for a large proportion of energy usage, such as in the UK where 65 % of total residential energy usage is used for heating [13]. However, as the middle class in countries with hotter climates grow, the demand for residential cooling is expected to increase which has led to the international energy agency (IEA) forecasting that energy usage for residential cooling may be triple that of residential heating by 2050 [2]. Hence, understanding the thermal performance of buildings is of great importance to identify where energy-saving interventions can be taken which

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/988831/Q1_2021_Smart_Meters_Statistics_Report.pdf

will lead to a significant reduction in GHG emissions.

When the RdSAP is used to update the thermal performance component of the EPC, a surveyor conducts an evaluation of features of the building such as the year it is built, the type of window glazing, etc. and uses these data-points to infer the thermal performance [11]. However, this has been shown to be inaccurate and it does not accurately represent the real-world thermal performance of the building [11, 131, 73]. The current best practise is to evaluate the thermal performance of a building with a co-heating test which consists of vacating the home for up to two weeks in the winter [115, 114, 10]. Such a method is intrusive, expensive, and not fit for many real-world scenarios [29]. Thus, there is a need to accurately evaluate the thermal performance of residential houses non-intrusively and cheaply so that the EPC better captures the real-world thermal performance and houses with poor thermal performance can be identified at scale. This need is emphasised by the fact that the Department for Business, Energy, and Industrial Strategy (BEIS) launched the smart meter enabled thermal efficiency ratings (SMETER) innovation competition[3] in 2018, with the purpose of developing better methods for evaluating the thermal performance of residential buildings.

Interventions can reduce the amount of energy residential buildings require. For the remaining energy usage it is important to shift this to sources with low GHG emissions. There is a growing trend of on-site residential energy generation, mainly driven by solar photovoltaic (PV) panel, which shows a significant shift from consumers to prosumers, those who consume and generate energy [15]. From 2010 to 2021, the capacity of small (under 10kW) PV panels installations grew by 3GW [15]. To allow solar PV energy to continue growing as a generation source, grid operators need to ensure that there is sufficient distribution and transmission infrastructure to avoid curtailing renewable generation [9]. The availability of smart-meter data means that techniques can be

[3] SBRI Competition:TRN 1611/09/2018

developed to accurately and non-intrusively identify the PV generation from residential homes for the EPC. This can help individuals and energy providers by providing them with a better understanding of PV generation, and there is potential for grid operators to also use this for a better understanding of regional energy generation.

Surprisingly, there are no UK installation database that monitor local and regional small-scale PV panel installations. Current techniques rely on databases of existing installations - however public databases only record installations above 1MW and over 20% of installations are below 4kW [85]. Whilst there does exist a database of residential PV panel installations from the feed-in tariffs scheme, it stopped being updated on 31 March 2019 [4]. Hence, tools are needed to predict the presence of PV panels and their generation in the UK. Predicting the presence and generation of a PV panel from data is difficult as smart meters do not record the net energy usage of a house and typically return 0 kWh when there is net negative energy consumption. Hence there is a need for improved techniques to infer residential PV generation that do not rely on existing databases.

Another emerging trend in residential building sector is the increasing amount of energy storage being available which often comes in the form of electric vehicles or a battery [42]. The presence of energy storage provides great benefits to the resident as it can enable them to use more of the energy they generate and they can take energy from the grid when it is cheaper or has a lower level of carbon intensity [121]. Furthermore, it can provide a source of grid independence in the case of energy instability [4]. However, the presence of residential energy storage can also interfere with methods which work with smart-meter data as the energy taken from the grid no longer matches direct energy usage as there is a battery being charged and discharged. Hence, it is important that the

[4] `https://www.ofgem.gov.uk/environmental-and-social-schemes/feed-tariffs-fit/changes-fit-scheme`

presence of residential energy storage is accounted for before using smart meter data.

When using smart-meter data and other data from residential buildings it is important that the data collected respects the privacy of the occupants, and is compliant with regulations. For example, in the EU and the UK the collection and processing of data must comply with the general data protection regulation (GDPR). In particular this means that customers must be able to opt-out of data collection, they must be informed of what data is collected and how it is being used, and their rights must be protected if any automated decisions are made [132]. Furthermore, best efforts must be made to encrypt, pseudonymise or anonymise data. Hence, any real world implementations of the methods developed in this thesis must ensure they adhere to GDPR requirements, or the relevant regulations for where the household is based.

When evaluating the energy performance of residential buildings it is important that contextual information is used alongside the energy performance metrics to ensure the right conclusions are reached. For example, a larger house will typically use more energy than a smaller house to achieve thermal space comfort, but that does not necessarily mean that the building material or heating system is less efficient, just that a larger space is being heated/ cooled. Hence, in this scenario the building size would provide useful context for evaluating the thermal performance. The EPC can potentially provide contextual information as the survey data from previous evaluations may have gathered the required information [23]. Alternatively, if there is no contextual data for a house the tools developed in this thesis may be used as a first step to identify if the house potentially has a poor energy performance and guide more targeted information gathering. Furthermore, to ensure the right conclusions are reached it is important that the methods developed are used in scenarios which match the scenarios in which they have been developed and tested. The methods developed in this thesis have used UK and US half-hourly smart-meter and weather data, and whilst the methods should be adaptable to many other locations there may be a need for some localisation of the techniques to account for different energy usage and weather profiles in different regions [106].

This thesis presents a number of data-driven tools to augment the EPC. The methods proposed to infer properties of residential buildings work with smart-meter and weather data making them non-intrusive and scalable due to the national deployment of smart meters. These tools have the potential to enable more accurate, and cheaper measurements of the thermal properties and PV generation of houses. By providing more accurate insights, residents will benefit by an improved understanding of the energy usage and generation for which their home is responsible for. Better understanding may lead to them taking steps which reduce their GHG emissions and/or reduce their energy bill. Further to this, energy providers may also benefit from these additional insights as it provides them with the opportunity to better quantify how they can help customers they are obliged to help due to ECO, and with a better understanding of customer profiles which may help them identify the best energy tariffs and auxiliary services for their customers.

1.2 Research Questions

The current EPC evaluation relies on intrusive evaluations which are costly and time-intensive, or assumptions which can be inaccurate [73]. It is important that the properties of the residential buildings are correctly labelled to identify opportunities to reduce residential GHG emissions. With the wide-spread adoption of smart meters it provides a source of data which can be used to quantify the real-world energy performance of residential buildings when performing EPC evaluations. This thesis explores if the EPC can be augmented with scalable, accurate, and low-cost techniques using smart-meter and weather data and aims to augment the EPC by addressing the following three questions:

1. Can we infer the maximum power output of PV panels on residential households whilst accounting for censored smart-meter readings?

The current EPC evaluation process for PV panels relies on an in-person inspection to identify the orientation, tilt, shading, and labelled maximum power output, or if this is not possible, assumptions are made about the PV generation. Furthermore, the recent closure of the feed-in tariff scheme means that no reporting scheme currently operates for smaller PV panels. For more accurate results the evaluation should be performed with real data which captures the performance of the PV panel. Currently there do not exist techniques that can identify the presence of PV panels and their generation of energy from smart-meter data in all UK homes. This is caused by the fact that smart meters and PV panels are not correctly calibrated in over 50 % of homes leading to local energy generation being recorded incorrectly and no existing techniques address this.

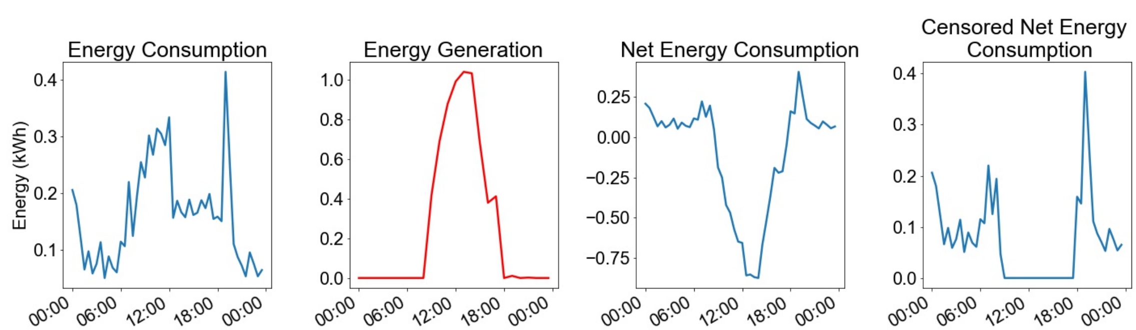

Figure 1.1: The leftmost plot shows the energy consumption for a random day from the Pecan Street dataset, whilst the second plot shows the corresponding energy generation for the day. The third plot displays the net meter reading, which most existing approaches to solar disaggregation are designed to use. The rightmost plot shows the corresponding readings of a censored smart meter - notice here the regions where net smart metering is censored, namely kept at zero when negative.

The incorrect calibration means that the smart meters are reporting the net energy taken from the grid, rather than the energy usage and the energy generated separately. This creates censored smart-meter readings meaning that if there is a net negative energy usage in a household the smart meter only records 0 kWh as it records the energy taken from the grid. Figure 1.1 provides a visualisation of how the true energy generation and usage are hidden in the smart-meter reading due to the censoring caused by measuring the energy taken from the grid rather than the two values separately. Hence, this thesis explores if an accurate, scalable approach can be developed to identify the maximum power output of PV panels from censored smart-meter data.

2. Can we accurately identify the thermal performance of a residential home non-intrusively?

The current EPC method to evaluate the thermal performance of a residential home is for a domestic energy assessor to conduct a survey of the physical features of a home and label its performance based on the predicted properties of each component. This method has been shown to be expensive and inaccurate [73]. Current alternatives are intrusive tests that are accurate but require the home to be unoccupied for over a week which is expensive and unfeasible [10], or non-intrusive approaches that are yet to show the accuracy required for commercial use [29]. Hence, this thesis explores if an approach can be developed that can measure the thermal performance accurately and non-intrusively at scale to provide a cheaper and improved alternative to use for evaluating thermal performance for the EPC.

3. Can we identify the scenarios in which the thermal performance can be accurately inferred and quantify the uncertainty?

To operationalise the large-scale, non-intrusive inference of the thermal performance of residential buildings, the scenarios in which the thermal performance can be accurately inferred need to be quantified [68]. The real world presents a number of scenarios with different sources of data available (i.e. an hourly or daily smart-meter, a smart thermostat, etc.), and different environmental factors and usage profiles. The differing scenarios create limitations on how accurately the thermal performance can be inferred and it is important for operators to understand when the results are reliable, and when more information is required to make an accurate inference. Furthermore, when inferring the performance of cooling and heating there are numerous sources of uncertainty that needs to be accounted for, which can potentially be achieved by developing a Bayesian framework to quantify the uncertainty in the thermal models.

When developing methods that may be used in the real-world as part of the EPC it is important that they work in as many scenarios as possible, that the scenarios in which they fail are identified, and uncertainty is quantified. To extend the method to work in more scenarios, it is explored if the thermal model can be adapted such that the cooling performance can be evaluated as well as the heating performance.

1.3 Publications

Each chapter in this thesis corresponds to a paper that has been published in a peer-reviewed journal, or presented at a top-tier international conference.

The work presented in Chapter 3 has also appeared as: J. Brown, A. Abate, and A. Rogers, "Disaggregation of household solar energy generation using censored smart meter data" Energy and Buildings 231 (2021): 110617.

The work presented in Chapter 4 has also appeared as: J. Brown, J. Chambers, A. Abate, and A. Rogers, "SMITE: Using Smart Meters to Infer the Thermal Efficiency of Residential Homes." Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. 2020. It was **awarded runner-up best paper**.

The work presented in Chapter 5 has also appeared as: J. Brown, A. Abate, and A.

Rogers, “QUILT: QUantify, Infer and Label the Thermal Efficiency of Heating and Cooling Residential Homes.” Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. 2021.

1.4 Thesis Overview

This thesis formulates a number of data-driven techniques to augment the EPC. The methods developed are non-intrusive and scalable to ensure they can be deployed nationally. The focus areas are identifying which buildings are thermally inefficient, identifying the scenarios in which this inference can be accurately made, and identifying which buildings have unreported PV generation or are not operating at their reported capacity. All three problems can be solved with levels of accuracy previously unavailable due to the data from smart-meters.

The thesis is structured as follows:

- **Chapter 2** presents the background understanding required for the thesis. It covers the broader environmental and energy trends which evidences the timeliness of this thesis. Then, an overview of the available sensors in residential buildings is provided before a deep-dive into the literature surrounding the current methods for identifying thermal performance in residential homes and PV generation.

- **Chapter 3** addresses the issue of smart-meters failing to accurately record PV generation. An algorithm is presented to deal with censored smart-meter readings, which occurs when energy generation is not separately recorded. The PV generation of residential buildings can then be inferred and this is validated on a dataset curated from real-world data.

- **Chapter 4** introduces a data-driven approach to infer the thermal performance of a residential building using only smart-meter readings and weather data. A thermal model for residential buildings is formulated, and methods to detect when the building is in a state of thermal equilibrium are developed, including identifying when the heating is on. Then inference can be performed on the model and it is is validated on real-world data.

- **Chapter 5** builds on Chapter 4 to extend the thermal model for residential buildings. The model is extended to account for factors that had previously been implicit assumptions. Furthermore, a Bayesian framework is then set up to infer the thermal performance whilst accounting for the uncertainty from different parameters in the model. A new model is also introduced to identify the thermal performance of a building when a cooling system is used, enabling the work to be used in more scenarios.

  Furthermore, a simulated dataset is presented to evaluate what data is required to accurately infer the thermal performance of residential buildings. The simulated dataset presents scenarios beyond what is available in the currently limited real-world datasets and allows for actionable insights to be made.

- **Chapter 6** concludes the thesis, highlighting the contributions to the field, the actionable insights, and future research directions.

2

Background

To frame the context of the rest of the thesis this chapter starts by providing an overview of the residential energy sector and the broader energy trends in Section 2.1. Then a description of the most common devices for residential data collection is provided in Section 2.2 as well as the privacy considerations. This is important to understand what data can be used when inferring the properties of a residential building. Following this, Section 2.3 provides an overview of methods used to infer and forecast the PV energy generation from residential buildings. Then in Section 2.4 the methods to evaluate the thermal performance of a building are discussed. Finally, Section 2.5 introduces the technical preliminaries of the methods used throughout the thesis.

2.1 Environmental and Domestic Energy Trends

To limit global warming to 1.5 °C, the 2018 IPCC report states that net human-caused emissions of carbon dioxide need to fall 45 % from the 2010 levels by 2030, and be net zero by 2050 [65]. In the UK, there has been a significant reduction in total GHG emissions, which have fallen from 610.5 million tonnes of carbon dioxide equivalent in 2010 to 414.1 million tonnes of carbon dioxide equivalent in 2020, a 32 % decrease [16]. This is predominantly driven by the reduction in coal consumption, as the UK has cut coal consumption from 50.8 million tonnes in 2010 to 8.2 million tonnes in 2020, an 84 % decrease[1]. Whilst this is an important first step, to continue decreasing greenhouse gas emissions additional measures need to be taken, which include shifting the remaining energy production to renewable energy and reducing the total energy consumption [40].

Low carbon sources of energy currently make up 18.4% of the UK energy supply in 2017, where nuclear contributes 7.9%, biomass 6.4%, wind 2.2%, solar 0.5% and hydro 0.3% [16]. Of these, PV generation is currently the most feasible renewable source of energy that can be deployed directly on residential buildings. This provides the opportunity for consumers to take partial control of their energy generation and become prosumers, someone who produces and consumes energy. PV panels have seen a huge increase in installations over recent years encouraged by government incentives and the falling price [25]. However, as PV panels generate energy when sun rays are incident on the panel, not when the energy is required, steps need to be taken to deal with the discrepancy between the times of generation and consumption. This will become prevalent if solar energy becomes a large proportion of national energy generation. To address this, accurate forecasts of solar energy generation are required and energy storage will be needed to stabilize the difference in time between when

[1] `https://www.statista.com/statistics/223493/uk-coal-consumption/`

energy is generated and when it is consumed [129].

The Directive of the European Parliament on energy efficiency has set a target of at least a 32.5 % improvement in energy efficiency by 2030 [46]. One of the largest sectors of energy usage is the residential sector, which accounted for 29 % of the final UK energy consumption in 2020 [14]. Further to this, thermal space comfort accounts for up to 76 % of residential sector energy usage, and unfortunately this demand is predominantly satisfied by gas [14]. Across the EU the picture is similar, where heating and cooling accounts for more than half of the energy consumption, and approximately 75 % of this is still generated by fossil fuel [47]. Hence, reducing the amount of energy required for thermal space comfort will play a significant role in preventing climate change [66, 67, 64].

2.2 Residential Data Collection

Across the UK smart meters are being deployed which offer opt-in reporting of half-hourly energy usage. It is believed that offering energy reporting at a finer granularity will allow consumers and energy providers to gain better insights into the energy usage of households and this will allow for energy reducing interventions to be implemented. Almost half (45 %) of UK houses now have a smart meter installed and the goal is to have a smart meter offered to every house by 2025 [12]. Hence this is a form of residential data collection that will be deployed across the majority of buildings in the country. It is important to note that there are currently two versions of smart meters, and how they work are specified by the Smart Metering Equipment Technical Specifications (SMETS). The major differences are that SMETS1 contacts the energy provider via a SIM card that is locked to the provider causing barriers when switching whilst SMETS2 uses a government data network to bypass the issue. However, both SMETS1 and

SMETS2 provide the energy data in the same way - with half-hour readings if the user has opted in. Whilst incorrect installations can cause data collection issues that need to be accounted for in some scenarios, smart meters will facilitate the collection of energy usage data nationally, and in some cases energy generation data.

Residential energy generation is most commonly achieved by PV panels, which convert solar radiation into electrical energy when solar irradiance is incident on the panel. The energy generated by the PV panel is direct current and an inverter is used to produce alternate current electricity at a voltage and frequency that match the specification of the region. There are approximately 1 million homes in the UK with PV panels[2], with cheaper, more efficient panels being developed that can be deployed in more scenarios [88, 53]. Beyond generating energy, PV panels also act as another data source as the generated energy is recorded by the smart meter, or used behind the meter which has an impact on the total energy consumption which can provide insight into local solar irradiance.

To develop scalable methods to gain insights into residential energy usage and generation the hardware used to collect data needs to be widely deployed, as is the case with smart meters. However, there is hardware that can provide more insight into residential usage which is not widely deployed. For example, the current smart meters being deployed in the UK record half-hourly energy usage and generation, but private companies have developed their own energy meters which monitor energy usage and generation at a much finer granularity, typically minutes or seconds [91]. Furthermore, the energy usage of individual appliances can be tracked with plug monitors, providing insight into how much energy each appliance uses [81]. Also, smart thermostats can be used to monitor the internal temperature and identify when the heating is on. The user can also control the heating with greater ease using such an

[2] `https://theswitch.co.uk/energy/guides/renewables/solar-energy`

appliance. These sources of data collection provide high fidelity data which makes it easier to infer properties of a house, however, due to the fact they are not widely deployed they are not a reliable source of data.

Data collection and usage must be in line with the regulatory requirements for the region, which means in the EU and the UK, GDPR must be adhered to. This means that residents must be informed of what the collected data is used for, to be able to opt-out of data collection, and their rights must be protected for any automated decisions. It is particularly important that the usage of the data is made clear to the users as insights about a house beyond its energy efficiency can be gained by monitoring energy usage [130]. This can be seen in recent research which shows that energy usage data can be used to identify if a house is unoccupied [30]. Also, with the presence of PV generation, the pattern of energy generation can be used to locate a house [33]. Hence, methods of non-intrusive load monitoring have been developed which use techniques such as differential privacy, or federated learning to try and protect this information [133, 134]. Furthermore, it has been explored how a residential battery can be used to store and release energy to cover up to add noise to residential energy usage to reduce the risk of privacy intrusions [86].

2.3 Solar Energy Generation Inference

PV panels are the largest source of residential energy generation, and their install base is rapidly growing, meaning they are making up a growing proportion of the grid's electricity supply [16]. PV energy generation can be intermittent since it is dependent on solar irradiance which is weather-dependent, thus, the grid operators who must balance the supply and demand of electricity in real-time to keep the grid at a stable frequency have an increasing challenge to estimate the quantity of energy provided by these PV installations [100, 101, 111, 45]. Ideally, the location of these panels and their capacity would be reported to the grid operator, and their generation would be measured and reported in real-time through smart meters. Unfortunately, in the UK, no reporting scheme currently operates, and the deployed base of smart meters (and those scheduled to be deployed to all UK homes in the future) do not always separate local generation and consumption. Rather, they simply record the net energy taken from the grid.

To establish the energy generated by a PV panel for the EPC, the peak power, orientation, tilt and shading are all identified. These values are either found by inspection, or by assumptions [23]. For example, the peak power is estimated based on the size of the PV panel, and if the orientation and tilt is not known it is assumed it is a south-facing panel at a 30°tilt with modest shading, and a lookup table for the average solar irradiance for each month in that region is used to infer PV generation [23]. However, it is often the case that the true peak generation of a PV installation does not match what is predicted due to faults or incorrect assumptions and hence it can lead to inaccurate labelling [74].

Research to better understand residential PV generation can be beneficial for energy consumers, energy companies, and grid operators as well as providing a better process for quantifying PV generation for the EPC. For consumers, it can provide better information of their actual generation and what they can expect to generate in the near future which allows for better planning of energy usage energy to minimize carbon emissions [135]. It can also be used to detect faults in the panels or inverters if energy is not being generated at the expected capacity [101]. These insights are also beneficial to energy companies, especially if they are considering selling energy as a service as they can better quantify the impact of installing a PV panel. For grid operators, they face the issue of having to match demand with an increasing proportion of energy generation coming from unreliable sources, better forecasting will allow for better planning of energy generation [44].

There are a number of areas of research to improve the understanding of residential PV generation. There is a need to infer what the maximum power output of the installed PV panels to understand how much energy can be generated by the house [32]. If the peak power of the PV panel is known, future PV generation can be predicted using weather forecasts to predict future solar irradiance [112]. Furthermore, the PV capacity and generation can be learned at a number of different levels: household, regional, or national depending on the granularity required for the task [71].

2.3.1 Forecasting Solar Irradiance

To infer the peak power of the PV system attached to a house, a signal is required, this is typically the solar irradiance incident on the panel [kW/m^2] or the energy generation. The most accurate approach is to measure the solar irradiance is using a pyranometer, however, this device is expensive, which restricts the scale at which they can be deployed [5]. A pyranometer measures the solar irradiance with a wavelength of 0.3μm to 3μm incident on the device. The measurements provided by a pyranometer can be used to predict solar energy generation accurately by using known calculations which relate solar irradiance to the expected PV generation [5]. The quality of the pyranometer data varies based on the equipment used with resolution varying from 0.001 kW/m^2 to 0.01 kW/m^2 and uncertainty in measurements can vary from 3% to 10%[3].

It is often the case that direct measurements can not be made or that future values of solar irradiance are needed to predict future solar generation. Typically to find these values the clearsky solar irradiance is required. This is the solar irradiance incident on the ground if there is no weather or atmospheric interference which only depends

[3]`https://www.campbellsci.com/blog/pyranometers-need-to-know`

on the location on Earth relative to the sun which can be calculated using existing techniques, such as the Bird model, as seen in the software package PVLIB [20, 116]. Then, the weather effects on reducing the clearsky solar irradiance (e.g. cloud coverage, precipitation) need to be calculated to find the actual solar irradiance incident on the PV panel. It has been shown that this factor is independent of the value of the solar irradiance, meaning they can be calculated separately [32].

To predict PV generation physical models of solar irradiance can be used. These methods usually focus on using cloud coverage imagery or measurements to predict the reduction of solar irradiance reaching the ground [99]. Physical models can fail once extended beyond laboratory settings due to the model not correctly representing the true complexity of the actual task being tackled. Recent success has been achieved combining physical and data-driven approaches, where the observed relationships between the actual solar irradiance reaching the ground and weather factors are used to train the models [112]. This relinquishes the requirement of building physical models which accurately reflect the real-world scenario.

Other approaches to predict PV generation have used machine learning to predict solar generation from weather forecasts [112]. Solar intensity is modeled as a function of the day, temperature, dew point, wind speed, sky cover, precipitation, and humidity. Linear least squares regression is used to provide interpretable coefficients for each variable to test the significance of each and a support vector machine with a radial basis function kernel is used to predict the solar intensity given the weather factors [112].

Approaches using neural networks with weather factors as inputs to forecast solar irradiance have achieved better accuracy, at the cost of explainability [105, 89]. It is hoped that data-driven models can learn the relationships that physical models can not, e.g. how the direction of the wind can contain information of future weather conditions and temperature change. Other approaches show the strong relationship that solar irradiance has with the previous 30 minutes, which can be used to improve short-term forecasting for the next 30 minutes [107]. Whilst neural network approaches can provide better predictions of solar generation, they do not provide an interpretable model as the function does not necessarily have real-world meaning. This means that they do not guide the user towards which weather factors are useful inputs for the predictive models.

2.3.1.1 Mapping Solar Irradiance to PV Generation

There are multiple factors to consider when mapping from solar irradiance to generation. Fortunately, the majority can be represented with simple physical models [32]. As the solar irradiance is measured in kilowatts per metre squared (kW/m^2), the size of the solar panel is simply a scaling factor for the amount of generation. Solar panels are typically sold with a reported peak power - corresponding to the power output at 25°C with 1 kW/m^2 of solar irradiance incident on the panel. The percentage of the solar irradiance converted to the power output is typically in the range of 15 %-25 % for commercial residential panels [32]. However, the reported peak power can be incorrect due to factors in the real-world that limit the performance (e.g. panel orientation and shading), or due to a fault. Data-driven methods may be needed to correct such errors in reporting to find the true peak power of PV panels.

The effect of the orientation and tilt of the solar panel on the generation of solar energy can be modeled using trigonometry. That is, the solar irradiance incident on the PV panel, I_{inc}, and the power output by the PV panel, P_s, is given by:

$$I_{inc} = \max\left(I\left(\sin(\theta_z)\sin(\beta)\cos(\gamma - \gamma_S) + \cos(\theta_z)\cos(\beta)\right), 0\right),$$

$$P_s = I_{inc}k.$$

Here I is the actual solar irradiance incident on the plane of land tangent to Earth (after reductions from weather and atmospheric factors) in (kW/m^2) and k is used to represent the other scaling factors such as size and efficiency. The zenith angle above the sun is given by θ_z, and γ_S is the azimuth angle of the sun. The tilt of the solar panel is given by β and the orientation (azimuth) of the panel is γ. All negative values are equal to 0 as the panel does not produce negative energy if the sun is behind it.

External factors can also interfere with the efficiency of a solar panel. A relationship between the efficiency of solar panels and their temperature has been observed and the following equation provides an estimate of the temperature of a solar panel [109]:

$$T_{\text{Panel}} = T_{\text{ex}} + I_{inc}\frac{\text{NOCT} - 20}{800},$$

where T_{Panel} is the temperature of the panel and T_{ex} is the temperature of the air, in degree Celsius. The nominal operating cell temperature, (NOCT), is a value in degree Celsius specific to a solar panel model. The solar irradiance, I_{inc}, is as above. The relation varies between panels but typically a solar panel drops in efficiency by approximately 0.5 % per 1°C increase [109].

