A Social Logic of Energy

A data science approach to understanding and modelling energy transitions of India's urban poor

André Paul Neto-Bradley

Fitzwilliam College

This thesis is submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

at the

University of Cambridge

1st April 2021

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

It is not substantially the same as any work that I have submitted, or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution.

The length of this thesis is approximately 56,032 words and 48 figures.

This thesis does not exceed the word limit set by the Engineering Degree Committee.

A. P. Neto-Bradley
1 April 2021

A Social Logic of Energy: A data science approach to understanding and modelling energy transitions of India's urban poor

André Paul Neto-Bradley

Abstract

Continued use of traditional solid biomass fuels for cooking in Indian households poses a serious public health risk. Particulate emissions in the form of soot contributed to approximately 600,000 deaths in 2019, a burden that falls disproportionately on women, children, and vulnerable populations. Despite over 95% of the population having access to clean cooking fuel distribution, following recent government initiatives to promote liquefied petroleum gas, biomass cooking fuel use is still widespread. This is the case even in cities, where low-income households have low levels of sustained clean cooking fuel use.

Interventions to promote transition to clean cooking often focus on cost and technology, informed by an economic-technical view of energy transition, but not all households benefit as expected from these interventions. Previous studies on socio-economic determinants of transition offer limited insight into the reasons for why some households can slip through the net of such interventions. The explanation lies in the socio-cultural and economic heterogeneity across households and the inherent spatial inequalities in urban India.

This thesis explores the influence of local socio-economic and cultural factors, and household practices and habits, on clean cooking transition with a view to understanding how the associated heterogeneity can be characterised, and integrated into quantitative energy models and methods. Public national survey and census data is supplemented with primary data collection, which provides valuable quantitative and qualitative data on low-income urban households.

Tree-based regression is used to investigate the influence of socio-economic and cultural factors within quantitative models. Determinants are found to exhibit non-linear trends, with thresholds for change in influence on transition. A statistical clustering reveals different typologies of household amongst clean cooking adopters, indicative of different enabling circumstances and pathways to transition. Continued use of biomass is found to be common across recently transitioned households.

The heterogeneity amongst low-income households, and the emergent transition pathways, are further investigated through data collected on low-income households in Bangalore. A novel method is used which combines mixed data in a two-stage clustering analysis, offering a means to characterise heterogeneity across households, identifying distinct transition pathways and associated barriers. The findings illustrate how wider socio-economic inequality is intertwined with access to sustained clean cooking.

A Bayesian multilevel microsimulation approach is proposed to model the spatial heterogeneity in clean cooking at a city scale. This approach combines publicly available data to generate a synthetic population, and estimates cooking fuel use and fuel stacking using a Bayesian multilevel model. The model takes into account household cooking practices, local spatial effects, and city level economic and policy context. The model reveals how low uptake of clean cooking fuel, and continued biomass use, is related to underlying spatial socio-economic inequalities in cities.

Acknowledgements

There is an old Chinese proverb that say that "A journey of a thousand miles begins with a single step". But the journey is made up of more than just that first step. It is the sum of countless individual steps, some of them backwards, all of them essential to arrive at the destination. The journey of this PhD has had many steps to it and I would be remiss not to acknowledge the many people who supported me along the way.

Nobody has taken more of the steps in this journey with me than my supervisor, Ruchi. She has gone above and beyond merely supervising, becoming a mentor on this journey. Whether supporting my creativity be it in visualising data or experimenting in the kitchen (the latter with the added encouragement of her husband Serge), she gave me the space to explore new ideas while never letting me go too far astray. This project would have never gotten off the ground without her active support, introducing me to interesting people, ideas, and even the Cape Verdian music scene. Thank you for hearing me out and taking a chance on me.

It was through Ruchi that I first met Amir Bazaz, who has in many ways taken on the role of an unofficial second supervisor. His belief in the ambitions of this thesis, and support both through hosting me at the Indian Institute of Human Settlements (IIHS) and constant intellectual stimulation has been instrumental. I often felt he understood my ideas better than I did.

I cannot mention my time at IIHS without thanking Sai Rama Raju Marella & Rishika Rangarajan. Ever welcoming and keen to share or hear a story, Sai went above and beyond to make sure I made the most of my time in Bangalore, whether introducing me to the team around the office over lunch, or travelling the width and breadth of the city to introduce me to regional delicacies. I look forward to getting the chance to raise a glass and break bread with him again. While I may have only met Rishika halfway into this journey, the outputs of this project would have been much diminished without her enthusiasm, curiosity, and constructive criticism. Not to mention her immeasurable patience in putting up with my emails!

Early on I was fortunate enough to get involved with the GENUS interdisciplinary group, which brought together researchers from across the university. There are a few people in particular who I must single out. Charlotte Lemanski has been a supportive and inspirational figure throughout this, always open to listening and offering honest advice. For a fresh take or idea I could always depend on Jaideep Prabhu, whose energy, curiosity, and general knowledge offered food for thought. Over the course of several fieldwork trips the camaraderie shared with fellow GENUS members Ruth Massey and Jiska de Groot made the experience all the more memorable. From them I learned a great deal not only about how to think about research but also have fun while doing it.

I was also fortunate that in the Energy Efficient Cities initiative (EECi) I had a research group that was more of a family than anything else. Whether troubleshooting or debugging some code, sharing a meal or a coffee, or more recently tuning into our weekly group zoom-meetings-turned-daytime-television-show, they are not just a great group of colleagues but great people.

My time at Cambridge has been shaped as much by my research as it has been by the memories. Fitzwilliam College has been my home for these past few years and staff, students, and fellows alike have made it so. There are too many people to thank individually here. But I spare a mention for my college boat club which often offered a welcome change of scenery. When I think of my time in Cambridge I will always remember the moments shared with the boys in my boat, Derek, Pete, Phil, Dan, Adam (the both of them), Stan, Eryk, and Emily.

The final steps in a journey can be some of the most gruelling, with the end tantalising in sight. This thesis would be a far more enigmatic read without the help of my editorial sherpas, Mel, Barbara, and Rory. And to Emily my eternal gratitude for shepherding my soul through this, and returning the favour of walks and food.

Last but certainly not least I am grateful for the love and support of my family who have had my back even when I couldn't decide whether to take that first step in this journey of a thousand miles.

Publications

Aspects of the research in this thesis have been published or submitted for publication as detailed below. Where contents of a chapter have been published or form part of a submitted manuscript this is noted at the start the respective Chapter.

Journal Articles

André Paul Neto-Bradley, Ruchi Choudhary, and Amir Bazaz **(2020)**. "Slipping through the net: Can data science approaches help target clean cooking policy interventions?" en. *Energy Policy* 144, p. 111650. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111650

André Paul Neto-Bradley, Rishika Rangarajan, et al. **(2021)**. "A clustering approach to clean cooking transition pathways for low-income households in Bangalore". *Sustainable Cities and Society* 66, p. 102697. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102697

André P. Neto-Bradley et al. **(2021)**. "Energy transition pathways amongst low-income urban households: A mixed method clustering approach". en. *MethodsX*, p. 101491. ISSN: 2215-0161. DOI: 10.1016/j.mex.2021.101491

André Paul Neto-Bradley, Ruchi Choudhary, and Peter Challenor **(n.d.)**. "A microsimulation of spatial inequality in energy access: A Bayesian multi-level modelling approach for urban India". *Environment and Planning B-Urban Analytics and City Science* In Review: Submitted April 2021 (Preprint on ArXiv). DOI: arxiv.org/abs/2109.08577

Conference Papers

A P Neto-Bradley et al. **(2019a)**. "Applicability of an 'uptake wave' energy transition concept in Indian households". In: *IOP Conference Series: Earth and Environmental Science*. Vol. 294. IOP, Tokyo, p. 012091. doi: 10.1088/1755-1315/294/1/012091

A. P Neto-Bradley et al. **(2019b)**. "Tailoring Residential Energy Provision Strategies in Fast-Growing Cities using Targeted Data Collection". In: *International Conference on Smart Infrastructure and Construction 2019 (ICSIC)*. ICE, Cambridge, pp. 151–160. doi: 10.1680/icsic.64669.151

Contents

List of Figures

List of Tables

Acronyms

BPL	Below Poverty Line
BRT	Boosted Regression Tree
GIS	Geographic Information System
ICAR	Intrinsic Conditional Autoregressive
IHDS	Indian Human Development Survey
INR	Indian Rupees
IPF	Iterative Proportional Fitting
LPG	Liquefied Petroleum Gas
MCMC	Markov Chain Monte Carlo sampling
MFI	Micro-Finance Institution
NSS	National Sample Survey
NUTS	No U-Turn Sampling
OBC	Other Backward Castes
OMC	Oil Marketing Companies
$PM_{2.5}$	Fine Particulate Matter
PMUY	Pradhan Mantri Ujwala Yojana
QDA	Qualitative Data Analysis
SC	Scheduled Caste
ST	Scheduled Tribe
USD	US Dollars

Glossary

Aadhaar	Unique biometric identity system that can be used for verification in transactions and benefit transfers. Can be obtained by Indian residents.
Bayesian inference	Statistical inference based on Bayes' Theorem, updating prior beliefs or data with empirical evidence.
Chulha	A traditional Indian solid biomass stove, styled after the traditional three stones open fire and made of ceramic/earthenware.
Clean cooking fuel	Cooking using a fuel with negligible particulate matter or other emissions hazardous to respiratory health when in use (i.e. LPG/Gas or Electricity)
Clustering analysis	An exploratory/descriptive analysis that seeks to identify groups where instances are more similar to each other than the other groups in a dataset. Sometimes called community detection.

Data science	An interdisciplinary field which brings together scientific methods, statistics, informatics, and data analysis to understand, analyse, and act upon insights from a broad range of data.
Descriptive analytics	Analysis which is concerned with understanding structure and describing the data.
Energy ladder	A concept of energy transition which assumes a preferential hierarchy of fuels which households climb as incomes rise. i.e. Households will switch to a more preferable fuel as incomes rise.
Energy transition	Defined in this thesis with respect to households as the transition from dependence on traditional biomass fuels with harmful emissions, to cleaner and more reliable alternatives such as gas or electricity.
Fuel stacking	The practice of a household using multiple different fuels concurrently, e.g. Using biomass for cooking, and electricity for other needs.
Fuel switching	The practice of a household switching their fuel use for a particular need or all household energy needs.

Global South	A term used to identify low-income countries on one side of the geopolitical North-South divide. This label replaces previously used terminology such as "developing counties" and "third world" which are seen as problematic, but is itself the subject of debate.
Logistic regression	A statistical model that uses a logistic function to model the relationship between determinant variables and a binary outcome.
Markov chain Monte Carlo	A sampling method that allows for estimation of a probability distribution without knowing all the distribution's properties.
Microsimulation	A simulation technique that involves modelling actions or characteristics of each individual in a population.
Multilevel model	A modelling approach that allows separation of within-group effects from between-group effects.
Non-notified slum	A low-income settlement area not officially recognised - inhabitants are not granted tenure rights. May feature temporary construction although this category covers a wide range of settlement types.
Notified slum	Officially recognised low-income settlement area - ensuring tenure rights for inhabitants.

Particulate matter	Particulate matter are microscopic airborne particles, such as soot or smoke emitted from combustion. These particles are small enough to get into the lungs or even bloodstream and have serious negative consequences for health.
Predictive analytics	Analysis that is concerned with extrapolating trends found in the data, often with a view to predicting an effect or outcome.
Prescriptive analytics	Analysis that is concerned with using data to identify interventions, actions, or solutions most likely to achieve a desired outcome.
Sangam	A local name for community-based microfinance institutions/savings clubs, which can serve as an alternative to informal money lenders in low-income communities
Synthetic population	A representative population for a geographic subdivision generated using anonymised data from Census or other sources.
Wage labour	A type of employment in India that is often informal or non-permanent and pays daily or weekly - typical examples would be construction labourers.

1 Introduction

> There is no logic that can be superimposed on the city, people make it, and it is to them ... that we must fit our plans.
> *- Jane Jacobs*

In a low-income neighbourhood in the city of Bangalore a family makes ends meet on daily wages, ever uncertain whether they will still have that source of income a month from now. Their family hasn't always lived here - they migrated to the city in search of a better livelihood, or perhaps their parents did. Surrounded by a burgeoning IT industry in this rapidly growing city, they have LED light bulbs and mobile phones, but they are still burning traditional solid biomass fuels to meet their cooking needs.

The use of traditional solid biomass fuels for cooking is a reality for many Indians, with just under 50% of the population still using solid biomass such as firewood, crop residue, charcoal, or dung to meet their cooking needs (International Energy Agency, 2020a). This is the case despite approximately 95% of the population having access to distribution of clean cooking fuels as of the end of 2019 following recent government policy initiatives to promote the uptake of liquefied petroleum gas (LPG) for cooking (International Energy Agency, 2020b). Urban areas have higher levels of LPG use than rural areas where today a majority of households still use some biomass (D. Sharma et al., 2020; Gould, Hou, et al., 2020), but even in urban areas biomass use persists with an estimated 10-15% of urban households still using biomass as of 2016 (Ravindra et al., 2019). This continued use of biomass in urban areas is most prevalent amongst low-income households (Ahmad and Puppim de Oliveira, 2015), who often face spatial inequality in their access to utilities (Bhan and A. Jana, 2015).

Use of solid biomass fuels in cooking stoves releases pollutants including particulate matter in the form of soot, and carbon monoxide (Pope Daniel et al., 2015; K. R. Smith, Bruce, et al., 2014). Solid biomass fuels account for an estimated 68% of total final energy consumption in the residential sector in India[1] (International Energy Agency et al., 2019), with this sector accounting for approximately 54% of fine particulate matter ($PM_{2.5}$) emissions in India (International Energy Agency, 2019). Emissions from cookstoves have been associated with increased risk of cardio-pulmonary diseases (Arku et al., 2018; Ranjan and Singh, 2020; Assad et al., 2016; P. Johnson et al., 2011). As a result solid biomass cooking fuels are a key contributing factor to annual deaths in excess of 600,000 from indoor air pollution in India (Pandey et al., 2021), and women and children face greater risk of exposure to indoor air pollution (K. R. Smith, Bruce, et al., 2014; Rohra and Taneja, 2016).

The use of solid biomass fuel has a wider impact on sustainable development, in particular hindering female empowerment - not only through women facing greater exposure to air pollution, but via the time and unpaid labour of collecting or carrying fuel, and the ingraining of stereotypes about the role of women (Brown and Lankford, 2015). Numerous studies have also linked solid biomass stoves to increased risk of Acute Respiratory Infection (ARI), and in turn mortality, in children who often are carried or kept around by their mothers or grandmothers tasked with the cooking (Joon and Chandra, 2016; Oguonu et al., 2018; Mandal et al., 2020). Longer term complications such as asthma have also been shown amongst

[1] Solid biomass fuel is used less efficiently than other fuels and so greater amounts need to be used to deliver the same useful energy. The proportion in the final energy consumption mix does not reflect the proportion of residential energy services/needs met.

children under the age of five who have been exposed to biomass cooking smoke (Oluwole et al., 2017). Improving access and provision of clean modern energy is intrinsically linked to socio-economic development (Kanagawa and Nakata, 2008). Indeed studies in India have shown that cleaner cooking technology reduces time spent cooking and collecting fuel which can have wider positive impacts for the women who often carry out these activities (Brooks et al., 2016), and children who grow up in homes without biomass fuel use have been shown to be half as likely to suffer from ARI (Mishra et al., 2005).

In the framework of the current Sustainable Development Goals (SDGs), access to clean and reliable energy is one of the 17 goals, but energy can directly or indirectly be an enabler of the other 16 goals with many theoretical synergies to be exploited (Pradhan et al., 2017). Improving clean cooking access is not just a public health issue but is also a keystone for socio-economic development. Recent studies have shown that synergies between SDG 7, which addresses provision of reliable and clean energy for all, and the other SDGs far outweigh any trade-offs (Fader et al., 2018). Indeed Fuso Nerini et al. (2018) detailed synergies with 143 of the 169 targets of the SDGs, noting the important role of energy in efforts to end poverty, improve healthcare provision, deliver reliable water and sanitation, and improve household incomes.

Promotion of LPG as a clean cooking fuel in India may seem at odds with a sustainable development agenda. LPG is a fossil fuel and its use does lead to more net CO_2 emissions than combustion of biomass fuels. However, it is considered a clean cooking fuel seeing as it offers a near total reduction in exposure to particulate matter emissions when compared to equivalent use of biomass fuels (Pillarisetti et al., 2019). As discussed by Sokołowski (2019) India's current policy on residential clean cooking represents a compromise. In the short term the aim is to address the immediate health risk, and the many associated sustainable development opportunities of moving away from solid biomass fuel use. This is done with the understanding that residential energy use including cooking can and will be decarbonised in the medium to long-term (2030s to 2040s) as renewable electricity generation capacity is added. In this thesis clean cooking will refer to cooking using fuels that do not produce particulate emissions (primarily LPG), regardless of their associated CO_2 emissions.

1.1 Clean cooking in India

Addressing the widespread use of traditional biomass fuels in India has been a focus of government policy for several decades although the focus has shifted from initiatives supporting adoption of improved cookstoves, to more recent efforts to promote the adoption of LPG (C. Jana and S. C. Bhattacharya, 2017). As the policy approach has shifted so has the prominence of this issue in the public sphere and accordingly recent initiatives have seen increased levels of funding and become flagship government policies (K. Smith, 2018).

1.1.1 Improved Chulhas

The main problem with the traditional chulha (biomass stove) is its inefficient and incomplete combustion of solid biomass fuels, which leads to high levels of particulate matter emissions and high consumption of solid biomass (M. Bansal et al., 2013). Efforts to promote clean cooking in India initially focused on improving the efficiency of these stoves, with the benefit of being a potentially cost effective way to address the problem using the existing fuel and thus requiring no new supply infrastructure. 'Smokeless' chulhas were first introduced in the 1950s, and in 1983 India launched the National Programme on Improved Chulhas (NPIC) to promote the uptake of improved cookstoves, however the programme had limited success and was eventually discontinued in 2002 (Kishore and Ramana, 2002). A re-boot of this initiative named the Indian National Biomass Cookstoves Initiative (NBCI) was launched in 2009 with the aim of getting high-efficiency stoves to 160 million households (Brooks et al., 2016). Efforts have also been furthered through the commercial approaches and social enterprises such as with the 'Oorja' stove - an improved biomass cookstove developed and marketed by a subsidiary of BP (Thurber et al., 2014).

Despite these initiatives over several decades, improved chulhas seem to have had little effect on the problem of indoor air pollution mortality, with use of improved chulhas remaining low (C. Jana and S. C. Bhattacharya, 2017; K. R. Smith and Dutta, 2011). K. R. Smith and Sagar, 2014 explain that improved cookstoves do not always match with cooking needs and practices which has discouraged uptake. Even when used these improved stoves do not eliminate emissions entirely, and by working with highly variable solid fuel the emissions reductions may not be as designed. Research has shown the reduction in emissions of two or three-fold versus a traditional stove is not enough to meaningfully reduce risk of mortality which varies non-linearly with exposure to emissions (Burnett, Richard T. et al., 2014).

1.1.2 LPG and Pradhan Mantri Ujjwala Yojana

While LPG has been used as a cooking fuel in India since the mid 20th century, it was mainly a fuel used by the upper classes, despite untargeted subsidies being in place (Gould and Urpelainen, 2018). As discussed above, clean cooking efforts largely focused on improved chulhas with subsidies for kerosene also in effect. But this has changed over the past decade with the Government introducing Pradhan Mantri Ujjwala Yojana (PMUY), a flagship policy aimed at getting 80 million households below the poverty line connected to LPG (Khan, 2017). This is in fact part of a trio of initiatives at play: The PAHAL scheme ensures LPG subsidies are paid directly to household's bank accounts, reducing illegal non-residential use of subsidies; The Give-it-up campaign encourages middle class households to transfer their subsidies to a poor household; which in turn has helped finance the PMUY subsidy scheme (K. Smith, 2018).

For a household transitioning to LPG there are two key costs: that of the connection which includes the first cylinder and stove (between INR 3000-7000, ca. USD 40-95), and the recurring cost of refill cylinders (between INR 700-900, ca. USD 10-12). Through the PMUY programme the government bore about half of these connection costs for below poverty line households (Kar et al., 2019). Additionally, households are also entitled to a subsidy of INR 200-300 (ca. USD. 3-4) per refill which is transferred to their bank account, however only after financing of the upfront subsidy is recovered by the gas companies (Aggarwal et al., 2018). This takes approximately 5-7 refills, and until this happens households pay the higher non-subsidised refill costs. A key feature of this policy is that it seeks to empower women, by requiring the

Current policy is based on the assumption that the barrier to clean cooking is cost, in this case both the upfront cost addressed by a subsidy for the connection costs and the running costs with a subsidy on offer for a refill. The government reports that PMUY has been successful in its goal of connecting over 80 million, however recent studies evaluating early roll out have noted some problems including not all households sustaining their use of LPG (Ranjan and Singh, 2020), not all truly poor households accessing the benefits (K. Smith, 2018), and high rates of concurrent biomass use (Mani et al., 2020). Some implementation challenges in the initial roll out included poor distribution among villages in certain areas limiting uptake through the scheme (Aggarwal et al., 2018), and the criteria for eligibility present a barrier to some households without suitable documentation or formal banking. Patnaik and Jha (2020) also note though that while recognising gender imbalances, the PMUY scheme reinforces the role of women as the primary cook of the household. Subsidies are paid out to the women of the household although they are not necessarily involved in the decision to purchase LPG. These outcomes and unintended consequences are hallmarks of the complexities of energy transitions.

1.2 Energy transition and the energy ladder

Energy transitions can refer to a number of phenomenon surrounding energy systems and energy use. In this thesis we shall focus on residential energy transitions, which refer to the transition of households from using traditional biomass fuels to cleaner more 'modern' alternatives like gas and electricity. Uptake of clean cooking represents such a transition, although the term can refer to transition away from traditional fuels for other energy needs too. While the focus of this thesis is clean cooking transition in India, the approach and methods have the potential to be applied to other residential energy transition challenges. On a note of terminology this thesis will often refer to energy modelling and residential energy use, on the understanding that cooking fuel use and modelling of cooking fuel use is a subset of these. While there is a substantial body of research on clean cooking specifically, methods and empirical research used in residential energy research more widely can be relevant and useful to a clean cooking context.

Policy approaches to addressing access to clean and reliable energy have largely drawn from conceptual models of residential energy transition rooted in quantitative economic and technical approaches. These can be traced back to the Energy Ladder conceptual model formulated by Leach (1992). It assumes that there exists a hierarchy of energy technologies (or fuels) as steps or rungs on the path to clean 'modern' energy use. Accordingly, different fuels on the ladder are ranked by perceived household preference in terms of cleanliness, convenience, and efficiency (Hiemstra-van der Horst and Hovorka, 2008). Underpinning this understanding of energy transitions is the assumption that individuals are rational utility-maximising consumers (Kroon et al., 2013), implying that as income rises they will switch to using ever more 'modern' fuels and associated technologies which are higher ranked on the ladder (Richard H. Hosier and Dowd, 1987).

This perspective on energy transition is problematic as it leads to policies which are overly focused on cost and are divorced from the needs of people and their households. This is not to suggest that addressing cost is not important. Indeed, recent studies make it clear that the upfront cost subsidies of the PMUY scheme for LPG connections has helped millions of households access LPG (Ranjan and Singh, 2020). But what is also clear as pointed out by Gould and Urpelainen (2018), is that this is only a first step and work remains to eliminate the use of solid biomass fuels. Long-term health and development benefits of clean cooking can only be achieved by full displacement of biomass fuels, which requires better understanding the needs and behaviour of households (Ruiz-Mercado and O. Masera, 2015).

Kroon et al. (2013) argue that the linear cost-driven view of the energy transition is an oversimplification, for example omitting the possibility of a household using multiple different fuels in parallel. This phenomenon of multiple concurrent fuel use is known as fuel stacking and is especially salient in the case of cooking fuels. O. R. Masera et al. (2000) addressed the importance of fuel stacking in understanding energy transition, highlighting how different fuel stacking practices result in different pathways to sustained use of clean fuels. Fuel stacking also implies that households that adopt clean cooking fuels will not necessarily abandon biomass fuels which can jeopardise the health and sustainable development aspirations of promoting clean cooking fuel.

In this thesis the term ‘transition pathway’ will be used to refer to the sequence of decisions, fuel consumption patterns, and cooking related energy use practices of a household in the course of switching from exclusive use of biomass fuels to exclusive use of clean cooking fuels. There are different transition pathways in a given area, and each pathway is defined by different challenges or barriers to access of clean fuel for the household following a particular pathway, which can lead to fuel stacking as households get stuck ’between fuels’. For example migrant workers in informal settlements lacking proper documentation to avail of LPG subsidies, relying on insecure wage labour as their main source of income, may resort to fuel stacking with collected biomass to supplement any non-biomass fuel they can buy. Their informal living arrangements and insecure daily wages pose barriers to sustained clean cooking fuel use. Understanding these barriers and challenges is crucial, but as demonstrated by Cheng and Urpelainen (2014) fuel stacking cannot simply be understood as an economic phenomenon. There is a need to look beyond income to understand residential energy transitions.

1.3 The social dimension

In his landmark book entitled ‘A Social Logic of Space’, Bill Hillier wrote that we "read space and anticipate a lifestyle" (Hillier, 1984). He was writing about what became known as ‘space syntax’, the idea that there is a logic to how we arrange space related to how we use it, there is a grammar that governs this at a local scale and by default people will arrange their space in certain preferred configurations. This idea is relevant in the sphere of energy use. To rewrite Hillier’s statement in terms of energy, we would ‘read energy use and anticipate a lifestyle’. Use of energy is related to lifestyles and often by knowing something about the habits and circumstances of an individual or household their choices and actions can be better understood. All too often habits, practices, and local socio-cultural context of the individual is not considered in the design of technology and policy aimed at delivering cleaner, more reliable, and more efficient energy to all (Benjamin K. Sovacool, 2011; Shove and Walker, 2014).

Approaches to understanding and characterising clean energy transition are often separated by disciplinary divides. In particular a divide between quantitative economic and technical approaches which assume rational utility maximisation and focus on predicting trends, and more exploratory qualitative and descriptive socio-technical and practice-based research approaches which focus on actors, behaviours, and local social factors. Geels et al. (2016) discuss how these different approaches are underpinned by different ontological assumptions which make integration of these approaches difficult.Turnheim et al. (2015) propose that a structured dialogue between these approaches is possible based on alignment of framing, bridging of data and metrics, and a process of iterative interactions with insights from different approaches informing the others.

The opportunities for bridging quantitative modelling approaches and socio-technical and practice based approaches in the study of residential energy transition, and the methodological gaps this leaves, might be best understood in terms of the components that form the focus of research on energy transition. Developing an understanding of energy transition involves people, data, and models, and Figure 1.1 shows a conceptual illustration of the relationship between these three components. While these three components in their own right are the subject of study, efforts to understand clean energy transitions can be characterised by the intersections between these components. The intersection of models and data is representative of the predictive quantitative analysis common in economic and engineering approaches. At this intersection the focus is on fitting models to the data at an aggregate scale, typically using regressions to test hypotheses on determinants of clean fuel use, relying on quantity of data to provide statistical power and make predictions at an aggregate level.

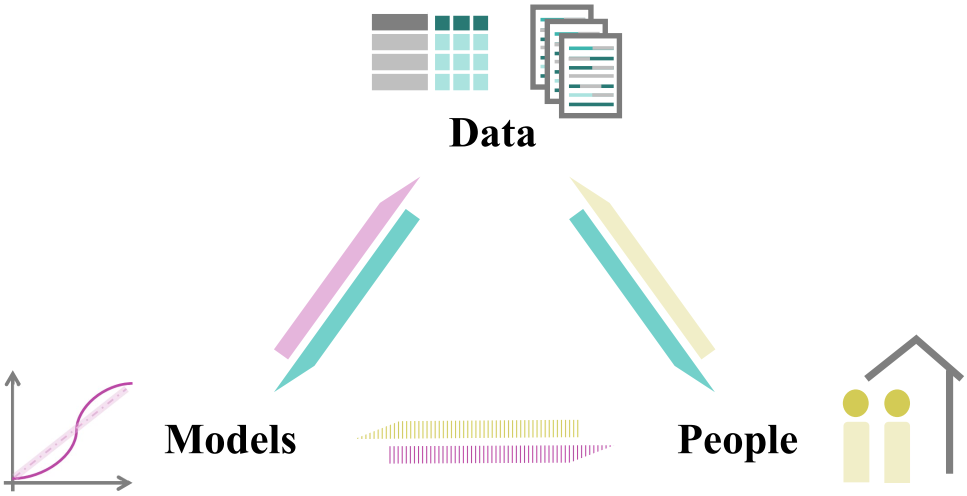


Figure 1.1: Conceptual illustration of relationship between focal components of residential energy transition studies. The arrows indicate the links and intersections between these, each characterising an approach to studying energy transitions.

The intersection of data and people represents descriptive analyses, often using socio-technical and practice-based qualitative approaches although sometimes including quantitative methods. These focus on qualitative methods and case studies, for example case studies of solar mirco-grids or cooking fuel practices in rural villages. By focusing on data collection from individual households these studies can explore individual narratives and practices, although insights can be context-specific and short-term (Geels et al., 2016). Some recent studies have explored the use of mixed methods approaches, which attempt to combine elements of statistical power of predictive quantitative models, with the information and explanatory ability of qualitative descriptive analysis in the context of a case study.

The remaining intersection in Figure 1.1 is that of models and people. Integrating models and people focuses on the needs, circumstances, and variability at a household scale. This can enable a descriptive prediction which distinguishes local effects of social, cultural, and economic factors on household energy transition, while also providing insight into spatial inequality at an urban scale. This approach enables the inclusion of a social dimension to the model. By modelling at a household level such an approach has a more intuitive correspondence with qualitative approaches focused on individual and local narratives. This also offers an avenue for the bridging that Turnheim et al. (2015) refers to, with an alignment of framing and data that can facilitate interpretation by stakeholders, and contribute to an inclusive process to designing, tailoring, and targetting interventions to promote clean energy transition. Unlike the intersection of models and data, and data and people, there are no studies to date on residential clean energy transitions in India which meaningfully integrate models and people.

1.4 Aims, objectives and research questions

This thesis considers how to bridge quantitative energy modelling and socio-technical approaches to enable a constructive dialogue, by considering the research question of *how can the influence of social factors and household practices on clean cooking transition be integrated into a quantitative energy model?* The underlying hypothesis being that a household's response to energy transition policies and incentives is conditioned by local socio-cultural circumstances and practices of the household, the interaction of which defines the energy transition pathway of a household. A logical implication of this statement is that different households can follow different transition pathways which are a function of the household's characteristics, practices, and local socio-cultural context.

An answer to this research question requires contributing to the knowledge on several key aspects in the field of residential energy transitions, including:

- Identifying non-income determinants of energy transition and the nature of their influence;
- Analysing limitations of current logistic regression models in predicting the influence of determinants on transition;
- Quantifiable characterisation of heterogeneous energy transition pathways using qualitative and quantitative data;
- Modelling the interaction between local socio-cultural context and household practices and circumstances related to energy transitions at a micro-scale;
- Accounting for sources of uncertainty in modelling and predicting residential energy transition.

Previous quantitative studies provide some insight into the influence of socio-economic characteristics of households, and case studies and qualitative approaches in studies across the Global South have shed light on individual practices and narratives of transition by households. However previous studies do not contextualise findings in terms of limitations of current quantitative predictive models, nor quantify implications of socio-cultural context and household practices for transition pathways, and no systematic consideration is given to uncertainty.

1.4.1 Research Objectives

This thesis aims to address the research question through the following objectives:

1. Understand the influence of socio-economic and cultural determinants of clean cooking transition, exploring and characterising the limitations of current quantitative predictive approaches in fully describing this.

2. Explore and characterise the heterogeneity in energy transitions among low-income households using mixed data and methods.

3. Develop a modelling approach capable of integrating the influence of, and uncertainty arising from heterogeneity in local socio-cultural context and household habits on clean cooking transition.

Each of these objectives addresses one of the intersections of models, data, and people shown in Figure 1.1. The first aim looks at the use of current quantitative predictive regression models and survey data for characterising socio-economic determinants of transition. The second aim is concerned with the intersection of data and people, using mixed household level data to characterise clean cooking transitions followed by people. The final aim addresses the gap in knowledge surrounding models and people by developing a modelling approach which integrates insights from addressing the preceding objectives.

1.4.2 Case Study

While exploring the issues relevant to low-income urban households across India, the case study region for this thesis is the southern Indian states of Karnataka, Tamil Nadu, and Kerala. Within this five cities were selected for data collection and modelling. The city of Bangalore is used as a case study in addressing the second research objective with both quantitative survey data collection and household interviews conducted. The cities of Coimbatore, Kochi, Tiruchirappalli and Trivandrum are used as test cases for the modelling approach developed in response to the third research objective. The rationale for choice of this case study region and selection of cities is discussed in Chapter 3.

1.4.3 Thesis Structure

This introduction is followed by Chapter 2 which provides background on key concepts related to energy transition, a review of quantitative and qualitative research characterising clean cooking transitions in India, and further details on India's clean cooking policy. The different methods and data required to address the research objectives call for an range of data sources and methods, and Chapter 3 provides an overview of the methodological strategy as well as detailing both primary and secondary data sources, and the choice of case study area. The research objectives are directly addressed by each of the three analytical chapters, Chapters 4, 5, & 6.

An analysis of determinants of clean cooking transition using a public panel dataset is presented in Chapter 4, which addresses research objective 1. Using the Indian Human Development Survey (IHDS) panel dataset an ensemble machine learning regression method is used to identify determinants of transition away from a biomass stove. The results are compared to a conventional Probit regression commonly used in the field and the predictive performance is assessed against a subset of data not used to fit the model. These are further explored through a descriptive clustering of LPG switching households highlighting the different typologies of transitioning households, and the bias of regression predictions towards some of these.

To address research objective 2 and develop a better understanding of clean cooking transition pathways amongst the urban poor, Chapter 5 details the results of a mixed method clustering analysis using survey data and interviews collected in Bangalore. Data from low-income households in seven different wards of the city are analysed using quantitative clustering and qualitative data analysis. A discussion of findings and commonalities between the quantitative community detection and the themes and concepts identified through the interviews outlines defining issues for clean cooking transition amongst this understudied population. A novel two stage clustering approach is used to match qualitative themes to the quantitative clusters to define distinct pathways for clean cooking transition.

An approach for modelling the effects of household practices and habits and local socio-cultural factors is proposed and demonstrated in Chapter 6 as a response to research objective 3. This chapter combines methodologies more commonly used in epidemiology and transport research to develop an urban household energy transition microsimulation model. This method combines multiple sources of publicly available data to estimate cooking fuel use and stacking between city wards, highlighting signs of differing transition pathways and barriers. By using a Bayesian multilevel model, uncertainty in inputs and specification are propagated through the model and factored into estimates. A demonstration of the utility of the model and its advantages and limitations concludes this chapter.

Finally, Chapter 7 shall discuss the contributions and outcomes of this work, with consideration of limitations and opportunities for further work.

1.5 Transcending Disciplinary Divides

The nature of the research question in this thesis transcends disciplinary divides, and that means fulfilling the aims stated above requires working at the intersections between engineering and other relevant disciplines. This can be viewed as a journey through the periphery of the engineering discipline to gain new perspective, insights, and cogent knowledge from related disciplines for quantitative energy modelling of clean cooking transitions. Figure 1.2 illustrates the conceptual journey of this thesis, intersecting with economics, social sciences, and business & policy. The analysis and contribution of each chapter accumulates insight from other disciplines as they pertain to quantitative modelling of energy transition, culminating in the development of a modelling approach that integrates household habits and local socio-cultural effects. The transdisciplinary nature of this work and the methodological strategy adopted to carry this out will be further addressed in Chapter 2 and 3.

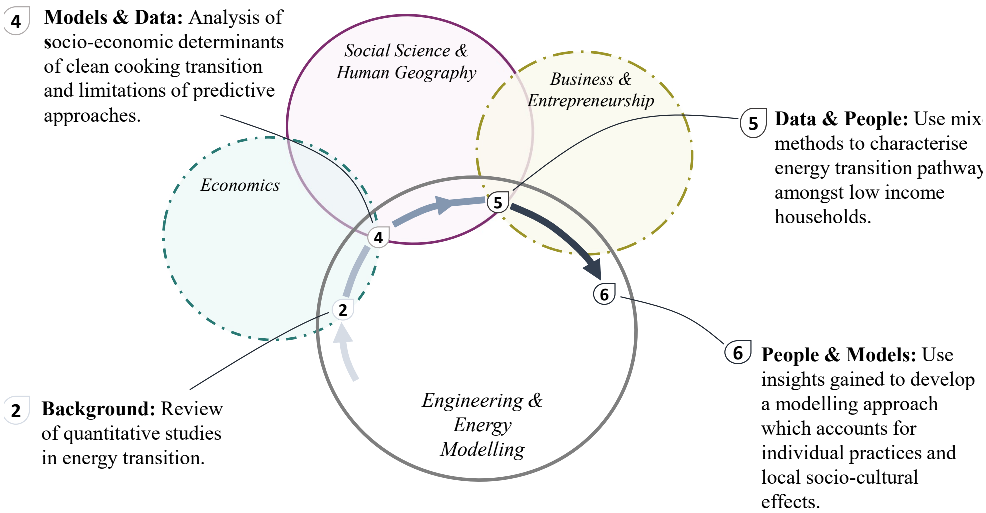


Figure 1.2: An overview of the journey of the analysis in this thesis, conceptualised as it relates to intersections between energy modelling and engineering and related energy research disciplines. The numbers refer to the relevant chapter.

2 Background

2.1 Introduction

The transition to clean cooking amongst India's urban poor is the focus of this thesis. Accordingly, consideration must be given to relevant conceptualisations and findings across different disciplines. Indeed Stephenson (2017) makes the point that single-discipline approaches to energy research may be ineffective. Imagine the simplistic example of trying to forecast monthly residential energy demand based on household incomes, without taking into account seasonal variation in demand for lighting and heating. The results would likely not be very useful. The structured dialogue between disciplines on energy research as envisioned by Turnheim et al. (2015) argues that it must rely on understanding of cogent knowledge from other disciplines to allow for alignment of framing, so that data from models and analyses can bridge the disciplinary divide. This chapter will thus review energy transitions from both a quantitative economic-technical point of view as well as considering concepts and findings from the social sciences.

Energy research in its broadest sense is concerned with addressing complex sustainability challenges, and clean cooking transition is a prime example of such a complex challenge. Complexity requires transcending not only traditional disciplinary approaches, but also interdisciplinary approaches focused on a division of labour between disciplines (Baumgärtner et al., 2008). Lang et al. (2012) describe such a transdisciplinary approach as involving collaborative problem framing, co-creation of solution-oriented knowledge, and applying that knowledge to the real world case. In such an approach the cogent knowledge Turnheim et al. (2015) refers to not only comes from academic disciplines, but from stakeholders whose input is crucial to the framing of the problem and understanding how best to deliver solutions. This participatory research approach can help ensure greater usability of findings although it is more time consuming and can clash with rigid institutional norms (Polk, 2015).

The approach taken in this thesis embodies such a transdisciplinary approach leveraging participatory data collection and an iterative approach to research phasing, to allow for outputs and policy recommendations to be better informed by, and suited to, the context of low-income urban households. This thesis takes as input cogent knowledge from other academic disciplines as reviewed in this chapter, but also from non-academic sources through the participatory data collection strategy and iterative feedback loops built in to the methodological strategy discussed in Chapter 3.

This chapter shall begin by introducing conceptualisations of residential energy transition in section 2.2. This will include both those rooted in quantitative economic-technical approaches, as well as briefly considering social science approaches to conceptualising residential energy transition. Section 2.3 will review findings from previous studies on residential energy transition across the Global South and in India in particular. Previous studies will be considered in the context of how and why they use their data, distinguishing between quantitative studies in subsection 2.3.1 and qualitative studies in subsection 2.3.2. Special consideration will be given to recent finding on clean cooking transition in India. Finally, in section 2.4 the Indian context shall be considered in more detail, not only with respect to clean cooking policy and access, but also the market structure and possible concerns about the policy.

2.2 Residential Energy Transitions

2.2.1 Overview

Different conceptual models have emerged to explain and describe a mechanism of household energy transition from traditional biomass fuels to cleaner more 'modern' fuels such as gas and electricity. Two prominent and long-standing models rooted in a quantitative economic perspective on transition will be introduced, before discussing more recent concepts developed in the social sciences.

To understand these conceptual models it is necessary to understand two important energy transition behaviours materially significant to residential energy transition: fuel switching and fuel stacking. Fuel switching refers to a household behaviour whereby for a given energy service or need of the household, the use of one fuel is discontinued and replaced by the use of a different fuel. Fuel switching is often considered to be unidirectional from traditional biomass fuels through to 'modern' electricity and oil and gas fuels, although it is recognised that this is not always necessarily the case (Heltberg, 2004). Fuel stacking is another strategy whereby households take on use of new and different fuels alongside those they already use. While continuing to use the existing fuel, the new fuel is used to deliver new or additional energy services (O. R. Masera et al., 2000).

The idea that energy transitions can, to a large extent, be understood by the dynamics of these two fuel use behaviours in developing countries arose from a series of observations in the 1980s regarding fuelwood use and the so-called 'fuelwood crisis' (Hiemstra-van der Horst and Hovorka, 2008). Empirical evidence indicated a link between economic development and use of modern petroleum fuels and electricity, and a shift away from biomass (Richard H. Hosier and Dowd, 1987). Furthermore, the observed differences across income groups and the urban-rural divide supported the notion that a switch or stacking from traditional biomass to modern fuels is simply an outcome of economic growth and urbanisation (Leach, 1992).

2.2.2 Energy Ladder

The 'Energy Ladder' concept as a hypothetical representation of the transition of household energy use from traditional fuels such as crop residue, animal waste, or firewood to 'modern' fuels such as electricity and petroleum products was formally postulated by Leach (1992), although the idea had been in existence before this (Richard H. Hosier and Dowd, 1987). The concept is illustrated in Figure 2.1. It assumes that there exists a hierarchy of energy technologies or fuels as steps on the path to clean 'modern' energy use. Accordingly, different fuels on the ladder are ranked by perceived household preference in terms of cleanliness, convenience, and efficiency (Hiemstra-van der Horst and Hovorka, 2008). Indeed, both fuel switching and appliance ownership studies have found that there is a preferential hierarchy of fuels and appliances (Richard H. Hosier and Dowd, 1987; Leach, 1992; B. S. Reddy, 1996; Khandker et al., 2010; Cabeza et al., 2014), and both B. J. v. Ruijven et al. (2011) and Rao and Ummel (2017) found a unique and distinct preferential hierarchy of appliances in Indian households. This view of energy transitions assumes that households behave like utility-maximising neoclassical consumers: i.e. they will choose the option with the best cost-benefit trade-off within their means, implying that as income rises they will switch to using ever more sophisticated fuels and associated technologies which are higher ranked on the ladder (Richard H. Hosier and Dowd, 1987).

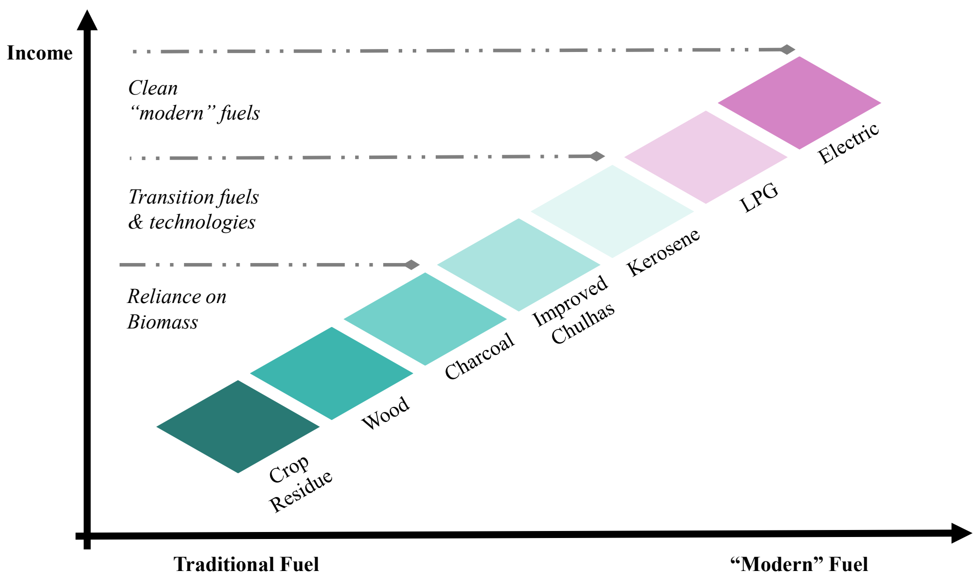


Figure 2.1: A conceptual illustration of the energy ladder concept of energy transition, adapted from Leach (1992)

However this original 'Energy Ladder' concept has certain limitations which have been discussed extensively in the literature. Firstly, and as pointed out by Kroon et al. (2013), the development of the concept was rooted in the need to progress from biomass to so called 'modern' fuels such as electricity and LPG, with the resulting oversimplification that biomass fuel is the 'fuel of the poor'. While a range of studies have found income to be an important factor in energy transitions (Richard H. Hosier and Dowd, 1987; R. H Hosier and Kipondya, 1993; Narasimha Rao and B. S. Reddy, 2007; Shonali Pachauri and Jiang, 2008; Rahut et al., 2016b; Rahut et al., 2016a), it is not the sole factor influencing the uptake of modern fuels. An implication of a hierarchy of fuel preferences in the ladder model was that energy transitions were driven "not by an emerging desire for modern fuels so much as by socio-economic changes, which help to break the constraints on their wider use" (Leach, 1992). While these constraints can include income, other factors related to the practicalities of access to the fuel can also be an issue. Leach mentions that the ‘larger lumpy payments’ for LPG (a standard 14.2kg cannister is usually approximately a month worth of fuel for a household) can lead households to not use LPG despite it being within their income range, and potentially cheaper in the long run. Furthermore, the energy ladder model assumes a clean switch between fuels whereby households use one technology at a time as they climb the ladder. This is somewhat problematic as it assumes fuel switching behaviour is dominant in energy transitions. In reality many households use multiple fuels to meet their energy needs, thus fuel stacking behaviour must also be considered (Saatkamp et al., 2000).

2.2.3 Fuel Stacking

The practice of fuel stacking, sometimes also referred to as multiple fuel use, involves a household using more than one fuel simultaneously, for example using both an LPG and biomass stove to meet cooking needs (Heltberg, 2004). Fuel stacking is particularly important to understanding residential cooking fuel use. A range of case studies from across the Global South have found fuel stacking to be a widespread practice with respect to cooking (Choumert-Nkolo et al., 2019; Medina et al., 2019; Ruiz-Mercado and O. Masera, 2015; Perros et al., 2021), and an obstacle to addressing the negative health impacts of biomass fuel use as many households continue using some biomass fuel after adopting a cleaner alternative (Shankar et al., 2020; Jewitt et al., 2020).

Ruiz-Mercado, O. Masera, et al. (2011) offer a rationalisation for cooking fuel stacking rooted in the literature on diffusion of innovations (Rogers, 1983). They view the adoption of a new fuel or stove as a process of innovation adoption, or more specifically a process of adoption of new cooking practices. This involves a complex web of interactions between the user, technology, fuel, and larger socio-economic contexts. The newly adopted stove will likely be introduced for the cooking practices which it is perceived to be best suited to, and old stoves are only phased out after an adjustment period in the medium to long-term. Multiple fuel use can also give the household flexibility, for example enabling cooking of multiple elements of the meal simultaneously (Perros et al., 2021), or to help manage risks with respect to finances and availability not unlike having a diversified investment portfolio (Choumert-Nkolo et al., 2019). Despite the importance of this phenomenon in understanding clean cooking transitions, Heltberg (2004) noted that it can often go uncharacterised in public datasets as many surveys only asked about primary cooking fuels.

2.2.4 Multiple Fuel Model

O. R. Masera et al. (2000) built upon the energy ladder taking account of households fuel stacking and using a mix of different fuels across different energy services, in a so-called multiple fuel model that is illustrated in Figure 2.2. Unlike Leach's energy ladder, the multiple fuel model accounts for fuel stacking, and sees households having a portfolio of energy services and fuels used to deliver them. Using data on household cooking energy use in Mexico, O. R. Masera et al. (2000) showed that as these households took on new more modern fuels they also continued to use traditional cooking methods and fuels simultaneously. Empirical evidence has found fuel stacking to be quite a common practice in low-income households where, for example, one or two alternative fuel options will be kept for cooking (Richard H. Hosier and Dowd, 1987; R. H Hosier and Kipondya, 1993; Saatkamp et al., 2000; B. M. Campbell et al., 2003; Baiyegunhi and Hassan, 2014). It can be a strategy for managing fuel availability, hedging against the risk of unreliable supply or unstable pricing of the newer fuel (Leach, 1992; R. H Hosier and Kipondya, 1993; Hiemstra-van der Horst and Hovorka, 2008). This continued use of previous fuels may also accommodate certain practices or habits that are best suited to these. Although this model does incorporate more realistic complexity of household energy use and transition, it is still focused on individual decision making (Yadav et al., 2021).

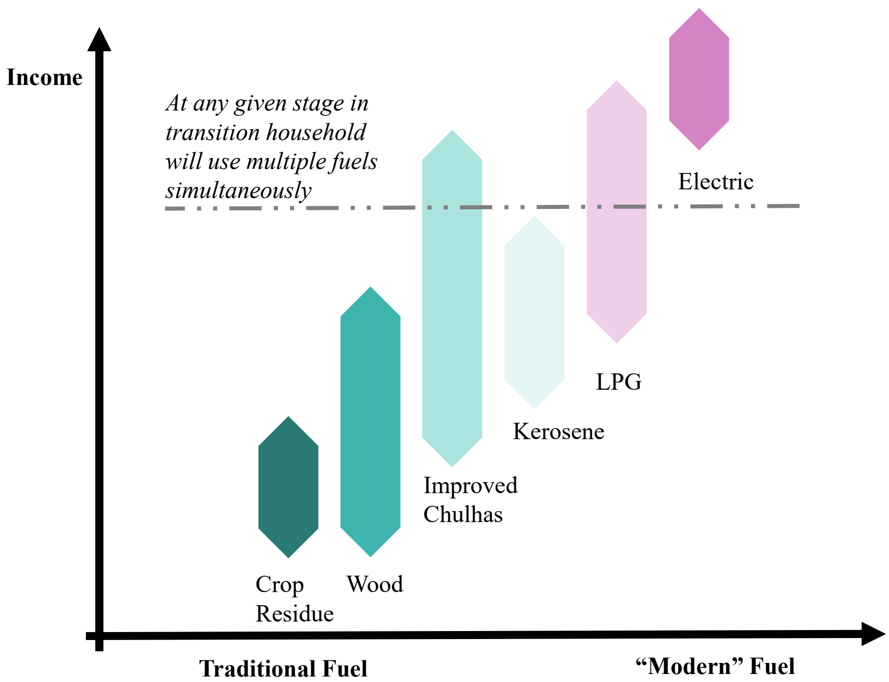


Figure 2.2: A conceptual illustration of the multiple fuel model of residential energy transition, adapted from O. R. Masera et al. (2000) and Yadav et al. (2021).

2.2.5 Social Science Approaches

Research continues to show that, in aggregate, there is a positive correlation between modern fuel use and income (Davis, 1998), although it is clear that there are other factors broaching social, cultural, spatial, and political dimensions that require further consideration (O. R. Masera et al., 2000; Hiemstra-van der Horst and Hovorka, 2008; Castán Broto et al., 2020). Few quantitative studies give thorough consideration to these non-income factors, beyond viewing these as determinants of fuel choice. Part of this is explained by the lack of widespread data on these factors which can make study of these difficult (B. M. Campbell et al., 2003; B. K. Sovacool et al., 2015). In any case many studies based on the energy ladder perspective show some contradiction with the basic assumptions of the model, contradictions that are often attributed to unexplained local factors (Hiemstra-van der Horst and Hovorka, 2008). For example, some studies have found that high income households in both urban and rural regions do not reduce their biomass use with increasing income (R. H Hosier and Kipondya, 1993; Kebede et al., 2002; B. M. Campbell et al., 2003; Khandker et al., 2010), while others have found that even the lowest income groups can make regular use of electricity and gas (Davis, 1998).

Studies using the energy ladder concept to study determinants of transition, such as Karekezi and Majoro (2002), have noted that some households have preferences due to cultural factors or matters of convenience and fit with the household's lifestyle and circumstances. The question of the 'problem of people in energy research' was discussed by Nader and Milleron (1979) as well as Stern (1986) who note that focusing on a narrow range of economic and technical determinants can lead to incorrect understanding of household decisions and behaviours. Energy provision also serves to satisfy a social need and this is often overlooked (Benjamin K. Sovacool, 2011); the economic and technical focus of concepts like the energy ladder and even the multiple fuel model do not offer a framework to consider the role of this social context.

Some interdisciplinary efforts have sought to integrate the social need for energy into the energy ladder, such as the energy services ladder postulated by Benjamin K. Sovacool (2011). This proposed expansion of the energy ladder through consideration of the related energy services begins to capture some aspects of social need for energy. Unlike water, food, and medicine, energy can not be directly consumed by humans. Indeed, it is the services that can be delivered that are of interest. When energy services are considered together with the energy ladder concept, different income groups are classed as using a particular mix of fuels and appliances to deliver energy services, as is summarised in Table 2.1. Across different income groups the motivation for their energy use changes, reflecting the changing socio-economic and cultural context of the households (Benjamin K. Sovacool, 2011).

Household Type	**Primary fuels**	**Primary energy services**	**Driving factors**
Low-Income	Wood, dung, kerosene, charcoal, biomass	Cooking and lighting, radio, mobile phone charging, television, space heating, hot water	Satisfying subsistence needs
Middle-Income	Electricity, LPG, Kerosene, Natural Gas, Fuel oil	All low-income services, plus space heating and cooling, entertainment and lighting, refrigeration, clothes washing and drying, IT	Convenience, comfort, and cleanliness
High-Income	Electricity, Natural Gas, Fuel oil	All middle-income services, air-conditioning in all rooms, bathroom with heated seat, sound system, multiple televisions	Conspicuous consumption and social signalling

Table 2.1: Sovacool's urban household energy services ladder. Adapted from Benjamin K. Sovacool (2011)

This view still assumes that income dominates energy use behaviours, but it does link fuels to the energy services they deliver. This is a move away from individuals as rational consumers and technology and cost, and somewhat closer to considering the context that energy is used in. Kroon et al. (2013) also reconsider the energy ladder from an interdisciplinary perspective suggesting that it may be more helpful to consider residential energy transitions in terms of the different decision making contexts a household operates in.

Social scientists have a rather different perspective on energy use and transition. Instead of focusing on technology and the individual as a rational consumer these approaches try to view residential energy use in relation to the social, cultural, geographical, and historical context (Lutzenhiser, 1992). In this view the focus on income is seen as misguided because human behaviour cannot be reduced to simple cost-utility maximisation, and energy technology itself can not be separated from the behaviours, habits, and practices that use it (Moezzi and Lutzenhiser, 2010). There has been growing interest in the use of social practice theory to understand energy transition, which views energy as part of social practices and emphasises emergent dynamics as opposed to causality of external determinants (Elizabeth Shove, 2010; Shove and Walker, 2014). This theoretical view of energy use and transition is suited to the methods used in the social sciences, but often energy research in these disciplines can struggle to scale findings, and contextualise these with respect to wider socio-economic and spatial trends. Also when carried out divorced from the reality of the technologies involved, these approaches can be as ineffective as purely cost-driven economic-technical ones (B. K. Sovacool et al., 2015). Approaches and theories across disciplines vary considerably, which is why several studies (B. K. Sovacool et al., 2015; Stern, 2014; Stephenson, 2017) have advocated for research into energy transition to take a more problem-focused approach. A shift away from the focus on individual choice and technologies can facilitate alignment of research questions between disciplines around the energy transition problem to be addressed.

This thesis, while grounded in engineering, will endeavour to understand and model residential energy transition by borrowing perspectives from across these conceptual approaches to focus on the problem of transition to clean cooking. Recognition that socio-economic determinants play a role in transition along a hierarchy of energy technologies, will feature alongside consideration given to the wider context. How and what energy is used for by individual households is key to understanding the problems of transition. Furthermore, individual households and their practices and behaviours must be considered within their local socio-cultural context.

2.3 Methods for studying Energy Transition

Residential energy transition in the Global South has been the focus of considerable research in the past two decades and there is a rich literature on this. There are many ways to categorise and classify energy research as pointed out by Benjamin K. Sovacool et al. (2018). This section shall adopt a relatively simple division between quantitative studies and qualitative studies. Within the framing of Chapter 1 the quantitative studies tend to cover methods that rely on data and models to interpret and analyse drivers of energy transition, while the qualitative studies tend to be more representative of methods that focus on data and people or stakeholders. The body of research on energy transitions within a Global South context features contributions from a multitude of disciplines.

Reviewing previous research from different disciplines with a view to develop useful models, necessitates a common framework for interpreting use of data across these studies. Categorisation of analyses in the field of urban data analytics provides such a framework. Urban data analytics involves using techniques from the data sciences to process large socio-economic and/or demographic datasets with mixed data types to inform better policy interventions, and there has been growing interest in their use for understanding the multi-dimensional nature of energy use (Bibri and Krogstie, 2017). Wang et al. (2019) describes a categorisation of analyses differentiated by their purpose:

1. Descriptive analysis, which is concerned with understanding the data;
2. Predictive analysis which is concerned with extrapolating the trends found in the data;
3. Prescriptive analysis which is concerned with using the data to identify the interventions likely to achieve desired outcomes.

The distinction offered by this framework is somewhat simplistic and admittedly does not fully capture the nuances of the breadth of qualitative methods. Nonetheless as outlined in Chapter 1 this thesis looks to learn from energy research in adjacent disciplines from the perspective and for the benefit of engineering energy modelling. Within this context the analytics trichotomy offers a lens through which to consider how and why data is used by a study regardless of discipline or methods. Both quantitative and qualitative studies on residential energy transition across the Global South will be reviewed and key findings highlighted. A review of research covering general residential energy use, looking at uptake of both LPG and electricity without a focus on cooking in particular, will be followed by an in-depth discussion of studies which investigated clean cooking fuel adoption in India.

2.3.1 Quantitative Studies

The majority of previous work on residential energy transition has used quantitative data and methods. While there is some variety in methods, there are several types of studies which are most common. First and foremost are studies looking to predict determinants and causes of transition and energy use, these are predominately predictive analyses. Quantitative policy evaluations instead focus on the effect of a particular policy or intervention on transition and energy use, combining aspects of predictive analysis with descriptive analysis. Willingness-to-pay studies are, as the name implies, concerned with identifying the optimal prices and subsidy levels required to achieve a desired level of uptake, an obvious case of prescriptive analysis. Finally, although less common in a Global South context, quantitative energy modelling studies are concerned with predicting future demand based on assumptions about households and their energy end uses, again a predictive analysis.

