THE DEVELOPMENT OF OBJECT ORIENTED BAYESIAN NETWORKS TO EVALUATE THE SOCIAL, ECONOMIC AND ENVIRONMENTAL IMPACTS OF SOLAR PV

A Doctoral Thesis

By Philip A. Leicester

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DOCTOR OF PHILOSOPHY OF LOUGHBOROUGH UNIVERSITY

2015

Abstract

Domestic and community low carbon technologies are widely heralded as valuable means for delivering sustainability outcomes in the form of social, economic and environmental (SEE) policy objectives. To accelerate their diffusion they have benefited from a significant number and variety of subsidies worldwide. Considerable aleatory and epistemic uncertainties exist, however, both with regard to their net energy contribution and their SEE impacts. Furthermore the socio-economic contexts themselves exhibit enormous variability, and commensurate uncertainties in their parameterisation. This represents a significant risk for policy makers and technology adopters.

This work describes an approach to these problems using Bayesian Network models. These are utilised to integrate extant knowledge from a variety of disciplines to quantify SEE impacts and endogenise uncertainties. A large-scale Object Oriented Bayesian network has been developed to model the specific case of solar photovoltaics (PV) installed on UK domestic roofs. Three specific model components have been developed. The PV component characterises the yield of UK systems, the building energy component characterises the energy consumption of the dwellings and their occupants and a third component characterises the building stock in four English urban communities.

Three representative SEE indicators, fuel affordability, carbon emission reduction and discounted cash flow are integrated and used to test the model's ability to yield meaningful outputs in response to varying inputs. The variability in the percentage of the three indicators is highly responsive to the dwellings' built form, age and orientation, but is not just due to building and solar physics but also to socio-economic factors. The model can accept observations or evidence in order to create scenarios which facilitate deliberative decision making.

The BN methodology contributes to the synthesis of new knowledge from extant knowledge located 'between disciplines'. As well as insights into the impacts of high PV penetration, an epistemic

contribution has been made to transdisciplinary building energy modelling which can be replicated with a variety of low carbon interventions.

Acknowledgments

I thank Loughborough University centenary fund, the Centre for Renewable Energy and Systems Technology, and the School of Civil and Building Engineering for the financial support to undertake this research. These are warmly thanked for their collegial working and social environment.

Many people have made valuable contributions to this research. I express gratitude to my research supervisors, Drs Chris Goodier and Paul Rowley for their engagement, persistent and constructive challenge and cajoling during this research, which, for all of us, was a new area of academic endeavour. I'd like to thank Drs Veronique Delcroix and Ali Ben Mrad (University of Valenciennes, France) for the enthusiastic and collaborative response to this new application of probabilistic modelling. I am grateful to Dr Ben Anderson (University of Essex/University of Southampton) for the generous sharing of spatially resolved household income data and to Dr Alastair Buckley and Aldous Everard (University of Sheffield) for the sharing of empirical solar PV yield data. I thank Dr Nick Doylend for assistance with the R statistical package and executing iterative proportional fitting.

Many research colleagues have shown great support and friendship over the years; the diverse, multi-cultural and international environment has been a pleasure to be amongst and delivers an optimism that humanity can indeed collaborate across borders, and between cultures, to solve our existential challenges. Or at least we can share some good food. There are too many to name, but *i due italiani*, Dr Biancamaria Maniscalco and Dr Fabiana Lisco, deserve special mention for the warmth of their friendship, and tolerance of the budgerigar.

Normal stochastic life events continued outside of University. The support of the Kerslakes for the Warner-Leicesters has been an invaluable safety net for which I am ever thankful. I thank Poppy and Rowan Warner-Leicester for their patience and tolerance of a uni-focussed father and the sacrifices they inevitably have made during some difficult years at school and University.

Lastly, and undisputedly mostly, I turn to, and thank, my soul-mate Susan Warner who has suffered, and enjoyed, with me, on this journey, whilst she has made a similarly difficult one of her own. The goals she has attained make me immensely proud and her achievement is all the more special because, during all this, she has still been my rock, my sounding board, my third supervisor, my crutch, my research assistant and umpteen other essential back office and front of house support roles, without which none of these pages would have been written.

No support was given by Marta Leicester nee Seemann, 'Vertriebene', economic migrant, Oma and Mum, whose world is now one of daily uncertainty, with no context for her priors with which to make posterior inferences. But I thank her anyway, with all my heart, for she bequeaths a hunger for understanding and knowledge which to her was always a form of nutrition. *„Iss", meinte sie, „dass aus dir was wird". Also Mutti, ich hab' nun was geschluckt und bin dabei beinah erstickt. Ich bin aber trotzdem was geworden – bin mir nur nicht so sicher was. Mal sehen.*

Philip Alexander Leicester, Loughborough, 23 September 2015

List of Publications

The following publications and research conference contributions have been made during the course of this research.

1. Leicester, P.A., Goodier, C.I. and Rowley, P.N. (2016). Probabilistic evaluation of solar photovoltaic systems using Bayesian networks: a discounted cash flow assessment. Prog. Photovolt: Res. Appl. DOI: 10.1002/pip.2754. (Principal author).
2. Leicester, P.A., Rowley, P.N. and Goodier, C.I., (2016). Probabilistic analysis of solar photovoltaic self-consumption using Bayesian network models. *IET Renewable Power Generation*, 10(4), pp.448-455. DOI: 10.1049/iet-rpg.2015.0360. (Principal author).
3. Mrad, A.B., Delcroix, V., Piechowiak, S., Leicester, P.A. and Abid, M. (2015). An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, [online] pp.1-23. Available at: http://dx.doi.org/10.1007/s10489-015-0678-6. (Contributing author).
4. Leicester, P., Rowley, P.N., Goodier, G.I., (2015), Probabilistic Evaluation of UK Domestic Solar Photovoltaic Systems Using Bayesian Networks: A Discounted Cash Flow Assessment, EU-PVSEC, September 2015. (Principal Author and accepted for poster presentation).
5. Leicester, P., Rowley, P.N., Goodier, C.I., (2015), Evaluating self-consumption for domestic solar PV: simulation using highly resolved generation and demand data for varying occupant archetypes, Proceedings of Conference C97 of the SOLAR ENERGY Society, PVSAT-11, 15-17/04/15, Leeds, UK, Eds: Michael Hutchins and Alex Cole, http://www.pvsat.org.uk, ISBN 0 904963 81 0. (Principal author and oral presentation).

6. Rowley, P., Leicester, P., Palmer, D., Westacott, P., Candelise, C., Betts, T., Gottschalg, R., (2015), Multi-domain Analysis of Photovoltaic Impacts via Integrated Spatial & Probabilistic Modelling, IET Renewable Power Generation, pp.1-8, ISSN 1752-1416. (Contributing author).
7. Mrad, A.B., Delcroix, V., Piechowiak, S., Leicester, P., (2014) From Information to Evidence in a Bayesian Network, Proceedings of 7th European Workshop on Probabilistic graphical Models, Lecture Notes in Artificial Intelligence/Computer Science, Volume 8754, pp 33-48. Published Springer, ISBN 978-3-319-11432-3 (Contributing author).
8. Rowley, P., Leicester, P., Thornley, P, Mander, S, Jones, C., (2014) WISE-PV: Whole System and Socio-Economic Impacts Of Wide Scale PV Integration, Proceedings PVSEC, Amsterdam, September 2014. (Contributing author).
9. Leicester P.A., Goodier C., Rowley P., (2014), Evaluating the contribution of PV to Social, Economic and Environmental Aspects of Community Renewable Energy Projects, Proceedings of Conference C96 of the SOLAR ENERGY Society, PVSAT-10, 23-25/04/14, Loughborough University, UK, Eds: Michael Hutchins, Alex Cole and Ralph Gottschalg, http://www.pvsat.org.uk, ISBN 0 904963 80 2. (Principal author).
10. Leicester P.A., Goodier C., Rowley P., (2013) Using a Bayesian Network to Evaluate the Social, Economic and Environmental Impacts of Community Deployed Renewable Energy, IN: Scartezzini, J.L. (ed.) Proceedings of CISBAT, Clean Technology for Smart Cities and Buildings, Lausanne, 4-6 September 2013, 10 pp. https://dspace.lboro.ac.uk/2134/14472. (Principal author).
11. Rowley, P., Gough, R., Doylend, N., Thirkill, A., Leicester, P.A., (2013). From Smart Homes to Smart Communities : Advanced Data Acquisition and Analysis for Improved Sustainability and Decision Making, Proceedings of the I-Society Conference, Toronto, June 2013. (Contributing author).

12. Leicester, p., Goodier, C.I. and Rowley, P., (2011) Evaluating the impacts of Community Renewable Energy Initiatives, ISES Solar World Congress, Kassel, Germany, 28 Aug.- 2 Sept. 2011, https://dspace.lboro.ac.uk/2134/9185. (Principal author and oral presentation).

Table of Contents

Table of Figures

Table of Tables

Glossary

Abbreviation	Meaning
BN	Bayesian Network
BOS	Balance of system
BREDEM	Building Research Establishment Domestic Energy Model
CAPI	computer assisted personal interviews
CHP	Combined heat and power
CPT	Conditional probability table
DAG	Directed acyclic graph
DCFA	Discounted cash flow analysis
DECC	Department for energy and climate change
DEFRA	Department for Environment, Food and Rural Affairs
EHS	English Housing Survey
EPC	Energy performance certificates
ESRI	Environmental Systems Research Institute
FiT	Feed-in tariff
GIS	Geographic information system
HEED	Home Energy Efficiency Database
IMD	Index of multiple deprivation
IPF	Iterative Proportional Fitting
JPD	Joint probability distribution
kWh	Kilowatt hour
LCF	Living Costs and Food Survey
LSOA	Lower super output area
MBE	mean bias error
MLR	multiple linear regression
NEED	National energy efficiency data
NPT	Node probability table
NPV	Net present value
OECD	Organisation for Economic Co-operation and Development
ONS	Office for national statistics
OOBN	object oriented Bayesian network
PDF	Probability density function
PDFT	Photovoltaic domestic field trials
PGM	Probabilistic graphical models
PMF	Probability mass function
PV	Photovoltaic
rdSAP	Reduced Dataset Standard Assessment Procedure
RMSE	Root mean square error
SAP	Standard Assessment Procedure
SE	Standard error
SEE	Social, environmental and economic
SMD	Sheffield Microgeneration Database
TAN	Tree augmented naïve
UML	Unified modelling language
VOA	Valuation Office Agency

1 Introduction

The only certainty is uncertainty

1.1 Motivation

Renewable energy technologies deployed in community contexts are seen as a valuable contribution to a number of energy policy objectives, and as such are benefitting from a range of financial support mechanisms both in the UK (Woodman and Mitchell, 2011) and internationally (IEA, 2012). In the UK, since 2010, a considerable increase in the rate of deployment and installed capacity of Solar PV (Photovoltaic) systems has occurred, incentivised by the feed-in-tariff (FiT) together with cost reductions (Cherrington et al., 2013).

The FiT has rendered Solar PV a sound financial investment such that by April 2015 the number of solar installations in the UK has reached 685,000 with a total installation capacity of 3.06GW (2015). However, significant uncertainty exists with regards to the potential impacts of community scale PV in terms of specific policy goals, including actual (as opposed to projected) greenhouse gas reductions, renewable energy generation capacity and socio-economic benefits, such as, for example fuel affordability. Such uncertainty represents a risk for decision and policy makers as well as investors (Rowley et al., 2015).

Sources of uncertainty with respect to Solar PV performance are due to technical factors, pertaining both to the renewable energy resource, and to the technologies developed to harness this energy (Goss, Gottschalg and Betts, 2012). However uncertainty also derives from the wide variability of social, economic and environmental (SEE) parameters which characterise solar PV within its deployment context. This interaction between the SEE, and technical variability, ensures that every

deployment context is different. This gives rise to the challenge of propagating uncertainty within, and between, disciplinary boundaries in a multi-disciplinary problem domain (Jiang et al., 2012).

It is possible to create conceptual models, abstractions of the real world, to explore the relationships between parameters in a systemic model. Often such approaches are qualitative and provide valuable insight into a multidisciplinary domain; causal mapping (Goodier et al, 2010) and Soft System Methodology (Checkland, 2003) are typical examples. Whilst such methodologies serve as valuable problem structuring and solving tools (Mingers and Rosenhead, 2004), they do not furnish decision and policy makers with the quantitative analysis often desired. As an alternative, a number of deterministic modelling environments exist which presume a mechanistic relationships between parameters. System Dynamics is one such approach, popularised by its use in the World3 model published in the Limits to Growth (Meadows, 1972).

Using such techniques, uncertainty can only be explored using sensitivity analysis approaches (Saltelli, 2008). Two problems persist; firstly a mechanistic (deterministic) relationship between parameters is required, empirical or otherwise, which is often impractical, particularly at the interface between knowledge domains. Secondly, for a sensitivity analysis the uncertainty is exogenous to the model as each parameter is varied outside of the model's definition.

A number of multidisciplinary research projects have treated this problem by endogenising the uncertainty into the model itself, by introducing variables as probability distributions. Moreover, the mechanistic relationship between parameters can be replaced by probabilistic relationships defined by conditional probabilities. These techniques also offer the benefit of the qualitative problem structuring methods by incorporating intuitive visualisation in the form of probabilistic graphical models (PGM) (Koller and Friedman, 2009). This marriage of qualitative and quantitative epistemologies has given PGM recognition for transdisciplinary knowledge integration (Duespohl, Frank and Doell, 2012).

The research presented here has a number of attributes which lend itself to such a combined quantitative approach using probabilistic parameters, and a qualitative structuring of the problem domain. As mentioned already, there are uncertainties in the parameters which define technology performance and those which define the SEE context of technology deployment. Thus this multidisciplinary problem domain is beset with probabilistic parameters and the requirement to interrelate these between various domains. Mechanistic relationships are absent between these domains, and probabilistic approaches, it will be shown (Chapter 2), offer a promising modelling solution.

The challenge, then, is quantify uncertainties and to propagate them between knowledge domains in a meaningful manner. The lack of attendance to this leads to a significant gap in the understanding of the SEE impacts of distributed renewable energy technologies and undermines the ability to test and predict policy objectives. The solution is to develop integrated modelling approaches which endogenise uncertainties in order to elicit meaningful perspectives on SEE impacts of distributed Solar PV. Such a solution could serve as a decision or policy support tool. It is to explore the development of such a solution in order to provide insight into the impacts of solar PV that is the overarching motivation for this research project.

To provide some context for the problem domain the next few sections in this introduction will introduce several core themes which explicate how this motivation has been developed into more formal research aims and corresponding objectives. These themes are:

(i) The energy transition, a euphemism to describe the decarbonisation of the energy system, and its policy drivers and incentives.

(ii) The impacts of this energy transition, particularly focusing on the technology of solar PV, and focussing on environmental, economic, social indicators.

(iii) The uncertainties in the measurement of such indicators and how this can be modelled. Here the main methodology utilised in this research is introduced.

These themes are corralled into a set of tangible research aims and objectives. The remainder of the introduction provides a summary of the structure of the whole thesis chapter by chapter.

1.2 The Energy Transition

The Energy Transition is a moniker for the techno-socio-economic processes to divest the energy system of a reliance on fossil fuels by replacement with alternative forms of energy generation, and energy efficiency (Strunz, 2014). The transition is the subject of academic research in a wide variety of disciplines from engineering, social sciences and geography (Bridge et al., 2013; Verbong and Geels, 2010). Relevant to this study are the drivers for the transition, since these influence the indicators which policy makers may wish use to measure impact, and the enablers, invariably market instruments, used to accelerate the adoption of new technologies, the creation of supply chains and development of new practices (Chmutina et al., 2014).

1.2.1 Drivers for 'clean energy'

The switch from an agrarian society to a major industrialised economy was accompanied by a dramatic increase in energy requirements, with UK demand soaring from less than 100 TWh/year to nearly 3000 TWh/year (Fouquet and Pearson, 1999). This energy demand has been satisfied by the extraction and consumption of carbon based fossil fuel. The close correlation between economic growth and the consumption of carbon (Jackson, 2012) has been characterised as 'carbon lock-in' (Unruh, 2000) suggesting that transition out of this dependency is fraught with socio-political difficulties (Unruh, 2002).

The drivers for just such a transition are manifold. Firstly there is the environmental impact resulting from the release of fossil fuel combustion products, chiefly carbon dioxide ($C0_2$). The latter is the

chief protagonist in the theory of anthropogenic climate change (IPCC 2013), upheld as the current scientific consensus (Cook et al., 2013). This holds that $C0_2$ traps long wavelength radiation emitted by the ambient Earth; with increasing concentration this barrier is increasingly lower in the atmosphere resulting in a thermal insulating effect (Arrhenius, 1896). At current rates of fossil fuel oxidation, climate models have predicted an average warming of between 2 and 6 °C by the end of this century (IPCC, 2013). The social and economic costs of climate change are said to far outweigh the cost of mitigating this damage (Stern 2005).

Secondly, there are increasing concerns over energy security (Mitchell, Watson and Whiting, 2013). At the time of writing, geopolitical instabilities in Eastern Europe and the Middle East are affecting oil and gas distribution. Lack of security of supply causes price shocks as occurred with the OPEC oil crisis in the 1970s, resulting in negative economic impacts (Helm, 2003). Attendant with such shocks is the inevitable search for secure supplies, satisfied in the UK by the extraction of North Sea oil and gas (Helm, opt cit.). The gradual depletion of North Sea reserves underscores the problem of peak oil. This is the contested moment in time after which oil recovery will only ever decrease, with diminishing supply unleashing ever increasing prices (Piercy, Granger and Goodier, 2010).

A third driver for an energy transition is the quest for affordability. The close-coupling between the Gross Domestic Product and the energy consumption of developed nations offers competitive advantage to economies with cheap energy supplies. As well as for business users, affordability has important ramifications for domestic consumers. With householders spending a significant proportion of their income on domestic fuel there is a smaller budget for other consumer spending. A more urgent problem, especially in the UK energy policy context is fuel affordability for low income households and the specific debate around fuel poverty (Boardman, 2012).

These drivers are interrelated and have been described as the energy trilemma (Hamakawa, 2002; Hammond and Pearson, 2013), based on the notion that it is difficult to mitigate one of the problems without impacting on the other two. Nevertheless, there has been focussed global, European and UK

efforts to mitigate climate change through, respectively, Kyoto protocol (UN, 1997), EU directives (EU, 2009) and the 2008 Climate Change Act, amongst others. These and related policies have translated into political support for the subvention of low carbon technologies which is discussed next.

1.2.2 Market Instruments

The liberalisation of the energy markets, which have evolved since the deregulatory utility Acts of the 1980s, has ensured that efforts to tackle the energy trilemma are pursued through the use of market instruments designed to give a pecuniary advantage to desirable technological and supply chain developments (Helm, 2002). Since 2000 the UK Government has introduced the Renewables Obligation (Woodman and Mitchell, 2011) to support low carbon generation technologies and a number of domestic energy efficiency programmes (Mallaburn and Eyre, 2013). In particular the latter, delivered, paradoxically by the large energy suppliers have been targeted at low income households in order to contribute towards fuel poverty objectives (Probert, 2014).

Two incentives are particularly relevant to small scale renewables. The Feed-In Tariff (FiT) was introduced by the Energy Act 2008 and became operational in April 2010. It guarantees a payment for all electricity generated by approved installers up to 5MW. The scheme provides payments for electricity generated plus additional payments for electricity exported to the grid. Tariffs are to be paid for 20 to 25 years and protected against inflation (Mendonça, Jacobs and Sovacool, 2009). The second is the renewable heat incentive, which came on stream in 2012. This is the first scheme to incentivise renewable heat generation which makes a guaranteed payment for each kWh of renewable heat generated over seven years (Connor et al., 2015).

The purpose of these schemes is to accelerate the adoption of microgeneration technologies. Germany, which adopted a FiTs scheme in 2000, witnessed the rapid diffusion of small scale

renewables (Jacobsson and Lauber, 2006). A similar outcome has occurred in the UK during the period of this research, with the capacity of grid-connected PV in the UK exceeding 7GW by April 2015 (DECC, 2015A). Affordability, security of supply and carbon emissions are key indicators for energy policy; thus the market instruments above had to be devised so that, as far as possible, they did not impact too heavily on the former, whilst delivering an adequate response to energy security and climate change. The next section expands the concepts of indicators in various policy domains which might be used to evaluate the deployments of renewable energy, particularly small scale domestic installations.

1.3 The Impacts of distributed renewables

Modern policy and decision making has adopted the concept of impact assessment (Lyytimäki et al., 2013). The approach requires a number of indicators that can be assessed or measured, and serve as benchmarks for comparative and evaluation purposes. In the context of the drivers for renewable energy, policy and decision makers frequently resort to multiple or composite indicators, in order to facilitate evaluation from a wide variety of stakeholder perspectives. There is a considerable academic discourse pertaining to the sustainability agenda (Chmutina et al. 2013) which proposes a move away from the pure economically grounded "financial bottom line" to a triple bottom line i.e. one which incorporates perspectives on social and environmental as well as the economic (SEE) impacts of socio-technical innovations (Elkington, 1998).