Whilst there exist physical models for the factors discussed in this section, which can be modeled with relatively accessible data, there are some factors that are difficult to account for. For instance, soiling and shading can have a significant impact on the efficiency of solar panels, but they are hard to model and require localised, difficult to obtain data [87, 54].

2.3.2 Solar Generation Prediction

An issue of using solar irradiance as an input to infer solar generation is that it requires solar irradiance and weather data, and whilst this is recorded by weather stations, the distance between the weather station and the PV panel adds a source of uncertainty. An alternative method is to measure solar generation directly using a large number of PV panels across the country and use this as the signal to forecast future PV generation or infer current generation from other panels [71]. One such approach uses the solar generation directly in a Bayesian ensemble to infer PV power given historic generation data and weather forecast data [27]. This approach focuses on predicting PV generation at an hour granularity for the next 24 hours. The approach uses an ensemble of naive Bayes, k-nearest neighbours, and motif predictors based on forecast weather conditions and known PV generation. Each component is weighted and a Bayesian model is used to average out the components. The approach demonstrates that solar irradiance does not need to be explicitly known to forecast future PV generation.

By using the PV generation as the training signal for the model rather than the solar irradiance there is the benefit that the model can learn some of the intricacies of how the panels' efficiency depends on weather and external factors which physical models do not currently express [26]. There are benefits to machine learning approaches, and physical models, hence hybrid approaches which set up physical models and learn the parameters of the physical models are viable to achieve the best results for the task, whilst being explainable [31]. Using physical models avoids over-complicating straightforward relationships and mappings, whilst the data-driven approaches are used when physical models can no longer correctly represent the problem.

2.3.3 Disaggregating Residential PV Energy Generation

If PV panels and smart meters are installed correctly, the smart meter can report the energy generated by the PV panel for each 30-minute period. However, they are often incorrectly installed and the actual PV generation is not reported, meaning that it needs to be disaggregated from other energy generation and/or usage in the house. For example, in the UK a smart meter records the electrical energy taken from the grid. This means that smart-meter readings can be thought of as the energy consumed by the house minus the energy generated by the house. However, if there is surplus energy generation the smart meter will only record 0 kWh as it does not record the energy fed back to the grid, only the energy that is taken. Techniques can be used to disaggregate the solar energy generation from the consumption to help provide better predictions of the current generation and to predict the actual maximum generation of the solar panel.

To disaggregate solar energy generation from the energy-meter readings, it can be achieved using hardware with a finer granularity than what is available with a smart meter [91]. With this data a variety of support vector machines with different kernel functions are implemented to disaggregate the solar energy generation from the energy meter readings [59]. Alternative methods proposed for disaggregating energy generation use physical models to predict solar irradiance and then to predict how the solar irradiance converts to solar energy [32]. The predictions of when there is clear sky and the base load of the smart meter is then used to disaggregate the generation. However, the above approaches do not deal with the situation where smart meter readings are censored and it requires additional monitoring hardware.

Existing work to understand national PV generation has been conducted by the Sheffield Solar research group [123, 124, 125, 126]. They have provided a new service for individuals to donate their solar generation data to create the microgen database of 7,000 PV systems in the UK. The microgen database is combined with the known capacity of PV panels in each region to predict the regional PV generation. This is predicted by taking the readings from pyranometers in each region and a weighted average based on the PV systems distance to the nearest pyranometers.

2.3.4 Predicting Solar Panel Presence and Size via Satellite Imagery

Satellite imagery has been utilized in approaches to detect solar panels and construct a database of the deployment of solar panels [136, 137, 82, 83]. The developments in deep learning have made it possible for image recognition techniques, using convolutional neural networks, to be deployed on large-scale tasks [34, 77]. Using satellite imagery allows for a scalable and non-intrusive approach to detecting the location of solar panels as long as there is access to data for the region. The detection of PV panels via satellite imagery can be achieved by adapting existing techniques to identify the presence of an object from an image [108].

A deep learning framework, DeepSolar, has been developed in the USA, which finds the GPS location of PV panels and the size of the installation from 30cm resolution satellite imagery data [136]. The framework has been used to create a database of PV panel locations around the USA, locating 1.47 million whereas alternative approaches by Google Sunroof[4] found 670,000 and the Open PV[5] voluntary survey located 1.02 million. On test data, precision of 93.1% and recall of 88.5% was achieved in residential areas [136]. The framework has shown the viability of satellite approaches and created a more comprehensive database of solar panel locations than any alternative approach currently available. The work extends to identify correlations between solar deployment and environmental/ socio-economic factors, culminating in SolarForest, a predictive model which estimates the solar deployment density based on census data [136].

[4] `https://www.google.com/get/sunroof`
[5] `https://openpv.nrel.gov/`

2.4 Thermal Performance of Homes

The residential sector makes up 29 % of all energy end-use cases, and within this sector thermal space comfort use the majority of energy, reportedly up to 76 % of it [14, 66, 67, 64]. Hence, reducing the amount of energy required to heat and cool homes has a significant role in reducing GHG emissions and in mitigating climate change. Whilst reducing emissions, the UK government and energy companies also have a duty to protect vulnerable consumers who are in fuel poverty; this is defined as those who cannot keep their home warm at a reasonable cost, as per the UK Warm Homes and Energy Conservation Act. To understand where energy-saving improvements to residential buildings can be made, it is important to quantify its thermal performance.

Currently, the thermal performance of a new house is evaluated using the SAP to produce an EPC, where a domestic energy assessor surveys the physical features of a home [11]. A simpler version, the reduced SAP (RdSAP) is used to evaluate existing homes and make inferences with a reduced amount of data from a survey. The evaluations for the EPC are used to rate each house and assigns an efficiency class and environmental impact rating. The data may also be used to advise on energy-saving retrofits. A major component of the EPC is to evaluate the thermal performance of the house and quantify the impact of retrofits. However, the current techniques performed in the SAP and RdSAP have been shown to be inadequate by independent investigations as the results do not reflect the thermal performance in a real-world setting whilst costing up to £120 per property [131, 73].

Metrics can be used to quantify how much energy is required to heat a building. For example, heating degree days (HDD) give how much the average external temperature is below a set value (18.5°C in the UK) and 0 if it is above [128]. This value is closely related to the amount of energy required to maintain thermal space comfort for a building.

Similarly, heating degree hours (HDH) measure how long the outside temperature is below a specific temperature [79]. Equivalent metrics can be found for cooling degree days and hours with similar calculations for when the external temperature is above a set temperature. These metrics can be used to identify general trends of expected energy usage for heating and cooling purposes. However, these measurements do not account for the specifics of an individual house and are based on averages.

The evaluation of the thermal performance of buildings has also been explored using different metrics, such as U-values, which measure the thermal transmittance through a surface. The U-value of a material can be measured directly with heat flux sensors where the sensors are set up in an unoccupied house for a number of weeks to acquire the U-value of the wall materials [78]. A quicker approach, the QUB method, has been proposed to calculate the U-values of the building materials in-situ [90]. However, it has been shown that the uncertainty in U-value measurements, and uncertainty in how they translate to the actual thermal performance of a home make them an unreliable metric [78]. Furthermore, it does not give an overall picture as it only measures the thermal properties of the walls in a static environment and it does not account for many other factors which may affect the real-world thermal performance of the building [18]. Unfortunately, these errors have real-world consequences, as a recent government report showed that cavity wall insulation was 38 % less effective than expected from lab tests, reiterating the importance of developing accurate techniques which reflect real-world performance [43, 78].

A similar approach of directly measuring the thermal performance of the building materials has been conducted in the PASLINK project [7]. The project was set up to improve the understanding of thermal and solar characteristics of houses under real conditions using test facilities. The performance of the building components are evaluated over a range of European climates using dynamic testing procedures to identify the parameters of a physical model representing the heat flow rate through the building component. The dynamic tests allow for the thermal performance of the materials to be evaluated in different scenarios. It was concluded that continuous-discrete time stochastic state-space models are more appropriate than the linear models that are typically used and that given a wall with a window, a nonlinear model is required for better performance [70]. However, it faces similar drawbacks to U-value based methods as the individual performance of the components do not necessarily give the overall performance of the building. Furthermore, these approaches require intrusive measurements and an insight into the structure of the property.

To provide a better understanding of the heat dynamics, models of thermal performance have been formulated instead of reducing it to a single metric [3]. By setting up a physical model which represents the situation, data-driven approaches are used to learn the parameters of the model. Using stochastic differential equations (SDEs), the rate of flow of heat energy through the rooms and the flow in and out of the building can be inferred [3, 6]. The physical models created provide real-world meaning to the parameters being learned, however it requires knowledge of the building infrastructure and internal temperature monitoring in each room.

For the metric to be accurate in practise, it is important that it reflects the thermal performance of the building in the real-world setting. Hence, data-driven approaches that measure the power required to keep the internal temperature fixed given the external temperature can provide real insight. To measure this directly, co-heating tests are conducted which require the home to be vacated for up to a week, and the internal temperature is maintained at a constant level by electric heaters whilst the external temperature and other factors are monitored [115, 114, 10]. The power required to keep the internal temperature constant given the external temperature is called the heating loss coefficient (HLC). The co-heating test is currently the best practise in terms of performance, and the uncertainty of the HLC is 10% in optimal settings [69].

Adaptations have been made to the co-heating test to allow the native heating system to be used, instead of installing temporary electric heaters [48], and alternative approaches to infer HLC have been developed, such as the P-STAR and the ISABELE methods [118, 127]. The thermal performance can also be inferred with specific kits of wireless sensors that are installed to estimate the thermal performance of a home in-situ [52]. Furthermore, internal temperature monitors can also be used in data-driven approaches to collect 10-minute data through nights and then an auto-regressive model is used to learn the thermal characteristics of a building, however, such internal temperature monitors are not currently widely deployed [104]. With 25 million homes in the UK alone, intrusive methods to evaluate the thermal properties of a home will take significant person-hours and incur notable financial costs [120]. Whilst these approaches can show improved performance or reduced testing time, they are still intrusive. As such, to achieve a large-scale roll-out of HLC evaluation, there is a need for non-intrusive methods to evaluate the thermal performance of homes nationwide, and globally.

2.4.1 Thermal Model

The HLC [kW/K] of a home is a value used to characterise the steady-state thermal performance of the building envelope. It is the power required to keep the home at a constant internal temperature, given a lower external temperature. The HLC is inferred through the data obtained from a home and measures how the difference between internal and external temperature, $\Delta T = T_{ex} - T_{in}$ [K], causes the average heat flow into the house i.e. the average heating power used, Q_{fa} [kW], to change

[115, 114], given by the equation

$$Q_{fa} = HLC\,\Delta T. \tag{2.1}$$

This equation concerns a sealed building where the only factors affecting the internal temperature are the external temperature and internal heat sources and the only transfer of heat is through the fabric of the building. In practise, ventilation losses, Q_{ve}, must also be considered, and additional sources of heat energy, such as external weather factors, Q_{we}.

The above terms alone would suffice if the home is unoccupied, however an occupied home has two additional factors that need to be considered: (i) heat from the occupants Q_o, and (ii) power supplied for tasks other than direct heating which may have a heating effect, Q_b (e.g. cooking, running electrical devices, etc.) [29]. Hence, the total flow of heat energy Q_h into the home is given by

$$Q_h = Q_{fa} + Q_{ve} + Q_{we} + Q_o + Q_b. \tag{2.2}$$

In this heat flow equation, the term representing the heat flow through the fabric of the building Q_{fa} follows from the initial definition of the co-heating test in Equation (2.1). The next term considered is the ventilation heat flow,

$$Q_{ve} = c_{air}\rho_{air}V_{air}(T_{in} - T_{ex}), \tag{2.3}$$

which denotes the transfer of heat energy via the flow of air moving freely between the external and internal environment. In co-heating tests, this is accounted for with experiments that track the air movement (e.g., by a tracer gas test or a blower-door test) [10]. This heat flow can be modeled as in Equation (2.3), with c_{air} as the specific heat capacity of air, ρ_{air} as the density of air, and V_{air} as the volume of air that flows between the external and internal environment.

The other factor which affects the internal temperature regardless of if the home is occupied is the external weather factor, Q_{we}. For simplicity the only weather contribution considered is the heat flow from solar irradiance,

$$Q_{we} = A_{sol} I_{sol}, \tag{2.4}$$

defined as the amount of solar irradiance, I_{sol} $[kW/m^2]$, incident on the effective aperture of the house A_{sol} $[m^2]$ [62]. Prior studies have shown that other high-level weather factors (i.e. regional wind speed and direction) do not correlate with energy usage, conversely, models with hyper-local data for factors such as wind have demonstrated an impact on internal temperature [28, 62]. With more precise weather data, both regional and with a finer time granularity, more complex ways in which the weather interacts with the internal temperature of a home can be modeled, however, this is currently not possible given the available data for non-intrusive approaches.

There are also the thermal contributions that are caused by occupants. Each person in the home emits heat energy,

$$Q_o = 0.06 N_{occ}, \tag{2.5}$$

where each of the N_{occ} occupants contributes 0.06 kW of heat energy [22]. When the home is occupied, appliances will be used for purposes other than heating, and a proportion, $0 \leq \eta_b \leq 1$, of the power used for other appliances, P_b, will convert to heat energy,

$$Q_b = \eta_b P_b. \tag{2.6}$$

To complete the heat flow model it is assumed that the heating system provides a heat flow in to balance the heat flows out and maintain the internal temperature,

$$Q_h = \eta_h P_h, \tag{2.7}$$

where η_h is the efficiency of the heating system, and P_h is the power supplied to the heating system. Hence the heat flow equation can be expressed as:

$$\begin{aligned} P_h = \frac{1}{\eta_h}(&HLC\,\Delta T + c_{air}\rho_{air}V_{air}(T_{in} - T_{ex}) + A_{sol}I_{sol} \\ &+ 0.06N_{occ} + \eta_b P_b). \end{aligned} \tag{2.8}$$

However, this equation assumes that the heating is always on, which is not necessarily true. Hence, it is proposed in the literature that the daily average internal temperature can be inferred based on the external temperature [80]:

$$T_{\text{in}} = F_{\text{T}}(T_{\text{ex}} - T_{\text{fix}}) + T_0, \tag{2.9}$$

where F_{T} is a dimensionless constant, set to 0.17 in the Deconstruct approach based on empirical data, and T_0 is the daily average internal temperature corresponding to a daily average external air temperature and $T_{\text{fix}} = 5°\text{C}$ [29]. However, this does not account for how long the heating is on.

This model is proposed as one which can be used with the low-fidelity data provided by smart meters and weather stations as it is realistic to expect that this data can be acquired non-intrusively. However, there are many intricacies that the thermal model will miss which higher fidelity models will capture. For example, this model assumes that a building consists of a single zone, whereas in reality buildings consist of numerous zones and the heat flow between different zones can be modelled [58, 51, 41]. Also, the model also assumes a single type of heating source, but in reality there may be numerous heating sources such as floor heating, central heating systems or electric heaters which behave differently [93]. Furthermore, more complex weather scenarios could be modelled to identify the effect of weather factors beyond solar irradiance on the internal temperature of a house, however, this typically requires high quality, localised, and granular data [62].

2.4.2 Non-Intrusive Methods to Evaluate Thermal Performance

To infer thermal performance across the country, it is important that accurate methods are developed which can be applied to the majority of homes with minimal cost and inconvenience. Non-intrusive methods which use readily available data have the potential to be deployed across the country with ease if they are accurate. This need is shown by the BEIS competition to evaluate the thermal performance of homes using smart-meter readings [37].

The original non-intrusive approach to infer thermal performance for homes was the PRISM model [50]. The method was developed to address the problem that existing predictions of energy savings using physical models were inaccurate and failing in the real-world. A model was built combining external weather factors, such as the daily temperature, with monthly billing data to infer the normalised annual consumption (NAC). This inferred value is then used to predict the energy consumption for a house given the external weather conditions. The PRISM model provided the foundation

Metric	Overview
HDD	The heating degree days measure how much the average external temperature is below a set temperature
HDH	The heat degree hours measure the number of hours the external temperature is below a set temperature
HLC/ HTC	The heating loss coefficient, also referred to as heating transfer coefficient, measures how the heating power for the house changes with external temperature
HPLC	The heating power loss coefficient also measures how the power supplied to the heating system changes with external temperature
NAC	The normalised annual cost predicts the annual cost of heating from monthly energy usage, external temperature and energy bills
PTG	The power temperature gradient calculates the rate that the power used by the house changes with external temperature
RdSAP/SAP	The standardised assessment procedure evaluates the thermal performance of a home by a survey of physical features is taken and the predicted properties are used
U-Value	The thermal transmittance through a single surface

Table 2.1: A summary of the metrics that are referred to in this section. The PTG, HLC, and HPLC are subtly different, the PTG does not isolate the contribution of heating so all the power is measured like HLC and HPLC. The HPLC also accounts for the heating system efficiency whilst the HLC considers the rate of change between external temperature and power output from the heating system.

to build non-intrusive methods to infer thermal performance.

With smart meters providing at least daily reports of energy usage, this information has been used to develop non-intrusive methods to infer thermal performance. The power temperature gradient (PTG) [kW/K], is the average daily power usage given the external temperature, and this metric is used to categorise the thermal performance of a house [119]. By using the average daily power for the home relative to the external temperature a general relationship can be learned of how the two factors relate, however, there are many other factors, such as how long the heating is on for and solar irradiance, which are unaccounted for and can lead to inaccurate results.

To address some of these shortcomings, the Deconstruct method creates a thermal model and uses it to identify days where the thermal contribution from factors other than the heating system and external temperature are small [29]. The thermal performance is measured using the heating power loss coefficient (HPLC), which is similar to the PTG except in theory it only measures the power used to heat the house rather than all of the power supplied. The HPLC is closely tied to the HLC as it is the HLC scaled by the efficiency of the heating system. By identifying days with small thermal energy contributions from other sources, the thermal performance of the building can be more accurately inferred. However, by not separating the duration that the heating is on from the thermal performance, the inferred HPLC will vary with the duration that the heating is on. In an attempt to address this, an assumption of a linear relationship between the external and internal temperature is made to infer the average internal temperature for the day [80]. However, as the relationship between internal and external temperature is not known for each home, a national average has to be taken, meaning that each house is treated as if it is the average home and useful information is thus lost. Hence, to accurately infer the HPLC, it is important to identify when the heating is on and remove this assumption. Using similar inputs, a smooth maximum function can identify the transition from weather-dependent to weather-independent heat consumption, and infer the HLC [103].

Unfortunately, theoretical measurements of thermal performance can fail to reflect real-world performance and can fail to quantify when they fail [76]. To address this, a Bayesian framework can be used to capture the uncertainty from the model and the data sources of the inferred thermal performance, as the WattScale approach demonstrates [68]. Bayesian models have also been used for heating disaggregation, namely to identify how much energy is used for heating [38]. However, whilst this captures model uncertainty, there may be scenarios where the model fails to infer thermal performance accurately, which need to be identified.

A number of metrics to evaluate the thermal performance of buildings have been introduced in this section and they are summarised in Table 2.1. It is important that the metrics are understandable, easy to compare, and can be measured at scale, such as HLC or HPLC. By selecting metrics that accurately represent the thermal performance of the building and are measured non-intrusively, it allows for scalable methods to be developed which can guide national decisions to reduce energy usage and for the thermal performance to be correctly labelled in the EPC.

2.5 Technical Preliminaries

To quantify energy properties in residential homes which are interpretable, models need to be constructed which represent the physical systems. In scenarios where the parameters of the models are not known, data-driven techniques are required which can accurately infer the parameters.

This section introduces the likelihood function that can be optimised to find the most likely parameters in many models, such as linear regression and hidden Markov models. Then Bayesian inference is introduced, which can use prior knowledge and data to capture the distribution of the inferred parameters. This can be implemented using probabilistic programming, which provides an easy-to-use framework to compute the values of parameters in the models and the associated uncertainty.

2.5.1 Likelihood

The likelihood function indicates how likely it is for a particular outcome, x, to be observed given the parameters of a probability distribution, θ [8]. It provides a function that can be optimised to find the parameters of a probability distribution that best match the observed data, and it can be expressed as,

$$\mathcal{L}(\theta|x) = P(x|\theta).$$

Given a dataset, $\mathbf{x}$, of n observations, the likelihood is the joint probability distribution of each observation in the dataset given the parameters of the probability distribution, which can be expressed as,

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i}^{n} P(x_i|\theta).$$

However, for large datasets, computing this directly can lead to numeric overflow and underflow issues, hence, log-likelihood is often used for optimisation tasks since the log function is a continuous strictly increasing function, maximising (or minimizing) the log-likelihood will also maximise the likelihood [57]. Given the logarithm rule that $\log(AB) = \log(A) + \log(B)$, the log-likelihood can be expressed as a summation of the individual observations, rather than the product, helping avoid the overflow and underflow issues,

$$\mathcal{LL}(\theta|\mathbf{x}) = \sum_{i}^{n} \log\left(P(x_i|\theta)\right).$$

2.5.1.1 Censored Likelihood

Datasets can have censored observations, which occur when the true observation is not known, and the only knowledge is that it is greater than, or less than, a limit. To account for censored observations in the likelihood function, the probability density function (PDF), $P(x|\theta)$, which represents the likelihood of a single observation, is replaced with a cumulative distribution function (CDF) $P(X \leq x|\theta)$, where X is a random variable, to represent the probability of the observation being greater than, or less than, the known limit [75]. Censoring can happen both above and below meaning the likelihood function becomes,

$$\mathcal{L}(\theta|X,R,L,I) = \prod_{i\in D} f(x_i|\theta) \prod_{i\in R} P(X \leq u_i|\theta)$$
$$\prod_{i\in L}(1 - P(X \leq l_i|\theta)) \prod_{i\in I} P(X \leq l_i|\theta) - P(X \leq u_i|\theta).$$

where X is the set of uncensored data, R is the set of censored above data, L is the set of censored below data, and I is the set of censored above and below data. The observed value is given by x_i, the upper bound for the censored above values is given by u_i, and the lower bound for the censored below values is given by l_i. Hence, the log-likelihood for censored readings is expressed as:

$$\mathcal{LL}(\theta|X,R,L,I) = \sum_{i\in D} log\left(P(x_i|\theta)\right) + \sum_{i\in R} log\left(P(X \leq u_i|\theta)\right)$$
$$+ \sum_{i\in L} log\left(1 - P(X \leq l_i|\theta)\right) + \sum_{i\in I} log\left(P(X \leq l_i|\theta) - P(X \leq u_i|\theta)\right).$$

2.5.2 Likelihood Optimisation

The likelihood is often used as a metric that is maximised to find the parameters of the probability distribution or statistical model that best represent the given data,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_i \mathcal{L}(\theta|x_i).$$

To find the parameters which maximise the likelihood a variety of methods can be used, depending on the properties of the distribution or model as factors such as if the likelihood function is differentiable, continuous, uni-modal, or multi-modal define how the likelihood can be found. In some cases, it can be found analytically and calculated directly, in other situations numerical methods are required to approximate the parameters which maximise the likelihood [57].

In this section, three scenarios of likelihood maximisation are outlined. Linear regression and hidden Markov models are introduced, and specific methods for each model that finds the most likely parameters by likelihood maximisation are explained. Then the golden section search is explained, which is a more generic, easy-to-use tool for optimization that approximates the likelihood optima in certain conditions.

2.5.2.1 Linear Regression

Likelihood optimisation can be used to find the parameters for a model which best describe the data, such as in linear regression, which is a model that assumes a linear relationship between the vector of inputs, $\mathbf{x}$, and the output, y. The probabilistic model assumes there is Gaussian noise on the output, and ordinary least squares can be used to find the maximum likelihood gradient and intercept. The linear relation is expressed as,

$$y = \boldsymbol{\omega}^T \mathbf{x},$$

where $\boldsymbol{\omega}$ is a vector of weights $\omega_{0:p}$, of which the dot product is taken with the inputs, $\mathbf{x}$, which is a vector of p features, $x_{0:p}$. A special case of the weights is ω_0 which is the bias term, and $x_0 = 1$. The purpose of linear regression is to learn the values of the set of weight terms by using a dataset, of known inputs and outputs. Across the dataset of n points, it can be expressed as

$$\mathbf{y} = \boldsymbol{\omega} \mathbf{X}$$

where $\mathbf{y}$ is the vector of outputs, $y_{0:n}$, and X is a matrix, where each of the n rows consists of p features. To learn the weights, $\boldsymbol{\omega}$, the ordinary least squares method can be used which minimises the Euclidean distance between the output and the output predicted by the model [92],

$$\boldsymbol{\omega} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} = ||\mathbf{y} - \boldsymbol{\omega}\mathbf{X}||^2.$$

This optimisation problem can be solved analytically by differentiating with respect to $\boldsymbol{\omega}$, and finding the turning points by solving for when the differential equals 0. This means the weights can be analytically found with the following equation,

$$\hat{\boldsymbol{\omega}} = (X^T X)^{-1} X^T y.$$

The theoretical computational complexity is $O(n^2 p + p^3)$, hence in practise on larger datasets, gradient descent methods may be preferred to get an approximate solution.

The covariance matrix of the parameters in the linear regression model is given by,

$$\text{cov} = \sigma^2 (X^T X)^{-1}$$

where,

$$\sigma^2 = (I - X(X^T X)^{-1} X^T) y,$$

and I is the identity matrix. Hence, the uncertainty of parameters in linear regression is typically explained by the standard error, capturing how dispersed the data used to find each inferred parameter is.

2.5.2.2 Hidden Markov Model

A hidden Markov model (HMM), can be used to model a process where there is an underlying cause that is not observable, generating observable outputs, and likelihood optimisation, can be used to find the parameters of the HMM that best capture the observed data. A Markov process is a stochastic model where the probability of the next state only depends on the previous state. An HMM, is a type of probabilistic graphical model with a Markov process on the unobservable states, $h_{1:T}$, and the only observable states, $x_{1:T}$, are based on emissions from the hidden state,

$$P(x_{t+1}|x_t) = P(x_{t+1}|x_t, x_{t-1}, ..., x_{t-T}).$$

The joint probability distribution of the HMM is defined as:

$$P(x_{1:T}, h_{1:T}) = P(x_1|h_1)P(h_1)\prod_{t=2}^{T} P(x_t|h_t)P(h_t|h_{t-1}).$$

Where the transition distribution, $P(h_t|h_{t-1})$, defines the probability of moving from the unobservable state h_{t-1} to state h_t, and the emission distribution, $P(x_t|h_t)$ defines the probability of observing state x_t given the hidden state h_t. Both of these distributions can be defined by discrete matrices of lookup values, or by continuous probability distribution functions.