Predicting determinants of transition

The inception of the Energy Ladder concept coincided with, and spurred, a number of studies mainly in sub-saharan Africa looking at determinants of residential energy transition. These studies mostly focused on using linear predictive models to identify socio-economic determinants of so-called 'modern' fuels, and the preferential hierarchy of fuels discussed above. Studies of households in Zimbabwe (Richard H. Hosier and Dowd, 1987) found income to be a predictor of fuel use along with household size, household location (urban vs rural), and access to electricity (B. M. Campbell et al., 2003). Similarly in South Africa Davis, 1998 found that income was associated with energy transition in rural households as well as access to electricity. Income and household size were identified as significant determinants alongside urban indicators in Ethiopia (Kebede et al., 2002).

Energy consumption and transition to clean fuels have continued to be predicted on the basis of socio-economic factors, with more recent studies expanding the range of socio-economic determinants considered. A plethora of socio-economic factors have been found to be determinants in household energy transition across the Global South including, education of household members (Chen et al., 2006; Alem et al., 2016; Castán Broto et al., 2020), costs of alternative fuels (Louw et al., 2008; Komatsu et al., 2011), dwelling type & ownership (Baiyegunhi and Hassan, 2014; Abbas et al., 2020; Ali, 2020; Castán Broto et al., 2020), female-headed households (Rahut et al., 2016b), duration of cooking (Baiyegunhi and Hassan, 2014), and occupation (Ali, 2020). Determinants of electricity and gas uptake in India have also been characterised by such approaches identifying similar determinants including education, occupation, location, as well as caste (Kemmler, 2007; Filippini and Shonali Pachauri, 2004; Ekholm et al., 2010; B. J. v. Ruijven et al., 2011). Studies specifically addressing clean cooking in India are discussed later on in this section.

Many studies looking at determinants of energy transition use large national datasets and predictive regression techniques. Although some studies instead use these datasets to compute measures of energy poverty, and by extension identify determinants and contributing factors to increasing the risk of energy poverty. This represents a shift in framing of the research question from use or access to poverty, nonetheless the trends identified are consistent. For example S. Pachauri et al., 2004 used a two dimensional measure of energy poverty to identify distribution in India, finding rural households to be particularly susceptible. Meanwhile S. Gupta et al., 2020 presented a novel method of measuring Household Energy Poverty Index, finding 65% of households still count as energy poor and that households in the north and north east were more likely to be energy poor.

More recently some alternative data science approaches and statistical techniques have been used to study determinants of energy transition, including a study by Rao and Ummel (2017) using boosted regression tree modelling to analyse socio-economic determinants of electrical appliance ownership. A separate study by Troncoso et al. (2013) used clustering methods to design metrics for improved cook-stove adoption and assessment of interventions. Such studies are often still trying to predict effects of determinants, albeit using different approaches, with different underlying assumptions (e.g. not assuming linear effects). A few studies have combined quantitative and qualitative data in mixed methods approaches which are discussed in the following section on qualitative studies.

Policy evaluations

Not all quantitative approaches focus on determinants of transition. A number of studies use quantitative data to evaluate policies and interventions identifying changes or trends occurring during a particular intervention. For example a study in Tanzania used an extensive dataset to consider the effectiveness of kerosene subsidy policy in the late 1980s, analysing household energy use changes over a 3 year period and considering the relative cost of the non-biomass fuels (R. H Hosier and Kipondya, 1993). Interestingly, this more descriptive study was one of the earlier studies to explicitly identify the vulnerability of the urban poor in sustaining transition to cleaner fuels.

A study by Rao (2012) looked at kerosene subsidies in the Indian state of Maharashtra, and found that the subsidy was relatively ineffective for rural households as it was based on consumption levels associated with cooking, but rural households seemed to use it primarily for lighting. They recommended targetting subsidies to kerosene-dependent urban households for greater efficiency. A more recent example by Gould, Schlesinger, et al. (2018) took a long term view of clean cooking policy in Ecuador comparing the effectiveness of a long running LPG subsidy and a more recent electric stove programme, noting how high continued levels of subsidy use suggest it may be dissuading households from availing of the electric stove promotion.

While these studies are still a form of predictive analysis, the analyses are often more detailed, with dedicated descriptions of policy mechanisms and resulting effects on distribution of household demand and decision making.

Willingness-to-pay surveys

Willingness-to-pay surveys offer unique insight into the decision making of households, using quantitative survey data to determine fuel or technology prices (and by extension subsidies) required to achieve a desired level of uptake. They are a good example of a more prescriptive analysis, looking to single out the optimal settings required to deliver a policy goal. These studies are by their very nature cost-centric but do also identify socio-economic factors that are associated with greater willingness-to-pay. For example, a study in Kenya (Kroon et al., 2014) investigated willingness-to-pay for different fuel-cookstove combinations, finding greater demand in urban areas. In rural India a study of willingness-to-pay for solar home systems found a gap of around INR 3200 (ca. USD 50) between the price households were willing to pay for such systems and their actual cost (Urpelainen and Yoon, 2015).

Quantitative energy models

A rather different predictive quantitative approach is that taken in studies using quantitative energy models to forecast and predict energy demand across a region or country. Admittedly there are not many examples of such studies looking at residential energy use in the Global South, and those that do often make predictions at an aggregated scale. These tend to forecast electricity needs for the purpose of optimising electrification strategies. Examples of such models include the REMG model which is based on assumptions about socio-economic determinants such as household size, population, expenditure, and electrification rates, and use these as determinants to make national estimates of future energy consumption for different end-uses (cooking, lighting, etc.) (Daioglou et al., 2012; B. J. v. Ruijven et al., 2011)

Some models attempt to integrate geospatial information to the model to map out variations such as the GIS-approach taken by Mentis et al. (2015), that determined electricity needs based on population density to inform cost effective electrification options for regions of Nigeria. Some limited consideration has been given to uncertainty with such residential energy demand modelling, for example B. v. Ruijven et al. (2010) found that uncertainty in model calibration using regional data in India could lead to variation in estimates from the TIMER 2.0 global energy model of 40%. Studies using these models are some of the few amongst quantitative studies on residential energy transitions to give any systematic consideration to uncertainty in the context of predictions at an aggregated scale.

2.3.2 Qualitative Studies

A recent review by Benjamin K. Sovacool et al. (2018) showed the phenomenal breadth of qualitative research methodologies available for energy research. It is not within the scope of this thesis to provide a detailed review of these. Nonetheless a brief summary of these approaches is instructive to understanding the insights these can offer. Many quantitative studies use large regional or national datasets, but qualitative studies often use case studies to explore energy transitions. Data from such case studies is collected through interviews and focus groups, and theories and hypotheses are derived inductively. Such studies employ a descriptive analysis using the data to describe energy transitions and dynamics surrounding them.

An example of a qualitative approach to studying clean cooking is the work of Williams et al. (2020) who used in-depth interviews with households in Peru to understand the socio-cultural dynamics of LPG use, finding that stove design and cooking methods as well as previous experience with alternative fuels were closely intertwined with observed energy use. Similar work by Eludoyin and Lemaire (2021) looked at the interaction between external and internal household context in rural Nigeria, showing the role of community context and how energy was intertwined with the daily habits of a household. Taking a slightly different approach the work of Khalid and Sunikka-Blank (2017) investigated electricity demand of middle class households in Pakistan using practice theories, and showed unpacking household practices relating to energy use can help explain peculiar patterns of demand. Qualitative case studies can feature creative methods to collect data from participants; a case study in Cameroon by Ronzi et al. (2019) used visual data collected by participants to allow households to critically reflect on the factors influencing their use of LPG for cooking.

Review studies are common in interdisciplinary energy journals, and could in some respects be likened to quantitative studies using secondary data. They provide a 'big picture' view of findings, concepts, themes, and practices in the field. Examples of such studies include recent work by O. W. Johnson et al. (2020) who reviewed academic literature on energy transition and intersectional impacts on gender and social equity, offering a description on the state of knowledge. Furszyfer Del Rio et al. (2020) reviewed literature on clean cooking stove interventions and behavioural change strategies, describing commonalities in strategies across research. They found that approaches shaping knowledge, or focusing on social support approaches were most common. They also noted that interventions lacked a focus on gender and did not involve working with stove and fuel suppliers.

Qualitative studies can explore dynamics of complex issues such as gender and inequality, going beyond viewing these as causal determinants in transition and unveiling how these are intertwined with energy use, practices, and behaviours. Indeed there is growing recognition that technologies alone do not fully address social injustice (O. W. Johnson et al., 2020). Musango et al. (2020) note that interest in gender equality in urban residential energy use has focused too much on counting numbers and identifying causal effects and there is a need for greater understanding of gendered differences in decision making. Meanwhile Ochieng et al. (2021) has looked into the role of male members of households in cookstove adoption in Kenya and showed that differences in perceptions (women valued the clean cook stove more but had less influence in decision making) ought to be factored in to the design of clean cooking programmes.

Mixed methods

This summary of previous studies on residential energy transition in the Global South has distinguished them as either quantitative or qualitative. However, there are some studies which have sought to explore the use of mixed methods, combining in-depth interviews with quantitative analysis of survey data. Al-Marri et al. (2018) took such an approach in their study of households in Qatar using quantitative and qualitative data to understand energy use behaviours and attitudes to sustainability, noting that high energy subsidies meant households were disinclined to change their energy use practices. In an Indian context Khosla et al. (2019) combined a regression analysis of survey data with in-depth interviews of low-income urban households showing that behavioural factors are intricately related to energy consumption, for example noting how a culture of socialising at night is reflected in low cooling demand even in new homes with more reliable electricity.

2.3.3 Studies on clean cooking in India

With the recent PMUY scheme leading to an increase in LPG connections, there has been a rise in studies assessing clean cooking transitions in India, particularly using predictive quantitative analysis to identify key drivers. Early studies in Bangalore by A. K. N. Reddy and B. S. Reddy (1994) and B. S. Reddy (1996) found that cooking fuels followed a preferential hierarchy from firewood, to charcoal, to kerosene and eventually LPG which was driven by income. More recent studies such as those by G. Gupta and Köhlin (2006), Farsi et al. (2007) and Narasimha Rao and B. S. Reddy (2007) in the early 2000s found that cooking fuel choice was influenced not only by income and size of household, but that wider socio-economic variables were also significant determinants including education, employment type, caste, and location. The most recent studies coinciding with the PMUY programme continue to show that a mix of socio-economic factors seem to be significant determinants.

	Income	**Household Size**	**Education**	**Fuel Price**	**Female Head of House**	**Urban**	**Low Caste**	**Piped Water**	**Minority Religion**	**Wage Labourer**	**LPG Availability**
G. Gupta and Köhlin, 2006	+	+	+				-		-		+
Farsi et al., 2007	+	+	+	-	+	+					
Narasimha Rao and B. S. Reddy, 2007	+	-	+				-		-	-	
Cheng and Urpelainen, 2014	+	+				+					
Ahmad and Puppim de Oliveira, 2015	+	-	+	-			-	+			
Saxena and P. C. Bhattacharya, 2018	+	+	+			+	-		-		
A. Sharma et al., 2019	+	-	+								
Mani et al., 2020	+	+	+								+
Gould, Hou, et al., 2020	+	+	+		+						

Table 2.2: Summary of significant determinants of LPG cooking from a sample of key quantitative studies using predictive regression type analyses.

Note: '+' Indicates a significant positive effect (i.e. higher value of that determinant is associated with greater LPG use), and '-' indicates a significant negative effect identified.

Table 2.2 shows the determinants of LPG use identified by key studies over the last 15 years in India. A '+' indicates a significant positive determinant, while conversely a '-' indicates a negative one. While data sources and research questions varied slightly, this illustrates the consensus and contradictions between previous quantitative studies on clean cooking in India. Almost all, if not all, of these studies found income, household size, and education to be significant determinants of LPG use by a household. Although it is worth noting that these studies do not agree on the positive or negative nature of the effect of household size. This is an issue seen more widely amongst such quantitative predictive studies on residential energy transition described by Chunekar and Sreenivas (2019), who pointed out that many quantitative studies contradict each other, and there is often little endeavour to compare and contrast results to previous studies.

A number of quantitative studies have not only sought to identify determinants of LPG transition and consumption, but have also looked to assess the effectiveness of the current PMUY programme. These studies have evaluated the roll out using survey data as well as sales data, find that the PMUY programme has been successful in providing millions of BPL households with a connection (Kar et al., 2019; A. Sharma et al., 2019; Giri and Aadil, 2018; Ranjan and Singh, 2020; Mani et al., 2020). However, these studies also find the current policy may have some problems. It does not seem to promote regular use of LPG Kar et al. (2020) and it can fail to help its intended beneficiaries, as up to 24% in some regions don't purchase a refill Kar et al. (2019). Households switching on the scheme are much more likely to not sustain their LPG use or sustain much lower levels than expected, indicating fuel stacking is taking place (Giri and Aadil, 2018; Mani et al., 2020). Some also noted that households that were not aware of hazards of indoor air pollution were more likely to stack use of LPG alongside biomass (Giri and Aadil, 2018).

An interesting study by Aggarwal et al. (2018) proposed a decision support system to aid the roll out of the PMUY that estimated remaining connections needed and associated expansion of LPG supply network required in each state. They identified the North and North East as being critical to the roll out. Some willingness-to pay studies have also been carried out in India with respect to cooking fuel. These studies have often focused on a specific state and in particular rural households. Their findings can be useful for informing pricing as in a study by Jeuland et al. (2015) who found a willingness-to-pay up to INR 1000 for an improved cookstove in northern rural households. More recently Chindarkar et al. (2021) found that willingness-to-pay for exclusive use of LPG was higher amongst those already using LPG or who saw it as better for their health. Paradoxically D. Sharma et al. (2020) found that rural households in Punjab could end up spending more on biomass fuels (paid in smaller frequent purchases) than they were willing to pay for LPG over the same period.

There are far fewer studies which use qualitative methods to study clean cooking transition in Indian households than quantitative ones, although several recent studies have coincided with the current PMUY scheme. Malakar et al. (2018) used in-depth interviews and focus groups to understand how cooking with biomass fuels is entangled with other day-to-day activities of these households, noting that altered gender roles and diversified income sources helped a household reject biomass fuels while noting that a sense of belonging to a community seemed to influence cooking practice more than caste. These issues of caste, class, and gender were further explored by Patnaik and Jha (2020), who described how the PMUY scheme has helped to overcome some caste, class and gender based barriers by recognising the inequality in its criteria, but that greater focus on inequality in affordability of the ongoing refills across groups is needed. A case study in Himachal Pradesh by Jagadish and Dwivedi (2018) deconstructed cultural beliefs around firewood use. They note the importance of availability, infrastructure, and household preferences as well as different cultural domains for fuels and stoves suggesting interventions had to target both aspects to promote clean cooking.

This review of previous studies on residential energy transition more widely, and clean cooking transition in India is not an exhaustive review. Instead this offers an overview of different approaches, particularly drawing a distinction between the quantitative approaches which predominantly use data to make predictions about the effect of different factors on transition, and qualitative approaches with a greater emphasis on using data to describe the state of energy use for particular groups, locations, policies, practices and intersections of these. Attention should be paid to the fact that a great deal of existing research on residential energy transition and clean cooking transitions focuses on the plight of rural households, or a country's population as a whole. Indeed Khosla et al. (2019) notes that the situation of the urban poor and their energy use in India has gone understudied.

2.4 Indian Context

An overview of India's clean cooking policy was provided in Chapter 1 including a brief history and description of current policy measures. The section details the different aspects of the policy in full as well as describes the structure of the market for LPG in India. An understanding of the current policy and its initiatives is necessary to constructively analyse challenges of transition to clean cooking facing low-income urban households today.

2.4.1 The LPG Market

The household LPG market in India is dominated by the three public sector undertaking oil companies (also known as Oil Marketing Companies or OMCs). The OMCs manage the supply chain for LPG from bottling to distribution & marketing, coordinating with a network of over 20,000 private distributors across the country (Aggarwal et al., 2018). Price and subsidy decisions are centrally overseen by the Ministry of Petroleum and Natural Gas (MoPNG). As the OMCs are responsible for administering subsidies they have been directly involved in the suite of consumer facing policies described below, but they have also carried out significant improvements to the supply chain to cope with the expected increases in demand including increased bottling capacity and new distributorships (Kar et al., 2019; A. Jain, Tripathi, et al., 2018).

2.4.2 Current LPG Policy

Current clean cooking policy in India has seen a trio of related initiatives at play. While the 80 million subsidised connections under PMUY may be the flagship measure, each of these elements have been critical in transforming support for clean cooking transition in low-income households in the past five years.

PAHAL

The PAHAL scheme involved a change in the way LPG subsidies were distributed; instead of being sold at subsidised prices LPG cylinders were instead sold at market prices and the subsidy was transferred directly to the bank account of the individual (A. Jain, Tripathi, et al., 2018). OMCs could cross check each other's databases to make sure nobody was claiming subsidies from multiple companies. This effort has been further helped by the growth of Aadhaar cards - India's citizen ID system - enabling verification of identity of subsidised connection beneficiaries. This has saved money lost due to 'ghost' connections or subsidies being illegally used for commercial purpose, and better enabled targetting of subsidies to poor households (K. Smith, 2018).

Give it Up (GIU)

The Give-It-Up campaign involved encouraging wealthier middle class households that had been supported in their transition to LPG two or three decades ago, to voluntarily give up their subsidy, a potential saving of about INR 2000 per household (K. Smith, 2018). The GIU campaign made use of social media and national advertising to encourage households to give up their subsidy so it can be transferred to a poor households. This has been seen as a clever initiative to redistribute subsidies from over 10 million wealthier households towards poorer ones without having to forcibly cut subsidies (K. Smith, 2018; Giri and Aadil, 2018).

PMUY

The PMUY scheme was introduced in 2016 to promote the uptake of LPG by households below the poverty line (BPL) by subsiding the cost of an LPG connection. Under the initiative only women qualified for the subsidised connection thus the policy also sought to enhance the status of women. The cost of connection includes the first cylinder and stove as well as a deposit (totalling between INR 3000-7000, ca. USD 40-95). Through PMUY, the government bears about half of these connection costs including the full deposit and administrative charges and subsidises the stove itself (Kar et al., 2019). Additionally, households are also entitled to a subsidy of approx. INR 200-300 (ca. USD. 3-4) on the recurring cost of refill cylinders (between INR 700-900, ca. USD 10-12) which is transferred to their bank account, after financing on connection costs is recovered by the gas companies (Aggarwal et al., 2018). It is worth noting that this takes approximately 5-7 refills, during which time households pay the higher non-subsidised refill costs.

K. Smith (2018) notes that it is hard to estimate the value of savings from PAHAL and GIU, but that a significant proportion of the cost of PMUY has likely been covered by the subsidies freed up through the GIU campaign.

Concerns

The use of Aadhaar and bank account details of customers for verification and identity checks have have been a key policy feature to save money by preventing fraudulent claims and connections and help better target the subsidy, but there are concerns that this might present a barrier for poor households (such as non-notified slum households). While there have been measures taken since the policy started to facilitate Aadhaar registration for PMUY applicants (MoPNG, 2017), Aadhaar saturation levels nationwide are estimated at around 91% (UIDAI, 2019), suggesting that some people are still unregistered.

While the reimbursement structure of subsidies for refills may be successful in reducing so called 'ghost' connections (K. Smith, 2018), it may also contribute to the relatively high levels of households connected through PMUY that have not sustained use of LPG (Kar et al., 2019). Concerns have been expressed that by requiring household to pay the full cost of the cylinder upfront those on particularly low incomes can be dissuaded from purchasing refills. Additionally if households pay for their connection with financing from the distributor they only avail of the ongoing refill subsidy once the distributor has recouped their costs of the upfront subsidy during which time households pay the higher non-subsidised refill costs, adding a deterrent to ongoing use (Dabadge et al., 2018).

Some concerns have also been raised about its targeting of beneficiaries using the Socio-Economic Caste Census (SECC) as a basis for a household's classification as BPL. This included the possibility of errors in the SECC and potential for relaxation of beneficiary identification to help distributors meet uptake targets (Dabadge et al., 2018). Early evidence in rural areas suggests that this concern has not been borne out (Giri and Aadil, 2018).

2.5 Conclusions

This chapter has presented a non-exhaustive and targetted review of approaches for conceptualising residential energy transition (of which clean cooking transition is an example). At one end of the spectrum the simplistic 'Energy Ladder' takes a quantitative linear view of transition, assuming households to be rational utility maximising consumers - an assumption implicit in many clean cooking programmes to date. At the other end of this continuum of concepts social science approaches such as those rooted in Social Practice Theory unpack the specific context and practices of the household. A key lesson from this is that as more complexity is accounted for, the conceptualisations are able to reveal and describe more heterogeneity across households. However, the trade-off is that this increased complexity makes these concepts harder to apply in an empirical setting at scale.

The review of previous research on residential energy transition brings to the fore the gulf of difference between quantitative and qualitative approaches. Most quantitative methods focus on finding trends and determinants at scale using simplistic linear models and assuming homogeneity across households. These studies characterise research at the intersection of Models and Data in Figure 1.1. Qualitative approaches offer great detail on the relationships between individuals, households, and their surrounding context, research characteristic of the intersection of Data and People in Figure 1.1. A clear gap exists between these extremes which is not well addressed. There are few bottom-up quantitative models, models that focus on heterogeneous individuals and represent research at the intersection of Models and People in Figure 1.1. Furthermore, no studies have looked at the problem of clean cooking transition in India using a truly integrated mixed methods approach.

The analysis in this thesis breaks with the rational utility maximising consumer assumption of other quantitative approaches, instead aiming to understand heterogeneity in clean cooking transitions. An exploration of the non-income factors influencing transition in Chapter 4 will be key to understanding the limitations of existing quantitative approaches at capturing heterogeneity, and informing the design of more suitable methods. By taking a participatory mixed methods approach and using the explanatory power of qualitative data to help understand quantitative characterisation of heterogeneity across urban poor households, Chapter 5 will develop a better understanding of how local context shapes clean cooking fuel transition. The lessons learnt from these analyses inform the design of a modelling approach presented in Chapter 6 that focuses on the spatial heterogeneity in clean cooking transition across a city, addressing the gap in research involving Models and People. The following chapter will expose in further detail how the methodological strategy is designed to achieve this.

3 Methods and Data Sources

3.1 Introduction

As discussed in the introduction to Chapter 2 this research takes a transdisciplinary approach to understanding how to incorporate the influence of socio-cultural heterogeneity into a quantitative energy model. This seeks to move beyond simply drawing on cogent knowledge from other disciplines to also include knowledge from stakeholders, and produce outputs and solutions that speak to the needs and context of those stakeholders. As a result of this the design of methodological strategy and approach to data collection are a key feature of this work.

This chapter begins by introducing an overview of the methodology of this research including a discussion of how research phases and data collection are linked, as well as detailing the timeline of these to expose the iterative, incremental, and participatory approach to this research. A review of key methods and tools used for analysis in this thesis are provided. Following the review of the methods, the case study area is introduced in detail justifying the choice of area and summarising features of the cities studied. A discussion of data used in this research begins with the secondary data sources used in this project, detailing sample features, and an overview of variables available, as well as noting shortcomings or limitations of each dataset. This is followed by a description of primary data collection, detailing the design of survey instruments and interviews, as well as the sampling approach, ethics, and data processing considerations.

3.2 Methods Overview

The analysis conducted in this thesis makes use of a range of different methods from predictive regression, to statistical clustering, interview coding, and microsimulation. It also involves using a wide array of data sources including both primary and secondary as well as quantitative and qualitative data. Both the choice of methods and integration of these are motivated by the transdisciplinary ambition of this thesis rooted in engineering modelling and data science.

The use of the broad range of methods is a result of considering clean cooking transition from the perspective of each of the three intersections of Models, Data, and People outlined in Figure 1.1. Recall that empirical research on residential energy use, can be characterised according to which pair of these elements are engaged with. Each intersection offers a different perspective on the challenges and nature of clean cooking and can yield unique insights, but scales of analysis often differ as does the type of data collected which necessitates that different methods are used for each.

The main contribution of this thesis, understanding how socio-cultural heterogeneity and inequality can be accounted for to better model cooking fuel use, addresses the gap in the knowledge on approaches combining Models and People. However to develop this requires developing a better understanding of socio-cultural heterogeneity through approaches identifying trends at scale using Models and Data, and qualitative people-focused approaches exploring Data on People. The literature review in Chapter 2 highlighted the utility of non-quantitative approaches and data in explaining and exposing the complex interactions and effects of household energy practices and context. Therefore, while the modelling outcome of this thesis is a quantitative one, considering insights from qualitative data and approaches is fundamental to create knowledge that is relevant to the context of stakeholders, and actually useful. It provides stakeholders, in this case specifically households, with a voice in this research.

The deployment and integration of these methods follows a deliberate order and structure. Concepts, findings and outputs from each analysis phase or data collection stage are used to shape, adapt, and inform future stages. This helps to involve not only knowledge from other disciplines, but also provides a means for participation of stakeholders and their cogent knowledge in the research process. The iterative feedback between analysis and data collection, and inclusion of mixed data types, make for a more participatory approach that values knowledge from different stakeholders, both academic and non-academic. Figure 3.1 shows a timeline of key research and analysis phases alongside fieldwork and data collection events, while Figure 3.2 provides a schematic outline of how findings, knowledge, and data flowed between fieldwork and key research phases.

As illustrated by Figure 3.1 the project began with analysis of existing secondary data which demonstrated the importance of understanding heterogeneity in residential cooking energy use. A workshop on energy use in low-income urban Indian housing with city planners and energy stakeholders in Bangalore provided first hand expert knowledge on residential energy use more widely and cooking fuel in particular. Knowledge from both this workshop and the secondary data analysis were used to design data collection for the first data collection phase in Bangalore. Secondary data analysis also informed design of a Mixed Data Clustering Analysis for collected data as shown in Figure 3.2. The preliminary analysis of the survey data from Bangalore was used to design the semi-structured interviews and select households to interview on the basis of issues identified from the survey data, while also informing adjustments to survey data collection for the second phase of surveys in the remaining cities.

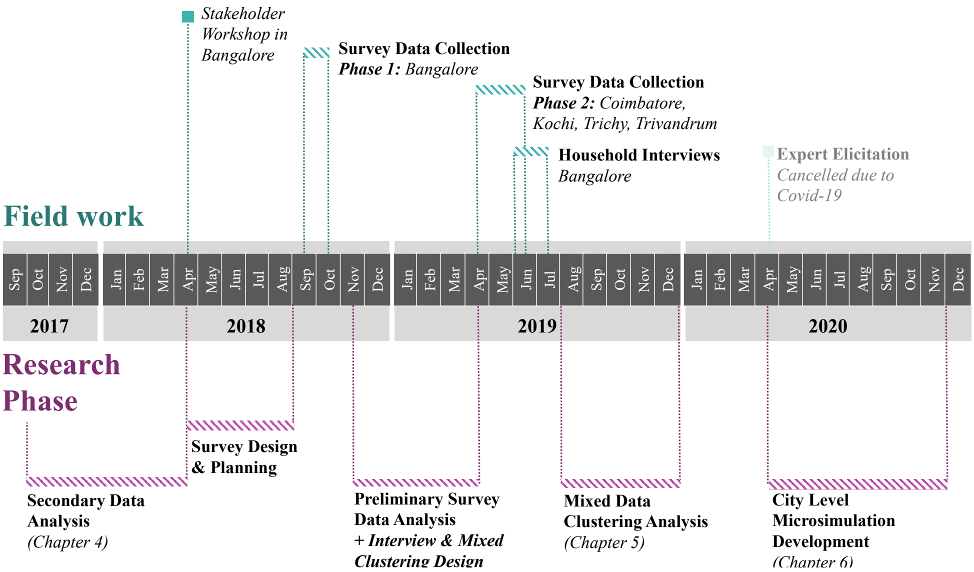


Figure 3.1: Timeline of main research phases and fieldwork phases.

The survey and interview data is jointly used in the Mixed Data Clustering Analysis and the knowledge on how household cooking fuel choices are shaped by socio-economic, cultural and local context informed the design of the City Level Microsimulation, which was validated using survey data from the second phase of surveys. In the original outline for this project a further data collection phase had been envisioned (shown as greyed out in Figure 3.1), which would have involved expert elicitation with community and government stakeholders to derive distributions for key parameters for the city level model. However, originally scheduled for April/May 2020 this had to be cancelled due to the Covid-19 global pandemic. The modelling approach was adjusted to use additional public survey data for parameter estimation, as detailed in Chapter 6.

Each of the three key research phases shown in Figure 3.2 address one of the three objectives of this thesis and by extension each of the three intersections of

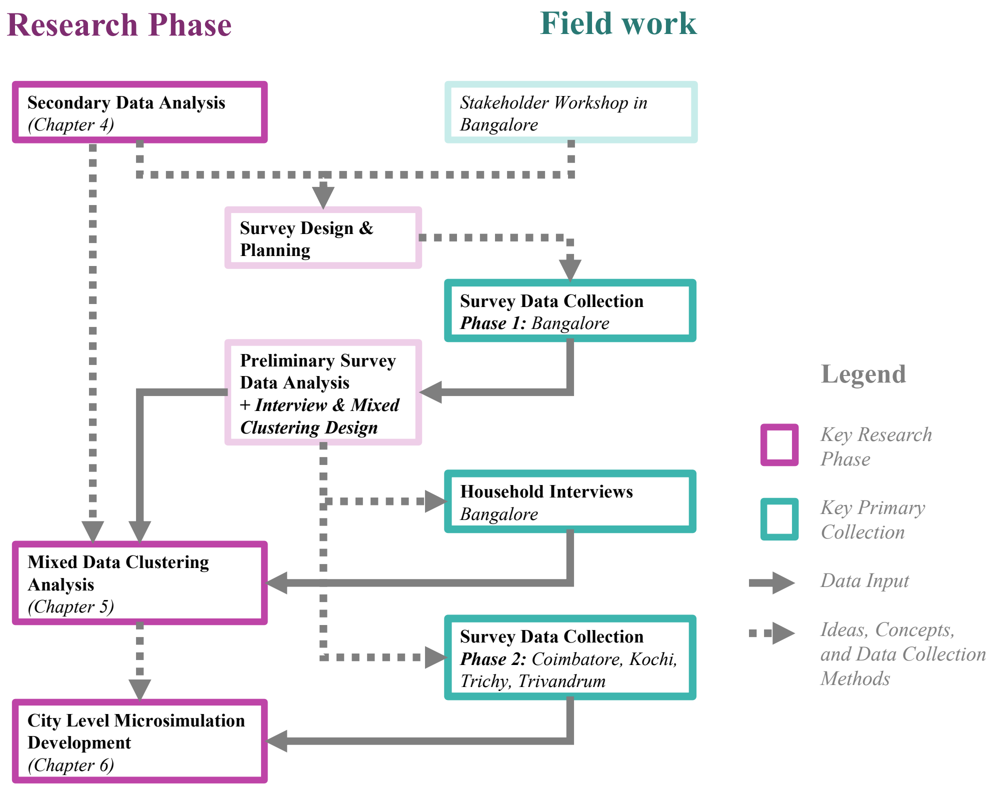


Figure 3.2: Schematic overview of research process highlighting feedback between research designs phases and successive fieldwork phases.

Models, Data, and People used to conceptualise the empirical energy research space in Chapter 1. The results of each phase are presented as a chapter, each using a different combination of data and methods to accomplish the stated research objectives. Table 3.1 summarises the methods and data used to address each of the three research aims across the respective chapters.

Table 3.1: Comparison of methods, data, and objectives addressed by analytical chapters

Chp.	Objective	Intersection	Methods	Data
4	Understand influence of socio-economic determinants & limitations of current predictive approaches	*Models-Data*	• Ensemble Learning Methods, • Regression (for comparison), • Clustering	• Indian Human Development Survey I & II
5	Explore and characterise nature of energy transition pathways among low-income households	*Data-People*	• Survey, • Clustering, • Qualitative Data Analysis,	• Household survey, • Semi-structured interviews
6	Develop modelling approach that integrates influence of local socio-cultural and economic context	*People-Models*	• Microsimulation, • Bayesian Hierarchical Modelling	• Census Tables, • Indian Human Development Survey II, • National Sample Survey R68, • Household survey

The remainder of this Chapter will give an overview of the different key methods and analytical tools as well as data sources used in this thesis that are listed in Table 3.1. Methods and models devised in response to insights and challenges identified through the research, and which constitute key methodological contributions of this work, are detailed in their relevant chapters.

3.2.1 Predictive Regression Methods

Regressions are a widely used method for identifying trends and determinants of an outcome of interest from quantitative datasets. As seen in Chapter 2 these are commonly used in research on energy transition, identifying determinants of clean fuel uptake informed by a linear energy ladder perspective. In this thesis the results of a probit regression and Boosted Regression Tree (BRT) model (an ensemble learning method) are compared to the results of clustering analysis to identify non-income determinants of transition and limitations of these regression approaches.

A probit regression was carried out for comparison with the BRT, as this is a commonly used model for studies on energy transition concerned with a binary outcome. Assuming that the individual's decision on a particular outcome (e.g. use of an LPG cookstove) is based on a latent variable which represents some measure of utility, then this variable can be defined as a linear function of the independent variables, as shown in equation 3.1 where X_i is a vector of all the independent variables for an individual household, β is a vector of coefficients, and u_i captures the uncertainty.

$$y_i^* = X_i\beta + u_i \tag{3.1}$$

The BRT is a tree based ensemble learning technique that combines a large number of simple categorisation trees using gradient boosting to build ensembles of decision trees that are fit to the remaining model residuals. Unlike a probit model there is no a priori specification of the functional form and the BRT analyses the influence of the variables capturing non-linear effects and complex interactions. A challenge of the BRT model is the specification of the hyper-parameters which include the number of trees, the learning rate, and the tree complexity. N-fold cross validation was used to determine the optimum number of trees, and the recommendations of Elith et al. (2008) were followed to optimise the remaining parameters to produce an accurate model and minimise risk of over-fitting. In this thesis BRT was implemented using the 'gbm' (Greenwell et al., 2019) and 'dismo' (Hijmans et al., 2017) packages in the R programming language.

A note on measuring model performance

Assessing the predictive ability of predictive regression models is an important, and often overlooked part of using such methods in research. While many studies often report R^2 values which indicate how well the regression model fits the data used to produce it, far fewer test how well it predicts outcomes for 'unseen' data not used to develop the model. This is represents a true test of real world performance and applicability. It is good practice in data science more widely, but when using regressions in particular, to partition any dataset into a training and testing portion. The training data is used to 'train' or build the model, and the test data is set aside to test how many true and false results the model predicts - such an exercise can flag the possibility of over-fitting.

The binary outcome of interest for such models applied to cooking fuel choice, i.e. whether a household uses clean fuel or not, is not unlike the binary outcomes in medical models assessing patient outcomes (although in this case the outcome is use of clean cooking or not, instead of life and death, at least in the short run). In such cases there is a need for the models to not only perform well on average but also to perform well in distinguishing borderline cases. In the field of medicine when assessing models for patient outcomes, it is good practice to report the calibration and discriminatory ability of the model (Steyerberg et al., 2010). The Brier

Score is an overall performance measure of calibration and discrimination for binary outcomes whose scoring rule is shown in equation 3.2 where N is the number of instances, f is the outcome from the model, and o is the actual outcome. The concordance statistic c, identical to the area under the Receiver Operating Characteristic (ROC) curve for binary outcomes offers a measure of how well the model distinguishes outcomes. Both of these measures were calculated for each model using base packages in R (R Core Team, 2018).

$$BrierScore = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{3.2}$$

3.2.2 Clustering

Clustering methods are concerned with finding groups of similar instances within a dataset, optimising for similarity between instances in the cluster and dissimilarity between clusters (Rousseeuw and Kaufman, 2009). There are different ways of identifying such latent groups in data broadly classified as partition based approaches or model based approaches (Hennig and Liao, 2013). There is a substantial body of literature on clustering methods for mixed data, looking at how different datatypes such as categorical and numerical data, which are common in socio-economic datasets, can be jointly clustered. A long-standing and popular method involves the use of the Gower similarity measure with hierarchical clustering (Gower, 1966), although many approaches to using mixed data often involve coding or discretization of data (A. H. Foss et al., 2019). Other more recent methods require the weighting of different datatypes and do not require conversion or coding of data, although implementation of such methods can be more involved and decisions on weighting can be subjective (Pierpaolo D'Urso and Massari, 2019; A. Foss et al., 2016). Selection of appropriate clustering methods is difficult and as discussed by (Hennig, 2015) when using real world social data latent groups are not necessarily clear cut, and choice of clustering method is highly context dependent.

Another important distinction in clustering methods is between those that can be described as 'crisp' which assign an instance to a single cluster, versus those that are 'fuzzy' and quantify degrees of uncertainty in assignment of an instances to a cluster (P. D'Urso, 2015). Fuzzy clustering can be particularly useful in a decision-making context, although implementation and interpretation can be less straightforward.In this thesis a relatively simple approach to clustering is taken, using hierarchical 'crisp' clustering using partition based approaches. This choice is guided by the motivation of this work to bridge disciplinary divides and as such a straightforward and simple to interpret approach is favourable.

Agglomerative hierarchical clustering is used in for secondary data analysis in Chapter 4 and as part of mixed data clustering in Chapter 5. The Gower distance measure is used to enable inclusion of categorical variables (Gower, 1966), and Ward's linkage criterion is used for clustering. Ward's linkage criterion identifies clusters to merge based on the lowest lack-of-fit sum of squares (Ward, 1963). A consideration with hierarchical clustering methods is that as the size of datasets increase the computational cost increases super-linearly (Contreras and Murtagh, 2015). A non-hierarchical clustering algorithm used as part of the mixed data clustering in Chapter 5 is k-means clustering. This is a popular partition based clustering approach. K-means clustering is not as well suited to the non-spherical overlapping distribution of clusters typical in socio-economic data, but is simple and fast to use for instances where clusters are relatively distinct and globular and evenly distributed (Mirkin, 2015). Both of these clustering methods require careful specification of an appropriate number of clusters.

Selecting optimal number of clusters

The optimal number of clusters in analysis for this thesis are determined using the silhouette width method (Rousseeuw, 1987) alongside the elbow method and the slope statistic of Fujita et al. (2014). While there are several other methods such as the gap statistic (Tibshirani et al., 2001) or the CH index (Caliński and Harabasz, 1974), Fujita et al. (2014) found the combination of the silhouette and slope statistic to be relatively simple and effective when used together to identify the optimum number of clusters. The key is to use more than one method or approach given the overlapping nature of clusters when using high-dimensional data, as any one method may be ambiguous as to the optimal number. The slope statistic is given in equation 3.3 where k is the current number of clusters, $\hat{k}$ is the optimal number of clusters, s is the silhouette value, and p is a positive integer tuned to weight importance of either the subsequent slope (small p) or the silhouette value of the current k (large p). In this thesis this value is set to 1, for equal weighting.

$$\hat{k} = max[-[s(k+1) - s(k)]s(k)^p] \tag{3.3}$$

The slope statistic identifies the optimum number of clusters $\hat{k}$, where a large silhouette value is given for k clusters followed by a significant decrease in the silhouette value for the subsequent $k+1$ clusters. Figure 3.3 plots an example of silhouette width and slope statistic curves for the survey data from Bangalore (the results of which are presented in Chapter 5, and shows the average silhouette width having a first local maximum at $k = 5$, supported by a high positive value of slope statistic indicating 5 as an optimal number of clusters. Hierarchical and k-means clustering was implemented using the base packages in R (R Core Team, 2018) as well as the 'dendextend' (Galili, 2015), 'fpc' (Hennig, 2018), and 'nbClust' (Charrad et al., 2014) packages.

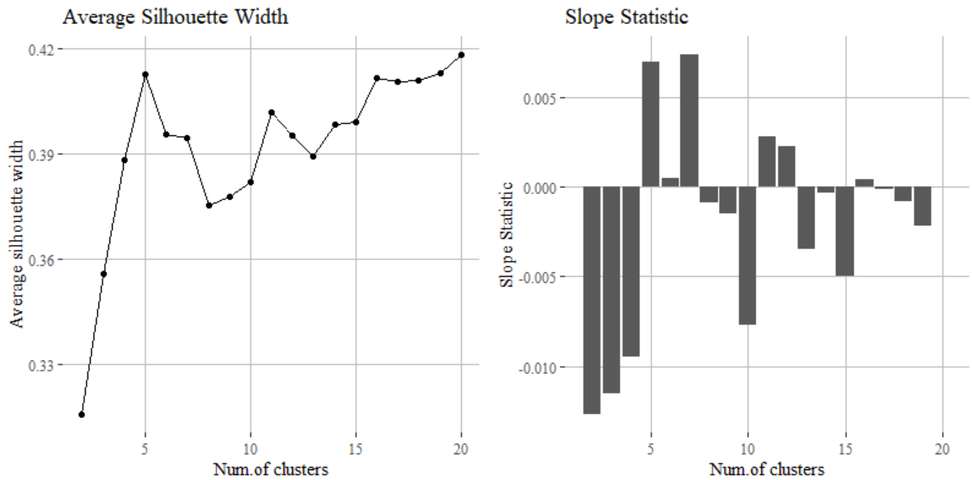


Figure 3.3: Silhouette Width and Slope Statistic to determine optimal number of clusters for a sample of households in Bangalore. While there are maxima on the silhouette width plot at 5 and 16 the slope statistic indicates 5 as preferable.

Correlation clustering

To identify different clusters amongst the interviewed households in Chapter 5 a correlation clustering approach is required given that data for clustering is in the form of a correlation network graph. This approach was first introduced by N. Bansal et al. (2004), and involves clustering a set of instances, in this case interviewed households using an adjacency matrix (Prevos, 2016). The correlation matrix used to produce the network graph is the adjacency matrix and is clustered using a form of hierarchical clustering known as 'fast greedy' clustering. This detects communities within the graph by directly optimizing modularity, as explained by Clauset et al. (2004). Setting a correlation threshold value below which values are set to zero helps to remove weak and negative links. All remaining positive non-zero values indicate a connection between two vertices. The definition of modularity used by this algorithm is shown in equation 3.4, where Q is the modularity, m the number of edges in the graph, v and w are a pair of vertices being considered, and A_{vw} is 1 where vertices v and w are connected, and 0 otherwise. Meanwhile k_v is the degree of vertex v defined as the number of edges incident upon it, δ is a function $\delta(i, j)$ which equals 1 when $i = j$, and 0 otherwise, and c_v is the community to which vertex v belongs.

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \tag{3.4}$$

Unlike the hierarchical clustering methods using Gower's distance and Ward's linkage criterion, modularity optimizing correlation clustering does not require specification of, or additional calculation to determine the optimal number of clusters. This analysis was implemented in R using the 'igraph' package (Csárdi and Nepusz, 2006), which allows for graphical visualisation where linkages between clusters can help indicate clusters which may have some features in common, or indicate members of a cluster which may have commonalities with members of other clusters.

3.2.3 Bottom-up Models and Microsimulation

Microsimulations are a form of bottom-up models that simulate features and actions of synthetic representative individuals in a population (Frayssinet et al., 2018). Such methods have been used in transport research and epidemiology, for example to investigate cardiovascular disease risks throughout a national population (Knight et al., 2017). More recently such methods have begun to be used in urban-scale energy studies with the use of synthetic building stocks where the individuals are buildings rather than people (Booth et al., 2012; Nageli et al., 2018; Zakhary et al., 2020).

Due to reasons of privacy and cost, individual-level representative datasets at urban or district scale are often not available (Casati et al., 2015). Microsimulation studies in epidemiology and transport research generate synthetic populations in the absence of actual individual level data using available public aggregate datasets, either using a single dataset if all required information is contained, or using a combination of datasets (Barthelemy and Toint, 2013). Such approaches have recently been used at a city scale to determine CO_2 emission density (Tirumalachetty et al., 2013), and household gas and electricity use in US cities (Zhang et al., 2018). By generating a population of individuals this enables simulation of outputs based on individual level features and behaviours, better capturing heterogeneity and inequalities across the area of study.

Uncertainty in modelling

Microsimulation modelling naturally embodies uncertainties in the estimation of the synthetic population. These are challenging to quantify methodologically. The absence of test data generally renders any form of uncertainty management of the model parameters impossible. However, it is feasible to include and propagate the uncertainties arising from the variabilities across households in the dataset used for generating the synthetic population. As such, these represent the first order uncertainties, or the heterogeneity across similar clusters of households. The advantage of including these in the analysis are twofold: Firstly, it helps understand the variability of outputs across different population groups, and second, it enables calibration of parameters as new data becomes available.

3.3 Case Study Region

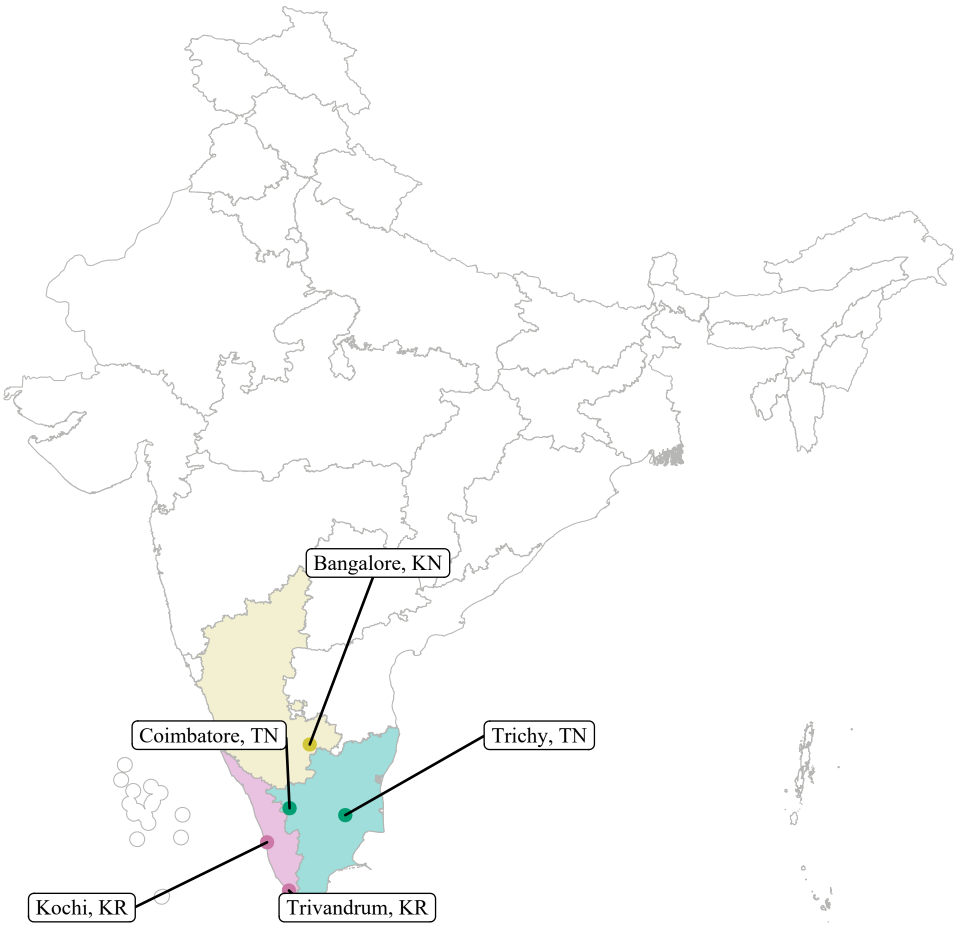


Figure 3.4: Cities covered in survey sample - all cities located in a relatively homogenous climate and geography.

This survey was conducted in five different cities in South India of differing sizes, and with different defining characteristics. The location of these cities is shown in Figure 3.4, notice that relative to the country as a whole the geographic coverage of the cities surveyed is relatively narrow. In a country as large and diverse as India there are considerable climatic, cultural, and regional factors that vary widely from one end of the country to the other and can have a marked effect on energy use, behaviours, and practices. Selecting the cities from the most Southern States of

Kerala, Karnataka, and Tamil Nadu ensures relative homogeneity of these wider factors.

As well as relatively homogenous geographic features, Table 3.2 shows all of these states have above average per capita consumption of petroleum products, and uptake of PMUY connections has been below the national average pointing to a level of existing LPG use above the national average. This offers an opportunity to single out and identify the late adopters, those households that may be struggling most with the adoption and transition to newer fuels.

Table 3.2: LPG and PMUY Uptake Statistics for Case Study States Compared to National figures. Based on data from Government of India (2011) and MoPNG (2021) accurate up to June 2019.

State	**Mean Annual per Capita Petroleum Product Consumption (kg)**	**PMUY Connections**	**Connections as Proportion of Households (%)**
Karnataka	216.0	2,820,262	21.4
Kerala	189.9	209,826	2.7
Tamil Nadu	199.9	3,147,742	17.0
India	*157.3*	*71,912,113*	29.1

With respect to the choice of individual cities within these case study states, the cities in Tamil Nadu and Kerala, namely Coimbatore, Kochi, Tiruchirappalli, Trivandrum, represent 'Tier 2' cities. As detailed in Table 3.3 the two Tamil cities display socio-economic and demographic characteristics representative of the national average, while the Keralan cities both have lower levels of SC households and above average asset ownership and access to banking. This state-wise difference will provide an opportunity to test the ability of the microsimulation model to distinguish the impact of different local context. As for the choice of city in Karnataka, Bangalore is one of India's megacities and altogether a very different type of city from the average urban centre in India. The rapid growth, large population, economic diversity, and greater levels of inequality of such megacities make Bangalore a particularly suitable area of study for an in-depth characterisation of heterogeneity in clean cooking transitions amongst urban poor households. Details of data collection design and an infographic summary of the survey data collected are presented later in this chapter.

Table 3.3: Socio-demographics of cities selected for modelling, compared to national average for urban areas. Based on data from Government of India (2011)

City	Population	SC (%)	Home Ownership (%)	Banking Access (%)	Asset Ownership[a] (%)
Bangalore	8,443,675	11.8	38.4	68.7	25.9
Coimbatore	1,601,438	16.5	59.4	69.5	18.6
Tiruchirappalli	916,857	17.8	71.4	67.5	12.4
Trivandrum	752,490	11.9	75.6	74.9	26.9
Kochi	677,381	8.3	75.5	82.7	27.5
India Urban	-	*14.3*	*69.1*	*67.8*	*12.2*

[a] Census asset ownership tables counts household that own all of the following: a TV, phone/mobile, computer, and scooter/car.

3.4 Data

Data is key to addressing the research question, and while there may be different methodologies employed in fulfilling each of the three objectives of this project, these analyses share a pool of datasets. This research makes use of free publicly available secondary data as well as primary data. Collection of primary data serves the dual purpose of addressing limitations in national datasets and also providing local-scale data for the purposes of model validation. The five different datasets used throughout this thesis are listed in Table 3.4 which summarises differences in coverage and scale. While the secondary datasets are nationally representative, the primary data collection focuses on the main area of study in South India.

Table 3.4: Overview of datasets used in this thesis, including both primary and secondary data and summarising differences in scale and coverage.

Source	**Dataset**	**Coverage**	**Scale**	**Frequency**
Secondary	IHDS	National	Household, District, State	7 years
Secondary	NSS	National	Household, State	3 years
Secondary	Census	National	Ward, District, State	10 years
Primary	Household Survey	South India	Household, Ward	-
Primary	Semi-structured Interview	Bangalore, KN	Household	-

3.5 Secondary Data

The secondary data sources used for this project are not an exhaustive list of all national surveys in India that include energy use questions. Rather the datasets used, namely the IHDS, NSS Consumer Expenditure Survey, and the Census are the three nationally representative surveys which are most commonly used in energy research in India[1]. They are collected periodically meaning that methods and models developed can be re-applied in future with updated inputs. Each of these secondary data sources are collected by a different organisation or government agency and differences in sampling, variables, and scale further distinguish them from each other.

3.5.1 Indian Human Development Survey

The IHDS is a nationally representative panel survey on a range of topics including health, education, income, infrastructure, employment, and energy use. The

[1]Out of the top 20 most cited results on a Web of Science search for the terms "india","energy transition", and "household" the majority (55%) use some combination of NSS, Census, or IHDS data. The NSS is most commonly used although it is worth noting the IHDS data has only been widely available since 2010. The IEA and World Bank statistics are a common data source for studies comparing trends between countries, and studies also collect their own data using the NSS or Census for reference.

first IHDS was conducted in 2004-2005 (referred to as IHDS-I) (Desai, Vanneman, and National Council Of Applied Economic Research, 2010) with a second follow-up survey in 2011-2012 (IHDS-II) (Desai and Vanneman, 2015), which returned to survey the same households originally surveyed for IHDS-I. The most recent round of the survey, IHDS-III, carried out over 2018-2019 is not yet available and is only set to be released in 2023. The surveys were conducted by means of two one-hour interviews with the whole household or the head of the household, and comprised of a nationally representative sample of 41,554 urban and rural households across all Indian territories excluding the Andaman Isles and Lakshadweep. This sample included 1503 villages and 971 urban city blocks across 383 districts in 33 different states. IHDS-II covered 85 percent of the original households, with those households not surveyed the second time either having been unreachable, having moved, or been struck by natural disaster (Desai and Vanneman, 2015). In this thesis where the IHDS is used as a panel dataset, to control for the effect of changes in built environment and constitution of the household due to a move or split of the household only instances where the original household was re-contacted are included. Thus the 6,911 households from the IHDS-I that could not be re-contacted are excluded as well as the 1,721 households which had split into multiple different households. Chapter 4 uses the resulting subset of 32,922 households which were surveyed both in the IHDS-I and IHDS-II. In Chapter 6 where the IHDS is not used as a panel dataset the most recent IHDS-II round is used with all 42,152 households from that round.

As pointed out by Khandker et al. (2012) and Ahmad and Puppim de Oliveira (2015), the energy related questions in the IHDS are more comprehensive than those in comparable studies including the Living Standards Measurement Studies coordinated by the World Bank, and the NSS Consumer Expenditure Surveys. The IHDS dataset is disaggregated by housing type and various demographic features such as gender, religion, caste, occupation, and education (Desai and Vanneman, 2015; Desai, Vanneman, and National Council Of Applied Economic Research, 2010). Additionally the IHDS includes some information on time spent carrying out certain energy related practices in the household, including time spent watching television, time spent collecting firewood, and hours of stove usage. Recommendations from the authors of the dataset were followed with regards to weightings and variable selection (Desai and Vanneman, 2015). All weightings used were the 'SWeights' specified for the households in the IHDS-I, and values for relatively unchanging variables (e.g. Caste and Religion) were taken from the IHDS-II.

Some variables were engineered from the dataset either to make variables more comparable, to create a dummy variable for a particular characteristic, or to characterise change in a variable between surveys. Table 3.5 summarises the engineered variables and the original variables used to compute these. Energy consumption values in the IHDS are given in units of cost (INR) as opposed to units of energy which makes comparisons difficult, and collected biomass fuel is not properly accounted for. These values were converted to estimated energy consumption in kWh using local price data available in the IHDS and collected from government sources (Government of India Planning Commission, 2012). Detail of the process for calculating fuel use, including estimating collected fuel, and auxiliary unit price data are included in Appendix A. In addition, appliance ownership was grouped according to associated household activity: cooking (Pressure Cooker, Mixer/Grinder, Microwave, Refrigerator), and IT (Television, Telephone, Mobile Telephone, Computer, Laptop).

3.5.2 National Sample Survey

The National Sample Survey Office (NSSO) is a branch of the Ministry of Statistics and Programme Implementation and carries out large-scale sample surveys on

Table 3.5: List of engineered variables and their constituent variables from the IHDS.

Engineered Variables	**Constituent Variables**
Firewood Consumption	FU7A: Firewood source, FU7B: Firewood expenditure, VP7B: Firewood price per kg
Kerosene Consumption	FU10A: Kerosene source, FU10B: Kerosene expenditure, VP8B: Kerosene price per kg
LPG Consumption	FU11A: Kerosene source, FU11B: Kerosene expenditure, VP9: LPG price per cylinder, VP9A: LPG kg per cylinder
Electricity Consumption	FU1C: Electricity Expenditure, State Electricity Price Table
Change in Female TV Hours	MM4W: Women TV hours per day (IHDS-II), MM4B: Women TV hours per day (IHDS-I)
Change in Fuel Collection Time	FU13A: Fuel Distance (Minutes) (IHDS-II), FU11A: Fuel Distance (Minutes) (IHDS-I)
Cooking Appliances	CG7: Own Mixer/Grinder, CG18: Own Refrigerator, CG19: Own Pressure Cooker, CG27: Microwave
IT Appliances	CGTV: Owns TV, CG16: Own Telephone, CG17: Own Cell Phone, CG24: Own Computer, CG25: Own Laptop

a regular basis. These include nationally representative household surveys which cover a multitude of socio-economic topics. The Consumer Expenditure Survey (CES) has historically been conducted at roughly five year intervals and includes data on socio-economic status of the household, as well as consumption expenditure broken down by commodity group including not only food, but also non-food items such as fuel and lighting, clothing, and durable domestic goods (appliances) (National Sample Survey Organisation, 2013). The NSS CES reports on the Monthly per Capita Expenditure (MPCE) of the household.

While the NSS has similarities to the IHDS, both including basic socio-economic variables and fuel consumption data as well as featuring household level data, there are some key differences. The NSS is collected by the government and surveys are more frequent than the IHDS (historically collected every five years, in the past decade CES have taken place every 3 years). Importantly the sample for the NSS is designed to be representative at both the national and State/UT level using the latest census data to stratify rural-urban split in the sample within each district (National Sample Survey Organisation, 2015). However, Maiti et al. (2016) discuss how the NSS has some issues in capturing informal economic activity, for example black markets for kerosene fuel, and we have previously discussed how the energy use questions in the NSS are not as detailed as the IHDS for example not including questions on time spent using appliances (Ahmad and Puppim de Oliveira, 2015).