Such an approach requires the development of indicators and agreed methodologies for their assessment across a number of disciplines. This is an emerging research agenda with attempts to harmonise indicators at an EU level (EERA, 2013). This thesis will lean towards the development of a new integrated methodology, rather than to solely add to the body of knowledge on actual indicators, but it is worthwhile to discuss SEE indicators in order to contextualise this research.

1.3.1 Social Indicators

There are a number of studies which evaluate the social impacts of distributed renewable energy technologies (Rogers et al., 2012). Civil society and political campaigners argue that investment in green technologies will create new employment (CACC, 2015), though this, it is argued, should be balanced against job losses in other sectors (Edenhofer et al., 2013). The affordability of fuel to business and householders is an important concern. With numerous claims on the weekly budget, increases in fuel costs can take householders below the minimum income standard with consequences for the wider local economy (Hirsch, 2015). This debate is frequently cast in terms of fuel affordability and the percentage of income spent on fuel as a percent of total household income. A UK household is said to be in fuel poverty if it spends, or would need to spend, 10% or more of its household income on fuel in order to maintain adequate comfort levels (Boardman, 2012). This has been revised to apply to households whose income is 60% or less than the median household income, the so-called low-income high-cost model (Hill, 2012). A consequential issue of energy affordability is ill-health due to poor housing standards partly caused by inadequate heating. Whilst social research such as employment availability and health and wellbeing impacts are vital indicators resources available did not permit their inclusion in this research; a focus on net benefit to household income as a contribution to living standards, and a measure of the potential of renewables to lift people from fuel poverty were achievable indicators.

1.3.2 Economic Indicators

Economic indicators are defined as those which influence investment decisions and drive the diffusion of new technologies. Classical indicators for such purposes are the return on investment which gives a percentage return or profit made on a sum invested, and net present value (NPV), which discounts the lifecycle expenditure and income streams to a value in the present day

(Campoccia et al., 2014). Policy makers are particularly keen to have indicators which facilitate comparisons between technologies so for example the lifecycle or levelised cost of energy discounted to the present day (Darling et al., 2011).

The role of uncertainty in financial modelling is particularly pertinent to investment risk. Using the approaches adopted in this research it was hoped to bring insight into investment decision making. Many of the financial terms can be related and it was recognised that by solving one metric such as NPV, other metrics could easily be derived. In this work therefore a strong emphasis was given to developing metrics based on NPV.

1.3.3 Environmental Indicators

The environment has both material and immaterial meanings; the former can be impacted physically, for example by having artefacts or pollutants introduced into it that effect its behaviour or ecology. The latter may be impacted metaphysically by concepts of ownership, sense of place and aesthetics. There is significant academic research into how renewable technologies change the immaterial qualities of the environment. This is particularly pertinent to gain insight into why people reject or support the energy transition and the adoption of renewables (Devine-Wright, 2008). It is however difficult to translate these concepts into generalizable quantitative or qualitative indicators.

More tangible is the impact of renewable technologies in the displacement of fossil fuel generation and thereby a commensurate reduction in carbon emissions. Indeed this is the single most sought indicator for renewable energy projects, usually presented as compound indicators such carbon emission savings per kWh of useful energy generated, or per unit cost invested. More advanced techniques consider a life cycle analysis which considers, amongst other things, the carbon cost of manufacture, transport, installation and decommissioning to give an estimate of the embedded carbon (Nugent and Sovacool, 2014).

1.4 Modelling under Uncertainty

1.4.1 Uncertainty

Several indicators pursued in this work have been set in the context of the three significant decision domains pertinent to sustainability agenda. As mentioned in section 1.1, this work was motivated by a desire to develop quantitative approaches which endogenise uncertainty, the measurement of which makes recourse to the mathematics of probability (De Finetti, 1974). This has two aspects. Firstly the rigorous use of statistical methods to quantify the uncertainty of a result derived using empirically measured variables. The second starts with a probabilistic perspective on the (input) variables by the use of parametric or empirical probability distribution functions. These are then employed in mathematical models which yield probabilistic results (Savage et al., 2012).

The latter approach endogenises uncertainty and is appropriate for this research where a great number of variables were derived as empirical or parametric probability distributions and subsequently used to deliver probabilistic outputs. This is a significant epistemological approach known as Bayesian statistics which contrasts to the more classical frequentist statistical approaches (Iversen, 1984).

1.4.2 Probabilistic Modelling

Probabilistic modelling is a discipline which has advanced steadily over the last century within several disciplines such as artificial intelligence, decision theory and environmental modelling (Fenton and Neil, 2012). Models use algorithms which employ probabilistic variables in ways which conform to the axioms of probability theory (Fenton and Neil, opt. cit.). Their utility is by virtue of the fact that uncertainty, encoded as probabilities, can be propagated through the model in order to obtain a

measure of the uncertainty of output parameters given one or more probabilistic predictor parameters.

The combination of graph theory with probability theory has led to the development of probabilistic graphical models (PGM) within the Artificial Intelligence and computer science fields which are now finding increasing application in a range of disciplines (Koller and Friedman, 2009). Such methodologies are attractive since they fulfil the dual role of a subjective visualisation of the problem domain using graphs, and the quantitative encoding and propagation of probability through the model.

1.4.3 Towards an integrated modelling paradigm using Bayesian Networks

A key challenge in system modelling is to integrate knowledge domains such that the influence by a parameter in one discipline on a parameter in another can be quantified. The requirement to evaluate simultaneously indicators in social, environmental and economic (SEE) domains is a classic research problem in this regard. Bayesian networks (BN) are one type of PGM which can be constructed to address this type of problem (Koller and Friedman, opt. cit.). The model is constructed using a directed graph which is a collection of nodes to represent system variables linked by directed arcs which denote a relationship between the variables. This is expanded upon in Chapter 3.

BNs are a powerful tool for multidisciplinary modelling because they can be componentised and interfaced together to create a whole system model, a feature known as object orientation (Armstrong, 2006). They have been applied in many fields, including to model a national energy system (Cinar and Kayakutlu, 2010), but not to explore the SEE impacts of community deployed renewables. This research has explored in depth the utility of BNs to investigate this multidisciplinary problem using quantitative empirical and theoretical input parameters and output indicators.

1.5 Research Questions, Aims and Objectives

The motivation for this work has been set in the context of three main themes: the energy transition, the impacts of distributed renewables and modelling under uncertainty. Arising from this is a number of more specific research questions:

1. Is it possible to integrate knowledge across several disciplines in the context of renewable energy deployment in UK communities, particularly social, economic and environmental?
2. Can the uncertainties of solar PV generation, and those of its deployment context, be employed to predict the uncertainties in a number of measured indicators?
3. Is it feasible to create insights into more plausible decision- and policy-support tools which have predictive and diagnostic qualities?
4. Is it possible from this to make some inferences regarding the deployment of PV of use to policy and decision makers?

This leads to a formal research aim and set of objectives for this PhD:

To develop a whole system modelling paradigm that endogenises uncertainties for key performance indicators in the deployment of solar PV in UK communities in order to evaluate its social, environmental and economic impacts.

In order to deliver this aim several specific objectives were established:

1. To develop a representative set of social, economic and environmental parameters to serve as key performance indicators that can be integrated in to a probabilistic model of the whole system.
2. To characterise and model the uncertainty of solar PV yield, electricity self-consumption and electricity exported to the grid when deployed in UK domestic housing stock.
3. To model the contribution of solar PV to domestic electricity consumption.

4. To evaluate the outputs of the models for several distinct geographic areas for the chosen KPIs

1.6 Structure of the Thesis

There are a number of disparate themes which are brought together, with the aim of constructing an integrated probabilistic model, to answer the research questions above.

Chapter 1 [this chapter] sets out the motivation for this research, presenting key themes and, from this, synthesises research questions and formal aims and objectives. Chapter 2 provides a research justification, the purpose of which is to support the argument for a gap in the knowledge addressed by this research and the research questions as posited. A critique of other approaches by which research has been conducted to address these gaps is presented. A review of modelling approaches is conducted which concludes with the choice of the Bayesian Network methodology employed in this research. Chapter 3 presents more details about the methodology. Theory, concepts, and terminology used in Bayesian Networks are discussed, and software tools used are introduced.

Domestic Solar PV deployed to UK communities under the feed-in tariff has been selected as the case study for this research. Chapter 4 provides an overview of the adoption of solar PV in the UK. The geographical unit of analysis for the study is elucidated. A conceptual model for the development of an integrated BN is scoped; four components are identified which are the subject of the following four chapters.

Chapter 5 develops the geographic unit of analysis and presents the acquisition and processing of building stock and socioeconomic data. A BN to model the geography is presented. Chapter 6 considers the performance of domestic Solar PV and proposes a BN to model the system yield based on a number of predictor variables. Chapter 7 presents a BN which models domestic building energy consumption based on acquired data and a statistical analysis. Chapter 8 presents a component

which predicts direct self-consumption of solar PV generated electricity. Chapter 9 discusses the integration of the BN components in chapters 5 to 8 in to a unified model, and presents components which model indicators selected for the study.

Finally chapter 10 presents a summing up of the contribution to knowledge with conclusions and further work.

For those who prefer a graphical, rather than a narrative outline, a graphical model of this thesis is presented in Figure 1-1. The reader can use the colour coding to identify the key components of the thesis. Chapters depicted by yellow nodes are the key thesis elements of introduction, justification, methodology and scoping, with a discussion and conclusion at the end.

The main body of research is represented by the blue and green nodes. The blue nodes are the key research outputs for 4 knowledge domains identified to construct an integrated model. These are referred to in the thesis as the model 'cornerstones'. The objective of each of these pieces of work was to create a BN sub-model for the knowledge domain. Each cornerstone chapter has a similar structure as illustrated in Figure 1-2. Firstly following an introduction, the domain ontology is researched using a review of literature in order to understand the key parameters and their dependencies. The objective is to be able to construct a BN from the position of a 'domain expert'. Data sources are key constraints in modelling of each cornerstone, and their sourcing and processing is discussed in section 3 of each chapter. Finally the BN sub-model is presented in section 4, followed by a discussion and conclusion in section 5.

Lastly the green nodes represent the integrative Chapter 9 and present the creation of the whole-system model and key performance indicator components which the model is designed to inform.

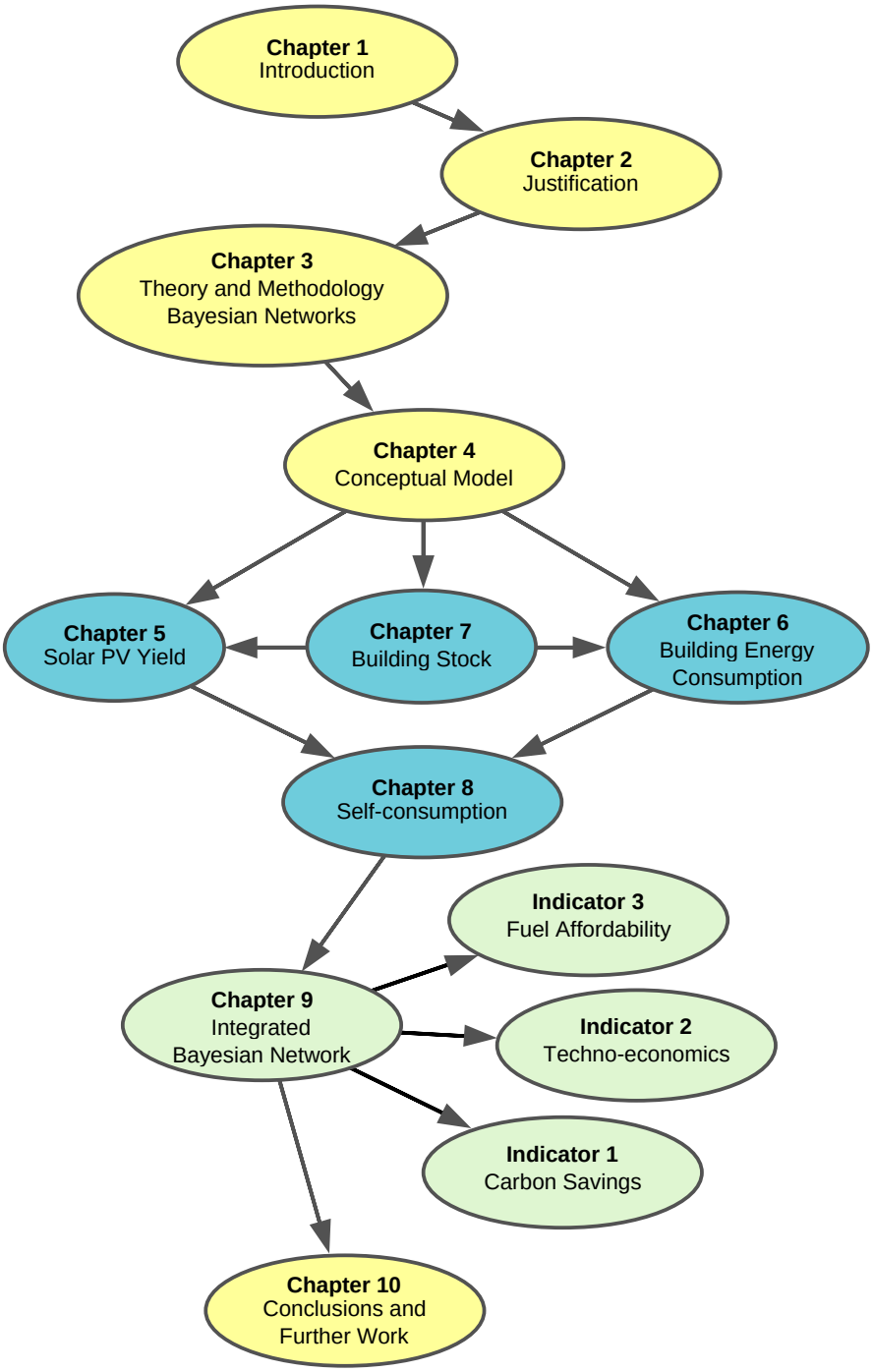


Figure 1-1 Graphical structure of the thesis

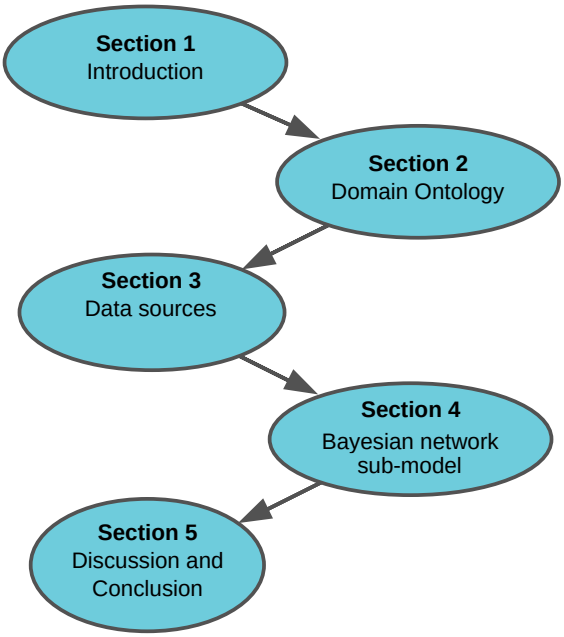


Figure 1-2 Structure of BN-submodel cornerstone chapters 5-8

2 Justification

All models are wrong – George Box

2.1 Introduction

The problem domain - the evaluation of technical solutions proposed, or adopted, to deliver a transition to a low carbon, sustainable energy system – is a prime motivation for this thesis. Above that presented in Chapter 1, it is not the purpose of this Chapter to further justify a research project into low carbon transition; there are enough scientist/engineers, learned scientific bodies, NGOs and whole nations pursuing this agenda to be able to safely, '*stand on the shoulders of giants'* (Cook et al., 2013; Cook and Cowtan, 2015). Indeed, 146 nations of the world are to meet in Paris, and, hopefully, on the 11th December 2015, reach a new international climate agreement (COP21, 2015).

The emphasis in this work, which will be justified in the chapter, is the development of a new modelling approach to a complex multidisciplinary domain with the commensurate and concomitant evaluation of SEE impacts. This has been advocated, by the author, with a view to facilitating deliberative policy and decision making to help satisfy multiple (and often conflicting) stakeholder perspectives. It is also not the prime purpose here to justify research towards sustainability and the benefit(s) of multi-stakeholder perspective decision making, as this is already an active and ongoing research discipline (Burgess et al., 2007).

It is the need for a new modelling approach, and the utility of probabilistic approaches advocated in this thesis, which requires further justification. More specifically, the claim that Bayesian networks offer a promising solution is critically evaluated. To this end, in the next section, the flaws of deterministic modelling in the context of this research topic are explored followed by a discussion on integrated modelling and Bayesian networks.

2.2 The Flaws of Deterministic Modelling

A starting point for the evaluation of the impact of PV is to consider its energy contribution to the built environment, and specifically, the building stock on which it is installed. The research theme is not new; Gadsden et al. (2003) investigated the contribution of PV to the building energy balance using GIS systems, which provided a reduced data set for input into a building physics model (Rylatt et. al. 2003). Both the estimation of PV, and the building energy demand, used deterministic models. Generally, building stock energy modelling is required to estimate baseline energy demand, prior to the implementation of any low carbon intervention, such as PV (Kavgic et al., 2010). 'Bottom-up', as opposed to 'top-down' econometric modelling approaches are suggested as the only way of ascertaining the impact of new technological interventions or policies (Kavgic et al., opt. cit.). Swan and Urgusal (2009) identified the two main approaches: building physics methods which use empirically and scientifically derived relationships to predict building energy demand, and statistical models which utilise data sets of domestic energy consumption data alongside a number of other data points.

In the UK, the building physics models of choice have been from the BREDEM family of estimation tools (Shorrock and Anderson, 1985), in particular the standard assessment procedure (BRE, 2014). A major virtue of these is the simplicity which renders them usable by both non-experts and building energy professionals alike (Kelly et al. 2012). Kavgic et al. (2010), state that the most important shortcoming of such models was the lack of quantification of inherent uncertainties in the data, and that their effects should be investigated through sensitivity analysis. One-at-a-time sensitivity analysis measures the effect of input variations on the predicted model outputs. Uncontroversially (and unsurprisingly) it was found that, *"various input parameters have widely varying effects on the prediction outputs"* (Firth et al. 2010).

Since BREDEM and similar physics models are deterministic, the variability of the input parameters is external to the model, which, even though the sensitivity is measurable, the true variability of the input (and therefore the output) remains unknown. Shipworth (2013) underscores the pitfall of externalising errors when such bottom up models are calibrated using top-down statistical data – all the errors are coerced into a single variable, such as the internal temperature for example. One solution to this is to employ Bayesian calibration to adjust the value of deterministic model parameters to produce a closer approximation of the model estimates to the observed data and produce realistic distributions for variables (Booth and Choudhary, 2013). However, this approach, whilst producing endogenised distributions for input variables, is computationally intensive and only suitable for single buildings (Heo et al. 2012).

In the construction sector it has become apparent that many new buildings, designed to achieve high levels of environmental performance, fail to deliver substantial reductions in energy use (Menezes et al., 2012; Doylend, 2015). This performance gap is also observed in domestic properties which have been assessed using site survey data. Thus deterministic models frequently deliver results which do not concur with empirical data, and, any single estimate of total energy consumption is subject to considerable inaccuracy (Hughes et al., 2013). Furthermore, research has shown that the survey data itself may be subject to significant uncertainties (Tronchin and Fabbri, 2012). As well as technical factors, Tweed (2003) highlighted the socio-technical factors – the interaction of occupants with the technical system to influence the energy consumption. De Wilde (2014), in his analysis of the differences between model-estimated and observed energy measurements alluded to the notion of a probabilistic problem, with a wide range of predictor parameters. In essence, outside of the confines of the '*laboratory*' there is, in a real-world context, a complex array of aleatory and epistemic uncertainties which deterministic models externalise (Shipworth, 2013).