In the instances when the parameters of the HMM are unknown, the Baum-Welch algorithm can be used to find the most likely initial distribution, transition and emission probabilities with dynamic programming [8]. The joint probability of the observed data up to time t, $x_{0:t}$, and the state, h_t, is expressed as,

$$\alpha(h_t) = P(x_{0:t}, h_t) = \sum_{h_{t-1}} \alpha(h_{t-1})P(h_t|h_{t-1})P(x_t|h_t),$$

and the conditional probability of the observed data for the rest of the sequence, $x_{t+1:T}$, and the state h_t, is expressed as,

$$\beta(h_t) = P(x_{t+1:T}, h_t) = \sum_{h_{t+1}} \beta(h_{t+1}) P(h_{t+1}|h_t) P(x_{t+1}|h_{t+1}).$$

These can be calculated recursively with the starting conditions,

$$\alpha(h_0) = P(h_0, x_0) = p(x_0|h_0)p(h_0),$$

$$\beta(h_T) = 1.$$

By calculating the α, and β values the probability of a hidden state given the observations can be expressed as,

$$\eta(h_t) = P(h_t|x_{0:T}) = \frac{\alpha(h_t)\beta(h_t)}{P(x_{0:T}),}$$

and the joint probability of a hidden state and the next hidden state given the observations can be expressed as,

$$\xi(h_t, h_{t+1}) = P(h_t, h_{t+1}|x_{0:T}) = \frac{\alpha(h_t)\beta(h_{t+1})P(h_{t+1}|h_t)P(x_{t+1}|h_{t+1})}{\sum_{h_t} \alpha(h_t)\beta(h_{t+1})P(h_{t+1}|h_t)P(x_{t+1}|h_{t+1})}.$$

Then, these functions are used to express the most likely values for the initial distribution,

$$\pi_0^* = \eta(h_0),$$

the transition probabilities,

$$a_{ij}^* = P(h_t = j|h_{t-1} = i) = \frac{\sum_t \xi(h_t = j, h_{t-1} = i)}{\sum_t \eta(h_{t-1} = i)},$$

and the emission probabilities,

$$b^*_{ij} = P(h_t = j|x_t = i) = \frac{\sum_t \eta(x_{t=i})\mathbb{I}_{x_t=j}}{\sum_t \eta(h_t = i)}.$$

Where a_{ij} represent entries in the transition probability matrix, b_{ij} represents entries in the emission probability matrix, and π_0 represents the initial states. These terms can also be replaced with probability distribution functions.

Once the most likely parameters of the algorithm have been inferred the Viterbi algorithm can be used to find the most likely sequence of hidden parameters in the HMM [8]. This is also achieved via dynamic programming, using the recursive relation

$$\mu(h_t) = \max_{h_{0:t-1}} P(h_{0:t}, x_{0:t}) = \max_{h_{t-1}} \mu(h_{t-1})P(h_t|h_{t-1})P(x_t|h_t).$$

and the arguments which maximise this recursion can be selected as the most likely path.

2.5.2.3 Golden Section Search

A basic method for finding the extremum value of a function is the golden section search, which can be applied to a likelihood function as long as the values can be calculated. The golden section search can be cheap to calculate as it does not require any gradient calculations. However, if the function is not uni-modal it may get stuck at a local optimum. The function is evaluated at four points, which are initially selected by setting an outer bound of where the maximum can be found and taking two points using the golden ratio inside those bounds. All four points are evaluated and the two points at either side of the maximum of the four points are selected as the new outer bounds and this process is repeated until convergence.

More specifically, for the interval [a,b], and the golden ratio, $\gamma = \frac{1+\sqrt{5}}{2}$, we calculate the points $c = b - \frac{b-a}{\gamma}$ and $d = a + \frac{b-a}{\gamma}$. For the function we are trying to maximise, $f(x)$, if $f(c) < f(d)$ then $b = d$, else $a = c$. This process is iterated until convergence is achieved by the difference between c and d being less than a selected tolerance value. The inequality can be flipped to find the minimum instead of the maximum.

2.5.2.4 Bayesian Inference and Sampling

In some cases, inferring parameters via likelihood maximisation, and using standard error to capture the uncertainty is reductive and it can cause useful information to be lost [21]. Instead, Bayesian inference can be used, which uses the likelihood, $P(y|\theta)$, to update the prior probability, $P(\theta)$, which is ones belief before any new data is accounted for. Then Bayes rule can be used to infer the posterior distribution, which is the updated beliefs given the likelihood and prior,

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}.$$

The posterior distribution is returned by propagating through the probability distributions of the prior and likelihood distributions, and it can capture the probability distribution of the parameter being inferred [8]. In some cases the posterior distribution can be calculated analytically, however, it is often the case that it is intractable, or difficult to solve, and sampling can instead be used to approximate the posterior distribution [56].

The premise behind the sampling approach is that if the distribution is difficult to calculate, it may be easier and more computationally efficient to draw samples from it and use these samples to compute an approximation of the posterior distribution. This can be achieved with Markov chain Monte Carlo (MCMC) methods, which create a Markov chain that has a stationary distribution that matches the distribution we want to sample from [56]. Then random sequences of states from the Markov chain are simulated to generate samples.

The Markov chain is constructed by samples being drawn from the probability distribution, and the next sample being drawn is dependent on the last sample. To obtain sequences of random samples, the Metropolis-Hastings algorithm can be used to draw random samples from probability distributions that could be difficult to sample from otherwise [36]. Improvements have been made on this algorithm such as the proposal of the Hamiltonian Monte Carlo (HMC), which uses Hamiltonian dynamics and gradient information from the likelihood to produce distant proposals from the previous one, increasing the speed of exploration [17]. Furthermore, HMC can be improved further with the no-U-turn sampler (NUTS) which addresses issues with tuning the step-size during simulations [60].

To perform Bayesian inference via sampling, probabilistic programming can be used, such as the Python package PyMC3 [110]. It provides a simple framework where models can be constructed with syntax similar to the programming language being used. Automatic Bayesian inference can be performed over the probabilistic model created by the user, with the data provided by the user, to find the posterior distribution of the parameters of interest. PyMC3 uses HMC, with the NUTS sampler to find the posterior distributions without specialised knowledge of the model. By using syntax that is easy to interpret for Python users and statisticians alike, and sampling methods that work across a broad range of models, it makes Bayesian inference more accessible to perform in computational settings. To demonstrate the simplicity of using probabilistic programming, an example of a Bayesian linear regression implementation in PyMC3 is given.

To create the prior distributions, a stochastic random variable which is drawn from a probability distribution can be created,

$\omega_0 \sim \mathcal{N}(0, 10) \Leftrightarrow$ `omega0 = pm.Normal("omega0", mu=0, sigma=10),`

$\omega_1 \sim \mathcal{N}(0, 10) \Leftrightarrow$ `omega1 = pm.Normal("omega1", mu=0, sigma=10).`

Many probability distributions are built-in, and if they are not they can be constructed as a function with mathematical operators in Python. Once the priors are defined, the model can be constructed as a deterministic relationship specified over the random variables and observed data,

$\mu = \omega_0 + \omega_1 * X \Leftrightarrow$ `mu = omega0 + (omega1 * X).`

Then, the output data can be stated as an observed variable, these variables can not be changed by the fitting algorithm, and represents the likelihood of the model,

```
Y_obs = pm.Normal("Y_obs", mu=mu, sigma=sigma, observed=Y).
```

With the model constructed and data ready, the final step is to run the Bayesian inference process to find the posterior distributions. PyMC3 comes with a number of in-built sampling and optimisation methods, with the default setting being a HMC and NUTS sampler with automatic initialisation based on a tuning phase. It can be run with one line of code,

```
pm.sample(1000).
```

After the sampling has been completed, it returns the posterior distributions which are shown in Figure 2.1, and the complete code for the implementation is as follows:

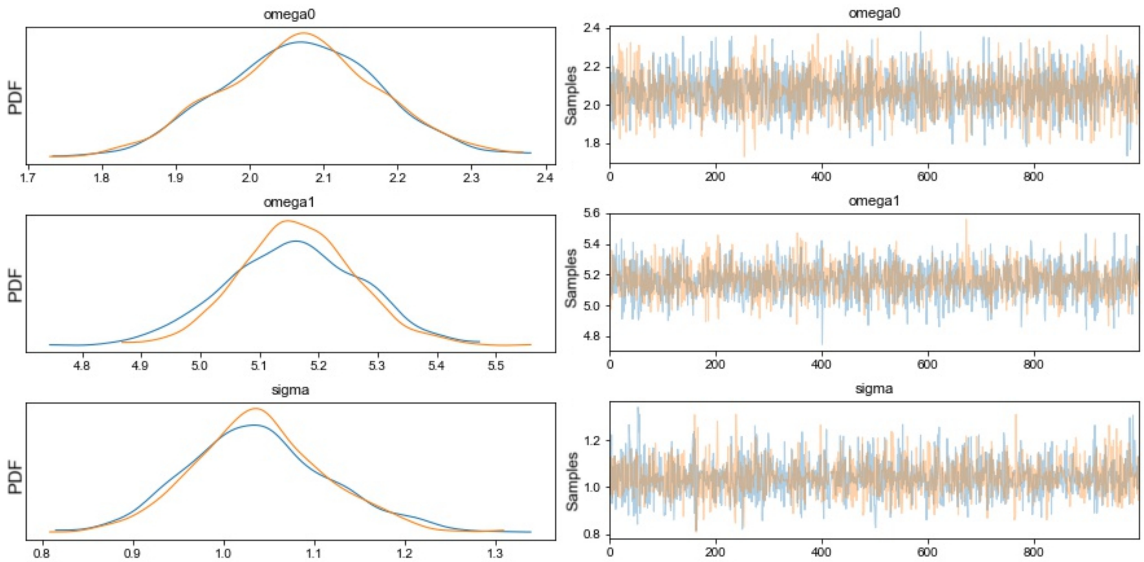

Figure 2.1: The left column show the smoothed histograms of the posterior distribution for each random variable, whilst the right column shows the Markov chain samples which can be used to spot issues in the model or sampling process. The different coloured lines each represent a run of MCMC, and using multiple chains can be used to identify problems. The expected value across all chains are then used.

```
import numpy as np
import pymc3 as pm


num_datapoints = 100
# True parameter values
sigma, omega0, omega1 = 1, 2, 5
# Simulate dataset
X = np.random.randn(num_datapoints)
Y = omega0 + (omega1 * X) + (np.random.randn(num_datapoints) * sigma)


model = pm.Model()
with model:
    # Prior distributions
```

```
omega0 = pm.Normal("omega0", mu=0, sigma=10)
omega1 = pm.Normal("omega1", mu=0, sigma=10)
sigma = pm.HalfNormal("sigma", sigma=1)
# model of how expected values are generated
mu = omega0 + (omega1 * X)
# Likelihood, given the observed data
Y_obs = pm.Normal("Y_obs", mu=mu, sigma=sigma, observed=Y)
# Use the built in HMC & NUTS sampler
trace = pm.sample(1000)
```

3

Learning Residential PV Generation from Censored Smart Meters

Reducing the GHG emissions from residential buildings requires shifting energy generation to lower emission generation sources and currently PV panels are one of the most viable methods of onsite generation. Unfortunately, the installation of PV panels are not always known due to the fact that schemes no longer operate in the UK to report PV panels. The maximum PV panel generation is reported in the EPC, however it can be up to 10 years between EPC evaluations which means there can be a significant delay between when the PV panel is installed and when it is reported [23]. Furthermore, it relies on an in-person inspection or survey which makes approximations about the PV generation [74]. It is possible to measure the residential energy generation with smart meters and this data could be used to evaluate the maximum PV generation. However, they fail to accurately record PV generation in over 50 % of cases due to the PV panels and smart meters not being correctly connected during installation. Quantifying small scale domestic PV generation from energy consumption is increasingly important as the install base of small residential PV panels rapidly grows.

If the PV panel and smart meter are not correctly connected, they only record the net energy taken from the grid and do not disaggregate energy taken from the grid from the energy fed back to the grid. Furthermore, they also fail to report the net exported energy to the grid, and simply report 0 kWh of consumption over each half-hour during these periods. In statistical terms, these observations are said to be censored. Thus, there is a need to develop approaches to identify homes with PV panels installed, and to estimate their real-time net contribution to overall grid generation, by disaggregating domestic energy generation from censored smart meter data.

In this chapter, an approach to disaggregate the PV generation and energy consumption from censored smart-meter readings is presented. To achieve this, the smart-meter data is combined with a solar proxy, which is defined as the known solar generation from a local house, and with these two sources of data the PV capacity can be inferred. Using a solar proxy to infer the PV generation of a house has been successfully demonstrated in other bodies of research, and it is assumed that there are enough local recordings to provide a reading with the correct tilt and orientation [122, 32]. Whilst over 50 % of PV panels are incorrectly installed, there is still a large proportion which are installed correctly and can be used to provide this information. To find the PV capacity, the most likely joint probability distribution of the PV capacity and the energy consumption for each time period across a year of weekdays is found. Only weekdays are used as the energy usage profile of weekdays and weekends are typically different.

This chapter is structured as follows. Section 3.1 outlines the model to infer PV capacity from censored smart-meter readings. Section 3.2 illustrates the implementation of the algorithm to infer PV capacity, which is extended in Section 3.3 to not require the solar proxy in some scenarios. Then Section 3.4 introduces the datasets, Section 3.5 outlines the experiments and results, and Section 3.6 concludes the chapter by summarising the contributions.

3.1 Censored Solar Disaggregation Model

To infer the capacity of the PV panel, the energy usage and generation need to be modeled. The smart-meter readings are considered, r_{htd} [kWh], where $h \in [1...H]$ denotes the H households, $t \in [1...T]$ denotes the T time-steps of the day, and $d \in [1...N]$, denotes the N days of data. A smart meter records only the energy supplied to the house from the grid. This means if behind-the-meter energy generation, g_{htd} [kWh], is larger than the energy consumption, x_{htd} [kWh], there is a censored reading, $r_{htd} = \max(0, x_{htd} - g_{htd})$. A diagram is provided in Figure 3.1.

3.1.1 Energy Generation Model

The PV panel is assumed to be the only source of possible energy generation in the household. A PV panel generates energy according to $g_{htd} = \tau c_h f_{td}$, where τ [hours] is the constant time step size and c_h [kW] is the PV capacity, namely the maximum power input into the home from the PV panel. In particular, $c_h = 0$ kW implies that there is no PV panel present. A solar proxy, is used as an input, which is defined as the known solar generation from a local PV panel [122]. As there is a large proportion of PV panels that are installed correctly, it is realistic to believe that a solar proxy would be a viable data source. The solar proxy factor, $0 \leq f_{td} \leq 1$, is the proportion of the

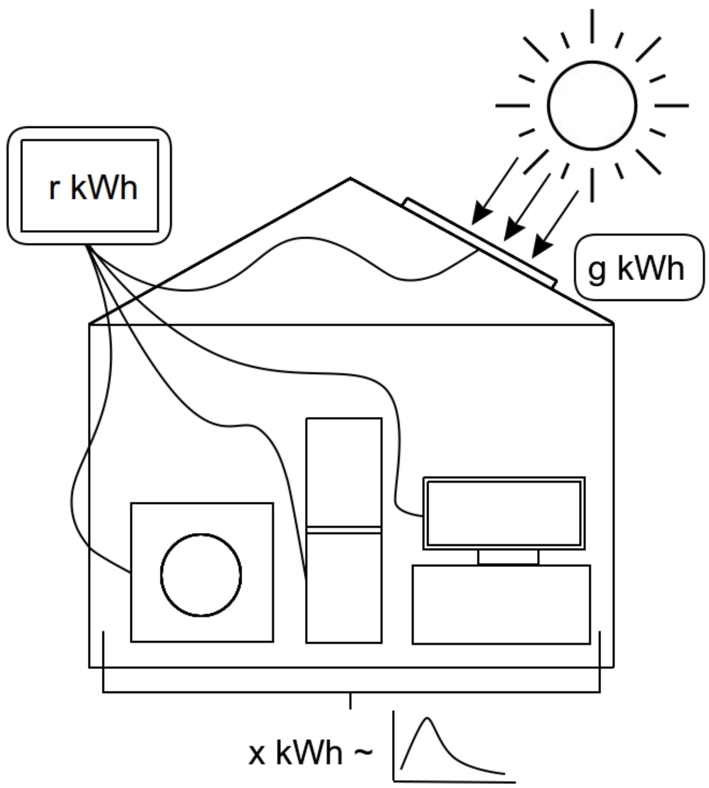

Figure 3.1: The smart-meter reading and solar proxy are the only observations made, and the energy generation and consumption need to be inferred. The fact that energy consumption follows a probability distribution for a given time period, and the solar proxy is known is used to disaggregate the energy generation from the consumption.

solar proxy recorded divided by the maximum recorded solar proxy for the year. The accuracy with which the solar proxy correctly represents the solar energy generation of the house in question will improve as the scale of smart meter installations increases, as it will reduce the distance to the nearest PV panel.

3.1.2 Energy Consumption Model

It is assumed that the energy consumption, x_{htd}, for a given time step (e.g., 10:00-10:30, with $\tau = 0.5$ h) comes from a probability distribution, P, parameterised by the set Θ_{ht}, valid for weekdays, d, as,

$$X_{htd} \sim \mathrm{P}(\Theta_{ht}). \tag{3.1}$$

The gamma distribution has been selected as the probability distribution to model energy consumption for a given time step in a residential household. This is because it is a continuous probability distribution that can model variables that are always positive and have asymmetric distributions. Only weekdays are used as weekends and weekdays typically have different energy usage profiles meaning it would be incorrect to assume the energy usage for each can be drawn from the same probability distribution. Furthermore, it is preferred over similar distributions such as the exponential, Erlang, and chi-squared distributions as they are all special cases of the gamma distribution, meaning this choice is more generalisable which allows for different energy usage profiles.

An indicator function is defined as $\mathbb{I}_{htd} = 1$ if $r_{htd} > 0$, and equal to 0 if censored for the respective time step, day, and household. A point is defined as censored if $r_{htd} = 0$ kWh. For the uncensored observations, the energy consumption is:

$$x_{htd} = r_{htd} + \tau c_h f_{td}, \tag{3.2}$$

whereas in the censored case it is known that the unknown energy consumption, x_{htd}, is bounded above by the energy generated, as:

$$x_{htd} \leq \tau c_h f_{td}. \tag{3.3}$$

To succinctly describe quantities over whole populations, sets of data corresponding to each house and time step are defined. Also sets comprising values of indicator functions, energy consumption values, smart meter readings and solar proxy factors for a time step, are introduced as: $\mathbb{I}_{ht} = \{\mathbb{I}_{htd}, d \in [1...N]\}$, $\boldsymbol{X}_{ht} = \{x_{htd}, d \in [1...N]\}$,

$\boldsymbol{R}_{ht} = \{r_{htd}, d \in [1...N]\}$, and $\boldsymbol{F}_t = \{f_{td}, d \in [1...N]\}$, respectively. Also defined are the sets of alpha and beta parameters for the gamma distribution, PV capacity, indicator function values, energy consumption values and smart-meter readings for a specific time step and day across houses, as follows: $\boldsymbol{\alpha}_t = \{\alpha_{ht}, h \in [1...H]\}$, $\boldsymbol{\beta}_t = \{\beta_{ht}, h \in [1...H]\}$, $\boldsymbol{C} = \{c_h, h \in [1...H]\}$, $\mathbb{I}_{td} = \{\mathbb{I}_{htd}, h \in [1...H]\}$, $\boldsymbol{X}_{td} = \{x_{htd}, h \in [1...H]\}$ and $\boldsymbol{R}_{td} = \{r_{htd}, h \in [1...H]\}$.

3.2 PV Capacity Inference with Solar Proxy

A novel algorithm to infer the PV capacity, c_h, is presented using a smart meter with censored readings, r_{htd}. For a given house, the algorithm aims to find the maximum power input into the home that can be expected from the installed PV panels. Notice that there is a counter-intuitive nature about the algorithm, as the PV capacity, which is the main point of interest, is found in the process of estimating the most likely distributions for energy consumption, X_{htd}.

The algorithm finds the most likely gamma distribution representing the energy consumption for each time slot. First consider dealing with uncensored smart-meter readings, namely when all the smart-meter readings are known, then by finding the most likely energy consumption, the most likely PV generation is also found, g_{htd}, via Equation ((3.2)). This approach is extended to when the smart-meter readings are censored and there is not a direct relation between the energy consumption and PV generation, the only relation known is the inequality seen in Equation ((3.3)).

We work with half-hourly smart meter data ($\tau = 0.5$ h) and only use the time steps when the sun is shining since a solar proxy factor equal to 0 does not help with the task of inferring PV generation. The algorithm is trained using a years worth of half-hourly readings of weekdays from historic smart meters and the recorded solar proxy.

It is worth reiterating that whilst the smart meter readings are known, the energy consumption values are unknown and are being inferred. Once the PV capacity, c_h, is known, it can be used to calculate the PV energy generation for any time step with the solar proxy, f_{htd}. Also, note that the inferred value of the PV generation is constrained as a scalar value (the PV capacity) is being inferred, which is shared across all time steps for the house, assuming the PV capacity remains constant, and this value is multiplied by the solar proxy, f_{htd}, for that specific time step to give the energy generation.

In this work, the energy consumption is assumed to follow a gamma distribution across days for each half-hour period in a specific house. Other distributions, such as the log-normal distribution, lend themselves to analytical solutions and may prove to be better assumptions when working with different datasets where energy consumption patterns are different. The unknown parameters of the gamma distribution, denoted as α_{ht}, and β_{ht}, are found by maximum likelihood estimation. When the true value in some cases is not observed due to the smart meter censoring readings below 0 kWh, a specific likelihood function can be used, which deals with data that is censored below [75]. To find both the PV capacity and the parameters of the gamma distribution representing the energy consumption, these parameters are iteratively updated in turn to maximise the likelihood.

3.2.1 Log-Likelihood of Censored Smart Meter Readings

A standard likelihood function is the product of probability density functions (PDF) for all the given data points. The PDF at a specific value corresponds to the probability that a data point with that value is observed. When the smart meter reading is 0 kWh, the energy consumption is unknown and there is not a data point to calculate the corresponding PDF. However, it is known that for the censored data points the energy generated by the PV panel is an upper bound on the energy consumption. Hence, the cumulative distribution function (CDF) of the upper bound is used when the smart meter reading is censored, as it describes the probability that the energy consumption value is less than the upper bound [75]. In summary, the proposed model uses the CDF when the smart meter reading is censored and the PDF when it is uncensored. The PDF, ϕ, and CDF, Φ, for the gamma distribution are defined respectively as follows where the gamma function is represented as Γ, whereas γ denotes the lower gamma function:

$$\phi(x|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

$$\Phi(x|\alpha,\beta) = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}.$$

For a specific house and time, the gamma distribution parameters, α_{ht} and β_{ht}, can be inferred, parameterising the set of energy consumption values, $\boldsymbol{X}_{ht}$, given the calculated set of indicator functions $\mathbb{I}_{ht}$, the smart meter readings $\boldsymbol{R}_{ht}$, and the solar proxy readings $\boldsymbol{F}_t$. The likelihood of the energy consumption data can then be found by factorising across the days as follows:

$$\mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau) \prod_{d=1}^{N} \phi(x_{htd}|\alpha_{ht}, \beta_{ht})^{\mathbb{I}_{htd}} \Phi(\tau c_h f_{td}|\alpha_{ht}, \beta_{ht})^{(1-\mathbb{I}_{htd})}. \tag{3.4}$$

Note that the PDF of $x_{htd} = r_{htd} + \tau c_h f_{td}$ (as per Equation ((3.2))) when the smart-meter reading is uncensored and the CDF of predicted PV generation when censored. By maximising this likelihood function the most likely gamma parameters representing half-hourly energy consumption, α_{ht} and β_{ht}, and PV capacity, c_h, are found. The likelihood function is factorised across all weekdays (N) in the year.

As the log function is monotonically increasing, finding the maximum of the likelihood is equivalent to finding the maximum of the following log-likelihood expression:

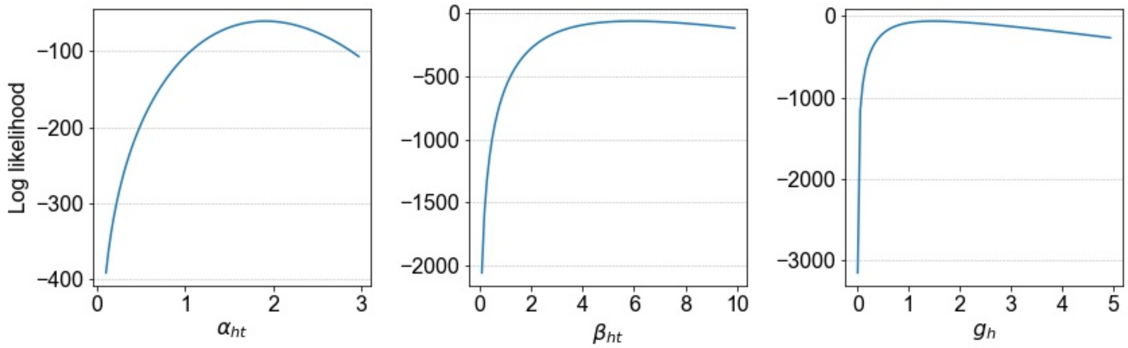

Figure 3.2: The log-likelihood function is uni-modal with respect to α_{ht}, β_{ht} and c_h, with the other parameters kept fixed.

$$\log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau) = \sum_{d=1}^{N} \mathbb{I}_{htd} \log(\phi(x_{htd}|\alpha_{ht}, \beta_{ht})) + \sum_{d=1}^{N} (1 - \mathbb{I}_{htd}) \log(\Phi(\tau c_h f_{td} | \alpha_{ht}, \beta_{ht})). \tag{3.5}$$

The golden section search is a standard technique used to find the extrema of a function. Since the log-likelihood of the gamma distribution with respect to each variable is uni-modal (see Figure 3.2), this technique is used to find the parameters that maximise the log-likelihood, with the other parameters being fixed. The benefit of using the golden section search is that there is no need to revise the approach if the distribution is changed, as only the PDF and CDF for that distribution need to be calculated. The golden section search finds the parameters of the gamma distribution, α_{ht} and β_{ht}, maximising the likelihood as follows:

$$(\alpha^*_{ht}, \beta^*_{ht}) = \underset{\alpha_{ht}, \beta_{ht}}{\operatorname{argmax}} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau). \tag{3.6}$$

Furthermore, the PV capacity can also be found using the golden section search since it is independent of the time of the day, and also the value of c_h that maximises the likelihood across all gamma distributions, corresponding to single time intervals of the day, namely:

$$c_h^* = \underset{c_h}{\operatorname{argmax}} \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau). \tag{3.7}$$

Algorithm 1: Iterative likelihood maximisation to find PV panel capacity

```
Result: Returns c_h
input : τ, r_htd, f_td ∀ t ∈ [1,...,T], d ∈ [1,...,N], h = 1
c_h = 1
// Identify censored readings
for t in 1:T; d in 1:N do
    if r_htd > τ c_h f_td then
        I_htd = 1
    else
        I_htd = 0
    end
end
// Find joint likelihood for all timesteps
// LL = Σ_{t=1}^T log L(α_ht, β_ht, c_h | R_ht, I_ht, F_t, τ) // Summation of Eqn (3.5)
// Finds parameters which maximise likelihood
while LL is not converged do
    for t in 1:T do
        (α*_ht, β*_ht) = argmax_{α_ht, β_ht} log L(α_ht, β_ht, c_h | R_ht, I_ht, F_t, τ) // Eqn (3.6)
        (α_ht, β_ht) ← (α*_ht, β*_ht)
    end
    for h in 1:H do
        c*_h = argmax_{c_h} Σ_{t=1}^T log L(α_ht, β_ht, c_h | R_ht, I_ht, F_t, τ) // Eqn (3.7)
        c_h ← c*_h
    end
    LL = Σ_{t=1}^T log L(α_ht, β_ht, c_h | R_ht, I_ht, F_t, τ)
end
```

To find the most likely value of the PV capacity, c_h, Algorithm 1 is used to maximise the joint log-likelihood of the gamma distributions representing the half-hourly periods of energy usage. The algorithm first identifies the censored data points. Then, the log-likelihood is maximised by finding the α_{ht} and β_{ht} parameters for each gamma distribution that best represents the energy consumption at each time step. Then for each house, the value of the PV capacity, c_h, that maximises the likelihood of the inferred energy consumption values being from their respective gamma distributions can be found. This is then repeated until the log-likelihood converges.