Table 3.6: Correspondence between NSS CES household socio-economic variables from Data Block 3 and IHDS equivalents

NSS Code	Variable Description	IHDS Equivalent
Sector	Rural or urban	URBAN2011
Cooking_code	Primary cooking fuel	FU6: Main chulha/stove *(only partial equivalence)*
HH_size	Number of people in household	NPERSONS: No. in household
Dwelling_unit_code	Ownership status of dwelling	CG1: House own/rent
Possess_ration_card	Household has a ration card	RC1: Ration card
MPCE_URP	Monthly per capita expenditure	COPC: Household expenditure/-capita
HH_Type	Principal occupation	ID14: Main income source
Religion	Religion	ID11: Religion
Social_group	Caste	ID13: Caste category

The variables of interest for this analysis will primarily be the fuel consumption values which similarly to the IHDS are recorded as monthly expenditure with unit prices included. The same process is used to convert these values to units of energy using expenditure and unit price variables in the NSS CES. Unlike the IHDS, the NSS does report quantity in kilograms of collected biomass obviating the need to estimate this. An important feature of the socio-economic variables in the NSS is their equivalence to variables in the IHDS, as this enables combined use of datasets with common household determinants that can be used in model specification. Table 3.6 lists the NSS CES socio-economic variables of interest and their IHDS equivalents. There is a suitable equivalent variable for all of these except the primary cooking fuel variable which has only a partial match in the IHDS. This thesis makes use of the most recently freely available NSS Consumer Expenditure dataset collected as part of NSS Round 68 between July 2011 and June 2012. These common socio-economic variables will be key to the microsimulation approach in Chapter 6.

3.5.3 Indian Census

The Indian Census is a decennial population census in the 33 states and union territories of India. The census has been collected since 1872 and is conducted by the Office of the Registrar General and Census Commissioner, part of the Ministry of Home Affairs (Goswami, 1989). This thesis makes use of the most recently conducted 2011 census (data collection for the 2021 census will be undertaken this year). While the census collects data on the whole population this micro-data at household and individual level is not made publicly available. Instead census tables are published which tabulate data at either ward and town, district, or state level.

Although the NSS and IHDS provide a source of micro data, and do distinguish the state and district of an individual household, they do not disaggregate households at the city scale. The census offers an advantage in this respect by providing aggregated data at a ward level, featuring counts of households grouped according to categorical variables. The census ward level house listing tables HH-14 are used in this thesis, which tabulate proportion of households by amenities and assets. The values from this table are used both to support sample design for the primary data collection as well as to constrain the population synthesis in Chapter 6.

3.6 Primary Data

The publicly available secondary data offers nationally representative samples and covers a range of energy related variables including cooking fuel choice, fuel use, and appliance ownership. However the available secondary data also has some drawbacks with respect to research objectives 2 and 3 both of which require detailed household level data at a city scale. In addition, the undersampling of low-income informal households in these public datasets can make it difficult to gain insight into their energy use. To address these limitations and complement these secondary data sources, primary data was collected through a household survey and follow-up semi-structured interviews. The primary data features both quantitative and qualitative data collected between September 2018 and July 2019. Table 3.7 lists the number of surveys and interviews conducted in each case study city.

Table 3.7: Breakdown of primary data collected across the five case study cities

City	No. of Household Surveys	No. of Semi-Structured Interviews
Bangalore	421	23
Coimbatore	431	-
Kochi	422	-
Tiruchirappalli	421	-
Trivandrum	433	-

3.6.1 Survey of Low-Income Households

This survey was designed to gather information on socio-economic characteristics of households, fuel use, appliance ownership, decision making and energy related practices of low-income households in Southern India. By focusing on the experience of low-income households the data from this survey will contribute to filling the knowledge gap around the energy use of low-income urban households who unlike their rural counterparts are rarely the subject of energy research. The survey also directly addresses some of the limitations of existing nationally representative datasets, as a result of which there are the three key benefits:

- *Resolution*: By selecting specific districts and low-income wards within cities, and surveying a statistically significant number of households in each, there will be sufficient resolution to draw comparisons between these different neighbourhoods. This resolution will also allow ward level out-of-sample validation of model results.
- *Energy Use Breakdown*: Detailed questioning on energy use will enable collection of data on the patterns of energy use, the services driving these, and reasons for fuel stacking by households.
- *Non-income phenomena*: By asking a wide range of questions on practices, lifestyle and socio-cultural characteristics alongside the energy use and socio-economic questions, phenomena such as aspirations, time of use profiles, and convenience can be investigated.

Design of Survey Instrument

The design of the survey instrument is key to ensuring data collection delivers the key benefits outlined above. An important principle of survey design, stressed repeatedly throughout the literature, is the recommendation of starting from the end product. The implication is to start by considering the final analysis and outputs sought, and then work backwards to define the necessary data. Reverse engineering survey instruments from the desired output is especially helpful in the design of questions, both in terms of wording of questions and encoding of answers (Krosnick, 1999; Ornstein, 2013). As detailed by Marsden and Wright (2010) and Ornstein (2013), successful design of survey questions should follow three basic tenets:

questions should be understood, questions should be answerable, and answer categories must fit with the intent of the question. These considerations are discussed below, explaining how recommendations have been incorporated into the survey instrument.

Wording of Questions

The understanding of any question is very much a function of wording, as choice of wording can affect how a question is perceived and interpreted and this in turn can affect the reliability of the answers (Krosnick and Alwin, 1987). If the question is being interpreted differently by different communities, answers will not be comparable. This is especially relevant in surveys across geographic regions with different languages or large cities with migrant labourers who may speak different languages. The different enumerator's interpretation and translation of the question and answer could vary if the wording is unclear.

Practices to overcome this problem include using questions from existing surveys and using simple unambiguous language. It is important to select nouns and adjectives that most accurately address the subject matter but can be understood by people of all levels of education (Payne, 1980; Fink, 2016). It is common in census-style surveys to borrow questions from previous studies which have been shown to provide consistent and reliable answers. This has two benefits: First, it saves time and resources needed to test questions. Second, and more importantly, using the same question as a previous survey allows the results to be compared against it and thus the new data can be checked for consistency and/or anomalies. This survey draws upon question wording from the IHDS and used pre-testing (discussed below) to identify and resolve any difficult wording.

Avoiding Ambiguity

Questions must be easily answerable. A question becomes unanswerable if the question is ambiguous, or if the respondent does not feel they can confidently answer the question due to lack of knowledge (Tourangeau et al., 2000). Cognitive theory approaches to survey research point out that difficult to answer questions can also lead to a behaviour known as 'satisficing' where respondents shortcut the response thought process and use biases and cues from the question or interviewer to select what they think will be a plausible answer (Krosnick, 1999; Tourangeau et al., 2000). Thus, designing answerable questions requires considering what might constitute an ambiguous question for a typical respondent and the knowledge or understanding they are likely to have on the subject. Pre-testing of questions used in the design of the survey instrument can help identify issues of ambiguity, as discussed below.

Categorisation of Answers

If the wording of the question is an important aspect, so is the categorisation of answers. In the case of multiple-choice questions, this includes the subset of possible answers and the inclusion of an "Other", "Unsure", or "No answer" option. The respondent may infer meaning or framing of the question from the available response categories (Krosnick, 1999), or the available responses may simply lead to an unconscious bias due to short-term cognitive processes (Ornstein, 2013). Thus it is important to avoid 'favouring' or 'indicating' a particular response. This effect, however, is usually small provided respondents understand the question (Krosnick and Alwin, 1987). Similarly, 'no answer' or 'unsure' options can prompt the sort of satisficing behaviour discussed previously. Respondents who don't feel confident, or feel that the interviewer is 'testing' their knowledge may choose a 'no answer' or 'unsure' response as a way out (Fink, 2016). In the survey an 'other' option is included for several questions, categorical answers are not placed in any order of preference and enumerator training emphasised the importance of not appearing to favour any answers.

Ordering of Questions

A final important consideration in the design of the questions in the instrument is their order, which has been shown to impact the responses (Knol et al., 2010; Krosnick and Alwin, 1987). It is common practice to order questions so as to allow participants to ease into the survey, leaving more complex questions towards the middle to end. However, and as Krosnick (1999) points out, respondents can become fatigued as a survey progresses so one cannot expect the same cognitive focus. Thus it is recommended to place questions which require the most cognitive outlay in the third quarter of the survey. The key energy use and decision making questions in the survey were identified as those requiring greatest cognitive outlay, and accordingly these are placed mostly in the second half of the survey.

On a note or structure of questions, it is often useful to breakdown complex questions into parts. For example, instead of asking 'how many units of electricity does the household use monthly?', one could ask 'how much did you spend on electricity this past month?', and then 'what is your electricity tariff/rate?'. The units of electricity (in kWh) could be calculated from their responses and would avoid mistakes made by the respondent if they do not think of their electricity use in terms of kWh, but instead perhaps on the basis of expenditure. The survey instrument makes use of this approach with respect to energy consumption questions.

Pre-testing of Questions

Pre-testing of the survey instrument on a small group of respondents can identify questions that are difficult to understand or are interpreted in a manner other than intended by the researcher, and allow for the above criteria to be assessed before widespread data collection. Typically a small group of 15-25 respondents will be used for pre-testing with a debriefing session following the survey to understand respondents' experience (Bischoping, 1989; Krosnick, 1999). More recently, variations on this approach have been introduced such as behaviour coding (Fowler and Cannell, 1996) in which an observer monitors the pre-test interview and takes note of problems in interpretation, re-reading of questions, and misunderstandings and relevant questions are reviewed. Another approach - which was employed for this survey - is cognitive pretesting where respondents "think aloud" while answering questions to allow interviewers to understand how the question is being interpreted (Bickart and Felcher, 1996; DeMaio and Rothgeb, 1996).

Data Compatibility

An underlying criteria for the design of the survey instrument is to ensure the potential for compatibility and cross-referencing with the existing IHDS and to a lesser extent other NSS and Census data. To allow for this, a selection of questions characterising the household in terms of household composition, dwelling type, caste, religion, expenditure, and education were taken from the IHDS-II (2011) survey (Desai and Vanneman, 2015). Table 3.8 details the structure of the survey in terms of composition of questions and data sought. The questions developed from scratch for this survey were built around probing energy use behaviours and decisions. Table 3.8 summarises the sections of the survey questionnaire.

Table 3.8: Overview of survey sections, data types collected and origin of questions.

Section	**Heading**	**Question Source**	**Type of Data**
1	Household identification	Abridged from IHDS-II	Socio-cultural indicators
2	Household roster	Abridged from IHDS-II	Demographic indicators
3	Occupation and salary	Abridged from IHDS-II	Economic indicators
4	Education	Abridged from IHDS-II	Socio-economic indicators
5	Appliance ownership	Expanded from IHDS-II	Appliance ownership
6	Fuel use	Developed from scratch	Fuel use magnitude
7	Energy use habits	Developed from scratch	Energy use practice

Survey enumerators were recruited from the local area who were fluent in the commonly used local languages and familiar with the areas being surveyed. A training session was held with the enumerators prior to data collection to ensure understanding of the questionnaire, and to address potentially problematic questions or missing answer options. The survey instrument was encoded as an ODK XForm that can be used interactively via the ODK Collect app to enable survey responses to be directly recorded in a digital format. Example screenshots of the app are shown in Figure 3.5. Enumerators completed surveys on tablet computers, this ensured that data was immediately encoded, tabulated, and uploaded to the server reducing the risk of any mistakes in transcribing from paper surveys. Appendix B contains a copy of the full survey instrument in English.

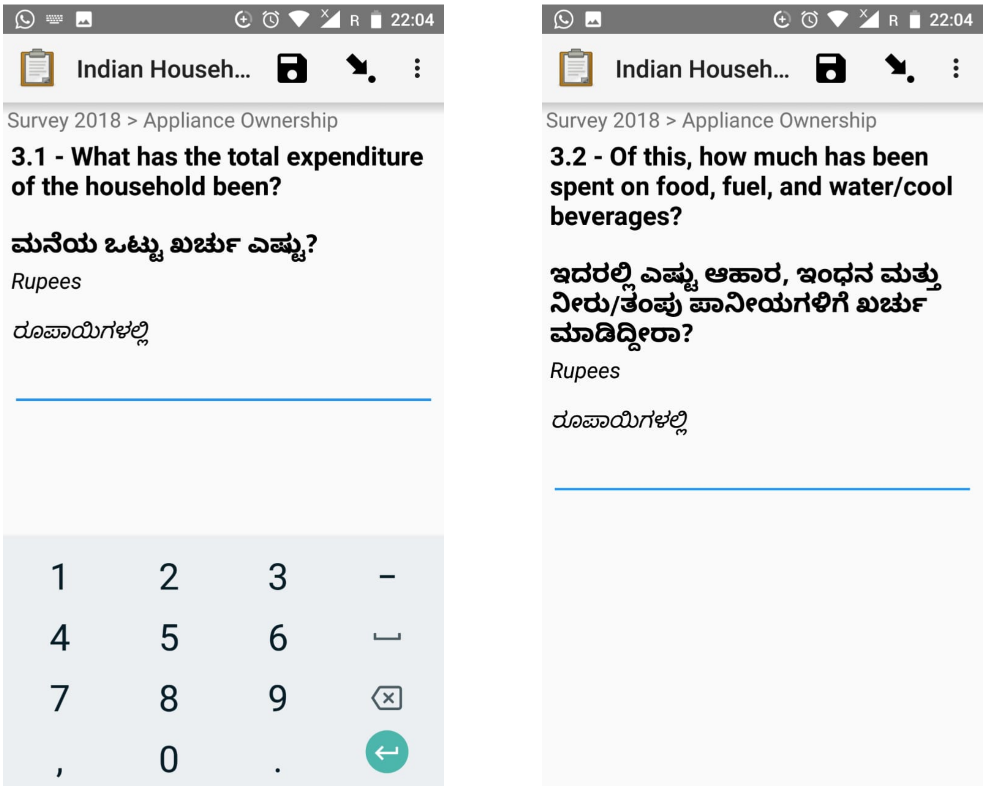


Figure 3.5: Example screenshots of household expenditure related survey questions on ODK Collect app. English/Kannada version shown, similar translations were made into Tamil and Malayalam

Sample Area & Sizing

The selection of sample wards within the the case study cities is important as cities in India can exhibit substantial spatial inequality and certain city wards will have a far greater proportion of low-income households than others. A subset of variables from the ward level 2011 census data (Government of India, 2011) listed in Table 3.9 were used for ward selection. Selection of wards was performed using a rank score on the variables of interest, to identify wards with socio-economic features which suggest a high proportion of low-income energy poor households. The equation 3.5 below gives the rank score used for shortlisting of wards, where n is the number of variables used for ward selection, W is the total number of wards, and x_i is the rank of the given ward for the ith variable. Final selection of wards was based on the shortlisted wards and local knowledge of enumerators familiar with logistical and political characteristics of the local area. Either 6 or 7 wards were selected given resource constraints on individual household survey numbers related to the sample sizing discussed below. Appendix C details the wards selected in each of the five cities surveyed.

Table 3.9: Table of Census City Ward Level Variables used for Ward Selection.

Variable	**Rank Ordering**
Access to Banking (%)	Descending
Home Ownership status: Rented (%)	Ascending
Primary Cooking Fuel: Kerosene (%)	Ascending
Primary Cooking Fuel: LPG (%)	Descending
Overall Asset Ownership (%)	Descending
Lighting Energy Source: Electricity (%)	Descending
Lighting Energy Source: Solar (%)	Ascending

Note: The rank ordering indicates whether high values of this variable (ascending) received a high rank score, or low values (descending).

$$Score = \sum_{i=1}^{n} \frac{(W+1) - x_i}{W} \tag{3.5}$$

Sample Sizing

The survey comprises a range of quantitative questions whose purpose is to determine population means, as well as qualitative questions with categorisation which will not follow a normal distribution. The selection of sample size for qualitative surveys cannot be obtained purely by calculation and often relies on precedent and best practice (Kelley et al., 2003; Knol et al., 2010), although some studies have sought to employ statistical test power measures of sample size based on theme prevalence (Fugard and Potts, 2015). Others have pointed out a trade-off between higher information power of small samples and greater statistical power of larger sample sizes (Malterud et al., 2015).

The sample can be definitively sized for the key quantitative data, in this case magnitude of fuel use, and in practice this is likely to be the limiting sample size criteria. There are several approaches that can be taken, one method is to define the width of confidence interval for the mean of the parameter of interest and calculate the sample size required to deliver this, or the power of a test hypothesis on the parameters of interest can be calculated to determine the minimum sample size to attain a certain power of test value (Lenth, 2001). There is a Bayesian approach which is well suited in cases where we have a prior distribution of the desired parameter, and can use this in place of making a guess (Sadia and Hossain, 2014).

In the case of this survey, while there are prior distributions for some cities they are several years out of date and thus using an estimate for the expected mean value offers a sensible method for determining sample size. The aim is to have a representative sample within each urban ward surveyed. Inferring key transition pathways and providing supporting quantitative data for these pathways motivated the criteria of estimating mean fuel use in a ward with a margin of error of +/- 10 percent at a 95% confidence level. Assuming that energy use is approximately normally distributed the sample size n is calculated using equation 3.6, where σ is the standard deviation, $Z_{\alpha/2}$ is the value of Z providing an area of $\alpha/2$ in upper tail of normal distribution, and E is the margin of error.

$$n = \frac{(Z_{\alpha/2})^2 \sigma^2}{E^2} \tag{3.6}$$

Obviously one of the problematic values in this calculation is the standard deviation of the expected data. Anderson (2009) discusses several possible approaches, including using the standard deviation from a previous study, the standard deviation from a pilot study, or a 'best guess' approach which involves estimating upper and lower bounds of the population. Given that there is existing data for Indian cities (albeit out-dated, from the IHDS-II), the initial sample sizing calculation has been based off the values from this. Table 3.10 below shows the indicated sample size for different magnitude of error on the mean biomass monthly use. In this case a ward sample size of 60 was chosen (rounded up from 58). Following a review of data from Bangalore the first city surveyed this sample was increased to 70 for the remaining cities. This resulting sample size for each city fell between 420-440 households, with a total sample of 2129 households.

Sample Validity

When collecting or generating data, wherever possible efforts should be made to check the validity of the data by comparing it to known expected data. The survey data was checked against the census ward level proportions to ensure the samples

Table 3.10: Table of required sample size for Bangalore survey case study given different margins of error in estimated mean LPG fuel use, at a confidence level of 95 %. Based on IHDS-II levels of urban LPG use.

Margin of Error, E	**Required Sample Size, n**
1.9 kWh (1.0%)	5674
4.8 kWh (2.5%)	892
9.6 kWh (5.0%)	226
14.4 kWh (7.5%)	102
19.3 kWh (10.0%)	58
24.1 kWh (12.5%)	39

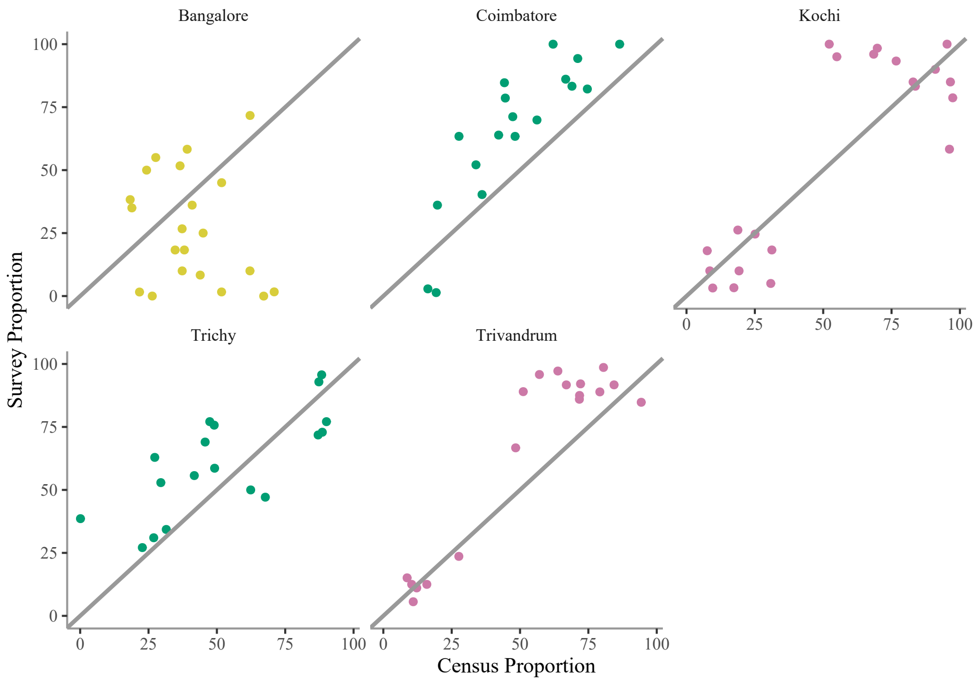


Figure 3.6: A visual check of survey data from wards plotting census and survey proportions of home ownership, households with multiple rooms, and LPG stove ownership for each ward. Grey line indicates parity between census and survey values.

were reasonably representative of the wards surveyed. Many of the primary variables of interest are not detailed at a ward level in the census data, however for purposes of a consistency check home ownership status, LPG stove ownership, and size of dwelling (rooms) were used to check the data given they are recorded in both survey and census. Figure 3.6 visualises a consistency check by comparing the proportions of these categorical variables in the census and survey. Each point represents a variable for a single ward and the census value is plotted against the survey value, where the grey line represents parity between the survey and census values. In any real world data collection situation a perfect match between datasets is extremely unlikely, instead the aim is for reasonable agreement indicated by the points being grouped around the parity line and following the slope. In this case Figure 3.6 shows a good agreement for the cities in Tamil Nadu and Kerala.

Bangalore is the one city where the survey data seems to be at odds with the census data. This is likely due to a problem with the sampling method as a result of larger size of the city and the density of population within Bangalore's wards. A simple random sample approach was used employing a random walk approach. That random walk approach relied on moving a set number of houses before knocking and basing decisions to turn onto another street on whether recent households called upon had agreed to be surveyed or not. In Bangalore extremely dense slums exist in the midst of wealthier neighbourhoods, by starting random walks within the slum and with only 60 households to be surveyed it was likely that the enumerators would not cover many if any households outside the slum resulting in an unrepresentative sample of the ward. This means that for purposes of model output verification in Chapter 6 the Bangalore subset of the data is not sufficiently reliable. However the Bangalore sample does meet the objective of collecting a sample of low-income households in Bangalore, and the issue of sample representativeness does not prevent the use of the survey subset in mixed method clustering analysis in Chapter 5.

3.6.2 Semi-Structured Interviews

The qualitative dataset consists of in-depth semi-structured interviews with a separate sample of households in Bangalore from the same subset of wards selected for the survey. This form of data collection is common in the social sciences for studying issues ranging from urban inequality to gender studies (J. L. Campbell et al., 2013). The anonymity of surveyed households was safeguarded by not collecting their addresses. Therefore it was not possible to return to the specific households covered by the quantitative surveys. Instead, from the seven wards where the survey was conducted in Bangalore, a purposive sample informed by the different types of households identified in the preliminary survey analysis were interviewed.

As previously discussed, sizing of samples for qualitative studies is typically based on previous experience and best practice. In this case the sizing was based on the advice of Guest et al., 2016, and the precedent set within the field that more than 12 interviews provides a reasonably high probability of identifying key issues (Galvin, 2015). 23 interviews were carried out (24 attempted although one respondent declined to continue interview). Selection of interviewees was targeted to feature a higher proportion of households representing outlier clusters in the analysis of the survey data while ensuring representation of all cluster types in the survey cluster analysis which is detailed below. Expert knowledge from the survey enumerators helped inform the selection of these households.

The 20 to 30-minute semi-structured interviews allowed flexibility in identifying and discussing issues important to participants. The interview was structured to cover four broad topics, namely household energy consumption preferences and practices, finances, social networks/community, and political networks/interactions. Table 3.11 details the issues the interviewer sought to discuss under each of these topics. The interviews were conducted in Kannada and Tamil and transcribed to English and stored as digital text files for coding and further analysis

Table 3.11: Structure of semi-structured interview based on four key topics, the table indicates the issues which the interviewer sought to discuss in connection to each topic

Topic	**Issues discussed**
Energy consumption	Preference of cooking fuels, appliance usage, aspirations, knowledge of health impacts of fuels;
Financing	Expenditure and budgeting habits of households related to energy costs, frequency of replacement of LPG cylinders, access to formal/informal loans;
Social network	Participants' involvement with community networks and ability to rely on the same for support, financial or otherwise (e.g., community lending or savings clubs or sharing information regarding schemes);
Political network	Participants' relationship with existing local government, involvement with political associations, experiences in community mobilisation;

detailed in the Chapter 5. Collection of this data was carried out by Ms. Rishika Rangarajan, a collaborator at the Indian Institute for Human Settlements, Bangalore. Design of the interview and coding of the data was carried out collaboratively with Ms. Rangarajan.

The interviews were coded and analysed using the 'RQDA' package in R (Huang, 2014), which provided a graphical interface for the coding process and facilitated export of datasets to the R environment for analysis alongside the quantitative survey data. The interview codings were peer-reviewed in collaboration with Ms. Rangarajan to eliminate bias of the individual researchers, and coding was then analysed to identify key themes. The interview coding data resulting from this collaboration was used as the qualitative dataset for the mixed methods approach in Chapter 5.

3.6.3 Ethics and Data Availability

Collection of survey and interview data in the study received ethical approval from the respective lead institute on the data collection. The household survey received ethics approval from the University of Cambridge Department of Engineering Ethics Committee and the ethics review can be found in Appendix B. As part of compliance with this participants were provided with information regarding the research before deciding to consent to the interview. Participants were informed that they could refuse at any time to be surveyed, and that no personal data would be shared. Only fully anonymized and processed datasets are made publicly available and all personal identifying data (e.g. names) was deleted as soon as checks had been performed on the collected data and the raw dataset was finalised.

As the interviews were conducted, translated, and transcribed by our collaborator Rishika Rangarajan at the Indian Institute for Human Settlements, ethics approval for the interviews was obtained from the IIHS Ethics Committee. Full interview transcripts could be used to identify individuals, therefore only anonymized excerpts were used in publications and only anonymised coding data was shared. This anonymised coding data was used for the analysis addressing research objective 2 and topics, themes, and issues identified through the joint coding analysis are discussed alongside the results in Chapter 5.

All data collected through this project has been made publicly available through the University's Apollo repository subject to the anonymity requirements outlined above, with additional R code and example analysis scripts on Github.

Bangalore Survey & Interview Data (processed)

https://doi.org/10.17863/CAM.59870

Tamil Nadu and Kerala City Survey Data (processed) https://doi.org/10.17863/CAM.66449

Microsimulation Model Code & Data:

github.com/anetobradley/urban_energy_microsimulation_india

3.6.4 Survey Infographic

The infographic overleaf summarises key variables for each of the five cities surveyed as part of this research and is intended as a visual aid for the reader highlighting features and differences, as opposed exact tabulation of values.

Notes on Socio-Economics & Demographics

Dwelling type: The distinction is made between notified and non-notified slums, and all other dwelling types. Non-slum housing was almost entirely bungalow/low-rise housing with almost no flat dwellers. Some households officially listed as non-notified slum but were in fact immediately adjacent to a notified slum area but outside the official boundary.

Caste: Caste was self-reported by the member(s) of the household present.

Time since migration: This indicates the time a household has lived in their present dwelling/area. Following the IHDS convention, households living in place for more than 90 years are assumed to be permanent residents of the area and influence of migration negligible.

Income Frequency: This was determined from categorical responses.

Monthly Expenditure: This was self-reported by the member(s) of the household present for the typical month.

Note on Appliance Ownership

Based on questions relating to household ownership of appliances. The survey did not ask what number of each appliance was owned.

Note on Monthly Fuel Consumption

The bars summarise fuel consumption across all households, with the bar representing the minimum and maximum value range and the circles representing the median household. Numbers on the right hand side are mean values including non-user households.

Note on Energy Use Habits and Decisions

Daily use profiles do not indicate magnitude of fuel use but rather proportion of households using energy for the given end use for each hour of the day.

SOCIO- ECONOMICS & DEMOGRAPHICS
APPLIANCE OWNERSHIP
MONTHLY FUEL CONSUMPTION
ENERGY USE HABITS & DECISIONS
Bangalore, KN
7 WARDS SURVEYED
421
HOUSEHOLDS
57% SLUM (NOTIFIED)
40% SLUM (NON-NOTIFIED)
3% NON-SLUM
DWELLING TYPE
42% SCHEDULED CASTE
8% SCHEDULED TRIBE
49% OTHER BACKWARD CASTES
CASTE
9% WERE NOT RECENT MIGRANTS
YEARS
TIME SINCE MIGRATION
DAILY 61%
WEEKLY 23%
MONTHLY 8%
INCOME FREQUENCY
INR
MONTHLY EXPENDITURE
12% 39% 0% 23% 86%
LPG STOVE
KEROSENE STOVE
IMPROV. CHULHA
CEILING FAN
MOBILE PHONE
ELECTRICITY 52 kWh
LPG 12 kWh
KEROSENE 23 kWh
BIOMASS 260 kWh
MEAN
kWh
MONTHLY CONSUMPTION
COOKING FUEL STACKING * 41%
ACQUIRED LPG IN LAST 3 YEARS 9%
COLLECT BIOMASS 4%
* AMONG HOUSEHOLDS OWNING NON-BIOMASS STOVES
LIGHTING
COOKING
TIME OF DAY
DAILY USE PROFILE
Coimbatore, TN
6 WARDS SURVEYED
431
HOUSEHOLDS
2% SLUM (NOTIFIED)
40% SLUM (NON-NOTIFIED)
56% NON-SLUM
29% SCHEDULED CASTE
13% SCHEDULED TRIBE
45% OTHER BACKWARD CASTES
4% WERE NOT RECENT MIGRANTS
DAILY 39%
WEEKLY 58%
MONTHLY 2%
98% 2% 80% 84% 99%
ELECTRICITY 135 kWh
LPG 96 kWh
KEROSENE 1 kWh
BIOMASS 29 kWh
COOKING FUEL STACKING * 37%
ACQUIRED LPG IN LAST 3 YEARS 71%
COLLECT BIOMASS 73%
Trichy, TN
6 WARDS SURVEYED
421
HOUSEHOLDS
0% SLUM (NOTIFIED)
43% SLUM (NON-NOTIFIED)
57% NON-SLUM
62% SCHEDULED CASTE
4% SCHEDULED TRIBE
32% OTHER BACKWARD CASTES
1% WERE NOT RECENT MIGRANTS
DAILY 47%
WEEKLY 38%
MONTHLY 2%
100% 0% 64% 93% 100%
ELECTRICITY 149 kWh
LPG 98 kWh
KEROSENE 0 kWh
BIOMASS 90 kWh
COOKING FUEL STACKING * 85%
ACQUIRED LPG IN LAST 3 YEARS 92%
COLLECT BIOMASS 40%
Kochi, KR
7 WARDS SURVEYED
422
HOUSEHOLDS
14% SLUM (NOTIFIED)
4% SLUM (NON-NOTIFIED)
82% NON-SLUM
15% SCHEDULED CASTE
<1% SCHEDULED TRIBE
68% OTHER BACKWARD CASTES
2% WERE NOT RECENT MIGRANTS
DAILY 69%
WEEKLY 19%
MONTHLY <1%
89% 23% 29% 90% 99%
ELECTRICITY 177 kWh
LPG 101 kWh
KEROSENE 66 kWh
BIOMASS 113 kWh
COOKING FUEL STACKING * 90%
ACQUIRED LPG IN LAST 3 YEARS 27%
COLLECT BIOMASS 66%
Trivandrum, KR
6 WARDS SURVEYED
433
HOUSEHOLDS
77% SLUM (NOTIFIED)
1% SLUM (NON-NOTIFIED)
22% NON-SLUM
68% SCHEDULED CASTE
3% SCHEDULED TRIBE
21% OTHER BACKWARD CASTES
0% WERE NOT RECENT MIGRANTS
DAILY 66%
WEEKLY 14%
MONTHLY 3%
94% 5% 23% 97% 97%
ELECTRICITY 108 kWh
LPG 106 kWh
KEROSENE 5 kWh
BIOMASS 168 kWh
COOKING FUEL STACKING * 93%
ACQUIRED LPG IN LAST 3 YEARS 24%
COLLECT BIOMASS 8%

4 Models and Data

The contents of this chapter have been published in the following article in Energy Policy; A P Neto-Bradley worked on conceptualisation, methodology, data curation, investigation, and visualisation; R Choudhary and A Bazaz provided supervision and with inputs on conceptualisation and methodology.

André Paul Neto-Bradley, Ruchi Choudhary, and Amir Bazaz **(2020)**. "Slipping through the net: Can data science approaches help target clean cooking policy interventions?" en. *Energy Policy* 144, p. 111650. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111650

4.1 Introduction

Chapter 2 reviewed the wealth of quantitative studies into the determinants of residential energy transition across the Global South, including the growing number of recent studies on cooking fuel in India. The overwhelming majority of such studies employ predictive regression models and large quantitative datasets to identify determinants of energy transition and the magnitude of their effect. Such models usually rely on the assumptions that effects of determinants vary linearly and that households are utility maximising consumers, i.e. that all households will have the same response to the socio-economic determinants. Such an approach can lead to an over-emphasis on cost and the resulting cost-centric policies, the most recent of which being the PMUY scheme which subsidises LPG stove connections, allow some households to slip through the net and get left behind using biomass.

Regression models and large survey datasets provide a powerful tool to predict the magnitude of effect and statistical significance of socio-economic determinants of energy transition across a population as a whole. However this approach can oversimplify the nature of that effect, and negates the interaction of local socio-cultural context with household practices and decision making. This chapter addresses the first of the three research objectives set out in Chapter 1, concerned with understanding the influence of socio-economic and cultural determinants of clean cooking transition, and the nature of that influence, while also exploring and characterising the limitations of current quantitative predictive approaches in fully describing this.

Ensemble machine learning regression methods and clustering algorithms are employed to conduct a combined predictive and descriptive analysis of the panel data from the IHDS. The predictive analysis characterises the non-linear relationship of different determinants that influence clean cooking adoption and contrasts the results of the boosted regression tree approach with a more conventional probit regression. The descriptive analysis demonstrates that groups of households follow different cooking transition pathways delineated by specific combinations of socio-economic circumstances. The combined analysis identifies differences in key policy relevant social, economic and cultural features between groups of households for which targeted interventions could be designed to address the needs and challenges of households that might currently be under-served by cost-centric policies.

The remainder of this chapter is organised as follows; section 4.2 describes the analytical method, with section 4.3 presenting results of two different regression models and comparing the performance of these. This is followed by section 4.4

which presents the results of a descriptive clustering analysis of households that did switch to a non-biomass stove, discussing the existence of different types of switching household. Consideration is given to limitations and caveats of this analysis in section 4.5 and conclusions are presented in section 4.6.

4.2 Methods

This analysis will make use of the IHDS panel dataset detailed in Chapter 3 with the main dependant variable of interest being a constructed binary variable indicating whether the household had switched from primarily using a biomass stove in 2004/5 to a non-biomass stove in 2011/12. This was constructed using the variables indicating the main stove used for cooking in the IHDS-I and IHDS-II respectively. The stove options included 3 types of biomass solid fuel stoves, and a general 'modern stove' category which could represent kerosene, LPG, or electric stoves. In the IHDS dataset 5358 households switched from using a biomass stove as their primary stove in 2004-5 to using a 'modern' non-biomass stove in 2011-12 representing 16.27% of households (14.94% when adjusted by sampling weights).

Studies on energy transition often use some form of logit or probit regression model to perform a predictive analysis identifying the trends and effect of a given set of variables on appliance ownership (Dhanaraj et al., 2018), or adoption rates of electricity or LPG (Kemmler, 2007; Farsi et al., 2007; A. Sharma et al., 2019). Recently Rao and Ummel (2017) used a form of ensemble technique called a Boosted Regression Tree (BRT) model to analyse the effect of a range of household characteristics on the uptake of so-called 'white good' appliances. A comparison of the predictive capability of these two modelling approaches found that the BRT model on the whole outperformed the logit model in predicting appliance ownership (Rao and Ummel, 2017).

The aim of the approach in this Chapter is to provide a greater level of descriptive or explanatory analysis using a dual approach to identify the existence of different transition pathways amongst households. The first stage involves predictive modelling using a Boosted Regression Tree model - an ensemble machine learning technique - to identify factors that are determinants of clean cooking transition and assess performance of this model. The second stage focuses on descriptive data analysis using hierarchical clustering, where the key determinants identified in the predictive modelling are used to identify the different groups of households that did switch stove and the different combinations of features that characterise each group. The first stage of the analysis uses a training subset of 25,000 of the 32,922 households from the IHDS to identify the influence of variables on the propensity of a household to switch from a solid fuel biomass stove to a cleaner 'modern stove' as their main cooking stove. The predictive performance of the ensemble learning regression and a conventional probit regression are assessed and compared using the remainder of the dataset which was not used to train the model. The secondary stage of analysis uses agglomerative hierarchical clustering to cluster the 5,358 households that did switch from biomass to a 'modern' non-biomass stove. By comparing the effect of key determinants identified by the predictive modelling and the defining characteristics of the clusters of stove-switching households, it is possible to identify the different combinations of key determinants enabling stove transition in each cluster.

Variable selection was carried out using both correlation and random forest analysis to identify the most relevant variables. Given the inter-related nature of the socio-economic and cultural variables of interest in the dataset, it was important to identify and address any significant multi-collinearity in the dataset before performing any analysis. A Farrar-Glauber test was conducted to identify and address any multi-collinearity (Farrar and Glauber, 1967). In particular, fuels used exclusively for cooking showed cross-dependent correlation with one another, so redundant fuels were removed from the selected variables. In addition, the number of different region categories was reduced by reassigning households in states in the centre region to the neighbouring eastern region as there was little distinction between these two. The descriptive statistics of the resulting independent variables are shown in Table 4.1 (except profession, caste, and region which are non-continuous, and non-binary).

Table 4.1: Descriptive statistics for variables

Independent variable	**Mean**	**Median**	**Min.**	**Max.**
Income per capita (INR/month)	2401	1363	0	346750
Urban	0.330	-	0	1
Time in Place (years)	78.41	90.00	0.00	90.00
Female Education (years)	5.395	5.000	0.000	16.000
Permanent House	0.704	-	0	1
Flush Toilet	0.392	-	0	1
Piped Water Availability (hours/day)	1.845	0.000	0.000	24.000
Dairy Spend (INR/month)	196.40	100.00	0.00	8600
Electricity Availability (hours/day)	13.11	14.00	0.00	24.00
Electricity Consumption (kWh/month)	93.68	54.50	0.00	1977.40
Kerosene Consumption (kWh/month)	28.01	24.79	0.00	587.00
Change in fuel collection time (min)	-3.475	0.000	-320.000	450.000
Cooking appliance ownership	0.291	0.250	0.000	1.000
IT appliance ownership	0.310	0.429	0.000	1.000
Change in Female TV Time (hours/day)	0.687	1.000	-12.000	14.000

Chapter 3 detailed the use of Probit and BRT regressions used for the analysis in this chapter. The Probit regression provides a baseline of the typical quantitative predictive model, assuming that the individual's decision to switch from a biomass stove to a non-biomass stove is based on a latent variable which represents some measure of utility, then this variable can be defined as a linear function of the independent variables. In contrast to this recall that a BRT model has no a priori specification of functional form, and does not assume linear effects. As explained in Chapter 3 specification of BRT hyperparameters (No. of trees, tree depth/complexity and learning rate) can be difficult. N-fold cross validation and the recommendations of Elith et al. (2008) were followed to optimise parameters to produce an accurate model and minimise risk of over-fitting. The BRT model presented used a tree complexity of 5, and a learning rate of 0.01, with 4100 trees fitted.

For the second stage of the analysis hierarchical clustering was used. This is an unsupervised machine learning method designed for community detection, i.e. it can be used to identify subsets within a dataset that have similar characteristics based on the connectivity between data points. A benefit of hierarchical clustering algorithms for such descriptive analysis is that the iterative process produces a clear tree like structure of clusters which offers a more intuitive view of the clustering process and easier analysis of results, although the iterative nature of the algorithm makes it inefficient for extremely large datasets (Kassambra, 2017). An agglomerative hierarchical clustering algorithm was used with Ward's linkage criterion (Ward, 1963) and the Gower distance measure (Gower, 1966) for categorical variables as it produced a clear and distinct cluster structure. Cluster number determination used the enhanced average silhouette width method of Fujita et al. (2014) described in Chapter 3.

4.3 Predictive Modelling Results

4.3.1 Boosted regression tree model

The BRT analysis outputs both the relative importance of variables shown in Figure 4.1 and the marginal effects of the independent variables shown in Figures 4.2, 4.3, and 4.4. Figure 4.1 shows all independent variables were found to have nonzero relative influence ranging from 1-12%. Use of kerosene and electricity both have an influence of around 11%, while cooking equipment ownership shows an 8.5% influence, and IT appliance ownership a 7.1% influence. The region a household is in has a 10.7% influence and the profession of the head of the household has an influence of 9.3%. Income per capita of the household does have an influence of 8.5% but the BRT shows it is not the dominant determinant of a household's switch to non-biomass stoves. The marginal effects for each variable shown in Figures 4.2, 4.3, and 4.4 exhibit one of three different types of response: either a constant response (for categorical variables), a threshold response, or a multiple threshold (multiple regime) response.

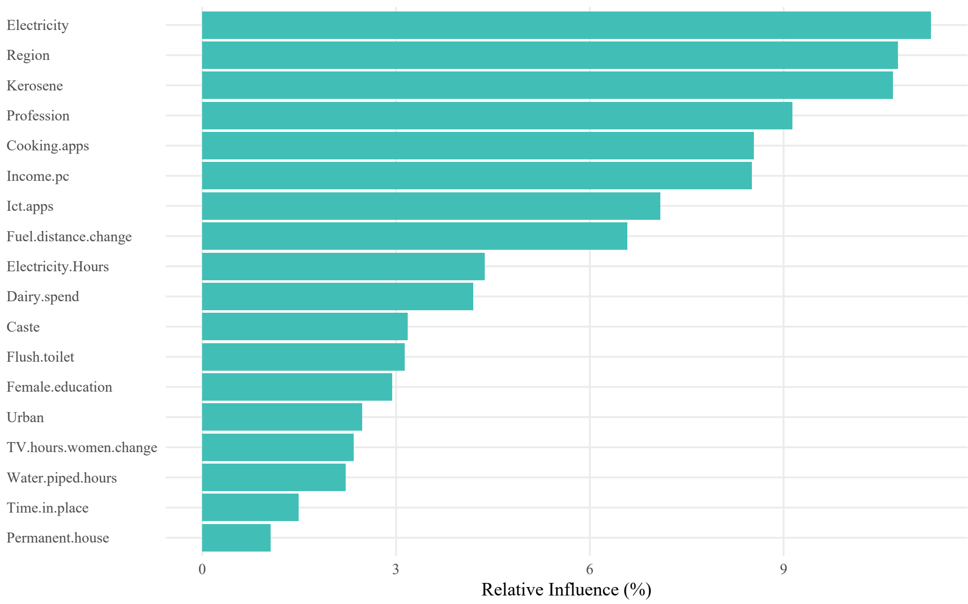


Figure 4.1: Relative influence of independent variables in BRT Model

The constant marginal effects observed for categorical variables shows that these variables will be key determinants of modern stove switching for only some households. For example, region is one of the more relatively influential variables, with North-Eastern states being associated with a markedly higher probability of switching stove, while households in the South have a slightly higher chance of switching than households in the East, North and West where region is a determinant of minor influence. This difference could be the result of local policy or climate differences; for example, the southern states are typically wealthier relative to the national average, and southern states such as Tamil Nadu and Karnataka have led development in renewable energy infrastructure in India (Schmid, 2012). North Eastern states have lower incomes and with historically lower access to infrastructure (Ghosh and De, 1998), the geography of this region also results in greater local availability and dependency on biomass fuel compared to other regions (Bhatt et al., 2016). LPG distribution infrastructure development under 'Vitrak Yojana' between 2005 and 2011 benefited many poorly serviced settlements in North Eastern states. In the work of Sankhyayan and Dasgupta (2019) a statistically significant relationship between region and LPG use was not found, however the coefficients from their model are compatible with the marginal effects from this analysis.

Figure 4.2: Marginal effect of constant effect independent variables on probability of a household switching from Biomass to LPG

The profession of the head of the household was also found to be of greater relative influence, although the marginal effects were only significant for some professions as shown in Figure 4.2. Those in skilled trades, artisans, salaried employment, or collecting pensions or rent all had a greater probability of switching, whereas those in agricultural wage labour, and unskilled work were less likely to switch. Kemmler (2007) found that more labour intensive and 'daily wage' type employment was associated with lower electricity use, and Sehjpal et al. (2014) found that, in rural India, households whose head was in more formal employment had a greater likelihood of the household transitioning to clean cooking. This may be related to the frequency of payment with the former group of jobs being associated with regular monthly or weekly pay whereas income can be more erratic for the latter group.

A measure of household infrastructure is provided through variables measuring permanent house construction, and availability of flush toilets shown in Figure 4.2 and both show a small positive increase in marginal effect on the switch to a non-biomass stove with greater levels of access. Rao and Ummel (2017) similarly found that better dwelling quality had a positive relationship with ownership of refrigerators and TVs, and Ahmad and Puppim de Oliveira (2015) showed that access to piped water was associated with clean cooking. Permanent housing, while having the lowest relative influence of the variables in the dataset, did have a positive marginal effect on the switch to a non-biomass stove. These findings suggest that access to public utilities and quality of the household's immediate built environment are important, as Debnath et al. (2019) found in their study of rehabilitated slum housing in Mumbai.

Figure 4.3 shows the marginal effects of variables which exhibit a threshold response, namely hours of electricity supply and years of education of the head female of the household. The marginal effect of hours of electricity supply on switching behaviour shows a constant effect up until 15 hours of electricity supply per day, after which the marginal effect increases with hours of electricity. Rao and Ummel (2017) similarly found that hours of electricity supply had a positive relationship with ownership of refrigerators and TVs, and Ahmad and Puppim de Oliveira (2015) showed that access to electricity was associated with clean cooking uptake. The threshold observed at 15 hours could be indicative of the added convenience or reliability of having electricity available for two thirds of the day, encouraging investment in appliances or changing household practices related to cooking.

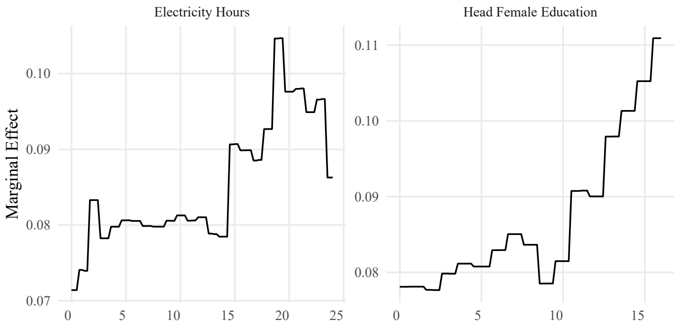


Figure 4.3: Marginal effect of threshold response independent variables on probability of a household switching from Biomass to LPG

Education of the head female of the household also displays a threshold response as seen in Figure 4.3. Households whose head female has 10 or more years of schooling, i.e. completing some level of secondary or tertiary education, has a greater probability of switching to a 'modern stove'. A recent study by A. Sharma et al. (2019) found a significant relationship between education and LPG uptake for households in the eastern states of Chattisgarh and Jharkhand, while Ahmad and Puppim de Oliveira (2015) found female education to be a significant determinant of non-biomass cooking in non-slum households. In their study, Sankhyayan and Dasgupta (2019) found that in urban areas there was a stronger positive association between female literacy and LPG use, especially for households where the female head of the household had more than 9 years of schooling, and they suggest this difference is a result of female literacy not translating into female empowerment as effectively in rural households.

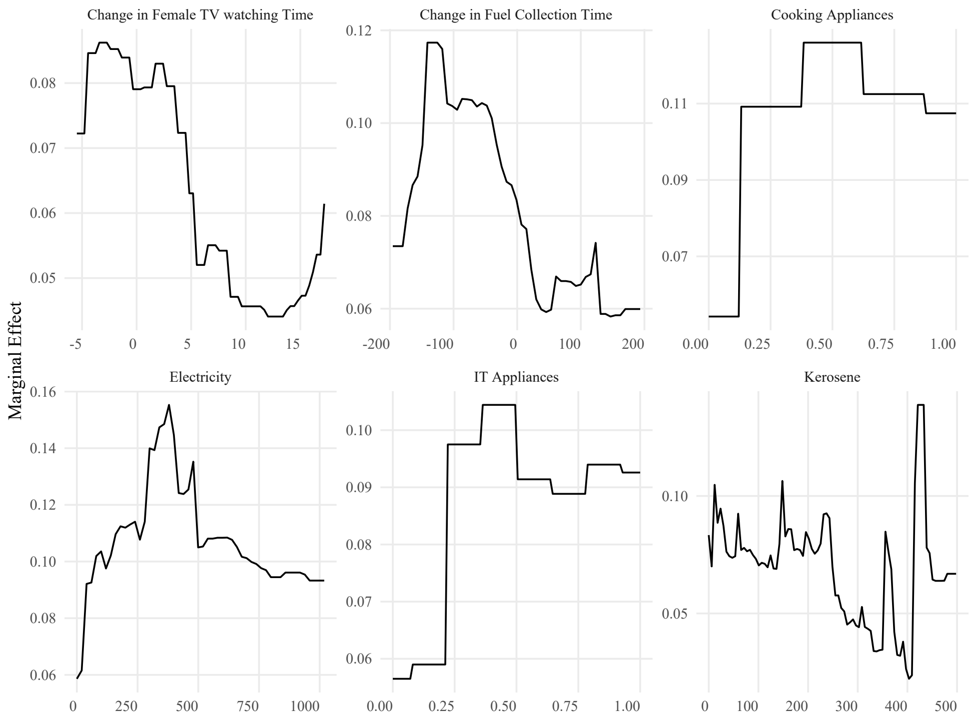


Figure 4.4: Marginal effect of multiple threshold response independent variables on probability of a household switching from Biomass to LPG

Figure 4.4 shows the marginal effects of variables with multiple thresholds, or different regimes, where marginal effect follows different trends within given ranges. LPG and biomass fuels are used fairly exclusively for cooking. In contrast electricity and kerosene have a range of different end uses. Use of these fuels can indicate transition to cleaner energy for other household activities which offers an explanation for the high relative influence of these variables. Figure 4.4 shows that low levels of electricity consumption are associated with a low marginal effect on the probability of a household switching, but this marginal effect increases to a higher level with increasing electricity consumption up to a level of 500kWh/month. Beyond this electricity has a negligible effect on the probability of switching as households using more electricity than that almost certainly have transitioned to clean cooking, with over 80% of such households in the IHDS using no biomass fuel at all. Similarly, kerosene use up to 200 kWh leads to a greater probability of a household switching whereas above 200 kWh the marginal effect is negligibly low, indicating reduced chance of switching. Households using more than 200 kWh of kerosene are likely using it for cooking, hence already have a non-biomass stove. The noisy behaviour between 350 kWh and 500 kWh is likely due to households switching from a biomass stove to a kerosene one, which counts as a 'modern stove' switch in the IHDS. The different marginal effect thresholds show how related energy practices of the household shape the observed energy consumption and how these practices have inter-dependencies, as Bisaga and Parikh (2018) found in their study.

Appliance ownership can serve as a proxy for energy use by a household as appliances are used to deliver a particular energy service. Figure 4.4 shows how increasing ownership of IT and cooking appliances increases the probability of a household having switched to a non-biomass stove. Rao and Ummel (2017) found that refrigerator and television ownership was associated with greater LPG use by a household, which suggests households with such appliances also have clean cooking facilities. Greater appliance ownership could also signal better access to markets or shops, as well as better availability of electricity. However there are two thresholds, as the marginal effect plataeus for households with average ownership, and drops off at high ownership levels. Households with very high levels of appliance ownership are more likely to already use LPG and thus the greatest marginal probability of switching occurs for households with middling levels (40-60%) of ownership.

Time spent collecting fuel and watching TV in a household, shown in Figure 4.4, offers some quantification of household practices as a measure of time allocation to given practices. A decrease in time spent collecting fuel of up to 130 minutes is associated with a greater probability of a switch to a 'modern stove', and decreases in time spent collecting fuel beyond 130 minutes have a relatively low marginal effect on the chance of a household transitioning. An increase up to 50 minutes is associated with a decreasing probability of switching and increases in fuel collection time above 50 minutes see the lowest probability of switching. Similarly the change in number of hours spent watching TV by the adult women of the household has a small positive association for small decreases and increases, but larger increases beyond 5 hours of TV viewing are associated with a lower probability of a household stove switching. The marginal effect of changes in energy practices surrounding energy use and clean cooking transitions are characterised by multiple thresholds. Additionally the marginal effects of these two variables quantitatively indicate that there is a change in the time allocated to energy related practices in a household that switches stove. This is important as it implies that characteristics of the stove and its usage have an impact on the practices of a household. Debnath et al. (2019) found that characteristics of household appliances in Mumbai slums had a significant effect on the practices of the household.

An aside on the role of women in decision making

While this thesis does not take a specific focus on the role of gender, this national dataset does offer an opportunity to gain some insights into women's role in clean cooking transition. As shown above, changes in women's activities are associated with changes in likelihood of transition. The BRT analysis for this chapter used education of the head female of the household as one of the independent variables. However this aside shall explore the relative influence of women in household decision making by using the same BRT model substituting female education for the difference between male and female education in the household (education of the head male of the household - education of the head female of the household).

This alternative independent variable has a similar relative influence in the model as female education (3-4%), and the marginal effect curve is shown in Figure 4.5.

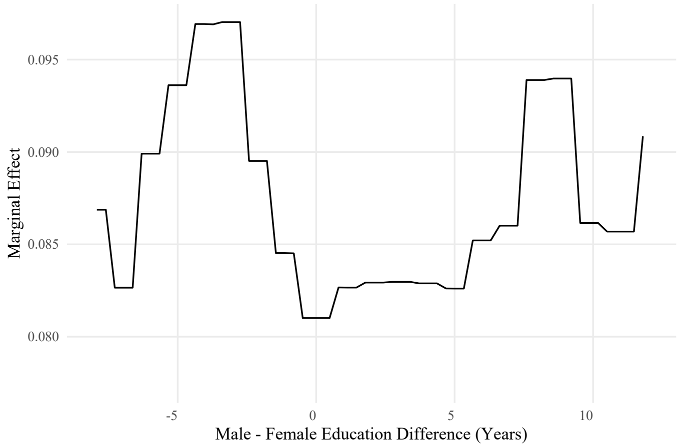


Figure 4.5: Marginal effect of Male-Female Education Differential on probability of a household switching from Biomass to LPG

Positive values of the education differential variable represent the number of additional years of education of the head male of the household relative to the head female of the household and vice-versa. Figure 4.5 shows that households where the male head of the household has several more years of education have a lower likelihood of adopting a non-biomass stove, while in households where the woman has several more years of education there is an increased likelihood of non-biomass stove adoption. It is important to note that the effect difference is very small, but this could relate to women's role in household decision making. Might women with more years of education than their husbands be better able, or more likely, to advocate for acquiring a cleaner non-biomass stove?

Obviously this variable can't be considered in isolation; professions or livelihood activities of these women may also play a role in shaping cooking fuel decisions in the household (detailed information which is not included in the IHDS). Furthermore, correlation does not necessarily imply causation, for example greater female education relative to male education could be common to households with higher overall levels of education. If that is the case, is the overall education level the determinant of clean cooking adoption and not the woman's education level relative to the head male of the household? Despite these caveats the results of the BRT with respect to women's relative education and daily activities reflect an association between differences in women's roles and clean cooking adoption that could be indicative of the impact of women's changing influence in household decision making. Gender dynamics in decision making around clean cooking merit further research, particularly given that as noted in Chapter 1 the PMUY scheme targets the head female of the household as the beneficiary. Understanding what influence women have and how it manifests itself is important to design successful and inclusive interventions to eliminate solid biomass fuel cooking.

4.3.2 Probit model

The coefficients of the probit regression model are shown in Table 4.2. The same subset of 18 variables used in the BRT model were included in this model, although several profession categories encompassing smaller proportions of the population, as well as caste, were not found to have a significant effect (at $p<0.1$) on fuel switching and are not included in the table. A key difference between the outputs of the probit and BRT models is that while the BRT provides relative importance and marginal effect plots, the probit model provides coefficients, standard errors, and confidence intervals denoted by statistical significance levels which can make the process of evaluating the model more straightforward. Comparing the coefficients in Table 4.2 with the relative importance and marginal effect plots from the BRT model in Figures 4.1, 4.2, 4.3 and 4.4 shows that many of the coefficients and marginal effects for the categorical variables such as region, permanent housing, profession, and flush toilet availability show compatibility with respect to influence on stove switching.

Table 4.2: Table of coefficients from probit model of switch to a non-biomass stove

	Dependent variable	
Independent Variable	**Coefficient**	**Standard Error**
Region: North	-0.008	(0.046)
Region: North East	1.078^{***}	(0.089)
Region: South	0.494^{***}	(0.045)
Region: West	-0.091^{**}	(0.045)
Income per capita	-0.000006^{*}	(0.000003)
Time in place	0.003^{***}	(0.001)
Female education	0.003	(0.004)
Profession: Agricultural wage labourer	0.811^{*}	(0.478)
Profession: Artisan/Skilled	1.036^{**}	(0.483)
Profession: Pension/Rent	0.847^{*}	(0.477)
Profession: Petty shop	1.038^{**}	(0.477)
Profession: Salaried	0.949^{**}	(0.477)
Profession: Wage labourer	0.958^{**}	(0.478)
Pucca house	0.352^{***}	(0.037)
Flush toilet	0.266^{***}	(0.035)
Water piped hours	0.007^{**}	(0.003)
Dairy spend	-0.00004	(0.0001)
Electricity hours	0.010^{***}	(0.002)
Electricity	0.00001	(0.0001)
Kerosene	0.001^{**}	(0.0004)
Fuel distance change	-0.002^{***}	(0.0004)
Cooking appliance ownership	0.311^{***}	(0.085)
IT appliance ownership	0.916^{***}	(0.111)
TV hours women change	-0.010	(0.008)
Constant	-2.817^{***}	(0.540)
Pseudo R^2	0.115	
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

However there are some key differences between the outputs, particularly those which have a non linear effect in the BRT model. For example, while the BRT identified the use of complimentary fuels as being significant, the probit regression does not find any significant effect. Similarly, the marginal effect plot for electricity use in Figure 4.4 shows that the marginal effects vary with the level of respective fuel use. This non-linear relationship cannot be captured by the probit regression. Conversely, while the probit regression correctly identifies significant effects for variables such as cooking and IT appliance ownership, distance travelled for fuel, and hours of electricity supply, it does not capture the threshold identified by the BRT beyond which the marginal effects of these variables are reduced or negligible.

4.3.3 Comparison of predictive performance of BRT and Probit Models

Using the test subset of the dataset as inputs to each of the two models, predictions of whether a household would switch to a non-biomass 'modern stove' or not were calculated and compared to the actual stove switching outcome in the dataset. Table 4.3 shows the classification tallies of each model as well as three measures of predictive performance: the percentage of correctly classified households (a higher score indicates better predictive ability); the AUC score indicating discriminative ability of the model (a higher score indicates better predictive ability); and the Brier score which is an indication of both calibration and discriminative ability of the model (a lower score indicates better predictive ability).