2.3 Towards Integrated Modelling

Since the energy consumption characteristics of the residential sector are complex and inter-related, comprehensive models are needed to assess the techno-economic impacts of adopting energy efficiency and renewable energy technologies suitable for residential applications (Swan and Urgusal, 2009). One shortcoming of energy demand building stock models is that final energy demand is the only output; whereas other important factors may demand a more comprehensive modelling approach which evaluates, for example, carbon emissions (Heeren et al., 2013). In recent years life-cycle assessments, which measure impacts over the entire lifecycle, and consideration of embodied energy of a product, have become popular (Cetiner and Edis, 2014). Such modelling requirements suggest a need for a more integrative modelling paradigm which can knit disciplines together. This, coupled with increasing demands for decision and policy oriented tools, has given rise to the new discipline of integrated environmental modelling (Laniak et al., 2013). Such a paradigm has a focus on multiple stakeholder perspectives and transdisciplinary research in order to model socio-techno-economic systems. Five main modelling approaches are used: systems dynamics, Bayesian networks, coupled component models, agent-based models and knowledge-based models (also referred to as expert systems) (Kelly (Letcher) et al., 2013). Of these, only Bayesian networks are explicitly able to deal with uncertainty in the interpretation of data and may even have elements of other modelling approaches, such as expert systems (Lecklin et al., 2011) or agent modelling, when dealing with specific interactions between system components (Lehikoinen et al., 2013). These unique properties, which render Bayesian networks a powerful tool for the integrated modelling of complex multi-domain problems, are summarised in Table 2-1.

Table 2-1 Properties of Bayesian networks for integrated environmental modelling

Endogenisation of uncertainty
Transdisciplinarity
Incorporate expert opinion/qualitative and/or quantitative data
Decision, management and policy applications

2.4 Probabilistic Modelling using Bayesian Networks

The preceding two sections have sought to highlight two issues; namely uncertainties are externalised, and trans-disciplinary problems are difficult to solve using deterministic modelling approaches. Bayesian networks are proposed as a solution when trans-disciplinarity and uncertainty are key issues (Fenton and Neil, 2012; Jensen and Nielsen, 2007; Pearl, 1985; Smith, 2010). This is because, unlike other modelling approaches, a Bayesian network uses probabilistic relationships between input and output parameters, rather than a deterministic relationship (Chapter 3). Thus, in the former, the value of an output variable is equal to a probability vector, conditional on the value of its input variables. The relationship between them is represented by a conditional probability distribution. In the latter, input variables are entered into a series of equations, either physics based or empirically determined, in order to yield the value of the output variable. Table 2-2 contrasts these equalities. This is discussed in greater depth in Chapter 3.

Table 2-2 Comparison of probabilistic and determinist relationships between an output variable A, and two input variables, B and C.

$\boldsymbol{P}(\boldsymbol{A}) = P(\boldsymbol{A}\|B, C)$	Probabilistic
$A(A, B) = f(B, C)$	Deterministic

Since their more academic and theoretical beginnings, Bayesian networks have recently found increasing real world application (Pourret et al., 2008). Examples include water quality studies, where the perspectives of multiple stakeholders are required to model the causes and solutions to pollution (Borsuk et al., 2004), particularly involving participatory methods (Carmona et al., 2013). A tool to aid the diagnosis of component defects in complex manufacturing systems was developed by Przytula and Thompson (2000). Complex social, economic and environmental impacts of industry were modelled to create a triple bottom line BN model. This served as an adaptable tool to enable informed assessment, dialogue and negotiation of strategies at a global level (Buys et al., 2014). In energy research disciplines Telenko and Seepersad (2014) modeled energy consumption of lightweight vehicles, and the inter-annual variability of wind speed has been modelled as a feasibility analysis for the installation of wind turbines (Carta et al., 2011). Cost effective greenhouse gas emissions reduction in the British agricultural sector has been modelled, thus enabling farmers to make better land-use, fertilizer and renewable energy investment decisions (Pérez-Miñana et al., 2012). BNs have also been used in risk assessment for industrial process control and enabled reasoning under uncertainty, presenting users with recommended corrective actions along with explanations of the root causes of problems (Weidl, et al., 2005).

The benefits of BN modelling approaches are manifold. Molina et al. (2013) state that BNs are powerful tools for assessing the interests of the multiple stakeholders. Duespohl et al. (2012) commended the participatory characteristic in model elucidation and claimed, due to many favourable characteristics, that BNs have the potential to become a core method of transdisciplinary knowledge integration. BN modeling can facilitate the integration of information from diverse sources (Johnson and Mengersen, 2011) and are gaining popularity due to their mathematically coherent framework, with the explicit accounting for uncertainty (Uusitalo, 2007). Model frameworks to facilitate the development of a BN in a multi-expert and multi-field domain serve as a powerful communication tool with stakeholders and collaborators (Johnson et al., 2010). Substantial

insight into many real-life problems can be imparted due to the graphical representation of model structures and probabilistic outputs, which are useful in communicating both theories and results to colleagues, students, and decision-makers (Uusitalo, 2007).

2.5 Bayesian Networks applied to the problem domain

The previous section presented a small sample of the real-world applications for which BN models have been developed, and the advantages that researchers and practitioners claim to benefit from. However, the approach has seen little application in the building energy and applied urban energy communities despite this area being a complex, multidisciplinary multi-scale system beset with uncertainties (Rowley et al., 2015).

The complex topic of sustainability and decarbonisation of the energy system in the built environment has naturally generated myriad interpretations of the problem domain and proposed modelling solutions in recent years. In a recent review of over 200 studies, six key areas of practice were identified: technology design, building design, urban climate, systems design, policy assessment, and, land use and transportation modelling (Keirstead et al., 2012). It was suggested that despite the number of approaches, four common challenges prevailed: complexity, data availability and uncertainty, model integration, and policy relevance. From the previous section, it is clear that BNs can provide some answers to all of these challenges; no mention was made by Keirstead et al. however, of this method, or probabilistic graphical models In general. A Monte Carlo approach however, which can accommodate input parameter variability, has been applied by Keirstead and Calderon (2012; 2014) to an urban area of Newcastle, UK, to model the uncertainty in urban energy and carbon models following various low carbon measures, including demand side measures and microgeneration technologies . Monte-Carlo simulation however, whilst exploring a broad parameter space, is still a deterministic modelling approach using mathematical equalities. It

does not capture the extent of probabilistic relationships, both direct and indirect, which a BN modelling approach can do.

Keirstead and Schulz (2010) have advocated the need for a suite of analytical tools that go to the local urban context to assess the unique local energy needs. They label these tools as part of a nascent field of urban energy and climate modelling, and have applied these ideas to the city of Newcastle. However, the lack of an endogenised uncertainty modelling practice provides a strong justification for this research, and the methodology proposed. Others have also taken up this challenge; BNs have been deployed to investigate the uncertainties in the performance of solar thermal systems in urban domestic contexts (Thirkhill, 2015) and exploring the performance gap in new non-domestic properties (Doylend, 2015).

2.6 Conclusion

In this chapter, the application of BNs to the research problem has been argued. Two needs were addressed – firstly, the need to endogenise uncertainty, and secondly to create an integrative model.

The benefits of a probabilistic approach have been juxtaposed with deterministic models, and it has been argued that both the above requirements are satisfied by the BN approach.

Finally, it has been shown that within the nascent area of urban energy modelling, BN can help satisfy all of the main challenges identified by the research community (Table 2-1). To reiterate, the research aim of this project is:

> *To develop a whole system modelling paradigm which endogenises uncertainties for key performance indicators (KPIs) in the deployment of solar PV in UK communities in order to evaluate its social, environmental and economic impacts.*

This has been shown to be an important gap in the academic literature, which this work aims to contribute to. The epigraph to this chapter states all models are wrong (Box, 1976). This pertains to both the structure of a model, and the input data. Whilst a probabilistic model can endogenise uncertainty, there is still a requirement to build a model which is, as far as possible, a good representation of the real world. Chapter 3 takes a deeper look at the theory and practice of BN to achieve this.

3 Bayesian Networks - The Theory and Method

The best we can do is to be less wrong

3.1 Introduction

In Chapter 2, it is proposed that statistical graphical models are an efficacious method of eliciting new knowledge in a complex interdisciplinary knowledge domain characterised by uncertainty. Specifically, Bayesian Networks (BNs) are reviewed and discussed as a method of choice to encapsulate this knowledge and facilitate inference and decision making. In this chapter the foundational statistical modelling (Section 3.2) and graph theory (Section 3.3.) are expounded. BNs as a general class of statistical models are theoretically explained in Section 3.4 and their construction and utilisation is described. In Section 3.5, object oriented BN are introduced for working with large and complex multidisciplinary domains. Section 3.6 looks at the practical use BN.

3.2 Statistical Modelling

Increasingly, statistical modelling approaches are used in many disciplines to explicate complex multi-parametric domains. A model establishes an ontological boundary and describes the relationships between the model's constituent entities. Unlike deterministic models where the relationships are described by mathematical equations (either physics based or empirically derived), in statistical models the relationship between variables is probabilistic. In such models uncertainty of a parameter is propagated through the model's interconnected parameters to augment or diminish the uncertainty of another. In the following sections the fundamentals of probability and graphical models are introduced which underpin these methods.

3.2.1 Probability

Probability is fundamentally a measure of uncertainty, or, since there is usually more interest in the probability that an event will happen, of certainty. A classical interpretation of probability is a measure of the number of times a unique event occurs, when compared with the number of times other mutually exclusive, or *disjoint*, events occur. Consider, for example, an experiment performed many times where each time the outcome of interest is one of three possible, but mutually exclusive, events A, B or C and the number of occurrences of an event is denoted by n. *The probability of event* A is given by Equation 3-1.

$$P(A) = \frac{n_A}{n_A + n_B + n_C} \qquad \text{Equation 3-1}$$

Such a definition of probability, only calculable when numerous repeat experiments are practical, is referred to as the frequentist interpretation. Alternatively a measure of probability may be required in situation where it is not feasible, or it is impractical (e.g. prohibitively expensive), to perform repeat experiments. For example, if one were outside an unfamiliar restaurant and wanted to know the probability of being served a good steak. In this situation one might solicit the subjective views of others and posit a 40% chance of being served a good steak. This is a Bayesian interpretation of probability and refers to the degree of belief about events in world. In the more positivist scientific context of the 20th century such admission of subjectivism into probability theory was, for many years, considered controversial and counter to the scientific method (Vallverdú, 2003).

Regardless of which interpretation is employed, a Bayesian or frequentist probability measure obeys the three fundamental axioms of probability calculus (Kolmogorov, 1933). Firstly a probability cannot be greater than unity or less than zero (Equation 3-2). If it is unity then the event is certain to happen; zero means the event will never occur.

$$0 \leq P(A) \leq 1 \qquad \text{Equation 3-2}$$

In a sample space S, consisting of a finite number of elementary events there is unit probability that one of the elementary events will occur (Equation 3-3).

$$P(S) = 1 \qquad \text{Equation 3-3}$$

And, where events are mutually exclusive, or *disjoint*, the total probability of one or other of the events occurring is given by the sum of their individual probabilities (Equation 3-4)

$$P(A \cup B) = P(A) + P(B) \qquad \text{Equation 3-4}$$

For *joint events*, that is events which can both occur (are not mutually exclusive) it can be shown, given the above axioms, that the probability of any one or both events occurring is given by Equation 3-5.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{Equation 3-5}$$

Where $A \cap B$ represents the intersection between A and B which is the event that both A and B occur.

3.2.2 Conditional Probability

Conditional probability is defined as the probability of an event given that another event has happened. The probability of the event A, given that B has occurred is expressed as $P(A|B)$ and is defined by Equation 3-6.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{Equation 3-6}$$

This is known as the fundamental rule of conditional probability, designated as a fourth axiom by Finetti (1937), is often expressed as a function of the joint probability as in Equation 3-7.

$$P(A \cap B) = P(A|B) \cdot P(B) \qquad \text{Equation 3-7}$$

The fundamental rule can be extended for multiple joint events as in Equation 3-8.

$$\begin{aligned} P(A \cap B \cap C) &= P(A|(B \cap C) \cdot P(B \cap C) \\ &= P(A|(B \cap C) \cdot P(B|C) \cdot \mathrm{P(C)} \end{aligned} \qquad \text{Equation 3-8}$$

The final form of Equation 3-8 is known as the chain rule and is expressed in Equation 3-9 for any number of joint events n. The rule is very important for factorising probability calculus in Bayesian Networks.

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i| \cap_{j=1}^{i-1} A_j) \qquad \text{Equation 3-9}$$

Bayesian probability asserts that all probability is conditional upon the context in which measurements are made or experiments executed. As discussed above, proponents of frequentist definitions of probability need to account for subjective assumptions inherent in the data (Koch, 2007).

3.2.3 Bayes Rule

Since, from the axioms of probability $A \cap B \equiv B \cap A$, and from the fundamental rule the relationship between conditional probabilities is given by Equation 3-10.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$ Equation 3-10

This equation, known as Bayes rule, published posthumously in a narrative form in 1763 (Price, 1763), allows the updating of a belief to give a *posterior* probability, $\mathrm{P(A|B)}$, given some new information, B, and having previously known the prior probability, $P(A)$ the likelihood of B given A, $P(B|A)$, and the probability of B.

3.2.4 Independence

Two events are independent of each other if our belief in the occurrence of one event is not influenced by the occurrence of another event. So if the probability of event A is not affected by the occurrence of event B, then their independence is demonstrated by Equation 3-11.

$$P(A|B) = P(A)$$ Equation 3-11

Using the fundamental rule Equation 3-12 follows.

$$P(A \cap B) = P(A) \cdot P(B)$$ Equation 3-12

3.2.5 Random Variables

Thus far probability has been considered in terms of events. For the purposes of modelling a problem domain, a set of discrete random variables is considered. A discrete random variable is one which can take one of a finite number of discrete values or disjoint *states*. The variable has a probability of being in each state. This gives rise to a probability distribution for the random variable which is correctly termed the *probability mass function* (PMF). For a variable $\boldsymbol{A}$, which has n discrete states, a_1, a_2, a_3... a_n, the PMF is a set represented as a set of probabilities (Equation 3-13).

$$\mathrm{P}(\mathbf{A}) \equiv \{\mathrm{p}(\mathbf{A} = \mathrm{a_i})\ \forall \mathrm{i} = 1,2,3 \ldots \mathrm{n}\} \qquad \text{Equation 3-13}$$

A continuous random variable is one which can have value on a continuous range. The probability of the value at any point on the range of possible values is described by a probability density function (PDF). The probability of having a specific value is infinitesimally small and therefore the PDF, $f(\boldsymbol{A})$ is integrated over a finite range Equation 3-14.

$$p(i \leq \boldsymbol{A} \leq j) = \int_i^j f(\boldsymbol{A}) \qquad \text{Equation 3-14}$$

The probability theory discussed above for events is applicable to variables whereby probability calculus is applied individually to each discrete state. Thus in agreement with the axioms of probability, the sum of probabilities over all possible states a_i of $\boldsymbol{A}$ is unity (Equation 3-15).

$$\sum_{i=1}^{n} p(\boldsymbol{A} = a_i) = 1 \qquad \text{Equation 3-15}$$

Similarly for a continuous variable Equation 3-16 applies.

$$\int_{-\infty}^{\infty} f(\boldsymbol{A})\ = 1 \qquad \text{Equation 3-16}$$

If $A = \{a_1, a_2 \cdots a_n\}$ is conditional on $B = \{b_1, b_2 \cdots b_m\}$, then to calculate the conditional probability $P(\boldsymbol{A}|\boldsymbol{B})$, each state of A must be conditioned separately on each state of B to generate $n \cdot m$ conditional probabilities such that for each state b_j the sum of the probabilities of $\boldsymbol{A}$ is one in accordance with the axioms of probability (Equation 3-17).

$$\forall\ \mathrm{b_j} \in \mathbf{B}, \sum_{i=1}^{n} p\big(\boldsymbol{A} = a_i | \mathbf{B} = \mathrm{b_j}\big) = 1 \qquad \text{Equation 3-17}$$

For variables the conditional probability denoted by $P(\boldsymbol{A}|\boldsymbol{B})$ is shorthand for a *conditional probability table* (CPT). For example if $n = 2\ and\ m = 3$ such a table is represented by Equation 3-18:

$$\begin{array}{cccc} & \boldsymbol{b_1} & \boldsymbol{b_2} & \boldsymbol{b_3} \\ \boldsymbol{a_1} & p(a_1|b_1) & p(a_1|b_2) & p(a_1|b_3) \\ \boldsymbol{a_2} & p(a_2|b_1) & p(a_2|b_2) & p(a_2|b_3) \end{array}$$

Equation 3-18

Similarly, the joint probability distribution, expressed as $P(\boldsymbol{A}, \boldsymbol{B})$[1] can be obtained using the fundamental rule (Equation 3-7) for each possible state combination, generating $n \cdot m$ joint probabilities (Equation 3-19).

$$\begin{array}{ccccc} & & \boldsymbol{b_1} & \boldsymbol{b_2} & \boldsymbol{b_3} \\ \boldsymbol{a_1} & p(a_1, b_1) = & p(a_1|b_1) \cdot p(b_1) & p(a_1, b_2) = p(a_1|b_2) \cdot p(b_2) & p(a_1, b_3) = p(a_1|b_3) \cdot p(b_3) \\ \boldsymbol{a_2} & p(a_2, b_1) = & p(a_2|b_1) \cdot p(b_1) & p(a_2, b_2) = p(a_2|b_2) \cdot p(b_2) & p(a_2, b_3) = p(a_2|b_3) \cdot p(b_3) \end{array}$$

Equation 3-19

Each state combination can be regarded as an elementary event. In order to satisfy the axioms of probability the sum of all the (joint) probability of each elementary event must sum to one as exemplified by Equation 3-20.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} P\left(\boldsymbol{A} = a_i, \mathbf{B} = b_j\right) = 1$$

Equation 3-20

The chain rule, as a logical extension of the fundamental rule, equally applies to discrete variables. Thus the joint probability distribution for any number of discrete variables is represented by Equation 3-21.

[1] For variables and their states it is common practice to express the joint probability as $P(\boldsymbol{A}, \boldsymbol{B})$ rather than $P(\boldsymbol{A} \cap \boldsymbol{B})$ as is the practice for events, though the two forms are equivalent.

$$P(\cap_{i=1}^{n} \boldsymbol{A}_1) = \prod_{i=1}^{n} P(\boldsymbol{A}_i | \cap_{j}^{i-1} \boldsymbol{A}_j) \qquad \text{Equation 3-21}$$

Each variable $\boldsymbol{A}$ represents a set of discrete states. The total number of elementary events, N_e, is the product of the cardinality, $n\{\boldsymbol{A}\}$, of each variable $\boldsymbol{A}$'s set of discrete states (Equation 3-22).

$$N_e = \prod_{i=1}^{n} n\{\boldsymbol{A_i}\} \qquad \text{Equation 3-22}$$

For example if there are three variables $\boldsymbol{A,B,C}$ then the chain rule is used to calculate the **JPD (**Equation 3-23**)**.

$$\mathrm{P}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = P(\boldsymbol{A}|\boldsymbol{B}, \boldsymbol{C}) \cdot \ P(\boldsymbol{B}|\boldsymbol{C}) \cdot \mathrm{P}(\mathbf{C}) \qquad \text{Equation 3-23}$$

If each variable as 10 states the **JPD** has 10^3 values.

3.2.6 Marginalisation

Consider the **JPD** where one might want to calculate the total probability of each value $P(\boldsymbol{A}=a)$ by removing the variable b (Equation 3-24).

	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	$\boldsymbol{b_3}$	$P(\boldsymbol{A})\ marginalised$
$\boldsymbol{a_1}$	$p(a_1, b_1)$	$p(a_1, b_2)$	$p(a_1, b_3)$	$p(a_1)$
$\boldsymbol{a_2}$	$p(a_2, b_1)$	$p(a_2, b_2)$	$p(a_2, b_3)$	$p(a_2)$

Equation 3-24

The probability distribution of $\boldsymbol{A}$ is given by Equation 3-25.

$$P(\boldsymbol{A}) = \sum_{i=1}^{m} P(\boldsymbol{A}, b_i) \qquad \text{Equation 3-25}$$

This process is termed marginalisation - so called since the process can be considered as adding up all the values in single row or column of the **JPD** and writing the sum in the margin. Marginalisation

can be carried out in a multivariate **JPD** to extract the marginal distribution of any one variable (Equation 3-26).

$$P(\boldsymbol{A_j}) = \sum_{\boldsymbol{A_i}\forall i \neq j} P(\cap_i \boldsymbol{A_i} \forall i) \qquad \text{Equation 3-26}$$

Marginalisation can be used to calculate the PMF for each variable in the **JPD**. For large numbers of variables this is processing intensive. The next section looks at how multivariate systems can be encoded and probability calculus made more tractable using graph theory and probabilistic graphical models.

3.3 Graph Theory

A graph, G (Equation 3-27), in this context is a set of vertices (nodes) V (Equation 3-28), and a set of edges (connectors) E (Equation 3-29), which are used to model the relationships between pairs of objects, usually variables, in a collection.