3.3 PV Capacity Inference with Only Clearsky Solar Irradiance

The above approach is extended to the inference of PV capacity only using clearsky (solar) irradiance (as opposed to a solar proxy for each panel), under the assumption that there is a group of houses in close proximity (e.g., sharing the same postcode). The clearsky irradiance, is the solar irradiance that would be incident on the ground if there was no atmospheric or weather interference. To identify how much of the solar irradiance is actually incident on the ground, the clearsky solar irradiance is multiplied by the clearsky irradiance factor, $f_{td} \leq s_{td} \leq 1$, defined as the solar irradiance at a particular time compared to the maximum recorded solar irradiance of the year.

Assuming all houses receive the same solar irradiance, an additional step is added in the previous algorithm to find an estimate of the solar proxy factor, f_{td}, for each time step across all houses, so that it maximises the likelihood that the inferred energy consumption values come from their corresponding distributions. Consider the new likelihood function which factorises the probability distribution across houses and is no longer conditioned on f_{td} as it is now being inferred:

$$\mathcal{L}_{\text{irr}}(f_{td}|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau) = \prod_{h=1}^{H} \phi(x_{htd}|\alpha_{ht}, \beta_{ht})^{\mathbb{I}_{htd}} \Phi(\tau c_h f_{td}|\alpha_{ht}, \beta_{ht})^{(1-\mathbb{I}_{htd})}. \quad (3.8)$$

This likelihood function is factorised over houses (H) in close proximity, as opposed to Equation ((3.4)) that factorises over days (N) for a single house. To infer the solar proxy, the likelihood is conditioned on the predicted gamma distribution parameters. This is due to the close houses sharing the same solar proxy, f_{htd}, but having unique PV capacity, c_h, and unique gamma distribution parameters representing energy consumption, α_{ht} and β_{ht}. The value of f_{td} maximising this likelihood is given by:

$$f_{td}^* = \underset{f_{td}}{\operatorname{argmax}} \log \mathcal{L}_{\text{irr}}(f_{td}|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau). \quad (3.9)$$

The approach finds the solar proxy factor that gives either the most likely energy consumption or an upper bound on the energy consumption in the case of censoring. This becomes an additional step in the algorithm compared to Algorithm 1, as the most likely gamma distribution parameters for each house, α_{ht} and β_{ht}, the most likely PV capacity for each house, c_h and the most likely solar proxy values, f_{td}, are inferred iteratively and until convergence.

Notice that this approach is realistically not feasible over a single house, as the inferred solar proxy factor would likely overfit, being dependent on a single data point. On the contrary, finding the solar proxy factor across a batch of houses (assumed to be adjacent) regularises the most likely solar proxy factor and allows for a prediction of its true value. This process is outlined in Algorithm 2. To emphasise, this is not intended to be an improvement of the above algorithm, but rather an approach that works with more limited data and an exploratory result towards alternative methods to using weather to predict solar irradiance for inferring PV generation.

Algorithm 2: Iterative likelihood maximisation to find PV panel capacity with only clearsky solar irradiance

---

Result: Returns c_h

input : $\tau,\ r_{htd}\ \forall\ h \in [1, ..., H],\ t \in [1, ..., T],\ d \in [1, ..., N]$

$c_h = 1\ \forall\ h \in [1, ..., H],\ f_{td} = 1\ \forall\ h \in [1, ..., H],\ t \in [1, ..., T],\ d \in [1, ..., N]$

`// Identify censored readings`

for *h in 1:H; t in 1:T; d in 1:N* **do**

  if $r_{htd} > \tau c_h f_{td}$ **then**

    $\mathbb{I}_{htd} = 1$

  else

    $\mathbb{I}_{htd} = 0$

  end

end

`// Find joint likelihood for all timesteps`

// $\mathcal{LL} = \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$ `// Summation of Eqn (3.5)`

`// Finds parameters which maximise likelihood`

while *$\mathcal{LL}$ is not converged* **do**

  for *h in 1:H; t in 1:T* **do**

    $(\alpha^*_{ht}, \beta^*_{ht}) = \underset{\alpha_{ht}, \beta_{ht}}{\operatorname{argmax}} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$ `// Eqn (3.6)`

    $(\alpha_{ht}, \beta_{ht}) \leftarrow (\alpha^*_{ht}, \beta^*_{ht})$

  end

  for *h in 1:H* **do**

    $c^*_h = \underset{c_h}{\operatorname{argmax}} \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$ `// Eqn (3.7)`

    $c_h \leftarrow c^*_h$

  end

  for *t in 1:T; d in 1:D* **do**

    $f^*_{td} = \underset{f_{td}}{\operatorname{argmax}} \log \mathcal{L}_{\text{irr}}(f_{td} | \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{C}, \boldsymbol{R}_{td}, \mathbb{I}_{td}, \tau)$ `// Eqn (3.9)`

    $f_{td} \leftarrow f^*_{td}$

  end

  $\mathcal{LL} = \sum_{t=1}^{T} \log \mathcal{L}(\alpha_{ht}, \beta_{ht}, c_h | \boldsymbol{R}_{ht}, \mathbb{I}_{ht}, \boldsymbol{F}_t, \tau)$

end

3.4 Datasets

To evaluate the approach a subset of the US Pecan Street dataset and a constructed UK dataset are used.

The Pecan Street dataset provides energy data for houses with PV panels [113]. The data has been cleaned and censored smart meter readings have been created for each house, obtained from their energy consumption and generation. From this dataset, 30 houses have been identified to have solar panels and selected and the PV Capacity is in the range 2.5 kW to 10.2 kW for the Pecan Street dataset.

Alongside this, a dataset is constructed representing smart meter readings from the UK, comprising solar energy generation from locations across the UK from the Sheffield Solar microgen dataset [126], and smart meter readings from London households[1]. Curation of this dataset is required as there are no large-scale public UK datasets that record energy usage, PV energy generation, and local solar irradiance. To create the dataset a selection of smart meter readings for 260 houses are taken from the London smart meter dataset and treated as energy consumption. A manual check was conducted to ensure that there is no PV energy generation on the selected houses. The fact that there is less than 30 minutes difference between the daily peak clearsky solar irradiance in London and Sheffield means the energy usage does not need to be realigned.

Each house has been randomly assigned one of the 50 available PV generation profiles from the microgen dataset. The maximum power output for each PV system in the microgen dataset has been re-scaled to demonstrate how the algorithm works across a range of judiciously selected PV capacity values in the range 0.5 kW - 6.5 kW with 20 values drawn randomly in each 0.5kW band, additionally, 20 houses were included with no PV panel (0 kW). The values are selected to cover the range of likely PV

[1] `data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households` (Accessed 01/09/21)

capacity values. Then, the half-hourly smart meter readings for the combined dataset are calculated using the formula $r_{htd} = \max(0, x_{htd} - \tau f_{td} c_h)$. Only the daylight hours of each day have been used, hence times of the day when there is no solar irradiance are not considered. As the assignment of the PV panels to the smart meters is random this can be considered as a worst-case scenario as there is typically a correlation between the energy usage of a home and the PV capacity. It is important to note that the solar proxy is required to closely match the energy generation profile of the house and that the algorithm does not currently correct for PV orientation which may need to be addressed when this is deployed if there are not enough local PV panels facing the same way.

To select the houses, those with erroneous readings, recorded as null, are removed and so are houses with a series of 0 kWh readings followed by a reading over 10 kWh, as this indicates that the smart meter has failed to report for a period of time and has recorded a backlog of energy usage. In practise, this would be dealt with in a data cleaning step, however, due to the abundance of data from the London households, it is possible to select houses with clean data for the experiments. Furthermore, only weekday data is used, as resident behaviour is typically different on a weekend which would affect the assumption that the energy usage follows the same distribution for each time step. In terms of the typical values of energy usage, over 85 % of the readings are under 0.5 kWh. This is expected as during the sunlight hours of a day residents are often out of the house and not much energy is used. There is also the case of data points being incorrectly labelled as censored when the energy usage is 0 kWh and solar generation is 0 kWh, however, this makes up less than 0.1 % of data-points and hence does not have a significant impact for the selected dataset but it is something that may be worth considering for other datasets.

The solar proxy has been implemented as seen in previous research by taking the known proportion of the PV energy generation through the day [122]. There is potential for transposition errors, where there is a difference between the plane of incidence for the solar proxy to the PV system that is being evaluated. However, it has been shown that the geometry of the proxy system can be recovered and mapped to the geometry of the PV system being evaluated removing this source of error [32]. Furthermore, it is shown that the factors other than orientation only affect the maximum power output, hence they can be absorbed into the PV capacity term which represents the true power input to the house from the PV system.

3.5 Evaluation

To evaluate the performance of the algorithm four metrics are used. The root mean square error (RMSE) measures the accuracy of the inferred values by taking its distance from the true value, regardless of the true PV capacity. RMSE is a useful metric in a situation when the total prediction of PV generation across all houses is of interest,

$$\text{RMSE} = \sqrt{\frac{\sum_{h=1}^{H}(\hat{c}_h - c_h)^2}{H}}.$$

Note that, $\hat{c}_h$ denotes the inferred PV capacity and c_h is the real PV capacity. To measure the precision the mean absolute percentage error (MAPE) is also calculated. The MAPE calculates the absolute error as a percentage of the real PV capacity,

$$\text{MAPE} = \frac{1}{H}\sum_{h=1}^{H}\left|\frac{\hat{c}_h - c_h}{c_h}\right|.$$

Furthermore, to identify the bias in the estimations the mean normalised bias error (MNBE) is calculated,

$$\text{MNBE} = \frac{1}{H}\sum_{h=1}^{H}\frac{\hat{c}_h - c_h}{c_h}.$$

The classification rate is also measured, which is the number of houses correctly predicted to have a solar panel. In order to classify the presence of a PV panel, a PV capacity threshold is selected, and assumed that if the inferred generation is above the given threshold, then a PV panel is present.

3.5.1 PV Capacity Inference with Solar Proxy

For the constructed UK dataset 260 houses are selected from the smart meter readings from the London households dataset that do not have a PV panel installed, and have treated the readings as energy consumption. They are then combined with PV generation as described in Section 3.4 to create the dataset used for the experiments. For the 30 houses from the US Pecan Street dataset the algorithm is run twice: once with the censored smart meter readings as described above, and once with the solar generation removed to emulate if the houses did not have solar panels.

3.5.1.1 Inference of PV Capacity

The first key insight from Table 3.1 is the 100% classification rate, meaning that the algorithm correctly identified every incident where there was a PV panel present and every incident where there was not. This was achieved by setting a threshold at 0.05 kW, and any value below this was reported as not having a PV panel. Furthermore, the results indicate that the performance of the algorithm does not vary with the size of the panel and there does not appear to be a range of values where the approach

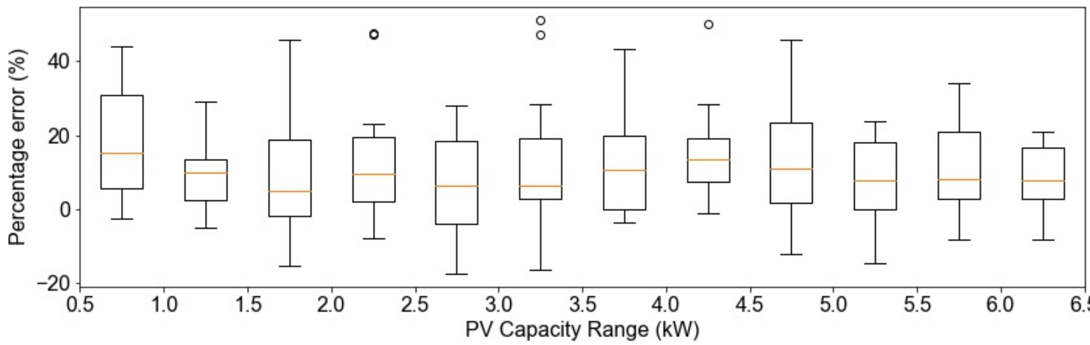

Figure 3.3: Each boxplot represents the distribution of percentage error for the 20 houses in each 0.5 kW band of PV capacity. The performance of the approach appears to be independent of the PV capacity, however, there is a bias and the approach consistently over-estimates the PV capacity which the approach can be adjusted for.

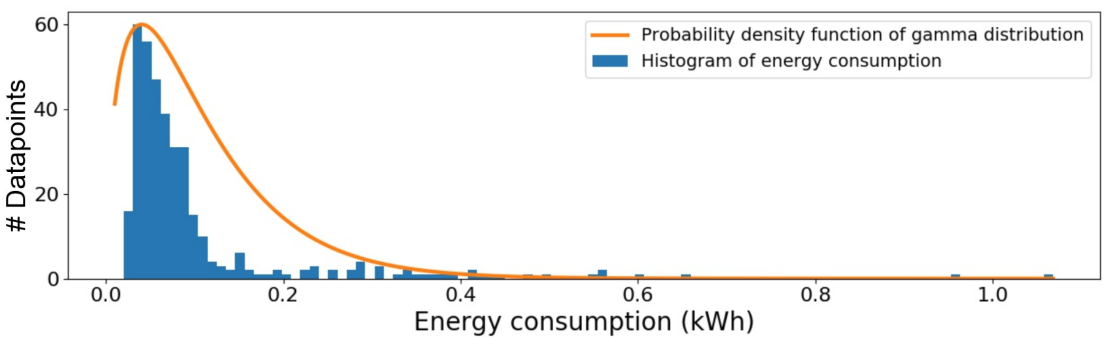

Figure 3.4: Half-hourly energy usage from a selected house and time-step. The corresponding gamma distribution is plotted over the data. This example demonstrates how the distribution can fail to accurately represent the energy usage distribution which leads to the bias in the results by over-estimating the energy usage and hence over-estimating the energy generated by the PV panels. It also shows that through the day of an unoccupied house the majority of data points correspond to states of low energy usage.

fails, showing the algorithm performs in the range of expected PV capacity values.

However, there is a consistent bias in the results as the MNBE and Figure 3.3 show. The algorithm typically over-estimates the PV capacity, which is caused by the distribution not correctly representing the energy usage in some cases and over-estimating energy usage, which is shown in Figure 3.4. Then for the censored readings the expected value is taken which is larger than the true value, leading to an over-estimate of energy usage and over-estimate of PV energy generation. This can be addressed by finding a distribution that better represents the energy usage in the specific scenario, or if the bias is known for the dataset the inferred values can be corrected for the bias.

	MAPE (%)	MNBE (%)	RMSE	Classification (%)
UK dataset				
0 kW	-	-	0.01	100
0.5 - 1.0 kW	18	17	0.14	100
1.0 - 1.5 kW	10	9	0.16	100
1.5 - 2.0 kW	13	9	0.32	100
2.0 - 2.5 kW	14	12	0.44	100
2.5 - 3.0 kW	13	6	0.41	100
3.0 - 3.5 kW	14	11	0.62	100
3.5 - 4.0 kW	15	13	0.73	100
4.0 - 4.5 kW	15	15	0.81	100
4.5 - 5.0 kW	15	12	0.92	100
5.0 - 5.5 kW	12	7	0.71	100
5.5 - 6.0 kW	13	11	0.96	100
6.0 - 6.5 kW	10	8	0.76	100
Average	**13**	**11**	**0.64**	**100**
Pecan Street				
Average	29	-18	1.68	100

Table 3.1: The MAPE, MNBE, RMSE, and classification metrics for each band of 20 PV capacity values on the UK dataset and across the Pecan street dataset. The performance is relatively consistent across all ranges of values with the RMSE increasing as expected due to the larger PV capacity values.

There is also a noticeable drop in performance from the constructed UK dataset to the Pecan Street dataset, with the MAPE across the dataset going from 13% to 29%. The drop in performance on the Pecan Street dataset is expected as it typically has PV systems with larger PV capacity, and if the PV generation is large relative to the energy usage it can be difficult to accurately infer the PV capacity as more information is censored. This explanation is validated by the MNBE showing an 18% under-estimate of PV capacity in the Pecan Street dataset on average due to the high energy generation censoring a large proportion of the smart meter readings. This is a difficult issue to address with censored smart meters and providing a lower bound on the PV capacity for these households may have to suffice. Perhaps, combining this approach with satellite imagery will provide an upper and lower bound on the potential PV generation.

3.5.1.2 Failure without Consideration of Censoring

To show the importance of correctly handling censored observations, it is demonstrated in Table 3.2 what happens if censored smart meter readings ($r_{htd} = 0$ kWh) are treated as a net reading and ignore censoring. Following Algorithm 1, and set $\mathbb{I}_{htd} = 1$ for all data points to achieve this result. The algorithm fails to detect the presence of PV panels or correctly infer the PV capacity, demonstrating the importance of using an appropriate likelihood function to deal with censored observations.

	UK Dataset		
	PV	No PV	All
MAPE	99.6%	-	-
RMSE (kW)	1.23	0.01	0.87
Classification	0.4%	98.2%	49.3%

Table 3.2: Error metrics and classification rate of PV capacity inference without censoring

3.5.2 PV Capacity Inference with Clearsky Solar Irradiance

For this experiment the constructed UK dataset is used, and the outcomes are presented in Table 3.3. A sample of 250 houses with PV panels and 250 houses without PV panels is taken. The smart meter readings have been calculated similarly to the previous experiment, and the actual PV generation from one of the microgen dataset houses is normalised and randomly assigned a PV capacity between 1.5 kW and 3.0 kW to each house with a PV panel, whereas the houses without a PV panel have been left as is. The solar irradiance and PV capacity data have then been disregarded and the algorithm is implemented to infer the PV capacity for each house. In the implementation, once the inference of a PV capacity has gone below 0.1 kW, it is classified that the house does not have a solar panel and it is removed from the next iteration as it would no longer contribute useful information to the problem.

	UK Dataset		
	PV	No PV	All
MAPE	33.0%	-	-
RMSE (kW)	0.55	0.01	0.39
Classification	100%	100%	100%

Table 3.3: Error metrics and classification rate of PV capacity inference with clearsky solar irradiance

Whilst the RMSE is similar to the previous results for the constructed UK dataset, the MAPE is larger (more than twice as large), indicating that the approach without solar irradiance incurs a larger error in houses with solar panels of smaller PV capacity. This is the first approach to infer PV capacity that does not require solar irradiance or weather data as an input.

Whilst in practice if the location is known then weather data could be acquired, the purpose of this implementation and experiment was to show the feasibility of alternative approaches to solar dissagregation, in particular approaches that do not inherit errors related to inferring solar irradiance from weather data. Weather-based approaches to solar disaggregation to infer PV capacity are still considerably more accurate, however, this shows the feasibility of an alternative approach and could inspire related research efforts that use local clusters of households to infer PV generation [122].

3.5.3 Runtime and Scalability

The runtime on a single Intel 7$^{\text{th}}$ gen i7 CPU core to detect the presence and capacity of a PV panel averaged at 9.32 seconds per house (8.54 seconds per house excluding data loading) across the 240 houses with PV panels and at 5.74 seconds per house (5.06 seconds per house excluding data loading) for the 20 houses without PV panels.

There are approximately 25 million homes in the UK, of which about 1 million homes have PV panels, which means to infer the PV capacity for all buildings across the UK it would require 16000 CPU-hours. As each house uses a separate instance of the algorithm, the workload could be parallelised across multiple CPUs, and based on current AWS EC2 T2 prices could be conducted for under $400. Further speed-ups are expected by employing GPU-based computations.

A nationwide scan only needs to be run once to detect panels and their associated PV generation. The algorithm can be re-run at set intervals at the discretion of the user to update their dataset of solar panel locations and capacity. Once the capacity of the solar panel is known - the inference of expected solar generation for the next time step is in the magnitude of seconds to infer the generation for all houses with solar panels, as it is a single calculation using the solar irradiance. The run times of the algorithm and inference show that this approach is fit for purpose at scale and affordable with cloud computing services.

3.6 Summary

This chapter presents the first approach to dissagregate solar (PV) generation from energy consumption given censored smart meter readings and to infer the PV capacity. The method presents an approach which can be used to augment the EPC by enabling the PV generation of the residential building to be inferred by smart-meter readings. Furthermore, it can be used to help reduce GHG emissions from residential buildings by providing more accurate information on residential energy generation to ensure that grid operators can correctly manage energy distribution and plan the appropriate infrastructure required. A better understanding of PV generation helps homeowners understand their true PV capacity, it can help identify when their PV panel is potentially faulty, and it can be used to nudge others in the neighbourhood to adopt PV panels with better data communication. This is possible due to smart-meters enabling non-intrusive data collection and the method developed is robust against common installation errors of smart meters and PV panels.

Shifting energy generation to renewable sources is fundamental to reduce the GHG emissions from residential buildings. However, as wind and solar are leading sources of renewable energy, and their time of generation can not be controlled, only predicted, steps need to be taken to minimize energy usage and shift energy usage to align with energy generation.

4

Using Smart Meters to Identify Residential Thermal Performance

For residential buildings thermal space comfort is the biggest source of energy usage, hence, it is important that buildings have a good thermal performance [46]. Buildings which better retain thermal energy require less energy to reach the desired temperature and then the desired temperature can be maintained with less energy. A more accurate and easier to deploy process of quantifying thermal performance of a building can be used to help identify buildings with poor thermal performance and to help the residents take appropriate actions to improve it, particularly if they are fuel poor. Unfortunately, the current methods used by the EPC to evaluate the thermal performance have been shown to be inadequate by independent investigations [131, 73].

In this chapter SMITE is proposed, a non-intrusive method to infer thermal performance of a residential building with only smart meter and weather data (particularly, external temperature and solar irradiance). This is a scalable, non-intrusive, and accurate tool that can be deployed across all houses with a smart meter. A thermal model is proposed to represent how heat energy flows in and out of a building, and a hidden Markov model is proposed to identify periods where the building is in thermal equilibrium. Then a series of filters are used to remove situations which are difficult to model or do not contain useful information. Then the hidden Markov model is used on the remaining data to identify the periods when the heating is on. Inference can then be performed to evaluate the thermal performance of the building on these remaining data-points.

The rest of the Chapter is formatted as follows. Section 4.1 introduces the modulated heat flow model for a residential home. Section 4.2 outlines the algorithm implemented to identify the relevant data and infer the HLC and HPLC. The dataset is described in Section 4.3 and Section 4.4 presents the results achieved. The impact of this work, potential interventions to reduce carbon emissions, and benefits of HPLC over HLC are discussed in Section 4.5 followed by the conclusions in Section 4.6.

4.1 Residential Home Thermal Performance Model

This section introduces the models used to explain how heat energy flows in and out of a home, and the thermal equilibrium state that needs to be identified to infer HLC. The practicalities and limitations of thermal modeling in the non-intrusive setting are discussed, which leads to the definition of HPLC and how it is inferred. Furthermore, a model is defined to identify the periods when the home is in thermal equilibrium from the smart-meter readings. Comparisons with traditional co-heating tests are made throughout this section to emphasise how it replicates the environment required to conduct a co-heating test, whilst extending it to a non-intrusive situation where occupants are present and there is no control over how the home is being used or heated.

4.1.1 Thermal Equilibrium Model

The thermal equilibrium model builds upon the heat flow model introduced in Section 2.4.1, as it breaks the heat flow into components that are independent and can be modeled with data-driven approaches:

$$Q_h = Q_{fa} + Q_{ve} + Q_{we} + Q_o + Q_b. \tag{4.1}$$

The individual components of this equation are expressed the same as in Equation (2.8), where Q_{fa} is the heat flow through the fabric of the building, Q_{ve} is heat flow via ventilation, Q_{we} is the thermal contribution due to weather factors, Q_{occ} is the thermal contribution from occupants, and Q_b is the heat energy from other appliances.

However, it is not made explicit in Equation (4.1) that the building may not be in thermal equilibrium and therefore there is a temporal component that needs to be considered, rather than assuming that the internal temperature is fixed due to the heating being on [29]. Hence, a term is required in the model to account for the net heat flow into or out of the building from the previous timestep, $Q_{\Delta t}$, which is added to the thermal model,

$$Q_h = Q_{fa} + Q_{ve} + Q_{we} + Q_o + Q_b + Q_{\Delta t}. \tag{4.2}$$

Previous models have failed to explicitly consider that the building may not be in thermal equilibrium and assumed that the internal temperature is fixed due to the heating being on [29]. If the previous time-step has a net flow of heat in or out then this needs to be represented in the heat flow equation. This is represented by a function of the internal temperature at the current and previous time-step, and the thermal mass, M_T, of the building,

$$Q_{\Delta t} = f(T_{in_t}, T_{in_{t-1}}, M_T). \tag{4.3}$$

To get a state where the thermal properties of the building can be inferred, a state of thermal equilibrium is required. This means a heating or cooling system must be active to ensure the internal temperature is kept at a constant value. If this state is achieved, $Q_{\Delta t} = 0$ and it can be removed from the equation.

Hence, to infer the HLC, only time periods where the home is in thermal equilibrium should be considered. When the home is not in thermal equilibrium there is a difference between the heat energy flowing in and out of the home, resulting in a change in internal temperature. With current approaches, the ability to infer the internal temperature from external weather data and smart-meter readings is limited [80]. By restricting the model to only consider times when the home is in thermal equilibrium the flow of heat energy provided by the heating system is equal to the net flow of heat energy out of the home, assuming the thermal mass of the building remains constant. When the heat flows in and out are equal, the heat flow equations for the home is given by:

$$\eta_h P_h = Q_h, \tag{4.4}$$

which means P_h, the power of the heating system to maintain thermal equilibrium, is given by:

$$P_h = \frac{1}{\eta_h}(HLC\,\Delta T + c_{air}\rho_{air}V_{air}(T_{in} - T_{ex}) + A_{sol}I_{sol} + 0.06N_{occ} + \eta_b P_b), \tag{4.5}$$

where η_h is the efficiency of the heating system.

The HLC is only supposed to account for the heat flow through the fabric of the building [29]. In the non-intrusive setting, if the ventilation loss is unaffected by the behaviour of the residents, i.e. it is due to a draft through window sealing rather than due to it being open/closed, then this cannot be separated from the heat flow through the fabric as there are no direct measurements of ventilation heat loss being made. This type of ventilation heat loss from a property, not due to resident behaviour, is of importance when labeling the thermal performance of a home as without intervention this source of heat loss will remain.

Furthermore, to infer the HLC the heating system efficiency, η_h, must be known for the home or estimated based on manufacturer guidelines of the typical heating system efficiency. As Equation (4.5) shows, the HLC and heating system efficiency, η_h, are compounded which means a non-intrusive method cannot separate the two values when only considering periods where the home has the heating on and is in thermal equilibrium.

4.1.2 Heating Power Loss Coefficient (HPLC)

Since in general the heating system efficiency cannot be separated from HLC, it makes sense to develop non-intrusive approaches to infer the HPLC instead. The HPLC is defined as the rate of supplied power to maintain the home at a constant internal temperature, where the supplied power is the average energy consumption to heat the home (i.e. energy input to the heating system).