The BRT model outperforms the probit model on all three measures particularly on its discriminatory ability, although the results are comparable. This is a reflection on the ability of the tree-based ensemble method to model non-linear effects. Indeed many of the independent variables had non-linear marginal effects. Thresholds for non-zero effects are a reflection of the non-linear nature of prac-

Table 4.3: Results of indicators for comparison of predictive performance of BRT and Probit Model

	Model	
	BRT Model	**Probit Model**
Correct classification	84.9%	83.5%
AUC	0.823	0.731
Brier Score	0.108	0.126
True positive	214	103
False negative	1138	1249
False positive	89	116
True negative	6866	6839
Note:	test subset of 7922 households	

tices and decision-making concerning household energy use. Figure 4.6 illustrates this difference between the probit and BRT model using the example of cooking appliance ownership. As shown, both models follow the same positive trend with greater appliance ownership and have similar marginal effects at the mean. However, for specific households the probit model will either under- or overestimate the effect of appliance ownership compared to the BRT model. Nevertheless, the probit regression offers the benefit of simplicity, which can make communicating results to a non-technical audience straightforward. Additionally, the assessment of compatibility of results via statistical significance can help easily validate and compare results - although statistical significance is not necessarily the most useful approach to assessing results (Amrhein et al., 2019). On the other hand, it should be noted that the outputs of the BRT offer a visual and intuitive way of conveying the variation in marginal effect and the existence of threshold levels.

While the measures of performance provide a metric for the calibration and discriminatory ability of each model, the rates of true and false positives and negatives for each model shown in the bottom half of Table 4.3 point to a problem of such models. For both the probit and BRT models the number of false negatives, that

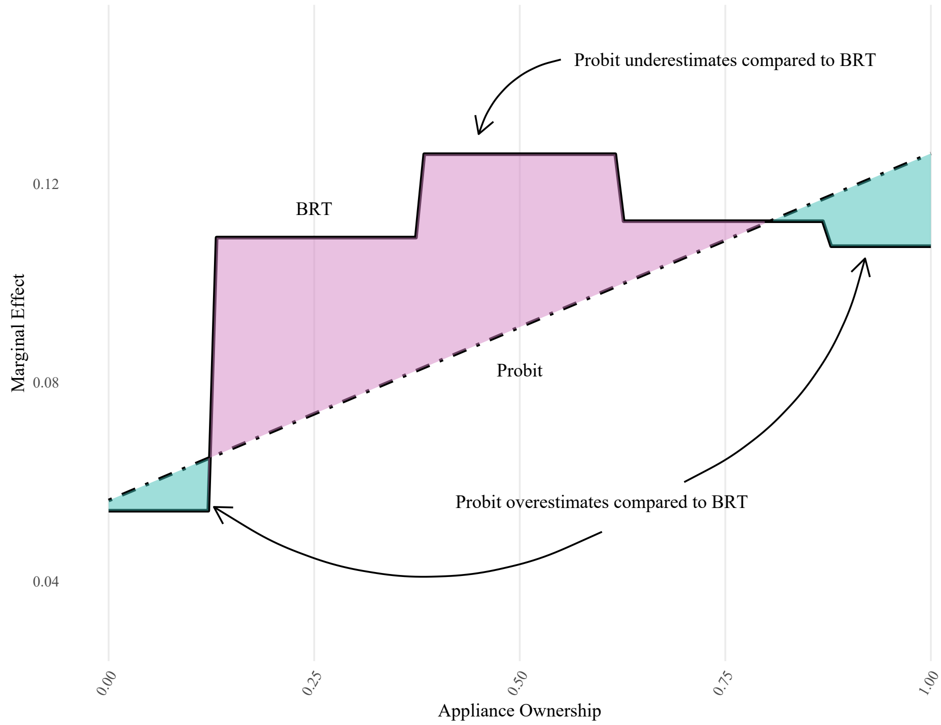


Figure 4.6: Illustration comparing the Marginal Effect of Cooking Appliance Ownership from Probit and BRT Models. The figure is meant for illustrative purposes only and values may not exactly match the results listed above.

is the households that the model predicted would not switch but did switch in reality, accounts for over 84% of switching households in the BRT model and 92% of such households under the probit model. This suggests that while these models are good at predicting households that did not switch (true negatives compared with false positives), they perform poorly at predicting households that do transition. Households that transition against the expectation of the model point to the existence of alternative transition pathways not captured by either model. These pathways are defined by characteristics that individually would ordinarily not be sufficient to be drivers of transition, but when present in specific combinations can allow households to overcome other barriers.

4.4 Descriptive Modelling Results

To better understand these differing transition circumstances, the variables shown in Table 5.3 were used in an agglomerative hierarchical clustering analysis on the subset of households that did switch their main stove from solid fuel biomass stoves to a clean non-biomass stove between 2004-5 and 2011-12. The clustering analysis identifying nine distinct clusters of households all of which had transitioned away from primarily using a biomass stove but with different combinations of defining characteristics. The resulting dendrogram is shown in Figure 4.7, and the mean characteristics of each cluster are shown in Table 4.4.

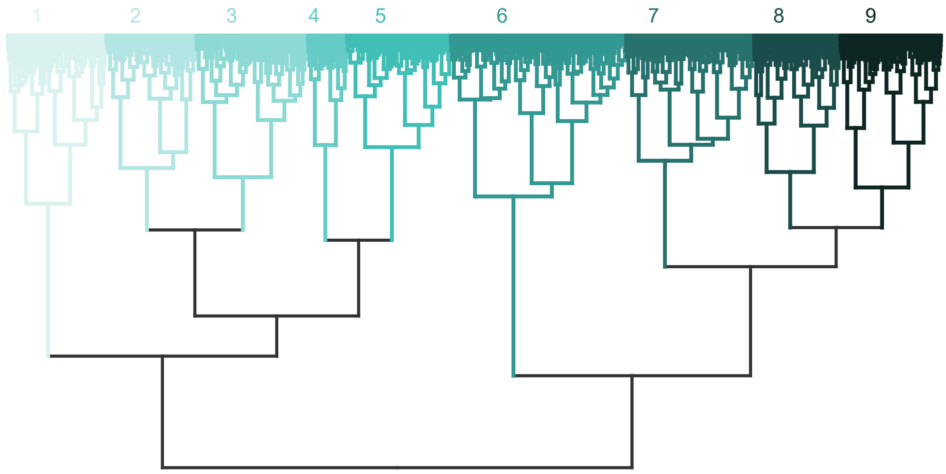


Figure 4.7: Dendrogram of Hierarchical Clustering with IHDS Biomass to LPG switching households

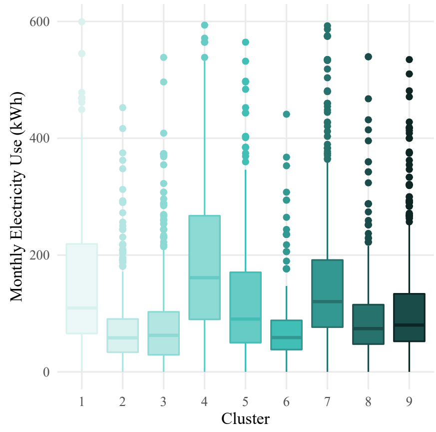
The diversity of characteristics between clusters is notable as it suggests that there is no one single combination of determinants that results in a transition to clean cooking fuels, and points to the different and complex transition pathways that Kroon et al. (2013) discussed. A comparison of clusters 1 and 2 detailed in Table 4.4 and shown in Figure 4.8 serves to illustrate a rural case of such different transition pathways: households in cluster 1 have a mean income of 55,058 INR, and are nearly all Northern rural households. They have good provision of water and electricity, with near ubiquity of flush toilets, permanent housing, above average appliance ownership and electricity use, as well as above average levels of female education. This group represents households that score highly on most of the key determinants, and a higher proportion of these households were correctly predicted to have switched stove by the BRT model.

Table 4.4: Mean characteristics of clean cooking transition clusters

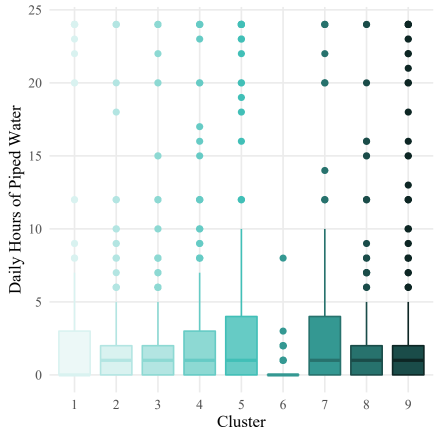
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Number of Households	597	640	565	735	497	223	577	518	1006
Region (most represented)	North	South	South	North	North	North East	East	West	South
Income	55058.20	33054.90	27915.67	46401.94	29561.38	48916.25	40480.51	35244.72	37380.36
Caste	Fwd/Gen	OBC	OBC	OBC	OBC	Fwd/Gen	Fwd/Gen	OBC	OBC
Time in Place	85.18	82.87	79.91	64.58	70.67	75.81	63.07	84.60	71.86
Urban (%)	0.01	0.02	0.38	0.99	0.96	0.54	0.98	0.01	0.48
Female Education	8.52	5.89	5.57	7.96	5.98	10.05	8.46	6.92	8.51
Pucca House	0.98	1.00	0.00	0.99	0.96	0.78	1.00	0.94	0.98
Flush Toilet	0.99	0.02	0.41	1.00	0.06	0.79	0.57	0.44	0.99
Water Piped Hours	2.53	1.70	1.98	2.22	4.91	0.35	3.04	1.86	3.23
Monthly spend on dairy	461.24	220.90	162.93	301.77	194.66	355.35	186.90	185.29	154.35
Electricity Hours	15.14	12.55	12.96	15.79	15.56	6.67	19.61	16.59	17.98
Electricity	166.48	74.36	80.94	228.55	130.85	96.90	167.94	95.86	110.81
LPG	172.18	136.10	132.90	181.68	141.11	215.12	145.49	128.91	130.25
Biomass	466.67	522.18	423.21	159.73	327.45	360.39	131.21	598.30	380.47
Kerosene	18.00	21.64	31.77	22.61	29.11	47.52	32.84	34.54	16.43
Change in Fuel Distance	-6.46	-9.67	-9.78	-2.30	-4.32	9.86	-4.50	-21.76	-5.06
Cooking Appliances	0.54	0.36	0.32	0.50	0.37	0.43	0.46	0.41	0.52
IT Appliances	0.45	0.38	0.34	0.42	0.38	0.43	0.43	0.37	0.46
Female TV Viewing Hours	0.91	0.85	0.85	0.63	0.42	1.55	0.26	0.98	0.21
Correct BRT Prediction	12.8%	8.1%	10.0%	4.7%	13.7%	76.6%	3.6%	12.6%	28.2%
Correct Probit Prediction	0.6%	0.6%	0.0%	1.0%	1.6%	64.1%	0.0%	0.0%	24.2‘%

In contrast households in cluster 2 have a lower mean per capita income of 33,054 INR, lower female education levels, lower prevalence of flush toilets, fewer hours of piped water and electricity access, lower electricity and higher average biomass consumption while having lower appliance ownership. However, households in cluster 2 all have permanent housing, have been settled for, on average, over 80 years and still have better than average availability of electricity and water. This suggests that despite their lower income these households still have access to a better than average level of physical infrastructure, but their high biomass use relative to cluster 1 suggests that there is a higher prevalence of fuel stacking in households of cluster 2.

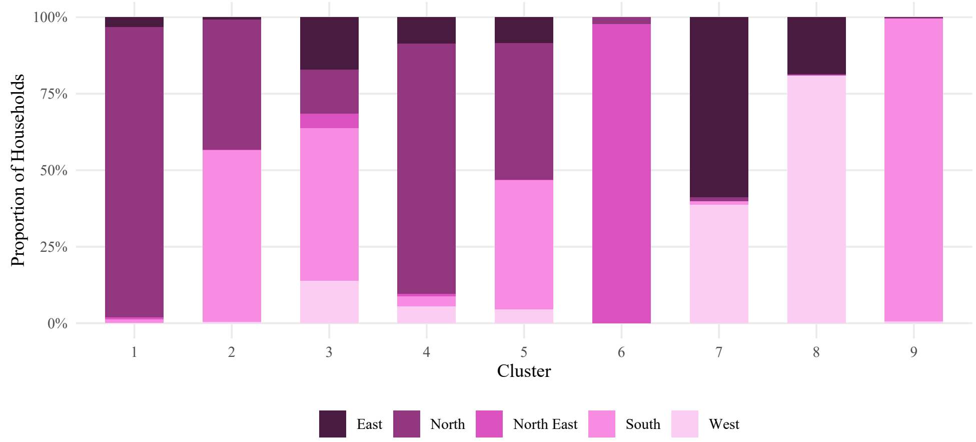
The existence of different transition pathways is also exemplified in urban clusters 4 and 5. Cluster 4 represents above average income households with a per capita annual income of 46,401 INR, and above average education of the head female of the household, access to flush toilets, hours of electricity, and appliance ownership. Cluster 4 also largely represents northern urban households. Conversely, households in cluster 5 are also urban, but have markedly lower mean per capita annual income of 29,561 INR, and low prevalence of flush toilets, lower electricity consumption, lower levels of head female education, appliance ownership, while mean biomass consumption is double that of cluster 4. Cluster 5 have a high proportion of households employed in stable jobs, and have equally good availability of water and electricity as those households in cluster 4, as well as being settled in their current neighbourhood for longer and containing more Southern households. These longer established households with steady employment are likely to have stronger communities with good 'social infrastructure', with better relationships and sharing of information between neighbours. A greater proportion of households in cluster 5 were correctly predicted to transition by the BRT model as they score highly on key determinants such as region, profession, and



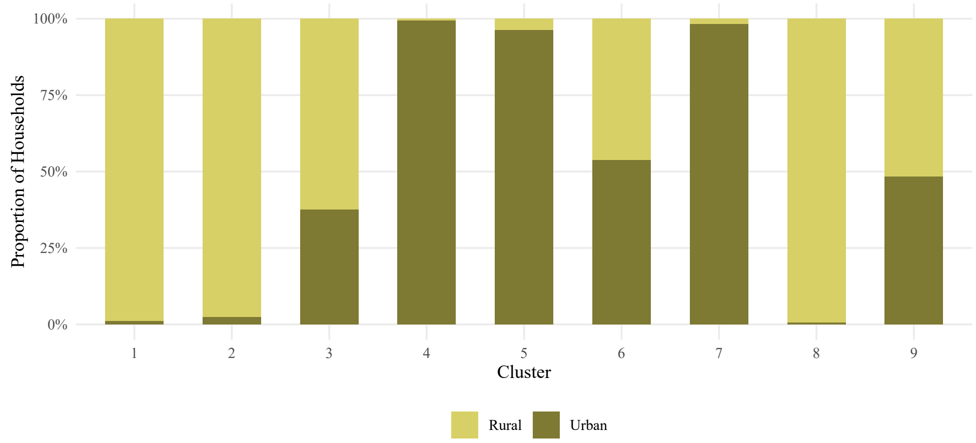
(a) Electricity Use



(b) Piped Water Supply



(c) Region



(d) Rural - Urban Split

Figure 4.8: Key explanatory variables by cluster for households that have switched from Biomass to LPG

change in fuel distance while not lagging too far behind the mean on other key determinants. These households are likely to have a higher prevalence of fuel stacking as evidenced by the higher mean biomass use, where biomass fuels may offer a back up fuel when LPG is not available, or in months when household income needs to be spent on other priorities.

While the predictive models had good performance in predicting non-switching households they did not perform well identifying households that did switch. It is interesting to note the uneven distribution of correct model predictions across the clusters; the rate of true positives for each cluster varies considerably. The probit model fails to predict a significant proportion of transitions in any cluster but 6 and 9 where it correctly predicted 64.1% and 24.2% of stove transitions respectively. These are the clusters which score highly in nearly all the determinants and are easy identification targets for the model. The BRT model does correctly identify a low percentage of stove switching in several other clusters but similarly performs best at identifying stove switching households in clusters 6 and 9. This bias in the predictive regression models would also imply that interventions based on the findings of such models would exhibit some bias in the households they support versus those they miss out.

Access to both local physical infrastructure - indicated by variables including housing quality, hours of electricity, piped water availability, and flush toilet availability - and/or social infrastructure - indicated by variables including years since migration, caste, and profession - seem to be important in transition to clean cooking. Clusters other than 6 and 9 do not score highly on nearly all determinants, but score highly on different combinations of determinants related to physical and social infrastructure. All of these combinations facilitated a transition albeit following a different pathway. This analysis sheds light on the defining features of each of these transition pathways, and these can indicate key challenges for policy to address to promote adoption and continued use of clean cooking fuel, and reduce dependency on solid biomass fuels. Table 4.5 summarises four distinct typologies or transition pathways observed across the nine clusters, and likely policy challenges associated with households on each pathway.

LPG consumption in rural and peri-urban households such as those on pathway B in Table 4.5, as well as urban and peri-urban households on average incomes in pathway D suggests that canisters are being refilled less than once a month (a regular 14.2kg canister has approximately 200kWh worth of LPG, so monthly use below this indicates non-monthly refills). This indicates that LPG is not being used to meet all of a household's cooking needs, and this could point to either issues of affordability and cash flow with households unable to afford more frequent refills, or issues of supply and delivery which may particularly affect rural and peri-urban households. A. Sharma et al. (2019) observed that doorstep delivery of LPG canisters in an area increased LPG usage in Chhattisgarh, and noted that often rural households had to collect their cylinders from the distributors. This analysis indicates that further policy intervention may be needed to address this issue for rural and peri-urban households.

Table 4.5 also shows that some households may require less policy intervention to encourage uptake of LPG because they exhibit many shared features with households that already use LPG as described for pathway A. Their good access to infrastructure within an urban setting and stronger financial position means that they may only need a small nudge to switch to LPG. These groups also appear more likely to discontinue biomass use once using LPG, and thus need less policy intervention to transition away from biomass stove use. Similarly, some households may face region specific challenges that require additional local intervention, such as north-eastern households of pathway C in Table 4.5, who despite above average socio-economic circumstances and regular LPG use are still dependant on biomass fuel. This could be, in part, due to different practices with biomass being used for non-cooking needs, or biomass serving as a more reliable alternative to less reliable supply of cleaner fuels.

Table 4.5: Key pathway features of households that transitioned to LPG and relevant policy issues

Path	Key Pathway Features	Cluster	Policy Issues
A	Urban pucca households, above average income, appliance ownership, electricity use and good infrastructure access. Low residual biomass use	4,7	Many households of this group will have already switched, and do not continue use of biomass fuels
B	Rural pucca households with average to above average income, electricity use, and good infrastructure. Above average reduction in fuel collection	1,2,8	LPG uptake has not stopped regular biomass use. Infrequent LPG refill suggests supply or cost issue. LPG has delivered notable time savings.
C	North eastern households, above average income and appliance ownership but poor infrastructure and low electricity use. Substantial use of biomass fuel.	6	Heavy continued reliance on biomass, perhaps due to regional fuel supply challenges or local practices.
D	Urban or peri-urban households with average income, either in pucca or non-pucca housing but good infrastructure access, and continued biomass or kerosene dependence	3,5,9	Biomass and kerosene may be relied on as backup. Possible cash flow or supply challenges persist.

An important policy challenge common to several clusters is the continued regular use of biomass fuels amongst a majority of households that have adopted a non-biomass stove. Across the country it appears that rural, and peri-urban households that switch to non-biomass stoves still use considerable amounts of biomass fuel, as do urban households on below average income. For many of these households it may be a way of coping with unreliability in supply or access to LPG, or as a means of reducing consumption of LPG to manage household cash flow. It could also arise due to energy related behaviours which favour biomass use, such as a preference for rice over bread. Crucially, the prevalence of fuel stacking to manage energy services as described by Kroon et al. (2013), means it is important to recognise that even once non-biomass stoves are adopted, further intervention to change energy related behaviours or improve reliability of LPG supply may be needed to reduce biomass use by these households. Understanding the energy practices and decisions leading to such fuel stacking behaviours requires an understanding at a local household level such as demonstrated by Khalid and Sunikka-Blank (2017) in order to enable policy interventions to promote greater uptake of sustained clean cooking among such households.

4.5 Limitations and Caveats

This analysis does face a limitation due to the nature of the IHDS dataset which is representative at the national level. It serves to make some crucial comparisons between regions and states, but cannot locate where households on a particular transition pathway are beyond their region and rural or urban designation. Differences in the non-income drivers that determine clean cooking transitions and the interaction between physical and social infrastructure and household energy practices all take place at a local scale. Thus, while the analysis above does indicate the existence of differing transition pathways or typologies, it cannot accurately characterise these at the more local scale necessary for stakeholders and local government to act upon.

Another issue with this analysis arising from the use of the IHDS dataset is that it does not consider temporality of energy use. There are two different aspects to this limitation, one concerns daily or monthly patterns of use, and the other concerns seasonal effects. While this analysis uses a panel dataset it does not offer time series data which forestalls any analysis of daily use profiles. With respect to seasonality, the IHDS does not include relevant data on seasonal patterns of energy use and questions on consumption were phrased so as to ask about consumption in the previous 30 days. While offering a coarse temporal view of change in energy use over a six year period, the lack of higher resolution temporal data is an important limitation of this analysis.

Larger sample size surveys at a city scale could be used to identify and characterise the different transition pathways of different groups of households in that city. Additional data on the current fuels used, different energy end uses within a household and time of use, as well as aspirations of households would be invaluable. In addition, such analysis would benefit from some qualitative interviews with households discussing their energy practices and decisions to provide context to the data. For example, this could provide an understanding of the non-monetary trade-offs considered by households when switching to LPG. Chapter 5 shall explore the use of such mixed city scale data.

It is also worth noting that while the selection of variables for the regression models in this analysis ensured multicollinearity assumptions were reasonably adhered to, the model variable selection was better suited to the BRT model than the probit model given that the latter was primarily presented for comparative purposes. Acknowledgement must also be made that there are methods of including more complex effects in linear regressions including the use of interaction terms, however these tend to make interpretation of results less intuitive, particularly when using a large number of predictors. Similarly, other quantitative methods for analysing residential energy transitions have been used such as structured equation modelling, willingness to pay, and utility models, but these fundamentally do not differ much from regression approaches and rely on similar assumptions relating to linearity and rational utility maximising consumers that have been shown here to be oversimplifications.

4.6 Conclusions

Current quantitative approaches to characterising residential energy transition in models used in the study of energy transition tend to rely on use of predictive regression methods. While these are powerful tools for identifying trends and statistical significance of determinants, applications of such methods assume homogenous utility maximising behaviour and often linearity of effects. This can lead to an overemphasis on the role of cost in energy transitions, that can overlook the influence of local socio-cultural context, household practices, and behaviours. This chapter used ensemble machine learning predictive modelling and descriptive clustering methods in a two part analysis to identify households that might be missed by current predictive regression models and policies informed by these. This characterised the influence of socio-economic and cultural determinants of clean cooking transition, and the nature of that influence, while also exploring and characterising the limitations of current quantitative predictive approaches in fully describing this.

A range of socio-economic and cultural factors were found to be relevant determinants of non-biomass stove adoption including region, profession, complementary fuel consumption, appliance ownership, income, changes in time spent carrying out energy related activities, infrastructure access, and education. Several determinants displayed a threshold relationship with stove switching, and were only influential determinants of stove switching beyond a given value - for example availability of electricity above 15 hours a day - was associated with increased stove switching. The influence of other determinants was characterised by multiple thresholds or regimes, for example appliance ownership of both cooking and IT appliances had a plateau of greatest marginal effect for households with ownership between 10 and 50% with slightly lower probability of fuel switching for households with higher appliance ownership and negligible chance of switching below this range.

By being able to handle the non-linear influence of determinants the BRT model performed better than the more conventional probit model in predicting whether households switched. However, both models performed relatively poorly in identifying the households that did switch compared to those that did not. Furthermore, the clustering analysis showed that while there were nine clearly distinguishable groups of household that had switched, merely two of those clusters accounted for nearly all the households correctly identified by the predictive models. Each of these clusters is defined by different combinations of key determinants, and the households not in those two clusters represent those more likely to slip through the net and be missed out by predictive models and policies informed by such models. While they all adopted a non-biomass stove these households had followed different pathways to transition.

Different household energy transition pathways require different policy interventions. Location specific variation in infrastructure, climate, and cultural factors influence adoption of clean cooking fuels, such as seen in the North East of India, indicate a need for locally adapted interventions to promote household transition to cleaner energy. This supports a conclusion of Kebede et al. (2002) that local variations must be factored into the design and tailoring of clean energy policies. Amongst households that do adopt non-biomass stoves, those in rural areas, as well as urban households on lower incomes, are likely to continue using considerable amounts of biomass. This finding adds to recent voices (Gould and Urpelainen, 2018; Mani et al., 2020; A. Sharma et al., 2019) which have been making the case that policy interventions promoting uptake of LPG for cooking in such households must pair LPG promotion with interventions to reduce biomass use in order to deliver on the public health aims of clean cooking. This is an issue that shall be revisited in later chapters.

A few lessons can be drawn from this chapter which are relevant to addressing the main research question of this thesis. First, the influence of socio-economic determinants does not follow a linear trend for all households and across the range of values, instead featuring threshold values defining different regimes - although within a given threshold defined regime effects do exhibit more a linear relationship. Secondly, there are different typologies of transitioning household featuring specific combinations of socio-economic characteristics - they represent different transition pathways for households. This finding shall be important to the development of the people-centric quantitative modelling approach in Chapter 6. The analysis of this chapter was limited by the resolution of the data which was aggregated at a state and national scale, while the influence of socio-economic variables of interest takes place at a more local level. To precisely characterise and distinguish different transition pathways in a way that may be instructive to stakeholders and decision makers necessitates greater resolution of data at a local scale. The following chapter describes a novel cluster based approach to characterising residential transition pathways using quantitative and qualitative data, and examines the results of such an approach applied to low-income households in the city of Bangalore.

5 Data and People

The contents of this chapter have been published in a journal article in Sustainable Cities and Society and a methods paper is under review. A P Neto-Bradley worked on conceptualisation, methodology, data curation, investigation, and visualisation; R Rangarajan collaborated on qualitative methodology, data curation, and investigation; R Choudhary and A Bazaz provided supervision and with inputs on conceptualisation.

André Paul Neto-Bradley, Rishika Rangarajan, et al. **(2021)**. "A clustering approach to clean cooking transition pathways for low-income households in Bangalore". *Sustainable Cities and Society* 66, p. 102697. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102697

André P. Neto-Bradley et al. **(2021)**. "Energy transition pathways amongst low-income urban households: A mixed method clustering approach". en. *MethodsX*, p. 101491. ISSN: 2215-0161. DOI: 10.1016/j.mex.2021.101491

5.1 Introduction

Current quantitative modelling efforts do not effectively model the role of needs, practices, and local context of individuals on energy use and clean cooking transitions in Indian households as shown in the previous chapter. Chapter 4 shows how a range of socio-economic factors and household practices influenced clean cooking transition, with many of these determinants having non-linear influence and displaying thresholds for affecting transition. More pertinently the comparison of descriptive and predictive analyses shows that transitioning households exhibit a clustering tendency where different typologies of households are grouped based on their particular combination of socio-economic features. These groups indicate the existence of different transition pathways for clean cooking that current predictive modelling approaches are unable to discern. However while national datasets indicate the existence of these in the previous chapter, the effects of socio-economic and cultural context act at a local scale and so to accurately characterise such pathways requires analysis with higher resolution of data at a local scale.

As discussed in Chapter 2, quantitative technical and economic approaches tend to treat all households the same way which is an oversimplification that can lead to biased decision making, as shown by the analysis in Chapter 4. On the other hand social science approaches consider households on an individual basis offering great detail and understanding, but such an approach is impractical to apply at scale. Clustering offers a pragmatic compromise. Where there is rich high-resolution data on households available, clustering can characterise patterns of heterogeneity at scale across households by identifying clusters with distinct circumstances, practices, barriers. Each cluster represents a segment of households that will follow a particular transition pathway to sustained clean cooking. The definition of each segment or transition pathways as they are referred to in this thesis, can in turn support tailoring and targetting of interventions to the needs and challenges of each specific group of households by policy makers.

This chapter will analyse primary data collected in Bangalore with a novel approach based on clustering methods to explore and characterise the nature of energy transition pathways among low-income households using mixed quantitative and qualitative data. This addresses the second research objective outlined in Chapter 1. Previous studies such as those by Yu, Haghighat, et al. (2011), Yu, Fung, et al. (2011) and Carmo and Christensen (2016) have made use of clustering methods to distinguish building occupant energy use profiles. Clustering has also proved valuable for identifying energy user profiles for electricity using smart meter data (Wei et al., 2018; Kwac et al., 2014). In other applications clustering has been used to tailor policy recommendations such as a study by Qin et al. (2019) investigating differences in drivers of carbon dioxide emissions between cities. A two stage clustering approach has been recently used to characterise residential gas use in the UK (Yuan and Ruchi Choudhary, 2020), and the previous chapter showed, in an Indian context, how clustering approaches could identify different types of residential gas adopter using socio-economic data.

Building on these approaches, a clustering approach is detailed that integrates quantitative survey and qualitative interview data to identify dominant socio-economic characteristics, and practices, that act as barriers to clean cooking across a community of households within a uniform range of income. The city of Bangalore is used as a case study, for which the different clean cooking transition pathways followed by low-income households are defined. The dataset consists of the subset of 389 Bangalore households from the survey dataset introduced in Chapter 3 (those with incomplete expenditure responses were excluded for this analysis), as well as follow-up interview data from a different sample of 23 households. The names and locations of the 7 wards surveyed in Bangalore are shown in Figure 5.1. Households surveyed represent low socio-economic backgrounds, with per capita incomes below the median income for Bangalore and mostly in the bottom quartile of incomes with a mean of 4458 INR/pp/per month (ca. USD 50).

Use of mixed data yields in-depth insights into the energy practices, decisions, and circumstances of individual households. This will show some groups of low-income households will be more at risk of failing to transition to clean cooking and face barriers to access. The variations arise due to their socio-economic status, lo-

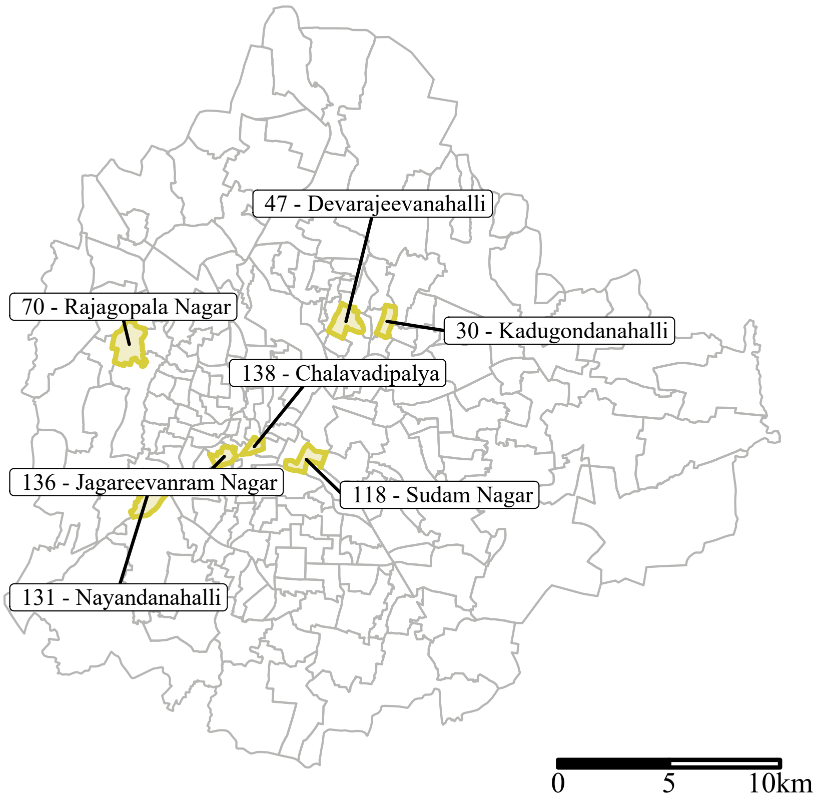


Figure 5.1: Bangalore city wards where low-income households were surveyed and interviewed.

cal socio-cultural context, community infrastructure, and practices, and these features are used to infer different transition pathways. How the detailed characterisation of these pathways can inform the design of tailored policy interventions will be explored with a view to facilitate clean cooking transition amongst low-income households. Consideration will also be given to what this means for developing a quantitative energy model that incorporates local social context.

The remainder of this Chapter follows with section 5.2 which introduces in detail the novel mixed data pathway clustering method. Section 5.3 presents descriptive statistics of the survey sample, before analysing the results of the first stage of the mixed data pathway clustering analysis. This is followed in section 5.4 with a presentation of the transition pathways identified and a discussion of the characteristics and insights offered by these. Consideration of limitations of the method and findings are given in section 5.5 and conclusions are presented in section 5.6.

5.2 Method

5.2.1 Mixed Data Pathway Clustering

Studies on drivers of clean energy transitions and energy access amongst low-income households in the Global South typically make use of either purely quantitative methods (such as predictive regression analysis discussed in Chapter 4), or qualitative methods (such as in-depth interviews), to the exclusion of the other. However energy access and the practices and decisions related to a household's energy use involve a complex interaction of local social, economic, cultural, and community features which can only be understood through both qualitative and quantitative data. To address this a simple yet efficient approach was developed combining qualitative data analysis with statistical clustering to identify links between qualitative information and quantitative data and thus infer different energy transition pathways followed by low-income urban households.

This method involves two separate stages of community detection using two datasets collected from the same geographic area. The schematic in Figure 5.2 provides an overview of the method which uses both a quantitative dataset containing household level socio-economic and energy use data, and a second qualitative dataset consisting of semi-structured interviews on household energy practices and decision making. Each dataset is clustered to identify common groups, and then the interview respondents are matched to a quantitative survey cluster. A second stage of clustering is performed on the quantitative and qualitative cluster membership of the interview respondents to identify the distinct energy transition pathways amongst these households, defined by socio-economic and energy use characteristics and associated narratives.

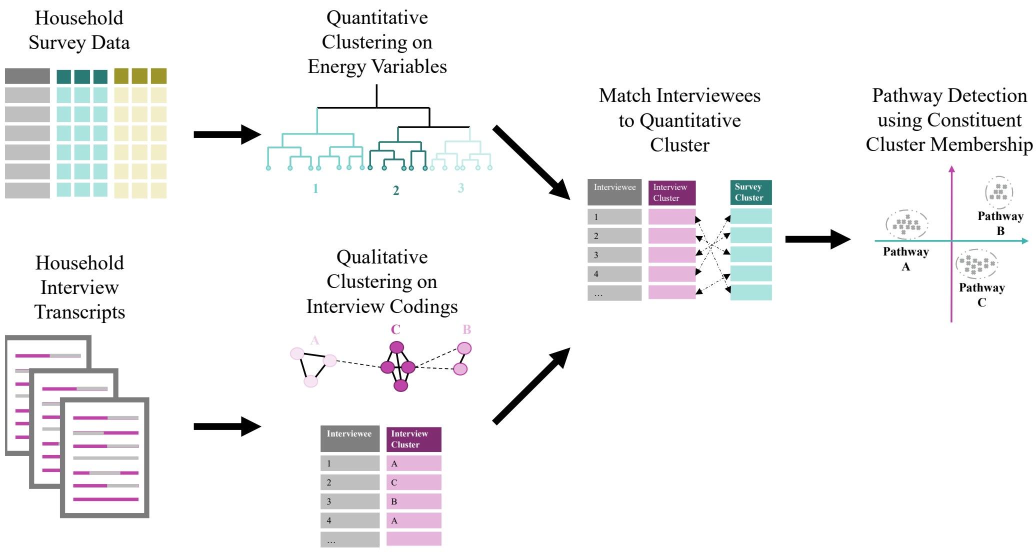


Figure 5.2: Schematic overview of mixed method cluster analysis - illustrating how first stage statistical clustering of quantitative and qualitative data is combined to identify distinct transition pathways.

This method requires a dataset consisting of quantitative survey data and in-depth qualitative interviews collected from the same geographic area of interest - the case study in this chapter will use the primary data collected in Bangalore detailed in Chapter 3, although this method could also make use of suitable secondary data.

5.2.2 First Stage Clustering Analysis

The first clustering step involves a separate cluster analysis of both the qualitative interview data and quantitative survey data to identify common groups or communities amongst respondents on the basis of their energy use habits, decisions, and socio-economic and cultural circumstances. Hierarchical clustering methods were used for first step analyses, although slightly different approaches were required given the difference in data types. Figure 5.3 shows the structure of the survey and interview datasets. The survey dataset contains a wide set of socio-economic and energy use variables, although only energy use variables will be used for clustering with the former used to characterise clusters. A conventional agglomerative hierarchical clustering method is used for community detection in the quantitative survey data. A grounded theory approach (Glaser, B.G and Strauss, A.L, 1968) is used to analyse the interview data, inductively identifying themes and concepts. Codification of the transcripts provides data for a graph-based correlation clustering analysis, as shown in Figure 5.3 the qualitative dataset for this analysis takes the form of a table of interview codes.

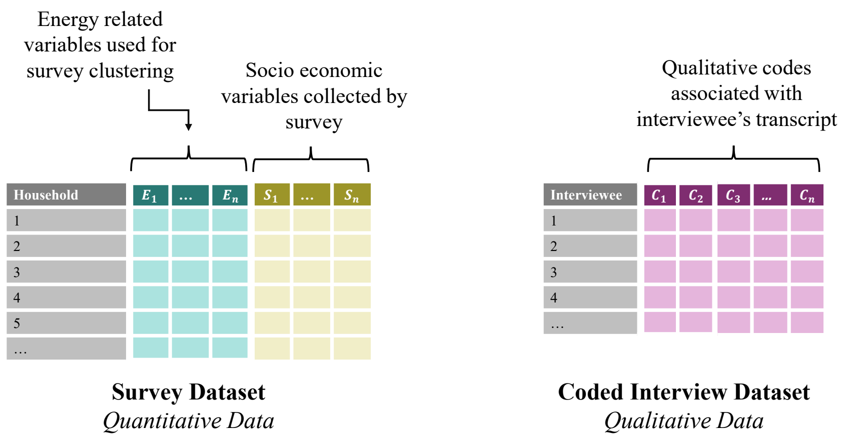


Figure 5.3: Schematic of mixed data structure. Two distinct datasets are required: one contains qualitative household level survey data with a mix of energy and socio-economic variables, the other dataset consists of a table of codings from household interviews.

Quantitative Survey Clustering

Variable Selection & Engineering

Variable selection is carried out to single out relevant variables and address multicollinearity in the dataset which can make it difficult to identify relevant variables and quantify their effect. Correlation coefficients help select variables which have a significant correlation with clean fuel use. A Farrer-Glauber test (Farrar and Glauber, 1967) is used to identify multi-collinearity and where variables have a causal relationship, the less relevant variable was excluded from the dataset. Some variables from the survey data were combined to facilitate clustering analysis and reduce the number of binary variables in the dataset. Specifically this involved creating compound appliance ownership variables where appliances were grouped by type (IT & Communication, Cooking), and variables were created to denote the percentage of each type owned. Table 5.1 shows the energy use related variables selected from the survey data in the Bangalore case study on which to cluster the surveyed households.

Table 5.1: Selected energy use variables from survey dataset for use in Bangalore case study clustering.

Variable	**Unit/Type**	**Name in dataset**
Monthly LPG Use	kWh/month	*lpg_kwh*
Monthly Electricity Use	kWh/month	*electric_kwh*
Monthly Kerosene Use	kWh/month	*kerosene_kwh*
Daily Electricity Availability	Hours/day	*electricity_hours*
Hours of Cooking	Hours/day	*cooking_hours*
Hours of Lighting	Hours/day	*lighting_hours*
IT Appliance Ownership	%	*it_appliances*
Cooking Appliance Ownership	%	*cooking_appliances*
Government LPG Support Awareness	Binary	*programme_awareness*
Cooking Location	Categorical	*cooking_location*

Hierarchical Clustering

The primary clustering of the survey data is performed using hierarchical clustering. Agglomerative hierarchical clustering produced more balanced clustering in this case, although it is advisable to try both agglomerative and divisive methods to determine which produces a more balanced set of clusters with a clearer optimal number of clusters. Details of the clustering algorithm and determination of the optimal number of clusters are explained in Chapter 3. For analysis of the Bangalore survey data, five clusters were identified as the optimal number of clusters.

Qualitative Interview Analysis

Qualitative Data Analysis

The analysis of the qualitative interview data uses a grounded theory approach to qualitative data analysis, which as defined by Glaser and Strauss (2017) is concerned with the systematic discovery of theory from data that both fits real world scenarios and can be easily understood by stakeholders. This approach is particularly relevant in this analysis, because as explained by Corbin and Strauss (2014) it can provide a common language of concepts which stakeholders can engage with to address energy access issues, and is key to the identification and characterization of energy transition pathways and problems.

A codified approach to analysing qualitative data in an inductive manner is important to convey credibility and understand how narratives and pathways are derived from the data (Glaser and Strauss, 2017) using coding of the interview transcripts to quantify key discussion points and content. This is a form of quantizing as described by Sandelowski (2000) and involves reducing interview transcripts into variables that can be associated with each interviewee. This allows for the combined clustering of the qualitative and quantitative data through the second step clustering. Despite this quantizing, selected quotes from the interviews are still included in the analysis and discussion to support the clustering results.

A first run of coding interview data is sometimes referred to as open coding and is carried out to identify concepts from the data (Corbin and Strauss, 2014), using line-by-line analysis. Following this initial coding the concepts identified are analysed to determine the categories that these concepts might fall under. Detailed codes are deduced from the open coding to form a list of second level codes, while common properties of certain concepts are used to help define broader first level codes. The transcripts are labelled with these first level codes indicating categories, and then a second run of coding is carried out on the interview transcripts by the team of researchers using the refined set of more detailed second level-codes to narrow in on a more specific categorization of the coded section of the transcript (J. L. Campbell et al., 2013).

The interviews were coded and analysed using the 'RQDA' package in R (Huang, 2014), which provided a graphical interface for the coding process and facilitated export of datasets to the R environment for analysis alongside the quantitative survey data. The coded interviews were peer-reviewed to eliminate bias of the individual researchers, and disagreement between coders was addressed by an additional round of coding assessment drawing on subject-specific expertise in line with recommendation from previous studies (J. L. Campbell et al., 2013; Kondracki et al., 2002) with new codings being decided on by consensus. This coding exercise was carried out in collaboration with Ms. Rishika Rangarajan at the Indian Institute for Human Settlements.

Correlation Clustering

To identify different clusters amongst the interviewed households on the basis of transcript codings a correlation clustering approach is used. The details of this method including the 'fast greedy' modularity optimising clustering algorithm are detailed in Chapter 3. In this analysis a correlation threshold of 0.3 is set such

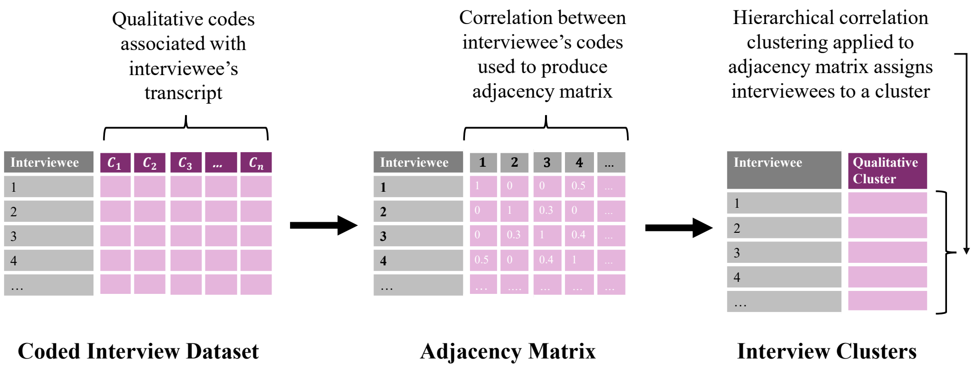


Figure 5.4: Schematic of interview coding clustering process. Coded interview table is transformed into an adjacency matric by calculating correlation between respondents. This adjacency matrix is then clustered to assign interviewees to respective clusters.

that any interviewee correlation below this was set to zero, increasing clarity in the graph by removing weak and negative links.

5.2.3 Second Stage Pathway Identification

The second stage clustering analysis combines the information gained through the separate community detection of the qualitative interview and quantitative survey data. This is done to identify commonalities between these that characterise distinct energy transition pathways with unique combinations of socio-economic characteristics and narratives. This involves first matching interviewed households to a quantitative survey cluster before performing a secondary clustering of these households on the basis of their quantitative and qualitative cluster membership.

Interviewee Survey Cluster Matching

The qualitative and quantitative data analysed in stage one yields two sets of clusters defined by different features and variables. In order to map one set of clus-

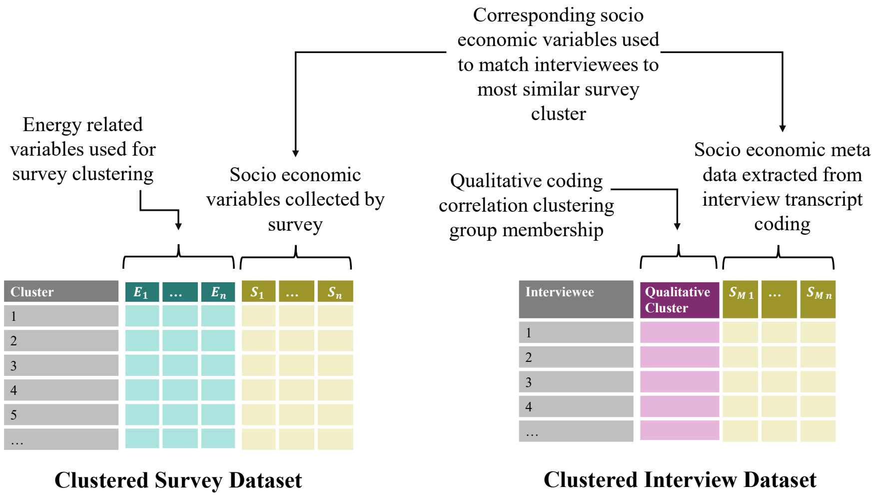


Figure 5.5: Schematic of interviewee survey cluster matching on the basis of socio-economic variables.

ters to the other, one set of clustered respondents must be matched to their closest clusters in the other dataset. To do this each of the interviewed households is matched to one of the survey clusters such that there are a set of households which have both an interview and survey cluster assigned. In theory this could be done the other way around, however it would require interviewing all the survey households and would be impractical for large sample sizes. To match the interviewed households to a survey cluster, common categorical variables need to be extracted from the interview transcript to create a metadata tag with household characteristics which can be compared to the survey cluster centroids. This is performed during the qualitative data coding described above, and the schematic in Figure 5.5 shows how the variables listed in Table 5.2 are used for household matching.

Using these categorical matching tags each interviewee is assigned to the most similar survey cluster using Euclidean distance to measure similarity. This can accommodate any number of matching variables n, using equation 5.1 where d is the Euclidean distance, n is the number of dimensions, and x and y are a pair of

Table 5.2: Table of socio-economic variables used for interviewee to survey cluster matching.

Variable	Unit/Type
Time since Migration	Years
Income Frequency	Categorical
Primary Cooking Fuel: LPG	Binary
Primary Cooking Fuel: Kerosene	Binary
Primary Cooking Fuel: Biomass	Binary
Majority Religion	Binary
Legal Electricity Connection	Binary

points representing the interviewee and the cluster centroid. The variables used to match the interviewees to a survey cluster are listed in the Table 5.2 and cluster centroid values used for distance measurement are based on mean values for each quantitative survey cluster.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5.1}$$

Cluster Membership Clustering

Once assigned to a quantitative survey cluster each of the interview households will have membership of both a quantitative and qualitative cluster. The interview households are then clustered on their qualitative and quantitative cluster membership. A k-means clustering approach is used to perform this clustering – the same agglomerative hierarchical clustering used in the first step could also be used but given the tendency for these clusters of clusters to group into distinct non-overlapping and spherical clusters, the simpler k-means approach is well suited. Figure 5.6 shows the process of second stage clustering, and exemplifies how pathways can be characterised by drawing upon the quantitative variable ranges of the associated quantitative clusters, and codes, concepts, and quotes from the interviewees in the respective interview cluster. Together these can define a transition pathway and provide a narrative and common language of concepts for understanding the energy access challenges of the given pathway.

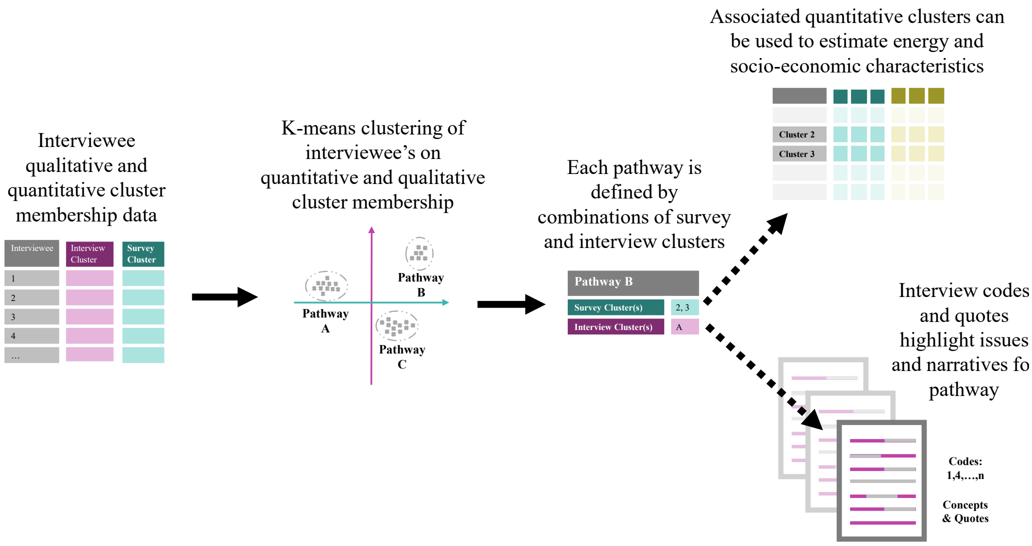


Figure 5.6: Schematic of second stage clustering and pathway characterisation. K-means clustering of the interview households using their cluster membership as variables identifies pathway groups with unique pairs of cluster membership

In order to determine the ideal number of clusters for the second stage clustering the elbow method is used alongside the silhouette width method as discussed previously, although given the low number of variables assessing a scatter plot of the cluster membership of these households will also show the clear divisions between groups of households. By using descriptive statistics for each of the constituent quantitative clusters a summary of likely mean values and ranges for energy use patterns and socio-economic characteristics can be calculated for each pathway. This information is augmented by linking these to the key narratives identified in the associated qualitative interview cluster.

5.3 Results and Analysis

5.3.1 Descriptive Statistics

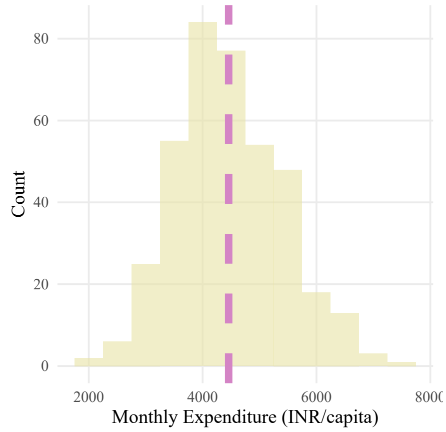
Despite the homogeneously low income levels of households surveyed there is a wide range of socio-economic characteristics, appliance ownership, and fuel use amongst the households in the Bangalore survey sample as shown in Table 5.3 and Figures 5.7. The majority of households surveyed were either 1st, 2nd, or 3rd generation migrants to Bangalore, and just over 30% migrated to Bangalore in the last 30 years. 96% of all households surveyed live in slums, with 54% living in notified slums with legal tenancy and 42% in non-notified slums without legal tenancy.

Table 5.3: Descriptive statistics for continuous variables

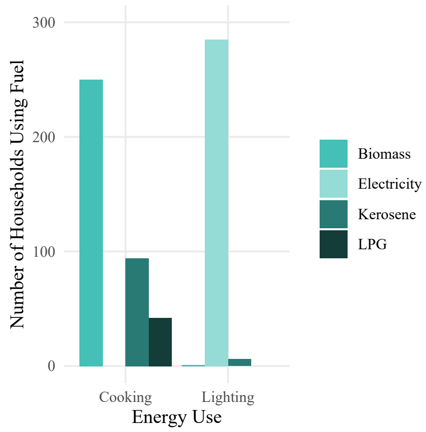
Independent variable	**Mean**	**Median**	**Min.**	**Max.**
Expenditure per capita (INR/month)	4458	4400	2000	7500
Time in Place (years)[a]	35.29	31.00	1.00	90.00
Electricity Availability (hours/day)	13.11	14.00	0.00	24.00
Electricity Consumption (kWh/month)	93.68	54.50	0.00	1977.40
Kerosene Consumption (kWh/month)	28.01	24.79	0.00	587.00
Cooking appliance ownership[b]	0.291	0.250	0.000	1.000
IT appliance ownership	0.685	0.500	0.000	1.000
Hours of Cooking	3.42	4.00	1.00	7.00
Hours of Lighting	5.21	5.00	1.00	9.00

[a] Followed the IHDS convention of counting years in place up to 90 years. 90 indicates a household has been in place for 90 or more years.
[b] Cooking appliance ownership of 0.000 indicates household uses open fire or 'three stones' stove.

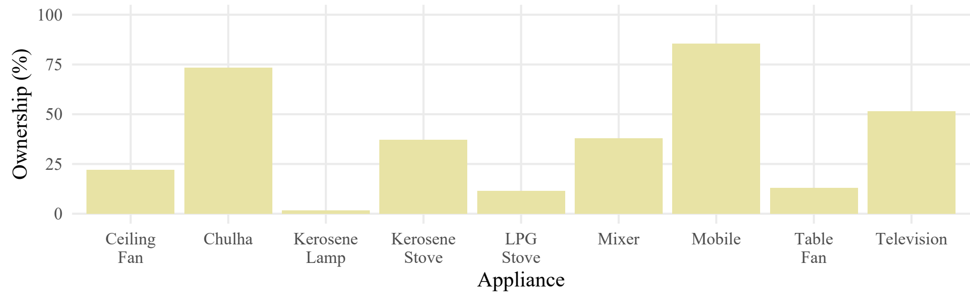
With respect to fuel use, all households used biomass for some of their cooking needs with the majority still being primarily dependant on such fuels at the time of our survey. Over 20% used kerosene to meet some portion of their cooking needs and a mere 9% of household primarily used LPG. This is in stark contrast to lighting, where over 65% of households used electric lighting and almost no household uses kerosene or biomass/candles - although over a third of households did not have some form of reliable powered lighting in their homes. This distinction in



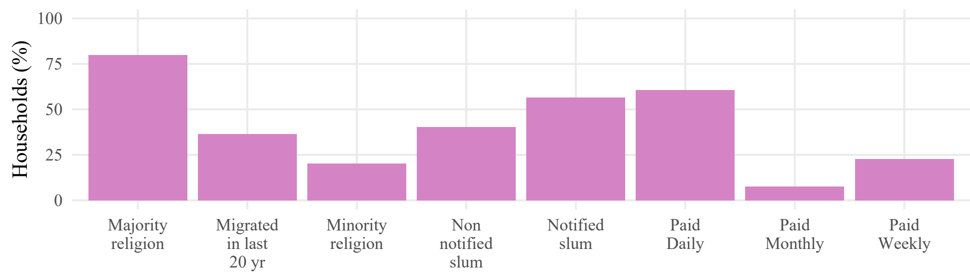
(a) Monthly exp. distribution & mean



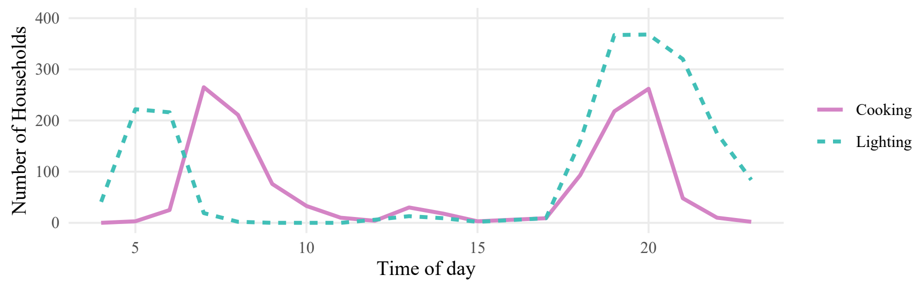
(b) Primary fuel for lighting and cooking



(c) Appliance ownership levels



(d) Socio-economic characteristics of households



(e) Time of use for cooking and lighting

Figure 5.7: Summary plots of key characteristics, appliance ownership, and energy use of all survey households.

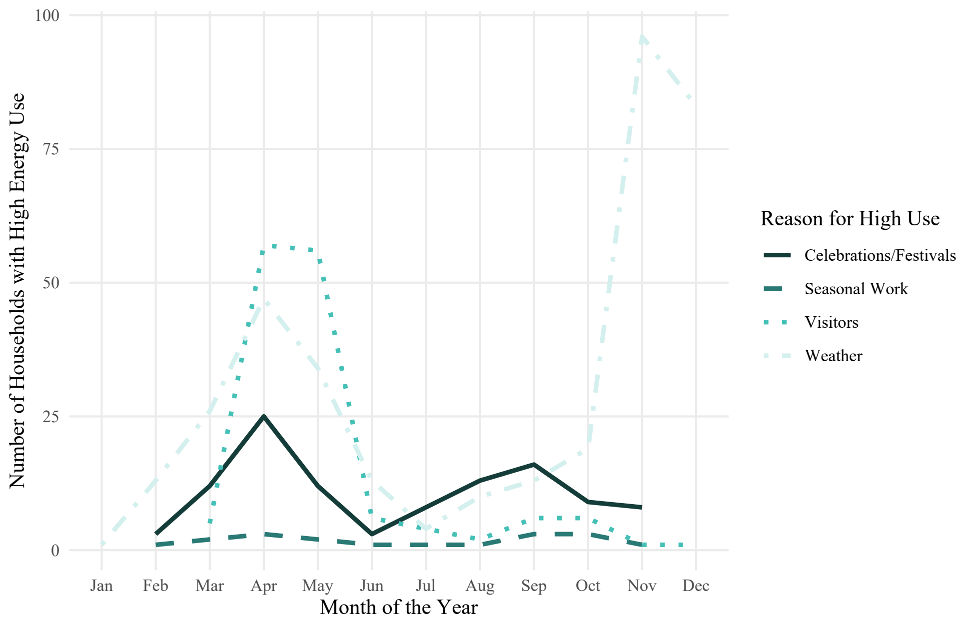


Figure 5.8: Summary of seasonal variations showing months of reported highest energy consumption differentiated by primary driver of increased consumption

level of clean fuel use for these two activities is reflected in appliance ownership. The vast majority of household's ownership of electrical powered IT appliances (particularly mobile phones) was relatively high with 82% of households owning a mobile phone, however fewer households owned cooling equipment such as table fans. For cooking appliances - only 39% of households owned a mixer and 38% owned a kerosene stove.