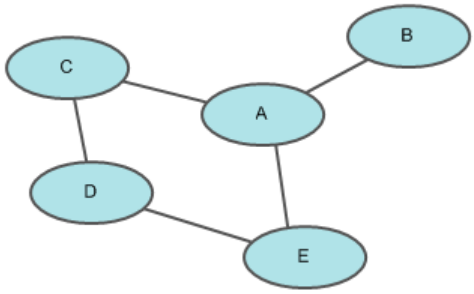
$$G = \{V, E\} \qquad \text{Equation 3-27}$$

$$V = \{V_1, V_{2,} V_3 \cdots V_n\} \qquad \text{Equation 3-28}$$

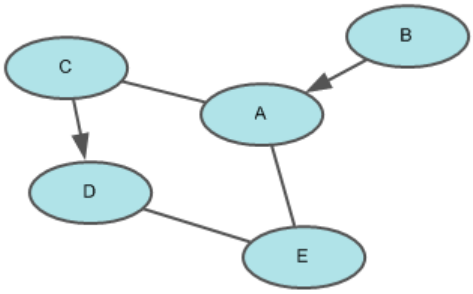
$$E = \{E_1, E_{2,} E_3 \cdots E_m\} \qquad \text{Equation 3-29}$$

Figure 3-1 shows types of graphical model (Koller and Friedman, 2009) each with 5 nodes and 5 edges. Graphical models may have undirected or directed edges. The latter serve to indicate a hierarchical relationship between the connected vertices. A third type of graph, consisting of a mixture of directed and undirected edges is called a chain graph. Two types of directed graphs may be defined. Firstly, by following directed edges it may be possible to arrive back at the starting

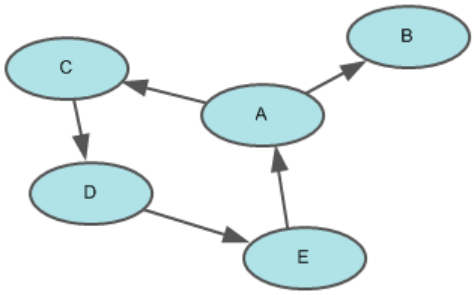
vertex; the edges trace a *cycle*. Such a graph is a directed cyclic graph. A graph constructed so there are no cycles, is a *directed acyclic graph* (DAG).



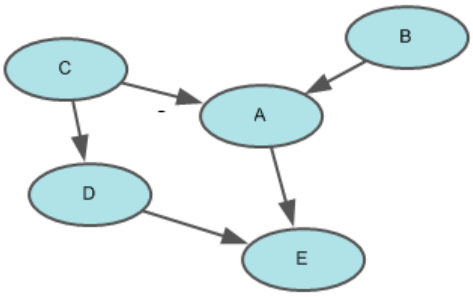
A. Undirected graph



B. Chain Graph



C. Directed cyclic graph



C. Directed acyclic graph

After Koller and Friedman, 2009

Figure 3-1 Different type of model represented by graphs or graphoids

3.3.1 Probabilistic Graphical Models

A probabilistic graphical model (PGM) is constructed using a graph as a conceptual model for the visual representation of a complex system. The vertices or nodes of the graph conveniently represent the parameters or variables used to describe the modelled system, and the edges between the vertices represent relationships or associations between the connected nodes. The graph communicates the structure of the problem domain to communities of experts and stakeholders.

The utility of graphs is further augmented by the encoding of uncertainty for each variable represented by a vertex. A software language can encode both the graph and mathematical representations of uncertainty to enable probability calculus over the PGM. There are several useful and convenient elements to such an encoding, discussed next.

Elementary events

In general, for a model with n random variables, V_1, V_2, V_3,..., V_n, a set of atomic states (v_1, v_2, v_3,..., v_n), forms an elementary event for the modelled system. The total number of elementary events, N_e, is the product of the cardinality, $n\{V\}$, of each variable's set of discrete states Equation 3-30).

$$N_e = \prod_{i=1}^{n} n\{\boldsymbol{V_i}\} \qquad \text{Equation 3-30}$$

This defines the number of exhaustive elementary events in the sample space S, used in classical statistics. Each elementary event has a probability as defined in the JPD.

Joint Probability Distribution

Each elementary event has a joint probability $p(v_1, v_2 v_3, \cdots v_n)$. For all events this is known as the joint probability distribution (JPD). The JPD over all of the graph's variables has N_e values. For models with large N_e naïve storage of the JPD creates a large storage and eventual data processing requirement rendering probability calculus intractable for software systems.

Marginalisation

A graph can be used to store the marginalised probabilities of variables of interest or these can be calculated from the JPD using the chain rule.

Dependence Relationships

The edges in the model encode dependence relationships between connected variables. The semantic of this relationship depends on the model type and may imply a causal relationship or a correlation (see 3.4.4).

Propagation

The model's calculus can be used to propagate the probabilities between nodes through connecting edges using a calculus and the edge semantic implicit in the model type. The probability distribution of a target variable of interest can, in this way, be updated as a new distribution is learnt for another variable in the model. For example the PMF for a node may be replaced for hard evidence, which is to say, the variable is instantiated or set to a particular member of its set of states. How this propagates through the model and its impact on the probabilities of other variables renders a PGM as very valuable tool for prediction and diagnostics in complex systems.

Two common types of PGM are (i) a Markov random field, which is a model based on an undirected graph (Kindermann and Snell, 1980), and (ii) a Bayesian Network, which is modelled over a DAG. This work, as discussed in Chapter 2 uses Bayesian Networks as the tool to model uncertainty in the problem domain of this work. The theory and calculus used in these is discussed next.

3.4 Bayesian Networks

The term Bayesian Network (BN), coined by Pearl (1985), is used since Bayesian probability calculus based on conditioning underpins the algorithms used in the propagation of probabilities in this type of network. In this section the definition of, theory behind and the practical construction of BNs is discussed.

3.4.1 Definition

A Bayesian Network (BN) is a specific type of PGM which is encoded over a DAG in which the vertices correspond to random variables and directed edges represent direct dependencies between them. A directed edge from a node $\boldsymbol{A}$ to $\boldsymbol{B}$ implies that $\boldsymbol{A}$ has a dependence on $\boldsymbol{B}$. $\boldsymbol{A}$ is termed a parent of $\boldsymbol{B}$ and $\boldsymbol{B}$ *is a child of* $\boldsymbol{A}$.

Each node is encoded with a prior probability distribution. In the case of root nodes - one without any parents - this is the variable's PMF. Because the model is derived within a specific context this is referred to as the marginal probability where conditional dependencies on variables outside of the scope of the model are deemed to have been marginalised. For each child node the prior probability distribution is given by a CPT in which the variable's probability conditional on all the parent variables is given. The strength of a relationship is thus encoded in the CPT.

For each node, a PMF of the encoded variable can be expressed, directly for root nodes, or derived through marginalisation for child nodes. The prior distributions of each node are therefore readily available. This renders the BN as a powerful knowledge base for the problem domain where the relationship between nodes and their prior probability distributions can be visualised and communicated to model users. A BN is regarded as having two components, the graph, which is a qualitative component which is a subjective conceptual model of the world being modelled, and a quantitative component which is the probability distribution data entered for each node. The latter can be based on subjective probabilities as in the Bayesian probability interpretation, or derive from quantitative empirical data using a frequentist interpretation.

3.4.2 The Joint Probability Distribution of a Bayesian Network

As discussed above, the joint probability distribution for a multivariate system of random variables is readily calculated using the chain rule. It can be shown that for a BN $\{\boldsymbol{V}, \boldsymbol{E}\}$ in which the vertices

correspond to a set of variables, $\{V_1, V_2, \cdots, V_n\}$, this factorises to the much simpler form in Equation 3-31 where $\boldsymbol{\pi}_{V_i}$ represents the set of variables which are parents of the variable V_i in the DAG.

$$P(V_1, V_2, \cdots, V_n) = \prod_i P(V_i|\boldsymbol{\pi}_{V_i}) \qquad \text{Equation 3-31}$$

This factorisation of the chain rule, which allows only the parent nodes of each node to be considered, simplifies the processing requirements for probability calculus over the BN. For example, for a BN represented by the DAG in Figure 3-1C, the JPD is simplified to Equation 3-32.

$$P(A, B, C, D, E) = P(A|B, C)P(B)P(C)P(D|C)P(D|C)P(E|A, D) \qquad \text{Equation 3-32}$$

It is frequently noted that a BN reveals as much about variables that are not connected as those that are (Smith, 2010). Variables for which there are no connecting directed edges are conditionally independent of each other, given other network variables and this has been subjectively declared in the construction of the model. Variables which are connected are directly dependent. Variables may be indirectly connected and may be independent (d-separated) or dependent depending on the instantiation (setting of evidence) of the intermediary variables and relative configuration. It is first and foremost the missing edges signifying conditional independency assumptions which allow the major simplification of probability calculus over a BN (Pearl, 1985).

3.4.3 Propagation of Probabilities

The dependency relationships allow the propagation of probabilities. It is this property which makes a BN a powerful tool for reasoning and inference making in complex multivariate systems. If evidence is learnt about a particular variable, its PMF can be adjusted to reflect the observed evidence. The chain rule can be used to recalculate the new JPD and thus readjust all the

distributions for each variable in the network. The new JPD, and the distribution for each variable, are called posterior distributions.

In practice the naïve use of the chain rule in this manner is not computationally tractable for large networks with granular variable state distributions. Such operations have been found to be NP-Hard, meaning that the number of arithmetic operations increases exponentially as the number of variables increases (Dagum, 1993). A key achievement of the early pioneers of Bayesian Networks was the factorisation of conditional probabilities encoded within a Bayesian Network to enable the development of efficient algorithms for the propagation of probabilities without having to calculate the entire JPD (Pearl, 1986; Lauritzen and Spiegelhalter 1988; Jensen et al. 1990; Shenoy and Shafer, 1990). This has made the updating of BNs computationally tractable.

3.4.4 Constructing Bayesian Networks

There is no formal or exact method to construct a BN. Two main approaches are documented in the literature: firstly a network can be constructed using domain knowledge to establish dependencies between variables, or the network can be learnt or discovered from domain data (Daly et al., 2011).

The latter approach involves computer algorithms to construct dependency relationships between the variables in a tabulated empirical dataset (Neapolitan, 2004). In practice a large number of samples must be available in order to have a high confidence in the elicited graph structure since the number of potential graphs varies exponentially on the number of variables. This approach is most often used in data mining applications where patterns are sought in vast datasets. The algorithms require a convergence on a proposed graph using maximum likelihood estimation. This ‘best fit’ process also learns the node **CPT**s in the process of discovery of the graphical structure.

In the absence of large datasets, or on the understanding that model discovery may not yield intuitive models, model construction using domain knowledge or experts is preferred. Domain

knowledge of experts is used to infer the required parameters and the dependency relationships between them – often referred to as a causal map or web (Marcot et al., 2006; Şahin et al., 2006; Nadkarni and Shenoy, 1999; 2004). Hybrid methods incorporating both data and prior domain knowledge have been employed (Heckerman et al., 1995; Zhou et al., 2014).

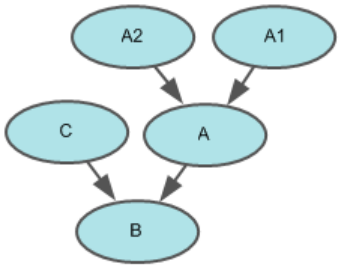
There are three steps involved in the causal map approach:

i. Determine the number and meanings of variables in the domain to be modelled
ii. Determine the causal relations between variables
iii. Determine the conditional probability tables for each variable dependent on its parents

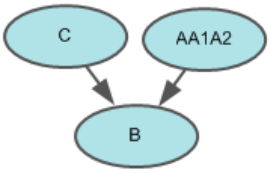
The following paragraphs discuss each of these in turn.

Determining the Variables

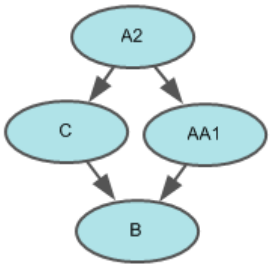
Usual practice is to elicit the required variables from domain experts and stakeholders. This can be via formal facilitated causal mapping workshops, from the domain literature, or individual expert groups. The objective of modelling is to abstract real world concepts into the model. It is feasible to start with a deterministic extreme where the microscopic scale is abstracted in all its causal detail. This may be unworkable since the model has to be populated with quantitative data associated with each and every variable. Moving up a level of abstraction, in effect aggregating variables, *introduces probabilities to summarise omitted variables* (Pearl, 2000). Figure 3-2A and Figure 3-2B show this abstraction with two variables $A1$ and $A2$ subsumed into A. A variable so subsumed must not, however, be required influence another variable. In Figure 3-2C, $A2$ is retained since it separately influences C. If the abstraction is taken too far (for example representing Figure 3-2A by just one random variable) the properties of causation may be lost in the probability distributions which summarise aggregated variables.



A. Microscopic model



B. Macroscopic model with A1 and A2 subsumed into A



C. A2 retained to model influence on C

Figure 3-2 Different type of model represented by graphs or graphoids

The object is to create a model which will help users of the model understand the problem domain. Implicit in this, as with all sense making models it is required to be parsimonious, that is not to include variables in which there is no interest (or resources) and lie outside the problem boundary. Common practice is to elicit variables of interest and then map related variables which have a dependency relationship.

Causal relations between variables

The process of constructing the graph, once the required variables have been determined, is that of creating direct dependencies represented by directed edges. The directed edges are often presumed to represent causal relationships; however, this need not be the case. The construct of causation is a philosophically contested concept. It is well know that *the rooster's crowing does not make the sun rise and association does not prove a cause and effect relationship* (Pearl, 2000).

In defining a causal relationship the first difficulty lies in overcoming preferred human models of abductive reasoning. Thus in is Figure 3-3A, the sour taste results in the inference that the milk is off; the flow of reasoning is from the sour taste to the milk being off. But this is not the cause of the milk going off which is due to the activity of bacteria producing lactic acid. The milk's being off caused the

sour taste so the direction of causation is counter to that used in reasoning as in Figure 3-3B which is often the case with goal oriented humans (Fenton and Neil, 2012).

A. Abductive reasoning B. Deductive reasoning

Figure 3-3 Abductive and deductive reasoning

Fenton and Neil, (2012) relax causality to mean 'strengthens belief in' which in the Bayesian interpretation means an increased probability that an event has occurred. This idea can usefully convey both abductive and deductive inference.

The second area of difficulty is where the variables of interest, or which are accessible to the problem are correlated, but do not have a causal relationship, for example if they share a common cause. In Figure 3-4A the off milk causes a stomach upset and a sour taste; the latter was not the cause of the former. In the absence of an acidity tester, Figure 3-4B shows how the two accessible variables are used to model the same problem. *Sour Taste* has become a proxy for *Milk Off* variable and an observational dependency of *Stomach Upset* on *Sour Taste* can be modelled. It not too remote from conventional human reasoning to argue the sour taste caused the stomach upset.

Fenton and Neil (2012) argue that the best strategy when constructing BNs is to use the arrows in direction of causation, though this is not always straightforward. For example Smoking causes cancer, but often the data collected is that of cancer suffered by smokers. Mathematically it can be showed that the direction is equivalent, all that is needed is to reverse the CPTs so that instead of *(Cancer|Smoking)* the table gives *(Smoking|Cancer)*.

A. B.

Figure 3-4 Common cause variable

Determining the conditional probability tables

It can be seen how the construction of the model can be quite subjective with regard to the variables chosen to be included, the levels of abstraction and the direction of dependency. Another factor may be the availability and the choice of data to include - the quantitative component.

Using Idioms

The use of common structural forms or *Idioms* as a method to construct BNs was suggested by Fenton and Neil (2012). Idioms reflect common patterns of human reasoning prevalent in real world problems. The method has been since taken up by other researchers in the field and was used in this research to construct BNs. Four idioms, discussed below, are proposed:

i. *Cause consequence*
ii. *Measurement idiom*
iii. *Definitional/synthesis idiom*
iv. *Inductive idiom*

Cause consequence idiom

This idiom models the cause effect dependency between variables and often involves a temporal relationship - 'effect follows cause'. The variables are often at opposite ends of a process, with the parent node preceding or contemporaneous with the child node. The process is not represented itself by anything other than the child node's CPT.

Measurement idiom

This is used when a variable represents the actual value of an attribute (a measurement) but is modified by the known assessment accuracy of the measurement implement or method. This delivers the assessed value. The idiom is employed for modelling test processes with a specific accuracy which yields the final result.

Definitional/Synthesis Idiom

This structure which combines many nodes into one is found to be very common. It does not represent a causal association but is one of definition. It might be used to create a categorical indicator which is a composite of two or more parent nodes. For example a variable *safety* may be defined in terms of variables the *frequency* and *severity* of an incident (Figure 3-5). The child node may be calculated using a deterministic function or axiomatic relations between modelled ideas.

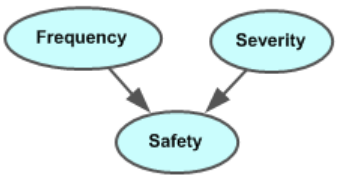


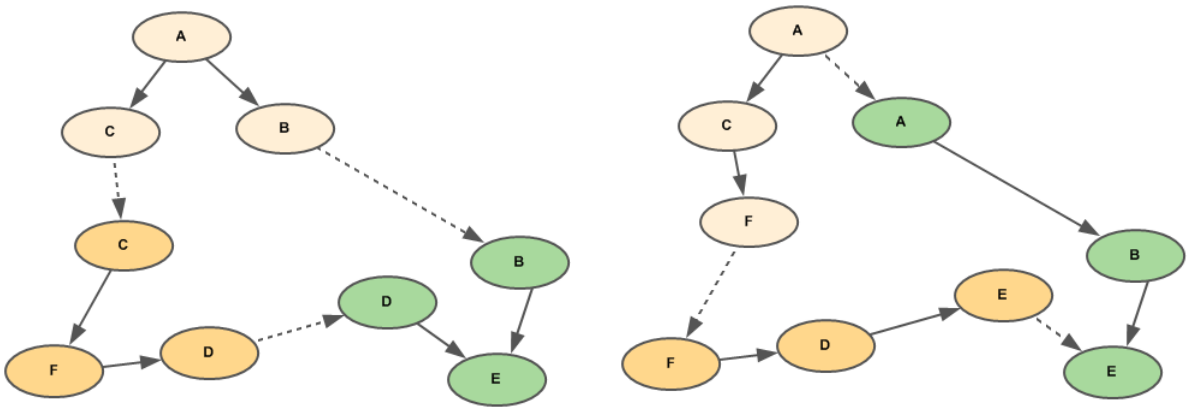
Figure 3-5 The Definitional or synthesis idiom

Inductive idiom

When a network models one idea it may be required to make an inference about another similar idea. The first node and a node to represent the similarity of the second to the first are used as parents of the node about which it is desired to make an inference. The semantic is quite subtle; the first idea does not cause the second, it is simply used inductively to make predictions about it.

3.5 Object Orientated Bayesian Networks

Individual domains of large multi-domain problems can be modelled by separate BN components or sub-models, which are linked by shared or common variables (Fenton and Neil, 2012). This approach was termed object oriented Bayesian network (Koller and Pfeffer, 1997) by analogy to object oriented software design. This approach is compatible with the research objective to where knowledge from different domains is knitted together to create a transdisciplinary knowledge representation. Each sub-model can be regarded as an autonomous BN, each of which requires its own inputs and has one or more outputs (Figure 3-6A). This shows three objects, each with its own colour scheme, ABC, CDF and BDE. For an object to be autonomous it must be capable of being abstracted from the wider BN model, and still functioning as a unit. This requires the duplication of variables in each sub-model which are then joined to represent an interface between them (Armstrong, 2006). Thus in Figure 3-6, the object ABC exposes parameters C and B. C is an input parameter for CDF and B is an input parameter for BDE. The objects can be joined together through this interface. The BN semantic whereby C in CDF is a child of C in ABC does not in practice apply thus dashed lines are used to indicate that the two nodes labelled C have an exact equivalence.



The dashed lines represent interfaces between compatibile variables. A and B represent two possible configurations of variables within components to represent the same model (see text)

A. **B.**

Figure 3-6 Separate components or sub-models of an object oriented bayesian network

An object and those parameters it encapsulates is a subjective choice. Thus the three objects in Figure 3-6A could be reconfigured as Figure 3-6B. The three colours are used to denote three autonomous objects, ACF, ABE and DEF, but this time the interfaces are between different variables. However, the network has the same joint probability distribution as given by Chain Rule.

The concept of the sub-system boundaries becomes important when considering the internal dependencies of the model and when more than one variable is used in an interface (Figure 3-7). If D is dependent on C i.e. there is a conditional probability relationship such that, $\mathrm{P(D|C)} \neq \mathrm{(P(D)}$, then there is an option to encode the conditional probability relationship $\mathrm{P(D|C)}$ or $\mathrm{P(D'|C')}$ or both.

This problem arises in Chapters 6 and 7 and has been termed the 'dependency ownership dilemma' of the object oriented model.