With the less strict definition of HPLC compared to HLC, the ventilation loss is split into two sources: behavioural ventilation loss, i.e. when heat is lost through open windows; and home ventilation loss, i.e. when heat is lost through the frame of windows. These two sources are separated by the volume of air lost by each source: V_{air_b} is the volume of air that leaves the house due to behavioural causes, whereas V_{air_d} is the volume of air leaving due to the structure of the home. The ventilation loss related to the home is absorbed into the HPLC term as it cannot be separated from the heat transfer through the fabric of the building. The HPLC is then expressed as follows,

$$HPLC = \frac{1}{\eta_h}(HLC - c_{air}\rho_{air}V_{air_d}), \tag{4.6}$$

and its relation with the power supplied to the home is given by

$$P_h = HPLC\,\Delta T + \frac{1}{\eta_h}(c_{air}\rho_{air}V_{air_b}(T_{in} - T_{ex}) + A_{sol}I_{sol} + 0.06N_{occ} + \eta_b P_b). \tag{4.7}$$

It is assumed that when the building is in thermal equilibrium there is no loss due to the behavioural ventilation.

4.1.3 Detection of Heating Periods

Co-heating tests are conducted with the heating continuously on [10]. This is not the case when occupants are in their home as the heating is set on and off to suit their thermal space comfort needs. Previous approaches have tried to account for the fact that the heating is not always on by inferring the internal temperature given the external temperature, as seen in Equation (2.9) [80]. However, this is an assumption that does not account for how long the heating is on and assumes every house operates like the average house. Hence, to replicate the conditions required for a co-heating test non-intrusively, the heating-on periods of the day need to be identified. The heating-on periods of the day can be recorded by devices such as smart thermostats, however, they are yet to be adopted at a large-scale. As such, until smart thermostats or similar devices are widely adopted, the heating-on periods of the day need to be inferred via smart-meter readings and local weather data. The SMITE method uses a hidden Markov model to identify these heating-on periods.

As thermal space comfort accounts for a majority of energy consumption in homes [16], it is assumed that energy consumption at any time t can be split into two modes: (i) when heating is off ($h_{on}^{(t)} = 0$), and (ii) when heating is on ($h_{on}^{(t)} = 1$) [13]. The power used to heat the home at time step t is denoted by $P_h^{(t)}$, and the power used for all other appliances at time step t by $P_b^{(t)}$. The ξ terms represent the noise associated with each scenario and are assumed to follow a Gaussian distribution. Then, $P_{all}^{(t)}$, the total power used by the home at time t, can be expressed as a function of the heating mode,

$$\mathbb{P}(P_{all}^{(t)} | h_{on}^{(t)}) \sim \begin{cases} P_b^{(t)} + \xi_1 & \text{if } h_{on}^{(t)} = 0, \\ P_h^{(t)} + P_b^{(t)} + \xi_2 & \text{if } h_{on}^{(t)} = 1. \end{cases} \tag{4.8}$$

The state of the heating system at any particular time t is unknown, however, based on prior knowledge of how a heating system is used and how it contributes to the power output, a model can be constructed to infer when the heating is on or off. As per Equation (4.8), the likelihood of the power recorded given each state of the heating system can be inferred, $\mathbb{P}(P_{all}^{(t)} | h_{on}^{(t)})$. A temporal component to heating systems makes them more likely to stay in the same state (on or off) for consecutive half-hour periods, rather than fluctuating between the two states. This is encompassed by the transition probabilities between the hidden states representing the probability

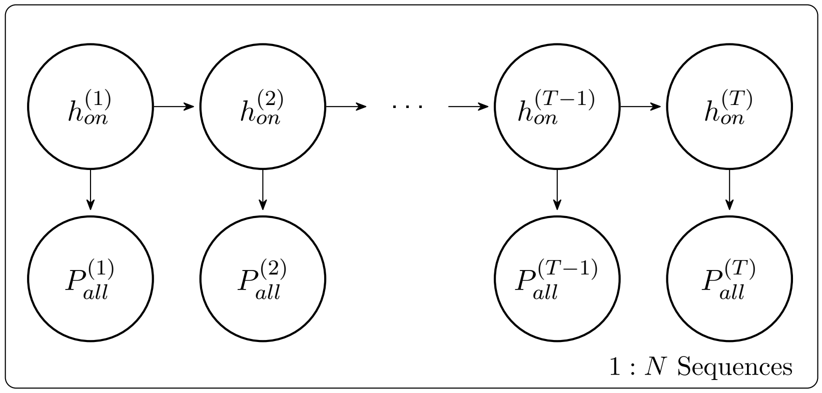

Figure 4.1: The hidden Markov model used to identify the periods when heating is on or off when the half-hourly average power is observed. This is performed for all N sequences of consecutive smart-meter readings that have passed through the filtering process. The transition matrix between hidden states and the parameters of the Gaussian distributions representing the emission probabilities are learned via the Baum-Welch algorithm.

of the heating switching between on or off, as $\mathbb{P}(h_{on}^{(t)}|h_{on}^{(t-1)})$. This scenario can be appropriately modeled with a hidden Markov model and the expression for the joint probability is given as

$$\mathbb{P}(h^{(1:T)}, P_{all}^{(1:T)}) = \prod_t \mathbb{P}(P_{all}^{(t)}|h_{on}^{(t)})\mathbb{P}(h_{on}^{(t)}|h_{on}^{(t-1)}), \tag{4.9}$$

where $\mathbb{P}(h_{on}^{(t)}|h_{on}^{(t-1)})$ is defined by a transition matrix, and the emission function $\mathbb{P}(P_{all}^{(t)}|h_{on}^{(t)})$ is defined in Equation (4.8). The joint probability is taken across all time steps, $t \in 1...\mathrm{T}$, where T is the number of time steps (T=48 for a day of half-hour periods). By maximising the joint probability in Equation ((4.9)) it is possible to find the most likely heating pattern for each day, identifying the heating-on periods. The model is outlined in Figure 4.1.

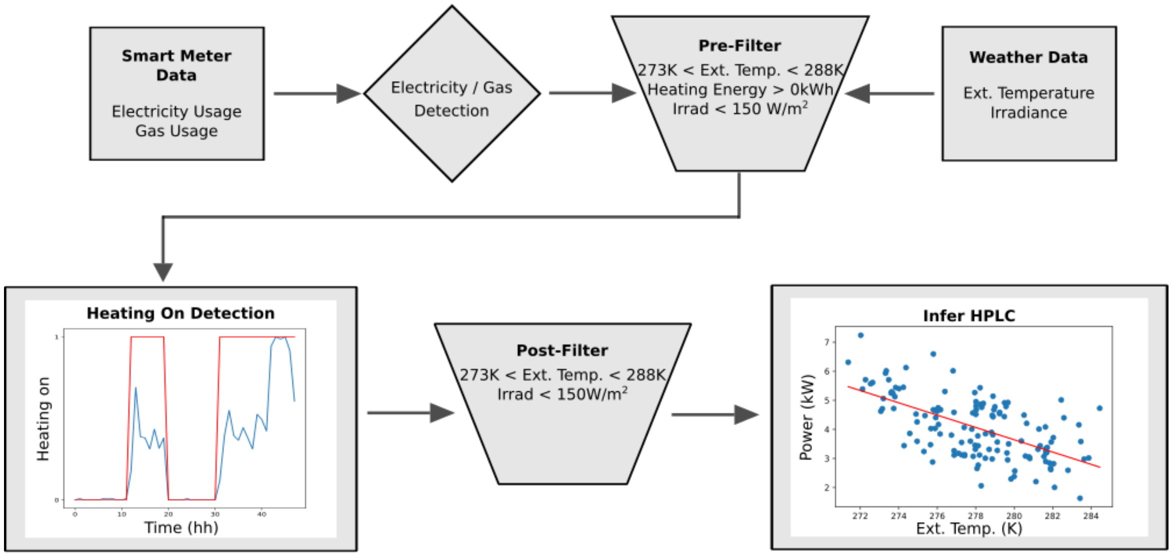

Figure 4.2: This flowchart outlines the steps of the SMITE method. The boxes in the upper right and left corners give the required input data. A check is performed on the smart-meter readings to infer if the home has gas or electric heating then the pre-filter removes days with high solar irradiance and external temperatures outside the range of accepted values. The cleaned data is used to infer the heating-on periods. The filters are applied again on the heating-on periods to ensure they do not coincide with a moment of high solar irradiance or external temperature. The HPLC is then inferred from the remaining data. The reason the gradient is negative is due to the external temperature, not the temperature difference being plotted as shown in Equation (4.10).

4.2 The SMITE Method

The thermal models and the hidden Markov model for detecting when the heating is on are implemented in the SMITE method to construct a post-hoc experiment to infer the thermal performance. This is done by extracting a year of data from the smart meter and following the SMITE method to select the data when the home is in thermal equilibrium and the energy used to heat the home is known and using this selected data to infer the HPLC (and HLC when heating system efficiency is available). To infer the HPLC, a three-step method is proposed: (i) identify the data points that have minimal factors affecting the internal temperature (other than the heating system and external temperature); (ii) identify the heating-on periods; and (iii) infer the HPLC and HLC from the selected data points. The aim is to reduce the model to a situation where the only parameter that needs to be inferred is the HPLC, this process is outlined in Figure 4.2.

The intrusive data collection process in the co-heating test is replaced by smart-meter readings and regional weather data which is a process that can be easily scaled across the country as it requires no additional hardware installations. A smart meter records energy usage every 30 minutes compared to being recorded every 40 minutes for a co-heating test. The similar frequency of recording of the smart meter means that the granularity of the data will not be a performance bottleneck when compared to the co-heating test.

4.2.1 Identification of Usable Data

Sampling is performed to reduce noise and remove terms that cannot be accurately monitored in a non-intrusive setting. With the current limited ability to model the thermal effects of factors such as solar irradiance [61], these steps are taken to identify and remove periods when these factors have a significant contribution to the heat flow into the home.

4.2.1.1 Data Cleaning

The first step is to remove days with corrupt data, missing data, and outlier energy readings. These outlier days are identified by calculating a Z-score to identify days where the smart-meter readings are in excess of what is realistically expected and days where 0 kWh of energy are recorded as it is either a faulty reading or provides no information. If there is an hour or more of consecutive readings that are missing, corrupt, or outliers, the data for the day is split into two sequences. However, if it is detected that the heating-on period begins or ends immediately before or after, it is discarded as there is no certainty that the full period is accurately recorded. If there is a single missing weather observation with recordings immediately before and after, the missing value is inferred linearly.

4.2.1.2 Gas or Electric Heating Detection

After the data is cleaned whether the source of heating energy is gas or electricity is identified. The heating energy source can be identified with 100% accuracy on the dataset used by identifying which source consumes the most energy through the winter months.

4.2.1.3 Temperature Filter

To ensure there is sufficient difference between the external and internal temperature, days with an average external temperature above 15°C (283.15 K) are removed. Further to this, a filter to exclude extremely cold days is added, if the average external temperature is below 0°C (273.15 K) it is also removed. This avoids situations where the heating system is running at maximum power for the duration of the heating-on period and the building does not reach a state of thermal equilibrium which was observed in the dataset used.

4.2.1.4 Solar Irradiance Filter

All days with an average solar irradiance above 0.15 kW/m^2 are filtered out to reduce the heat energy flow into the home from solar irradiance. This value is selected based on empirical results on the dataset and may be altered for different scenarios. Methods have been proposed to estimate the effects of solar irradiance on the temperature of a home, however, it introduces modeling uncertainty which is unnecessary if there is enough data after the filtering process to infer the thermal performance of a home [61].

4.2.1.5 Repeated Filters

The temperature and solar irradiance filters are applied twice. First against the average for the whole day; and again, on the average of the heating-on periods once they are identified. This filters out the days with large sources of external heat energy when detecting the heating-on periods and afterward on the half-hourly data-points to avoid scenarios where there are spikes in temperature or solar irradiance during a heating-on period.

4.2.2 Detection of Periods when Heating is On

To detect the heating-on periods a hidden Markov model is used, as outlined in Figure 4.1, with emissions affected by an additive Gaussian distribution. The only observations required are the half-hourly values of average power from the smart-meter readings. The hidden binary state represents whether the heating is on or off. Using the filtered data, each complete day (and the segments of a day split around unusable data) is used as a sequence and the Baum-Welch algorithm is applied across all sequences to maximise the joint probability distribution in Equation (4.9) by finding the most likely parameters for the Gaussian emission distributions and transition matrix [8].

Once the parameters are inferred, the Viterbi algorithm is applied to find the most likely heating state for each time step of each day [8]. All the identified sequences of at least two hours of consecutive heating are extracted as a heating-on period. The first hour of each heating-on period is ignored to allow the house time to reach thermal equilibrium, and the average of the remaining power and external temperature

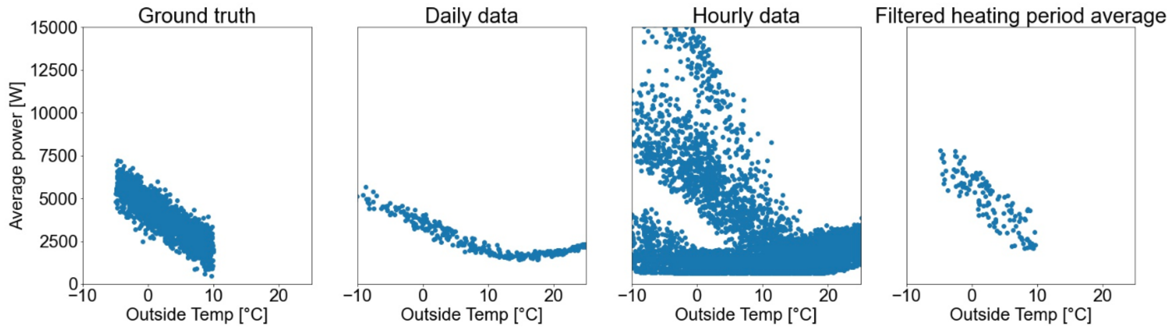

Figure 4.3: The data used to infer the ground truth is presented alongside the daily average energy usage, hourly energy usage, and the average energy usage of each heating period. The daily data has little variance as the energy usage is averaged across the day, however, the gradient does not accurately reflect the HPLC as it is influenced by the duration that the heating is on. The hourly data is noisy due to variance of energy usage between hours and different modes of energy usage (heating off, heating on and stable internal temperature, and heating the building from a colder state). Using the average energy usage for each heating period ensures that the correct data is collected for inferring the HPLC.

recordings in the heating-on period are used to infer the HPLC.

Notice that as part of the steady-state assumption, it is assumed that the internal temperature is kept constant, which is true for the heating-on periods after the first hour is omitted, however, the exact value is unknown. Under this assumption, the actual value of the internal temperature does not particularly matter as it is a fixed value that will only provide a constant shift, assuming the home is always heated to the same temperature. The value of interest, the HPLC, is not affected by this as it is inferred through the gradient of the heating power against the external temperature. To demonstrate the importance of exclusively considering heating-on periods, Figure 4.3 shows how taking daily data leads to a gradient (which represents the HPLC) that is different from the ground truth, whereas hourly data is noisy and splits into different usage modes. This plot also shows that it is important to remove the first hour of heating data as the first hour corresponds to a higher energy usage as the building is not in a state of thermal equilibrium.

4.2.3 Inference of the HPLC

After identifying the heating-on periods for a home, the average power and external temperature for each period are found. The process of finding these periods is outlined in Figure 4.2. By considering Equation (4.7), it shows that the rate the average power used to heat the house changes with the external temperature gives the HPLC if the other terms are accounted for.

It is assumed that there is the same number of occupants in the home for all of the heating periods, this causes a constant shift in the energy usage, and as the gradient is the value of interest it can be ignored. Note that across the data used in this study the average power supplied to heat a home is 6.5kW, the contribution of heat energy per occupant is $< 1\%$. Similarly, power for other purposes has been shown to not correlate with the external temperature and can be treated as noise [28]. Furthermore, it is assumed that there is no behavioural ventilation loss during heating-on periods as they are assumed to be identified by the outlier filter. Finally, the solar filter ensures there is only a minimal contribution from the solar irradiance and this is also absorbed into a noise term, meaning the relation between the heating power and external temperature can be expressed as

$$P_h = -HPLC\ T_{ex} + C_p + \xi_h. \tag{4.10}$$

To infer the HPLC, linear regression is used to find the expected value of the HPLC, and the variance is calculated from the external temperature and heating power data points. Note the negative sign in front of the HPLC as the external temperature is used instead of the difference between the internal and external temperature and C_p represents the constant contribution of other sources of heat energy.

Whilst the HPLC is the metric of interest, for comparisons against existing approaches the HLC is required. This is calculated from the HPLC by multiplying it by the standard heating system efficiency, η_h, of 0.85, which is the mean boiler efficiency found in a Department of Energy and Climate Change study [98].

4.3 Dataset

To conduct this experiment the ideal dataset consists of homes where the HLC has been inferred via a co-heating test on a vacated house as the ground truth, the heating system efficiency is known to infer the HPLC, half-hourly smart-meter recordings for gas and electricity usage for at least a year whilst the home is occupied, and half-hourly weather data, particularly external temperature and solar irradiance. This presents a major difficulty for obtaining a large-scale dataset as each home in the dataset requires an expensive, intrusive, and time-consuming co-heating test prior to being occupied and monitored for a year.

To address this challenge, the dataset selected for the experiments is a subset of homes from the solid wall insulation field trial originally conducted by the UK Energy Saving Trust and then processed and anonymised by another study [102, 96, 19, 117]. This data was collected with the intent to evaluate the effectiveness and experience of customers who decided to install solid wall insulation, however, the study recorded the data required for the experiments. For each home, the HLC was measured after the installation of the solid wall insulation along with half-hourly smart-meter readings of gas and electricity usage for an extended time period after. Unfortunately, the HPLC or the heating system efficiency were not reported hence the HLC must be used to evaluate the performance of the SMITE method. Furthermore, the relevant weather data for each home is available as an anonymised location of each home was provided and the data could be collected from the nearest weather station provided by the Met Office [94].

To ensure that each home in the subset has the data required to infer the HLC it must have enough days recorded after filtering out the corrupted data and be within 15km of the nearest weather station in the Met Office dataset. From the solid wall dataset, 14 homes have sufficient data and, when combined with the Met Office dataset, 7 of these homes are within the required proximity of a weather station. The 7 homes fitting the requirements use gas heating. Whilst the experiments show the feasibility of the SMITE method on the available data, there is a pressing need for publicly available large-scale datasets to enable research into evaluating the thermal performance of homes.

4.4 Experiments

To evaluate the SMITE method it is implemented on the 7 homes selected from the solid wall insulation field trial dataset and compared to the ground truth HLC calculated by a co-heating test [102, 96, 19, 117]. For comparison, the Deconstruct method is also implemented and the two are compared across a number of metrics [29].

4.4.1 Experimental Setup

The SMITE method is implemented as described in Section 4.2 and the co-heating inferred HLC values are used as the ground truth. The standard deviation values for the co-heating test are not reported and are hence inferred based on the known uncertainty in co-heating HLC measurements [69].

For the Deconstruct, method the only parameter changed is the solar irradiance filter from 0.05 kW/m^2 to 0.15 kW/m^2 as this minimised the MAPE of the Deconstruct method on this dataset. The SMITE method uses the same filters as Deconstruct to find the viable days. Then the heating-on periods are found using the proposed detection algorithm and the filters are re-applied.

As all the 7 homes in the dataset have gas heating, the experiments are set up to show that the approach works when metering of the gas usage is recorded. Another scenario is constructed, adding the gas and electricity smart-meter readings together, as the combined meter readings are what would be expected when a home only has electric utilities and no supply of gas. The purpose of this is to demonstrate that the SMITE method works in a scenario similar to what is expected in a home that has electric heating.

As previously discussed, the HPLC is the HLC multiplied by the heating system efficiency. Hence without any insight into each house, an assumption has to be made on the heating system efficiency to infer the HLC from the HPLC.

4.4.2 Evaluation Metrics

To evaluate the performance of the SMITE method a selection of metrics are used to demonstrate how accurately it can infer the HLC of a home. A number of metrics are implemented to compare the approaches. The mean absolute percentage error (MAPE) calculates the absolute error as a percentage of the real HLC, measuring how accurate the predictions are and is given by,

$$\mathrm{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\mathrm{H}_{real}^{(n)} - \mathrm{H}_{pred}^{(n)}}{\mathrm{H}_{real}^{(n)}}\right|,$$

where the HLC value inferred through the co-heating test for house n is $H_{real}^{(n)}$, and $H_{pred}^{(n)}$ refers to the HLC value inferred through the SMITE or Deconstruct method. The Z-score is used to measure how the uncertainty of the prediction captures the ground truth, measuring how many standard deviations away from the ground truth HLC is from the predicted HLC. The σ term denotes the standard deviation inferred through linear regression for home n with the SMITE or Deconstruct method,

$$\text{Z-score} = \frac{H_{pred}^{(n)} - H_{real}^{(n)}}{\sigma_{H_{pred}}^{(n)}}.$$

Also, Pearson's correlation coefficient (PCC) is used to measure how the predicted values correlate with the ground truth,

$$\rho\left(\mathbf{H}_{real}, \mathbf{H}_{pred}\right) = \frac{\operatorname{cov}\left(\mathbf{H}_{real}, \mathbf{H}_{pred}\right)}{\sigma_{\mathbf{H}_{real}} \sigma_{\mathbf{H}_{pred}}},$$

where The bold $\mathbf{H}_{real}$ and $\mathbf{H}_{pred}$ represent the vector of all the HLC values inferred via the method in question. whilst the σ with a bold subscript represents the standard deviation across the vector of HLC values. The 'cov' function finds the covariance between the two vectors.

4.4.3 Evaluation

The SMITE method shows a significant performance improvement compared to the Deconstruct method when using gas smart-meter data and the combined gas and electricity smart-meter readings to infer the HLC as shown in Table 4.2. The accuracy improvement across all settings makes the SMITE method the state-of-the-art non-intrusive approach. Additionally, It can be seen in Table 4.1 that the SMITE method also has a larger variance than the Deconstruct method. This is expected as the SMITE method uses the average of shorter periods of data and whilst there is more variance in the data it removes a modeling bias that is present in the Deconstruct approach. Furthermore, Figure 4.4 and Figure 4.5 present the results in the form of a bar chart to allow for a visual comparison of the results.

			Gas meter readings				Gas and electric combined meter readings			
	Co-heating		Deconstruct		SMITE		Deconstruct		SMITE	
HLC [W/K]	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
house 1	304	10.1	252	28.0	264	45.4	271	41.1	186	43.5
house 2	223	6.3	244	14.6	248	31.2	258	14.7	235	34.2
house 3	198	11.7	501	33.8	177	59.7	517	35.4	195	64.9
house 4	351	6.3	436	40.2	387	42.6	433	63.8	524	47.2
house 5	214	7.1	235	30.8	211	45.1	302	37.9	239	71.4
house 6	189	7.4	209	37.8	148	51.2	205	40.5	218	70.9
house 7	189	6.6	195	31.0	219	64.2	206	32.3	240	90.1

Table 4.1: The mean and standard deviation of the HLC values inferred from the Deconstruct and SMITE approaches are presented alongside the HLC values inferred for each home from the co-heating tests (ground truth). The standard deviation was not provided for the co-heating tests, instead, it has been inferred based on the known uncertainty of co-heating tests [69].

First, consider the setting where gas smart-meter readings are used. There is a significant gain in accuracy as the MAPE decreases from 32.6% for the Deconstruct method to 12.0% for the SMITE method. This performance increase should be expected as by identifying the heating-on periods the data used to infer the HLC is more similar to the data collected in a co-heating test. Furthermore, the reduction in Z-score when moving from the Deconstruct method to the SMITE method shows that the predicted values and the uncertainty inferred through the SMITE method better capture the true value. For the SMITE method, the ground truth HLC is within 1 standard deviation of the inferred HLC for all homes. The consistency of the accurate predictions is valuable in this setting as incorrect predictions may result in costly interventions for the wrong homes. On the point of reducing cost, the SMITE method is shown to have a strong correlation with the intrusive co-heating test, with a PCC of 0.91 compared to a PCC of 0.35 for the Deconstruct method, meaning the SMITE method is a viable replacement for a co-heating test as it achieves similar results whilst bringing the benefits of being non-intrusive.

	Gas heating				Gas and electric combined meter readings			
	Deconstruct		SMITE		Deconstruct		SMITE	
	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score
house 1	-17.1	-1.86	-13.4	-0.90	-11.0	-0.81	-38.9	-2.72
house 2	9.4	1.44	11.4	0.81	15.6	2.38	5.5	0.36
house 3	53.0	8.97	-10.5	-0.35	160.8	9.00	-1.7	-0.05
house 4	24.3	2.12	10.2	0.84	23.2	1.28	49.2	3.66
house 5	10.0	0.69	-1.3	-0.06	41.4	2.34	11.6	0.35
house 6	10.9	0.55	-21.7	-0.80	9.0	0.42	15.5	0.41
house 7	3.2	0.20	15.7	0.46	8.8	0.52	26.8	0.56
Average	27.7	1.73	**-1.3**	**0.00**	35.4	2.16	**9.7**	**0.37**
Average of absolute values	32.6	2.26	**12.0**	**0.60**	38.9	2.39	**21.3**	**1.15**

Table 4.2: The table shows the percentage difference between the non-intrusive methods and the ground truth, and the Z-score gives how many standard deviations away the ground truth is from the mean of the inferred value for each non-intrusive method. The average of absolute values gives the average magnitude of the percentage differences (MAPE) and Z-scores for each method in each setting. The SMITE approach shows a significant improvement across all measurements. The average without taking the absolute of each value shows that there is less bias in the SMITE method, whilst the Deconstruct method over-predicts.

In the setting with combined energy recording, a slight drop in performance is expected, as there are two sources of energy usage combined making it more difficult to separate the energy used for heating. There is again a significant improvement in accuracy when moving from the Deconstruct method to the SMITE method with a MAPE of 38.9% and 21.3% respectively. There is also a similar decrease in Z-score, from 2.39 for Deconstruct to 1.15 for SMITE, and the PCC increases from 0.31 for the Deconstruct method to 0.70 for the SMITE method. This shows that the SMITE method improves on the current approaches to infer HLC non-intrusively in all settings and as algorithms to disaggregate heating energy usage from other uses of energy improve it is expected that the performance can be in line with what is achieved in the gas-only setting.