The time-of-use for cooking and lighting shown in Figure 5.7e have some similarities in that for both there is a morning and evening peak with little or no daytime use - this fits with residential energy demand profiles in other studies on India (Sandwell et al., 2016). It is curious to note however that while the evening cooking and lighting peaks coincide, the morning cooking peak lags the lighting peak by about two hours. Given that sunrise in Bangalore is usually within an hour of 6:00am, this explains why the lighting peak dwindles and indicates that households are awake and active for a while before doing any cooking on a stove.

Chapter 4 noted how as well as a lack of time series data, the IHDS had no information on seasonality of energy use by households. Figure 5.8 shows the months of highest energy consumption reported by households surveyed, disaggregated by the primary reason for this high consumption. There is a distinct seasonal variation with peaks in March-May and again in the August and September due to Festivals/Celebrations, and slightly later in November and December due to weather. A major driver of high consumption reported in the spring peak was having visitors staying in the household. Seasonal work accounted for a relatively minor proportion of household with increased consumption. This demonstrates how household energy use habits fluctuate throughout the year and cannot be assumed to be static.

5.3.2 Survey and Interview Analysis

Figure 5.9 shows the dendrogram of the first stage agglomerative clustering of survey data in which households are split into five distinct clusters of different sizes, with three large clusters accounting for 76% of households. A further two smaller clusters of similar size account for the remaining 24% of households. While the clusters are not evenly sized they are distinct from one another. Further separation would result in less distinctive clusters, while less separation would result in a larger and more heterogeneous 4th cluster with greater in-cluster variance. The distinct subsets of households identified through the clustering show features which can be interpreted to identify possible transition behaviours, defined by distinctly different socio-cultural characteristics despite having the same financial means. Variable means for each cluster are shown in Table 5.4, and Figure 5.10 plots ranges of appliance ownership, hours of electricity supply, and time in place as well as proportions of households aware of subsidies, living in slums, and in daily wage jobs for each cluster.

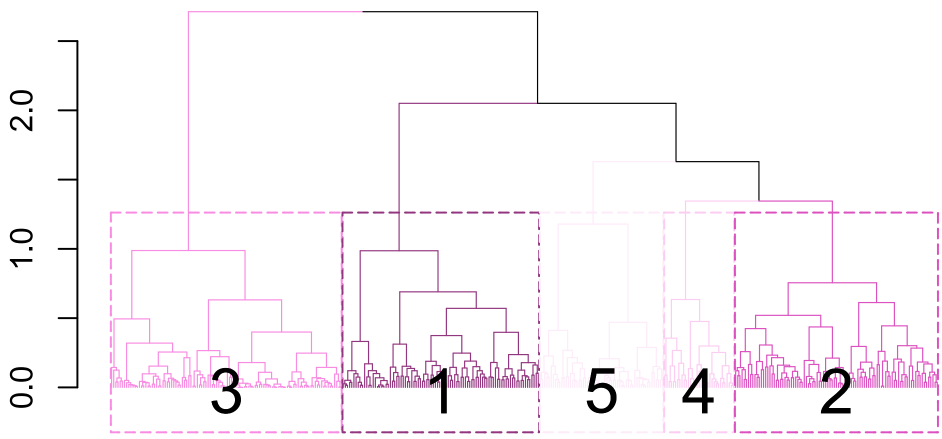


Figure 5.9: Dendrogram of Agglomerative Clustering Analysis with resulting five clusters identified.

Alongside these distinct groups of households identified through the clustering, the analysis of the 23 in-depth interviews revealed several key issues common to subsets of households. Survey clusters 3 and 5 were more heavily represented in the interview samples. Seven common issues emerged from the interviews; some of these were corroborated by findings from the survey analysis, while others were not directly identifiable from the survey data. These seven issues are discussed under the respective headings below. All names of interviewees mentioned here have been changed to protect the privacy of respondents.

LPG access

More than half the interviewed households were using LPG as their primary cooking fuel, although only two of them benefited from an upfront cost subsidy. Across all clusters the majority of households were not aware of government support and subsidies available to them. Only the households in clusters 3 and 5 had any significant awareness of such support (see Figure 5.10b), with over 80% aware of government subsidies in these clusters. Clusters 3 and 5 also have a greater proportion

Table 5.4: Descriptive statistics of the survey clusters

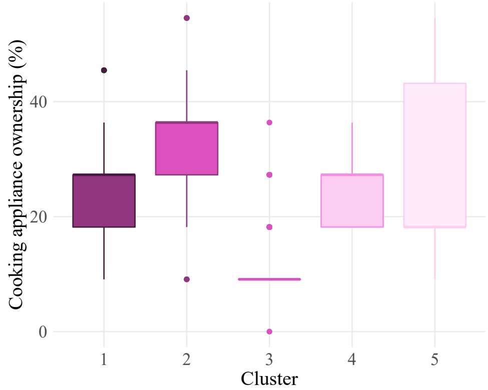
Independent variable	**1**	**2**	**3**	**4**	**5**
No. of Households	92	95	108	33	58
Persons per Household	3.6	3.7	3.6	4.1	3.9
Exp. per capita (INR/month)	4413	4216	4458	5188	4509
Time in Place (years)	34.0	35.4	11.4	88.5	51.4
Daily Wage Earner (%)	42.3	37.9	94.4	63.6	62.1
Electricity Availability (hours)	12.4	18.8	8.9	17.8	23.2
Electricity Consumption (kWh)	60.6	66.8	16.7	46.5	82.5
Biomass Consumption (kWh)	431.7	275.5	257.6	105.3	86.0
LPG Penetration (%)	1.1	16.9	2.8	12.1	32.8
Kerosene Penetration (%)	30.4	70.5	4.6	57.6	51.7
Appliance ownership (%)	24.1	34.4	13.9	27.2	38.2
Hours of Cooking	3.9	3.7	2.6	3.9	3.3
Hours of Lighting	6.3	5.8	3.4	7.6	4.7
Majority Religion (%)	82.6	92.6	99.1	45.5	37.9
SC/ST[a] (%)	77.2	86.3	33.3	42.4	31.0
OBC[b] (%)	22.8	10.5	66.7	57.6	69.0
Notified Slum (%)	9.8	7.4	97.2	0.0	60.3
HoH[c] Literacy (%)	41.3	49.5	38.3	18.1	50
Awareness of Gov. Programmes (%)	0.0	1.1	82.4	0.0	89.7
Beneficiary of Financial Support (%)	7.6	17.9	0.0	3.0	22.4

[a] Scheduled Caste/Scheduled Tribe
[b] Other Backward Castes
[c] Head of Households

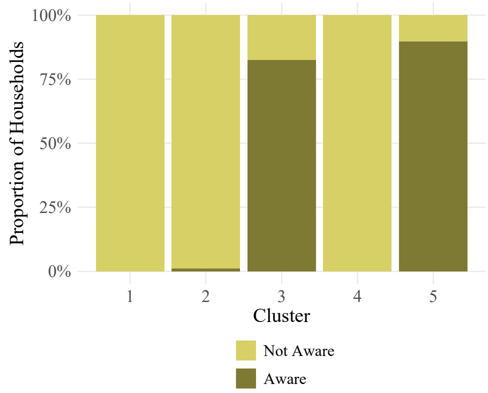
of households paid daily or weekly, living in non-notified slums [1], and in the case of cluster 3 they are also recent migrant arrivals to the city. Cluster 5 constitutes households that have received the highest level of financial support (22.4%) and LPG penetration of 32.8%.

Frequency of payment is an important factor in household energy decision making. Many low-income households are paid on a day-to-day basis which restricts cash flow, limiting purchasing power, as well as resulting in lower income security. Households in cluster 3 who are largely migrant labourers on daily wages have the lowest LPG penetration rates, 2.8% as shown in Table 5.4 and Figure 5.10d, de-

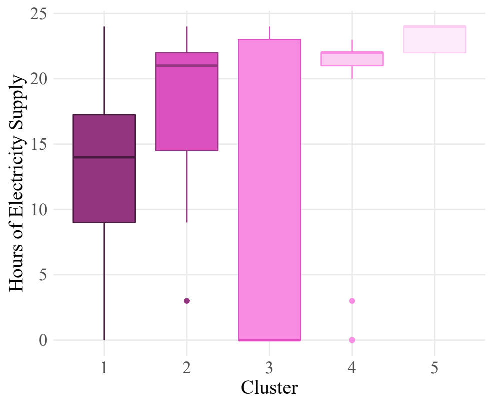
[1] Technically some of those in Cluster 5 were officially listed as non-notified as they were outside the official boundary of a notified slum but were immediately adjacent to one.



(a) Cooking appliance ownership

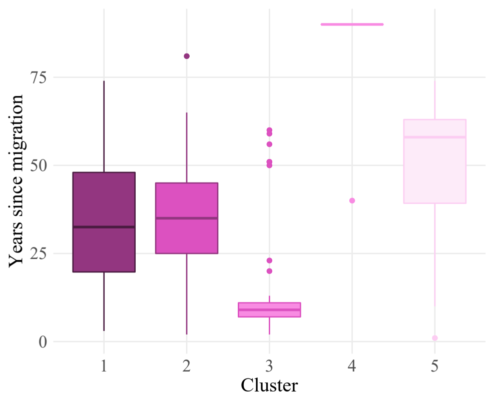


(b) Awareness of government support initiatives

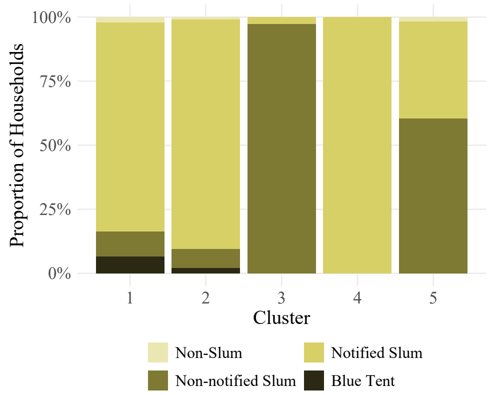


(c) Hours of Daily Electricity Supply

(d) Income frequency



(e) Years since migration



(f) Slum type

Figure 5.10: Plots by cluster of socio-economic characteristics

spite the high level of awareness of government subsidies generally available. The informal nature of their circumstances including lack of tenure and poor access to banking and loans, bars them from accessing the government support as discussed by A. Jain, Agrawal, et al. (2018) and also restricts their access to financing to address their cash flow problems.

Lack of awareness of the available support forces households to turn to other financial arrangements such as savings clubs or informal finance if they still want an LPG connection, such as households in clusters 1 and 2. Two household interviewed took out an informal loan, loans which are often on high interest rates but are commonly used by low-income households (Collins et al., 2009). One household borrowed the money from a family member and one household saved for months to be able to make the payment. Rizwana[2], a native of Bangalore, switched to LPG because of pressure from neighbours who complained of the constant smoke emerging from their house. Her family paid INR 8000 (ca. USD 110) for a double cylinder connection for which they received a loan from a local money lender. Despite getting the connection three months prior to the interview, they were still paying the lender up to INR 1000 per week. They are receiving a sum of INR 100 in subsidies after each refill. Her husband, a daily wage labourer, supports her and their three children.

Table 5.4, shows clusters 1 and 2 had a significant proportion of households like Rizwana's who despite little awareness of government clean cooking initiatives did avail of financial support in the form of subsidies for refills and a free stove. Other households interviewed mentioned that despite promises of regular subsidies, many of the households have not been receiving any. Interestingly, these households expressed no inclination or interest to receive the subsidy either, as they felt the INR 100-300 they may receive would do little to help with the costs.

[2]Names of respondents have been changed to protect privacy.

Electricity access

According to the World Bank (2018), India has an urban electrification rate of over 99% and indeed all interviewed households had regular access to electricity of some sort. However, this statistic often ignores the difficulty in getting metered access to electricity, and one third of our interviewed households relied on illegal connections. Despite the difficulty expressed by some households in accessing a legal metered connection, lack of access to a regular source of electricity was not an issue. Indeed the only group of households in the survey who reported limited access to electricity were those in cluster 3 with a high proportion of recent migrants living in non-notified slums, as seen in Table 5.4. Only 41% of households in this group had regular electricity access.

During an interview with Meena at Nayandahalli, a ward with a high migrant population due to its proximity to industrial areas, she stated that the authorities have been coercing the community to pay for metered connections. However, they are resisting until they are granted land rights. Similar stories of respondents being asked to pay up to INR 16,000 (ca. USD 225) for a metered connection came up during several of the interviews. Tangled wires precariously hanging from a street light providing electricity to the surrounding houses are a common sight.

Table 5.4 shows that clusters 1,2 & 4 whose households migrated more than a generation ago use more electricity than recent migrants in cluster 3. The more established a household is the more time they will have had to acquire electrical appliances, as well as gain legal tenancy or move to a property with recognised tenancy, facilitating access to a metered connection. Ahmad and Puppim de Oliveira (2015) similarly found that access to utilities such as electricity and water were associated with the uptake of LPG among slum households.

The cost of electricity was not seen as a major barrier for most households, with metered households paying monthly bills between INR 100-500. Another interviewee expressed that despite difficulty in making the payment every month, with two school age children at home, they don't have any choice. She is a tea seller, and her husband is a rickshaw driver and together they earn between INR 200-400 daily, which they use to support their family of six.

Involuntary transitions

Lack of affordable alternative fuels was a key narrative for uptake of LPG in interviews, but this was not apparent from the survey data. Historically, kerosene was subsidised by the state and distributed through the Public Distribution System (PDS) to enable access to low cost cooking fuels. However, subsidies were rolled back in an attempt to encourage the uptake of cleaner cooking fuels (Shine Jaocb, 2017), and there have also recently been restrictions on the sale of kerosene in PDS shops. Additionally, the availability of firewood is also dwindling owing to rapid urbanisation. With the two most accessible and affordable sources of traditional cooking fuels restricted, households are forced to switch to LPG. While studies have addressed the impact of kerosene subsidy withdrawal elsewhere in South East Asia (Lam et al., 2016), the impact of such measures on cooking fuel choice in India have not been studied in detail.

Nearly two thirds of interviewees who used LPG, cited a lack of affordable kerosene as a driver of their transition. The emerging black market prices for kerosene of INR 70-100 (ca. USD 1-1.50) per litre, were considerably higher than the INR 15 per litre at the PDS shops. While this may appear to have had the desirable effect of nudging households towards cleaner cooking fuels, it does so at a cost. As highlighted above, households often lack access to the financial tools that can make this transition easier, and none of these households made use of available, albeit insufficient, financial support. In fact, a novel finding in this analysis was that such a measure can even have the unintended consequence of forcing some households to switch back to using firewood. For example, Parvathy who lives with her husband and four children in a self-constructed temporary dwelling, has been using firewood for the last four years since her local PDS shop stopped providing kerosene. She also stated that her children were falling sick frequently due to the smoke.

Frugal energy practices

Households are required to constantly compromise on their energy consumption. Past studies have shown that even if households 'switch' to cleaner fuels, they continue to use the existing fuel as well as the new fuel to meet their energy needs (O. R. Masera et al., 2000); this practice of fuel stacking was prevalent in the interviews and surveyed households. For example, despite having both electricity access as well as an LPG connection, some households opted to use firewood for water heating which is both cheaper and quicker. Janaki, who lives with her ailing husband and works as a daily wage labourer, uses a portable kerosene stove which requires a refill every five days, which costs her INR 100 per refill. To try and reduce this recurring expense, they cook rice over rotis, which cooks faster using less fuel. Sometimes, they go up to two days relying on outside food such as plain buns, when kerosene is either not available or they are unable to meet the cost.

The hours spent cooking or using light in the household provide a proxy for time spent on energy related practices in the household. Table 5.4 shows that across all clusters except for cluster 3, households spend between 3 to 4 hours cooking each day, however there is greater variation in hours of lighting. While cooking satisfies a basic need to eat, lighting may be a proxy for other activities in the household. Whether children doing homework, watching television, or doing small jobs, households can adjust their practices and routines to the resources available. To meet her and her husband's daily expenses, Mariamma, a resident of Rajagopal Nagar, relies entirely on their son's income. She noted that "I don't use the fan nor light during the day and spend my time sitting outside the house to reduce my electricity bill". She uses firewood for water heating despite having an LPG connection and does not own a refrigerator because she says it would increase the bill further.

These patterns show low-income households adapting their energy practices in aid of managing their finances. This is interesting in light of the work of Collins et al. (2009), who found low-income households often made use of a range of financial tools, both formal and informal, to manage their 'financial portfolio' in light of their circumstances. Our findings suggest that households may also have a range of different energy related practices which they use to frugally manage their 'energy consumption portfolio' in response to their circumstances.

LPG Safety and Health Concerns

The harmful health impact of firewood and kerosene was not a cause for concern amongst most interviewees. When asked whether they were worried about their own health or the health of their children, interviewees often showed little concern. Pachiamma, who has been relying on kerosene for over 40 years, spoke of her fear of using LPG since it might burst or cause a fire in the house; a finding seen in other recent studies by Osano et al., 2020 and D. Sharma et al., 2020. She affirmed that she is not worried about the smoke caused by kerosene. She shares a small pucca house (a pucca house is a permanent house usually built of concrete, stone, metal or clay brick) with two others and the cooking is done indoors. Another interviewee, Urmila, who has been using LPG for ten years, said that she in fact preferred using kerosene since the food tasted better and she did not see indoor smoke as an issue either.

Interestingly findings in a recent study by Mani et al., 2020 suggested that prior awareness of health benefits of LPG over biomass were not a driver of uptake, and more likely households are made aware of health benefits upon adoption. Only one household interviewed showed active knowledge of the health benefits LPG can have in a household. She spoke of how 7 years earlier, the local community leaders in collaboration with a private gas company, organised a meeting in their community to discuss the health benefits of switching to LPG. However, she did point out that it was the INR 1000 (ca.USD 14) subsidy along with the free stove, that ultimately drove households to switch to an LPG stove. Nonetheless these findings do align with those of Gould, Urpelainen, and Hopkins SAIS (2020) that messaging, perceptions, and education would appear to have a role to play in clean energy transitions.

Difficulty in prioritising clean energy

Households are often unable to prioritise clean energy in their financial decisions. One interviewee named Rama told of how "We don't even have toilets, where will we get the money to buy LPG", this was met by nods of agreement from neighbours sitting nearby. Discussions were often centred around stories of mounting health care bills, school related fees and loans that are being cleared as the 'more important' household expenses. LPG was not a key priority for the households interviewed.

Although a few households spoke of how they had discussed the possibility of purchasing LPG, such ambitions were quickly dismissed due to the lack of money and other spending priorities such as household improvements or acquiring a refrigerator. The 'lumpy' payments for LPG were seen as a major barrier. Naseema who relies on firewood and kerosene said, "We just took a loan of INR 25,000 for festival and other household expenses, we cannot afford to take more loans". For Ramadan, an Islamic festival celebrated in the month of May, she had to purchase new clothes and gifts for her family of six, leaving very little behind for other expenses.

Lack of strong community and political networks

Finally, a recurring theme that emerged almost across all households was the lack of any political or community network. "Nobody visits us. During election time people will come for votes, but that is it.", said one widowed housewife living with her late husband's family. Socio-cultural groupings along lines of religion and caste can often play a role in functioning and organisation of a community. As shown in Table 5.4, clusters 4 and 5 have over 50% of households belong to religious minorities, and 57.6% and 69.0% of households are classed as OBC (Other Backward Caste) respectively. They also display relatively low biomass consumption, and the majority of households use Kerosene or LPG. This may point to differing social norms and community perceptions about the use of these fuels for cooking. It could also reflect access to resources related to their standing in society, or community network. This supports the findings of Saxena and P. C. Bhattacharya, 2018, who found that structural inequalities experienced by certain caste and religious communities impacted access to energy goods.

When asked about the role of sangams, a local term for informal community-based Micro-Finance Institutions (MFIs) and Savings Clubs, households often lacked interest in participating in these networks. Note that these sangams are different from local informal money lenders that many of the households had borrowed from. The reasons for the lack of interest in sangams ranged from lack of time due to work commitments to lack of trust in these groups. Latha, who lives with her parents and her daughter, told of how despite being aware of the local MFI, she does not trust them. "In this community, all the 'big' people are selfish and enjoy the benefits themselves", she said. This sentiment came up in four other interviews, with two women having lost their money to people who claimed to represent a sangam. A few women also spoke about the fear of taking loans from sangams due to the fear of inability to pay the money back. Only five households spoke of helpful neighbours who are willing to share information and mobilise financial support in times of need.

5.4 Transition Pathway Characterisation

5.4.1 Qualitative Clustering

The previous section presented the results of the clustering of the quantitative survey data, identifying 5 clusters, and discussed these in the context of key issues identified in the qualitative interviews. The mixed method pathway clustering requires both the quantitative and qualitative data to be clustered, before matching and second stage clustering can identify different pathways. In contrast to the five survey clusters, clustering of the correlation network of the coded interviews indicates six different groups of interviewees with common narratives of energy use identified as groups A to F shown graphically in Figure 5.11. Table 5.5 has a detailed breakdown of key discussion themes and issues for each of six clusters. The households in these six different groups discussed combinations of the issues discussed in the previous section. For example interview group A are affected by difficulty prioritising clean energy but are also affected by issues surrounding electricity connection, and lack of a strong community or political network. Conversely interview group B have reliable metered access to electricity, and have good community and political networks, managing their finances to pay for cleaner fuels, while also being aware of health issues related to fuel use.

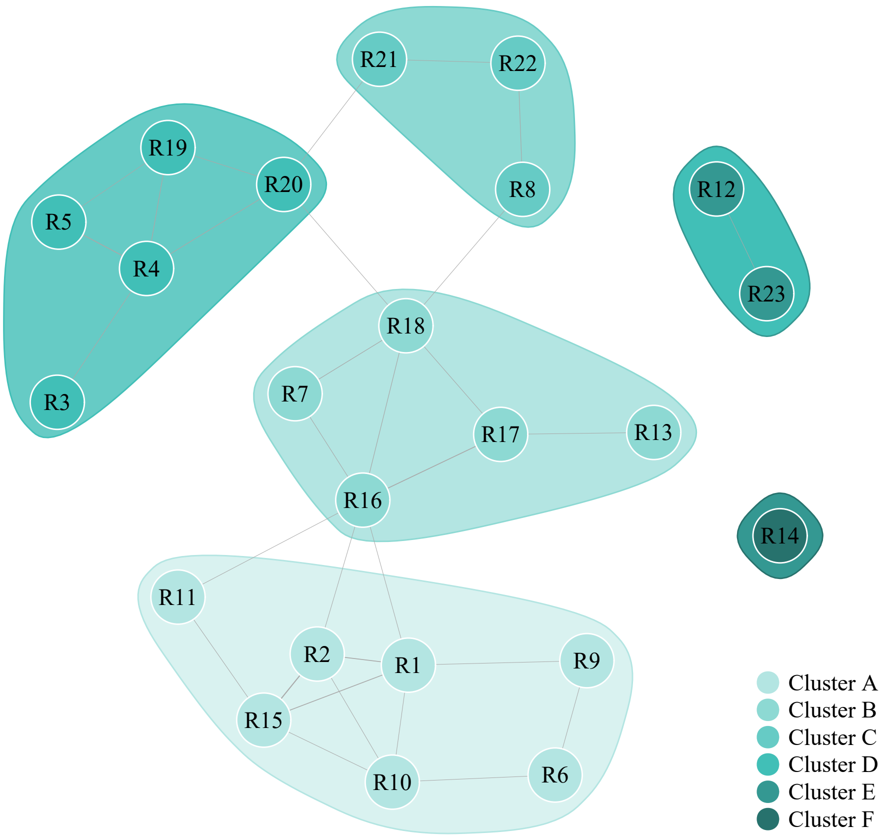


Figure 5.11: RQDA Network Correlation Graph - Links indicate high level of similarity between respondents

Table 5.5: Key discussion themes and issues broken down by interview cluster

	Cluster					
	A	B	C	D	E	F
Energy Access						
Regular access to electricity	X	X	X	X	X	
No electricity bill	X		X			X
Use portable (small) cylinder				X		X
Firewood used for water heating	X				X	X
Attitudes & Practices						
Prefer Firewood			X			
Prefer Kerosene			X			
Prefer LPG					X	
Fear of LPG			X			
Health concerns		X				X
Frugal practices					X	X
Finances, Markets & Subsidies						
Paid for own LPG connection		X			X	
Pressure to pay for electricity meter	X	X				
LPG not a financial priority	X					X
Lack/Cost of kerosene		X		X		X
Not aware of financial incentives	X				X	X
Receiving subsidy		X		X		
Informal market for fuel		X				X
Difficulty managing finances	X		X		X	
House in poor condition/saving for repair	X		X			X
House building aspirations		X				
Community & Politics						
Fridge aspirations				X	X	
Active/supportive community		X			X	
Community pressure				X		
Limited community	X			X		X
Mistrust of Sangam/Savings Clubs		X	X	X	X	X
No government/politician interaction	X		X	X	X	
Good government/politician interaction		X				

5.4.2 Transition Pathways

Recall from the method that each interviewee is also associated with the survey cluster they are found to be closest to. As the next step, each interviewee is further clustered on these two dimensions - their interview grouping and their survey cluster. Doing so, infers plausible and distinct pathways followed by households with similar characteristics and narratives for their transition to clean cooking fuel. Through this, four distinct pathways are identified across interviewed households. The four distinct transition pathways each have a unique data driven definition of the challenges and barriers to clean cooking transition that they present, which are described in Table 5.6, and the proportion of each constituent cluster in each pathway is shown in Figure 5.12. These pathways descriptions offer policy makers and practitioners with rich descriptions and criteria to devise tailored and targetted interventions and solutions for the challenges of households on each of these pathways.

Transition pathway P1 detailed in Table 5.6 concerns households in mostly notified slums, who are at least second generation migrant families, with an even split between daily and non-daily wage earners. This pathway consists of households in survey clusters 1 and 3, and interview groups D, E, and F. These households are most unlikely to use LPG and have moderate prevalence of kerosene use. There is a low proportion of religious minorities in this group, and these households have a lack of strong community networks, with many being unaware of available support for clean cooking fuel. They often have reliable electricity with a metered connection, but have frugal fuel stacking practices for example using collected firewood to boil water or cook rice to save money. The interviewed households in this group discussed how cost and availability restrictions on kerosene had driven them to informal markets, and that they had financial aspirations other than using LPG, such as purchasing a refrigerator.

Transition pathway P2 consists of mostly second or third generation households in notified slums, where a majority of households are paid weekly or monthly. These households are part of survey clusters 1 and 2 and interview groups A, B, and C. They are the most prolific users of biomass fuels and few such household use LPG. Safety concerns regarding LPG and a preference for kerosene may help explain this. These households have been forced to use LPG or revert to firewood as a result of kerosene shortage and cost increase. These households have good community networks but are sceptical of semi-formal MFIs and savings clubs, and switching to LPG is not usually an aspiration or financial priority. Rather these households would prioritise acquiring a fridge or improvements to their home.

LPG penetration was highest amongst households on transition pathway P3. These are households in survey cluster 5 and interview groups B and D. These households are surprisingly mostly in non-notified slums (or undefined areas adjacent to notified slums) and daily wage earners. However they have the lowest biomass use and highest use of petroleum fuels for cooking as well as metered electricity access. These households often belong to minority religions and OBCs and have good community and political networks. Head of household literacy rate is higher than amongst the mean household. They have felt pressure to switch cooking fuel due to kerosene shortage and cost, and some have been able to access some subsidies and benefits to switch to LPG.

The final transition pathway P4 is analogous to the situation of daily wage labourers, many of whom have migrated to the city for work. They are mostly living in non-notified slums or sometimes 'blue tent' informal settlements, and have limited awareness or difficulty accessing benefits such as subsidies or financial rebates for LPG use. LPG is often not a financial priority amongst this group, and their irregular electricity access is of more concern with many being pressured to pay for a meter. These households often lack community or political networks and have limited access to even semi-formal finance. The precarious nature of their daily wage employment can make managing finances difficult, and their lack of tenure makes access to support a problem.

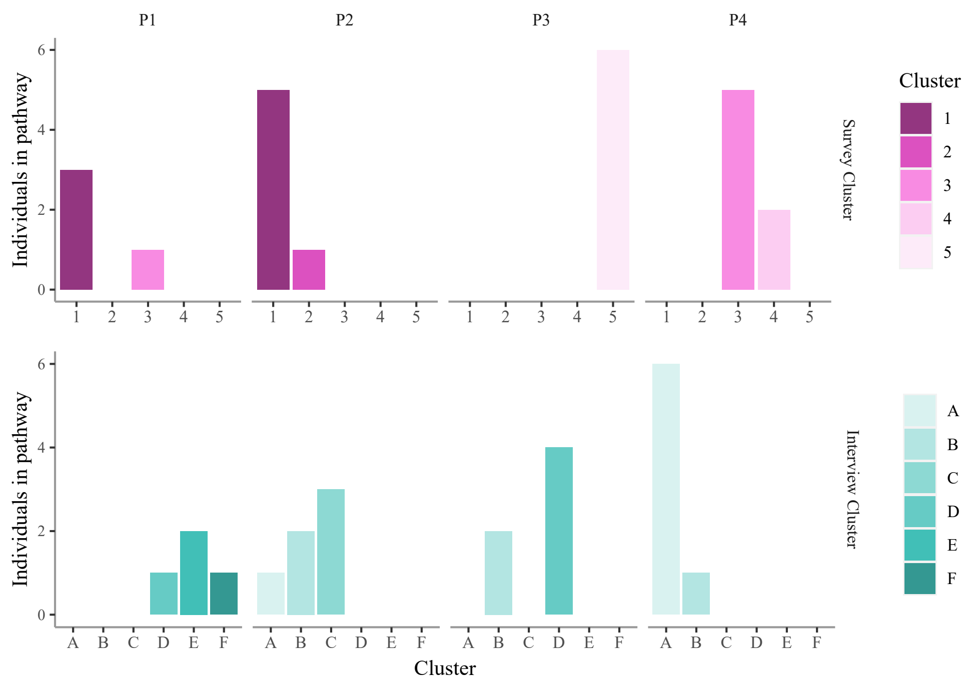


Figure 5.12: Proportion of interview and survey clusters represented in the four second-stage clusters.

While this method is able to identify key determinants of clean cooking adoption which align with those found in other studies, the transition pathways identified by this second stage clustering provide important additional information about the specific combinations of barriers to clean fuel, and household characteristics, that were not apparent from the interview or survey analysis alone. Unlike rural areas, amongst low-income households in urban areas there is far greater heterogeneity and this mixed method pathway clustering is able to distinguish these different typologies of low income household, and identify their needs and context to inform policy design. This is a salient novelty with respect to more straightfor-

Table 5.6: Key attributes and narratives for transition pathway clusters

Clust.	Mean Attributes	Key Narrative
P1	Notified slum: 61.8% Daily wage earners 55.3% Biomass use: 388 kWh LPG prevalence 1.5% Kerosene prev. 24.0% Minority religion 13.3% Migrated 28 years ago	• Reliable electricity access with metering; • Frugal energy practices with multiple fuels; • Unaware of available support for clean fuel; • Fridge ownership aspirations; • Kerosene restriction leads to involvement in informal markets; • Lack of strong community networks.
P2	Notified Slum 82.8% Daily wage earners 41.6% Biomass Use: 406 kWh LPG prevalence: 3.8% Kerosene prev.: 37.2% Minority religion 15.8% Migrated 34 years ago	• Fear of LPG safety/preference for kerosene; • Cost of kerosene has driven uptake of LPG and firewood; • Have other financial priorities such as home improvement of purchasing a fridge; • May have good community network but skeptical of informal finance initiatives.
P3	Notified Slum 37.9% Daily wage earners 63.6% Biomass Use: 86 kWh LPG prevalence: 32.8% Kerosene prev. 51.7% Minority religions 37.9% Migrated 51 years ago	• Electricity access and metered connections; • May have paid for own connections but do receive some subsidies for fuel; • Have community or political networks; • Have felt pressure to use LPG from lack of availability/cost of kerosene.
P4	Notified Slum 30.6% Daily wage earners 85.6% Biomass Use: 214 kWh LPG prevalance 5.4% Kerosene Use 20.2% Minority religions 16.3% Migrated 20 years ago	• Electricity access concerns, pressure to pay for meter; • LPG not a financial priority; • Lack of awareness or access to support; • Lack of community and political network; • Recent migrant wage labourers; • Informal community/no tenure rights.

ward quantitative clustering or predictive analytical methods. This allows tracking of how different factors work together to define different pathways to clean fuel and highlight the barrier or challenge present with each of these.

These transition pathways offer insight into how a household will respond to changes in energy provision and policy in different ways. An example of this concerns pathway P1 and P2's response to kerosene restrictions. Pathway P1 have poor community networks, and little knowledge of existing government support for clean fuels. Pathway P2 do not financially prioritise LPG and may have safety concerns about LPG, as well as distrusting informal finance initiatives. With the restrictions on kerosene at ration shops, households on pathway P1 were more likely to respond by turning to informal markets for kerosene, while households on pathway P2 either adopted LPG or if they could not access financing reverted to using biomass.

The integrated qualitative and quantitative dataset has enabled inference of these pathways and identification of households with different socio-economic circumstances that may share a pathway in transitioning to clean cooking. For example pathway P4 is characterised by lack of awareness and access to government support for clean cooking, lack of community and political networks, and poor access to other utilities. However two very different groups of households, from different survey clusters, were associated with this pathway. Both recent migrant wage labourers and longer term retired or widowed residents in notified slums appear to face some similar barriers to accessing clean cooking.

Interestingly there are some parallels between the pathway analysis in this chapter and market segmentation analysis in the business world. In this comparison clean cooking might be seen as the product, and the policy or intervention would represent the rest of the 'marketing mix'. This comparison is not necessarily a trivial one - it suggests that there might be a role for entrepreneurship and innovation in addressing the challenges of particular pathways. This may also say something about the how stakeholders wanting to support low-income households in clean cooking transition need to operate. Entrepreneurs pursuing a new market segment have to be responsive and willing to experiment, to optimise their offering for the target market. Reaching low-income households and supporting their transition to clean cooking might require government and stakeholders to operate on some of the same principles. This shall be discussed further in Chapters 6 and 7

5.5 Limitations and Caveats

The descriptive quality and high resolution of the survey data provide a greater level of detail on energy use habits among low-income urban households in India than existing national datasets, and make an important contribution to the limited body of knowledge on low-income urban households. However, this data is limited in applicability by pertaining to a specific agro-climatic, social, and political geography: namely Bangalore. Further studies across different states would be required to test the extent to which one could borrow strength across households located in different physical and socio-political conditions. While some of the trends and patterns could be identified in other cities using survey data, detailing the narratives, practices, and motivations would require further interviews.

In terms of practicality, this approach while providing rich detailed descriptions of transition pathways and associated issues is also labour, data, and resource intensive requiring collection of a not insignificant amount of data, and a laborious analysis process. This is appropriate for case studies but may not be suited to widespread use. Furthermore, this work has focused on descriptive analysis of the data to gain information on the variations of barriers faced by households on their path towards clean cooking transitions. Whilst these outcomes can be used to tailor and fine-tune energy policies to maximize their reach across households, they cannot be used directly to predict their impacts.

5.6 Conclusions

This Chapter explored and characterised the nature of energy transition pathways among low-income households using mixed data and methods, addressing the second research objective of this thesis. Using data from low-income households in Bangalore exposed behaviours and decision making surrounding clean cooking transitions in poor urban households. Energy use and fuel choices vary considerably due to the interaction of a wide range of socio-economic and behavioural determinants. Firewood was used by all households surveyed to some extent and all households used more than one fuel in the household. This residual use of biomass regardless of access to cleaner cooking was consistent with the findings in Chapter 4. However, in the survey sample from Bangalore just over 10% of households primarily made use of LPG for cooking.

The four transition pathways identified in this case study demonstrate that some households are at greater risk of being trapped using biomass fuels, particularly those in non-notified slums and without strong community networks. Each of the four pathways highlight different barriers to transition and an integrated strategy of interventions is required to address these. For some households better access to more reliable financing is needed to help deal with the issue of cash flows resulting from being in daily or weekly wage employment. Interventions could implement alternative financing arrangements that allow households to pay for their cylinder in daily or weekly instalments, such as methods employed with mobile enabled pay-as-you-go solar home systems in Kenya (Rolffs et al., 2015).

There is also a need for behaviour changing and awareness raising initiatives for more established households and communities. These communities can have less awareness of available support and inaccurate perceptions of risks and safety of biomass and LPG use. These could be targeted to relevant wards and combined with promotional efforts for the LPG uptake scheme. Households in non-notified slums can struggle to access much needed support for cleaner fuels. Their precarious living circumstances and limited access to infrastructure need to be addressed to facilitate their adoption of cleaner fuels. Lack of tenure rights and formal banking prevent these households from accessing current initiatives. Targeted interventions could either use specific criteria for access to subsidy schemes for those in such circumstances or indeed this could be addressed as part of a wider initiative to improve living standards and legal status of those in non-notified slums. This might be thought of as improving the local socio-economic context the household interacts with, and would support the argument of Sankhyayan and Dasgupta (2019) that access to energy cannot be viewed in isolation of other socio-economic development indicators.

What is clear from this characterisation of low-income clean cooking transition pathways in Bangalore is that a household's clean cooking transition is influenced by three key components: The interaction with the local socio-economic and cultural context a household is in, such as the case of migrant labourer households in non-notified slums; the wider policy and economic context of the city, such as the restrictions on kerosene sale in the PDS shops; and energy related practices and habits of the household, as evidenced by the case of households cooking rice with biomass fuels when they need to save money. This builds upon the analysis in Chapter 4 that showed that households could follow different pathways to transition, and shows how the interaction of the household with its local circumstances condition its transition pathway. This insight will form the basis of the microsimulation model presented in the following chapter.

6 People and Models

The contents of this chapter have been submitted to Environment and Planning B: Urban Analytics and City Science for review; A P Neto-Bradley worked on conceptualisation, methodology, data curation, investigation, and visualisation; R Choudhary and P Challenor provided supervision and with inputs on methodology.

A. P. Neto-Bradley et al. **(2021)**. "A microsimulation of spatial inequality in energy access: A Bayesian multi-level modelling approach for urban India". *arXiv:2109.08577 [physics]*. arXiv: 2109.08577

6.1 Introduction

Access to clean cooking at an urban scale in India is subject to considerable variation, with households following different pathways in transition to clean cooking, each characterised by specific barriers and challenges. Addressing these requires tailored and targetted interventions and solutions. Targeting efforts to address different sets of barriers at an urban scale is limited by the lack of available data and modelling methods suited to a developing country context (Bai et al., 2018). Previous quantitative modelling approaches assume all households will respond in the same way to economic stimuli, but this is an oversimplification. Furthermore interpretation of such model outputs as they relate to local context may not be intuitive or feasible and can require additional learning, limiting the real world usefulness of such models (Nochta et al., 2020). There is a need for models of residential energy use which better reflect, represent, and locate the reality of energy use by people in cities, while making efficient use of current data.

In the preceding Chapter 5, a mixed method approach was used to characterise different clean cooking transition pathways amongst low-income households in Bangalore. The in-depth mixed method clustering approach is well suited to characterising heterogeneity in cases where there is a representative and high resolution data set available. However in most instances in India (and indeed across the Global South more widely) such data does not exist, and so for stakeholders interested in the city-scale distribution of clean cooking access and fuel stacking there will be little to no information available to support decision making and targetting of solutions to the challenges faced by households in adopting sustained clean cooking fuel.

The work in Chapter 4 drew attention to the importance of considering each household and its unique combination of characteristics and surrounding context individually to better understand patterns of cooking fuel use and stacking. This implies that to effectively model cooking fuel use at an urban scale requires a representative sample of individuals for that area, for which some socio-economic characteristics are known. The creation of synthetic or simulated populations, also referred to as a microsimulation approach to modelling, offers a solution to this problem. These methods have been applied in fields such as transport research and epidemiology for several decades and can synthesise a representative population of individuals from census data.

Each synthetic household has a set of socio-economic characteristics, that can be used by a model to estimate likely fuel use and fuel stacking. In the previous chapter three key influences on a household's energy transition pathway were identified: The interaction with the local socio-economic and cultural context a household is in; the wider policy and economic context of the city; and energy related practices and habits of the household.This Chapter develops this insight into a modelling approach that integrates people and the nature of their energy use to model clean cooking access at a city scale. This addresses the third research objective of this thesis detailing a model integrating the influence of, and uncertainty arising from local socio-cultural factors as well as household habits on clean cooking transition.

A microsimulation approach is proposed which uses a Bayesian multilevel model to estimate cooking fuel consumption and fuel stacking across city wards, and thus provides a disaggregated urban scale view of spatial inequality in clean cooking access, with quantification of uncertainties. This approach is based on the hypothesis of this thesis that household level energy use is conditioned by household habits and practices, local socio-economic and cultural context and spatial effects. This is the first attempt to use such methods to characterise spatial distribution and inequalities in energy access in Indian cities. This has the potential to be an invaluable tool for policy makers capable of identifying city wards at greatest risk of fuel stacking and with low levels of sustained LPG use. This can provide a cost and resource efficient way of tailoring and locally targetting interventions and solutions. Knowledge of where worst affected wards are can also encourage focused engagement with these households and stakeholder to help design well informed and adapted policies.

The approach combines publicly available data from the census and nationally representative surveys, to generate a synthetic population of individual households. Markov Chain Monte Carlo (MCMC) sampling is used to estimate parameters for a multilevel model, which predicts fuel use and fuel stacking prevalence (use of multiple fuels i.e. both biomass and LPG) at a household scale while accounting for group effects of primary cooking fuel choice, and spatial and non-spatial heterogeneity. Primary data collected from selected wards in four modelled cities is compared to model outcomes for a consistency check. The model outputs enable the identification of those wards within a city that are worst affected by solid biomass fuel, and offer insight into how wards compare to each other as well as to the average urban household in the state. With a new Indian census being conducted this year this approach will provide a means to update and track the state of clean cooking transition in urban India in the coming decade.

This chapter will briefly introduce microsimulations and the study area, before detailing the methodology and model structure. Survey data will be used to assess model performance and a discussion of model outputs will consider the insights these can offer, before concluding with a consideration of the features, utility, and policy relevance of this approach.

6.1.1 Modelled Cities

Four cities will be modelled to evaluate and demonstrate the modelling approach. These are located across the southern Indian states of Kerala, and Tamil Nadu. The city locations and population are detailed in Figure 6.1. Notice that the population (not counting wider metropolitan area) of these so-called Tier 2 cities sits a little under one million for all except Coimbatore which is slightly more populous than the rest. All cities have larger populations than the average urban area in their states, and have seen growth in recent decades. While there are differences between these states in terms of policy and even socio-economic landscape, they all fall within a narrow geographic region with similar climate and geography which ensures a relatively homogeneous physical landscape with minimal differences in effect of climate on energy use, or availability and type of biomass.

Figure 6.1: Map showing location and and size of four cities modelled.

6.2 Methodology

The methodology combines multiple data sources at differing geographic scales to model the spatial inequality in access to clean cooking across a city. The use of multiple data sources is driven by the scale of analysis and availability and detail of public data at these different scales; the aim is to understand differences across city wards by modelling household level energy use but household level data only exists per state and district, while available ward level data includes only ward-level tallies and does not provide details on fuel consumption, with only limited socio-economic variables.

The objective is to estimate cooking fuel use and energy choices at a household level taking into account household practices, locally acting socio-economic and community factors, and spatial effects. This requires generating a representative population of households whose combination of socio-economic characteristics is known. Iterative Proportional Fitting (IPF) is used with census data with national household surveys to first generate a synthetic population of households representative at the ward level. For each of these synthetic households the primary fuel choice group it belongs to is assigned using a categorical logistic regression based on socio-economic predictors. The fuel consumption and fuel stacking prevalence are then estimated for each household using a multilevel model with model coefficients varying on basis of the primary cooking fuel choice of the household and taking into account the ward-level effects. Figure 6.2 provides a schematic overview of the model structure.

6.2.1 Data

Model Inputs

As shown in Figure 6.2 there are several different data sources used as model inputs. The household level data from the Indian Human Development Survey 2011 (IHDS-II) (Desai and Vanneman, 2015) and the 2011 census (Government of India, 2011) ward-level tables of household asset ownership are used to generate a synthetic population. GIS maps of the Census 2011 ward boundaries, and National Sample Survey (NSS) consumer survey data from Round 68 (National Sample Survey Organisation, 2013) are used to estimate coefficients of economic determinants in the multilevel model. While the boundary map must feature the ward divisions of the year of the census, this approach can use the most recently available consumer survey data, which is at a household scale (although only identified by state), to provide up to date estimates between the decennial census. Table 6.1 details the different scales and variables covered by each dataset. Note that none of these individual datasets contain all the information needed.

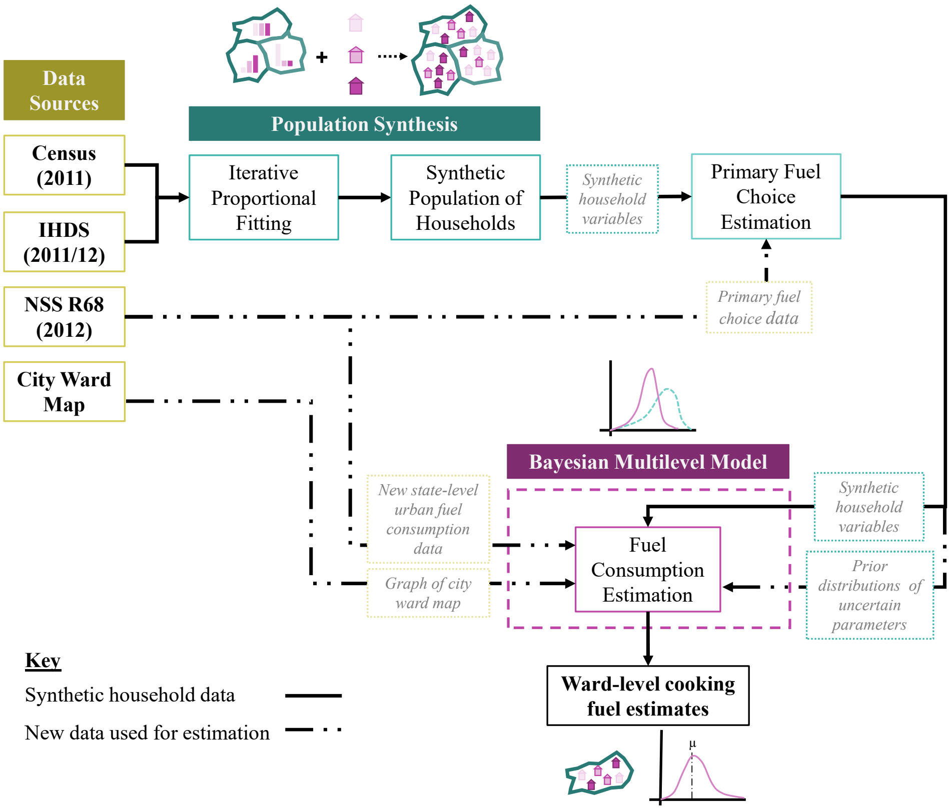


Figure 6.2: Schematic overview of micro simulation workflow detailing how different data sources are combined to generate a synthetic population of households and estimate energy use across the city

Table 6.1: Scale and variable types available in each of the input datasets for microsimulation

Data Source	Scale	Year	Exp.	Soc-econ.	Fuel choice	Fuel use	Geo
Census [1]	W	2011	-	Yes	Yes	-	-
IHDS-II [2]	HH	2011/12	Yes	Yes	-	Yes	-
NSS R68 [3]	HH	2011/12	Yes	Yes	Yes	Yes	-
GIS	W	2011	-	-	-	-	Yes

Ward-level Survey Data

To examine the performance of the model, survey data from 24 low-income wards across the four cities is used. This data comes from the primary data collection described in Chapter 3. Recall that this dataset contains a random sample of 60-70 households within each ward, and was conducted between October 2018 and May 2019, and 2011 census ward boundaries were used.

6.2.2 Household population synthesis

Generating a synthetic population is done using IPF, a method also known as matrix raking first introduced by Deming and Stephan (1940). While this approach does have some drawbacks including that it relies on categorical data, and can only control for household or individual level attributes (Casati et al., 2015), this is in fact well suited to the data publicly available on Indian households. As the unit of interest is the household and not the individuals in it a simple implementation of IPF will meet the needs of this modelling approach. IPF offers the benefit of being computationally efficient (Mosteller, 1968), simple to use (including practical guidance on performing IPF in R (Lovelace and Dumont, 2016)), and converges to a single solution (Fienberg, 1970).

IPF requires two different types of data, the first being a contingency table of categorical variables for each geographic subdivision, in this case a ward, indicating the totals for each categorical feature within the ward. The census ward level data is used for this and the categories of constraints are detailed in Table 6.2. The IHDS is used as the microdata seeing as it contains instances of individuals in the wider area, a sample of representative individual households. IPF allocates these individuals from the micro dataset to the zones or, in this case wards, through a weight matrix that indicates how representative each individual in the microdata is of the specific zone (Lovelace and Dumont, 2016), a process illustrated schematically in Figure 6.3.

Table 6.2: Categories of constraint variables used for IPF contingency table

Variable	Categories
Caste	Scheduled Caste, Scheduled Tribe, Other
Roof Material	Thatch, Tile, Stone, Plastic, Metal, Concrete, Brick, Other
Wall Material	Thatch, Mud, Plastic, Wood, Brick, Metal, Stone, Concrete, Other
Floor Material	Mud, Wood, Brick, Stone, Concrete, Tiles, Other
Rooms	1 to more than 6
Members of household	1 to more than 6
Home ownership	Rent, Own
Primary cooking fuel	Firewood, Crop Residue, Dung, Coal, Kerosene, LPG, None
Cooking location	Outside, Inside, No Kitchen
Asset Ownership	Bank access, TV, Moped

IPF will produce fractional weights for each individual in the microdata, and although these would be suitable for producing small area statistics, a motivation for generating a synthetic population is to have actual synthetic individual households with a particular combination of features. Integerisation is the process of essentially rounding these fractional weights and differing approaches to this exist, although it should be noted that integerisation can result in minor discrepancy between the original and synthetic marginal data (A. Smith et al., 2017). To integerise the fractional weights the approach proposed by Lovelace and Ballas, 2013 was implemented which is based on truncating, replicating, and sampling. Lovelace et al. demonstrated this method outperformed in accuracy other commonly used methods such as simple rounding, inclusion threshold, counter-weight, and proportional probabilities.

It is important to remember that population synthesis assumes that the micro data (in this case the IHDS) is representative of the study area, dependence between variables not constrained for and the constraint variables (the variables present in the census ward level data that we used to determine weightings, shown in figure 6.2) is relatively constant, relationships between constraint variables are

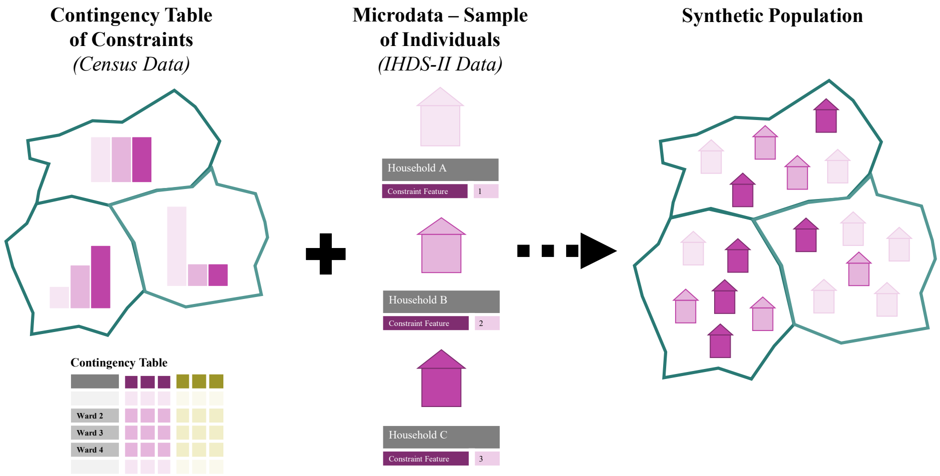


Figure 6.3: Schematic overview of process of synthetic population generation using IPF with a contingency table of constraints and a micro dataset. Adapted from Lovelace and Dumont, 2016

not spatially dependent, and the micro data is sufficiently detailed to represent full diversity of the region of study (Lovelace and Dumont, 2016). These assumptions are clearly oversimplifications, in particular that constraint variables are not spatially dependent - while this does not invalidate the results it is important to understand how this approach simplifies reality.

Validation of the synthetic population is by its very nature tricky as the synthetic population serves to estimate unknown data (D. M. Smith et al., 2009). As a preliminary check, validation was carried out internally against the variables in the census contingency table used to determine weights to ensure the IPF has produced an output faithful to the contingency table. External validation can be carried out by aggregating features of the synthetic population to a geographic level for which there are known values to compare against (Ballas and Clarke, 2001). Synthetic population data was aggregated at the city level to compare constrained and unconstrained variables to those from the Socio Economic and Caste Census of India from 2011. Checking model estimates against the ward level survey data discussed in section 6.3 below also provides a form of external validation on variables which were not used in generating the synthetic population.

Households are then assigned to a primary cooking fuel group representative of current cooking practices. A categorical logistic regression of the form shown in equation 6.1 is used to infer the primary cooking fuel j of a household i based on the socio-economic characteristics of the household. The predictors used are household expenditure, majority religion membership, ration card possession, caste category, and income frequency. The coefficients for this regression are estimated from the NSS data for the respective state. The household is assigned to one of six primary cooking fuel choice groups: No cooking, Biomass, Kerosene, Low LPG (uses less than a cylinder every month), High LPG (uses more than a cylinder every month), or Electricity.

$$j[i] = \beta_0 + \beta_1 exp_i + \beta_2 relig_i + \beta_3 rationcard_i + \beta_4 caste_i + \beta_5 incfreq_i \tag{6.1}$$

Generation and validation of the synthetic population for each of the five cities was done using base packages in R as well as the 'ipfp' package (Blocker, 2016).

6.2.3 Multilevel model specification

A multilevel model is used to estimate the magnitude of biomass and LPG fuel use for each synthetic household on the basis of its socio-economic features and location. The model uses two household level predictors, household expenditure and household size (no. of persons). The coefficients on these predictors vary by cooking fuel group. These represent the intrinsic component of cooking fuel use, dependant on the features of the household.

A novelty of this study is to quantify the influence of local socio-economic and cultural features on fuel use and access. These are extrinsic features which relate to the household's interaction with its wider neighbourhood and community and spatial effects from its location within the city. Quantifying the impact of local features is complicated by the unmeasurable nature of some socio-cultural factors influencing clean cooking access. To address this we adapt the Besag-York-Mollie model (Besag et al., 1991) used in spatial epidemiology studies (DiMaggio, 2015) which is a lognormal Poisson model with an Intrinsic Conditional Auto-Regressive (ICAR) and a random effects coefficient. Similar approaches have been used in district-wise building energy use models (R. Choudhary and Tian, 2014), and recent studies have explored improved approaches for implementation (Morris et al., 2019). In this case the model is a linear normal one with added components for spatial and random effects.

Basic model: No location based effects

The basic model starts from an idealised linear relationship for a household's mean fuel use shown in equation 6.2. Both household income and household size have been found to be determinants of fuel use across a range of studies (Farsi et al., 2007; Ahmad and Puppim de Oliveira, 2015; Cheng and Urpelainen, 2014). For a household i, fuel use μ_{fuel_i} (kWh/month) is a linear combination of household expenditure and household size. Importantly, any given household i belonging to group j may use more than one cooking fuel.

$$\mu_{fuel_i} = a_0 + a_1 exp_i + a_2 size_i \tag{6.2}$$

Coefficients for the household level predictors are allowed to vary by primary cooking fuel group (Gelman and Hill, 2007). This is a non-spatial, and non-random group effect. The statistical model for household fuel use is specified in equation

6.3, where for a household i in cooking fuel group j fuel use is assumed to follow a normal distribution with a mean given by the relationship in equation 6.2, and a precision σ.

$$fuel_{[i]} \sim N(a_{0\,j[i]} + a_{1\,j[i]}exp_{[i]} + a_{2\,j[i]}size_{[i]}, \sigma) \tag{6.3}$$

Whose coefficients are defined by parameters as shown:

$$a_{0[i]} \sim N(\mu_{a0}, \sigma_{a0})$$
$$a_{1[i]} \sim N(\mu_{a1}, \sigma_{a1})$$
$$a_{2[i]} \sim N(\mu_{a2}, \sigma_{a2})$$

Parameters for the distributions of coefficients $a_{0[i]}$, $a_{1[i]}$, and $a_{2[i]}$ are estimated using the NSS consumer survey data. A prior is set on the parameters using the fuel use values embedded in the synthetic population which have been derived from the IHDS micro dataset. This mean fuel use (the synthetic household's historical fuel use), $\mu_{fuel\,h}$ is assumed to have the same linear relationship with household predictors that we use to estimate current use in equation 6.3. A statistical model for $\mu_{fuel\,h}$ is shown in equation 6.4. As the synthetic population does not have corresponding primary fuel choice embedded (this is assigned by the categorical logistic regression) the prior coefficients do not vary by primary cooking fuel group. Note that for this prior data, σ_h is not an estimated parameter as it is a known value and determined from the synthetic population.

$$fuel_{h[i]} \sim N(a_{0h[i]} + a_{1h[i]}exp_{h[i]} + a_{2h[i]}size_{h[i]}, \sigma_h) \tag{6.4}$$

The coefficients in the linear relationship for mean prior fuel use $\mu_{fuel\,h}$, are related to the overall parameters through the following reparameterisation:

$$a_{0h[i]} = \mu_{a0} + \sigma_{a0} z_i$$
$$a_{1h[i]} = \mu_{a1} + \sigma_{a1} z_i$$
$$a_{2h[i]} = \mu_{a2} + \sigma_{a2} z_i$$
$$z \sim N(0, 1)$$

Setting a prior using the data embedded in the synthetic population ensures consistency, and provides an initial 'best guess' for coefficients which are then updated using more recent NSS survey statewide data. Most importantly, this prior has fuel use data for each individual ward, and is thus the only data available for estimation of location dependent effects described in the subsequent models. Note that this basic model will account for less uncertainty than the subsequent models given that it does not consider spatial effects or local heterogeneity. The precision terms for the coefficients σ_{a0}, σ_{a1}, and σ_{a2} represent the uncertainty in the coefficients, while the σ in the base model (equation 6.3) represents chance variability.