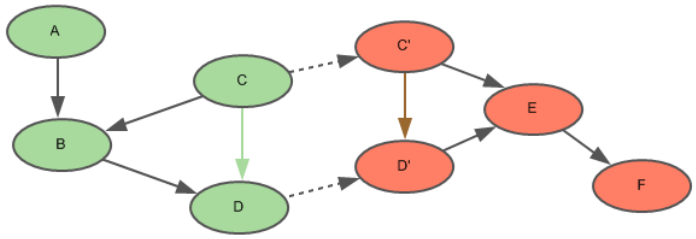


Figure 3-7 Demonstrating the choice of where a dependency should be encoded

3.6 Using a Bayesian Network

A fully functional working network is used by applying evidence to one or more nodes of interest. The literature has a rather obfuscated array of definitions and here the recently proposed clarified definitions are used (Mrad, 2015). The first type of evidence is hard evidence or a hard finding. This represents the instantiation of a variable X, to given state x, given evidence e, such that $P(X = x|e) = 1$. Once this is set, all the probabilities of other variable are adjusted to their posterior distributions.

The second type of evidence is uncertain evidence. There are two types of uncertain evidence. The first of these is likelihood evidence which models the case where the observation is uncertain. It is specified as a belief in the current observation on a variable. The second type is probabilistic evidence which invokes a new probability distribution on a variable. This may be of two types, fixed, which cannot be modified by evidence applied to other nodes, and non-fixed which can be modified by further evidence on any variable in the model.

Not all BN software enables probabilistic evidence to be entered with ease and in this work it has been achieved using a proxy node to update dependent variables with new distributions based on the selected census area (see Chapter 4). More frequently hard evidence is applied to one or more nodes of interest in order to instantiate the variable to a particular value of interest and then observe the posterior distributions on target nodes.

3.7 Norsys Netica

Networks in this research were developed in Norsys Netica which is a popular Bayesian network development software created by the commercial company Norsys (Norsys 1995). It is free for small networks (12 nodes or less) and has a reasonable single user price of £250 which allows the development of larger models.

It has the following redeeming features for BN researchers:

- Intuitive GUI for creating networks with nodes and arcs.
- Tabular CPTs allowing easy manual entry of probabilities.
- Able to incorporate deterministic functions.
- In built repertoire of statistical functions.
- Able to import case file data from Excel spreadsheets or Access databases for CPT learning
- Ability to dynamically link with excel for display and reporting purposes.
- Ability to learn CPTs using a choice of three different algorithms with imported data.
- Able to learn simple naïve Bayes nets using TAN learning.
- Inexpensive $250 for single user price.
- Completes Bayesian inferences very quickly compared to other software on the market.
- Rapid entry of hard evidence and likelihood evidence.

A major weakness is that it does not easily allow probabilistic evidence to be entered. Netica has three algorithms for learning CPS, these are the count algorithm, expectation maximisation (EM), and gradient learning. Only the count algorithm was used in this research since the other techniques are useful when there is missing data which was never the case.

3.7.1 CPT Count Learning

The count algorithm is the simplest method of learning CPTs from data held in a spreadsheet or database. It generates CPTs automatically by counting the number of occurrences for each of the child node states for each combination of parent node states; a frequency table is generated and then normalised to generate a CPT for the child node. Each occurrence of a combination of states

represents one counted case. A weighting column can be used to increment the counting of cases by a number other than one, automatically.

3.7.2 TAN Learning

Netica can learn a tree augmented naïve (TAN) Bayes net from a case file (Friedman et al. 1997). A node is selected as the classifier and the strength of the relationships between the node and all other nodes is learnt using the case file data. In addition Netica will add relationships on between other nodes if they are detected.

TAN learning is very powerful for discovering the strength of influence between a classifier and all the variables in the network.

3.7.3 Deterministic Nodes

A node in Netica can have its conditional probability tables determined by an equation. As an example Figure 3-8 shows a deterministic node $C = f(A, B)$. Netica is furnished with a whole set of standard functions such as common probability density functions. A CPT for the deterministic node is calculate using a Monte Carlo simulation; each value of the input nodes is sampled a pre-set number of times (e.g. 1000) and the deterministic node calculated according the encoded function. A frequency table is constructed on the fly during the simulation and a CPT generated.

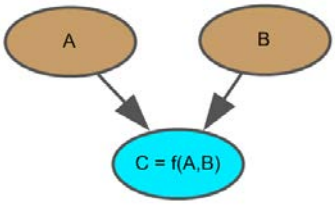


Figure 3-8 A deterministic node with two input variables, A and B.

This is a powerful method of integrating deterministic relationships in to the BN which will respond to evidence on other nodes in the network.

3.8 Summary

Statistics and graph theory, which underpin probabilistic graphical modelling and BN have been briefly presented; the reader however will appreciate this has only scraped the surface of a large knowledge base on the theory, building and use of BN. Some of the key features of Netica, which was used in this work, have been described – CPT learning using the counting algorithm, TAN learning, the use of deterministic nodes, and a general discussion of some key issue concerning the building of BN.

Some of these methods will be applied in subsequent chapters to build BN using causal mapping techniques. Not all the direct dependencies represented by edges, however, are causal in nature, and thus it should not be assumed proposed BN are causal networks.

The next chapter will set out the preliminary context for this research and begin to work towards a causal map from which to create a BN using domain expertise.

4 Conceptual Model

4.1 Introduction

Chapter 2 highlighted the lack of integrated approaches for the assessment of SEE impacts of renewable energy. The Bayesian network was proposed as a tool which fills this gap. Theory and methodology for constructing BNs were outlined in the previous chapter. The first stage is to construct a 'causal map', a diagram similar to a DAG (Chapter 3), which defines parameters of interest as nodes and directed edges to show directions of influence. Also discussed was the technique of deconstructing a large BN, particularly one which models multiple knowledge domains, into components to create an object oriented Bayesian network (OOBN).

The aim of this chapter is to describe the construction of a causal map to serve as a conceptual model for such an OOBN. This is supported by research outputs from an analysis of the diffusion of community deployed renewables, particularly in the context of the Feed-in tariff introduced in April 2010. As well as the creation of a conceptual model, there are two further key objectives. The first is to define a unit of analysis used for the assessment of SEE impacts and the second is to articulate how case studies for which to source quantitative data for the model were sourced.

The structure of the chapter is as follows. Section 4.2 looks at the diffusion of renewable technologies under the FiT. This is analysed according to the type of technology and who is adopting the technology. This provides insights in to the scoping of an integrated model. Section 4.3 takes the SEE impacts presented in the literature review and develops ideas about KPIs to assess these impacts. This influences the construction of the model which is expanded in section 4.4. The scope is narrowed in section 4.5 to consider solar PV as the technology in the domestic context and section 4.6 considers the acquisition of data from representative case studies in defined geographic areas.

4.2 Diffusion of Renewables under the Feed-in Tariff

Several technologies have been subsidised under the UK FiT introduced in April 2010. This section uses the official register of renewable installations supported under this regime to analyse the rate of diffusion of renewable technologies. This purpose is to gain insight in to the requirements for integrated modelling.

4.2.1 Fits Register

The regulatory authority for the electricity market, OFGEM, is the responsible body for the administration of the FiT Register which lists all installations eligible to receive the FiT (OFGEM, 2013). An anonymised reduced dataset is published every quarter which details the date, technology type, capacity and locality, as well as other meta-data of all installations on the register. A data dictionary[2] for the dataset is shown in Table 4.1. The final version of the register used in this work was published in April 2014 (OFGEM, 2013).

[2] The term data dictionary is used to document a data tables column names and their semantic descriptions and is produced for all datasets. Sample data for each dataset used in this work are included in the appendices.

Table 4-1 Data dictionary for the OFGEM Feed-in Tariff Register

Field	Description
FIT ID	Unique identifier in the OFGEM register
Postcode District	Outward postcode
Technology Type	Technology type PV, Wind, Hydro etc.
Installed Capacity (kW)	The boiler plate generating capacity of the installation
Declared Net Capacity (kW)	Capacity for which FiTs are claimed
Application Date	Data Fits were applied for
Commissioned Date	Data installation was commissioned by approved agent
Export Status Type	Export deemed or measured
Tariff Code	Determines the tariff the installation is eligible for
Description	Description of the tariff
Installation Type	Community, domestic, commercial or industrial
Country Name	England, Northern Ireland, Scotland or Wales
Local Authority	Principle local authority
Government Office Region	Name of region (England only)
Accreditation No	MCS number
Supply MPAN No (first 2 digits)	Electricity meter number
LSOA Code	Unique ID of lower super output area

4.2.2 Diffusion of Renewable Technologies

An analysis of the register shows a marked increase in the uptake of smaller scale electricity generating renewables in the UK since 2010. This has been reported previously (Leicester et al., 2011). Between 2010 and 2014 the cumulative installed capacity is dominated by Solar PV (Figure 4.1), at over 2GW, which compares with 215MW for wind, and significantly less (68MW) anaerobic digestion and (46MW) for micro-hydro. Micro-CHP has only 500kW of installed capacity. The total capacity and number of installations at March 31st 2014 for each technology is shown in Table 4.2.

The dominance of Solar PV is also illustrated by the number of installations, which has risen to over 460,000. There were 5359 wind turbines on the register and for the other technologies, only a few 100s.

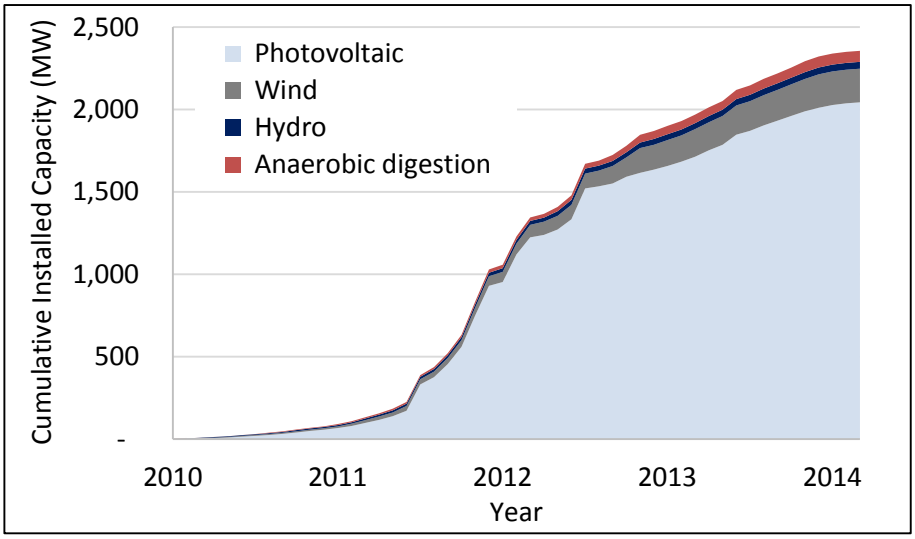


Figure 4-1 Cumulative installed capacity of the main technologies supported by the FiT

Table 4-2 Installed Capacity and Number of installations at 31st March 2014

Technology	Installed Capacity (MW)	Number of Installations	Average Installation Capacity (kW)
Anaerobic digestion	68	84	810
Hydro	46	452	102
Micro CHP	0.5	477	1
Photovoltaic	2,056	464,522	4
Wind	215	5,359	40
Total	**2,386**	**470,894**	**5**

The number of installations illustrates a large number of discrete market actors each assigned to a particular sector. In this thesis a sector is referred to as an adopter vector. Figure 4-2 shows for each technology the percentage installed capacity attributed to each adopter vector. Table 4-3 gives the average capacity of each technology within each vector.

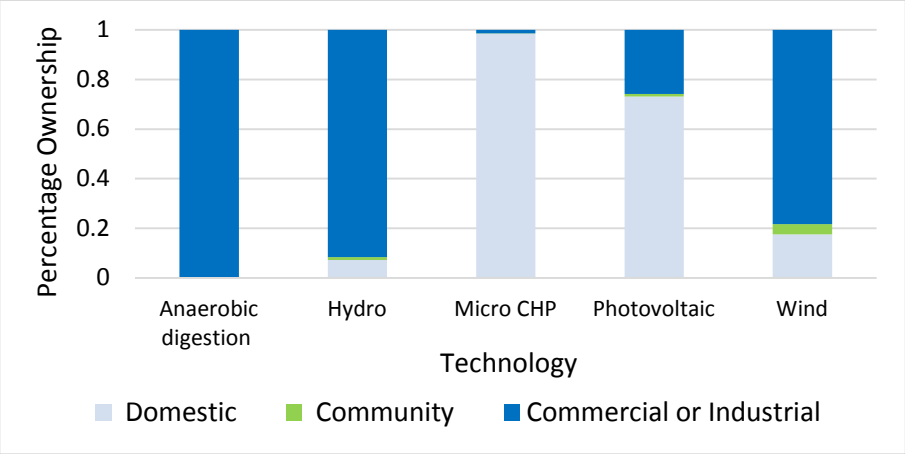


Figure 4-2 Percentage capacity installed by market sectors

It is observed that larger capacity technologies such as anaerobic digestion, hydro and wind, are predominantly in the commercial and industrial adopter vectors. In contrast, small capacity technologies like CHP, typically 1kW, and Solar PV, 3.3 kW, are predominantly within the domestic adopter vector.

Table 4-3 Average capacity of technologies by market sector (kW)

Technology	Domestic	Community	Commercial or Industrial
Anaerobic digestion	4.0		819.3
Hydro	12.8	27.9	249.0
Micro CHP	1.0	1.0	1.2
Photovoltaic	3.3	11.4	41.0
Wind	10.0	61.1	116.8

4.2.3 A Socio-economic perspective on the diffusion of renewables

In the previous section an analysis of the diffusion of renewables in the UK by technology type was examined. In this subsection, insight into the requirements of an integrated model was gained using geographic and socio-economic lenses. For this work the lower super output area[3] (LSOA) was selected as the geographic unit of analysis. LSOA are derived by the Office of National Statistics

[3] Sometimes referred to as the lower layer super output area (LLSOA).

(ONS) from UK census output areas and comprise, on average, 672 dwellings and 1614 residents[4]. Social geographers use algorithms to create socio-economically homogeneous output areas constrained by the population criteria and the need to be coterminous with district or unitary local authority level administrative areas (Martin et al., 2001).

A number of aggregated UK statistics are released at the LSOA level which ensures pockets of deprivation are captured and risk of disclosure of personal data is minimised. A widely used composite index presented at LSOA level is the index of multiple deprivation (IMD) (DCLG, 2010). This combines several domain deprivation indices such as household income, employment, health and disability, education, skills and training, housing, crime and the living environment. The IMD is widely used for the distributional impact assessment of policy and targeted interventions. In this context distributional refers to differing impacts of a policy based on baseline spatially resolved socio-economic conditions in order to highlight iniquitous outcomes. IMD was therefore was selected as a parameter for a socio-economic analysis of the diffusion of renewables.

An analysis of an early FiT register using the 2007 IMD was published in 2011 (Leicester et al., 2011). DECC subsequently performed a similar analysis to assess the impact of the FiT policy (DECC, 2012A). An updated version of the IMD was used to update these results using 4 years of FiT register data up to March 2014. The LSOA code column was used to perform a one-to-many left outer join between the IMD data and Fit register[5]. The IMD scores for each LSOA were used to calculate the IMD decile (10 being the most deprived). The rural urban classification for small area geographies was used to

[4] These figures are based on own analysis of the ONS LSOA population estimates. The literature rarely quotes the variability of these means which is represented by a standard deviation of 131 for households and 303 for residents.

[5] A one to many join is a technical database term to mean that matching field values from two tables are used to join the data. If the join is one to many then there may be many matching records (rows) on the right hand table. Left outer means that if no matching rows are found in the right hand table, the rows in the left table are still included, but will have empty values for the right hand table.

apply a rurality classification (Bibby and Shepherd, 2004) to each LSOA. In order to check for population effects the ONS population, dwelling and age profile estimates for each LSOA (ONS, 2011) were added to the dataset. For the purposes of later work ESRI[6] polygon data to represent the boundaries of each LSOA in the graphical information system (GIS) was included in the dataset (see Chapter 7). The data dictionary for the resultant dataset is shown in Table 4-4.

Table 4-4 Data dictionary for the derived LSOA dataset

Field	**Description**
LSOA Code	Unique ID
LSOA Name	Name
Residents	Total population
Household residents	Population in domestic dwellings
Communal residents	Population in communal dwellings (care homes, prisons etc.)
Households	Number of households
IMD Decile	Calculated decile from IMD
Rurality	Rural urban classification code
21-64 year olds	Number of 21-64 year olds
Shape data	ESRI Shape data for GIS

Figure 4-3 shows the total rates of adoption for the domestic adopter vector for solar and wind technology for each IMD decile for England and Wales. Normalisation by the LSOA population, and number of households yielded comparable results. This shows that above an IMD of 5 the rate of adoption of solar PV is approximately half the rate than in less deprived areas with an IMD 5 or less. The picture for wind turbines in contrast shows a peek in adoption rates at a median IMD decile, tailing off significantly at the extremities. It is clear that the adoption of renewable technologies shows a significant dependency on IMD.

[6] This is a proprietary format of the Environmental Systems Research Institute, which is now a recognised standard (ESRI, 1998).

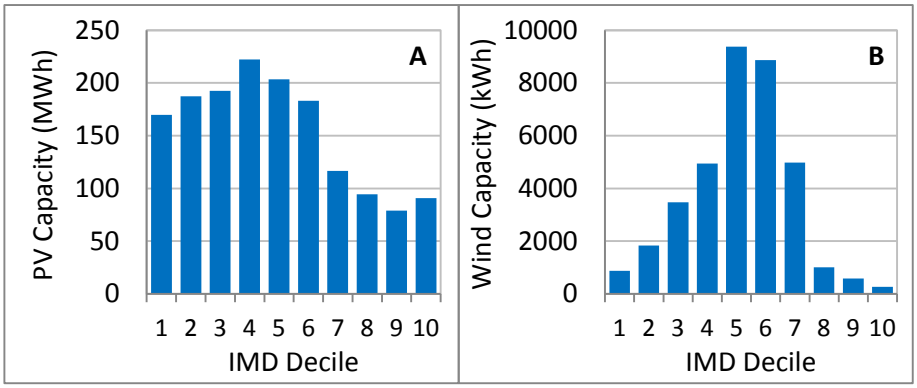


Figure 4-3 Capacity of domestic solar PV (A) and wind energy installations (B) installed in English census areas (LSOA) segmented by deciles of the index of multiple deprivation

Figure 4-4 shows the installed capacity of wind and PV technologies segmented by the rurality classification and normalised to population density. For both solar PV and wind there is a greater rate of adoption (i.e. increase in installed capacity per head) where populations are sparse (i.e. more rural), than in urban areas. This is more marked for wind energy.

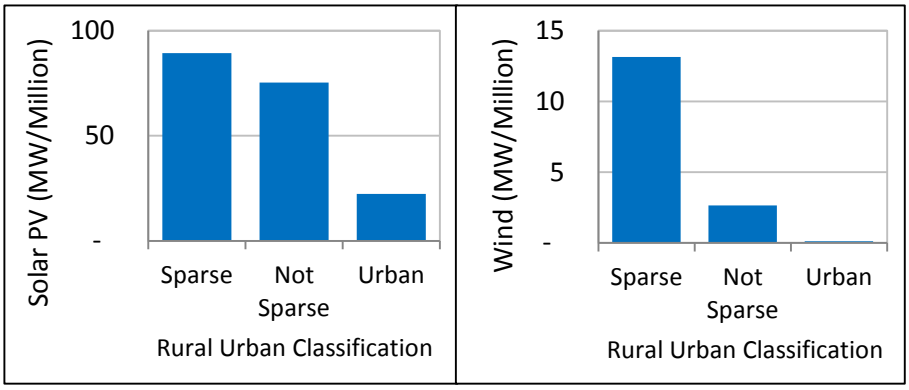


Figure 4-4 Installed capacity of PV by LSOA rurality classification per million population

If the absolute installed capacity is examined, however, the picture is very different. Figure 4-5, shows that wind technology is more likely to be installed in 'not sparse' areas and PV is far more likely to be installed in an urban setting due to relative proportions of population inhabiting areas with a particular rurality classification;. 1.6%, 16.5% and 81.9% of LSOA are classified as sparse, not sparse, and urban respectively.