Considering the average, rather than the average of absolute values, highlights how

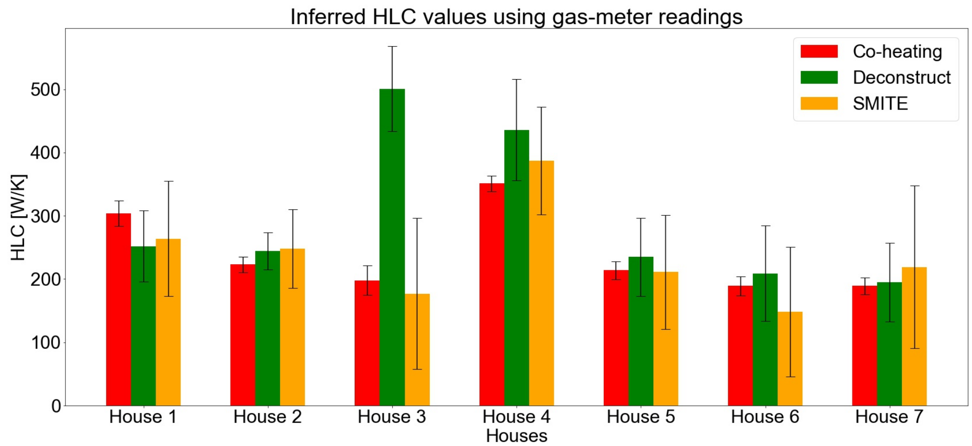

Figure 4.4: The inferred HLC values from the gas meter readings for the Deconstruct and SMITE methods compared to the co-heating test. The error bars represent a 95% confidence interval.

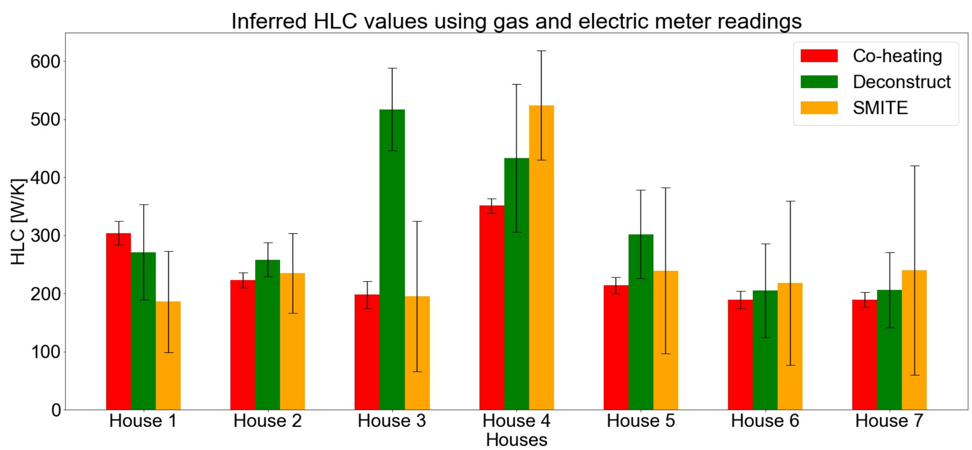

Figure 4.5: The inferred HLC values from the combined gas and electric meter readings for the Deconstruct and SMITE methods compared to the co-heating test. The error bars represent a 95% confidence interval.

the SMITE method removes a modeling error and source of bias, by using half-hourly rather than daily data. The average percentage difference is 27.7% for the Deconstruct method compared to -1.3% for the SMITE method. This is most likely caused by the fact that using daily data rather than half-hourly confounds the duration of heating with the HLC, an issue discussed further in Section 4.5.3, causing an over-estimate of the HLC. By removing this source of bias with the SMITE approach the predicted HLC values are more accurate and inferences made over large samples of homes are more likely to be accurate compared to the Deconstruct method.

The results of individual houses are compared to understand how switching to the combined gas and electricity smart-meter readings affects the value of HLC inferred by SMITE. For houses 1, 4, and 7 there is a large performance drop which can partly be attributed to all three houses having high electricity usage of other appliances. By improving heating energy disaggregation, this source of inaccuracy can be reduced. Considering the SMITE method for house 3 where the HLC is inferred from gas data two potential issues arise: (i) the worse MAPE when in the setting with the combined gas and electricity usage data indicates there may be a source of electric heating as well as gas heating in this house, which is currently unaccounted for; (ii) The gas data has systematic noise, indicating the presence of another system using gas energy, perhaps with improved disaggregation, the performance on this house could be improved.

To compare SMITE to Deconstruct, the differences between the methods can be seen by looking at individual houses. For example, on house 3 Deconstruct fails to infer the HLC due to the fact that the heating is on for a duration three times longer compared to the other houses. The fact that Deconstruct fails to separate the heating duration from the inferred HLC value leads to a large over-estimate, which the SMITE method handles correctly by identifying the heating periods and only using this data. However, the performance drop is lesser for Deconstruct when shifting from the gas meter readings to the gas and electricity meter readings, although it is still worse. This is due to the fact that Deconstruct averages energy usage across the day, meaning there is less variation in the data caused by energy usage for non-heating appliances as it is averaged across the day. This is particularly visible in Houses 1 and 4 where the usage of electricity causes a larger performance drop for SMITE compared to Deconstruct. However, with better energy disaggregation methods, there is significant room for the SMITE method to improve in this scenario and potentially reach the accuracy seen with just the gas meter readings.

4.5 Discussion

Following the successful experimental results, three concepts are discussed: (i) The merits of using HPLC instead of HLC and what is the most appropriate metric for evaluating the thermal performance of a home given the limitation of a non-intrusive setting; (ii) detecting when heating is used and what further insights into the home this can give to guide energy usage related policy; and (iii) the shortcomings of approaches using daily energy usage and how considering only daily data causes multiple factors to be mixed up.

4.5.1 Non-Intrusive Thermal Performance Metrics

The metric used for comparisons across methods is HLC as this was the only relevant metric labeled in the dataset used. To infer the HLC it requires an assumption to be made about the heating system efficiency. However, a more appropriate metric may be HPLC as it alleviates the requirement of this assumption and measures the compounded value of HLC and heating system efficiency instead, resulting in a more accurate evaluation when performing non-intrusive inference of the thermal performance of homes.

In the absence of sensors recording real-time data in the home, such as smart meters or smart thermostats, measuring the HLC of a house with an intrusive thermal performance measurement is required - such as a co-heating test. When the HLC is measured, specific procedures can be taken to ensure that the heating system efficiency does not influence the measurement of thermal performance. However, the advent of smart meters creates a significant paradigm shift in how the thermal properties of a house can be measured, and hence it may make sense to change the metric that is used. The HPLC may be a more appropriate metric to track in this new data-rich scenario. Whilst the HPLC compounds the HLC with heat loss through ventilation and heating system efficiency, it can be inferred through cheap non-intrusive techniques, which means it can be deployed at a scale that is not feasible for HLC.

Finding the HPLC instead of HLC could identify which homes have the lowest thermal performance and where the biggest impact from interventions could be made. In some cases, it may be more important to replace a heating system than to provide insulation. As any intervention would require engagement with the homeowner, the heating system efficiency can be measured, and then the HLC can be inferred with this new information.

4.5.2 Detection of Periods when Heating is On

In the process of finding the HLC, the SMITE method detects periods when the heating is on. When evaluating which homes require intervention, this information may also prove useful as studies have shown that presenting consumers with insights into their energy behaviour can have a significant impact on their energy usage [39, 139]. Homes with a high HPLC and extended heating periods may be identified as those that could see the largest reduction in their energy bill by taking energy-saving interventions, thus making such interventions more financially viable. Furthermore, as mentioned in Section 4.2.1.3, this method can be extended to detect periods when the heating system is at full capacity for extended periods of time, which can suggest that thermal

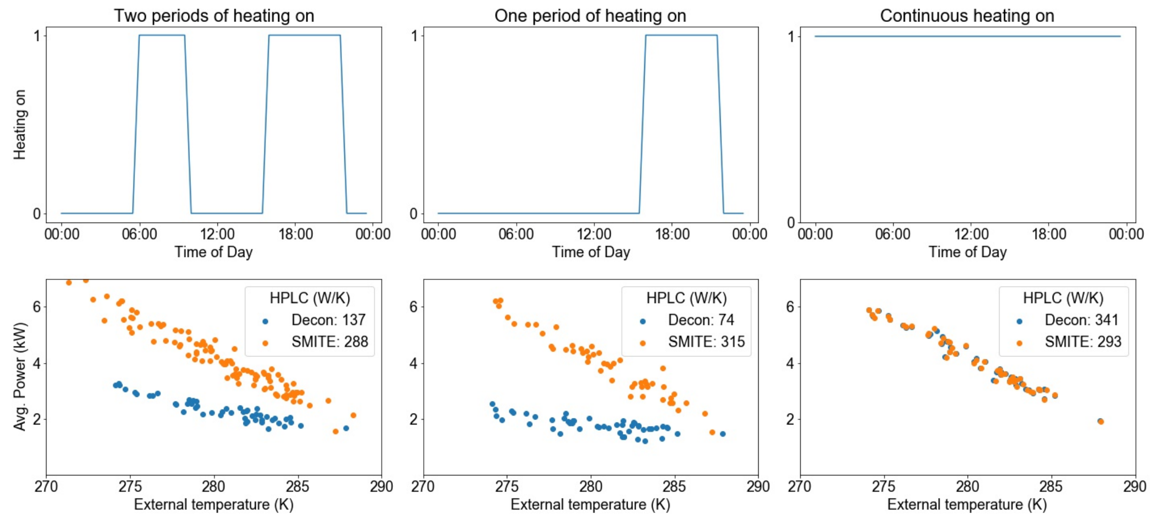

Figure 4.6: The plots demonstrate how the length of time heating is on affects the HPLC inferred with the daily approach. The top row shows their respective heating patterns for a 24 hour period whilst the bottom row shows the average power and external temperature for each period recorded by each method. The inferred HPLC remains consistent for the half-hour approach, However, the daily approach confounds heating usage behaviour with the home's thermal performance. Note that the HPLC for the Deconstruct approach is scaled by the factor F_T, introduced in equation (2.9), which accounts for the internal temperature depending on the external temperature when the heating is off. Hence, when the heating is always on it over-predicts, despite the gradient matching that of the SMITE method.

equilibrium is never reached which may identify inadequate heating systems. Whilst the proposed metrics alone will not provide enough information to make decisions, they can be a valuable tool when combined with wider socioeconomic factors. One such case is that if other factors indicate a household is at high risk of fuel poverty, these methods can be used to ensure that the energy company obligation is adhered to and that households in fuel poverty are assisted in taking energy-reducing measures.

4.5.3 Issues with Daily Energy Usage

A major issue with the daily approach is that it confounds the length of time the heating is on with the thermal performance of the home (HPLC). Generally speaking, the heating is on for longer intervals when the weather is colder. As expected, if the heating is on for longer periods of the day, the average heating energy usage for the day is greater and if the length of time the heating is on is correlated with the external weather, this has an effect on the inferred HPLC value. By not separating the length of time the heating is on and the thermal performance, the predicted HPLC will vary with the time heating is on, which should not be affecting the HPLC. This is a shortcoming addressed by taking the half-hourly smart-meter readings in the SMITE method.

To illustrate the point, a synthetic dataset is created. External temperature values are taken from the Met Office dataset and the energy usage when heating is on is generated by using Equation (4.10). The plots in Figure 4.6 demonstrate the point, as the heating schedule changes, the daily approach infers varying HPLC values, whilst the half-hourly approach accurately infers the HPLC consistently.

4.6 Summary

This chapter proposes SMITE, a non-intrusive method to evaluate the thermal performance of a residential building using only smart-meter readings and local weather data. SMITE performs in line with a co-heating test whilst being non-intrusive, presenting the first method to infer the HLC and HPLC accurately whilst being scalable. This enables insights into which buildings can benefit from thermal performance interventions to reduce the amount of energy used by residential houses.

The methods developed demonstrate that it is possible to evaluate the thermal performance non-intrusively. The next chapter will focus on expanding the number of scenarios in which this approach can work, particularly for residential cooling, improve the model to capture uncertainty, and identify which scenarios the model will not work in and will require more data.

5

Identifying Scenarios for Accurate Inference of Thermal Performance

The previous chapter demonstrates that it is possible to infer the thermal performance non-intrusively. However, theoretical measurements of thermal performance can fail to reflect real-world performance, and SMITE does not quantify when it fails to infer the thermal performance accurately, which needs to be identified [76]. Furthermore, the previous chapter only demonstrates an approach for inferring the thermal performance when heating a building, not when cooling.

The above issues are addressed in this chapter with the proposal of QUILT, which builds upon the approach of Chapter 4 to deliver a non-intrusive, data-driven framework to quantify the thermal performance of a home, grounded in physical models of how heat flows through a building. A Bayesian model of the heat flows in a building is used to infer loss coefficients (such as the HPLC), and to account for the uncertainty associated with data sources and parameters in the thermal model, and known distributions of parameters can be used as priors to inform the model. The model can also be adapted when a cooling system is used and can infer the cooling power loss coefficient (CPLC), which is a cooling equivalent of HPLC. This allows the approach to be deployed in more scenarios.

Furthermore, simulated data is generated with EnergyPlus using templates from the U.S. Department of Energy's Building Energy Codes Program[1] to create a broader range of scenarios to test QUILT beyond the limited real-world data. In particular, as there are no actual labeled CPLC datasets, the simulated data can be used to validate the feasibility of inferring the CPLC. The experiments across different scenarios highlight a number of key findings, in particular: 1) the importance of only using the data-points when the heating or cooling system is on; 2) holidays provide periods of low-noise data which allow for the HPLC or CPLC to be inferred with a week of data in the right external conditions, and; 3) as heating is such a significant source of energy use, energy disaggregation does not have a particularly large effect on accuracy when inferring HPLC or CPLC. By combining improved accuracy with knowledge of when the approach works and fails, whilst being non-intrusive and cheap to deploy, QUILT is a step towards an effective and general method that can be deployed in the real world.

The rest of the chapter is structured as follows. Section 5.1 introduces the thermal model, and Section 5.2 extends it to introduce the priors of the parameters in the model and the modeling choices that can be made. In Section 5.3 the algorithm used to infer the posterior parameter values is outlined. Section 5.4 outlines the data that the experiments

[1] `https://www.energycodes.gov/development/residential/iecc_models`

are performed on and Section 5.5 outlines the experimental set-up and results. Finally, Section 5.6 presents actionable insights, before the concluding remarks in Section 5.7.

5.1 Thermal Models

As in the thermal model in Chapter 4, the thermal performance of a building can be characterised by the HLC [kW/K], however, the actual heat energy in a house is not recorded, so it can be difficult to infer the HLC. Instead, the closest recorded proxy to the heat energy is the energy supplied to the heating system, which is measured by smart meters. Hence, instead of inferring the HLC we focus on inferring the heating power loss coefficient, HPLC [kW/K], and use this to characterise the steady-state thermal performance.

Furthermore, the cooling power loss coefficient, CPLC [kW/K], is introduced, which is the equivalent of the HPLC but for cooling systems. It is defined as,

$$P_{c_t} = CPLC\,(T_{ex_t} - T_{in_t}), \tag{5.1}$$

where P_{c_t} is the power supplied to the cooling system for the given time-step. This assumes an ideal scenario where the only heat flow is due to the cooling system and through the fabric of the building given the difference between the internal and external temperature. To infer CPLC and HPLC, a thermal model of the home needs to be constructed which accounts for the heat flows of the home beyond this ideal scenario.

5.1.1 Heat Flow Model

The heat flow model follows from Equation (4.1) in the previous chapter. However, in that equation, there are flows of heat energy that could not be accurately accounted for. One such term is the heat flow due to ventilation, Q_{ve}, as there is no way to measure the ventilation heat flow with non-intrusive methods. It is similarly the case for the impact of weather factors on the heat flow that can not be accurately modeled such as wind and precipitation [28, 62]. To address this, a noise term, ξ_Q, is introduced to account for these unmeasured sources of heat transfer, and the heat flow equation is now expressed as follows:

$$Q_h = Q_{fa} + Q_{we} + Q_o + Q_b + Q_{\Delta t} + \xi_Q. \tag{5.2}$$

5.1.2 Heating Model

The heat flow into the building from the heating system is given as the power supplied to the heating system, P_{h_t}, scaled by the efficiency of the heating system, η_h,

$$Q_{h_t} = \eta_h P_{h_t}. \tag{5.3}$$

Hence, the thermal equilibrium heat flow model is expressed as:

$$\begin{aligned} \eta_h P_{h_t} = HLC\,(T_{in_t} - T_{ext_t}) + 0.06 N_{occ_t} + \eta_b P_{b_t} \\ + A_{sol_t} I_{sol_t} + f(T_{in_t}, T_{in_{t-1}}, M_T) + \xi_Q. \end{aligned} \tag{5.4}$$

However, if the heating system efficiency is unknown this creates a scenario where each factor might be out by a scaling amount. To address this, we can divide by the heating system efficiency, η_h, and instead learn the ratio of each unknown variable to the heating system efficiency. As previously defined, this gives us the HPLC instead,

$$\begin{aligned} P_{h_t} = HPLC\,(T_{in_t} - T_{ex_t}) + 0.06\hat{N}_{occ_t} + \hat{\eta}_b P_{b_t} \\ + \hat{A}_{sol_t} I_{sol_t} + \hat{f}(T_{in_t}, T_{in_{t-1}}, M_T) + \hat{\xi}_Q. \end{aligned} \tag{5.5}$$

5.1.3 Cooling Model

Similarly, if the external temperature is greater than the internal temperature, the flow of energy to cool the building from the cooling system is given as the power supplied to the cooling system, P_c, scaled by the cooling system efficiency, η_c,

$$Q_c = \eta_c P_c. \tag{5.6}$$

Then to construct the thermal equilibrium heat flow model from Equation (5.2), Q_h is replaced with Q_c to represent the cooling situation we are now in. Then the equation is similar to Equation (5.4), except the flow of heat due to the difference in internal and external temperature is inwards rather than outwards, causing the sign to change so we have $(T_{ex_t} - T_{in_t})$ instead of $(T_{in_t} - T_{ex_t})$ also the signs of the other terms are turned negative as the energy provided by the cooling system is now being used to cool rather than heat,

$$\begin{aligned} \eta_c P_{c_t} = HLC\,(T_{ex_t} - T_{in_t}) - 0.06N_{occ_t} - \eta_b P_{b_t} \\ - A_{sol_t} I_{sol_t} - f(T_{in_t}, T_{in_{t-1}}, M_T) - \xi_Q. \end{aligned} \tag{5.7}$$

Similar to the scenario with heating, the terms are re-scaled by the cooling system efficiency, η_c, and the ratio of each variable is instead learned, this gives us a scenario where the CPLC can be inferred,

$$P_{c_t} = CPLC\,(T_{ext_t} - T_{in_t}) + 0.06\tilde{N}_{occ_t} - \tilde{\eta}_b P_{b_t} \\ - \tilde{A}_{sol_t} I_{sol_t} - \tilde{f}(T_{in_t}, T_{in_{t-1}}, M_T) - \tilde{\xi}_Q. \tag{5.8}$$

5.2 Bayesian Thermal Model

The theoretical model of the heat flow is constructed, as in Equation (5.5) for heating, and cooling in Equation (5.8). However, in practice some terms are not fully observed and may have to be inferred, which introduces sources of uncertainty and inaccuracy. As such, the model needs to reflect the practicalities of gathering the data required to infer the HPLC or CPLC and the uncertainty and inaccuracy in different scenarios needs to be quantified. It is assumed that all parameters are independently distributed.

5.2.1 Fixed Internal Temperature

When the heating is on and it is sufficiently cold outside, it is reasonable to assume that there is a significant difference between the internal temperature T_{in_t} and the external temperature T_{ext_t}, such that the heat flow through the fabric of the building, Q_{fa}, is large. This provides a scenario where it is easier to infer the HPLC. Furthermore, when the heating is on the internal temperature is relatively fixed, and the net heat flow into or out of the building from the previous time-step, $Q_{\Delta T}$, can be absorbed into the noise term, ξ_Q.

Hence, there are two ways in which modeling error can occur for the internal temperature, T_{in}, errors in detecting when the heating is on, and errors in identifying the temperature that the thermostat is set to. If a smart thermostat is installed in the home, the internal temperature can be directly observed, t_{in}, with small measurement noise. A smart thermostat would also provide a clear indication of when the heating is on. However, if a smart thermostat is not present, the internal temperature must be inferred and a model must be used to infer when the heating is on. A uniform distribution across expected thermostat temperatures can be used as a prior for the internal temperature, and it can be assumed that this value will be the same for all heating periods, thus in summary,

$$T_{in} \sim \begin{cases} \mathcal{N}(t_{in}, 0.5) & \text{if } t_{in} \text{ observed,} \\ U(19, 23) & \text{otherwise.} \end{cases} \tag{5.9}$$

5.2.2 External Temperature and Weather

With regards to the external temperature, it is assumed that it is known and the only uncertainty required to account for is the discrepancy between the temperature recording at the nearest weather station, t_{ex}, and the home,

$$T_{ex} \sim \mathcal{N}(t_{ex}, 0.5). \tag{5.10}$$

Similarly, it can be assumed that that solar irradiance is observed with a small amount of measurement uncertainty due to discrepancy between weather station recordings of solar irradiance,i_{sol}, and the solar irradiance incident on the house. A gamma distribution is used to avoid negative values,

$$I_{sol} \sim \text{gamma}(\mu = i_{sol}, \sigma = 10). \tag{5.11}$$

The contribution of heat flow from solar irradiance depends on how much is incident on the solar aperture of the home. The solar aperture for a building can not be larger than the size of the sun-facing walls, however, it also depends on the wall material, size of windows, and the structure of the building, hence the prior has a large amount of uncertainty [49, 84]. A gamma distribution is also used to represent this to avoid negative values,

$$A_{sol} \sim \text{gamma}(\mu = 10, \sigma = 10). \tag{5.12}$$

5.2.3 Power

If there is separate metering for heating and cooling power, then they are observed as p_h or p_c, respectively, alongside the power used for other appliances, p_b, and only measurement noise needs to be considered. However, if this is not the case, they need to be disaggregated and inferred as $\hat{p}_h$ or $\hat{p}_c$, and $\hat{p}_b$. The errors in the disaggregation process can propagate through and affect the inferred value of the HPLC,

$$P_h \sim \begin{cases} \mathcal{N}(p_h, \sigma_{p_h}) & \text{if } p_h \text{ observed,} \\ \mathcal{N}(\hat{p}_h, \sigma_{\hat{p}_h}) & \text{otherwise.} \end{cases} \tag{5.13}$$

$$P_c \sim \begin{cases} \mathcal{N}(p_c, \sigma_{p_c}) & \text{if } p_c \text{ observed,} \\ \mathcal{N}(\hat{p}_c, \sigma_{\hat{p}_c}) & \text{otherwise.} \end{cases} \tag{5.14}$$

$$P_b \sim \begin{cases} \mathcal{N}(p_b, \sigma_{p_b}) & \text{if } p_b \text{ observed,} \\ \mathcal{N}(\hat{p}_b, \sigma_{\hat{p}_b}) & \text{otherwise.} \end{cases} \tag{5.15}$$

Where the prior over the variance is given as,

$$\begin{aligned}
\sigma_{p_h} &\sim \left|\mathcal{N}(0,500)\right|, \\
\sigma_{p_c} &\sim \left|\mathcal{N}(0,500)\right|, \\
\sigma_{p_b} &\sim \left|\mathcal{N}(0,500)\right|, \\
\sigma_{\hat{p}_h} &\sim \left|\mathcal{N}(0,1000)\right|, \\
\sigma_{\hat{p}_c} &\sim \left|\mathcal{N}(0,1000)\right|, \\
\sigma_{\hat{p}_b} &\sim \left|\mathcal{N}(0,1000)\right|.
\end{aligned} \tag{5.16}$$

5.2.4 HPLC & CPLC

The HPLC and CPLC are unobserved parameters that must be inferred to identify the thermal performance of a home. The prior for the HPLC is based on the distribution of HPLC values from a previous study of 541 homes [29]. The CPLC is a newly introduced term so we base the prior on the fact that the buildings will have the same properties, however, cooling systems are typically more efficient,

$$\begin{aligned}
HPLC_\mu &\sim \mathcal{N}(400,200), \\
HPLC_\sigma &\sim \left|\mathcal{N}(0,300)\right|, \\
HPLC &\sim \mathcal{N}(HPLC_\mu, HPLC_\sigma).
\end{aligned} \tag{5.17}$$

$$\begin{aligned}
CPLC_\mu &\sim \mathcal{N}(350,200), \\
CPLC_\sigma &\sim \left|\mathcal{N}(0,300)\right|, \\
CPLC &\sim \mathcal{N}(CPLC_\mu, CPLC_\sigma).
\end{aligned} \tag{5.18}$$

Scenario	hardware	Observed	Unobserved
Daily data	smart meter (daily readings)	T_{ex}, I_{sol}, P_{all}	HPLC, T_{in}, A_{sol}, η_b
Hourly data	smart meter	T_{ex}, I_{sol}, P_{all}	HPLC, T_{in}, A_{sol}, η_b
Week holiday data	smart meter	T_{ex}, I_{sol}, P_h, P_b	T_{in} HPLC, A_{sol}, η_b
Filtered hourly data	smart meter	T_{ex}, I_{sol}, P_{all}	HPLC, T_{in}, A_{sol}, η_b
Heating-on data	smart meter, smart thermostat	T_{ex}, I_{sol}, P_{all}	T_{in} HPLC, A_{sol}, η_b
averaged filtered heating period data	smart meter, smart thermostat	T_{ex}, I_{sol}, P_{all}	P_h, P_b, T_{in} HPLC, A_{sol}, η_b
averaged filtered heating period data & heating energy known	smart meter, smart thermostat, energy disaggregation	T_{ex}, I_{sol}, P_{all}, T_{in}	P_h, P_b HPLC, A_{sol}, η_b

Table 5.1: The scenarios of data procurement created with the EnergyPlus simulator, and the corresponding variables which are observed or unobserved in the thermal models presented in Equation (5.5) and Equation (5.8).

Finally, to account for uncertainty in the model there is a general noise term,

$$\xi_Q \sim \mathcal{N}(0, 300). \tag{5.19}$$

5.3 Algorithm

To infer the HPLC or CPLC, filters are applied to identify data-points where the thermal contributions from hard-to-model terms are small, and then Bayesian inference is performed on the probabilistic model constructed in Section 5.2 to find the posterior distributions of the HPLC or CPLC.

5.3.1 Data Selection

If there is enough data it is possible to remove the data-points where there are high contributions from difficult to model sources of heat flow [29, 24]. If the solar irradiance is large, it can create a source of uncertainty as the linear model used can fail to capture the true dynamics of how it converts to heat energy, hence data-points with solar irradiance over 0.15 kW/m^2 can be removed [61]. Furthermore, if the external temperature is too high the heating will typically not be on, hence data-points with external temperatures above 15°C can be removed. Also, if the external temperature is too cold the heating system may run at maximum capacity without maintaining thermal equilibrium, hence external temperatures below 0°C are filtered out.

The reverse arguments can be made with regard to external temperature for the CPLC. If the external temperature is too low the difference between the internal and external temperature will not be large enough, hence data-points with an external temperature below 25°C are filtered out. And if the external temperature is too high, the cooling system may run at maximum capacity without maintaining thermal equilibrium, hence external temperatures above 35°C are filtered out. As a high external temperature is correlated with high solar irradiance, this typically relies on the cooling system running at night and on occasional days when there is high external temperature and low solar irradiance.

5.3.2 Inference

Bayesian inference is used to find the posterior distribution of the parameters in Equation (5.5) for heating, and in Equation (5.8) for cooling. The prior distributions for the parameters and the data source measurement uncertainty are introduced in Section 5.2, and they are used to inform the posterior distributions alongside the data observed in each scenario, which is outlined in Table 5.1. To generate the posterior distribution of the unknown variables, in particular, the HPLC and CPLC, a Hamiltonian Monte Carlo (HMC) algorithm is used to propose samples that are distributed according to the target distribution, where the samples are generated using the No-U-Turn Sampler (NUTS) to improve the efficiency of the HMC [60]. The construction of the Bayesian model is outlined in Algorithm 3, and it has been implemented using PyMC3 [110].