Random effects model: Non-spatial heterogeneity

This formulation of the model includes an ordinary random effects coefficient to account for non-spatial heterogeneity. As the data used for estimating linear coefficients of μ_{fuel_i} is representative of all urban households in the state, the non-spatial heterogeneity captured by this coefficient represents how the given city differs from the average urban area in the state. With respect to the objectives of the model formulation this component captures the influence of local socio-economic context.

The synthetic population's historical energy use is aggregated at the ward level which allows for estimation of these ward level coefficients. The formulation of this model in equation 6.5 takes the $\mu_{fuel\,j[i]}$ from the basic model (eq. 6.3) where coefficients vary by cooking fuel group, with the addition of a random effects coefficient $\theta_{w[i]}$ which varies by the ward a household is in.

$$fuel_{[i]} \sim N\left(\mu_{fuel\,j[i]} + \theta_{w[i]}, \sigma_h\right) \tag{6.5}$$

ICAR model: Spatial Auto-correlation

This model variant includes an Intrinsic Conditional Auto-Regressive (ICAR) component which accounts for the tendency for adjacent areas to share similar characteristics (Morris et al., 2019). ICAR models are based on an approach developed by Besag (1974), and in implementing this the advice of Morris et al. (2019) was followed with respect to formulation in Stan. In the ICAR model the spatial component ϕ_w is modelled as normally distributed with a mean equal to the average of its neighbours, where the number of neighbours is denoted by d_w. The conditional specification is shown in equation 6.6, where w is the current ward, $w \sim x$ is the relationship between ward w and its neighbours.

$$p(\phi_w|\phi_{w\sim x}) = N\left(\frac{\Sigma_{w\sim x}\phi_w}{d_w}, \frac{\sigma_w^2}{d_w}\right) \tag{6.6}$$

The ICAR component ϕ is estimated from the historical data of the synthetic population, and uses a fully connected ward graph as determined from the ward map shapefiles, as an input for the ICAR model. A prior is set on the mean ϕ providing a 'soft' sum-to-zero constraint (Morris et al., 2019). This formulation of the fuel use model is shown in equation 6.7.

$$fuel_{[i]} \sim N(\mu_{fuel\,j[i]} + \phi_w[i], \sigma) \tag{6.7}$$

Combined effects model

In the combined effect formulation both an ordinary random effects component and an ICAR spatial smoothing component are included along with the linear combination of household level predictors $\mu_{fuel\,j[i]}$, as shown in equation 6.8. Again these ward coefficients are estimated using the historical data from the synthetic population, and the ward graph for each city. This model accounts for the most uncertainty given it factors in variability in spatial effects, and local heterogeneity for each ward.

$$fuel_{[i]} \sim N(\mu_{fuel\,j[i]} + \phi_w[i] + \theta_{w[i]}, \sigma) \tag{6.8}$$

The four different versions of the model were implemented to compare the performance of the different components of the model. Models for each city were estimated separately, and NSS consumer survey data for the respective state is used. All expenditure and fuel consumption values are normalised and a square root transformation is also applied to fuel consumption values. Estimation of the parameters of the model is done using Stan which performs full Bayesian inference using Hamiltonian Monte Carlo. Stan's No U-Turn Sampling (NUTS) performs better than alternative Gibbs or Metropolis algorithms for models with complex posteriors (Hoffman and Gelman, 2014). The RStan interface for Stan in R enabled input and output data handling in R. It is important to note that as Gelman, 2006 points out, multilevel models are advantageous for making predictions as they can estimate the effect of individual predictors as well as the group-level mean, but these cannot be necessarily interpreted as causal and this should be kept in mind when examining outputs.

6.3 Model Performance

By comparing the model outputs to the survey data available for selected wards, the consistency of the model and its four variants (Basic, Random Effects, Spatial Smoothing, Combined Effects) can be examined. Figure 6.4 compares the model estimates for LPG and firewood use in each of the 24 wards included in the household survey data. The points indicate the mean and lines show the 95% confidence interval of each model and the survey data ward-by-ward across the five cities. The wards are ordered by their mean fuel consumption. Figure 6.4 shows that with a few exceptions the survey means fall within the confidence interval for the combined model estimates, and in a majority of cases those of the random effects model too. In particular, notice that the combined model estimates broadly follow the variation seen in the survey means, with wards with lower survey means likely to have a lower model estimate than those with higher survey means. This is noteworthy as these wards were selected for the survey on the basis of their low-income status and greater likelihood to have lower access to LPG and use greater amounts of firewood. Thus they do not represent the full spectrum of wards but rather the lower end of LPG and upper end of firewood consumption.

Comparing the different model variants, the basic model tends to have the narrowest confidence interval, and the interval increases for the random effects and spatial smoothing models, with the combined model featuring the widest interval. This reflects the added uncertainty the estimates from these models account for. Recall that the basic model infers coefficients for predictors from NSS data on all urban households in the state, including those not in these cities. It represents the estimate of fuel use for that ward's population of households assuming they are in an 'average' city in that state. The random effects model includes a coefficient that represents the impact of the 'city effect' versus the average city in the state. The spatial smoothing model represents how the ward compares to the wards around

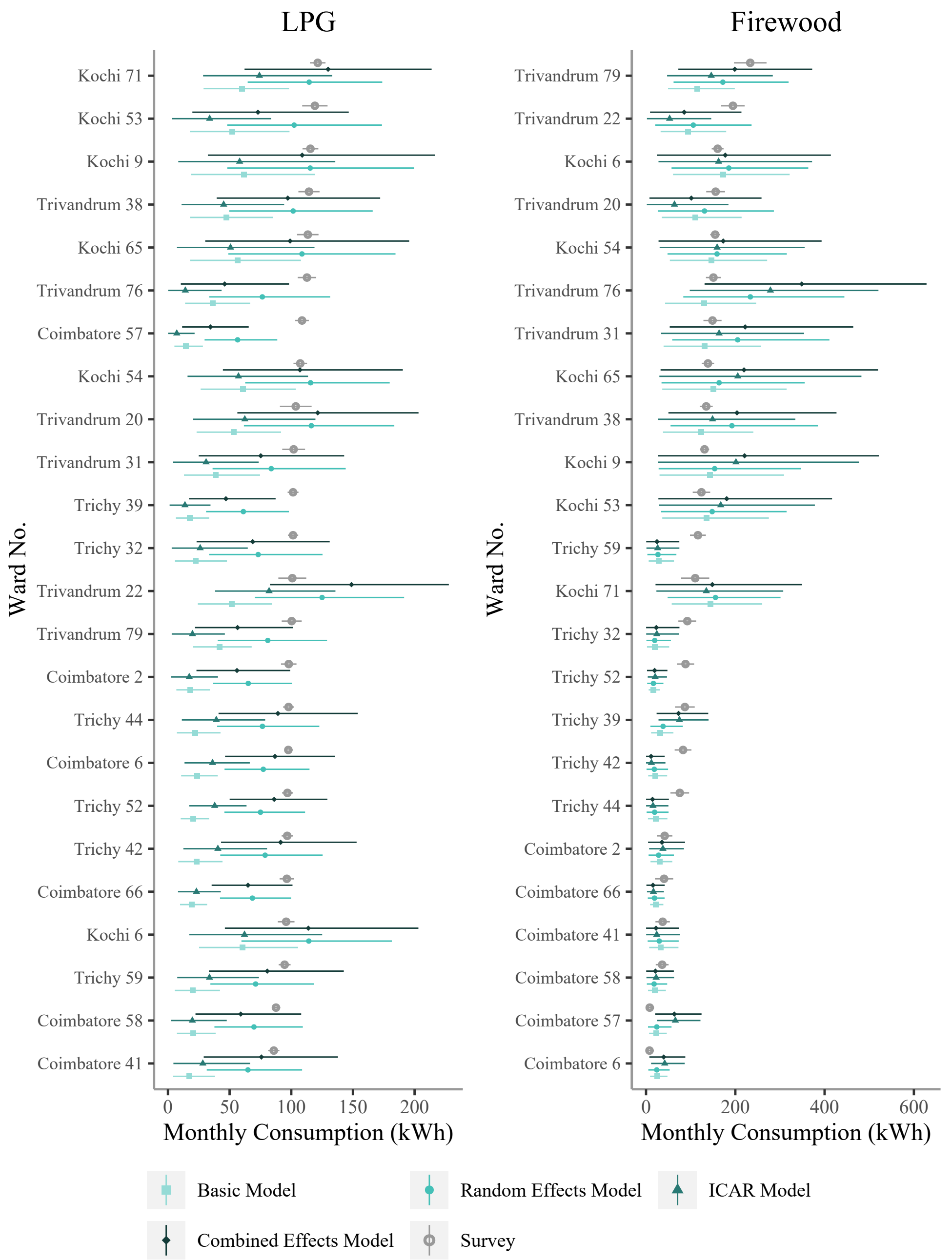


Figure 6.4: Comparison of model outputs and validation data for selected wards. Point values represent mean values from the model or validation samples, and lines indicate the 95% confidence interval. Solid point shapes indicate model variants, and hollow circle indicates validation data for comparison.

it. Accounting for each of these effects adds uncertainty to the estimate and so it is expected that the combined model would have the widest confidence interval.

It is worth noting the difference in performance versus the survey data between the four model variants. In the case of firewood consumption there is little difference in the mean estimates between the four model variants, and the survey data means fall within the confidence interval of model outputs for most wards particularly those with lower consumption levels. This indicates that the levels of consumption in each ward are similar to those in the average urban area in the state and that local spatial effects have a relatively modest effect on magnitude of consumption. The discrepancy between model estimates and survey data for some wards in Tiruchirapalli (abbreviated to Trichy in the figure) could represent a transient random effect, for example added use of fuel in the month surveyed due to a festival or greater availability of fuel than normal.

The situation with LPG is rather different, where the basic model underestimates LPG use, while the random effects and combined models are more compatible with the survey data as a result of both a higher mean estimate and wider confidence interval. In this instance the random effects coefficient provides the bulk of the adjustment from the basic model with the ICAR component adding only small additional adjustments. It is worth noting two points here; Firstly the fact that these cities are larger than the average urban centre within their states would suggest that they are likely to have a better network of LPG suppliers than the average city. In addition, the survey data was collected in 2018-19 coinciding with the government's PMUY LPG connection initiative which targeted certain groups of low-income households, which possibly explains why in wards such as Coimbatore's 57th or Trivandrum's 76th the models underestimate LPG consumption.

6.4 Results & Discussion

The aim of this microsimulation approach was to estimate spatial inequality in access to and use of clean cooking fuels across cities, with the expectation that variation at a ward level arises due to household context, local socio-economic and cultural factors, and spatial effects. The various components of the model outputs are instructive to understand the influence of the different components and how uncertainty is propagated. At the heart of the basic model is a multilevel normal linear model whose coefficients vary according to the primary cooking fuel group a household belongs to. Figure 6.5 shows the estimated linear coefficients of the LPG consumption model for Tiruchirapalli across the six different primary cooking fuel groups. The prior is represented by the dotted line. Due to absence of a comparable primary cooking group variable in the raw synthetic population data, the prior does not vary by group. Nonetheless, using the synthetic population data to form a prior ensures consistency and makes for a more reasonable prior.

The new data from the NSS consumer survey helps reduce uncertainty in the model coefficients with posteriors having smaller variances. While all groups start with the same prior, the new data as well as model design that allows variation by primary cooking fuel group greatly reduce uncertainty. Notice that for groups using kerosene and electricity as primary cooking fuels, the posteriors for some coefficients have greater variance than other groups. This is due to the fact that in the case of Tiruchirappalli the new consumer survey data had considerably fewer cases of households in these groups and there was some variation amongst these few cases. This resulted in posterior distributions for coefficients with slightly greater variance. In particular expenditure slope coefficient a2 for these groups shows uncertainty with respect to the positive or negative nature of this relationship covering a range of values from approximately -0.05 to 0.15 in the case of the kerosene group, and -0.1 to 0.1 in the case of the electricity group. Interestingly the coeffi-

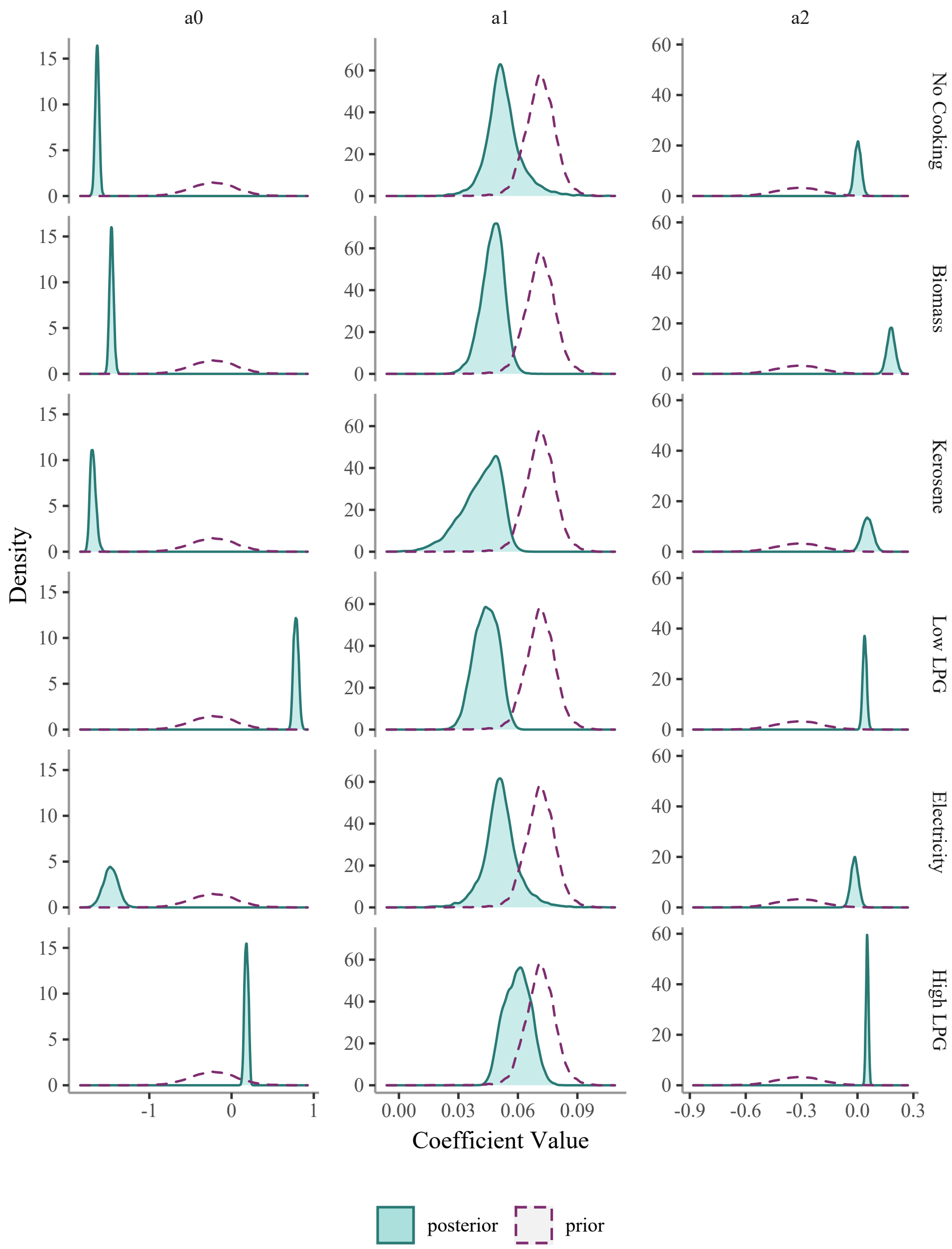


Figure 6.5: Linear coefficients of LPG basic model by primary cooking fuel group for the city of Tiruchirappalli. The curve shown by the dotted line represents the prior based on historical data without group level coefficients.

cients for household size a1, while having slightly different distributions between groups have similar values, within a range of 0.04 to 0.08, suggesting the effect of this predictor does not vary much with primary cooking fuel choice. The compo-

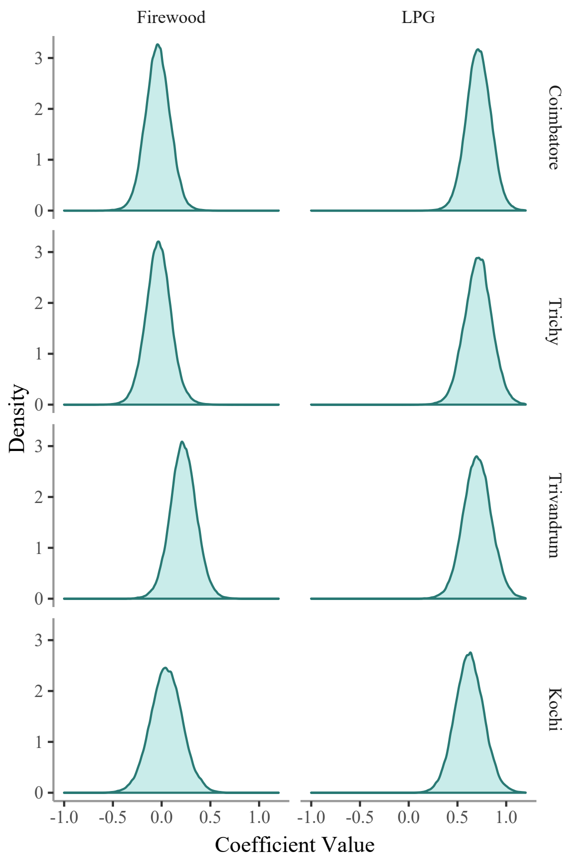


Figure 6.6: Distribution of random effects coefficients for both firewood and LPG across cities.

nents of the model that account for local socio-cultural features and spatial effects are the random effects and spatial smoothing coefficients. These location related coefficients offer valuable insight for the targeting of interventions as they quantify the impact of local socio-economic and community factors relative to the average ward in the state and in that city in particular. The random effects coefficients across the city as a whole help indicate whether the fuel consumption of that city is above, below, or about average for the state. Figure 6.6 shows the distribution of estimated random effect coefficients across all wards in each of the four cities for both firewood and LPG consumption. They indicate all cities are highly likely to have a positive random effect coefficient for LPG with some small difference in magnitude between them, nonetheless indicating LPG consumption above state urban area average. The same is not true for firewood consumption where with the exception of Trivandrum which is likely to be above the mean, the random effect coefficient value for firewood consumption across the cities is normally distributed around or close to zero indicating that firewood consumption levels are close to the state urban average. Note that the variance in the coefficient does vary slightly between cities and fuels, reflecting differences in uncertainty in the magnitude of difference between local fuel consumption compared to the average urban household in the state.

Figure 6.7 shows the mean ward level random effects and spatial smoothing coefficients for LPG consumption across the city of Tiruchirappalli. The mean random effect coefficient shows negligible variation between wards, and indicates that the city as a whole has a level of LPG use well above that of the average urban area in Tamil Nadu. Meanwhile the spatial smoothing coefficient shows how city wards compare to each other. Figure 6.7 shows that the central southern wards of the city have a positive coefficient indicating higher LPG use than the average ward, while the eastern wards under perform relative to the average ward. Analysis of the spatial distribution of ICAR coefficients provides a clear indication of which wards are below average in terms of uptake of LPG and may thus be in need of additional targeted intervention.

Putting these model components together they add up to produce the model fuel consumption estimates as shown in Figure 6.8 which shows the distribution of the monthly LPG consumption estimates for each of the four model variants in

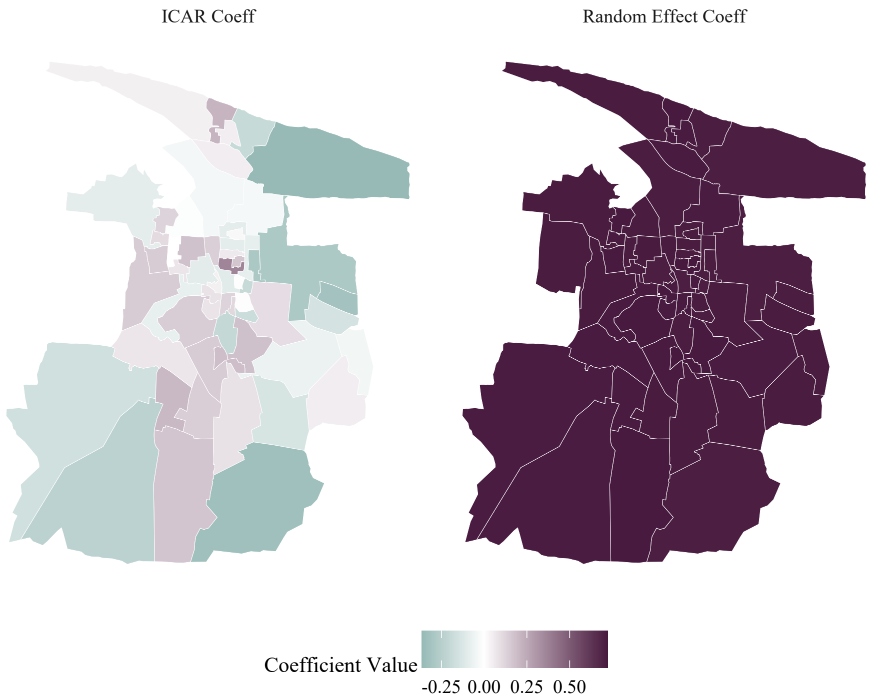


Figure 6.7: Ward level random effects and ICAR coefficient values for Tiruchirapalli.

Ward 38 of Tiruchirappalli. The ICAR model indicates that ward 38 is likely to have above average LPG consumption compared to the city average, and the Random Effects model indicates that Tiruchirappalli's LPG consumption is also likely to be above average. The contribution of these two coefficients are added together in addition to the basic model estimate in the combined model, resulting in a boost to the mean estimate. Note how the uncertainty from the various individual components adds up with the ICAR and Random Effects models having greater variance than the basic model, and this variance compounds in the combined model which accounts for the uncertainty in all three constituent model components.

The impact of the different models can be seen spatially in Figure 6.9, which shows the LPG consumption estimates across city wards for each of the four models. The estimates from the basic model show only small variation between wards

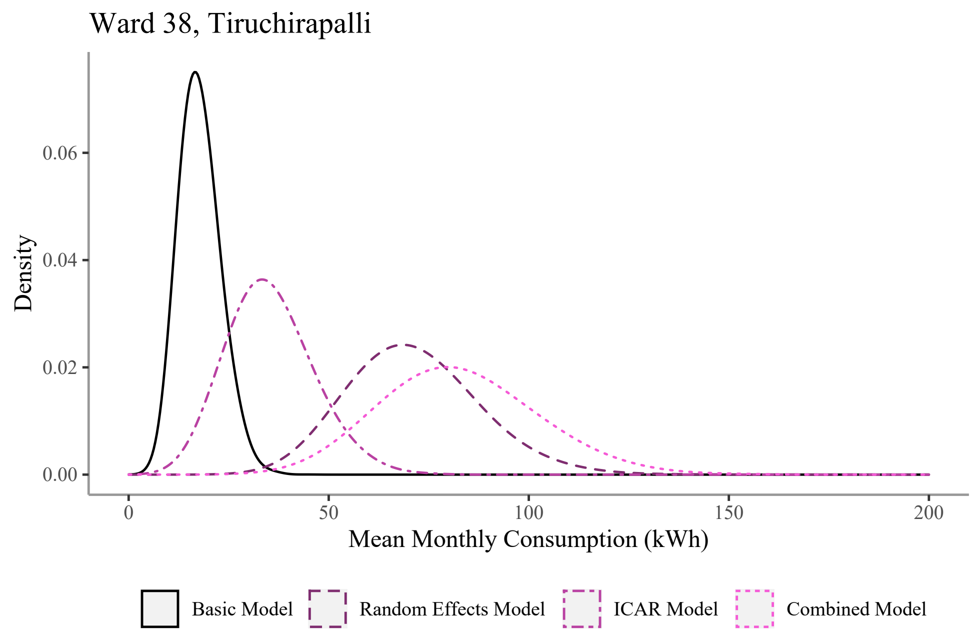


Figure 6.8: Distribution of LPG Consumption estimates for each of four model variants for Ward 38 in Tiruchirapalli

while the random effects model boosts the estimates of fuel consumption relative to the basic model - although still preserving the relative homogeneity between wards. The spatial smoothing model introduces greater variation between wards with either an increased or decreased mean estimate compared to the base model. The combined model features both the city wide increase in mean estimated consumption while also including greater variation between wards, more fully capturing the spatial variability.

6.4.1 Modelled City Outputs

This approach can provide valuable insight for policy makers and stakeholders by indicating spatial patterns of clean cooking use and identifying wards within a city most likely to have less access to sustained exclusive clean cooking. As discussed in section 6.3, while the model output means may vary from actual figures and

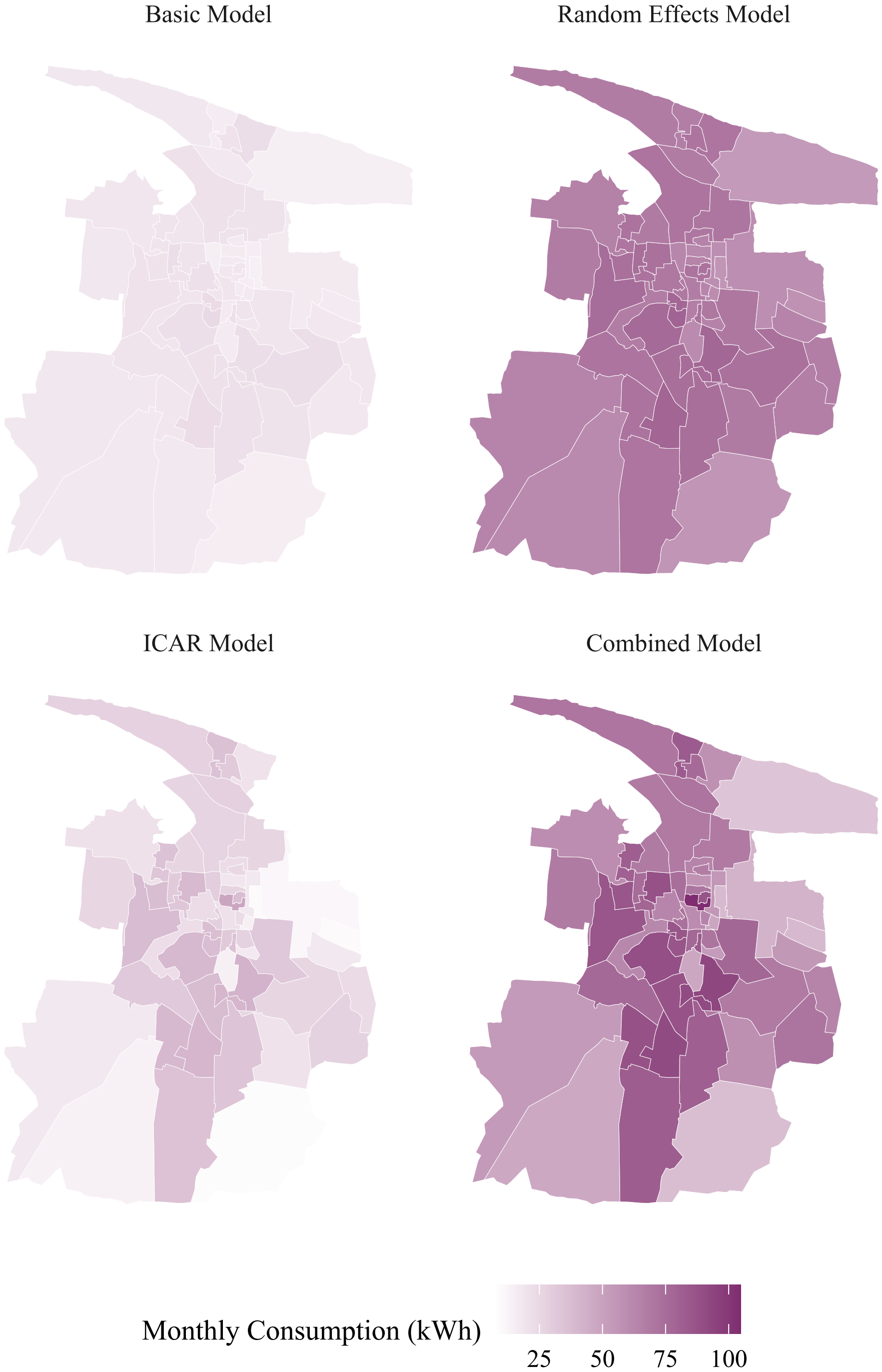


Figure 6.9: Comparison of model variant estimates for mean monthly LPG consumption in Tiruchirappalli mapped by city wards

be subject to relatively wide confidence intervals, the model does reliably distinguish wards with high levels of clean fuel use from those with low-levels of use. By using a representative population of synthetic households not only can this approach estimate ward level fuel use and stacking, but the socio-economic characteristics of the synthetic population can also be considered on a ward by ward basis to contextualise energy use patterns. Figure 6.10 shows the model estimates of LPG consumption and prevalence of concurrent biomass use as well as proportions of households paid daily or weekly as opposed to monthly, for wards in the city of Kochi, Kerala.

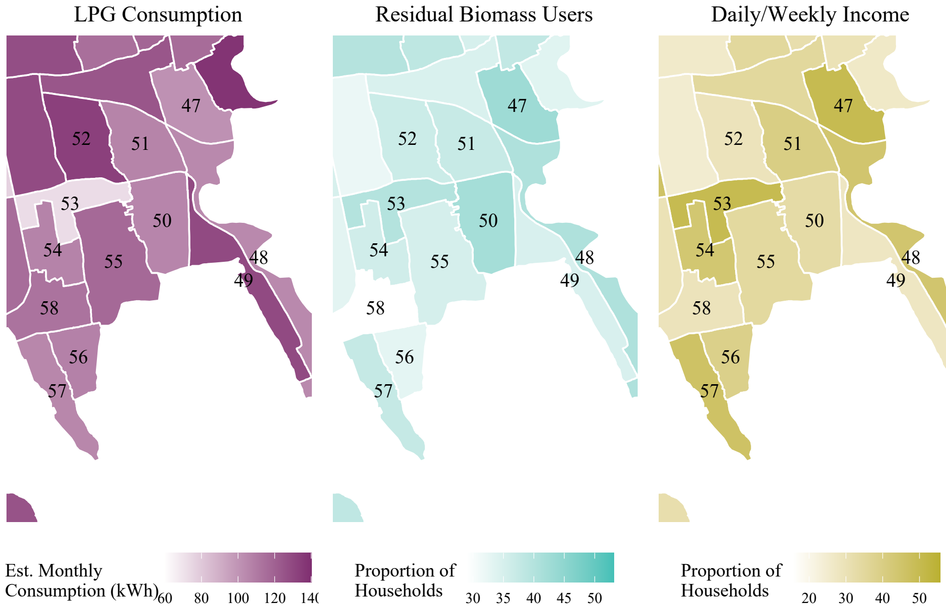


Figure 6.10: Model outputs for wards in south-eastern section of Kochi, Kerala.

In Figure 6.10 wards 47, 48, 50, 51, 53, and 54 have lower levels of mean estimated LPG consumption than surrounding wards, although the model outputs indicate that not all these wards with lower LPG consumption are equal. Amongst these wards 47, 48, 50 and 53 have a greater proportion of household with residual biomass use (continuing to use biomass despite having an LPG stove) suggesting that a different intervention may be required to encourage sustained and exclusive use of LPG in these wards. The microsimulation also provides some local socio-economic context as shown by the map of daily and weekly income earners by ward. Notice that many of the wards with high proportions of households paid daily or weekly are also those same wards with higher proportions of residual biomass users, such as wards 47 and 53.

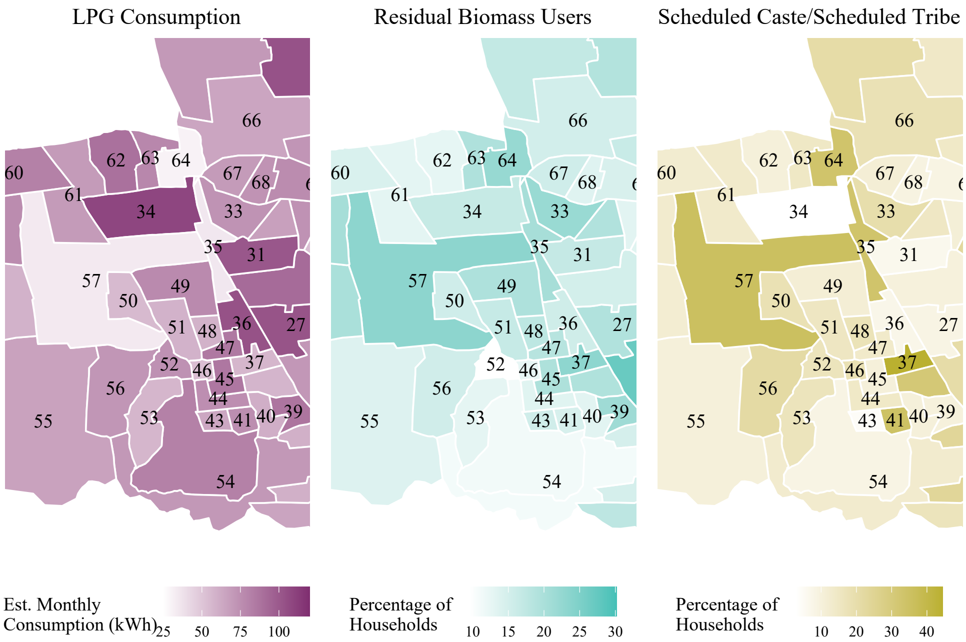


Figure 6.11: Model outputs for wards in the central-western section of Coimbatore, Tamil Nadu.

Figure 6.11 maps model outputs for a central-northwestern portion of the city of Coimbatore, Tamil Nadu. This area includes some of the older neighbourhoods and green areas surrounding the agricultural university. There is considerable variance in the spatial distribution of clean cooking uptake. Notice how clusters with higher proportions of lower caste households (e.g. 37, 57, 35, 64) tend to have a higher proportion of household fuel stacking, and low levels of LPG consumption. Adjacent to some of these wards are wealthier neighbourhoods with lower proportions of low caste households such as wards 31,36, and 27. These have relatively high levels of LPG consumption (average of over 100 kWh/month) and below average proportions of households with residual biomass use. Not all wards with greater proportions of low caste households have high residual biomass use though – ward 41 has a relatively average level of LPG use and residual biomass and this could be a reflection of factors related to its central location, such as having less availability of biomass fuels and greater proximity to LPG distributors.

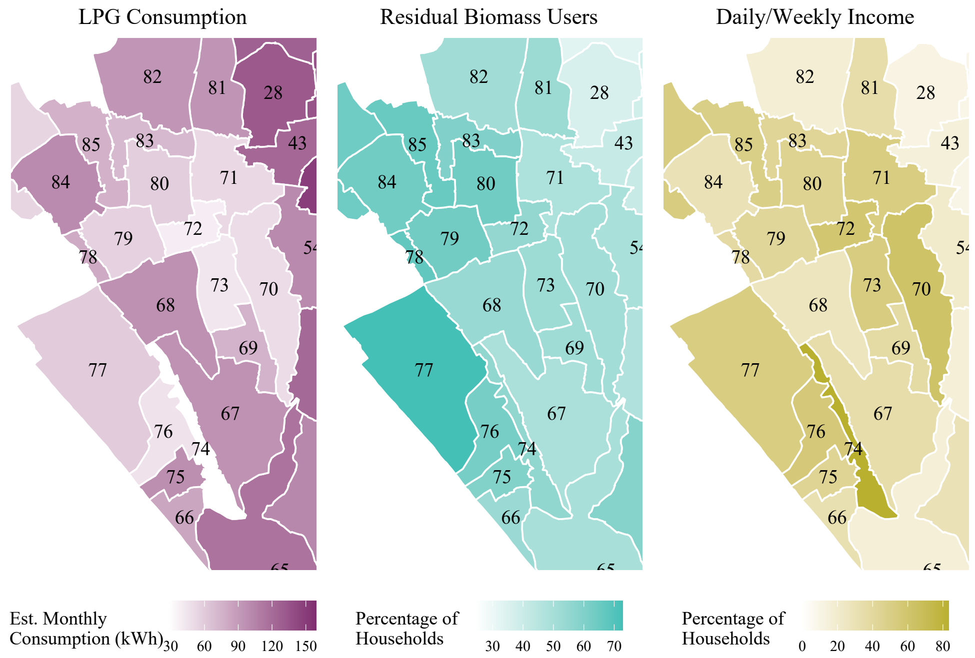


Figure 6.12: Model outputs for wards in the south-western section of Trivandrum, Kerala.

Model outputs in the central-western part of Trivandrum, Kerala are mapped out in Figure 6.12. The wards to the west in this map border the airport (airfield and airport are not counted as a city ward here) – notice that these wards (70-85) have lower levels of LPG consumption than those to their immediate East and South. These wards also have noticeably higher levels of households dependent on daily or weekly income. There is a distinct spatial pattern here, the closer the wards are to the airport along the western edge of this section the greater the proportion of households using some residual biomass – this may also represent the impact of local availability of firewood/biomass that makes it convenient to continue using some biomass after switching stove (for example for rice cooking which can often be cooked on a biomass stove even when households have LPG). This could also reflect a trend in affordability and desirability of housing in proximity to the airport runway.

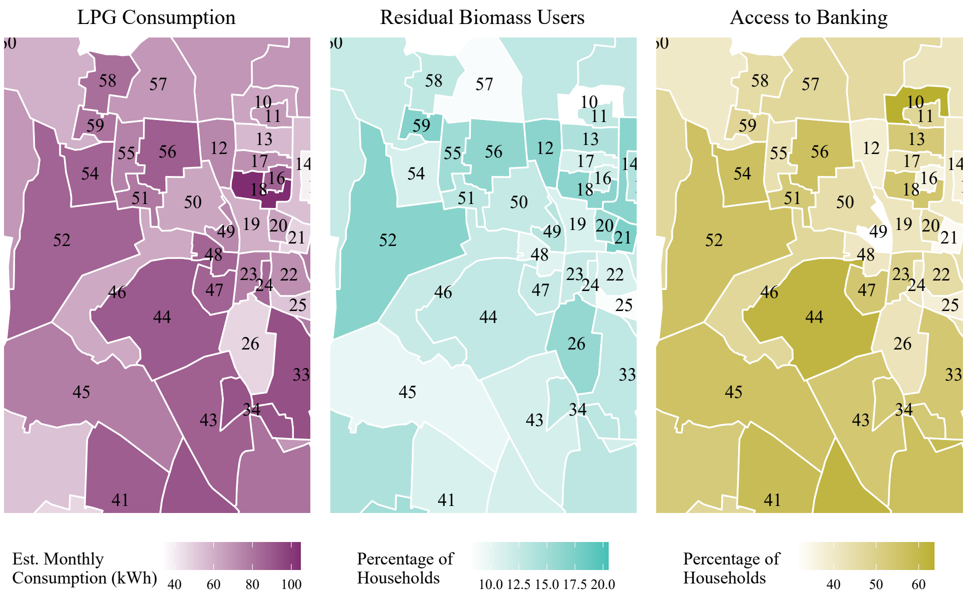


Figure 6.13: Model outputs for wards in the south-western section of Tiruchirappalli, Tamil Nadu.

Finally, Figure 6.13 illustrates model outputs for a section of central Tiruchirappalli in Tamil Nadu which displays some interesting spatial variation in cooking fuel use with respect to access to banking across wards. Many of these central wards have high levels of LPG use such as 18, 44, 52, 54. However some of these have relatively high levels of residual biomass use – in ward 52 greater availability of such fuels towards the edge of the city may be a factor in this continued use of biomass. The ward with the highest levels of access to formal banking, ward

10, has relatively average levels of LPG consumption in this model but some of the lowest levels of estimated residual biomass use suggesting that these households do not engage in much fuel stacking with biomass – it is possible the low LPG indicates use of electricity for cooking in this wealthier ward (recall the model does not estimate magnitude of electricity use).

Interestingly, recent studies have shown how multilevel models can help identify and locate residential inequality and segregation in cities in the US (Arcaya et al., 2018) and UK (Jones et al., 2018). While this model does not imply causality between variables it does contextualise spatial patterns and inequality in clean cooking access with local socio-economic factors. This could allow stakeholders to not only identify wards in need of further attention, but could provide a starting point for understanding the barriers and features of the transition pathways households in that ward may be following. Such data could inform the setting of criteria for a particular policy or transition support initiative, or even possibly support the local tailoring of such criteria for national policies.

Chapter 5 noted the parallels between market segmentation studies and the identification of clean cooking transition pathways, suggesting that this might point to an opportunity for entrepreneurial solutions to some of these barriers. The model outputs lend further support to that notion. The spatially uneven distribution of clean cooking use and fuel stacking could provide an opportunity for local entrepreneurs or LPG dealerships to devise locally targetted solutions such as tailored financing or payment arrangements. A case could also be made for local government and stakeholders exploiting the different physical locations of clusters of households with clean cooking transition problems to trial different targetted solutions before rolling these out in a more widespread manner.

6.5 Limitations and Caveats

As the result of an integration of two different methods, namely population synthesis and a Bayesian multilevel model, this modelling approach suffers from a combination of limitations of both of these. The assumptions of population synthesis were discussed with respect to the generation of a synthetic population, but it is worth revisiting the assumption regarding representativeness of the sample of individual households (the microdata). While a reasonably representative sample of several hundred urban households from the city's state were used to generate the synthetic population it is obvious that these few hundred households do not capture the diversity of the actual hundreds of thousands of households in the city. This means this model may fail to capture the full diversity of households and their energy use.

This relates to another important caveat with this approach, which is that it is dependent on available past data. This means that if a dataset systematically omits a certain type of household or omits certain types of fuel use this will create a systematic blind spot in this model - although given the model aggregate results at a ward level as opposed to at a household scale it is likely minor omissions will have a negligible impact on ward-level fuel estimates. A more pertinent limitation is that past data does not allow the model to foresee shocks such as policy changes - a good example of this can be seen in Figure 6.4. The LPG estimate intervals for Coimbatore ward 57 and Tiruchirappalli ward 39 underestimate observed consumption in the survey data - the data this model is based on predates the recent LPG promotion policies as well as restrictions on kerosene sale in Tamil Nadu. It is possible that these changes led to greater use of LPG by households in these wards but the model could not anticipate these given it did not have any data for households under these policies. An direction for future work could look to modify the model to include priors based on local expert knowledge, which might be offer a way to overcome this limitation.

Criticism of quantitative models applied to the context of clean cooking transitions and more widely clean energy transition in the Global South often centre on the oversimplification models make and the aggregation of individual experience. This reduces individual experiences and narratives to a number or regression line, and can hinder the identification of different transition pathways and challenges. It is important to point out that the outputs from this microsimulation do not imply causality between variables, nor are they designed to prescribe decision making, rather they are geared to support and facilitate an exploration of local energy access issues by stakeholders. This approach makes use of public data to offer a starting point for a process of tailoring and targeting interventions, as opposed to making a concrete recommendation or decision on a quantitative basis.

6.6 Conclusion

This chapter presented a microsimulation approach to estimate cooking fuel use across a city using publicly available data to help characterise spatial inequality in access to clean cooking. In doing so it addressed the need to target and tailor clean cooking interventions for low-income urban households, while having little or no city-scale data available. Using a multilevel model which accounted for household cooking practices, local city effects, and spatial effects produced outputs that showed reasonable consistency with real world data. Allowing model parameters to vary by primary cooking fuel group reduced uncertainty in these, while the model structure propagated uncertainty in ward level effects through to fuel use estimates.

The random effect and spatial components of this model capture the 'city effect' and 'spatial effect', that is to say the model offers insight into how fuel use in a given city compares to the average urban household in the state, and how the fuel use in a given city ward compares to the average fuel use in the city. By simulating a representative sample of households in each ward, spatial distribution of clean cooking access can be examined in context of socio-economic features across a city.

The spatial distribution of model outputs offers a previously unavailable level of detail on residential clean cooking transition across a city. The examples of the cities modelled show how not only disparities in LPG use can be related to underlying socio-economic inequality, but more importantly the risk of fuel stacking and non-exclusive use of clean cooking fuels displays strong spatial trends. Understanding the link between the underlying spatial inequalities and the prevalence of clean cooking and fuel stacking may offer a path to designing more effective targetted interventions.

With India's urban population set to expand in the coming decades, limited data on the effects and distribution of spatial inequality is a challenge for policy implementation, not only for clean cooking, but wider energy access issues. Design of effective interventions requires stakeholder involvement and consideration of local household needs. Cost-effective and data-efficient modelling approaches such as the one presented here offer an exploratory view of clean cooking access across a city that can facilitate such a process. The insights from this model and its design alongside the findings of prior Chapters 4 and 5 give rise to further research questions and implications for residential energy transition in India which are explored in Chapter 7.

7 Conclusions and Future Work

All models are wrong, but some models are useful.

- George Box

7.1 Context

This thesis has focused on the need to better understand and model the influence of social factors and household practices on clean cooking transition of low-income households. The purpose is to inform interventions and policies which address the needs of households, characterising barriers to access and inequalities at play amongst these households.

Recall that a suite of recent policies including the flagship PMUY have sought to tackle the issue of access to clean cooking and the associated negative health impacts of biomass fuel use. PMUY aims to provide subsidised LPG connections to BPL households across India, and has succeeded in its aim of providing new connections to 80 million households. But not all low-income urban households have been connected, and many have not sustained their use of LPG or have continued to use biomass fuels leaving the health benefits of switching to clean cooking tantalising out of reach. Studies on economic determinants of transition offer limited insight into the reasons for this, with the explanation for these unintended outcomes lying in social and local factors.

The overarching research question this thesis asked was, how can the influence of social factors and household practices on clean cooking transition be integrated into a quantitative energy model? In addressing this question the findings of this thesis can be distinguished between those that contribute to understanding energy transitions, and those regarding the design of a method and modelling approach for energy transition in urban households. While distinct, these are also two sides of the same coin; the understanding of energy transitions helps to describe a 'Social Logic', which both informs and is informed by the design of the modelling approach. This chapter presents the key findings as well as recommendations for clean cooking interventions in urban India. Limitations of this research as a whole and avenues for future research arising from this thesis are also discussed.

7.2 Understanding Energy Transitions

To better understand residential cooking energy transition, this thesis explored the influence of local socio-economic and cultural factors and household practices. This involved grappling with the inherent heterogeneity across populations, between households, and in spatial distributions. This is something that previous quantitative studies reviewed in section 2.3.1 have not thoroughly addressed. Data science approaches offer a way to quantitatively include the influence and nature of this heterogeneity.

The influence of socio-economic and cultural factors within quantitative models is explored using a BRT model. Relevant determinants include occupation, caste, location, income, and education, as well as appliance ownership, quality of utility access, changes in time spent on energy related practices, and electricity use.

The results show that socio-economic determinants do not necessarily follow a linear trend for all households, instead featuring thresholds for changes in influence on transition. Additionally a statistical clustering reveals different typologies of households that did transition. The clustering demonstrates that heterogeneity in socio-economic features of households manifests in different typologies of transitioning household, indicative of distinct enabling circumstances. Continued use of biomass fuels amongst transitioning households is shown to be widespread, highlighting the problem of ongoing fuel stacking which results in continued indoor air pollution from biomass cooking.

The heterogeneity amongst low-income households, and the transition pathways that emerge from this, were further investigated using a novel clustering method. The method used quantitative and qualitative data in two clustering stages to identify and match household typologies and narratives of people obtained via surveys. The method was applied to low-income households in Bangalore to identify clean cooking transition pathways. The findings illustrate that wider socio-economic inequality is intertwined with access to clean cooking. These also confirm the hypothesis that residential clean cooking transition is influenced by the interaction of three components: the influence of local socio-economic and cultural context a household is in, such as the case of households in non-notified slums; the wider policy and economic context, for example seen in the effect of reduced availability of subsidised kerosene at PDS shops; and the energy related practices and habits of that household, exemplified by households choosing to cook rice instead of bread to save money.

The different pathways and barriers indicated for different clusters of households demonstrate that heterogeneity across households and their local context can condition response to policies and incentives, leading to marked differences in uptake of sustained clean cooking even within the same city. This was illustrated by the response of low-income households to reduced kerosene availability; for some it incentivised their switch to an LPG stove, but for others they reverted to using biomass or turned to higher-priced black market kerosene. The primary data from Bangalore also showed fuel stacking was common among slum households that did have an LPG stove.

Findings from Chapters 4 and 5 inform the design of the microsimulation model presented in Chapter 6. The spatial distribution of clean cooking and fuel stacking is modelled at a city scale. The model shows that in addition to heterogeneity across households the local context also has specific significance. The model shows that there is significant variation in clean cooking fuel use and prevalence of fuel stacking across city wards. Wards with lower sustained use of clean cooking fuel and greater prevalence of fuel stacking tend to be consistent, often corresponding to the spatial distribution of socio-economic inequality in the city.

Distinct typologies of households with differences in incomes and locations were shown in Chapter 4 to transition to clean cooking, with distinct enabling circumstances. However the analysis of low-income households in Bangalore in Chapter 5 went a step further and showed that even for households within a given income range and in the same city, different transition pathways and barriers exist. To ensure sustained uptake of clean cooking, the targetting criteria for interventions and solutions cannot just rely on income measures. Instead measures need to be tailored to the barriers faced by households locally.

7.3 Modelling Urban Energy Transition

There are two methodological contributions of this thesis that widen the scope of quantitative models in understanding energy transitions. Both offer means of characterising the influence of socio-economic and cultural heterogeneity, although do so at different scales. The first uses statistical clustering methods to characterise household heterogeneity using mixed data. The second is a microsimulation model providing the means to characterise spatial heterogeneity in clean cooking transition.

The mixed data clustering analysis presented in Chapter 5 employs statistical clustering methods to jointly analyse qualitative interview data and quantitative survey data with a view to distinguishing different predominant energy transition pathways and their problems. This method requires high resolution data from cities to infer typologies of households with distinct energy transition pathways. In urban areas where households display considerable heterogeneity, being able to distinguish different barriers and challenges to access is key to supporting sustained transition. The approach presented provides a systematic way of combining qualitative and quantitative data on residential energy use.

The inclusion of qualitative data in the clustering analysis is an important feature, as through these individual narratives the interaction of a household with its local socio-economic and cultural context, and the wider policy and economic context, can be more fully described. For example interview data highlighted how many slum households have financial priorities other than LPG, and can struggle to make the relatively large monthly payments given their daily income. An added benefit of this approach for informing policy and intervention design is that the pathways identified can highlight cases where households with different socio-economic circumstances can face similar barriers and access problems, and vice versa. This was illustrated by the case of recent migrant wage labourers and older slum households who both had low access to electricity and LPG and had not accessed subsidies, the former due to lack of documentation and the later due to lack of awareness of eligibility.

Chapter 6 proposes a Bayesian multilevel microsimulation model to incorporate the influence of socio-cultural factors and household practices in estimating clean cooking fuel use and fuel stacking at a city scale. This microsimulation approach forms a key contribution of this research, with several distinguishing novelties. Unlike the mixed data pathway clustering which requires detailed household level data, the microsimulation model characterises city-scale distribution of clean cooking uptake in the absence of such detailed data. The model is able to generate estimates by combining multiple publicly available data sources, offering a resource efficient method to model residential energy transition at an urban scale.

The Bayesian multilevel modelling approach offers the distinct advantage of being able to combine different data sources, and propagate uncertainties in model parameters through to ward-level estimates. The use of a multilevel model is also well suited to addressing the heterogeneity across households and the local context, by including household level and ward level components that account for household cooking practices, local city effects, and spatial effects of the local ward. The ward level random effects and ICAR components can provide insight on how the city compares to the average urban area in the state, as well as how city wards compare to each other, highlighting wards in need of additional support or intervention. By simulating a representative sample of households in each city ward, the spatial distribution of clean cooking access can be examined in the context of spatial socio-economic inequalities. The outputs show how access to utilities or services such as banking correspond to patterns in clean cooking fuel use and stacking. They also demonstrate that local context not captured by socio-economic variables may sometimes be associated with greater chance of fuel stacking, such as local availability of fuels. This provides data that can facilitate communication and collaboration with other disciplines and stakeholders in addressing clean cooking transition problems.

These methods both focus on characterising the problems faced by households. The introduction of this thesis highlighted the need to bridge technical-economic and social approaches through common framing and bridging of data, and the background in Chapter 2 elaborated on the need for a problem-focused approach to promote such interdisciplinary cooperation. Both of these methodological contributions are concerned with identifying, describing, and characterising barriers to transition, and problems in sustaining clean cooking, rather than modelling cause and effect. The approaches developed in this thesis contribute problem-focused data science methods to the methodological toolbox for energy research in urban India and the wider Global South.

7.4 Recommendations for Policy and Practice

From the conclusions discussed above, arise a number of key recommendations for practitioners, decision making, and policy design addressing residential clean cooking transition in urban India.

1. For many low-income households LPG access alone does not eliminate the negative health impacts of biomass use due to high levels of fuel stacking. Clean cooking efforts must complement LPG access programmes with interventions to actively reduce dependence on solid biomass fuel.

2. Access to clean cooking cannot be viewed in isolation of other socio-economic development indicators. Low-income households, in particular non-notified slum households, can struggle with access due to the informal and precarious nature of their situation. This could be addressed as part of a wider

initiative to improve living standards and legal status which in turn could offer co-benefits for achieving other sustainable development goals.

3. The effects of spatial inequality in cities mean that similarly low-income urban households can face very different barriers to sustained clean cooking. Income alone does not distinguish the barriers a household is likely to face, and so criteria for programme eligibility and interventions need to be widened beyond purely income-based measures.

4. These different barriers to sustained clean cooking also mean that clean cooking programmes in cities cannot take a one-size fits-all approach and instead a multi-faceted approach is required with different measures tailored to each of these barriers.

5. Addressing the challenge of achieving transition to sustained clean cooking use among low-income urban households requires a better understanding of who faces what barriers and where. Spatially disaggregated data is needed to identify communities to engage with, and to help target tailored interventions. The microsimulation tools developed in Chapter 6 of this thesis could provide practitioners and policy makers with such data.

7.5 Limitations

As is the case with any research the analysis and approach taken in thesis is not without its limitations, and while specific limitations of methodologies have been noted in each chapter there are some more general limitations and caveats applicable to this research as a whole. First and foremost is a matter of geographical specificity; while the analysis in Chapter 4 did use a nationwide dataset to characterise the effect of socio-economic factors and determine a clustering tendency, detailed city-scale analysis in Chapter 5 and modelling in Chapter 6 relied on primary data collection in the southern states of Karnataka, Kerala, and Tamil Nadu. As a result this thesis has had a limited geographical focus on cities in these southernmost states, which have a distinct geography, climate, politics, and culture. Further work is needed to verify the widespread suitability of this approach, expanding this analysis and modelling to other regions.

This alludes to another issue with this research which is related to the scale of disaggregation. While analysing and modelling at a less aggregated scale than previous studies, using the city ward level as the level of aggregation instead of the overall city, district, or state, this is still a relatively coarse level of resolution in cities. Particularly in larger cities where wards can contain over 20,000 households with considerable heterogeneity and different neighbourhoods and communities within this. The choice was constrained by available data both for selection of survey sample locations as well as generation of the synthetic population, seeing as in both cases the census ward tables were used. Collection and use of GPS location data from surveyed households would present some serious ethical challenges, and hence the ward level aggregation provides a reasonable compromise between privacy concerns and disaggregation. Some city science and urban analytics methods discussed in section 7.6 might offer a means to analyse at a finer scale without needing to compromise on household privacy.

7.6 Future work

Three distinct avenues for further work are identified that build upon the approach and findings of this research. These focus on the role of innovation and entrepreneurship for addressing clean cooking transition problems, the applicability of the approach to other energy transition and demand reduction efforts, and the case for greater use of 'city science' approaches for addressing energy inequality in India's cities.

7.6.1 The role of innovation

While this thesis has studied energy needs and barriers to access of households and considered the role and fit of policy to these, as noted in Chapters 5 and 6 the findings may point to an opportunity for innovation and entrepreneurship. The different pathways identified may be likened to market segments, each presenting an opportunity for an innovation either of product or business model. The analysis has looked at how present technologies, policies, and business models fit with different household transition pathways. However, innovations which change the offering to better suit the needs and circumstances of households could prove effective to provide tailored solutions for different barriers to transition. For example, extrapolating from findings in this thesis, potential innovations could address costs, convenience for cooking staples like rice with clean fuel, or allowing for monthly LPG bills to be paid reliably in small instalments.

Barrie and Cruickshank (2017) talked about the role of disruptive innovations in helping to bridge the 'last mile' challenges in electricity access, giving the example of successful pay-as-you go solar home systems in Eastern Central Africa. The challenge of transitioning low-income households to clean cooking can itself be regarded as a 'last mile' problem. In addressing such last mile problems, a frugal approach to innovation might be able to succeed where other approaches can not - put simply frugal innovation is the idea of 'doing better with less resources for more people' (Prabhu, 2017). Prabhu and S. Jain (2015) illustrate examples of such frugal or *jugaad* innovations providing households with solutions for energy services, such as the MittiCool clay 'fridge' - a USD 50 fridge made of clay that can keep food cool without using electricity, perfectly suited to unelectrified rural households in India.

To illustrate this potential in a clean cooking context consider the example of cooking rice, in section 5.4 we noted that rice cooking was often seen as a frugal low-cost choice of staple and could be cooked with a small amount of biomass. The electric rice cooker was invented in Japan in the 1950s and helped facilitate a transition to clean cooking in postwar Japan (Nakano, 2009). The design of the devices was tailored to the needs and cooking practices of Japanese households, designed to be efficient and convenient. Might there be an opportunity for a Jugaad electric rice cooker? A low-cost, efficient and convenient device that is locally manufactured and designed to replicate the rice cooking method employed on a biomass stove.

Social Enterprises

While this potential for entrepreneurial involvement and innovation could be capitalised upon by any range of businesses, there is a particular type of enterprise which may be particularly suitable and could work in tandem with government to deliver targeted solutions, or offer a model for how local governments could better design and implement interventions. Social businesses or enterprises have as their primary purpose a drive to serve society, but they have products, customers, and markets just like any regular enterprise and operate on a no-loss, self-sustaining basis (Yunus et al., 2010).

As part of fieldwork for this project meetings were held with representatives from several social enterprises, including Selco who market solar home systems and appliances to rural households, and Pollinate who sell solar lighting kits to low-income urban households. Both of these have interesting business models relying on community level engagement, flexible financing arrangements, and crucially they were constantly fine tuning and experimenting with their product offering, with a keen interest in addressing inequality and poverty.