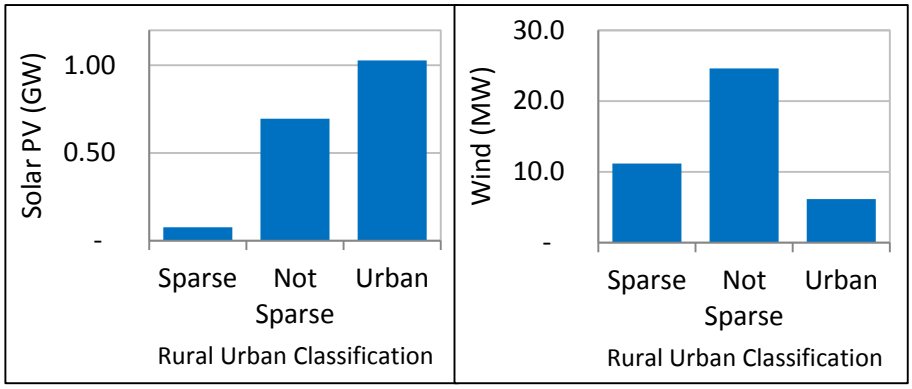


Figure 4-5 Absolute installation capacities of wind and solar PV

Other Fit supported technologies have also been investigated using this approach. These technologies have greater site specificity and so are located in proximity to favourable sites. Of course this is apparent for wind for which adoption is more rapid in sparsely populated areas and installations are less likely to be found in urban areas.

This section has highlighted that geo-spatial and socio-economic factors are highly significant in determining the rate of adoption of renewable technologies. The aim of this research is not to investigate barriers to adoption though it is clear the measures of deprivation and geography are serving as predictors of the rate of adoption due to the barriers of affordability and site specificity. For example low public acceptance of wind has played a role in adoption rates of wind in particular (Devine-Wright, 2010).

Special Case of Solar PV

The high adoption rate and low site specificity of solar PV affords a detailed socio-economic analysis. A count of the number of LSOAs cross-tabulated by their IMD and the banded number of installations is shown in Table 4-5. This demonstrates a propensity for large penetrations of solar PV

within an LSOA to have a high IMD; 71% of the 41 LSOAs with over 100 installations have an IMD of 8 or higher (shown in red type in the table). This is counter to the observation of lower diffusion rates in LSOA with high IMD. This is hypothesised as being driven by an agency[7] other than domestic actors, even though these clustered solar PV installations are registered as domestic with OFGEM. This phenomenon was identified with there being a number of purposeful community energy projects targeted at social housing or low income communities (Leicester et al, 2011).

Table 4-5 Count of LSOA cross-tabulated by the IMD and banded count of installations.

	IMD Decile									
Number of installations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0	1	1	1	0	1	2	1	1	6
1 - 25	2863	2760	2729	2667	2732	2831	2949	2974	2963	2780
26 - 50	308	386	394	437	364	277	180	139	140	172
51 - 75	14	29	49	69	66	58	33	39	30	77
76 - 100	1	1	6	10	16	11	9	7	10	37
101 - 125	0	1	0	0	4	4	3	**4**	**3**	**10**
126 - 150	0	0	0	0	0	0	0	**2**	0	**5**
151 - 175	0	0	0	0	0	0	0	0	**2**	**2**
176 - 200	0	0	0	0	0	0	0	0	**1**	0

4.2.4 Summary

The analysis of the OFGEM fits register and the socio-economic data afforded by the specially prepared LSOA dataset presents a matrix of options for a contemporary study of the SEE impacts of distributed renewables. There are five technology vectors and three adopter vectors each of which is contextualised by a range of socio-economic and spatial factors which also affect the rate of diffusion of the technology. Once in situ, it can be hypothesised that their SEE impacts will also vary (Leicester

[7] The term agency here refers to the sociological concept of the power or capacity of an agent (person or some other entity) to act.

et al, 2011). Predictors of the rate of diffusion are not in the scope of the research objectives. However this process provided insight into the highly varied contexts for impact assessment.

A valid question is whether an integrated model to explore social and economic impacts can be technology and/or adopter vector agnostic. Before answering this it is necessary to further develop the components of the model. To this end the next section considers the assessment of impacts using quantifiable indicators which the model must be able to report on.

4.3 Assessment of the Impacts of Distributed Renewables

This section discusses the selection of KPIs with which to assess SEE impacts for distributed renewables. This is commensurate with approaches to data collection for their quantification and seeks to answer the question as to whether the differing adopter vectors have a bearing on the scope of the model due to differing data requirements.

Table 4-6 Broad impact domains under the SEE sustainability framework

	Impact
Social	Fuel Affordability Aesthetics Community Cohesion Energy Attitudes and Behaviours Social Equality Employment
Environmental	Pollution Resource Depletion Biosphere impact
Economic	Energy Security Energy Resilience Competitiveness Return on Investment Growth

Table 4-6 presents several impact domains for renewable energy under the SEE sustainability framework. Impacts themselves are generally described in subjective terms. Any formal impact assessment requires the measurement of key performance indicators discussed in the next section.

4.3.1 Key Performance Indicators

A key performance indicator is a quantitative or subjective measure with which to objectively assess a tangible outcome (Deakin, 2012). It has been asserted, in Chapter 3, that as well as frequentist probabilities derived from quantitative data, a BN is able to incorporate subjective probabilities[8]. In theory, therefore, qualitative parameters can be integrated in to the model as long as subjective probabilities can be derived. This is an intensive undertaking which uses either expert opinion or an interpretive methodology demanding primary data. Resources did not permit this approach.

Incorporating even just one indicator per impact domain in Table 4-6 would escalate this research into an extended multi-disciplinary research programme. It was expedient, therefore, to select several impact domains, each of which presented readily quantifiable KPIs and closely relate to Government energy policy objectives. These are shown in Table 4-7.

Table 4-7 Selected impact domains under the SEE sustainability framework

Impact	**Indicators**
Fuel Affordability	Spending on fuel Percentage spending on fuel (fuel poverty)
Pollution	Carbon reduction
Return on Investment	Income Generated Discounted cash flow

[8] The Bayesian statistician would argue that frequentist probabilities have a subjective element which pertain to the assumptions in data collection, sample size etc.

The next section considers the measurement of these KPIs for renewable deployments in each of the adopter vectors.

4.3.2 Measurement of Key Performance Indicators

Consideration is given here to how data could be collected to measure or predict the indicators in Table 4-7 for each of the adopter vectors and to any vector specific approaches required.

Commercial and community renewable energy projects are delivered by registered companies, charities or community interest companies and as such are generally required to provide regulatory information (company reports and accounts) which provide a source of primary data about such projects including, potentially, renewable energy generation yield, carbon reduction and financial rates of returns for investors. A number of researchers have gathered large quantities of data on such projects by extensive engagement with a large number of practitioners (van der Schoor and Scholtens, 2015; Seyfang et al. 2014; Walker and Cass, 2007). In order to acquire sufficient data for a quantitative modelling approach a large number of representative projects would have to be surveyed. This is further complicated in the case where an entity were responsible for two or more renewable energy installations since regulatory information commonly presents aggregated information concerning all the responsible entity's activities.

Less tractable is the one-to-many relationship between the generating technology and its individual stakeholders. Data concerning the generating technology can be forthcoming from regulatory documents but no examples of reports could be found which presented a probabilistic distributional analysis of individual the stakeholder benefits – only a mean financial rate of return was available. Thus the distribution of socio-economic benefits accrued by individuals or households would be inaccessible without recourse to stakeholder surveys.

A focus on the domestic adopters of consumer oriented renewable generation technologies provides a more generalizable context whereby the energy generation technology is deployed in a domestic unit with a one to one, or at most one to several[9] relationship with stakeholders. The acquisition of data presents the challenge of acquiring generation data alongside socio-economic data for individual households. Generation data for FiT supported renewables has not been made publically available in the timeframe of this research. However many domestic users have been willing to share their generation data or samples can be obtained from available datasets (See Chapter 6). In most cases the socio-economic context of such data generation data is not available directly, necessitating recourse to surveys or modelling.

4.3.3 Summary

Each adopter vector is defined by either a one-to-one (for domestic), or a one-to-many (community and commercial) relationship between the energy system and stakeholders. This is modelled in Figure 4-6 where key components are represented by a UML class. The deployed energy system is a top-level class which is instantiated as a container object for a single chosen technology object, and one or multiple stakeholder objects. This difference in cardinality[10] between the components has ramifications for modelling the deployed energy system since the latter one-to-many relationship is inherently more complex. It has been argued that although OOBN is suited to modelling instantiated objects, it is unable to model multiple instantiations of the same object (Howard and Stumptner, 2009) suggesting that the development of a model which is sector agnostic is challenging.

[9] This would be the case where the energy technology were installed on a rented property where the owner is distinct from the occupants.

[10] This term, common in data modelling representations such as relational database design structures and unified modelling language (UML), means the number of entities in the relationship.

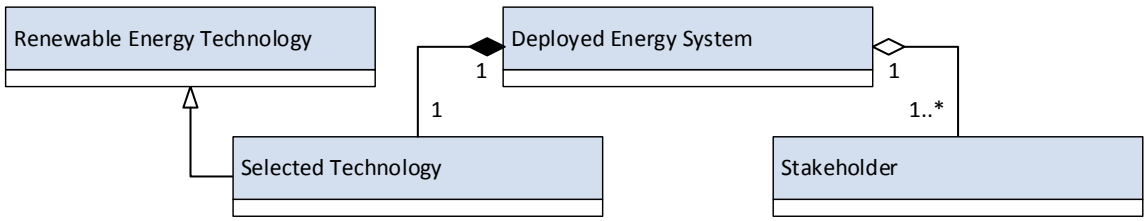


Figure 4-6 UML Object Class Diagram for a deployed energy system

The collection of socio-economic data for multiple stakeholders associated with a single deployed energy system, and for single stakeholders associated with a domestic energy system, presents distinct challenges. In the latter case socio-economic parameters are more predictable from the geographic context. In the former case the technology and stakeholders are not necessarily co-located. Indeed individual stakeholders may be geographically dispersed and therefore socio-economic parameters are less predictable from a geographic context. The only option would be to access sources of actual stakeholder data or collect such data directly by survey methods.

This discussion highlights the notion that the energy technology can be abstracted as separate component of the deployed energy system. In the UML diagram, Figure 4-6, this is emphasised by the representation of the selected technology as a class which inherits from a generic renewable energy technology. The model structure can therefore be considered to be technology agnostic whereby the chosen technology is substitutable in a true object orientated design pattern.

The ideas presented in this chapter were developed in order to propose a heuristic model to underpin an OOBN. At this juncture in the research the focus was given to the domestic vector – thus requiring the modelling of the one-to-one pattern only. This is because the context of the stakeholder is known with greater precision than the one-to-many pattern. This enables the geographic context to be used as a predictor of socio-economics parameters. The conceptual model for the domestic deployed renewable energy is further developed in the next section.

4.4 A Conceptual Model

Object oriented Bayesian networks (OOBN) were introduced in Chapter 3. In the previous section, components to represent the renewable energy system comprising the energy technology and stakeholders were introduced using a UML class diagram. Furthermore it was decided to focus on the domestic vector to constrain the task of creating an integrated model to a single design pattern. It has been argued that a technology agnostic model can be developed. The purpose of this section is to describe the process of constructing a conceptual model as a heuristic which represents those parameters, and the relationships between them, as a first stage in the construction of a formal OOBN.

4.4.1 Building an Object Oriented Bayesian Network

Researchers suggest the employment of an iterative approach when building a large multi-domain OOBN (Johnson, 2009). This highly formalised approach which involves workshops with domain experts was not adopted here. Instead, tacit knowledge, in-house[11] expertise, and knowledge documented in the academic and grey literature were used to support the development of the model. A heuristic approach was adopted based on causal mapping (Goodier et al., 2010; Nadkarni and Shenoy, 2004). Domain parameters and qualitative relationships between them are identified. As well as named variables, a class, which might encapsulate one or more quantitative variables, was often used as a proxy. For example, from tacit knowledge it can be normatively stated that the site where the renewable energy technology is located (class: site) is a predictor for the renewable

[11] Centre for renewable energy and systems technology and the school of civil and building engineering at Loughborough University

energy resource (class: energy resource). This is asserted with being explicit about neither the attributes of the site, nor the renewable resource, which quantify this relationship[12].

The first stage in the process of building a BN starts with the identification of the target nodes to be used in decision making (Varis and Kuikka, 1999). These represent variables of interest for which the BN updates posterior probability distributions following the presentation of evidence on input nodes. In this research the target nodes are to be used to assess the technology and answer the research questions. The variables they represent are the KPIs discussed below.

For each identified variable (or class proxy) there followed a process of working iteratively backwards, identifying in turn their predictor variables or classes. This continued until the system boundary was reached, which in this case was the renewable energy system deployed within its domestic context. Once reached the conceptual model is complete.

This is an inductive approach to defining an ontology of the whole problem domain. The second stage of the iterative process is to componentise the model in order to render it object oriented. This involves the deconstruction of any class elements used in the heuristic model to expose their encapsulated variables and relationships. These, alongside other already identified variables and newly identified variables are reconstructed into classes each of which can serve as Bayesian network components. The guiding principles are as those for object orientation in software design, principally substitutability, autonomy and abstraction (Armstrong, 2006). This iterative process requires further development and refinement of the ontology of each Bayesian network component.

[12] As a practical example of this consider a building roof as the installation site of a solar energy technology. The size, geometry and geographical location of the roof are predictors of the insolation – the solar energy resource – impinging on the energy technology.

The two stage process described above is a new method developed for this PhD research and is documented in Table 4-8. The next subsection discusses the finalisation of the conceptual model to feed in to the creation of the OOBN.

Table 4-8 Iterative design procedure for the object oriented Bayesian network

Steps to building the OOBN
Create heuristic model
1. Identify target nodes representing key variables of interest.
2. Identify variables or proxy classes which are predictors for these variables, building a causal map as each design pattern is identified.
3. Identify variables or proxy classes which are predictors for previously identified variables, building up the causal map.
4. Repeat previous step until system boundary reached.
Create formal OOBN
5. Deconstruct any classes in the heuristic model to expose their variables.
6. Collate all the variables into classes to serve as BN components.
7. Identify relevant dependencies and interfaces for each component.
8. Verify and validate Bayesian network components as far as possible. 9. Connect the components through their interfaces.
10. Verify and validate integrated OOBN as far as possible.

4.4.2 Building the Conceptual Model

The results of the stage one procedure described above are presented in this section as a causal map (Figure 4-7). This shows the relationships between variables or proxies represented as nodes. The directed edges indicate the direction of influence between the nodes. A key to the variables is shown in Table 4-9.

The target nodes, shown yellow in the causal map, are the percentage of household income spent on fuel, the reduction in carbon emissions, and the income generated by the energy system. The green nodes pertain to the renewable technology system, and the red nodes to the socio-technical context in which the energy system is located. This illustrates the concept of an object oriented design whereby these components can be substituted with equivalent components for a different technology or socio-technical context without having to redesign the rest of the model.

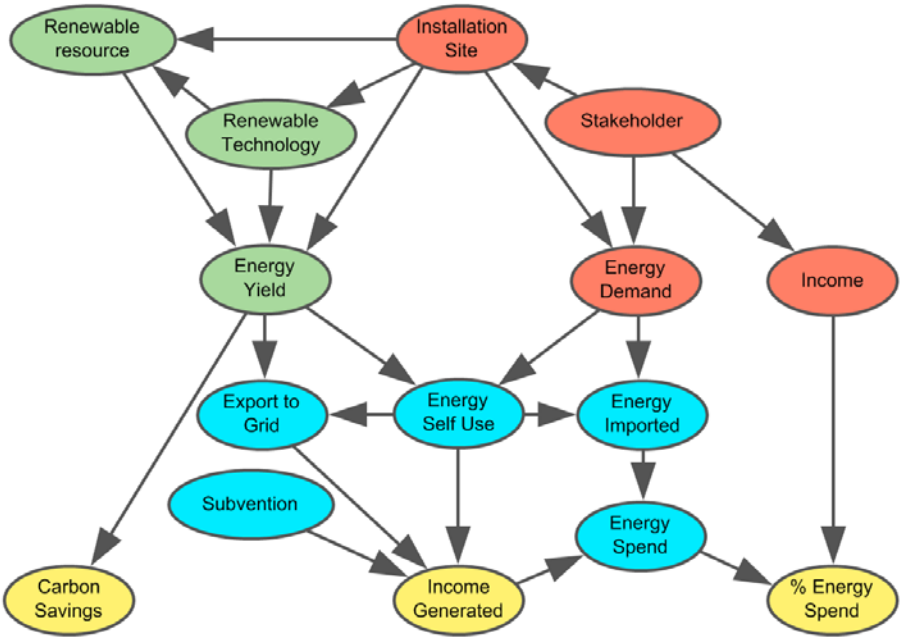


Figure 4-7 Causal map for key parameters for the domestic vector

Table 4-9 Parameters for conceptual model for the domestic vector

Parameter/Class	Comments
Renewable Resource	The available resource for the renewable energy system for example the amount of sunshine (insolation) or wind resource.
Installation Site	Factors which affect the energy potential of a site such as its physical size, geographic location, and any adverse factors such as shading. For this model the site is the domestic property.
Renewable Technology	A class which encapsulates the type of energy generator, the capacity or rating, and meta data which influence its capacity factor and efficiency
Energy Yield	The total energy generated by the renewable energy system (kWh/year)
Energy Demand	The energy demand of the site. For domestic properties this is electricity, gas and other fuel consumption. (kWh/year)
Stakeholder	This is a class representing the householders or occupants where the energy technology is installed.
Income	Household income
Export to Grid	The quantity of generated power exported to the grid (in the case of electricity generating technologies) (kWh/year)
Energy Self-consumption	The amount of energy generated which is consumed on site
Energy Imported	The amount of energy imported from the grid or suppliers to make up for shortfall by variable renewable generation.
Subvention	Subsidy i.e. FiT the technology attracts.
Energy spend	Amount of money spent on energy by the household
Carbon emissions reduction	This is a result of displaced carbon from the conventional energy system, and will assume a carbon intensity of contemporary grid electricity. (kg CO_2/year)
Income Generated	Amount of financial value the technology contributes to the household
Percentage Energy Spend	Percentage of household income spent on energy needs

4.5 The Choice of Technology

Solar PV, solar thermal, heat pumps, micro-CHP and micro wind are all renewable microgeneration technologies which have been considered suitable for domestic installation (Sudtharalingam et al., 2010). Each technology has the potential to be modelled as substitutable component, as shown in Figure 4-8, and inserted in to the conceptual model shown in Figure 4-7. The introduction to this thesis proposes solar PV as the technology of choice for a case study to demonstrate integrated modelling using BN. This section clarifies that choice.

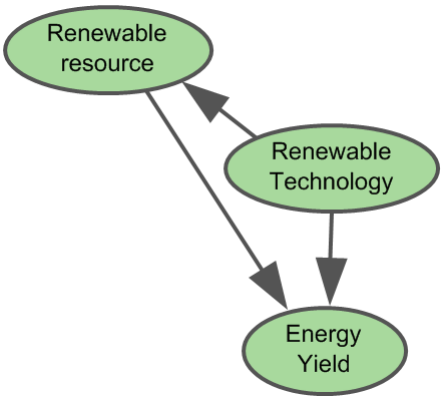


Figure 4-8 renewable technology components for the conceptual model

The installed capacity and number of installations for each technology is shown in Table 4-10. UK field trials delivered mixed results for heat pump systems and, so far, take up has been adversely affected by consumer confidence and lack of market awareness (Singh et al., 2010). As a heat technology it has only recently benefitted from subvention through the Renewable Heat Incentive (Rees and Curtis, 2014). Micro-CHP, despite being supported by the FiT, has had a low take up (Table 4-3). The technology is not suitable for all domestic properties, and has high site specificity[13], requiring space for both the generator and fuel store. Micro-wind technologies also suffer from site specificity inhibiting wide-scale penetration. They have proven not to be effective in the urban environment where turbulence reduces their performance (Heath et al. 2007).

[13] Site specificity refers to the number of conditions or criteria a potential site must satisfy before being suitable for a particular technology.

Table 4-10 Candidate microgeneration technologies for the OOBN model

Technology	UK installations	Installed Capacity (MW)	Date	Reference
Heat Pump	17760	Not available	2013	Rees and Curtis, 2014
Micro CHP	477	0.5	April 2014	See above
Micro-wind	Not available			
Solar PV	464,522	2,056	April 2014	See above
Solar Thermal	177,418	497	Dec 2012	Mauthner and Weiss, 2014

Only Solar PV and solar thermal technologies have been adopted on a scale numbering in the 100 thousands in the UK. PV benefiting from incentivisation by the FiT, has seen rapid rates of uptake. Solar thermal technology has several decades of market readiness (Sudtharalingam et al., 2010). Both technologies have been the subject of extensive field trials and large datasets of time resolved generation data are available to the research community. Solar thermal is perceived has having greater installation complexity and requires a hot water storage the space for which many UK domestic properties have forgone in the conversion to gas fired combi-boilers. In comparison, solar PV has the least site specificity, requiring only a suitable roof as the installation site. The context for this research was the launch of the Fit in 2010. In order for the integrated model to make a greater impact solar PV technology was chosen as a prime candidate for the development of an object oriented BN.

The next task in developing a conceptual model was to choose case studies for the estimation of solar PV yield commensurate with available socio-economic data. This is discussed in the next section.