Algorithm 3: Inferring HPLC with QUILT. The CPLC can be inferred with minor adjustments to align the algorithm with Equation (5.8)

input : T_{ex}, i_{sol}, $((p_h, p_b) \vee (\hat{p}_h, \hat{p}_b))$, optional: t_{in}

$T_{ex} \sim \mathcal{N}(t_{ex}, 0.5)$

$I_{sol} \sim \text{gamma}(\mu = i_{sol}, \sigma = 10)$

$HPLC_{\mu} \sim \mathcal{N}(400, 200)$

$HPLC_{\sigma} \sim |\mathcal{N}(0, 300)|$

$HPLC \sim \mathcal{N}(HPLC_{\mu}, HPLC_{\sigma})$

$A_{sol} \sim \text{gamma}(\mu = 10, \sigma = 10)$

$\xi_Q \sim |\mathcal{N}(0, 500)|$

`// Assign priors based on if internal temperature observed`

if t_{in} *observed* **then**

  $T_{in} \sim \mathcal{N}(t_{in}, 0.5, \text{Observed} = t_{in})$

else

  $T_{in} \sim \mathcal{U}(19, 23)$

end

`// Assign priors based on if heating energy disaggregated`

if p_h *observed* **then**

  $p_{h_\sigma} \sim |\mathcal{N}(0, 50)|$

  $p_{b_\sigma} \sim |\mathcal{N}(0, 50)|$

  $P_h \sim \mathcal{N}(p_h, p_{h_\sigma})$

  $P_b \sim \mathcal{N}(p_b, p_{b_\sigma})$

else

  $p_{h_\sigma} \sim |\mathcal{N}(0, 200)|$

  $p_{b_\sigma} \sim |\mathcal{N}(0, 200)|$

  $P_h \sim \mathcal{N}(\hat{p}_h, p_{h_\sigma})$

  $P_b \sim \mathcal{N}(\hat{p}_b, p_{b_\sigma})$

end

`// Use the data and priors to sample posterior distribution`

$\text{Sample}(\mathcal{N}(\mu = HPLC(T_{in} - T_{ex}) + A_{sol}I_{sol} + \eta_b P_b, \sigma = \xi_Q, \text{Observed} = P_h))$

5.4 Dataset

Two datasets are used to evaluate the performance of QUILT in different situations. The first dataset used is a selection of houses from the solid wall insulation field trial conducted by the UK Energy Saving Trust [102, 96, 19, 117]. For each home in the dataset, the HLC was measured and half-hourly electricity and gas meter readings were provided for a year with the house in normal use. Unfortunately, the HPLC can not be inferred as the heating system efficiency is not provided, hence for these experiments the results will be compared on HLC with an assumed heating system efficiency of 0.85. Also, there are no cooling systems installed, so there is no CPLC to infer. The weather data of each house is available from the nearest weather station provided by the Met Office [94].

5.4.1 Simulated Data

To identify the scenarios in which the HPLC or CPLC can be accurately inferred, the EnergyPlus simulator is used to generate a number of different energy usage scenarios on a set of buildings. By using the EnergyPlus simulator, it enables the evaluation of the performance of QUILT in scenarios where no real-world data exists and provides exploratory results which may be beneficial in targeting future data collection efforts.

The simulated data is created using the residential prototype building models produced by the Pacific Northwest National Laboratory (PNNL) which is used for the U.S. Department of Energy's Building Energy Codes Program[2]. The simulations are run in a variety of climate settings and energy usage scenarios to evaluate the performance of QUILT and benchmark methods to identify when the algorithm can reliably infer the HPLC or CPLC. By simulating different energy usage scenarios on

[2] httpS5://www.energycodes.gov/development/residential/iecc_models

the same set of buildings it allows for direct comparisons to be made.

The scenarios that are simulated are presented in Table 5.1, they are selected to represent different set-ups of data collection which are possible with different hardware to identify what data is required to accurately infer HPLC and CPLC. The use of hardware to acquire the heating-on periods and energy disaggregation is assumed, however, research has shown this can be acquired non-intrusively from smart-meter data [72].

The datasets selected validate the performance of the algorithm on real data and show the potential of it working on a broader range of scenarios with the simulated data. However there is a huge need for publicly available, large-scale datasets as a better understanding of how building thermal performance can lead to significant reductions in global energy usage.

5.5 Experiments

The experiments demonstrate that QUILT outperforms the current state-of-the-art approaches. Furthermore, the experiments identify the data and models required for the HPLC and CPLC to be accurately inferred. The experiments are implemented on both the real and simulated data described in Section 5.4.

5.5.1 Experimental Setup

For the seven real houses, the HLC values are known and the uncertainty is inferred from the known accuracy of co-heating tests. On the simulated dataset, the HPLC and CPLC are inferred with a scenario that is set up to replicate a co-heating test, and the uncertainty in this process is quantified. Then the EnergyPlus simulations are run a number of times to create the scenarios outlined in Table 5.1 for each building in each climate. In each scenario, unless specified otherwise, the schedules of energy usage within the house and occupancy follow the guidelines of the PNNL prototype residential buildings[3]. For the real houses, the results are taken from the previous chapter as it is performed on the same dataset [24], and for the simulated data, both methods are re-implemented.

5.5.2 Evaluation Metrics

To reflect the broad range of applications that QUILT can be used for, a number of evaluation metrics are used. The mean absolute percentage error (MAPE) is used to calculate the absolute error as a percentage of the true HPLC or CPLC value, which is given by,

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\text{H}_{real}^{(n)} - \text{H}_{pred}^{(n)}}{\text{H}_{real}^{(n)}}\right|.$$

Whilst the MAPE does not account for the uncertainty it is a good benchmark for how accurate the algorithm is when the most likely predicted value is selected, which may be how it is used in practise. To evaluate how the uncertainty in the predictions are capturing the true value, we use the Z-score, to measure how many standard deviations (σ) away the predicted HPLC or CPLC is from the true HPLC or CPLC,

$$\text{Z-score} = \frac{H_{pred} - H_{real}}{\sigma_{H_{pred}}}.$$

5.5.3 Evaluation

QUILT shows performance improvement compared to the previous state-of-the-art approach, SMITE, in the real world settings, with an overview of the inferred HLC values by all methods available in Table 5.2. The evaluation metrics are presented

[3]https://www.energycodes.gov/prototype-building-models

			Gas meter readings						Gas and electric combined meter readings					
	Co-heating		Deconstruct		SMITE		QUILT		Deconstruct		SMITE		QUILT	
HLC [W/K]	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
house 1	304	10.1	252	28.0	264	45.4	**246**	**28.4**	271	41.1	186	43.5	**159**	**29.1**
house 2	223	6.3	244	14.6	248	31.2	**212**	**20.7**	258	14.7	235	34.2	**179**	**24.9**
house 3	198	11.7	501	33.8	177	59.7	**220**	**32.8**	517	35.4	195	64.9	**170**	**42.4**
house 4	351	6.3	436	40.2	387	42.6	**299**	**23.7**	433	63.8	524	47.2	**373**	**32.5**
house 5	214	7.1	235	30.8	211	45.1	**215**	**28.6**	302	37.9	239	71.4	**196**	**42.1**
house 6	189	7.4	209	37.8	148	51.2	**188**	**29.7**	205	40.5	218	70.9	**175**	**41.5**
house 7	189	6.6	195	31.0	219	64.2	**244**	**33.2**	206	32.3	240	90.1	**208**	**54.9**

Table 5.2: The HLC values inferred via Deconstruct, SMITE and QUILT for the seven real households alongside the ground-truth HLC inferred via a co-heating test.

	Gas meter readings						Gas and electric combined meter readings					
	Deconstruct		SMITE		QUILT		Deconstruct		SMITE		QUILT	
	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score	% Diff	Z-score
house 1	-17.1	-1.86	-13.4	-0.90	**-19.3**	**-2.07**	-11.0	-0.81	-38.9	-2.72	**-47.7**	**-4.96**
house 2	9.4	1.44	11.4	0.81	**-5.0**	**-0.54**	15.6	2.38	5.5	0.36	**-18.7**	**-1.68**
house 3	53.0	8.97	-10.5	-0.35	**11.0**	**0.66**	160.8	9.00	-1.7	-0.05	**-11.8**	**-0.56**
house 4	24.3	2.12	10.2	0.84	**-14.8**	**-2.20**	23.2	1.28	49.2	3.66	**6.2**	**0.71**
house 5	10.0	0.69	-1.3	-0.06	**0.1**	**0.05**	41.4	2.34	11.6	0.35	**-4.7**	**-0.22**
house 6	10.9	0.55	-21.7	-0.80	**-0.1**	**-0.04**	9.0	0.42	15.5	0.41	**-9.8**	**-0.41**
house 7	3.2	0.20	15.7	0.46	**29.0**	**1.66**	8.8	0.52	26.8	0.56	**10.3**	**0.37**
Average	27.7	1.73	-1.3	0.00	**0.00**	**-0.35**	35.4	2.16	9.7	0.37	**-10.9**	**-0.96**
Average of absolute values	32.6	2.26	12.0	0.60	**11.5**	**1.03**	38.9	2.39	21.3	1.15	**15.6**	**1.27**

Table 5.3: The percentage difference and Z-score between the inferred HLC values and the ground truth values. The metrics indicate a performance improvement by QUILT, particularly in the setting with gas and electricity meter readings combined.

in Table 5.3, where most significantly the MAPE for the inferred HLC compared to the co-heating test is reduced from 21.3 % to 15.6 % in the setting with the combined gas and electricity smart-meter readings, used to represent a scenario with electric heating. There was also a marginal improvement from 12.0 % for the SMITE approach to 11.5 % by QUILT with only gas-meter readings.

A significant change between SMITE and QUILT is the use of a Bayesian model to allow for prior knowledge to inform the inferred HLC value and to more accurately account for additional terms in the thermal model. This is demonstrated in Figure 5.1 which shows the HLC inferred by both methods for House 4 from the real dataset.

	MAPE (%)	
Data collection & model	HPLC	CPLC
SMITE	18.7	-
Daily data	46.7	-
Hourly data	42.0	-
Filtered hourly data	42.0	-
Heating on data	64.5	-
averaged filtered heating period data	**15.8**	**18.9**
averaged filtered heating period data & heating energy known	**15.6**	**17.4**
Week holiday data (all)	24.9	30.3
Week holiday data (>100 data-points)	**13.1**	**16.1**

Table 5.4: The MAPE of the inferred HPLC and CPLC in different scenarios from the simulated dataset.

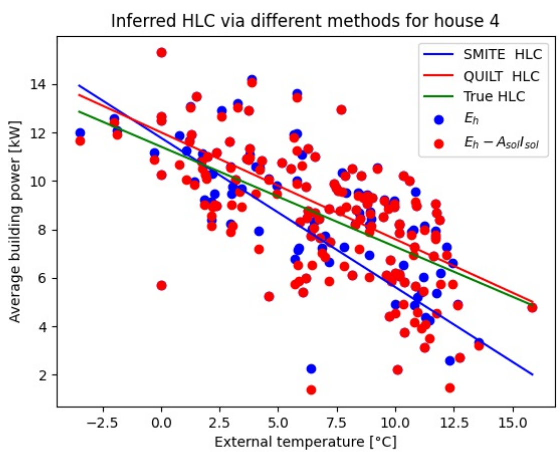

Figure 5.1: How the HLC inferred by SMITE and QUILT relates to the data. Notice that there is only a marginal shift due to the solar irradiance, mainly due to the larger values of solar irradiance being filtered out, and that the prior information helps inform QUILT to better match the true HLC.

Here the HLC is more accurately inferred by QUILT than SMITE, as the new approach accounts for terms beyond the energy usage and the external temperature. Furthermore, the usage of a Bayesian approach means that the prior knowledge of how the HLC values are distributed can inform the inferred HLC for each house. Figure 5.1 also shows that there is a marginal correction when solar irradiance is accounted for, as it is worth remembering in the filtering process that data points with >0.15 kW/m^2 of solar irradiance are removed, hence only a small correction is expected.

In order to compare the accuracy of the inferred HPLC in different scenarios with different algorithms, the EnergyPlus simulator is used to create 60 building scenarios, and the MAPE of the HPLC and CPLC is calculated under different data collection techniques and with different models. The experiments demonstrate the importance of the different steps taken when inferring HPLC and CPLC, and highlight how it is more accurate than the existing state-of-the-art SMITE in a broad range of simulated settings. Across all houses, QUILT shows a reduction in MAPE to 15.6 %, compared to SMITE which achieves a MAPE of 18.7 %. When inferring the CPLC, there was a worse performance with a slightly larger MAPE. This was expected, as with a cooling system there are terms in the thermal equilibrium model that are hard to model, e.g. the higher solar irradiance and the presence of open windows. This leads to less data-points being collected and noisier data-points. Beyond the improved accuracy, the simulated data experiments also allow for a number of scenarios to be tested to identify where QUILT works. The MAPE for the inferred HPLC and CPLC in these different scenarios are presented in Table 5.4.

5.5.3.1 Filtering, Clustering, and Disaggregation

To identify what data is required and which models need to be used to accurately infer the HPLC, the MAPE in a number of different simulated scenarios are compared. As a starting point, when hourly energy usage data is used with no filters, the MAPE is 70.6 %, which is unsurprising as steps need to be taken to get the correct data. When filters are applied to only take data-points when the external temperature is low enough that the heating could feasibly be on, and the solar irradiance is small such that it doesn't have a significant impact on the thermal energy in the building, the MAPE decreases to 46.4 %, this is however still inaccurate and not useful in practise.

As a next step, only using hourly data when the heating is detected to be on gives a worse MAPE of 68.6 %. This counter-intuitive result is caused by the fact that when the heating is first on, the building is not in thermal equilibrium and excess energy is required to heat up the building before it reaches thermal equilibrium. This can be seen in how the data-points split into 3 modes below 15°C in the plot of hourly data in Figure 4.3. Hence, the reason why the filtered-only data appears to be doing better is that the two sources of error are slightly canceling each other out. This indicates that the heating periods of data need careful treatment.

Hence, the start of the heating period is removed to account for the period where the building is not in thermal equilibrium, and the average is taken for the energy usage, external temperature, and solar irradiance for the duration of the rest of the heating period. By using these values to infer the HPLC, the MAPE significantly reduces to 15.8 %. By applying the correct filters and models to select the data, it ensures the data being input into the model aligns with the modeling assumptions that are made (i.e. that the building is in thermal equilibrium). Also, by taking the average across the heating period it reduces the variation in data-points caused by energy usage for other appliances.

With existing disaggregation techniques, the energy usage for heating can be identified and used in the algorithm. However, this did not provide an increase in MAPE when inferring the HPLC, as when the heating is on it typically consumes a large proportion of the energy used by the house and a significant proportion of energy usage for other appliances is relatively constant across time-steps [14]. Hence energy usage disaggregation did not provide a significant increase in accuracy when inferring HPLC on the simulated data and other steps made a much more significant impact. However, it is worth noting that this may be a limitation of the simulated data.

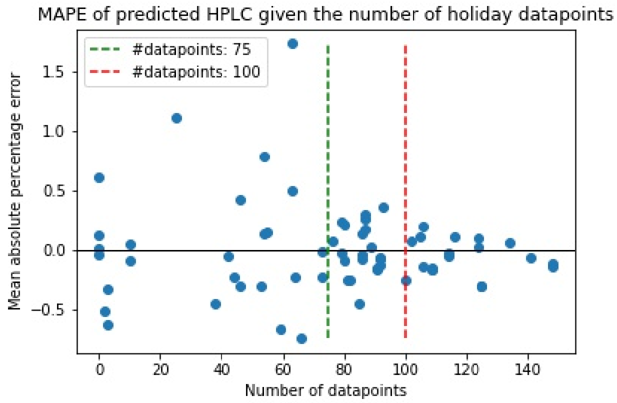

Figure 5.2: A random week from winter is selected as the holiday period for each house. The average MAPE is 24.9 % across all houses. However, when there are over 100 data-points the MAPE is 13.1 % which is more accurate than all other scenarios. This shows that inferring the HPLC is viable for periods when the homeowner is on holiday if the external conditions allow for enough valid data-points.

5.5.3.2 Using Holiday Data

The experiments show that it is feasible to infer the HPLC of a home accurately from holiday data when there are enough data-points collected. Whilst the average MAPE across all situations is 24.9 %, for houses with over 100 data-points the MAPE reduces significantly to 13.1 %. This demonstrates that a viable method of inferring the HPLC is to monitor the energy usage whilst the building is unoccupied for a short duration. However this comes with the caveat that there must be enough data-points that are accepted by the filter stage, i.e. there need to be at least 100 hours of temperatures below 15°C and with solar irradiance below 0.15 kW/m^2. Figure 5.2 outlines the relation between the MAPE and the number of data-points. Further real-world experiments may identify that this can be inferred to a sufficient level of accuracy with fewer data-points, and it may be possible with 100 half-hour data-points instead, which would half the amount of time the house needs to be in eligible conditions through the holiday. Furthermore, real-world experiments could explore how low the temperature on the thermostat can be set whilst still accurately inferring the HPLC, to minimize the energy used in this HPLC labeling process. These results show that using data when the home is unoccupied can provide an accurate inference of the HPLC with approximately a week-worth of data with the right conditions. As such, this can prove to be a useful tool when there is not enough data to infer the HPLC via other methods, or when the inferred HPLC via the other methods is inaccurate.

5.5.3.3 Cooling and CPLC

QUILT accurately infers the CPLC with a MAPE of 17.4 % compared to 15.6 % when inferring HPLC. Since the thermal model is fundamentally the same when the building has a cooling system or heating system running, it should not be a surprise that the accuracy levels are similar. The slightly worse performance can be attributed to the fact that the cooling system typically uses less power, and there is typically more solar irradiance and windows opened when a cooling system is operating which means there are less viable data-points and more noise in the remaining data-points.

When the occupants are on holiday the CPLC was inferred with a MAPE of 16.1 % when there was enough data. This is a useful insight as it is more likely that the occupants will take a holiday over the summer and there is potential to learn the HLC if the cooling system efficiency is known. Further work could explore scenarios beyond holiday periods where the house is empty and identify scenarios where heating and cooling systems are run in situations with low noise and solar irradiance. For example, it is common for people to run a cooling system through the night when there is typically no solar irradiance, the number of occupants is fixed and energy usage for other applications is low. This reiterates the need for more labeled real-world data to help identify these scenarios where the HPLC and CPLC can be accurately inferred non-intrusively.

5.6 Actionable Insights

QUILT can be used to quantify the thermal performance of buildings which may be used to help create an EPC, and be used by energy companies, governing bodies, and homeowners to identify energy-saving interventions. Studies have shown that the energy efficiency of a building can increase its price, providing an economic incentive beyond energy savings, which often do not cover the cost of the energy-saving interventions [63]. Hence, correctly labeling the thermal performance of buildings in the EPC is important to align environmental impact with financial incentives. This is now achievable non-intrusively, accurately, and at a low cost thanks to the development of QUILT.

The experimental results highlight how effective QUILT can be at inferring HPLC and CPLC. Our experiments have demonstrated the feasibility of QUILT on real-world data and on a broader set of scenarios using simulated data. However, steps need to be taken to operationalise this research in the real world. There is a need for more real-world data on the thermal performance of buildings, and such efforts of creating real-world datasets are underway[4]. Deploying QUILT on larger real-world datasets is an important next step to demonstrate how robust it is and to identify scenarios where adaptations are needed to work beyond the datasets it has been tested on.

If QUILT is deployed at scale, the operators of the algorithm can combine it with other sources of data to identify the most suitable interventions. Houses with high heating energy usage and low HPLC may benefit from learning to reduce their thermostat temperature, and learning how more careful heating scheduling can lead to reduced energy usage. Whilst survey data has been shown to not be useful for inferring the thermal performance, it does identify what potential interventions can be made to improve insulation, and combined with an accurate HPLC measurement

[4] httpS5://www.gov.uk/guidance/smart-meter-enabled-thermal-efficiency-ratings-smeter-innovation-programme

more targeted interventions can be made to the buildings with the highest HPLC [73]. Furthermore, if QUILT is deployed widely, it will uncover a large number of inefficient heating systems, which can be replaced by modern solutions such as heat pumps.

If the evaluation of QUILT on a broader range of real homes identifies it to be inaccurate in certain scenarios then steps can be taken to improve the data collection process. For example, smart thermostats could remove a potential source of inaccuracy by providing information of when the heating is on, and what the internal temperature is. As discussed in Section 5.5, identifying the energy usage for heating does not provide any noticeable benefit when identifying HPLC in our simulated scenarios, however, real-world scenarios may be more complex, meaning this could still be a valuable tool.

Furthermore, if the energy usage data for the house is too noisy, or there is not enough data, a targeted data collection process may be beneficial. The results show that the HPLC can be inferred if the house is unoccupied, and there are at least 100 viable smart-meter readings where the solar irradiance and temperature filters are satisfied. This requires the heating system to be on for the duration of the holiday. In real-world experiments, different thermostat set-points could be explored to minimise the cost of inferring the HPLC. Also, as inferring the HPLC is a function of the difference between internal and external temperature, if there is not enough variation in external temperature the thermostat set-point can be varied.

5.7 Summary

In this chapter we have proposed QUILT, a Bayesian approach for inferring the HPLC and CPLC of buildings non-intrusively using only smart-meter readings and weather data. The new approach improves on the state of the art, as demonstrated by a number of experiments on real data, and on simulated data to represent scenarios with unavailable real-world data. Furthermore, the experiments conducted explore the importance of different features when inferring HPLC and CPLC, and identify that correctly clustering heating periods are of great importance, that the HPLC and CPLC can also be inferred when the occupants are on holiday if there are enough valid data points, and that energy disaggregation for heating and cooling does not make a big impact on the accuracy of inferred HPLC and CPLC. These findings emphasise how important a successful smart-meter roll-out is to enable accurate inference of the thermal performance of buildings across the country.

6

Conclusions

This thesis proposes new methods to identify the PV generation of a residential building, its thermal performance, and to quantify where the thermal performance can be accurately inferred. The methods developed can be deployed at scale cheaply, and the accuracy is shown to be in line with, or better than, the state-of-the-art. These proposed methods can be used to augment the EPC to better capture the real-world energy performance of buildings. This has been made possible by the wide adoption of smart meters and publicly accessible weather data.

1. Can we infer the maximum power output of PV panels on residential households whilst accounting for censored smart-meter readings?

PV panels are the most common method for on-site residential energy generation. Unfortunately, incorrect installations cause PV generation to be measured incorrectly in the majority of cases causing the smart-meter readings to be censored and not reflect the true energy usage and generation. In Chapter 3, this problem is addressed with a method proposed to disaggregate PV generation from energy consumption given censored smart-meter readings and to infer the PV capacity. This is achieved by an algorithm that finds the most likely joint probability of energy consumption for each time period and the PV capacity. The approach is evaluated on the Pecan Street dataset, and due to a lack of UK data, a dataset is created by combining PV generation data from the Sheffield Solar microgen dataset and smart-meter readings from London households. It is shown that the presence of a PV panel can be detected 100 % of the time, and the PV capacity can be inferred with a mean absolute percentage error of 13 % on the UK dataset. This is the first method to account for censored smart-meter readings to ensure the PV generation can be inferred when smart-meters are not correctly calibrated with PV panels, enabling the inference to be performed across more houses with PV panels.

2. Can we accurately identify the thermal performance of a residential home non-intrusively?

The existing methods to evaluate the thermal performance of residential properties are intrusive and expensive to perform, or inaccurate. To address this, SMITE is proposed in Chapter 4, a method to infer thermal performance non-intrusively using only smart-meter readings and local weather data. To achieve this, a thermal model which represents the heat flow through the building is proposed. Then a hidden Markov model and set of filters are used to identify when the building is in a state of thermal equilibrium, identifying the data-points that can be used to infer the thermal performance of the building. Then the HPLC can be inferred using these data-points. The SMITE method is evaluated on a dataset of 7 real homes selected from the solid wall insulation field trials, and achieves a MAPE of 12.0 % when inferring the thermal performance with gas heating, and 21.3 % with electric heating. This is compared to a MAPE of 32.6 % and 38.9 % respectively for the previous state-of-the-art approach [29]. This improved method was deployed by EDF Energy in their bid for the BEIS SMETER innovation competition. As thermal space comfort is the largest proportion of energy usage in a home, identifying which homes have the potential for the biggest increase in thermal performance can contribute to a significant reduction of energy usage in the residential sector.

3. Can we identify the scenarios in which the thermal performance can be accurately inferred and quantify the uncertainty?

Buildings have different amounts of data available for collection and are situated in different environments. Hence, it is important to understand the limitations of inferring thermal performance to identify when it can reliably be inferred and to make it work in as many scenarios as possible. Traditionally, thermal performance has corresponded to how efficiently a building is heated, not how efficiently it is cooled. In Chapter 5, QUILT is proposed to extend the heat flow model from Chapter 4, to work when a cooling system is being used as well as a heating system, and to use Bayesian inference to allow prior information to inform the parameters of the model and to quantify the uncertainty from different parameters in the model and data sources. Then a variety of scenarios are created with EnergyPlus, a building energy simulator, which identify the scenarios where the thermal performance can be accurately inferred, and in which scenarios more data will be required. This method achieves a reduction of 27 % in MAPE compared to SMITE in some scenarios and demonstrates that the thermal performance of a building can be inferred to a similar level of accuracy with a cooling system.

6.1 Real World Impact and Challenges

For new residential buildings the SAP is used to produce the EPC. It is typically conducted before the residents are in the building which means there is no real data for the methods to be performed on. However, the proposed methods could be used to evaluate the thermal performance and identify the true PV generation a year after the residents move in. This can be performed non-intrusively at low-cost whilst giving an accurate evaluation of the real-world thermal performance and maximum PV generation. If these methods are implemented across enough households, it may be possible to create a dataset which can be used to re-calibrate the predictions of thermal performance in the SAP based on the surveys conducted. Alternatively, the methods proposed in this thesis may be used during the RdSAP which is conducted to update energy certificates when there are residents in the building. The RdSAP can be improved to be more accurate and cheaper with the proposed methods, and it may even be possible to conduct it remotely.

Operationalising the research conducted in this thesis will enable a more accurate, quicker, and more scalable process to label homes with an EPC. The reduced cost will be beneficial to all parties involved, and the improved accuracy will ensure that more homes are labelled correctly to align incentives for energy saving measures correctly. Beyond augmenting the EPC, the methods developed could be deployed for residents to independently check how energy efficient their home is. This will enable home-owners to better understand if energy-saving and cost-saving measures can be taken in their house. The proposed methods may also be used by energy companies to better understand the energy usage profiles of their customers and ensure they are compliant with ECO by ensuring they do their best to protect fuel-poor customers.