Success in scaling up such social business hinges on increasing the number of customers and expanding the offering either through product development or diversification (Bocken et al., 2016). To this end it is important to explore to what extent a city-scale model such as the one presented in this thesis could facilitate the identification and experimentation with new solutions, and support scaling up of social enterprises. To do so such a model must not only yield credible outputs but must also be presented in forms that are accessible to their end-user.

7.6.2 Beyond clean cooking

While the context of this thesis has been clean cooking transitions, there may be scope for the approach to be used for other household energy use, such as cooling demand and appliance use which is likely to contribute to forecasted growth in energy demand in the residential sector (International Energy Agency, 2019). The need to reduce carbon emissions to limit the effects of climate change motivates a push to improve efficiency of residential energy use as well as understand how to meet energy needs while decarbonising electricity infrastructure.

Any effort to adapt the proposed microsimulation modelling approach for such a purpose would likely require more data, given that current public data has limited information on electricity consumption patterns. An analysis along the lines of that undertaken in Chapter 5 would also be needed to identify different pathways and typologies of household to adapt model design. The model would also have to take into account the temporality of use, which while not of great relevance in the analysis of cooking fuel use, would be critical in understanding electricity demand. There could be an opportunity to integrate different types of data and analysis to enhance the model - for example studies have explored the use of imagery and ranging data to map out building features across cities (Wu et al., 2019), as well as the use of social media data to map out spatial trends in creative activity across a city (Reuschke et al., 2021).

7.6.3 A case for city science approaches

L. Bettencourt and G. West (2010) wrote a comment piece in Nature on what they termed "A unified theory of urban living" where they made the case for a science of cities to understand how the complexity at play in cities impacted society and the environment. It built upon work by G. B. West (2017) on the idea of simple scaling laws that can make predictions of complex systems, belying the underlying network properties of cities and emergent dynamics of the complex interactions of that network. While there is still debate on exactly what this nascent field of urban analytics and city science is and how to define it (Kang et al., 2019), it is interested in the complexity at play in cities and using the growing amount of data and data science tools to deliver answers to the big questions of our time, such as inequality, environmental degradation, and aging (Batty, 2019).

Recent work by Sahasranaman and L. M. Bettencourt (2019) applied a city science scaling law approach to look at urban income levels in India, and compared how cities performed versus expectations. These approaches have also been applied to consider inequality in cities, with Sarkar (2019) noting that specialised industries in smaller cities could lead to worse than expected inequality, and that larger cities faced greater challenges with inequality due to the greater concentrations of high-income earners.

The models and methods proposed in this thesis arguably encompasses elements of a city science approach and could be extended applying other methods and theories in this field to perhaps predict energy access challenges that might arise as cities grow. But there is another aspect to this which is that current theories of cities focus on the long term, but as Batty (2019) points out when discussing the meaning of urban analytics, there is plenty of data in real time but there is a lack of robust theory of the urban that embraces this data and considers the questions of our time here and now. Could approaches such as those proposed be used to lend insights and develop such theories?

This thesis began with the story of a family in Bangalore, whose story is simultaneously unique yet similar to countless others. Whatever advances modelling and city science approaches offer to tackling the challenge of sustainable clean energy transition, one must not forget that behind every data point in a model or prediction are real households and people. Models will only be useful if they address their problems.

A IHDS Data Handling, Pre-processing, & Auxiliary Data

Data Handling and Pre-Processing

For ease of analysis the dataset was rearranged, with errors, anomalous, and missing data pre-processed. All pre-processing was done using Python 3.6 and the the resulting dataset was saved as a csv file for analysis in R.

Handling Missing Values

Missing data values were an issue for several entries in the IHDS dataset. A key challenge was that of estimating the quantity of biomass fuels used by households that collected rather than purchased their fuel, for these households there was no indication of the magnitude of fuel use. In addition to this the data contained a number of anomalies and outliers that needed to be dealt with before analysis of trends.

Estimating Collected Fuel

The energy related questions in the IHDS only quantified spending on fuels, as demonstrated by the IHDS dataset a majority of Indian households using biomass fuels in rural areas collect their own fuel. Thus an estimate of the quantity of fuel collected for use by the household had to be made from the available information.

In conducting this estimation three key assumptions were made:

1. The quality and energy content per unit mass of the biomass fuel collected is assumed to be the same as for that purchased;
2. Households in a similar region, and of a similar income will collect the same amount of fuel as that purchased by non-collecting households;
3. The distribution of purchased fuel with income can be approximated by a truncated inverse exponential distribution.

The first of these assumptions can be reasonably assumed, given that purchased biomass fuel is likely to be local and perhaps even sourced from the same place that it is being collected. The second assumption is important as it allows for collected fuel to be estimated on the basis of location and income only, ignoring any possible impact on collected fuel use from other features, which without further analysis, data, or assumptions would not be possible to quantify.

The recorded data on purchased biomass fuels was used to fit an inverse exponential function to the distributions for each type of biomass fuel. Households which had collected fuel were identified and incomes input to the purchased fuel distribution function to calculate an estimated quantity of collected fuel.

Errors and Outliers in Dataset

While the dataset is provided relatively free of errors by the IHDS, there are a number of anomalies which can hamper statistical analysis of the data. These include missing values where an answer simply was not entered, or more commonly a categorised missing response code which distinguishes between answers that are not applicable, and cases were answers were refused by the interviewee. Where the case was simply a null response and treating it as a zero would not impact the data this was converted accordingly, however where this was not the case the offending instance was removed.

Auxiliary Data

To analyse the data on fuel consumption and energy use behaviour in the IHDS Survey, ancillary data was required to convert spending on electricity into energy consumption.

State by State Electricity Tariff Table

An electricity tariff reference table was constructed with government data on mean electricity tariffs for household consumers (Government of India Planning Commission, 2012). It was not possible to capture the different subsidisation and tariff structures from state to state with the mean tariff values used, however at a statewide level the mean tariffs still provide a sufficiently accurate estimate of electricity consumption.

Table A.1: Table of Mean Electricity Tariffs by Indian State

	State ID	State Name	Domestic Tariff (INR)
1	28	Andhra Pradesh	2.82
2	18	Assam	3.93
3	10	Bihar	2.29
4	22	Chhattisgarh	1.91
5	7	Delhi	2.82

continued on next page

Table A.1() —continued from previous page

	State ID	State Name	Domestic Tariff (INR)
6	24	Gujarat	3.72
7	6	Haryana	3.38
8	2	Himachal Pradesh	2.93
9	1	Jammu and Kashmir	1.61
10	20	Jharkhand	1.03
11	29	Karnataka	3.62
12	32	Kerala	1.98
13	23	Madhya Pradesh	3.83
14	27	Maharashtra	4.17
15	17	Meghalaya	3.04
16	21	Odisha	1.35
17	3	Punjab	3.20
18	8	Rajasthan	3.84
19	33	Tamil Nadu	1.77
20	9	Uttar Pradesh	2.95
21	5	Uttarakhand	2.30
22	19	West Bengal	3.92
23	12	Arunachal Pradesh	3.23
24	30	Goa	1.85
25	14	Manipur	2.60
26	15	Mizoram	1.92
27	13	Nagaland	2.29
28	34	Puducherry	0.99
29	11	Sikkim	1.60
30	16	Tripura	3.16
31	35	Andaman and Nicobar Islands	1.70
32	25	Daman and Diu	1.30
33	26	Dadra & Nagar Haveli	1.30

Fuel to Energy Conversion Factors

While fuel mass to energy content conversions (the heat of combustion) for standard fuels such as LPG and to a lesser extent kerosene are defined within a fairly narrow range (Çengel and Boles, 2011), biomass energy content can vary greatly between different locations. However this was simplified for this analysis by taking a single nationwide value for each type of biomass fuel. The fuel mass to kWh conversion values used are detailed in table A.2.

Table A.2: Table of fuel mass to energy conversions used in converting IHDS fuel expenditure data. Adapted from Çengel and Boles (2011)

Fuel	**Energy Density (kWh/kg)**
LPG	13.721
Firewood	4.356
Kerosene	9.630
Crop Residue/Dung (estimate)	3.900

B Household Survey Instrument & Ethics Approval

This Appendix contains ethics review considerations and approval as well as a reference copy of the survey instrument developed for a survey of Indian Household Energy Use. Translated and interactive xml versions of this form were used to conduct the survey between October 2018 and June 2019.

Ethics Approval

The following summarises the information on this study submitted for ethics approval by the Department of Engineering Ethics Committee.

Description of the Purpose of the Research

The research involves surveying 2100 Indian households to obtain information on their energy use habits and patterns, to identify energy transition trends driven by non-income factors that could impact upon infrastructure and technology development strategies.

Surveys will be conducted in person via a 30min (approx.) interview with the head of the household or a representative. Data collected will include income and expenditure data, occupation and education of household members, and information on appliance ownership, and magnitude and time of use of energy in the household.

Data will be collected in India and a copy of the dataset will stored by academic partners at the Indian Institute of Human Settlements in Bangalore, India. Data will be collected exclusively outside the EEA and may be stored on cloud services with servers outside the EEA.

Personal identification data in the form of name and date of birth will only be collected for the purpose of facilitating the interview with the respondent, and sorting household members into ages. All data will be input through a digital app after each interview, and names and dates of birth will be omitted from the official data input to the final digital dataset, with the only record being on the form used by the interviewer. Once data has been input digitally names will be deleted after which the household response will only be referred to by an anonymous randomly generated alphanumeric code and corresponding census ward/district code.

Participant Consent

Consent will be obtained upon arrival at the household prior to commencing the interview. The interviewer will read out an information statement, which prospective respondents will be given time to consider, and consent to. The first questions on the survey pertain to obtaining explicit consent, and the respondent and interviewer will sign/digitally mark as witnesses to this. For each household visited consent will be obtained personally prior to commencing the survey.

Households will be selected at random within urban districts of interest, selected on the basis of target income ranges, and type of low-income and social housing, and zoning.

A copy of ethics approval letter for the survey is included below and the Participant Consent Form is included with the survey instrument below.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Dr Giovanna Biscontin
Chairman of the Ethics Review Committee

André Neto-Bradley

Division D

13 August 2018

Dear Mr Neto-Bradley

Ethical Approval for your Research Project 'Social Logic Of Energy Use In Indian Households'

The Department's Research Ethics Committee has considered the documentation you provided in support of your research project in line with recommended procedures concerning ethical approval of research.

I am able to inform you that, with respect to ethical considerations, approval has been given to your project. Please note that this approval is based on the documentation you provided. You must re-submit your application to the Committee should you subsequently make any substantive changes relating to matters reviewed by the Committee.

We are content for this letter to be forwarded to your grant sponsors or to any partner institutions you may be working with if appropriate.

Yours sincerely

[Signature removed for GDPR reasons]

Giovanna Biscontin

Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
research-ethics@eng.cam.ac.uk

Survey Instrument

Surveys were conducted in person via a 20-40min (approx.) interview with the head of the household or a representative. In compliance with the ethics review data was collected in India and a copy of the dataset was stored by academic partners at the Indian Institute of Human Settlements in Bangalore, India. Data was collected exclusively outside the EEA and may be stored on cloud services with servers outside the EEA.

Personal identification data in the form of name and date of birth was only collected for the purpose of facilitating the interview with the respondent, and sorting household members into ages. All data was input through a digital app after each interview. Once raw data had been verified name data was deleted after which the household response was only referred to by an anonymous randomly generated alphanumeric code and corresponding census ward/district code.

INDIAN HOUSEHOLD ENERGY USE SURVEY

September 2018

Indian Institute for Human Settlements
&
University of Cambridge

[INTERVIEWER: PLEASE IDENTIFY YOURSELF BY NAME & ORGANISATION AND THEN READ THE FOLLOWING STATEMENT EXACTLY AS IT IS WRITTEN]

CONSENT STATEMENT

Before you decide to take part in this study it is important for you to understand why this research is being done. Please take the time to consider the following information carefully and discuss with others if you wish.

This study is being conducted by the Indian Institute for Human Settlements (IIHS), Bangalore and the University of Cambridge, UK to investigate trends in energy use in Indian households and identify factors that may affect the uptake of new technologies. We are asking many people in cities across the country to participate in this very same survey.

The interview is voluntary, and you are free to refuse or withdraw at any time. During our visit we would like to ask a series of question related to your energy use and some other aspects of your life such as education, family, and work.

If you choose to not reply to any questions in this questionnaire you are free to do so. If you decide to answer some or all of the questions the information you give us will be used only for the purpose of research and publication. This study will allow people to learn about the energy use habits and trends of people in India but will not reveal what you personally said.

<u>Your name and any other personal information will be retained by IIHS and the University of Cambridge in a confidential manner and will not be disclosed with any third parties.</u> All data used in any outputs will be identified by an anonymised code number only. Results may be published in academic journals and presented at conferences, where results will usually be presented for a group, and any individual responses will be completely anonymous.

For information about the study please contact Dr. Amir Bazaz
at IIHS Bangalore City Campus, 197/36, 2nd Main Road, Sadashivanagar, Bangalore 560 080
Phone: 80 6760 6666 | Email: abazaz@iihs.co.in

André Paul Neto-Bradley *at Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. | Email: apn30@cam.ac.uk*

0.1 Do you agree to be interviewed (if you answered yes to all extra consent questions overleaf)?

Yes=1
No=0 [ ]

Respondent's Initials ____________

Interviewer's Initials ____________

Consent Questions Continued…

Please tick the appropriate boxes	**Yes**	**No**
Taking Part		
I have read and understood the project information sheet dated 09/2018.	☐	☐
I have been given the opportunity to ask questions about the project.	☐	☐
I agree to take part in the project. Taking part in the project will include being interviewed.	☐	☐
I understand that my taking part is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	☐	☐
Use of the information I provide for this project only		
I understand my personal details such as phone number and address will not be revealed to people outside the project.	☐	☐
Use of the information I provide beyond this project		
I agree for the data I provide to be archived in anonymous form in an open data repository for use in further research not related to this project.	☐	☐
I understand that other authenticated researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
So we can use the information you provide legally		
I agree to assign the copyright I hold in any materials related to this project to Dr. Amir Bazaz, Dr. Ruchi Choudhary, and André Paul Neto-Bradley	☐	☐

__________________________ _______________________ _________

Name of participant [printed] Signature Date

________________________ __________________ ________

Researcher [printed] Signature Date

1 Household Identification

ENTER 1.1 TO 1.4 BEFORE GOING TO HOUSEHOLD

1.1 Name of state or Union Territory – 2001 Census CODE

____

1.2 District Name – 2001 Census 2001 CODE

____

1.3 Current District Name – 2011 Census EDITOR 2011 CODE

____

1.4 District Name – 2001 Census CODE

____

1.5 What is your religion?

Hindu = 1	Sikh = 4	Tribal = 7
Muslim = 2	Buddhist = 5	Others = 8
Christian = 3	Jain = 6	None = 9

1.6 Which caste do you belong to?

Caste

1.7 Is this General/Forward, OBC, SC, or ST?

General/Forward = 1	SC = 3
OBC = 2	ST = 4

1.8 What is the principal source of income for the household? CODE

Cultivation=01	Organized Trade/Business=07
Allied agriculture=02	Salaried employment=08
Agricultural wage labour=03	Profession NEC=09
Non agricultural wage labour=04	Pension/Rent/Dividend, etc.=10
Artisan/Independent=05	Other (specify)=11
Petty shop/Small business=06	

1.9 How many years ago did your family first come to this city? YEARS

IF MORE THAN 90 (includes 'forever') ENTER 90, AND GO TO QUESTION 1.12

1.10 From where did the family come?

Same state, same district = 1
Same state, another district = 2
Another state = 3
Another country = 4

1.11 Was this a village or a town/city?

Village = 1
Town = 2

1.12 Are you a member of any local clubs or associations?

No = 0
Yes = 1

[1.13 to 1.17 TO BE FILLED IN BY INTERVIEWER]

1.13 Rural/Urban/Peri-Urban (Current)

Rural = 0
Peri-Urban = 1
Urban = 2

1.14 What is the building/house type? CODE

Bungalow, no shared walls=01
House with shared walls=02
Flat=03
Chawl=04
Slum Housing=05
Others=06

1.15 What is the predominant wall type? CODE

Thatch=01
Mud Brick=02
Plastic=03
Wood=04
Kiln-Fired Bricks=05
Metal Sheet=06
Stone=07
Concrete=08
Others=09

1.16 What is the predominant roof type? CODE

Thatch, Mud, Wood=01
Tile=02
Slate=03
Plastic=04
Metal, Asbestos=05
Cement=06
Brick=07
Stone=08
Concrete=09
Others=10

1.17 What is the predominant floor type? CODE

Mud=01	Cement=05
Wood, Bamboo=02	Brick=06
Brick=03	Stone=07
Stone=04	

2 Household Roster

Household = all those who live under the same roof and share the same kitchen for 6+ months during the last year.									
2.1	2.2	2.3	2.4		2.5		2.6	
ID code	*Please tell me the names of all the people who live and take meals in this house* *WRITE IN ENGLISH OR HINDI CAPITAL LETTERS* *NAMES (start with head of household)*	*Sex* *F=1* *M=2*	Relationship to head of household		*AGE* *How old is he/she?*		*Primary activity status*	
01			0	1				
02								
03								
04								
05								
06								
07								
08								
09								
10								
11								
12								

Head=01	Grandchild=05	Nephew/Niece=09
Wife/Husband=02	Father/Mother=06	Other relatives=10
Son/Daughter=03	Brother/Sister=07	Servant/others=11
Daughter-in-law/ Son-in-law=04	Daughter-in-law/ Son-in-law=08	Daughter-in-law/ Son-in-law=12

Cultivation=01	Organised Trade/Business=07	Looking for work/unemployed=13
Allied agriculture=02	Salaried employed=08	
Agricultural wage labour=03	Profession NEC=09	Too young/Unfit to work=14
Non agricultural wage labour=04	Retired=10	
Artisan/Independent work=05	Household work/housewife=11	Others=15
Petty shop/small business=06	Student=12	

3 Occupation and Salary

Now, besides work in the household business/farm, what work did [NAME] DO LAST YEAR for payment in cash or kind?

ASK SEPARATELY FOR EVERYONE. IF A PERSON WORKED AT DIFFERENT TYPES OF JOBS, ENTER EACH ON A SEPARATE LINE.

Household = all those who live under the same roof and share the same kitchen for 6+ months during the last year.

3.1	3.2	3.3	3.4	3.5	3.6
Household Roster ID code	*What kind of work did [NAME] do? DESCRIPTION OF WORK*	*Occupation Code*	Industry Code	*How many hours did [NAME] work in a usual day?*	*Frequency of Payment* Daily=1 Monthly=2 Fixed=3

Now, I would like you to think about your expenditure in the last 30 days...

3.1 What has the total expenditure of the household been?

RUPEES | | , | | , | | |

3.2 Of this, how much has been spent on food, fuel, and water/beverages?

RUPEES | | , | | , | | |

4 Education

Now, I would like to ask you some questions about the education of each member or your family?

DETACH THE HOUSEHOLD ROSTER CARD FROM THE ROSTER, ALIGN WITH THIS PAGE AND FOR EACH INDIVIDUAL ASK.

			ASK ONLY FOR THOSE WHO EVER ATTENDED	ASK ONLY FOR THOSE WHO HAVE STUDIED 10th CLASS & MORE	
4.1	*4.2*	*4.3*	*4.4*	*4.5*	*4.6*
Household Roster ID code	*Can [NAME] read and write a sentence?* *No=0* *Yes=1*	*Has [NAME] ever attended school?* *No=0* *Yes=0*	How many standard years has [NAME] completed? <1 Class=55 5th Class=05 Bachelors=15	*After secondary school what did [NAME] do?* *USE CODES BELOW*	*What is the highest level of education [NAME] has completed?* *USE CODES BELOW*
01					
02					
03					
04					
05					
06					
07					
08					
09					
10					
11					
12					

Codes for 4.5	Codes for 4.6
Attend regular School/College=1	Incomplete/Did not receive degree=1
Studying privately=2	BA/BSc/B.Com/BCA/BBA=2
Distance education/open university=3	B.Tech/BE=3
Technical Degree/Diploma < 3 years=4	Master's Degree/PhD=4
Technical Degree/Diploma > 3 years=5	Professional Degree=5
Dropped Out/Nothing=6	Others=6

5 Appliance Ownership

5.1 What is the ownership status of the house in which you live?

Owned = 1 Office accommodation = 3
Rented = 2 Other = 4

Now I would like to ask you about what things your household owns…

5.2 Pressure Cooker
Yes=1
No=0

5.3 Mixer
Yes=1
No=0

5.4 Grinder
Yes=1
No=0

5.5 Roti Maker
Yes=1
No=0

5.6 Cast Iron Cookware
Yes=1
No=0

5.7 Stainless Steel Cookware
Yes=1
No=0

5.8 Electric Fan (Table)
Yes=1
No=0

5.9 Electric Fan (Ceiling)
Yes=1
No=0

5.10 Air Cooler
Yes=1
No=0

5.11 AC
Yes=1
No=0

5.12 Traditional Chulha
Yes=1
No=0

5.13 Improved Chulha
Yes=1
No=0

5.14 Kerosene Stove
Yes=1
No=0

5.15 LPG Stove
Yes=1
No=0

5.16 Electric Cooker
Yes=1
No=0

5.17 Water Pump
Yes=1
No=0

5.18 Electric Geyser
Yes=1
No=0

5.19 Gas Boiler
Yes=1
No=0

[CONTINUED FROM PREVIOUS PAGE…]

5.20 Solar Water Heater
Yes=1
No=0

5.21 Inverter/Battery Storage
Yes=1
No=0

5.22 Television
Yes=1
No=0

5.23 Mobile Phone
Yes=1
No=0

5.24 Photovoltaic Solar Panel
Yes=1
No=0

5.25 Roti Machine
Yes=1
No=0

5.26 Solar Lamp
Yes=1
No=0

5.27 Kerosene Lamp
Yes=1
No=0

6 *Fuel Use*

6.1 How many rooms are there in your house?

DO NOT COUNT BALCONIES, CORRIDORS, AND BATHROOMS BUT INCLUDE HALL

6.2 Where is the cooking, generally done for this household?

Cooking is Outdoors = 0
Cooking is Indoors = 1
Outhouse = 2

6.3 Is there a chimney/extractor for the stove?

No = 0
Yes = 1
NA / Cooking is outdoor = 2

6.4 Does this house have electricity?

No = 0
Yes = 1

IF YES:

6.4a How many hours per day do you generally have power?

(in a season like this)

6.4b Do you have electricity around the same time(s) each day?

No = 0
Yes = 1

6.4c How do you pay for the electricity you use?

No Bill/Govt. scheme=1
Bill from State Elec. Board/ Company=2
Fee to neighbour=3
Part of rent=4
Own generator=5
Own solar panels=6
Paid by Employer/Officer=7
Pension/Rent/Dividend, etc.=8
Other means=9

6.4d How much do you typically pay for electricity in a 30 day period?

(in rupees) , ,

6.4e What is the typical rate you pay for electricity?

(in rupees/unit-kWh) , ,

6.5 Which month of the year does your household spend the most on fuel?

CODE

January=01	July=07
February=02	August=08
March=03	September=09
April=04	October=10
May=05	November=11
June=06	December=12

6.5a What is the primary reason for increased fuel spending during this period?

CODE

Cold weather=01	Celebration=05
Hot weather=02	Visitors=06
Fuel unavailability=03	Seasonal work=07
Volatile prices=04	Other=08

SPECIFY ___________________________

6.6 Are you aware of any government energy and fuel provision programmes (such as Pradhan Mantri Ujjwala Yojana), or financial support for fuel or appliances available in your area?

No = 0
Yes = 1

6.6a Who/what made you aware of this these programmes?

CODE

Household member=01	Public Official=05
Relative=02	Radio/TV advert=06
Neighbour=03	School Teacher=07
Work colleague=04	Other=08

SPECIFY ___________________________

NOTE ANSWERS ON USE, PROCUREMENT AND PRICE FOR ONE FUEL AT A TIME.

	6.7	6.8	6.9		6.10	
Does Your Household use…	*Household uses fuel?* *No = 0* *Yes = 1*	*Where do you get most of your fuel from?*	How often do you get fuel? *Daily=01* *More than once a week=02* *Weekly=03* *Every two weeks=04* *Monthly=05*		*How far do you have to travel to obtain fuel?* *[MINUTES]*	
Firewood/Twigs?						
Dung Cake?						
Crop residue?						
Kerosene?						
LPG?						
Coal/Charcoal?						

Codes for 6.8

Purchase from Market=1
Collect from own land=2
Collect from village/other place=3
Both (purchased and collected)=4
Ration shop=5
Purchase from Local Vendor=6
Gas Company=7

IF FUELS ARE PURCHASED, ASK FOR EACH FUEL.

	6.11	6.12	6.13
	How much did you pay for what you used in the last 30 days?	*What rate do you pay for the fuel at?* *CAN BE GIVEN AS PRICE PER KILOGRAM, LITRE, OR BUNDLE/UNIT IF BUNDLE UNIT ASK ABOUT BUNDLE APPROX. WEIGHT*	*Unit* *SEE CODE LIST*
	RUPEES	*RUPEES*	
Firewood/Twigs	, ,	, ,	
Dung Cake	, ,	, ,	
Crop residue	, ,	, ,	
Kerosene	, ,	, ,	
LPG	, ,	, ,	
Coal/Charcoal	, ,	, ,	

	6.14
What is the bundle/unit weight for (if applicable)…	
Firewood/Twigs?	
Dung Cake?	
Crop residue?	
Kerosene?	
LPG?	
Coal/Charcoal?	

Codes for 6.13

Per kilogram=1
Per litre=2
Per bundle*=3

*IF PER BUNDLE ASK BUNDLE WEIGHT

7 *Energy Use Habits*

7.1 How many meals including breakfast do you usually take in your household per day?

(meals in a day) [ ]

7.2 What type of chulha does the household mainly use?

Open fire (i.e. three stones) = 1
Traditional Chulha = 2
Improved Chulha = 3
Kerosene = 4
LPG Stove = 5
Electric Stove = 6

7.3 Who does most of the cooking in the household?

HOUSEHOLD ROSTER CODE

7.4 Have you bought or received a new stove in the last 3 years?

No = 0
Yes = 1

IF YES

7.4a What was your main reason for changing stove?

Cost of Fuel = 1
Availability of Fuel = 2
Subsidy = 3
Convenience/Time Saving = 4
Neighbours had one = 5
Health benefits = 6

IF NO

7.4b What was your main reason for not changing stove?

Cost of Fuel = 1
Availability of Fuel = 2
Lack of Subsidy for Equipment = 3
Convenience/Time Saving = 4
Daily routines = 5

7.5 Have you bought or received new lighting equipment in the last 3 years?

No = 0
Yes = 1

IF YES

7.5a What was your main reason for changing lighting equipment?

Cost of Fuel = 1
Availability of Fuel = 2
Subsidy = 3
Convenience/Time Saving = 4
Neighbours had one = 5
Health benefits = 6
Quality of lighting = 7

IF NO

7.5b What was your main reason for not changing lighting equipment?

Cost of Fuel = 1 Convenience/Time Saving = 4
Availability of Fuel = 2 Neighbours had one = 5
Subsidy = 3 Health benefits = 6

7.6 Do you use a solar panel for lighting, or small appliances in your house?

No = 0
Yes = 1

IF YES

7.6a What made you decide to get a solar panel ?

Reliability = 1 Convenience/Time Saving = 4
Cost Saving = 2 Neighbours had one = 5
Subsidy = 3 Government Scheme = 6

7.7 Have you bought or received new water heating and/or pumping equipment in the last 3 years?

No = 0
Yes = 1

IF YES

7.7a What was your main reason for changing equipment?

Cost of Fuel = 1 Convenience/Time Saving = 4
Availability of Fuel = 2 Neighbours had one = 5
Subsidy = 3 Health benefits = 6
Quality of lighting = 7

IF NO

7.7b What was your main reason for not changing equipment?

Cost of Fuel = 1 Convenience/Time Saving = 4
Availability of Fuel = 2 Neighbours had one = 5
Subsidy = 3 Health benefits = 6

7.8 Do you use energy in your household for business activities, e.g. food preparation, sewing, equipment repairs, etc.?

No = 0
Yes = 1

IF YES

7.8a What times of day are these activities typically carried out?

05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00

7.8b What fuel do you primarily use for these activities?

Firewood = 1
Dung cake = 2
Kerosene = 3
LPG = 4
Coal/Charcoal = 5
Electricity = 6

IF NO

7.8c Would you be likely to start a small business such as bread making, or sewing if you had access to a different fuel?

No = 0
Yes = 1

7.9 When you go to bed is the room sometimes too hot for you to sleep?

No = 0
Yes = 1

7.10 Are there times of day when it is too hot in your house for daily activities?

No = 0
Yes = 1

IF YES

7.10a What hours of the day are too hot?

05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00

7.11 What is the main source of drinking water for your house?

Piped (public supply) = 1
Tube well = 2
Hand pump = 3
Open/covered well = 4
Bottled = 4
River, stream, or pond = 5
Tanker truck = 6
Rainwater = 7

NORMALLY SUMMER

7.12 If piped water: How many hours per day of water do you get?

NORMALLY SUMMER

7.13 Is the availability of drinking water adequate?

No = 0
Yes = 1

NORMALLY SUMMER

	7.14		7.15		7.16																	
	What fuel does the household PRIMARILY use for this?		What ADDITIONAL fuel does the household use for this?		What hours of the day do you typically use these equipment/appliances? *CHECK EACH HOUR WITHIN RANGES INDICATED.*																	
					05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00
Cooking																						
If you have recently changed stove what was time of use before?																						
Lighting																						
If you have recently changed lights what was time of use before?																						
Water Heating																						
If you have recently changed water heater what was time of use before?																						

Codes for 7.14 & 7.15

Firewood/Twigs=01
Dung Cake=02
Crop Residue=03
Kerosene=04
LPG=05
Coal/Charcoal=06
Electricity=07
Solar=08

Now I would like to ask you if you were not limited by financial, availability, or other barriers, what fuel would you aspire to use.

	7.17		*7.18*		*7.19*																	
	What fuel would the household PRIMARILY use for this?		What ADDITIONAL fuel would the household use for this? (if applicable)		What hours of the day would you typically use these equipment/appliances? *CHECK EACH HOUR WITHIN RANGES INDICATED IF DIFFERENT FROM CURRENT USE.*																	
					05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00
Cooking																						
Lighting																						
Water Heating																						

C Household Survey Sample Ward Maps

This Appendix contains maps of selected wards for the five cities surveyed through primary data collection. The ward boundaries used are based on 2011 Census ward boundaries. These maps are for reference purposes only, and exact ward boundary definitions may vary. Ward boundaries are regularly updated and redrawn as city populations change, and current ward boundary definitions will vary from those shown below. Where major changes to boundaries had occurred and previous boundary locations were unclear enumerators used best judgement as to whether a household fell within a given ward or not. Wards boundaries are drawn to have similar populations although there is variation in population between wards within a city and variation between mean ward population between different cities.

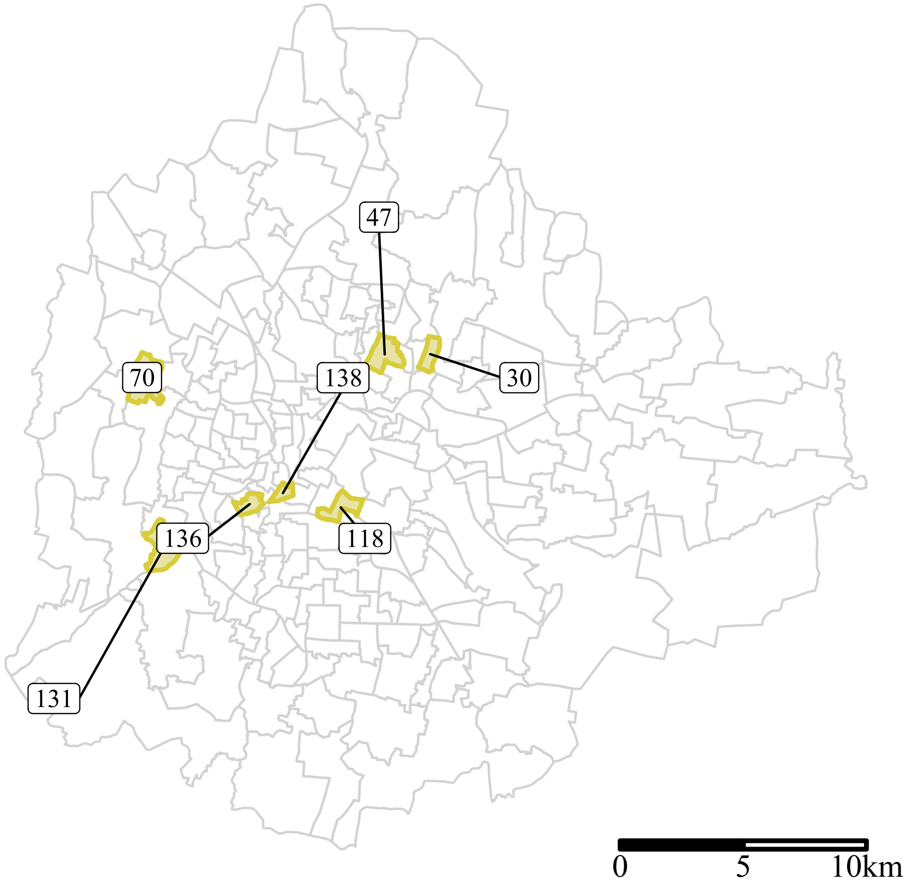


Figure C.1: Wards selected for survey in Bangalore based on 2011 Census Ward Boundaries

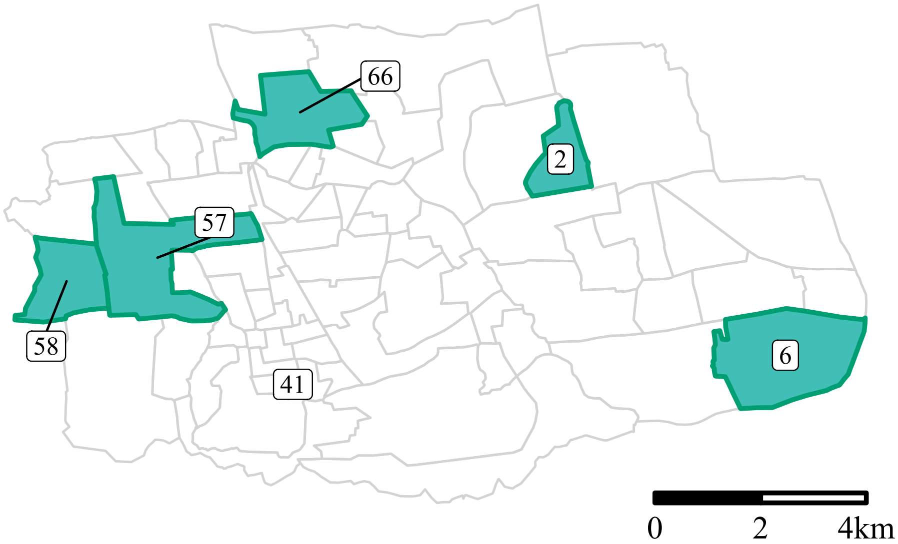


Figure C.2: Wards selected for survey in Coimbatore based on 2011 Census Ward Boundaries

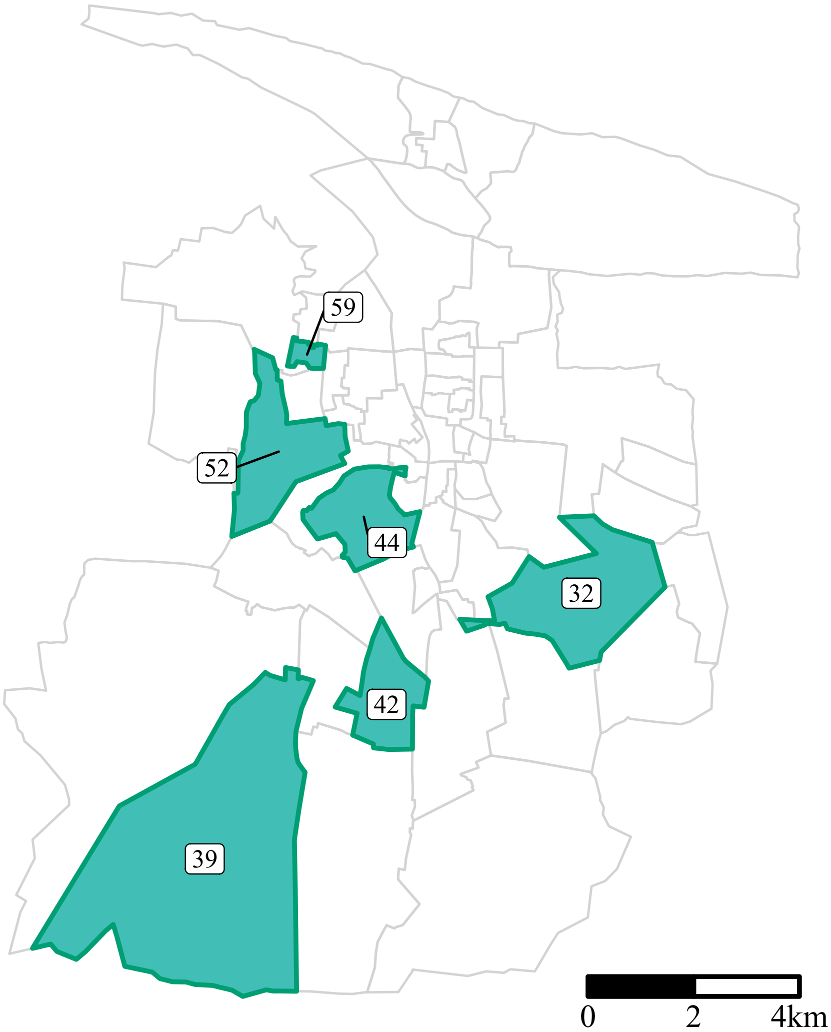


Figure C.3: Wards selected for survey in Trichy based on 2011 Census Ward Boundaries

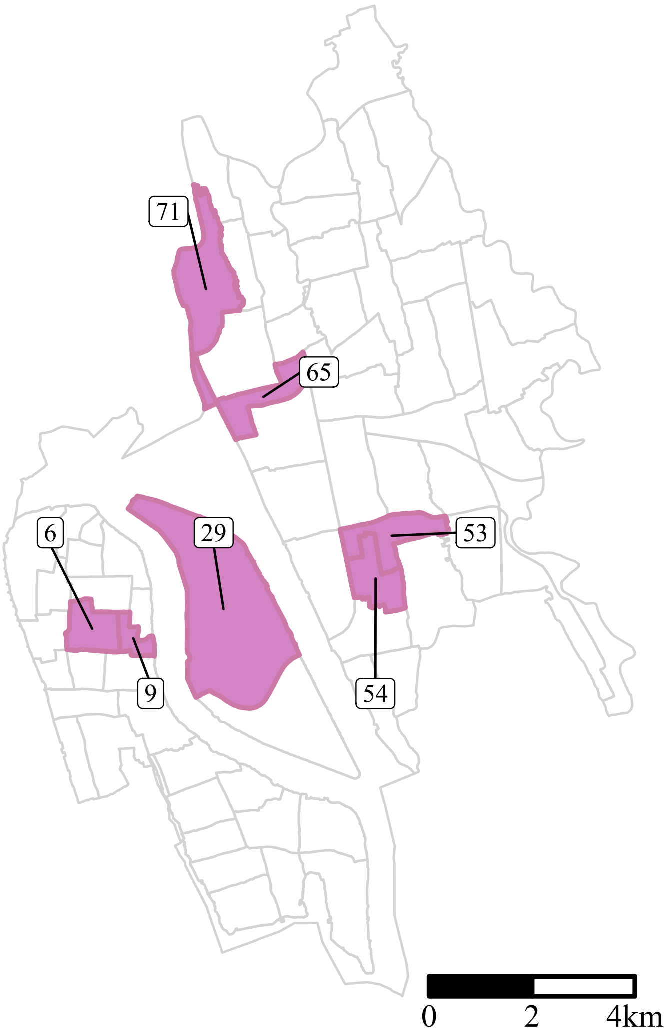


Figure C.4: Wards selected for survey in Kochi based on 2011 Census Ward Boundaries

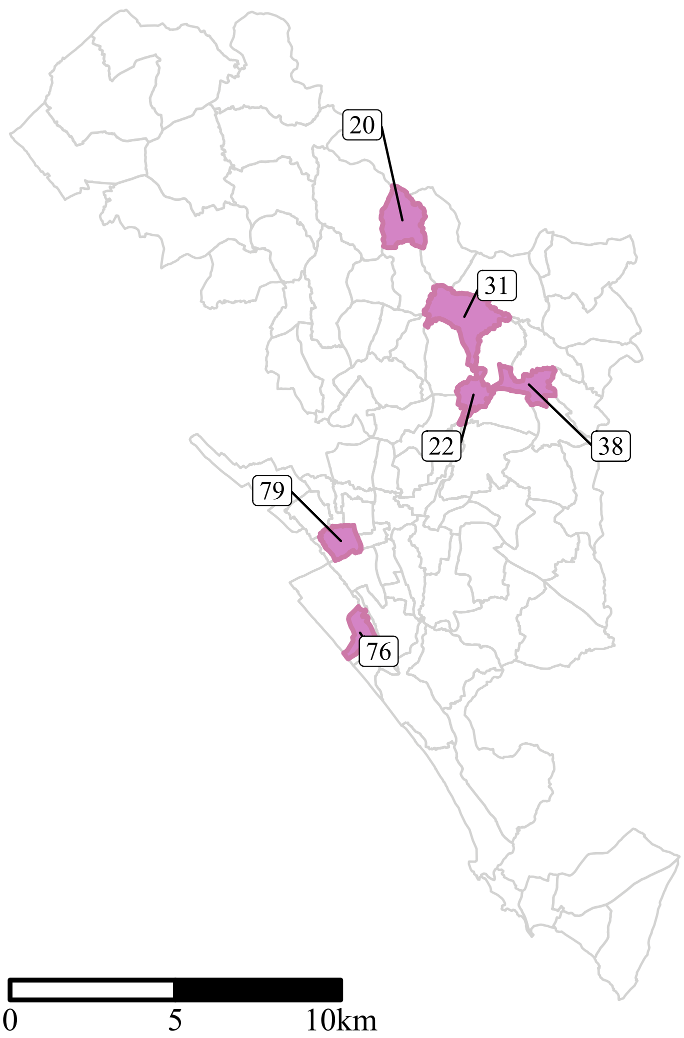


Figure C.5: Wards selected for survey in Trivandrum based on 2011 Census Ward Boundaries

D Stan Model Code

LPG Combined Model

The following is a sample of Stan model code for LPG fuel estimation using the 'Combined Model' which includes random effects and spatial ICAR components. Code has been simplified and commented for greater clarity.

```
// iheus_neapolitan_hier_BYM.stan
// This Stan program defines a model for estimating household LPG use 'y'
// with a multilevel structure using individual level predictors conditional
// on primary cooking fuel group, and ward level random and spatial effects.

// ICAR function for use in model
functions {
        real icar_normal_lpdf(vector phi, int N, int[] node1, int[] node2) {
                return -0.5 * dot_self(phi[node1] - phi[node2]);
        }
}

// Model inputs
data {
        int<lower = 0> N; // No. Households in NSS data
        int<lower = 0> M; // No. Households in Synthetic Pop
        int<lower = 2> K; // No. Predictors for Primary Cooking Fuel Group
        int<lower = 0> J; // No. Latent Primary Cooking Fuel Groups
        int<lower = 0> W; // No. of Wards in City
        int<lower = 1, upper = J> cook_choice[N]; // Primary Cooking Fuel NSS data
        int<lower = 1, upper = J> cook_choice_test[M]; // Synthetic Pop Primary Cooking Fuel Dummy
```

```
    int<lower = 1, upper = W> ward[M]; // Synthetic Pop Wards
    int<lower=0> nodes; // No. Ward Graph Nodes
    int<lower=0> N_edges; // No. Ward Graph Edges
    int<lower=1, upper=nodes> node1[N_edges];  // node1[i] adjacent to node2[i]
    int<lower=1, upper=nodes> node2[N_edges];  // and node1[i] < node2[i]
    vector[N] y; // Vector of fuel use NSS (LPG)
    vector[M] y_p; // Vector of prior fuel use (LPG)
    vector[N] HH_size; // No. Persons in Household (NSS)
    vector[N] HH_exp; // Household Expenditure (URP) (NSS)
    matrix[N,K] x; // Matrix of socio-economic inputs (NSS)
    vector[M] HH_size_test; // No. Persons in Household (Synthetic Pop)
    vector[M] HH_exp_test; // Household Expenditure (URP) (Synthetic Pop)
    real sigma_y_p; // Prior standard deviation for fuel (LPG Synth. Pop)
    matrix[M, K] x_test; // Matrix of socio-economic inputs (Synthetic Pop)
}

parameters {
    // Paramters for std normal reparamaterisation
    matrix[K,J] beta;
    vector[J] yay;
    vector[J] yee_1;
    vector[J] yee_2;
    vector[W] gaa;
    real mu_a; // Individual Intercept Parameter
    real<lower = 0> sigma_a;
    real mu_b1; // Individual HH Size Slope Paramters
    real<lower = 0> sigma_b1;
    real mu_b2; // Individual HH Exp Slope Parameters
    real<lower = 0> sigma_b2;
    real mu_beta_raw; // Cat Logit Coeff Parameters
    real<lower = 0> sigma_beta_raw;
    real mu_g; // Random effects parameters
    real<lower = 0> sigma_g;
    real<lower = 0> sigma; // Overall precision parameter
    vector[M] zee;
    vector[nodes] phi_l; // Spatial effects coefficient
```

```
}

transformed parameters{
        // Prior Individual coefficients
        vector[M] peta_1;
        vector[M] peta_2;
        vector[M] pay;

        // Actual Individual coefficients
        vector[J] a;
        vector[J] b1;
        vector[J] b2;

        // Random effects coefficient
        vector[W] g;

        real mu_beta; // Slope 2
        real<lower = 0> sigma_beta;

        matrix[N, J] x_beta = x * beta;

        // Reparameterisation for individual coefficients
        a = mu_a + sigma_a*yay;
        b1 = mu_b1 + sigma_b1*yee_1;
        b2 = mu_b2 + sigma_b2*yee_2;

        // Reparameterisation for random effects coefficients
        g = mu_g + sigma_g*gaa;

        // Synthetic population prior paramterisation
        for(p in 1:M){
                peta_1[p]=mu_b1+sigma_b1*zee[p];
                peta_2[p]=mu_b2+sigma_b2*zee[p];
                pay[p]=mu_a+sigma_a*zee[p];
        }

        mu_beta = 1*mu_beta_raw;
        sigma_beta = 1*sigma_beta_raw;
}
```

```
model {
        // Cooking fuel logit priors
        to_vector(beta) ~ normal(mu_beta, sigma_beta);
        mu_beta_raw ~ std_normal();
        sigma_beta_raw ~ std_normal();

        // Hierarchical priors
        zee~std_normal();
        gaa ~ std_normal();

        // Fuel use prior
        for(p in 1:M){
                y_p[p] ~ normal(pay[p] + peta_1[p].*HH_size_test[p] + peta_2
                    [p].*HH_exp_test[p] + phi_l[ward[p]] + g[ward[p]],
                    sigma_y_p);
        }

        // ICAR component prior and specification
        phi_l ~ icar_normal_lpdf(nodes, node1, node2);
        // soft sum-to-zero constraint on phi
        // more efficient than mean(phi) ~ normal(0, 0.001)
        sum(phi_l) ~ normal(0, 0.001 * N);

        // Standard normal priors for reparamterisation
        yay~std_normal();
        yee_1~std_normal();
        yee_2~std_normal();

        // Standard normal prior for fuel model variance sigma
        sigma ~ std_normal();

        y ~ normal(a[cook_choice] + b1[cook_choice].*HH_size + b2[
            cook_choice].*HH_exp, sigma);

        for(i in 1:N){
                cook_choice[i] ~ categorical_logit((x_beta[i])');
        }
}

generated quantities{
        vector[M] y_rhat_test;
```

```
        int cook_rhat_test[M];

        // Synthetic Population Output Estimates
        for(m in 1:M){
                cook_rhat_test[m] = categorical_logit_rng((x_test[m]*beta)')
                    ; // Posterior predictive
                y_rhat_test[m] = normal_rng(a[cook_rhat_test[m]] + b1[
                    cook_rhat_test[m]]*HH_size_test[m] + b2[cook_rhat_test[m
                    ]]*HH_exp_test[m] + phi_l[ward[m]] + g[ward[m]], sigma);
        }
}
```

Bibliography

Abbas, K., Li, S., Xu, D., Baz, K., and Rakhmetova, A. (2020). "Do socioeconomic factors determine household multidimensional energy poverty? Empirical evidence from South Asia". en. *Energy Policy* 146, p. 111754. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111754.

Aggarwal, S., Kumar, S., and Tiwari, M. K. (2018). "Decision support system for Pradhan Mantri Ujjwala Yojana". *Energy Policy* 118, pp. 455–461. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2018.04.011.

Ahmad, S. and Puppim de Oliveira, J. A. (2015). "Fuel switching in slum and non-slum households in urban India". *Journal of Cleaner Production* 94, pp. 130–136. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2015.01.072.

Alem, Y., Beyene, A. D., Köhlin, G., and Mekonnen, A. (2016). "Modeling household cooking fuel choice: A panel multinomial logit approach". en. *Energy Economics* 59, pp. 129–137. ISSN: 0140-9883. DOI: 10.1016/j.eneco.2016.06.025.

Ali, A. M. (2020). "Household Energy Use Among Female-Headed Households in Urban Ethiopia: Key Issues for the Uplift of Women". en. *Indian Journal of Human Development* 14:3. Publisher: SAGE Publications India, pp. 460–480. ISSN: 0973-7030. DOI: 10.1177/0973703020967897.

Amrhein, V., Greenland, S., and McShane, B. (2019). "Scientists rise up against statistical significance". EN. *Nature* 567:7748, p. 305. DOI: 10.1038/d41586-019-00857-9.

Anderson, D. R. (R. (2009). *Essentials of statistics for business and economics*. Ed. by Dennis J Sweeney and Thomas A. (Thomas Arthur) Williams. 5th ed. Statistics for business and economics. Mason, OH : Thomson South-Western, c2009., Mason, OH.

Arcaya, M. C., Schwartz, G., and Subramanian, S. (2018). "A multi-level modeling approach to understanding residential segregation in the United States". en. *Environment and Planning B: Urban Analytics and City Science* 45:6. Publisher: SAGE Publications Ltd STM, pp. 1090–1105. ISSN: 2399-8083. DOI: 10.1177/2399808318760858.

Arku, R. E., Birch, A., Shupler, M., Yusuf, S., Hystad, P., and Brauer, M. (2018). "Characterizing exposure to household air pollution within the Prospective Urban Rural Epidemiology (PURE) Study". *Environment international* 114, pp. 307–317. ISSN: 0160-4120. DOI: 10.1016/j.envint.2018.02.033.

Assad, N. A., Kapoor, V., and Sood, A. (2016). "Biomass smoke exposure and chronic lung disease". eng. *Current Opinion in Pulmonary Medicine* 22:2, pp. 150–157. ISSN: 1531-6971. DOI: 10.1097/MCP.0000000000000246.

Bai, X. et al. (2018). "Six research priorities for cities and climate change". en. *Nature* 555:7694, pp. 23–25. DOI: 10.1038/d41586-018-02409-z.

Baiyegunhi, L. J. S. and Hassan, M. B. (2014). "Rural household fuel energy transition: Evidence from Giwa LGA Kaduna State, Nigeria". *Energy for Sustainable Development* 20:Supplement C, pp. 30–35. ISSN: 0973-0826. DOI: 10.1016/j.esd.2014.02.003.

Ballas, D. and Clarke, G. P. (2001). "Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach". en. *Environment and Planning C: Government and Policy* 19:4. Publisher: SAGE Publications Ltd STM, pp. 587–606. ISSN: 0263-774X. DOI: 10.1068/c0003.

Bansal, M., Saini, R. P., and Khatod, D. K. (2013). "Development of cooking sector in rural areas in India—A review". *Renewable and Sustainable Energy Reviews* 17, pp. 44–53. ISSN: 1364-0321. DOI: 10.1016/j.rser.2012.09.014.

Bansal, N., Blum, A., and Chawla, S. (2004). "Correlation Clustering". en. *Machine Learning* 56:1, pp. 89–113. ISSN: 1573-0565. DOI: 10.1023/B:MACH.0000033116.57574.95.

Barrie, J. and Cruickshank, H. J. (2017). "Shedding light on the last mile: A study on the diffusion of Pay As You Go Solar Home Systems in Central East Africa". en. *Energy Policy* 107, pp. 425–436. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2017.05.016.

Barthelemy, J. and Toint, P. L. (2013). "Synthetic Population Generation Without a Sample". *Transportation Science* 47:2. Publisher: INFORMS, pp. 266–279. ISSN: 0041-1655.

Batty, M. (2019). "Urban analytics defined". en. *Environment and Planning B: Urban Analytics and City Science* 46:3. Publisher: SAGE Publications Ltd STM, pp. 403–405. ISSN: 2399-8083. DOI: 10.1177/2399808319839494.

Baumgärtner, S., Becker, C., Frank, K., Müller, B., and Quaas, M. (2008). "Relating the philosophy and practice of ecological economics: The role of concepts, models, and case studies in inter- and transdisciplinary sustainability research". en. *Ecological Economics* 67:3, pp. 384–393. ISSN: 0921-8009. DOI: 10.1016/j.ecolecon.2008.07.018.

Besag, J. (1974). “Spatial Interaction and the Statistical Analysis of Lattice Systems”. en. *Journal of the Royal Statistical Society: Series B (Methodological)* 36:2, pp. 192–225. ISSN: 2517-6161. DOI: https://doi.org/10.1111/j.2517-6161.1974.tb00999.x.

Besag, J., York, J., and Mollié, A. (1991). “Bayesian image restoration, with two applications in spatial statistics”. en. *Annals of the Institute of Statistical Mathematics* 43:1, pp. 1–20. ISSN: 1572-9052. DOI: 10.1007/BF00116466.

Bettencourt, L. and West, G. (2010). “A unified theory of urban living”. en. *Nature* 467:7318, pp. 912–913. ISSN: 0028-0836. DOI: 10.1038/467912a.

Bhan, G. and Jana, A. (2015). “Reading Spatial Inequality in Urban India”. English. *Economic and Political Weekly* I:22, pp. 49–54.

Bhatt, B. P., Rathore, S. S., Lemtur, M., and Sarkar, B. (2016). “Fuelwood energy pattern and biomass resources in Eastern Himalaya”. en. *Renewable Energy* 94, pp. 410–417. ISSN: 0960-1481. DOI: 10.1016/j.renene.2016.03.042.

Bibri, S. E. and Krogstie, J. (2017). “Smart sustainable cities of the future: An extensive interdisciplinary literature review”. en. *Sustainable Cities and Society* 31, pp. 183–212. ISSN: 2210-6707. DOI: 10.1016/j.scs.2017.02.016.

Bickart, B. and Felcher, E. (1996). “Expanding and enhancing the use of verbal protocols in survey research”. In: *Answering Questions*. Jossey-Bass, San Francisco.

Bisaga, I. and Parikh, P. (2018). “To climb or not to climb? Investigating energy use behaviour among Solar Home System adopters through energy ladder and social practice lens”. *Energy Research & Social Science* 44, pp. 293–303. ISSN: 2214-6296. DOI: 10.1016/j.erss.2018.05.019.

Bischoping, K. (1989). "An evaluation of interviewer debriefing in survey pretests". In: *New Techniques for Pretesting Survey Questions*. Survey Research Center, Ann Arbor, MI.

Blocker, A. W. (2016). *ipfp: Fast Implementation of the Iterative Proportional Fitting Procedure in C*.

Bocken, N. M. P., Fil, A., and Prabhu, J. (2016). "Scaling up social businesses in developing markets". en. *Journal of Cleaner Production* 139, pp. 295–308. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2016.08.045.

Booth, A. T., Choudhary, R., and Spiegelhalter, D. J. (2012). "Handling uncertainty in housing stock models". en. *Building and Environment* 48, pp. 35–47. ISSN: 0360-1323. DOI: 10.1016/j.buildenv.2011.08.016.

Brooks, N., Bhojvaid, V., Jeuland, M. A., Lewis, J. J., Patange, O., and Pattanayak, S. K. (2016). "How much do alternative cookstoves reduce biomass fuel use? Evidence from North India". *Resource and Energy Economics* 43, pp. 153–171. ISSN: 0928-7655. DOI: 10.1016/j.reseneeco.2015.12.001.

Brown, L. S. and Lankford, W. F. (2015). "Sustainability: Clean cooking empowers women". en. *Nature News* 521:7552. Section: Comment, p. 284. DOI: 10.1038/521284a.

Burnett, Richard T. et al. (2014). "An Integrated Risk Function for Estimating the Global Burden of Disease Attributable to Ambient Fine Particulate Matter Exposure". *Environmental Health Perspectives* 122:4. Publisher: Environmental Health Perspectives, pp. 397–403. DOI: 10.1289/ehp.1307049.

Cabeza, L. F., Urge-Vorsatz, D., McNeil, M. A., Barreneche, C., and Serrano, S. (2014). "Investigating greenhouse challenge from growing trends of electricity consumption through home appliances in buildings". *Renewable and Sustainable Energy Reviews* 36, pp. 188–193. ISSN: 1364-0321. DOI: 10.1016/j.rser.2014.04.053.

Caliński, T. and Harabasz, J. (1974). "A dendrite method for cluster analysis". *Communications in Statistics* 3:1, pp. 1–27. ISSN: 0090-3272. DOI: 10.1080/03610927408827101.

Campbell, B. M., Vermeulen, S. J., Mangono, J. J., and Mabugu, R. (2003). "The energy transition in action: urban domestic fuel choices in a changing Zimbabwe". *Energy Policy* 31:6, pp. 553–562. ISSN: 0301-4215. DOI: 10.1016/S0301-4215(02)00098-8.

Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). "Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement". en. *Sociological Methods & Research*. DOI: 10.1177/0049124113500475.

Carmo, C. M. R. do and Christensen, T. H. (2016). "Cluster analysis of residential heat load profiles and the role of technical and household characteristics". *Energy and Buildings* 125, pp. 171–180. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2016.04.079.

Casati, D., Mueller, K., Fourie, P. J., Erath, A., and Axhausen, K. W. (2015). "Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking". English. *Transportation Research Record* 2493. Place: Washington Publisher: Natl Acad Sciences WOS:000371378000013, pp. 107–116. ISSN: 0361-1981. DOI: 10.3141/2493-12.

Castán Broto, V., Arthur, M. d. F. S. R., and Guibrunet, L. (2020). "Energy profiles among urban elite households in Mozambique: Explaining the persistence of charcoal in urban areas". en. *Energy Research & Social Science* 65, p. 101478. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101478.