4.6 Selection of Cases

Where a sample of a population is to be analysed requires the selection of cases for inclusion. This sub-section presents the rational for the selection of cases. Firstly the unit of analysis – the definition

of a single case - is defined. The selection of cases using a whole LSOA is proposed and the selection of several LSOAs for this purpose is discussed.

4.6.1 Unit of Analysis

The conceptual model shows the components for the renewable technology, which, when in situ, is contextualised by socio-technical sub-system consisting of a physical location or site, and a stakeholder[14] (Figure 4-9). This can be regarded as a unit of analysis. For the domestic vector the site is the domestic property, and the stakeholder is the occupant.

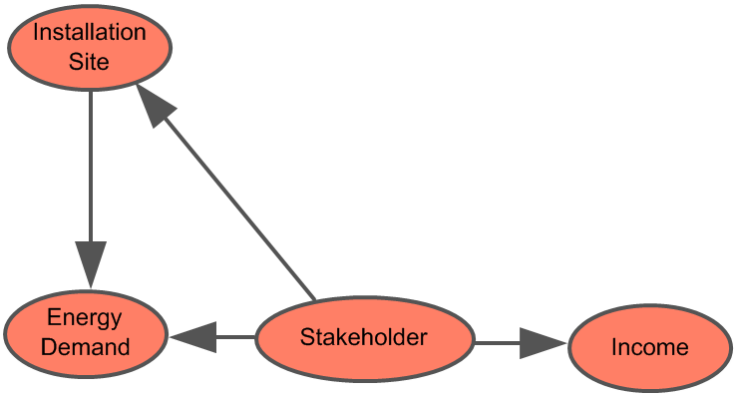


Figure 4-9 Unit of analysis as a socio-technical representation of the technology adopter

The model requires the acquisition of data for a representative sample of the population of UK households. As the sub-model suggests, the data required needs to include attributes of the site – i.e. the property - which influence the energy demand, and, which predict the renewable resource for, and the energy yield of, the chosen technology, solar PV. The refinement of the required the predictor variables required to quantify these influences is discussed further in chapters 5-8.

[14] The stakeholder may be a household which may comprise one or more persons.

A second requirement for each unit of analysis is the income which enables the prediction of percentage energy spend – the fuel affordability indicator.

Without recourse to extensive property and occupant surveys, for which this project had limited resources, consideration was given to the sourcing of secondary data. A review of secondary data sources pointed to the LSOA as a geographic unit of analysis for which to source domestic building stock data and commensurate socio-economic attributes of occupants. These data sources are discussed in detail in subsequent chapters. The selection of specific LSOA for which to source data is discussed in the next sub-section.

4.6.2 The LSOA as a Geographic Scale

The LSOA is used as a spatial unit for the presentation of aggregated statistical data by government departments and the ONS. By choosing the LSOA as a spatial unit for which to source data ensured the impact assessment methodology can be closely coupled with many other socio-economic impact assessments and policy initiatives such as energy efficiency interventions targeted at low income communities (Rosenow et al., 2013).

4.6.3 Purposeful Selection of Lower Super Output Areas

There are over 34,000 LSOA in England and Wales and of these, four areas were purposefully selected[15]. The rational for selection of these was as follows:

- A high IMD in order to be able to test the hypothesis that high solar penetration could have an impact on fuel affordability indicators.

[15] Purposeful selection means that the choice was not randomised sample, but was selected with particular criteria in mind with implications for the generalizability of the research.

- The LSOA were chosen to be at a representative range of latitudes in England, from Cornwall in the South, to Newcastle in the North, which influences the renewable resource.
- For the purpose of validation of building stock attributes at least one LSOA local to the research institute was chosen in order to be able to conduct a 'walk through' for data validation purposes.
- The LSOAs between them should provide a range of housing stock in order to test the hypothesis that stock would influence the SEE results.
- In order to have a large impact on more populous city and town local authorities urban LSOA where selected.
- For impact LSOA were selected where evidence was found of considerable participation in the renewable energy agenda in order to potentially facilitate co-operation and impact.

Table 4-11 Selected LSOAs

LSOA Code	Name	Town	Region	IMD
E01018870	Kerrier 008B	Camborne	Cornwall, South West	10
E01025703	Charnwood 002D	Loughborough	Leicestershire, East Midlands	7
E01011223	Kirklees 042B	Huddersfield	West Yorkshire	10
E01008380	Newcastle 008G	Newcastle	North East	10

The four selected LSOAs are shown in table 4-11, with the top one in the South, and progressing northwards. Cornwall is a County undertaking considerable promotion of renewable energy. The Charnwood LSOA is local to Loughborough where this research was conducted. Kirklees is another local authority actively engaged in community energy initiatives. Newcastle is the most Northern large city in England. A maps of each LSOA showing the building footprints and roofs and are shown in Appendix 1.

4.7 Summary

An analysis of the diffusion of renewable technologies in the UK has provided a picture of adoption rates segmented by technology and adopter vectors. This has given insight in to the scoping of a decision making tool which could account for multiple technology and adopter vectors and account for the build environment and deprivation. From a range of qualitative impacts three KPIs were selected as target nodes for a BN model which relate to government policy objectives: carbon reduction, economic impacts and fuel affordability.

How measures of these might be incorporated in to a model suggested different patterns for the domestic vector in comparison to and community or commercial adopter vectors. The former can be modelled with a one-to-one relationship between technology and stakeholder, whilst the latter requires a one-to-many relationship; these cannot be modelled in a unified way using a BN. For practical purposes the domestic vector was chosen.

A technology agnostic conceptual model was developed with the object oriented characteristic of substitutable components for the renewable technology. This conceptual model can thus be further developed into an OOBN using formal model building techniques. The model defined a socio-technical system perspective of the adopter as an installation site and stakeholder which for the modelling of SEE impacts can be regarded as a unit of analysis.

Solar PV was chosen as the technology which presents opportunities for impact. Given its greater rate of adoption and low site specificity makes it an appropriate technology across a wide range of socio-economic contexts as evidenced by the number of purposeful community energy projects in a significant number of LSOA.

The problem of finding case studies from which to acquire data to furnish the model was resolved by electing to use whole LSOA to provide representative properties and occupants. Four LSOA were

purposefully selected which met the criteria of low income urban areas with a variety of building types and spatially distributed north to south.

The choices made here narrow down the scope of building an OOBN for domestic solar PV deployed in four urban LSOA, whilst further developing the methodology to be applicable for other technologies.

5 Solar PV Yield

5.1 Introduction

This chapter presents the development of a BN sub-model which predicts the energy generated by domestic solar PV systems. The energy generated by domestic solar PV, called the yield, is very much influenced by the system technology and its spatial context. Thus there is a need to understand how these diverse parameters influence the yield, and to get some measure of their uncertainty.

The construction of the core BN sub-models in this thesis follows 'pattern' outlined in Chapter 4. For this chapter this is as follows. Firstly the ontology of the domain is developed to explicate predictor parameters for the yield in Section 5.2. This is derived from a literature review, including an overview of solar PV technology. This analysis informs data requirements for the model and the acquisition, provenance and processing of data sources which are critically discussed in Section 5.3. The available data, and the domain ontology from the literature review, determine the dependency relationships and thereby the constructions of the BN sub-model in section 5.4. Here the construction of a DAG and the derivation of conditional probability tables are presented. A discussion of the working Netica BN sub-model follows in Section 5.5.

5.2 The Domain Ontology

In order to appreciate the knowledge domains from which to derive model parameters, a brief explanation of a PV system is required. This consists of one or more PV modules, collectively an array, which, when exposed to sunlight, generate an electrical potential difference (voltage) explained by the photovoltaic effect (Becquerel, 1839 cited in Wenham et al 2011). This occurs when a semiconductor material with a p-n junction is irradiated with photons with a quantum energy

equal to or above the band gap energy of the material. This excites electrons from the valence band to the conduction band. The resultant charge carriers migrate in the junction zone to produce a potential difference which in turn can drive a DC current through a load in a closed circuit. The modules are wired to an inverter and transformer, the function of which is to convert and match the DC current to the single phase AC supply used by domestic appliances. The AC output of the inverter is connected to the domestic electricity supply such that, when the instantaneous electricity demand on the consumer side of the electricity meter is less than the power output of the array, any excess is exported in to the low voltage network. When the demand is more than the power output of the array the shortfall is met by importing electricity from the low voltage network.

From this brief exposition of PV technology[16] several key knowledge domains are pertinent to the magnitude and variability of the yield of a deployed solar PV system. Firstly the environmental parameters which govern the quantity of sunlight received by the PV modules system are considered in Section 5.2.1. The light receiving technology, i.e. the types of PV module, are considered in Sections 5.2.2 whilst Section 5.2.3 presents a discussion of the 'balance of system (BOS) components. Section 5.2.4 considers how PV systems are rated and Section 5.2.5 examines simulation methods which are used to estimate PV Yield. The section concludes with a summary (Section 6.2.6).

5.2.1 Operational Irradiance of Solar PV

The instantaneous power generated by solar PV is determined by the intensity of the solar radiant flux striking the PV module surface. This is called the irradiance, G_T measured in Wm^{-2}. The total energy H_T, known as the solar radiation or insolation (the latter term is used in this work), over a

[16] Further reading on grid connected solar PV can be found in Goss (2010) and Wenham (2011).

given time is given by integrating the irradiance with respect to time (Equation 5-1). This is measured in Jm^{-2} or kWh/m^2.

$$H_T = \int_{t_1}^{t_2} G_T \cdot dt \qquad \text{Equation 5-1}$$

The irradiance consists of three components (Figure 5-1):

(i) The direct or beam irradiance which results from light travelling in a straight line directly from the sun.

(ii) The indirect irradiance which results from the scattering of the beam irradiance as it travels through the atmosphere. This resultant diffuse irradiance arrives from all directions at the PV module.

(iii) Light is also reflected from the surface of the earth may ultimately arrive at the PV module.

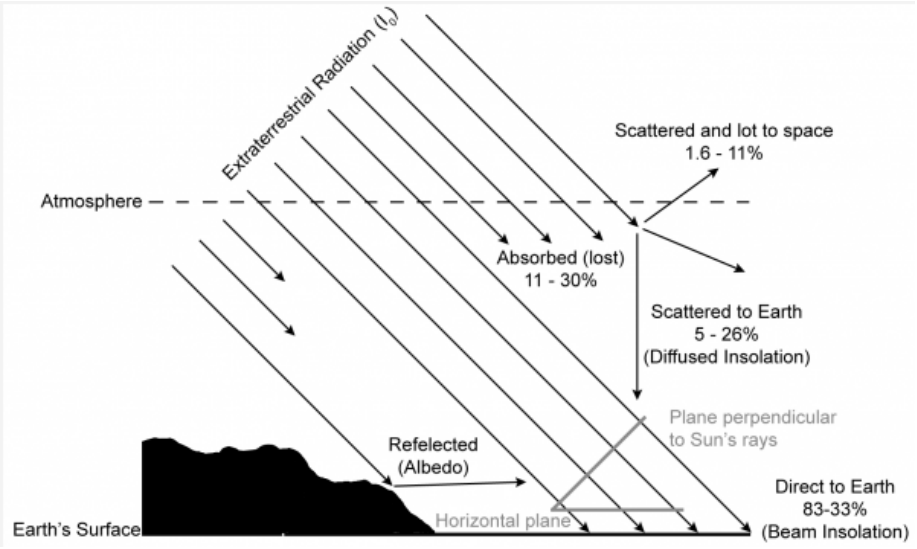


Figure 5-1 Direct, scattered and reflected solar radiation

G_T is partly governed by a complex, non-linear function of time which predicts where the sun is relative to an observer, and partly governed by stochastic meteorological conditions which determine the absolute and relative contributions of beam and diffuse components (Wieder, 1982).

Thus the integral in Equation 5-1 is non-trivial, which ensures that the precise prediction of solar PV yield is uncertain. These uncertainties are explored in the following sections.

Solar Radiation

The starting point in a prediction of irradiance striking a solar PV system is the irradiance impinging the upper atmosphere. The sun emits electromagnetic radiation with a frequency spectrum which approximates that of black body at a temperature of 5800K. Integrated over this spectrum, which spans from the UV to the far infrared, the irradiance striking the upper atmosphere has an average intensity of 1366 W/m^2, known as the solar constant G_{SC} (Iqbal, 1983). As light passes through the atmosphere this intensity is attenuated by molecular absorption processes (principally by carbon dioxide, oxygen, ozone and water), and scattering processes caused by dust particles and gas molecules. The intensity of the beam component is attenuated and the diffuse component increases. Since the diffuse component is distributed in all directions 50% of this is directed way from the planet's surface. Even on a clear day as much as 30% of the incident radiation is attenuated by these mechanisms, and with cloud cover significantly more.

The degree of attenuation depends on the air mass (AM) through which the light traverses on its way to the Earth's surface (Suri and Hofierka, 2004). Thus radiation travelling to a point with the sun directly overhead i.e. at a solar elevation of 90° from the horizontal, will traverse one atmosphere thickness (AM1), whereas if the sun is lower in the sky, at solar elevation of 30°, it will travel through approximately two atmospheres (AM2) thickness. Both scattering and absorption mechanisms are highly wavelength dependent thus the AM2 spectrum will have a different wavelength distribution than an AM1 spectrum. Figure 5-2 shows the AM1.5 spectrum in comparison to the solar spectrum incident on the upper atmosphere, which is called an AM0 spectrum. The integrated intensity of the

AM1.5 spectrum is 1003W/m^2. These spectra are agreed standards (ISO 1992) and are used for benchmarking solar panels.

As well as the air mass effect, the position of the sun determines the irradiance per square meter on the horizontal surface, explained by the cosine effect – a parallel beam of light of unit cross-section and power G^oimpinges on a larger projected area by a factor of the cosine of the oblique incident angle θ, resulting in a power density of $G^o cos\theta$.

This effect is the main driver for the seasonal variation in weather; in seasons when the sun is, on average, lower in the sky, the incident irradiation on the surface is, on average, spread over a larger area. This also has a stark effect on the seasonal and daily insolation striking the solar PV modules.

Irradiance therefore is highly dependent on the cosine effect and atmospheric attenuation which in turn depend on the sun's position. Equation 5-1 therefore has a high dependency of the sun's daily and seasonal motion across the sky, discussed in the next section.

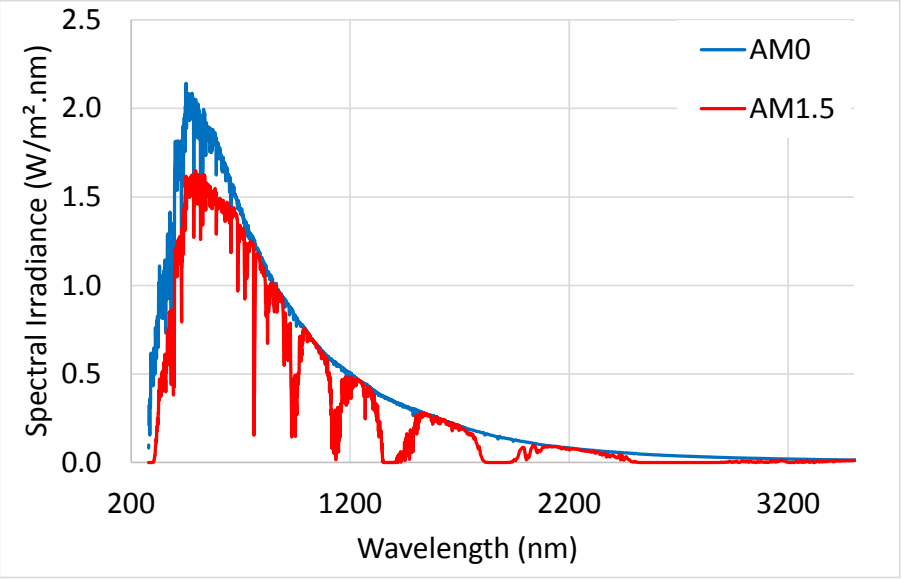
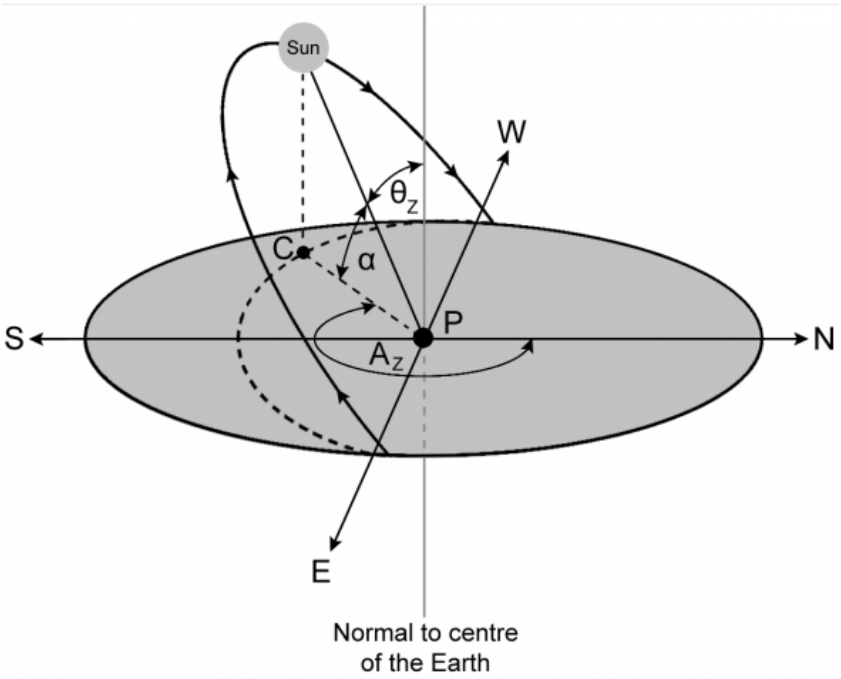


Figure 5-2 Comparison AM0 and AM1.5 solar spectra

The Apparent Motion of the Sun

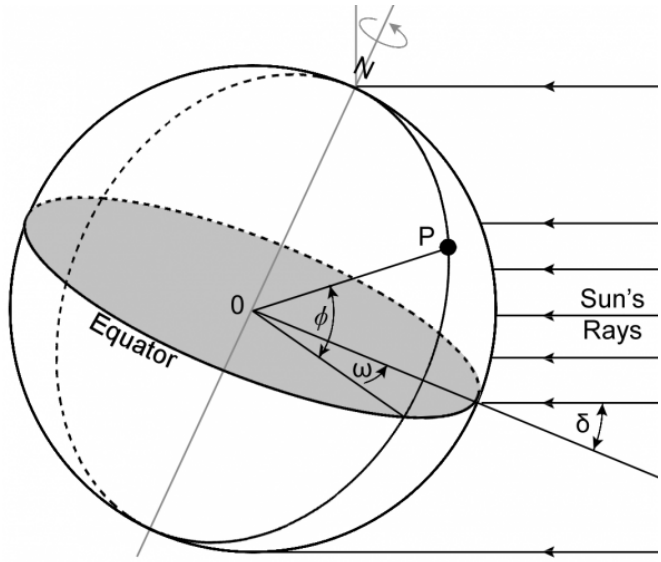
The motion of the sun can be understood from the perspective of the celestial sphere (Jenkins, 2013). To an observer at any point P on the Earth's surface, the position of the sun at any moment in time is defined by the solar zenith angle, θ_z (or its complimentary angle the solar elevation, α) and the solar azimuth angle A_z (Figure 5-3).



The solar elevation α is the angle subtended by the line between the point P and the sun, and the horizontal plane at P. The solar azimuth, A_z, is the angle on the horizontal plane between the plane defined by α and geographic North.