The methods presented in this thesis have the potential to augment the EPC by making it more accurate, cheaper, and scalable. However, to operationalise this research a number of steps are required. It is important that any implementations comply with GDPR or local regulations and give users the appropriate level of control and information regarding the data collected and information inferred from the data. In this thesis, it has been assumed that high-quality data is readily available, however, this assumption may not be true for a number of houses. For example, smart meters can fail to record for periods of time and report a backlog of energy usage when it next updates creating half-hourly energy readings which do not accurately reflect energy usage and generation. Furthermore, the resident can opt to not provide half-hourly data, which means these methods will not work and may have to be adapted for daily data. Also, as weather data is required the location of the house needs to be known and the smart-meter data needs to be aligned with the correct weather data which can introduce a potential source of error.

As well as ensuring the data is usable and regulatory compliant, there may be further research challenges that need to be addressed. For example, the presence of residential energy storage can change when energy is taken from the grid, which in turn affects smart-meter readings. Hence, if the residential energy storage is not correctly dealt with it may affect assumptions about the smart-meter readings, such as the assumption that half-hourly energy usage follows a specific probability distribution in Chapter 3. There may also be a temporal component to the metrics which are not currently captured as the maximum PV generation and thermal performance can change over time based on a number of factors such as renovations and degradation. Therefore, the proposed methods may need to be conducted at regular intervals to identify changes, or further methods may need to be developed, such as fault detection, to identify when there is a change in value.

6.2 Future Work Recommendations

The work completed in this thesis can be built upon in a number of different areas.

With a method to infer the thermal performance and PV generation of residential buildings, regular re-evaluation is possible and this may be used to **identify PV panel or thermal performance faults** via drifting HPLC, CPLC, or PV capacity values. This could potentially be achieved with an auto-regressive model or a Gaussian process to identify if the metrics are changing over time. Seasonality of when the data is typically collected may make the time-frames for this task long, and there is a lack of large-scale long-term datasets in this domain.

Extensions may be developed to **adapt the methods in this thesis to a broader range of building types** such as commercial buildings, or apartment blocks. These adaptations will have to address the significant challenge of creating new thermal models to represent the heat flows of the different buildings, or for PV capacity inference the probability distributions of energy usage will need to be identified for each building type. Furthermore, identifying the low-noise data-points will require new approaches as the energy profile is typically very different from a residential home. For example, commercial buildings typically have low occupation levels through the night, which may provide a good angle for identifying the thermal performance. However, commercial buildings also typically have multiple heating systems with complex control systems which will need to be accounted for.

This thesis has introduced a number of ways to get metrics from residential buildings, which may be used to inform other tasks, such as **managing energy usage across a district with reinforcement learning**. The learned HPLC, CPLC, and PV capacity may be used as prior values for each household, and reinforcement learning approaches can be used to schedule energy usage based on these metrics. It may also be possible to incorporate the models used in this thesis into a model-based reinforcement learning approach for such tasks. This presents a significant challenge as alternative approaches achieve very good performance levels that need to be surpassed, and successful implementation will require a deep understanding of reinforcement learning.

Energy suppliers currently lack knowledge on the size of the building the energy is being provided to, hence developing **tools to infer the size of buildings** would provide insight that would give context for other metrics. This is useful information as it correlates with how much energy is used to heat or cool the building and thermal performance could be conditioned on the size of the building. There is potential to achieve this via open-source satellite and street-view data as computer vision techniques improve, however, the task of separating terraced houses via such imagery is a difficult challenge.

Bibliography

[1] Warm homes and energy conservation act of 2000.

[2] International Energy Agency. The future of cooling. https://www.iea.org/reports/the-future-of-cooling/, 2018.

[3] K. Andersen, H. Madsen, and L. Hansen. Modelling the heat dynamics of a building using stochastic differential equations. *Energy and Buildings*, 31(1):13–24, 2000.

[4] S. Aznavi, P. Fajri, R. Sabzehgar, and A. Asrari. Optimal management of residential energy storage systems in presence of intermittencies. *Journal of Building Engineering*, 29:101149, 2020.

[5] A. Azouzoute, A. Merrouni, A. Gennioui, et al. Accuracy measurement of pyranometer vs reference cell for pv resource assessment. *Energy Procedia*, 157:1202–1209, 2019.

[6] P. Bacher and H. Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, 2011.

[7] P. Baker and H. Van Dijk. Paslink and dynamic outdoor testing of building components. *Building and Environment*, 43(2):143–151, 2008.

[8] D. Barber. *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[9] M. Basit, S. Dilshad, R. Badar, and S. Sami ur Rehman. Limitations, challenges, and solution approaches in grid-connected renewable energy systems. *International Journal of Energy Research*, 44(6):4132–4162, 2020.

[10] G. Bauwens and S. Roels. Co-Heating Test: A State-of-the-Art. *Energy and Buildings*, 82:163–172, 2014.

[11] BEIS. Standard assessment procedure. https://www.gov.uk/guidance/standard-assessment-procedure, 2013.

[12] BEIS. Smart metering implementation programme: progress report. https://www.gov.uk/government/publications/smart-metering-implementation-programme-progress-report-2018, 2018.

[13] BEIS. Department for Business, Energy and Industrial Strategy, Energy Consumption in the UK (ECUK). 2019. Online; accessed 21 July 2020.

[14] BEIS. Energy consumption in the UK. https://www.gov.uk/government/statistics/energy-consumption-in-the-uk, 2020.

[15] BEIS. National statistics: Solar photovoltaics deployment. https://www.gov.uk/government/statistics/solar-photovoltaics-deployment, 2021.

[16] BEIS. UK energy in brief. https://www.gov.uk/government/statistics/uk-energy-in-brief-2021, 2021.

[17] M. Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[18] P. Biddulph, V. Gori, C. Elwell, C. Scott, C. Rye, R. Lowe, and T. Oreszczyn. Inferring the thermal resistance and effective thermal mass of a wall using frequent temperature and heat flux measurements. *Energy and Buildings*, 78:10–16, 2014.

[19] S. Birchall, C. Pearson, and R. Brown. Solid wall insulation field trial-baseline performance of the property sample. *Bracknell: BSRIA*, 2011.

[20] R. Bird and R. Hulstrom. Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1981.

[21] G. Box and G. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

[22] BRE. Building Research Establishment, SAP 2009: The Government's Standard Assessment Procedure for Energy Rating of Dwellings. `https://www.bre.co.uk/filelibrary/SAP/2009/SAP-2009_9-90.pdf`, 2010. Online; accessed 21 July 2020.

[23] BRE. Standard Assessment Procedure (SAP 2012). 2012. Online; accessed 18 Dec 2021.

[24] J. Brown, J. Chambers, A. Abate, and A. Rogers. SMITE: Using Smart Meters to Infer the Thermal Efficiency of Residential Homes. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 21–30, 2020.

[25] M. Castaneda, S. Zapata, J. Cherni, A. Aristizabal, and I. Dyner. The long-term effects of cautious feed-in tariff reductions on photovoltaic generation in the uk residential sector. *Renewable Energy*, 155:1432–1443, 2020.

[26] M. Ceci, R. Corizzo, F. Fumarola, D. Malerba, and A. Rashkovska. Predictive modeling of PV energy production: How to set up the learning task for a better prediction? *Proceedings of the IEEE Transactions on Industrial Informatics*, 13(3):956–966, 2017.

[27] P. Chakraborty, M. Marwah, M. Arlitt, and N. Ramakrishnan. Fine–grained photovoltaic output prediction using a Bayesian ensemble. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 274–280, 2012.

[28] J. Chambers. *Developing a rapid, scalable method of thermal characterisation for UK dwellings using smart meter data.* PhD thesis, UCL (University College London), 2017.

[29] J. Chambers and T. Oreszczyn. Deconstruct: A Scalable Method of As-Built Heat Power Loss Coefficient Inference For UK Dwellings Using Smart Meter Data. *Energy and Buildings*, 183:443–453, 2019.

[30] D. Chen, P. Bovornkeeratiroj, D. Irwin, and P. Shenoy. Private memoirs of iot devices: Safeguarding user privacy in the iot era. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 1327–1336. IEEE, 2018.

[31] D. Chen and D. Irwin. Black-box solar performance modeling: Comparing physical, machine learning, and hybrid approaches. *Proceedings of ACM SIGMETRICS Performance Evaluation Review*, 45(2):79–84, 2017.

[32] D. Chen and D. Irwin. Sundance: Black-box behind-the-meter solar disaggregation. In *Proceedings of the Eighth International Conference on Future Energy Systems*, pages 45–55. ACM, 2017.

[33] D. Chen and D. Irwin. Weatherman: Exposing weather-based privacy threats in big energy data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1079–1086. IEEE, 2017.

[34] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: benchmark and state of the art. In *Proceedings of the IEEE*, volume 105, pages 1865–1883. IEEE, 2017.

[35] A. Chiabai, S. Quiroga, P. Martinez-Juarez, S. Higgins, and Tim Taylor. The nexus between climate change, ecosystem services and human health: Towards a conceptual framework. *Science of the Total Environment*, 635:1191–1204, 2018.

[36] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *Journal of the American statistical Association*, 49(4):327–335, 1995.

[37] J. Crawley, E. McKenna, V. Gori, and T. Oreszczyn. Creating Domestic Building Thermal Performance Ratings Using Smart Meter Data. *Buildings & Cities*, 2020.

[38] F. Culière, L. Leduc, and A. Belikov. Bayesian model of electrical heating disaggregation. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, pages 25–29, 2020.

[39] S. Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 2006.

[40] J. Deutch. Is net zero carbon 2050 possible? *Joule*, 4(11):2237–2240, 2020.

[41] X. Ding, W. Du, and A. Cerpa. Mb2c: Model-based deep reinforcement learning for multi-zone building control. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 50–59, 2020.

[42] S. Dong, E. Kremers, M. Brucoli, R. Rothman, and S. Brown. Improving the feasibility of household and community energy storage: A techno-enviro-economic study for the uk. *Renewable and Sustainable Energy Reviews*, 131:110009, 2020.

[43] S. Doran and B. Carr. Thermal transmittance of walls of dwellings before and after application of cavity wall insulation. *Energy Saving Trust*, 2008.

[44] Ben Elliston and Iain MacGill. The potential role of forecasting for integrating solar generation into the Australian national electricity market. In *proceedings of Solar 2010: The annual conference of the Australian solar energy society*, 2010.

[45] B. Espinar, J. Aznarte, R. Girard, A. Moussa, and G. Kariniotakis. Photovoltaic forecasting: A state of the art. In *Proceedings of the Fifth European PV-Hybrid and Mini-Grid Conference*, pages Pages–250. OTTI-Ostbayerisches Technologie-Transfer-Institut, 2010.

[46] European Commission. Directive (EU) 2018/2002 of the European Parliament and of the Council of 11 December 2018 amending Directive 2012/27/EU on energy efficiency (Text with EEA relevance.). *Official Journal of the European Union*, 2018. Online; accessed 09 July 2021.

[47] European Commission. Commission Delegated Regulation (EU) 2019/826 of 4 March 2019 amending Annexes VIII and IX to Directive 2012/27/EU of the European Parliament and of the Council on the contents of comprehensive

assessments of the potential for efficient heating and cooling. *Official Journal of the European Union*, 2019. Online; accessed 09 July 2021.

[48] D. Farmer, D. Johnston, and D. Miles-Shenton. Obtaining the heat loss coefficient of a dwelling using its heating system (integrated coheating). *Energy and Buildings*, 117:1–10, 2016.

[49] A. Fateh, D. Borelli, A. Spoladore, and F. Devia. A state-space analysis of a single zone building considering solar radiation, internal radiation, and PCM effects. *Applied Sciences*, 9(5):832, 2019.

[50] M. Fels et al. Prism: an introduction. *Energy and Buildings*, 9(1-2):5–18, 1986.

[51] A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay. State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews*, 23:272–288, 2013.

[52] M. Frei, C. Deb, Z. Nagy, I. Hischier, and A. Schlueter. Building Energy Performance Assessment Using an Easily Deployable Sensor Kit: Process, Risks and Lessons Learned. *Frontiers in Built Environment*, 6:222, 2020.

[53] A. Freundlich and A. Alemu. Multi quantum well multijunction solar cell for space applications. *physica status solidi*, 2(8):2978–2981, 2005.

[54] S. Gallardo-Saavedra and B. Karlsson. Simulation, validation and analysis of shading effects on a PV system. *Solar Energy*, 170:828–839, 2018.

[55] J. Gattuso, A. Magnan, L. Bopp, W. Cheung, C. Duarte, J. Hinkel, E. Mcleod, F. Micheli, A. Oschlies, P. Williamson, et al. Ocean solutions to address climate change and its effects on marine ecosystems. *Frontiers in Marine Science*, 5:337, 2018.

[56] C. Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.

[57] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[58] S. Goyal and P. Barooah. A method for model-reduction of non-linear thermal dynamics of multi-zone buildings. *Energy and Buildings*, 47:332–340, 2012.

[59] Y. He, A. Gupta, H. Cheng, and V. Garud. Energy disaggregation techniques for whole-house energy consumption data, May 21 2015. US Patent App. 14/543,824.

[60] M. Hoffman, A. Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[61] F. Hollick, V. Gori, and C. Elwell. Thermal Performance of Occupied Homes: A Dynamic Grey-Box Method Accounting for Solar Gains. *Energy and Buildings*, 208:109669, 2020.

[62] Z. Huifen, Y. Fuhua, and Z. Qian. Research on the Impact of Wind Angles on the Residential Building Energy Consumption. *Mathematical Problems in Engineering*, 2014.

[63] M. Hyland, R. Lyons, and S. Lyons. The value of domestic building energy efficiency—evidence from ireland. *Energy economics*, 40:943–952, 2013.

[64] IEA. International energy agency, world energy statistics and balances. `https://www.iea.org/reports/world-energy-balances-2019`, 2018. Online; accessed 21 July 2020.

[65] IPCC. Global warming of 1.5 °c special report. https://www.ipcc.ch/sr15/, 2018.

[66] IPCC. Intergovernmental Panel on Climate Change, 2018: Global Warming of 1.5°C. `https://www.ipcc.ch/sr15/`, 2018. Online; accessed 21 July 2020.

[67] Climate Change IPCC. Mitigation of Climate Change. *Contribution of working group III to the fifth assessment report of the intergovernmental panel on climate change*, 2014. Online; accessed 21 July 2020.

[68] S. Iyengar, S. Lee, D. Irwin, P. Shenoy, and B. Weil. WattScale: A Data-driven Approach for Energy Efficiency Analytics of Buildings at Scale. *ACM Transactions on Data Science*, 2(1):1–25, 2021.

[69] R. Jack. *Building diagnostics: practical measurement of the fabric thermal performance of houses*. PhD thesis, Loughborough University, 2015.

[70] M. Jimenez and H. Madsen. Models for describing the thermal characteristics of building components. *Building and Environment*, 43(2):152–162, 2008.

[71] E. Kara, C. Roberts, M. Tabone, L. Alvarez, D. Callaway, and E. Stewart. Disaggregating solar generation from feeder-level measurements. *Sustainable Energy, Grids and Networks*, 13:112–121, 2018.

[72] J. Kelly and W. Knottenbelt. Neural nilm: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments*, pages 55–64, 2015.

[73] S. Kelly, D. Crawford-Brown, and M. Pollitt. Building performance evaluation and certification in the uk: Is sap fit for purpose? *Renewable and Sustainable Energy Reviews*, 16(9):6861–6878, 2012.

[74] D. Kim, J. Youn, and C. Kim. Automatic fault recognition of photovoltaic modules based on statistical analysis of uav thermography. *The International*

Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 42:179, 2017.

[75] J. Klein and M. Moeschberger. *Survival analysis: Techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[76] E. Lambie and D. Saelens. Identification of the building envelope performance of a residential building: A case study. *Energies*, 13(10):2469, 2020.

[77] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):255–258, 1995.

[78] F. Li, A. Smith, P. Biddulph, I. Hamilton, R. Lowe, A. Mavrogianni, E. Oikonomou, R. Raslan, S. Stamp, A. Stone, et al. Solid-wall u-values: heat flux measurements compared with standard assumptions. *Building Research & Information*, 43(2):238–252, 2015.

[79] Hui Li. *Pavement materials for heat island mitigation: design and management strategies*. Butterworth-Heinemann, 2015.

[80] J. Love. *Understanding the interactions between occupants, heating systems and building fabric in the context of energy efficient building fabric retrofit in social housing*. PhD thesis, UCL (University College London), 2014.

[81] S. Makonin and F. Popowich. Nonintrusive load monitoring (nilm) performance evaluation. *Energy Efficiency*, 8(4):809–814, 2015.

[82] J. Malof, K. Bradbury, L. Collins, R. Newell, A. Serrano, H. Wu, and S. Keene. Image features for pixel-wise detection of solar photovoltaic arrays

in aerial imagery using a random forest classifier. In *Proceedings of the Fifth IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 799–803. IEEE, 2016.

[83] J. Malof, L. Collins, K. Bradbury, and R. Newell. A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery. In *Renewable Energy Research and Applications (ICRERA), 2016 IEEE International Conference on*, pages 650–654. IEEE, 2016.

[84] C. Marino, A. Nucara, and M. Pietrafesa. Thermal comfort in indoor environment: Effect of the solar radiation on the radiant temperature asymmetry. *Solar Energy*, 144:295–309, 2017.

[85] E. McKenna, J. Pless, and S. Darby. Solar photovoltaic self-consumption in the uk residential sector: New estimates from a smart grid demonstration project. *Energy Policy*, 118:482–491, 2018.

[86] S. McLaughlin, P. McDaniel, and W. Aiello. Protecting consumer privacy from electric load monitoring. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 87–98, 2011.

[87] S. Mehta, A. Azad, S. Chemmengath, V. Raykar, and S. Kalyanaraman. Deepsolareye: Power loss prediction and weakly supervised soiling localization via fully convolutional networks for solar panels. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 333–342. IEEE, 2018.

[88] A. Mei, X. Li, L. Liu, et al. A hole-conductor–free, fully printable mesoscopic perovskite solar cell with high stability. *Science*, 345(6194):295–298, 2014.

[89] A. Mellit and A. Pavan. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at trieste, italy. *Solar Energy*, 84(5):807–821, 2010.

[90] J. Meulemans, F. Alzetto, D. Farmer, and C. Gorse. Qub/e: A novel transient experimental method for in situ measurements of the thermal performance of building fabrics. In *Building Information Modelling, Building Performance, Design and Smart Construction*, pages 115–127. Springer, 2017.

[91] R. Mohan, T. Cheng, A. Gupta, V. Garud, and Y. He. Solar energy disaggregation using whole-house consumption signals. In *NILM Workshop*, 2014.

[92] L. Moutinho and G. Hutcheson. Ordinary least-squares regression. *and GD Hutcheson, The SAGE dictionary of quantitative management research*, pages 224–228, 2011.

[93] J. Myhren and S. Holmberg. Flow patterns and thermal comfort in a room with panel, floor and wall heating. *Energy and buildings*, 40(4):524–536, 2008.

[94] Met Office. Met office midas open: Uk land surface stations data (1853-current). centre for environmental data analysis. `http://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1`, 2020. Online; accessed 21 July 2020.

[95] Comittee on Climate Change. The sixth carbon budget the UK's path to net zero. https://www.theccc.org.uk/publication/sixth-carbon-budget/, 2020.

[96] T. Oreszczyn, J. Chambers, V. Gori, P. Biddulph, I. Hamilton, and C. Elwell. How solid is our knowledge of solid walls? Comparing energy savings through three different methods. In *Proceedings of CISBAT 2015 international conference on*

future buildings and districts-sustainability from nano to urban scale-vol 1, volume 2015, pages 107–112. EPFL Solar Energy and Building Physics Laboratory, 2015.

[97] World Health Organization et al. Cop24 special report: health and climate change. 2018.

[98] G Orr, T Lelyveld, S Burton, and I Summerfield. In-situ monitoring of efficiencies of condensing boilers and use of secondary heating. *Energy Savings Trust Report*, 2009.

[99] R. Perez, M. Beauharnois, K. Hemker, S. Kivalov, E. Lorenz, S. Pelland, J. Schlemmer, and G. Van Knowe. Evaluation of numerical weather prediction solar irradiance forecasts in the US. 2011.

[100] A. Peruffo, E. Guiu, P. Panciatici, and A. Abate. Aggregated Markov models of a heterogeneous population of photovoltaic panels. In *Proceedings of International Conference on Quantitative Evaluation of Systems*, pages 72–87. Springer, 2017.

[101] A. Peruffo, E. Guiu, P. Panciatici, and A. Abate. Safety guarantees for the electricity grid with significant renewables generation. In *Proceedings of International Conference on Quantitative Evaluation of Systems*, pages 332–349. Springer, 2019.

[102] D. Prince. Solid wall insulation study. *Energy Saving Trust*, 2014.

[103] C. Rasmussen, P. Bacher, D. Calì, HA Nielsen, and H. Madsen. Method for scalable and automatised thermal building performance documentation and screening. *Energies*, 13(15):3866, 2020.

[104] JP Real, C. Rasmussen, R. Li, K. Leerbeck, OM Jensen, KB Wittchen, and H Madsen. Characterisation of thermal energy dynamics of residential buildings with scarce data. *Energy and Buildings*, 230:110530, 2021.

[105] S. Rehman and M. Mohandes. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. *Energy Policy*, 36(2):571–576, 2008.

[106] J. Rhodes, W. Cole, C. Upshaw, T. Edgar, and M. Webber. Clustering analysis of residential electricity demand profiles. *Applied Energy*, 135:461–471, 2014.

[107] F. Rodríguez, A. Fleetwood, A. Galarza, and L. Fontán. Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control. *Renewable Energy*, 126:855–864, 2018.

[108] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[109] R. Ross Jr. Flat-plate photovoltaic array design optimization. In *Proceedings of the fourteenth Photovoltaic Specialists Conference*, pages 1126–1132, 1980.

[110] J. Salvatier, T. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

[111] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme. Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. *Nature Energy*, 3(2):119–126, 2018.

[112] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. Predicting solar generation from weather forecasts using machine learning. In *Proceedings of IEEE International*

Conference on Smart Grid Communications (SmartGridComm), pages 528–533. IEEE, 2011.

[113] C. Smith. *The Pecan Street Project: developing the electric utility system of the future*. PhD thesis, U. of Texas, 2009.

[114] R. Sonderegger, P. Condon, and M. Modera. In-Situ Measurements of Residential Energy Performance Using Electric Co-Heating. Technical report, California Univ., Berkeley (USA). Lawrence Berkeley Lab., 1980.

[115] R. Sonderegger and M. Modera. Electric co-Heating: A Method For Evaluating Seasonal Heating Efficiencies and Heat Loss Rates in Dwellings. 1979.

[116] J. Stein, W. Holmgren, J. Forbess, and C. Hansen. PVLIB: Open source photovoltaic performance modeling functions for MATLAB and Python. In *Proceedings of Forty-Fourth IEEE Photovoltaic Specialist Conference (PVSC)*, pages 1–6. IEEE, 2017.

[117] G. Stevens and J. Bradford. Do u-value insulation? england's field trial of solid wall insulation. In *ECEEE Summer Study Proceedings*, volume 2013, pages 1269–1280, 2013.

[118] K. Subbarao. PSTAR: Primary and Secondary Terms Analysis and Renormalization: A Unified Approach to Building Energy Simulations and Short-Term Monitoring. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1988.

[119] A. Summerfield, T. Oreszczyn, I. Hamilton, D. Shipworth, G. Huebner, R. Lowe, and P. Ruyssevelt. Empirical variation in 24-h profiles of delivered power for a

sample of UK dwellings: Implications for evaluating energy savings. *Energy and Buildings*, 88:193–202, 2015.

[120] A. Summerfield, T. Oreszczyn, J. Palmer, I. Hamilton, and R. Lowe. Comparison of Empirical And Modelled Energy Performance Across Age-Bands of Three-bedroom Dwellings in the UK. *Energy and Buildings*, 109:328–333, 2015.

[121] Y. Sun, H. Yue, J. Zhang, and C. Booth. Minimization of residential energy cost considering energy storage system and ev with driving usage probabilities. *IEEE Transactions on Sustainable Energy*, 10(4):1752–1763, 2018.

[122] M. Tabone, S. Kiliccote, and E. Kara. Disaggregating solar generation behind individual meters in real time. In *Proceedings of the 5th Conference on Systems for Built Environments*, pages 43–52. ACM, 2018.

[123] J. Taylor, A. Everard, J. Briggs, A. Buckley, S. Casement, H. Jones, J. Harwood, and J. Caplin. Estimating real time GB aggregated solar PV generation.

[124] J. Taylor, J. Leloux, A. Everard, J. Briggs, and A. Buckley. Quantifying PV fleet output variability in the UK: Consequences for distribution network operators.

[125] J. Taylor, J. Leloux, A. Everard, J. Briggs, and A. Buckley. Monitoring thousands of distributed PV systems in the UK: Energy production and performance. *Proceedings of the eleventh Photovoltaic Science, Application and Technology conference PVSAT-11, Leeds*, 2015.

[126] J. Taylor, J. Leloux, L. Hall, A. Everard, J. Briggs, and A. Buckley. Performance of distributed PV in the UK: a statistical analysis of over 7000 systems. In *31st European photovoltaic solar energy conference and exhibition*, 2015.

[127] S. Thébault and R. Bouchié. Refinement of the isabele method regarding uncertainty quantification and thermal dynamics modelling. *Energy and Buildings*, 178:182–205, 2018.

[128] Herbert Conrad Schlueter Thom. The rational relationship between heating degree days and temperature. *Monthly Weather Review*, 82(1):1–6, 1954.

[129] J. Torriti. Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. *Energy research & social science*, 25:37–47, 2017.

[130] A. Ukil, S. Bandyopadhyay, and A. Pal. Privacy for iot: Involuntary privacy enablement for smart energy systems. In *2015 IEEE International Conference on Communications (ICC)*, pages 536–541. IEEE, 2015.

[131] USwitch. Energy Performance Certificate. `https://www.uswitch.com/energy-saving/guides/energy-performance-certificates/`, 2020. Online; accessed 21 July 2020.

[132] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:3152676, 2017.

[133] H. Wang, C. Si, J. Zhao, G. Liu, and F. Wen. Fed-nilm: A federated learning-based non-intrusive load monitoring method for privacy-protection. *arXiv preprint arXiv:2105.11085*, 2021.

[134] H. Wang, J. Zhang, C. Lu, and C. Wu. Privacy preserving in non-intrusive load monitoring: A differential privacy perspective. *IEEE Transactions on Smart Grid*, 12(3):2529–2543, 2020.

[135] C. Weeks, C. Delalonde, and C. Preist. Investigation into the slow adoption of retrofitting. In *Proceedings of the twenty-ninth International Conference on Informatics for Environmental Protection (EnviroInfo 2015), Copenhagen, Denmark*, 2015.

[136] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal. DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. *Joule*, 2(12):2605–2617, 2018.

[137] J. Yuan, H. Yang, O. Omitaomu, and B. Bhaduri. Large-scale solar panel mapping from aerial images using deep convolutional networks. In *Proceedings of the Fourth IEEE International Conference on Big Data*, pages 2703–2708. IEEE, 2016.

[138] C. Zanocco, H. Boudet, R. Nilson, H. Satein, H. Whitley, and J. Flora. Place, proximity, and perceived harm: extreme weather events and views about climate change. *Climatic Change*, 149(3):349–365, 2018.

[139] Y. Zhang, X. Bai, F. Mills, and J. Pezzey. Rethinking the role of occupant behavior in building energy performance: A review. *Energy and Buildings*, 172:279–294, 2018.