Çengel, Y. A. and Boles, M. A. (2011). *Thermodynamics : an engineering approach / Yunus A. Çengel, Michael A. Boles ; adapted by Mehmet Kanoğlu*. Ed. by

Mehmet Kanoglu. Seventh edition in SI units. McGraw-Hill, Singapore. ISBN: 9780071311113 paperback.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set". *Journal of Statistical Software* 61:6, pp. 1–36.

Chen, L., Heerink, N., and Berg, M. van den (2006). "Energy consumption in rural China: A household model for three villages in Jiangxi Province". en. *Ecological Economics* 58:2, pp. 407–420. ISSN: 0921-8009. DOI: 10.1016/j.ecolecon.2005.07.018.

Cheng, C.-y. and Urpelainen, J. (2014). "Fuel stacking in India: Changes in the cooking and lighting mix, 1987–2010". *Energy* 76, pp. 306–317. ISSN: 0360-5442. DOI: 10.1016/j.energy.2014.08.023.

Chindarkar, N., Jain, A., and Mani, S. (2021). "Examining the willingness-to-pay for exclusive use of LPG for cooking among rural households in India". en. *Energy Policy* 150, p. 112107. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.112107.

Choudhary, R. and Tian, W. (2014). "Influence of district features on energy consumption in non-domestic buildings". *Building Research & Information* 42:1. Publisher: Routledge _eprint: https://doi.org/10.1080/09613218.2014.832559, pp. 32–46. ISSN: 0961-3218. DOI: 10.1080/09613218.2014.832559.

Choumert-Nkolo, J., Combes Motel, P., and Le Roux, L. (2019). "Stacking up the ladder: A panel data analysis of Tanzanian household energy choices". en. *World Development* 115, pp. 222–235. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2018.11.016.

Chunekar, A. and Sreenivas, A. (2019). "Towards an understanding of residential electricity consumption in India". *Building Research & Information* 47:1, pp. 75–90. ISSN: 0961-3218. DOI: 10.1080/09613218.2018.1489476.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). “Finding community structure in very large networks”. *Physical Review E* 70:6. arXiv: cond-mat/0408187, p. 066111. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111.

Collins, D., Morduch, J., Rutherford, S., and Ruthven, O. (2009). *Portfolios of the Poor*. Princeton University Press. ISBN: 978-0-691-14819-9.

Contreras, P. and Murtagh, F. (2015). “Hierarchical clustering”. English. In: *Handbook of Cluster Analysis*, pp. 103–124. ISBN: 978-1-4665-5189-3. DOI: 10.1201/b19706.

Corbin, J. M. and Strauss, A. L. (2014). *Basics of qualitative research : techniques and procedures for developing grounded theory / Juliet Corbin, International Institute for Qualitative Methodology, Anselm Strauss.* eng. Fourth edition. Thousand Oaks : SAGE Publications, Inc. ISBN: 978-1-4833-1568-3.

Csárdi, G. and Nepusz, T. (2006). “The igraph software package for complex network research”. *InterJournal Complex Systems*, p. 1695.

D’Urso, P. (2015). “Fuzzy clustering”. In: *Handbook of Cluster Analysis*, pp. 545–574. DOI: 10.1201/b19706.

D’Urso, P. and Massari, R. (2019). “Fuzzy clustering of mixed data”. en. *Information Sciences* 505, pp. 513–534. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.07.100.

Dabadge, A., Sreenivas, A., and Josey, A. (2018). “What Has the Pradhan Mantri Ujjwala Yojana Achieved So Far?” en. *Economic and Political Weekly* 53:20, pp. 7–8.

Daioglou, V., Ruijven, B. J. van, and Vuuren, D. P. van (2012). “Model projections for household energy use in developing countries”. *Energy*. 7th Biennial International Workshop “Advances in Energy Studies” 37:1, pp. 601–615. ISSN: 0360-5442. DOI: 10.1016/j.energy.2011.10.044.

Davis, M. (1998). "Rural household energy consumption: The effects of access to electricity—evidence from South Africa". *Energy Policy* 26:3, pp. 207–217. ISSN: 0301-4215. DOI: 10.1016/S0301-4215(97)00100-6.

Debnath, R., Bardhan, R., and Sunikka-Blank, M. (2019). "How does slum rehabilitation influence appliance ownership? A structural model of non-income drivers". *Energy Policy* 132, pp. 418–428. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2019.06.005.

DeMaio, T. and Rothgeb, J. (1996). "Cognitive interviewing techniques: in the lab and in the field". In: *Answering Questions*. Jossey-Bass, San Francisco, pp. 177–96.

Deming, W. E. and Stephan, F. F. (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". *The Annals of Mathematical Statistics* 11:4. Publisher: Institute of Mathematical Statistics, pp. 427–444. ISSN: 0003-4851.

Desai, S. and Vanneman, R. (2015). *India Human Development Survey-II (IHDS-II), 2011-12: Version 6*. eng. type: dataset. DOI: 10.3886/ICPSR36151.V6.

Desai, S., Vanneman, R., and National Council Of Applied Economic Research, N. D. (2010). *India Human Development Survey (IHDS), 2005: Version 12*. eng. type: dataset. DOI: 10.3886/ICPSR22626.V12.

Dhanaraj, S., Mahambare, V., and Munjal, P. (2018). "From Income to Household Welfare: Lessons from Refrigerator Ownership in India". en. *Journal of Quantitative Economics* 16:2, pp. 573–588. ISSN: 2364-1045. DOI: 10.1007/s40953-017-0084-5.

DiMaggio, C. (2015). "Small-Area Spatiotemporal Analysis of Pedestrian and Bicyclist Injuries in New York City:" en. *Epidemiology* 26:2, pp. 247–254. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000000222.

Ekholm, T., Krey, V., Pachauri, S., and Riahi, K. (2010). "Determinants of household energy consumption in India". *Energy Policy*. The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers 38:10, pp. 5696–5707. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2010.05.017.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). "A working guide to boosted regression trees". en. *Journal of Animal Ecology*, p. 22.

Elizabeth Shove (2010). "Beyond the ABC: Climate Change Policy and Theories of Social Change". en. *Environment and Planning A: Economy and Space* 42:6, pp. 1273–1285. ISSN: 0308-518X. DOI: 10.1068/a42282.

Eludoyin, E. O. and Lemaire, X. (2021). "Work, food, rent, television: The role of lifestyles and experiences on household energy behaviour in rural Lagos, Nigeria". en. *Energy Research & Social Science* 71, p. 101820. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101820.

Fader, M., Cranmer, C., Lawford, R., and Engel-Cox, J. (2018). "Toward an Understanding of Synergies and Trade-Offs Between Water, Energy, and Food SDG Targets". English. *Frontiers in Environmental Science* 0. Publisher: Frontiers. ISSN: 2296-665X. DOI: 10.3389/fenvs.2018.00112.

Farrar, D. E. and Glauber, R. R. (1967). "Multicollinearity in Regression Analysis: The Problem Revisited". *The Review of Economics and Statistics* 49:1. Publisher: The MIT Press, pp. 92–107. ISSN: 0034-6535. DOI: 10.2307/1937887.

Farsi, M., Filippini, M., and Pachauri, S. (2007). "Fuel choices in urban Indian households". en. *Environment and Development Economics* 12:6, pp. 757–774. ISSN: 1469-4395, 1355-770X. DOI: 10.1017/S1355770X07003932.

Fienberg, S. E. (1970). "An Iterative Procedure for Estimation in Contingency Tables". *The Annals of Mathematical Statistics* 41:3. Publisher: Institute of Mathematical Statistics, pp. 907–917. ISSN: 0003-4851.

Filippini, M. and Pachauri, S. (2004). "Elasticities of electricity demand in urban Indian households". *Energy Policy* 32:3, pp. 429–436. ISSN: 0301-4215. DOI: 10.1016/S0301-4215(02)00314-2.

Fink, A. (2016). *How to conduct surveys : a step-by-step guide*. Ed. by Jacqueline B Kosecoff. SAGE Publications Ltd, Newbury Park, Ca. ISBN: 978-1-4833-7848-0.

Foss, A., Markatou, M., Ray, B., and Heching, A. (2016). "A semiparametric method for clustering mixed data". en. *Machine Learning* 105:3, pp. 419–458. ISSN: 1573-0565. DOI: 10.1007/s10994-016-5575-7.

Foss, A. H., Markatou, M., and Ray, B. (2019). "Distance Metrics and Clustering Methods for Mixed-type Data". en. *International Statistical Review* 87:1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12274, pp. 80–109. ISSN: 1751-5823. DOI: https://doi.org/10.1111/insr.12274.

Fowler, F. and Cannell, C. (1996). "Using behavioural coding to identify cognitive problems with survey questions". In: *Answering Questions*. Jossey-Bass, San Francisco.

Frayssinet, L., Merlier, L., Kuznik, F., Hubert, J.-L., Milliez, M., and Roux, J.-J. (2018). "Modeling the heating and cooling energy demand of urban buildings at city scale". en. *Renewable and Sustainable Energy Reviews* 81, pp. 2318–2327. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.06.040.

Fugard, A. J. B. and Potts, H. W. W. (2015). "Supporting thinking on sample sizes for thematic analyses: a quantitative tool". *International Journal of Social Research Methodology* 18:6, pp. 669–684. ISSN: 1364-5579. DOI: 10.1080/13645579.2015.1005453.

Fujita, A., Takahashi, D. Y., and Patriota, A. G. (2014). "A non-parametric method to estimate the number of clusters". *Computational Statistics & Data*

Analysis 73, pp. 27–39. ISSN: 0167-9473. DOI: 10.1016/j.csda.2013.11.012.

Furszyfer Del Rio, D. D., Lambe, F., Roe, J., Matin, N., Makuch, K. E., and Osborne, M. (2020). "Do we need better behaved cooks? Reviewing behavioural change strategies for improving the sustainability and effectiveness of cookstove programs". en. *Energy Research & Social Science* 70, p. 101788. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101788.

Fuso Nerini, F. et al. (2018). "Mapping synergies and trade-offs between energy and the Sustainable Development Goals". en. *Nature Energy* 3:1. Number: 1 Publisher: Nature Publishing Group, pp. 10–15. ISSN: 2058-7546. DOI: 10.1038/s41560-017-0036-5.

Galili, T. (2015). "dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering". *Bioinformatics*. DOI: 10.1093/bioinformatics/btv428.

Galvin, R. (2015). "How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge?" *Journal of Building Engineering* 1, pp. 2–12. ISSN: 2352-7102. DOI: 10.1016/j.jobe.2014.12.001.

Geels, F. W., Berkhout, F., and Vuuren, D. P. van (2016). "Bridging analytical approaches for low-carbon transitions". en. *Nature Climate Change* 6:6. Number: 6 Publisher: Nature Publishing Group, pp. 576–583. ISSN: 1758-6798. DOI: 10.1038/nclimate2980.

Gelman, A. (2006). "Multilevel (Hierarchical) Modeling: What It Can and Cannot Do". *Technometrics* 48:3, pp. 432–435. ISSN: 0040-1706. DOI: 10.1198/004017005000000661.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models / Andrew Gelman, Jennifer Hill*. Ed. by Jennifer Hill. Data analysis using regression and multilevel models. Cambridge, Cambridge.

Ghosh, B. and De, P. (1998). "Role of Infrastructure in Regional Development: A Study over the Plan Period". *Economic and Political Weekly* 33:47/48, pp. 3039–3048. ISSN: 0012-9976.

Giri, A. and Aadil, A. (2018). *Pradhan Mantri Ujjwala Yojana: A demand-side diagnostic study of LPG refills*. en. Technical report Policy Brief. MicroSave, p. 24.

Glaser, B. G. and Strauss, A. L. (2017). *Discovery of Grounded Theory : Strategies for Qualitative Research*. en. eBook. Routledge. ISBN: 978-0-203-79320-6. DOI: 10.4324/9780203793206.

Glaser, B.G and Strauss, A.L (1968). *The discovery of grounded theory: strategies for qualitative research*. Ed. by Strauss, A.L. London : Weidenfeld and Nicolson, 1968., London.

Goswami, P. R. (1989). "The Census of India: A discussion of issues of data collection and dissemination". en. *Government Publications Review* 16:5, pp. 429–438. ISSN: 0277-9390. DOI: 10.1016/0277-9390(89)90072-1.

Gould, C. F., Hou, X., Richmond, J., Sharma, A., and Urpelainen, J. (2020). "Jointly modeling the adoption and use of clean cooking fuels in rural India". en. *Environmental Research Communications* 2:8. Publisher: IOP Publishing, p. 085004. ISSN: 2515-7620. DOI: 10.1088/2515-7620/abaca9.

Gould, C. F., Schlesinger, S., Toasa, A. O., Thurber, M., Waters, W. F., Graham, J. P., and Jack, D. W. (2018). "Government policy, clean fuel access, and persistent fuel stacking in Ecuador". en. *Energy for Sustainable Development*. Scaling Up Clean Fuel Cooking Programs 46, pp. 111–122. ISSN: 0973-0826. DOI: 10.1016/j.esd.2018.05.009.

Gould, C. F. and Urpelainen, J. (2018). “LPG as a clean cooking fuel: Adoption, use, and impact in rural India”. en. *Energy Policy* 122, pp. 395–408. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2018.07.042.

Gould, C. F., Urpelainen, J., and Hopkins SAIS, J. (2020). “The role of education and attitudes in cooking fuel choice: Evidence from two states in India”. en. *Energy for Sustainable Development* 54, pp. 36–50. ISSN: 0973-0826. DOI: 10.1016/j.esd.2019.09.003.

Government of India (2011). *Census of India*.

Government of India Planning Commission (2012). *Annual Report 2011-2012*. Annual Report. New Delhi: Government of India.

Gower, J. C. (1966). “Some distance properties of latent root and vector methods used in multivariate analysis”. en. *Biometrika* 53:3-4. Publisher: Oxford Academic, pp. 325–338. ISSN: 0006-3444. DOI: 10.1093/biomet/53.3-4.325.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. B. M. (2019). *gbm: Generalized Boosted Regression Models*.

Guest, G., Bunce, A., and Johnson, L. (2016). “How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability”. en. *Field Methods*. DOI: 10.1177/1525822X05279903.

Gupta, G. and Köhlin, G. (2006). “Preferences for domestic fuel: Analysis with socio-economic factors and rankings in Kolkata, India”. en. *Ecological Economics* 57:1, pp. 107–121. ISSN: 0921-8009. DOI: 10.1016/j.ecolecon.2005.03.010.

Gupta, S., Gupta, E., and Sarangi, G. K. (2020). “Household Energy Poverty Index for India: An analysis of inter-state differences”. en. *Energy Policy* 144, p. 111592. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111592.

Heltberg, R. (2004). "Fuel switching: evidence from eight developing countries". *Energy Economics* 26:5, pp. 869–887. ISSN: 0140-9883. DOI: 10.1016/j.eneco.2004.04.018.

Hennig, C. (2015). "What are the true clusters?" en. *Pattern Recognition Letters*. Philosophical Aspects of Pattern Recognition 64, pp. 53–62. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.04.009.

– **(2018)**. *fpc: Flexible Procedures for Clustering*.

Hennig, C. and Liao, T. F. (2013). "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification". en. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62:3, pp. 309–369. ISSN: 1467-9876. DOI: https://doi.org/10.1111/j.1467-9876.2012.01066.x.

Hiemstra-van der Horst, G. and Hovorka, A. J. (2008). "Reassessing the "energy ladder": Household energy use in Maun, Botswana". *Energy Policy* 36:9, pp. 3333–3344. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.05.006.

Hijmans, R. J., Phillips, S., Leathwick, J., and Elith, J. (2017). *dismo: Species Distribution Modeling*.

Hillier, B. (1984). *The social logic of space / Bill Hillier, Julienne Hanson.* Ed. by Julienne Hanson. Cambridge : Cambridge University Press, 1984., Cambridge.

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". en. *Journal of Machine Learning Research* 15, pp. 1593–1623.

Hosier, R. H. and Kipondya, W. (1993). "Urban household energy use in Tanzania: Prices, substitutes and poverty". *Energy Policy*. Urban energy and environment in Africa 21:5, pp. 454–473. ISSN: 0301-4215. DOI: 10.1016/0301-4215(93)90035-E.

Hosier, R. H. and Dowd, J. (1987). "Household fuel choice in Zimbabwe: An empirical test of the energy ladder hypothesis". *Resources and Energy* 9:4, pp. 347–361. ISSN: 0165-0572. DOI: 10.1016/0165-0572(87)90003-X.

Huang, R. (2014). "RQDA: R-based qualitative data analysis". *R package version 0.2–7*.

International Energy Agency (2019). *World Energy Outlook 2019*. en-GB. Technical report. Paris: International Energy Agency.

– **(2020a)**. *Access to clean cooking – SDG7: Data and Projections – Analysis*. en-GB.

– **(2020b)**. *India 2020 - Energy Policy Review*. en. Policy Review. Paris: International Energy Agency, p. 305.

International Energy Agency, International Renewable Energy Agency, United Nations Statistics Division, World Bank, and World Health Organization (2019). *Global Energy Progress Report 2019*. en. Technical report 136961. The World Bank, pp. 1–176.

Jagadish, A. and Dwivedi, P. (2018). "In the hearth, on the mind: Cultural consensus on fuelwood and cookstoves in the middle Himalayas of India". en. *Energy Research & Social Science* 37, pp. 44–51. ISSN: 2214-6296. DOI: 10.1016/j.erss.2017.09.017.

Jain, A., Agrawal, S., and Ganesan, K. (2018). "Lessons from the World's Largest Subsidy Benefit Transfer Scheme". In: *The Politics of Fossil Fuel Subsidies and their Reform*. Ed. by Harro van Asselt and Jakob Skovgaard. Cambridge University Press, Cambridge, pp. 212–228. ISBN: 978-1-108-41679-5. DOI: 10.1017/9781108241946.014.

Jain, A., Tripathi, S., Mani, S., Patnaik, S., Shahidi, T., and Ganesan, K. (2018). *Access to Clean Cooking Energy and Electricity: Survey of States 2018*. Technical report.

Jana, C. and Bhattacharya, S. C. (2017). "Sustainable cooking energy options for rural poor people in India: an empirical study". en. *Environment, Development and Sustainability* 19:3, pp. 921–937. ISSN: 1387-585X, 1573-2975. DOI: 10.1007/s10668-016-9774-y.

Jeuland, M. A. et al. (2015). "Preferences for improved cook stoves: Evidence from rural villages in north India". en. *Energy Economics* 52, pp. 287–298. ISSN: 0140-9883. DOI: 10.1016/j.eneco.2015.11.010.

Jewitt, S., Atagher, P., and Clifford, M. (2020). ""We cannot stop cooking": Stove stacking, seasonality and the risky practices of household cookstove transitions in Nigeria". en. *Energy Research & Social Science* 61, p. 101340. ISSN: 2214-6296. DOI: 10.1016/j.erss.2019.101340.

Johnson, O. W., Han, J. Y.-C., Knight, A.-L., Mortensen, S., Aung, M. T., Boyland, M., and Resurrección, B. P. (2020). "Intersectionality and energy transitions: A review of gender, social equity and low-carbon energy". en. *Energy Research & Social Science* 70, p. 101774. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101774.

Johnson, P. et al. (2011). "Prevalence of chronic obstructive pulmonary disease in rural women of Tamilnadu: implications for refining disease burden assessments attributable to household biomass combustion". eng. *Global Health Action* 4, p. 7226. ISSN: 1654-9880. DOI: 10.3402/gha.v4i0.7226.

Jones, K., Manley, D., Johnston, R., and Owen, D. (2018). "Modelling residential segregation as unevenness and clustering: A multilevel modelling approach incorporating spatial dependence and tackling the MAUP". en. *Environment and Planning B: Urban Analytics and City Science* 45:6, pp. 1122–1141. ISSN: 2399-8083. DOI: 10.1177/2399808318782703.

Joon, V. and Chandra, A. (2016). "Household Air Pollution from Cooking Fuels: An Environmental and Public Health Challenge". en. *Asian Journal of Wa-*

ter, Environment and Pollution 13:4. Publisher: IOS Press, pp. 33–39. ISSN: 0972-9860. DOI: 10.3233/AJW-160036.

Kanagawa, M. and Nakata, T. (2008). "Assessment of access to electricity and the socio-economic impacts in rural areas of developing countries". *Energy Policy* 36:6, pp. 2016–2029. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.01.041.

Kang, W., Oshan, T., Wolf, L. J., Boeing, G., Frias-Martinez, V., Gao, S., Poorthuis, A., and Xu, W. (2019). "A roundtable discussion: Defining urban data science". en. *Environment and Planning B: Urban Analytics and City Science* 46:9. Publisher: SAGE Publications Ltd STM, pp. 1756–1768. ISSN: 2399-8083. DOI: 10.1177/2399808319882826.

Kar, A., Pachauri, S., Bailis, R., and Zerriffi, H. (2019). "Using sales data to assess cooking gas adoption and the impact of India's Ujjwala programme in rural Karnataka". en. *Nature Energy* 4:9, pp. 806–814. ISSN: 2058-7546. DOI: 10.1038/s41560-019-0429-8.

– **(2020)**. "Capital cost subsidies through India's Ujjwala cooking gas programme promote rapid adoption of liquefied petroleum gas but not regular use". en. *Nature Energy* 5:2. Number: 2 Publisher: Nature Publishing Group, pp. 125–126. ISSN: 2058-7546. DOI: 10.1038/s41560-019-0536-6.

Karekezi, S. and Majoro, L. (2002). "Improving modern energy services for Africa's urban poor". *Energy Policy*. Africa: Improving Energy Services for the Poor 30:11, pp. 1015–1028. ISSN: 0301-4215. DOI: 10.1016/S0301-4215(02)00055-1.

Kassambra, A. (2017). *Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning (Multivariate Analysis I)*. Vol. 1. STHDA.

Kebede, B., Bekele, A., and Kedir, E. (2002). "Can the urban poor afford modern energy? The case of Ethiopia". *Energy Policy*. Africa: Improving Energy Services for the Poor 30:11, pp. 1029–1045. ISSN: 0301-4215. DOI: 10.1016/S0301-4215(02)00056-3.

Kelley, K., Clark, B., Brown, V., and Sitzia, J. (2003). "Good practice in the conduct and reporting of survey research". en. *International Journal for Quality in Health Care* 15:3, pp. 261–266. ISSN: 1353-4505. DOI: 10.1093/intqhc/mzg031.

Kemmler, A. (2007). "Factors influencing household access to electricity in India". *Energy for Sustainable Development* 11:4, pp. 13–20. ISSN: 0973-0826. DOI: 10.1016/S0973-0826(08)60405-6.

Khalid, R. and Sunikka-Blank, M. (2017). "Homely social practices, uncanny electricity demands: Class, culture and material dynamics in Pakistan". *Energy Research & Social Science* 34, pp. 122–131. ISSN: 2214-6296. DOI: 10.1016/j.erss.2017.06.038.

Khan, F. (2017). "Air Pollution: Why a clean cooking campaigner chooses the kitchen to save the planet". *The Economic Times*.

Khandker, S. R., Barnes, D. F., and Samad, H. A. (2010). *Energy Poverty in Rural and Urban India*. Working Paper 5463. The World Bank, Development Research Group.

– **(2012)**. "Are the energy poor also income poor? Evidence from India". *Energy Policy* 47, pp. 1–12. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2012.02.028.

Khosla, R., Sircar, N., and Bhardwaj, A. (2019). "Energy demand transitions and climate mitigation in low-income urban households in India". en. *Environmental Research Letters* 14:9, p. 095008. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ab3760.

Kishore, V. V. N. and Ramana, P. V. (2002). "Improved cookstoves in rural India: how improved are they?: A critique of the perceived benefits from the National Programme on Improved Chulhas (NPIC)". en. *Energy* 27:1, pp. 47–63. ISSN: 0360-5442. DOI: 10.1016/S0360-5442(01)00056-1.

Knight, J., Wells, S., Marshall, R., Exeter, D., and Jackson, R. (2017). "Developing a synthetic national population to investigate the impact of different cardiovascular disease risk management strategies: A derivation and validation study". *PLoS ONE* 12:4. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0173170.

Knol, A. B., Slottje, P., Sluijs, J. P. van der, and Lebret, E. (2010). "The use of expert elicitation in environmental health impact assessment: a seven step procedure". *Environmental Health* 9:1, p. 19. ISSN: 1476-069X. DOI: 10.1186/1476-069X-9-19.

Komatsu, S., Kaneko, S., Shrestha, R. M., and Ghosh, P. P. (2011). "Non-income factors behind the purchase decisions of solar home systems in rural Bangladesh". *Energy for Sustainable Development*. Special issue on off-grid electrification in developing countries 15:3, pp. 284–292. ISSN: 0973-0826. DOI: 10.1016/j.esd.2011.03.003.

Kondracki, N. L., Wellman, N. S., and Amundson, D. R. (2002). "Content Analysis: Review of Methods and Their Applications in Nutrition Education". *Journal of Nutrition Education and Behavior* 34:4, pp. 224–230. ISSN: 1499-4046. DOI: 10.1016/S1499-4046(06)60097-3.

Kroon, B. van der, Brouwer, R., and Beukering, P. J. H. van (2013). "The energy ladder: Theoretical myth or empirical truth? Results from a meta-analysis". *Renewable and Sustainable Energy Reviews* 20:Supplement C, pp. 504–513. ISSN: 1364-0321. DOI: 10.1016/j.rser.2012.11.045.

– **(2014)**. "The impact of the household decision environment on fuel choice behavior". en. *Energy Economics* 44, pp. 236–247. ISSN: 0140-9883. DOI: 10.1016/j.eneco.2014.04.008.

Krosnick, J. A. (1999). "Survey Research". *Annual Review of Psychology* 50:1, pp. 537–567. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.50.1.537.

Krosnick, J. A. and Alwin, D. F. (1987). "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement". *The Public Opinion Quarterly* 51:2, pp. 201–219. ISSN: 0033-362X.

Kwac, J., Flora, J., and Rajagopal, R. (2014). "Household Energy Consumption Segmentation Using Hourly Data". *IEEE Transactions on Smart Grid* 5:1. Conference Name: IEEE Transactions on Smart Grid, pp. 420–430. ISSN: 1949-3061. DOI: 10.1109/TSG.2013.2278477.

Lam, N. L., Pachauri, S., Purohit, P., Nagai, Y., Bates, M. N., Cameron, C., and Smith, K. R. (2016). "Kerosene subsidies for household lighting in India: what are the impacts?" en. *Environmental Research Letters* 11:4, p. 044014. ISSN: 1748-9326. DOI: 10.1088/1748-9326/11/4/044014.

Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., and Thomas, C. J. (2012). "Transdisciplinary research in sustainability science: Practice, principles, and challenges". English (US). *Sustainability Science* 7:SUPPL. 1. Publisher: Springer Japan, pp. 25–43. ISSN: 1862-4065. DOI: 10.1007/s11625-011-0149-x.

Leach, G. (1992). "The energy transition". *Energy Policy*. Energy and the Third World 20:2, pp. 116–123. ISSN: 0301-4215. DOI: 10.1016/0301-4215(92)90105-B.

Lenth, R. V. (2001). "Some Practical Guidelines for Effective Sample Size Determination". *The American Statistician* 55:3, pp. 187–193. ISSN: 0003-1305. DOI: 10.1198/000313001317098149.

Louw, K., Conradie, B., Howells, M., and Dekenah, M. (2008). "Determinants of electricity demand for newly electrified low-income African households".

Energy Policy 36:8, pp. 2812–2818. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.02.032.

Lovelace, R. and Ballas, D. (2013). "'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation". en. *Computers, Environment and Urban Systems* 41, pp. 1–11. ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2013.03.004.

Lovelace, R. and Dumont, M. (2016). *Spatial Microsimulation with R*.

Lutzenhiser, L. (1992). "A cultural model of household energy consumption". *Energy* 17:1, pp. 47–60. ISSN: 0360-5442. DOI: 10.1016/0360-5442(92)90032-U.

Maiti, P., Rao, T. J., and Ghosh, J. K. (2016). "The Indian Official Statistical System Revisited". *Sankhyā: The Indian Journal of Statistics* 78:2. Place: Kolkata, India : Publisher: Indian Statistical Institute, pp. 215–237. ISSN: 0976-8386.

Malakar, Y., Greig, C., and Fliert, E. van de (2018). "Resistance in rejecting solid fuels: Beyond availability and adoption in the structural dominations of cooking practices in rural India". en. *Energy Research & Social Science* 46, pp. 225–235. ISSN: 2214-6296. DOI: 10.1016/j.erss.2018.07.025.

Malterud, K., Siersma, V. D., and Guassora, A. D. (2015). "Sample Size in Qualitative Interview Studies: Guided by Information Power". en. *Qualitative Health Research*. Publisher: SAGE PublicationsSage CA: Los Angeles, CA. DOI: 10.1177/1049732315617444.

Mandal, S., Zaveri, A., Mallick, R., and Chouhan, P. (2020). "Impact of domestic smokes on the prevalence of acute respiratory infection (ARI) among under-five children: Evidence from India". en. *Children and Youth Services Review* 114, p. 105046. ISSN: 0190-7409. DOI: 10.1016/j.childyouth.2020.105046.

Mani, S., Jain, A., Tripathi, S., and Gould, C. F. (2020). "The drivers of sustained use of liquified petroleum gas in India". en. *Nature Energy*. Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2058-7546. DOI: 10.1038/s41560-020-0596-7.

Al-Marri, W., Al-Habaibeh, A., and Watkins, M. (2018). "An investigation into domestic energy consumption behaviour and public awareness of renewable energy in Qatar". *Sustainable Cities and Society* 41, pp. 639–646. ISSN: 2210-6707. DOI: 10.1016/j.scs.2018.06.024.

Marsden, P. V. and Wright, J. D. (2010). *Handbook of survey research*. 2nd ed. / edited by Peter V. Marsden, James D. Wright. Bingley : Emerald Group Pub., 2010., Bingley.

Masera, O. R., Saatkamp, B. D., and Kammen, D. M. (2000). "From Linear Fuel Switching to Multiple Cooking Strategies: A Critique and Alternative to the Energy Ladder Model". *World Development* 28:12, pp. 2083–2103. ISSN: 0305-750X. DOI: 10.1016/S0305-750X(00)00076-0.

Medina, P., Berrueta, V., Cinco, L., Ruiz-García, V., Edwards, R., Olaya, B., Schilmann, A., and Masera, O. (2019). "Understanding Household Energy Transitions: From Evaluating Single Cookstoves to "Clean Stacking" Alternatives". en. *Atmosphere* 10:11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 693. DOI: 10.3390/atmos10110693.

Mentis, D., Welsch, M., Fuso Nerini, F., Broad, O., Howells, M., Bazilian, M., and Rogner, H. (2015). "A GIS-based approach for electrification planning—A case study on Nigeria". *Energy for Sustainable Development* 29, pp. 142–150. ISSN: 0973-0826. DOI: 10.1016/j.esd.2015.09.007.

Mirkin, B. (2015). "Quadratic error and k-means". English. In: *Handbook of Cluster Analysis*, pp. 33–54. ISBN: 978-1-4665-5189-3. DOI: 10.1201/b19706.

Mishra, V., Smith, K. R., and Retherford, R. D. (2005). "Effects of Cooking Smoke and Environmental Tobacco Smoke on Acute Respiratory Infections in Young Indian Children". en. *Population and Environment* 26:5, pp. 375–396. ISSN: 1573-7810. DOI: 10.1007/s11111-005-0005-y.

Moezzi, M. and Lutzenhiser, L. (2010). "What's Missing in Theories of the Residential Energy User". en. *ACEEE Summer Study on Energy Efficiency in Buildings*, p. 15.

MoPNG (2017). *Ministry of Petroleum and Natural Gas Notification*. Hindi/English. Technical report No. P-20019/53/2014-LPG. New Delhi: Government of India Press.

– **(2021)**. *Indian Petroleum & Natural Gas Statistics 2019-20*. Economics & Statistics Division. New Delhi: Government of India.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan". en. *Spatial and Spatio-temporal Epidemiology* 31. Publisher: Elsevier, p. 100301. ISSN: 1877-5845. DOI: 10.1016/j.sste.2019.100301.

Mosteller, F. (1968). "Association and Estimation in Contingency Tables". en. *Journal of the American Statistical Association* 63:321, pp. 1–28. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1968.11009219.

Musango, J. K., Smit, S., Ceschin, F., Ambole, A., Batinge, B., Anditi, C., Petrulaityte, A., and Mukama, M. (2020). "Mainstreaming gender to achieve security of energy services in poor urban environments". en. *Energy Research & Social Science* 70, p. 101715. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101715.

Nader, L. and Milleron, N. (1979). "Dimensions of the "people problem" in energy research and "the" factual basis of dispersed energy futures". en. *Energy*.

Proceedings of a Conference on Non-Fossil Fuel and Non-Nuclear Fuel Energy Strategies 4:5, pp. 953–967. ISSN: 0360-5442. DOI: 10.1016/0360-5442(79)90025-2.

Nageli, C., Camarasa, C., Jakob, M., Catenazzi, G., and Ostermeyer, Y. (2018). "Synthetic building stocks as a way to assess the energy demand and greenhouse gas emissions of national building stocks". English. *Energy and Buildings* 173. Place: Lausanne Publisher: Elsevier Science Sa WOS:000442173200037, pp. 443–460. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2018.05.055.

Nakano, Y. (2009). *Where There Are Asians, There Are Rice Cookers: How "National" Went Global via Hong Kong*. Hong Kong University Press. ISBN: 978-988-8028-08-5.

Narasimha Rao, M. and Reddy, B. S. (2007). "Variations in energy use by Indian households: An analysis of micro level data". *Energy* 32:2, pp. 143–153. ISSN: 0360-5442. DOI: 10.1016/j.energy.2006.03.012.

National Sample Survey Organisation (2013). *India - Household Consumer Expenditure, Type 1 : July 2011 - June 2012, NSS 68th Round*.

– **(2015)**. *Energy Sources of Indian Households for Cooking and Lighting, 2011-12*. Consumer Expenditure Survey Report No. 567. New Delhi: Ministry of Statistics and Programme Implementation.

Neto-Bradley, A. P., Choudhary, R., and Bazaz, A. B. (2019a). "Applicability of an 'uptake wave' energy transition concept in Indian households". In: *IOP Conference Series: Earth and Environmental Science*. Vol. 294. IOP, Tokyo, p. 012091. DOI: 10.1088/1755-1315/294/1/012091.

Neto-Bradley, A. P., Choudhary, R, and Bazaz, A. B (2019b). "Tailoring Residential Energy Provision Strategies in Fast-Growing Cities using Targeted Data Collection". In: *International Conference on Smart Infrastructure and Construc-*

tion 2019 (ICSIC). ICE, Cambridge, pp. 151–160. DOI: 10.1680/icsic.64669.151.

Neto-Bradley, A. P., Choudhary, R., and Challenor, P. (2021). "A microsimulation of spatial inequality in energy access: A Bayesian multi-level modelling approach for urban India". *arXiv:2109.08577 [physics]*. arXiv: 2109.08577.

Neto-Bradley, A. P., Rangarajan, R., Choudhary, R., and Bazaz, A. B. (2021). "Energy transition pathways amongst low-income urban households: A mixed method clustering approach". en. *MethodsX*, p. 101491. ISSN: 2215-0161. DOI: 10.1016/j.mex.2021.101491.

Neto-Bradley, A. P., Choudhary, R., and Bazaz, A. (2020). "Slipping through the net: Can data science approaches help target clean cooking policy interventions?" en. *Energy Policy* 144, p. 111650. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111650.

Neto-Bradley, A. P., Choudhary, R., and Challenor, P. (n.d.). "A microsimulation of spatial inequality in energy access: A Bayesian multi-level modelling approach for urban India". *Environment and Planning B-Urban Analytics and City Science* In Review: Submitted April 2021 (Preprint on ArXiv). DOI: arxiv.org/abs/2109.08577.

Neto-Bradley, A. P., Rangarajan, R., Choudhary, R., and Bazaz, A. (2021). "A clustering approach to clean cooking transition pathways for low-income households in Bangalore". *Sustainable Cities and Society* 66, p. 102697. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102697.

Nochta, T., Wan, L., Schooling, J. M., and Parlikad, A. K. (2020). "A Socio-Technical Perspective on Urban Analytics: The Case of City-Scale Digital Twins". *Journal of Urban Technology* 0:0, pp. 1–25. ISSN: 1063-0732. DOI: 10.1080/10630732.2020.1798177.

Ochieng, C. A., Murray, U., Owuor, J., and Spillane, C. (2021). "The forgotten half: Men's influence over cookstove adoption decisions in Northern Kenya". en. *Energy Research & Social Science* 74, p. 101913. ISSN: 2214-6296. DOI: 10.1016/j.erss.2021.101913.

Oguonu, T., Obumneme-Anyim, I. N., Eze, J. N., Ayuk, A. C., Okoli, C. V., and Ndu, I. K. (2018). "Prevalence and determinants of airflow limitation in urban and rural children exposed to cooking fuels in South-East Nigeria". *Paediatrics and International Child Health* 38:2. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/20469047.2018.1445506, pp. 121–127. ISSN: 2046-9047. DOI: 10.1080/20469047.2018.1445506.

Oluwole, O., Arinola, G. O., Huo, D., and Olopade, C. O. (2017). "Biomass fuel exposure and asthma symptoms among rural school children in Nigeria". *Journal of Asthma* 54:4, pp. 347–356. ISSN: 0277-0903. DOI: 10.1080/02770903.2016.1227334.

Ornstein, M. D. (2013). *A companion to survey research*. SAGE Publications Ltd, Los Angeles. ISBN: 978-1-4462-0909-7.

Osano, A., Maghanga, J., Munyeza, C. F., Chaka, B., Olal, W., and Forbes, P. B. C. (2020). "Insights into household fuel use in Kenyan communities". en. *Sustainable Cities and Society* 55, p. 102039. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102039.

Pachauri, S., Mueller, A., Kemmler, A., and Spreng, D. (2004). "On Measuring Energy Poverty in Indian Households". *World Development* 32:12, pp. 2083–2104. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2004.08.005.

Pachauri, S. and Jiang, L. (2008). "The household energy transition in India and China". *Energy Policy*. Transition towards Sustainable Energy Systems 36:11, pp. 4022–4035. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.06.016.

Pandey, A. et al. (2021). "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019". English. *The Lancet Planetary Health* 5:1. Publisher: Elsevier, e25–e38. ISSN: 2542-5196. DOI: 10.1016/S2542-5196(20)30298-9.

Patnaik, S. and Jha, S. (2020). "Caste, class and gender in determining access to energy: A critical review of LPG adoption in India". en. *Energy Research & Social Science* 67, p. 101530. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101530.

Payne, S. L. B. (1980). *The Art of Asking Questions : Studies in Public Opinion, 3*. Princeton, N.J. : Princeton University Press, [1980].

Perros, T., Buettner, P., and Parikh, P. (2021). *Understanding Pay-As-You-Go LPG Customer Behaviour*. MECS Behaviour Change Project Report. Loughborough: Modern Energy Cooking Services.

Pillarisetti, A. et al. (2019). "Promoting LPG usage during pregnancy: A pilot study in rural Maharashtra, India". en. *Environment International* 127, pp. 540–549. ISSN: 0160-4120. DOI: 10.1016/j.envint.2019.04.017.

Polk, M. (2015). "Transdisciplinary co-production: Designing and testing a transdisciplinary research framework for societal problem solving". en. *Futures*. 'Advances in transdisciplinarity 2004-2014' 65, pp. 110–122. ISSN: 0016-3287. DOI: 10.1016/j.futures.2014.11.001.

Pope Daniel, Diaz Esperanza, Smith-Sivertsen Tone, Lie Rolv T., Bakke Per, Balmes John R., Smith Kirk R., and Bruce Nigel G. (2015). "Exposure to Household Air Pollution from Wood Combustion and Association with Respiratory Symptoms and Lung Function in Nonsmoking Women: Results from the RESPIRE Trial, Guatemala". *Environmental Health Perspectives* 123:4. Publisher: Environmental Health Perspectives, pp. 285–292. DOI: 10.1289/ehp.1408200.

Prabhu, J. (2017). "Frugal innovation: doing more with less for more". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375:2095. Publisher: Royal Society, p. 20160372. DOI: 10.1098/rsta.2016.0372.

Prabhu, J. and Jain, S. (2015). "Innovation and entrepreneurship in India: Understanding jugaad". en. *Asia Pacific Journal of Management* 32:4, pp. 843–868. ISSN: 1572-9958. DOI: 10.1007/s10490-015-9445-9.

Pradhan, P., Costa, L., Rybski, D., Lucht, W., and Kropp, J. P. (2017). "A Systematic Study of Sustainable Development Goal (SDG) Interactions". en. *Earth's Future* 5:11, pp. 1169–1179. ISSN: 2328-4277. DOI: 10.1002/2017EF000632.

Prevos, P. (2016). "The invisible water utility: employee behaviour and customer experience in service-dominant logic". en. PhD Thesis. Melbourne, Australia: La Trobe University. College of Arts, Social Sciences and Commerce.

Qin, H., Huang, Q., Zhang, Z., Lu, Y., Li, M., Xu, L., and Chen, Z. (2019). "Carbon dioxide emission driving factors analysis and policy implications of Chinese cities: Combining geographically weighted regression with two-step cluster". *Science of The Total Environment* 684, pp. 413–424. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2019.05.352.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahut, D. B., Behera, B., and Ali, A. (2016a). "Household energy choice and consumption intensity: Empirical evidence from Bhutan". *Renewable and Sustainable Energy Reviews* 53:Supplement C, pp. 993–1009. ISSN: 1364-0321. DOI: 10.1016/j.rser.2015.09.019.

– **(2016b)**. "Patterns and determinants of household use of fuels for cooking: Empirical evidence from sub-Saharan Africa". *Energy* 117:Part 1, pp. 93–104. ISSN: 0360-5442. DOI: 10.1016/j.energy.2016.10.055.

Ranjan, R. and Singh, S. (2020). "Household Cooking Fuel Patterns in Rural India: Pre- and Post-Pradhan Mantri Ujjwala Yojana". en. *Indian Journal of Human Development* 14:3. Publisher: SAGE Publications India, pp. 518–526. ISSN: 0973-7030. DOI: 10.1177/0973703020975045.

Rao, N. D. (2012). "Kerosene subsidies in India: When energy policy fails as social policy". *Energy for Sustainable Development* 16:1, pp. 35–43. ISSN: 0973-0826. DOI: 10.1016/j.esd.2011.12.007.

Rao, N. D. and Ummel, K. (2017). "White goods for white people? Drivers of electric appliance growth in emerging economies". *Energy Research & Social Science* 27, pp. 106–116. ISSN: 2214-6296. DOI: 10.1016/j.erss.2017.03.005.

Ravindra, K., Kaur-Sidhu, M., Mor, S., and John, S. (2019). "Trend in household energy consumption pattern in India: A case study on the influence of socio-cultural factors for the choice of clean fuel use". en. *Journal of Cleaner Production* 213, pp. 1024–1034. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2018.12.092.

Reddy, A. K. N. and Reddy, B. S. (1994). "Substitution of energy carriers for cooking in Bangalore". en. *Energy*. Energy Use in Asian Cities 19:5, pp. 561–571. ISSN: 0360-5442. DOI: 10.1016/0360-5442(94)90052-3.

Reddy, B. S. (1996). "Modelling of competing technologies in the residential sector". *Energy Conversion and Management* 37:1, pp. 117–125. ISSN: 0196-8904. DOI: 10.1016/0196-8904(95)00016-7.

Reuschke, D., Long, J., and Bennett, N. (2021). "Locating creativity in the city using Twitter data". en. *Environment and Planning B: Urban Analytics and City Science*. Publisher: SAGE Publications Ltd STM, p. 2399808320980745. ISSN: 2399-8083. DOI: 10.1177/2399808320980745.

Rogers, E. M. (1983). *Diffusion of innovations.* 3rd ed. / Everett M. Rogers. New York : Free Press, New York : London.

Rohra, H. and Taneja, A. (2016). "Indoor air quality scenario in India—An outline of household fuel combustion". en. *Atmospheric Environment* 129, pp. 243–255. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2016.01.038.

Rolffs, P., Ockwell, D., and Byrne, R. (2015). "Beyond technology and finance: pay-as-you-go sustainable energy access and theories of social change:" en. *Environment and Planning A*. Publisher: SAGE PublicationsSage UK: London, England. DOI: 10.1177/0308518X15615368.

Ronzi, S., Puzzolo, E., Hyseni, L., Higgerson, J., Stanistreet, D., Hugo, M. N. B., Bruce, N., and Pope, D. (2019). "Using photovoice methods as a community-based participatory research tool to advance uptake of clean cooking and improve health: The LPG adoption in Cameroon evaluation studies". en. *Social Science & Medicine* 228, pp. 30–40. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2019.02.044.

Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

Rousseeuw, P. J. and Kaufman, L. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. ISBN: 1-282-30805-X.

Ruijven, B. van, Vries, B. de, Vuuren, D. P. van, and Sluijs, J. P. van der (2010). "A global model for residential energy use: Uncertainty in calibration to regional data". en. *Energy* 35:1, pp. 269–282. ISSN: 0360-5442. DOI: 10.1016/j.energy.2009.09.019.

Ruijven, B. J. van, Vuuren, D. P. van, Vries, B. J. M. de, Isaac, M., Sluijs, J. P. van der, Lucas, P. L., and Balachandra, P. (2011). "Model projections for household energy use in India". *Energy Policy*. Clean Cooking Fuels and Technologies in Developing Economies 39:12, pp. 7747–7761. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2011.09.021.

Ruiz-Mercado, I. and Masera, O. (2015). "Patterns of Stove Use in the Context of Fuel–Device Stacking: Rationale and Implications". en. *EcoHealth* 12:1, pp. 42–56. ISSN: 1612-9210. DOI: 10.1007/s10393-015-1009-4.

Ruiz-Mercado, I., Masera, O., Zamora, H., and Smith, K. R. (2011). "Adoption and sustained use of improved cookstoves". en. *Energy Policy*. Clean Cooking Fuels and Technologies in Developing Economies 39:12, pp. 7557–7566. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2011.03.028.

Saatkamp, B. D., Masera, O. R., and Kammen, D. M. (2000). "Energy and health transitions in development: fuel use, stove technology, and morbidity in Jarácuaro, México". en. *Energy for Sustainable Development* 4:2, pp. 7–16. ISSN: 0973-0826. DOI: 10.1016/S0973-0826(08)60237-9.

Sadia, F. and Hossain, S. S. (2014). "Contrast of Bayesian and Classical Sample Size Determination". *Journal of Modern Applied Statistical Methods* 13:2. DOI: 10.22237/jmasm/1414815720.

Sahasranaman, A. and Bettencourt, L. M. (2019). "Economic geography and the scaling of urban and regional income in India". en. *Environment and Planning B: Urban Analytics and City Science*, p. 2399808319879463. ISSN: 2399-8083. DOI: 10.1177/2399808319879463.

Sandelowski, M. (2000). "Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies". en. *Research in Nursing & Health* 23:3, pp. 246–255. ISSN: 1098-240X. DOI: 10.1002/1098-240X(200006)23:3<246::AID-NUR9>3.0.CO;2-H.

Sandwell, P., Chambon, C., Saraogi, A., Chabenat, A., Mazur, M., Ekins-Daukes, N., and Nelson, J. (2016). "Analysis of energy access and impact of modern energy sources in unelectrified villages in Uttar Pradesh". *Energy for Sustainable Development* 35, pp. 67–79. ISSN: 0973-0826. DOI: 10.1016/j.esd.2016.09.002.

Sankhyayan, P. and Dasgupta, S. (2019). "'Availability' and/or 'Affordability': What matters in household energy access in India?" *Energy Policy* 131, pp. 131–143. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2019.04.019.

Sarkar, S. (2019). "Urban scaling and the geographic concentration of inequalities by city size". en. *Environment and Planning B: Urban Analytics and City Science* 46:9. Publisher: SAGE Publications Ltd STM, pp. 1627–1644. ISSN: 2399-8083. DOI: 10.1177/2399808318766070.

Saxena, V. and Bhattacharya, P. C. (2018). "Inequalities in LPG and electricity consumption in India: The role of caste, tribe, and religion". en. *Energy for Sustainable Development* 42, pp. 44–53. ISSN: 0973-0826. DOI: 10.1016/j.esd.2017.09.009.

Schmid, G. (2012). "The development of renewable energy power in India: Which policies have been effective?" en. *Energy Policy* 45, pp. 317–326. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2012.02.039.

Sehjpal, R., Ramji, A., Soni, A., and Kumar, A. (2014). "Going beyond incomes: Dimensions of cooking energy transitions in rural India". *Energy* 68, pp. 470–477. ISSN: 0360-5442. DOI: 10.1016/j.energy.2014.01.071.

Shankar, A. V. et al. (2020). "Everybody stacks: Lessons from household energy case studies to inform design principles for clean energy transitions". en. *Energy Policy* 141, p. 111468. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111468.

Sharma, A., Parikh, J., and Singh, C. (2019). "Transition to LPG for cooking: A case study from two states of India". *Energy for Sustainable Development* 51, pp. 63–72. ISSN: 0973-0826. DOI: 10.1016/j.esd.2019.06.001.

Sharma, D., Ravindra, K., Kaur, M., Prinja, S., and Mor, S. (2020). "Cost evaluation of different household fuels and identification of the barriers for the choice of clean cooking fuels in India". en. *Sustainable Cities and Society* 52, p. 101825. ISSN: 2210-6707. DOI: 10.1016/j.scs.2019.101825.

Shine Jaocb (2017). "Kerosene subsidy may go by 2020". English. *Business Standard*.

Shove, E. and Walker, G. (2014). "What Is Energy For? Social Practice and Energy Demand". en. *Theory, Culture & Society* 31:5, pp. 41–58. ISSN: 0263-2764. DOI: 10.1177/0263276414536746.

Smith, A., Lovelace, R., and Birkin, M. (2017). "Population Synthesis with Quasirandom Integer Sampling". *Journal of Artificial Societies and Social Simulation* 20:4, p. 14. ISSN: 1460-7425. DOI: 10.18564/jasss.3550.

Smith, D. M., Clarke, G. P., and Harland, K. (2009). "Improving the Synthetic Data Generation Process in Spatial Microsimulation Models". en. *Environment and Planning A: Economy and Space* 41:5. Publisher: SAGE Publications Ltd, pp. 1251–1268. ISSN: 0308-518X. DOI: 10.1068/a4147.

Smith, K. (2018). "Pradhan Mantri Ujjwala Yojana: Transformation of subsidy to social investment in India". *Making of New India: Transformation under Modi Government. Dr Syama Prasad Mookerjee Research Foundation and Wisdom Tree*, pp. 401–410.

Smith, K. R., Bruce, N., et al. (2014). "Millions Dead: How Do We Know and What Does It Mean? Methods Used in the Comparative Risk Assessment of Household Air Pollution". *Annual Review of Public Health* 35:1, pp. 185–206. DOI: 10.1146/annurev-publhealth-032013-182356.

Smith, K. R. and Dutta, K. (2011). ""Cooking with Gas"". en. *Energy for Sustainable Development* 15:2, pp. 115–116. ISSN: 0973-0826. DOI: 10.1016/j.esd.2011.05.001.

Smith, K. R. and Sagar, A. (2014). "Making the clean available: Escaping India's Chulha Trap". *Energy Policy* 75, pp. 410–414. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2014.09.024.

Sokołowski, M. M. (2019). "When black meets green: A review of the four pillars of India's energy policy". *Energy Policy* 130, pp. 60–68. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2019.03.051.

Sovacool, B. K. et al. (2015). "Integrating social science in energy research". *Energy Research & Social Science* 6:Supplement C, pp. 95–99. ISSN: 2214-6296. DOI: 10.1016/j.erss.2014.12.005.

Sovacool, B. K. (2011). "Conceptualizing urban household energy use: Climbing the "Energy Services Ladder"". *Energy Policy* 39:3, pp. 1659–1668. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2010.12.041.

Sovacool, B. K., Axsen, J., and Sorrell, S. (2018). "Promoting novelty, rigor, and style in energy social science: Towards codes of practice for appropriate methods and research design". en. *Energy Research & Social Science*. Special Issue on the Problems of Methods in Climate and Energy Research 45, pp. 12–42. ISSN: 2214-6296. DOI: 10.1016/j.erss.2018.07.007.

Stephenson, J. (2017). "What does energy mean? An interdisciplinary conversation". *Energy Research & Social Science* 26, pp. 103–106. ISSN: 2214-6296. DOI: 10.1016/j.erss.2017.01.014.

Stern, P. C. (1986). "Blind Spots in Policy Analysis: What Economics Doesn't Say about Energy Use". *Journal of Policy Analysis and Management* 5:2. Publisher: [Wiley, Association for Public Policy Analysis and Management], pp. 200–227. ISSN: 0276-8739. DOI: 10.2307/3323541.

Stern, P. C. (2014). "Individual and household interactions with energy systems: Toward integrated understanding". en. *Energy Research & Social Science* 1, pp. 41–48. ISSN: 2214-6296. DOI: 10.1016/j.erss.2014.03.003.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). "Assessing the performance of prediction models: a framework for some traditional and novel measures". *Epidemiology (Cambridge, Mass.)* 21:1, pp. 128–138. ISSN: 1044-3983. DOI: 10.1097/EDE.0b013e3181c30fb2.

Thurber, M. C., Phadke, H., Nagavarapu, S., Shrimali, G., and Zerriffi, H. (2014). "'Oorja' in India: Assessing a large-scale commercial distribution of advanced biomass stoves to households". *Energy for sustainable development : the journal of the International Energy Initiative* 19, pp. 138–150. ISSN: 0973-0826. DOI: 10.1016/j.esd.2014.01.002.

Tibshirani, R., Walther, G., and Hastie, T. (2001). "Estimating the Number of Clusters in a Data Set via the Gap Statistic". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63:2. Publisher: [Royal Statistical Society, Wiley], pp. 411–423. ISSN: 13697412, 14679868.

Tirumalachetty, S., Kockelman, K. M., and Nichols, B. G. (2013). "Forecasting greenhouse gas emissions from urban regions: microsimulation of land use and transport patterns in Austin, Texas". en. *Journal of Transport Geography* 33, pp. 220–229. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2013.08.002.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge. ISBN: 978-0-521-57629-1. DOI: 10.1017/CBO9780511819322.

Troncoso, K., Armendáriz, C., and Alatorre, S. (2013). "Improved cook stove adoption and impact assessment: A proposed methodology". en. *Energy Policy*

62, pp. 637–645. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2013.07.074.

Turnheim, B., Berkhout, F., Geels, F., Hof, A., McMeekin, A., Nykvist, B., and Vuuren, D. van (2015). "Evaluating sustainability transitions pathways: Bridging analytical approaches to address governance challenges". *Global Environmental Change* 35, pp. 239–253. ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2015.08.010.

UIDAI (2019). *Annual Report 2018-19*. Annual Report AADHAR. New Delhi: AADHAAR - Unique Identification Authority of India, p. 109.

Urpelainen, J. and Yoon, S. (2015). "Solar home systems for rural India: Survey evidence on awareness and willingness to pay from Uttar Pradesh". *Energy for Sustainable Development* 24, pp. 70–78. ISSN: 0973-0826. DOI: 10.1016/j.esd.2014.10.005.

Wang, Y., Chen, Q., Hong, T., and Kang, C. (2019). "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges". *IEEE Transactions on Smart Grid* 10:3, pp. 3125–3148. ISSN: 1949-3061. DOI: 10.1109/TSG.2018.2818167.

Ward, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58:301, pp. 236–244. ISSN: 0162-1459. DOI: 10.1080/01621459.1963.10500845.

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., and Zhao, X. (2018). "A review of data-driven approaches for prediction and classification of building energy consumption". *Renewable and Sustainable Energy Reviews* 82, pp. 1027–1047. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.09.108.

West, G. B. (2017). *Scale: The Universal Laws of Life and Death In Organisms, Cities and Companies*. Weidenfeld & Nicolson, London. ISBN: 978 1 78022 559 3.

Williams, K. N., Kephart, J. L., Fandiño-Del-Rio, M., Condori, L., Koehler, K., Moulton, L. H., Checkley, W., and Harvey, S. A. (2020). "Beyond cost: Exploring fuel choices and the socio-cultural dynamics of liquefied petroleum gas stove adoption in Peru". en. *Energy Research & Social Science* 66, p. 101591. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101591.

World Bank (2018). *"Access to electricity, urban (% of urban population)"*. Database.

Wu, Y., Blunden, L. S., and Bahaj, A. S. (2019). "City-wide building height determination using light detection and ranging data". en. *Environment and Planning B: Urban Analytics and City Science* 46:9, pp. 1741–1755. ISSN: 2399-8083. DOI: 10.1177/2399808318774336.

Yadav, P., Davies, P. J., and Asumadu-Sarkodie, S. (2021). "Fuel choice and tradition: Why fuel stacking and the energy ladder are out of step?" en. *Solar Energy* 214, pp. 491–501. ISSN: 0038-092X. DOI: 10.1016/j.solener.2020.11.077.

Yu, Z., Fung, B. C. M., Haghighat, F., Yoshino, H., and Morofsky, E. (2011). "A systematic procedure to study the influence of occupant behavior on building energy consumption". en. *Energy and Buildings* 43:6, pp. 1409–1417. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2011.02.002.

Yu, Z., Haghighat, F., Fung, B. C., Morofsky, E., and Yoshino, H. (2011). "A methodology for identifying and improving occupant behavior in residential buildings". *Energy* 36:11, pp. 6596–6608. ISSN: 0360-5442. DOI: 10.1016/j.energy.2011.09.002.

Yuan, M. and Choudhary, R. (2020). "A two-step clustering framework for locally tailored design of residential heating policies". en. *Sustainable Cities and*

Society 63, p. 102431. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102431.

Yunus, M., Moingeon, B., and Lehmann-Ortega, L. (2010). "Building Social Business Models: Lessons from the Grameen Experience". en. *Long Range Planning*. Business Models 43:2, pp. 308–325. ISSN: 0024-6301. DOI: 10.1016/j.lrp.2009.12.005.

Zakhary, S., Rosser, J., Siebers, P.-O., Mao, Y., and Robinson, D. (2020). "Using unsupervised learning to partition 3D city scenes for distributed building energy microsimulation". en. *Environment and Planning B: Urban Analytics and City Science*. Publisher: SAGE Publications Ltd STM, p. 2399808320914313. ISSN: 2399-8083. DOI: 10.1177/2399808320914313.

Zhang, W., Robinson, C., Guhathakurta, S., Garikapati, V. M., Dilkina, B., Brown, M. A., and Pendyala, R. M. (2018). "Estimating residential energy consumption in metropolitan areas: A microsimulation approach". en. *Energy* 155, pp. 162–173. ISSN: 0360-5442. DOI: 10.1016/j.energy.2018.04.161.