Figure 5-3 The apparent motion of the sun

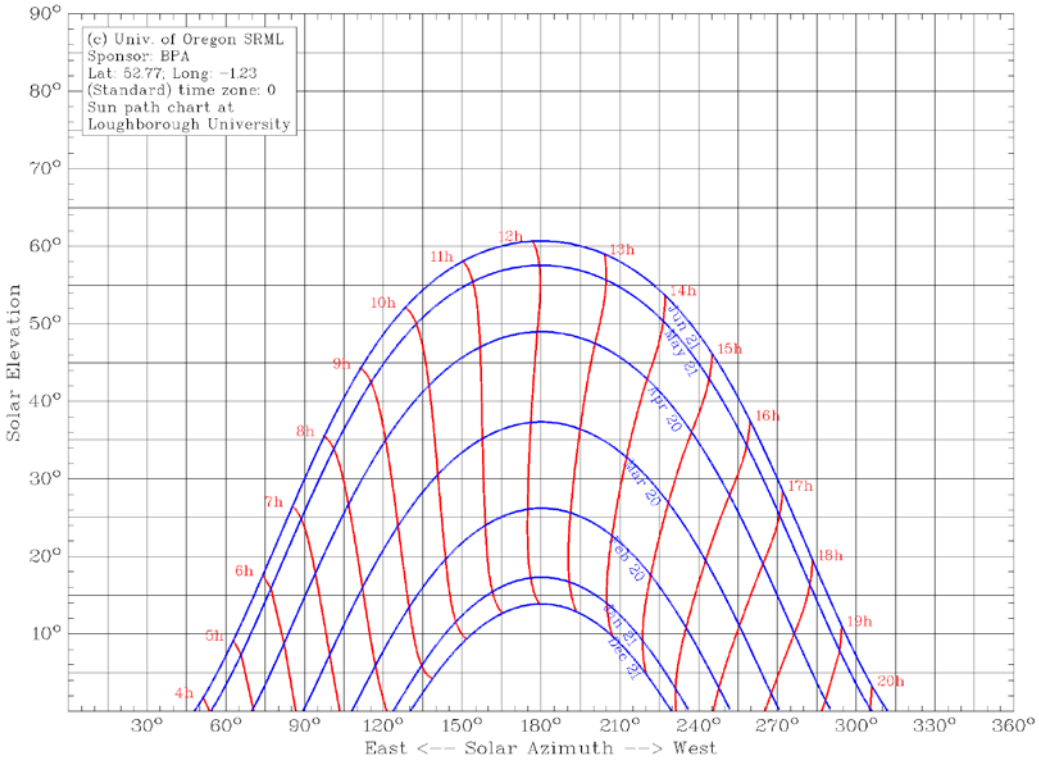
This position, as a function of time, is calculable using geometry and three observable parameters: the latitude of point P, the hour angle, and the declination angle (Figure 5-4) (Probst 2002).



ϕ is the latitude, ω the hour angle and δ the declination angle. The hour angle is calculated from the time of day with 24 hours (one rotation) equivalent to 2π $(-\pi\ to + \pi)$ radians and 12 noon set at zero. δ is the declination angle, due to the tilt of the earth's axis of rotation relative to the plane in which it orbits the sun (and the plane of the sun's radiation). When the Earth is titled towards the sun in summer the maximum value of δ is 23.24° at the summer solstice. The declination becomes zero at each equinox and -23.24° at the winter solstice. The declination angle can be calculated for any day of the year (Probst, 2002)

Figure 5-4 Observable parameters with which to calculate the position of the sun at a point P at any moment in time

Using these algorithms the apparent motion of the sun across the sky from the perspective of an observer at any point on the surface of the Earth can be determined. This can be shown using a sun path diagram (Figure 5-5). This shows the variation in solar elevation through the day at different times of the year. For the given location the diagram shows the sun achieves an elevation of only 15° in the winter and the day is seven and a half hours long. In contrast in summer the sun achieves an elevation of 60° and maximum 16 hours of daylight.



Sunpath calculated for geographic co-ordinates latitude 52.77 longitude -1.23 generated using the University of Oregon Solar Radiation Monitoring Laboratory's Sun Path Chart Program (UOSRML 2014). The blue lines plot the solar elevation against the solar azimuth for a day of the month from December (bottom) to June (top). The red lines show the time of day.

Figure 5-5 Sun path diagram at Loughborough University

The above discussion clarifies the variation in seasonal and daily irradiance. The apparent motion of the sun enables the prediction of beam irradiance on a horizontal plane on the Earth's outer atmosphere using geometry and the solar constant. In principle, taking into account the air mass and cosine of the angle of incidence, the irradiance on a horizontal surface on the surface of the earth should also be accessible. However, the stochastic behaviour of the weather as it affects cloud cover and the clearness of the sky, ensures that the prediction of beam and diffuse irradiation on a horizontal plane on the ground remains elusive to theoretical prediction, and empirical methods are therefore required. These are discussed in the next section.

Measurement and Prediction of Solar Irradiance

The first step in the prediction of terrestrial insolation in the plane of array of an operational PV system is to measure or predict the irradiance over time on the horizontal plane. Irradiance can be measured using a pyranometer, which uses a blackened thermopile to absorb all wavelengths in the solar spectrum incoming from all angles of the hemisphere above its plane of installation; i.e. it receives both the direct beam and diffuse irradiance from the sky and reflected from the environment (Scharmer and Greif, 2000). Using a specially positioned shading ring to block the beam irradiance as the sun moves across the sky a pyranometer can also be configured to measure only the diffuse component.

It is both expensive and complex to measure both components at every site where one might wish to install a PV system, particularly at the domestic scale. The challenge, then, for the prediction of PV yield, at any location (as required in this work), is the accurate estimation of irradiance and insolation at any geographic point of interest. Temporally and spatially resolved solar irradiance is an essential component in the toolkit for a number of diverse disciplines (e.g. climate science, agriculture, forestry and architecture), as well as solar energy conversion (Page, 2005). To meet this demand a number of solar irradiance products (SIP) with a wide territorial reach have been developed over a period of years. Table 5-1 lists several such products, but this is by no means exhaustive.

Table 5-1 Solar Irradiance Products

Database	**Area of coverage and spatial resolution**	**Comment**
European Solar Radiation Atlas (ESRA)	Europe and North Africa. 10km x 10km	Uses data collected from meteorological stations between 1981 and 1990 and some satellite data.
ESRA/r.Sun	Europe and North Africa. 1kmx1km	A GIS adaptation of ESRA with higher resolution - see text
Climate Modelling Satellite Application Facility (CMSAF)	Europe and North Africa. 1.5kmx1.5km	Derived from European satellite data collected by the Meteosat First Generation (MFG) (1998-2005) and Second Generation satellites (MSG) (2006-2011).
NASA Surface Meteorology and Solar Energy (SSE)	Worldwide. 111kmx111km (1° latitude and longitude grid)	Derived from NASA satellite data collected between 1983 and 2005
Meteonorm	Worldwide	Data collected between 1981 and 2010 from 8,325 weather stations from the World Meteorological Organization (WMO)
SolarGIS	Europe, North Africa and Southwest Asia	Data collected from the Meteosat MSG (2006-2011)

These SIPs have been reviewed by a number of scholars (Burgess 2014; Cros et al, 2004; Šúri, 2007). There are several general points to be made about the inherent uncertainty of derived irradiance and commensurate PV yield prediction. Firstly, their spatial resolution varies widely. In a maritime climate such as that of the UK, the irradiance and insolation is likely to vary over quite short distances due to variability in atmospheric conditions. Moreover the albedo (reflectivity) of the landscape, the elevation of the site and the horizon are all factors which affect the global horizontal irradiance. Therefore a database with a low spatial resolution, which homogenises complex terrain such as mountainous landscapes, mixtures of forest and water or built-up areas, is likely to yield average values rather than site specific irradiance in such areas (Huld, 2012).

All the SIPs use averaged historical data to derive site specific irradiance data and, whilst there is a well-documented variability in annual terrestrial irradiance of about five percent (Goss et al., 2012), there is an assumption that the long-term average is constant and their irradiance values are valid today. However, recent work has reported a long-term trend over a number of decades of an increase in global horizontal irradiance of 3% to 5% per decade (Betts and Gottschalg, 2013; Wild, 2009, Wild et al 2005) thus rendering this assumption suspect, particularly for products using data from the 1980s. This "global brightening" was attributed by Huld (2014) for the generally higher irradiance values delivered by the CMSAF in comparison to the ESRA irradiance product. Thus, the two sources give very different PV yields, which is important for the work presented in this thesis and is discussed further below.

The third general point about SIPs is to note that the underlying methodology to estimate the global horizontal irradiance (G_H) is of two distinct types. The first method involves the measurement of terrestrial global horizontal irradiance using a network of meteorological ground stations. An estimate of the irradiance at any point is then derived by interpolation between the empirical station data points. Interpolation algorithms involve three dimensional surface fitting techniques using spline functions (Hutchinson et al., 1984), weighted averages (Hulme et al., 1995, Perez, 1997) or kriging methods (Zalenka et al., 1992) which create a best-fit surface through the empirical data points (Nguyen and Pearce, 2010). The second method uses observations by geostationary satellites of the irradiance reflected by the Earth and its atmosphere to derive the global irradiance at ground level, first demonstrated by Yonder Haar and Ellis (1978) and continuously improved since (Lefèvre et al, 2004, Perez et al, 1997). Both methods use parametric models, each with a number of assumptions which introduce uncertainty into derived irradiance values.

Finally, whilst the core objective of the products is to estimate the global irradiance on the horizontal plane, G_H, many products also furnish the spatially resolved beam and diffuse irradiance on the

horizontal plane. This is achieved using empirical methods and, since its quantification is essential for the estimate of irradiance on a tilted plane, is discussed next.

Estimation of Diffuse and Beam Components

As discussed already, the beam component of solar irradiance is attenuated by various mechanisms including absorption and scattering. The degree of attenuation can be described by the ratio between the extra-terrestrial irradiance, I_o, and the global horizontal irradiance at ground level, I_G (Equation 5-2).

$$k_T = \frac{I_G}{I_0} \qquad \text{Equation 5-2}$$

The ratio is called the clearness index. The diffuse fraction, that is, the ratio of the horizontal diffuse irradiance component to the horizontal global irradiance, both at ground level, will be a function of the clearness index (Equation 5-3).

$$\frac{I_D}{I_G} = f(k_T) \qquad \text{Equation 5-3}$$

Using ground station measurements for global and diffuse components, and theoretical measurements of extra-terrestrial irradiance these relationships can be determined (Liu and Jordan, 1960). Known as diffuse fraction correlations they are often fitted with polynomial regression functions. However, because the scatter is so large, predictor variables other than the clearness index have been introduced such as the solar elevation, humidity and temperature (Reindl et al., 1990). This improved the predictive accuracy, as indicated by the residual sum of squares, by 14%.

The fundamental problem with this approach is that over the measurement period of an hour there are many values of diffuse fraction for a given clearness index, as demonstrated by McCormick and Suehrcke (1991). Thus the use of diffuse fraction correlations introduces a large uncertainty in to the

predicted diffuse component using a given global irradiance. This inaccuracy is exacerbated when longer periods of measurement such as daily irradiance or average monthly irradiance are used to construct diffuse fraction correlations.

Improving diffuse component estimations is still an active research endeavour, with the development of the BRL model and the use of muli-parameter logistic models (Boland et al., 2013). However, these techniques have not found their way into irradiance and solar PV performance estimation tools and simple models such as Muneer's diffuse fraction correlation polynomial, as used by ESRA and r.Sun (Celik and Muneer, 2013) which purports to be a standard for a number of world-wide locations.

Measurement and Prediction of Irradiance in the Plane of Array

It is necessary to predict the solar irradiance in the plane of the PV array in order to assess a site for an installation. In this study the focus is on roof-mounted domestic systems, which by their nature have a constrained azimuthal and inclination angle. As discussed above, most irradiance measurement and irradiance maps provide the global horizontal irradiance. The tilted plane irradiance G_T, is made up of three components the beam G_{bT}, diffuse G_{dT} and the ground reflected G_{rT} (Equation 5-4).

$$G_T = G_{bT} + G_{dT} + G_{rT} \qquad \text{Equation 5-4}$$

The beam tilted component, G_{bT} can be calculated from the beam horizontal irradiance G_b on the tilted plane using trigonometry (Equation 5-5), where σ_z is the elevation of the sun and ζ is the angle of incidence of the beam on the tilted plane.

$$G_{bT} = G_b \cdot \frac{\cos\zeta}{cos\sigma_z} \qquad \text{Equation 5-5}$$

The calculation of the diffuse component on the tilted plane depends on the anisotropy of the horizontal diffuse irradiance. Early assumptions of an isotropic diffuse irradiance profile (Perez et al., 1987; Saluja and Muneer, 1987) proved inaccurate. Anisotropic models were developed to take into consideration a circumsolar diffuse component (Perez et al., 1987) and horizon effects (Muneer, 1990).

The inaccuracies of these models are underscored by the findings of Šúri et al. (2008), who found a 21% increase in the standard deviation among models moving from a horizontal to a 34° south-oriented plane. Similarly, Betts and Gottschalg (2013) found that the range of in-plane irradiance variability on an optimally inclined plane was 6% greater than that of horizontal radiation. These findings are significant when considering the modelling and comparison of estimated and empirical yield data.

Summary

The amount of sunlight striking the solar panels is in the first instance tractable problem using geometry and the motion of celestial bodies. The effect of weather, however, renders this less than accurate due to photo-physical processes in the atmosphere which creates diffuse and direct components of solar radiation. To estimate these, without site specific measurement equipment, recourse has to be made to empirical models to estimate the global in-plane solar radiation. It has been shown that this necessarily introduces a large uncertainty in to the estimated insolation.

5.2.2 Semiconductor and Substrate Technology

A typical module is made up of semiconductor devices or cells which are hermetically sealed under toughened low-reflectivity glass. A range of different semiconductors materials and morphologies are available for use in commercial modules. The cells may be held in rigid, rigid thin film, or flexible thin film modules designed to suit a range of deployment applications such as roofs, building façades, and ground-based arrays. The main commercial types are shown in Table 5-2.

Table 5-2 Solar PV module technology, market share and efficiency

Semiconductor Material	**Market share (%)** [1]	**Efficiency$_{STC}$ (%)** [2]	**Maximum Efficiency recorded (%)**[2]
Silicon (monocrystalline)	24	17-20	23
Silicon (polycrystalline)	62	15-17	18.5
Amorphous silicon	2	10	10.5
Cadmium telluride (CdTe)	4	12-13	16.1
Copper indium [gallium] selenide (CIS/CIGS)	2	<13	15.7
Other	6		
1. Solarbuzz (2013). 2. Green et al. (2013)			

Due to the conversion efficiency of the photovoltaic effect, a joule of incident light energy produces a fraction of a joule of electrical energy. The module efficiency, η_m, is measured as the ratio of the total electrical power produced, P_T, per unit area to the total incident light power G_T per unit area (Equation 5-6) (Wenham, 2011).

$$\eta_m = \frac{P_T}{G_T} \qquad \text{Equation 5-6}$$

As shown in Table 5-2, η_m is typically 20% for mono-crystalline silicon but only 10% for amorphous silicon. Thus, when considering the specific yield of a PV module, the material is significant; PV

modules made of materials with higher conversion efficiencies will, all other things being equal, produce higher specific yields. This means when simulating specific yields for domestic properties it is pertinent to consider the market share of the different PV technologies.

An important characteristic which influences the conversion efficiency is the spectral response of the semiconductor material relative to the solar spectrum. Photons with an energy less than the band gap energy do not excite electrons into the conduction band and so do not contribute to useful irradiance. Wavelengths corresponding to photon energies greater than the semiconductor band gap energy have an excess energy which is wasted and becomes heat in the material. Thus the overlap between the band-gap absorption spectrum and the spectrum of the incoming light radiation is significant in the determination of the conversion efficiency since in Equation 5-6 all the incoming radiation contributes to the measured denominator, but only the band-gap energy of absorbed photons can contribute to the numerator. When comparing the same materials in different environmental or seasonal contexts spectral variations should be taken into account (Krawczynski et al., 2010).

Two further loss mechanisms occur which contribute to the lowering of efficiency. Optical losses occur due to the specular reflection of light at material interfaces. This can occur from electrical contacts on the upper surface of the PV cell, from the material substrate itself or from the rear contact. Recombination losses occur when an excited electron does not contribute to the power output of the cell, but returns to the valence band. The equivalent band gap energy is either converted to heat or re-emitted as a photon.

The operating temperature of a PV module is a significant factor in the efficiency of PV modules; the efficiency decreases linearly as the module temperature increases due to a reduction in the open circuit voltage, V_{OC}. This in turn caused by an increase with temperature of the recombination rate of carriers (Green 2003). This is characterised by the power temperature coefficient, γ, which, using

Equation 5-7, allows the calculation of the output power P_T at temperature T compared to a reference power $P_{T_{ref}}$ recorded at a reference temperature T_{ref}.

$$P_T = P_{T_{ref}} \frac{100 - \gamma(T - T_{ref})}{100} \qquad \text{Equation 5-7}$$

For crystalline silicon a typical γ value is 0.5 %/°C (Skoplaki and Palyvos, 2009). The loss factors discussed above, where some of the incident light energy is converted to heat within the module mass, means that at higher irradiances greater heating occurs resulting in lower efficiency. On a day with high insolation it is not untypical for a module to suffer a warming of 40°C above ambient temperature in real operating conditions. This would cause a 20% reduction in the efficiency of a typical crystalline silicon module.

5.2.3 Balance of System

The balance of system (BOS) refers to all the components (other than the PV modules) required to build a working PV system and includes wiring and connections, switches, module mounting systems and the inverter. The design and configuration of the BOS also influences the electricity yield of the PV system. Equation 5-6 can be reconsidered as a system efficiency, η_s, which, due to the occurrence of energy losses in BOS components, is less than the module efficiency (Equation 5-8).

$$\eta_s = \frac{P_{T_s}}{G_T} \qquad \text{Equation 5-8}$$

The source of these losses is manifold. Resistive losses (I^2R) occur in the transmission of DC currents to the inverter; the quality and rating of wiring and connectors are important to minimise this.

The inverter unit contains the inverter switching devices, a transformer or other voltage regulating devices, and control and safety electronics. The switching devices used to convert DC to AC, and the

transformer used to convert to mains voltage are not loss free. An inverter needs to have a power rating that is well matched to the power output of the PV system in order to minimise these losses (Notton et al., 2010). The system performs best when operating at its maximum current and voltage known as the maximum power point (MPP). The variable output of the PV system due to varying irradiance, often on a fast temporal scale (minutes or even seconds), renders such matching difficult (Wenham, 2011). Modern inverters include a MPP tracking device to optimise the load, thus maintaining as high efficiency as possible at every power output of the system. Despite this, the efficiency of the inverter will vary over the range of operational irradiance powers. The average inverter efficiencies in real operating environments have improved immensely in recent years from 84% to 90% in the 1990s (Decker et al., 1993) up to 95% to 98% in the mid-2000s (Navigant Consulting 2006).

As well as the electrical components of the system, the mounting systems are also critical components. Not least, this determines the modules' orientation relative to the sun and hence the quantity of incident light. The temperature behaviour of PV modules discussed above renders their rate of cooling significant when considering the yield and losses of the system. As well as the ambient temperature, the wind speed and air circulation around the modules is important in determining the operating temperature. Rack mounted systems can perform significantly better than building integrated PV modules where no air can circulate underneath.

5.2.4 Nominal Power Rating, Specific Yield and System Yield

Commercial PV modules are tested in standard test conditions (STC), a carefully controlled environment, whereby the PV module is maintained at a temperature of 25°C and irradiated with an AM1.5 solar spectrum light source, perpendicular to the module plane, at an intensity of 1kW/m^2. The resultant instantaneous power gives the nominal, or so-called W-peak (W_p), power rating of the

PV module. The rating of a PV system is simply the number of modules multiplied by their individual rating. Thus a typical domestic PV system consisting of 8 modules each rated at 250W_p has a system rating of 2kW_p.

The energy generated over a year is the annual system yield. If a 2kW_p rated system has an annual system yield of 2000kWh this is equivalent to 1000kWh/kW_p. This measurement is denoted the annual specific yield. The specific yield allows the comparison of PV systems with a different rating. It is useful to devise a model which predicts the annual specific yield and then, given known or estimated system ratings, the annual system yield of deployed PV systems can be predicted.

Whilst nominal power ratings are provided by the manufacturer's specification sheet it is well documented that the yield of solar PV systems reduces over time due to a degradation of the system components (Jordan and Kurtz, 2013). This can be due to morphological changes in the semiconductor material and the ageing of electronic components resulting in greater electrical resistance. A typical module may be operating at 80% of its original power rating after 25 years.

5.2.5 PV Yield Simulation

With an understanding of irradiance on the tilted plane and PV technology, it is possible to simulate the predicted annual specific yield for a system located in a known geographical location. There are a number of web-based and desktop applications and methodologies for PV yield estimation. Two approaches, PVGIS and SAP, discussed in this section, have been used in this work.

PVGIS

PVGIS has emerged as one of the most popular and free tools for rapid estimates of PV performance in Europe (Huld, 2012). PVGIS is essentially a solar PV performance estimation tool which furnishes

solar energy adoptors and implementers with a decision support system (Suri et al, 2007). It is a web-based application developed by the Institute of Energy and Transport of the European Commission Joint Research Centre which enables any user to obtain the estimation of the electricity production provided by any PV system (ECJRC, 2013).

For the purposes of yield estimation PVGIS utilises two irradiance databases which are based on the ESRA and CMSAF irradiance products discussed above. These have a strong European focus and have played a prominent role in European solar energy research for the last twenty years (Page, 2005).

The radiation atlas is a digital map developed by the European Commission as a resource to provide horizontal, diffuse and beam irradiance estimates at a 10 km resolution for the whole of Europe and parts of North Africa (Scharmer and Greif, 2000). Interpolation of ground station measurements from a network of 560 weather stations between 1981 and 1990 were used to provide 10 year average global irradiance values. Satellite measurements were then used to estimate the clearness of the sky and thereby estimate the diffuse and beam components (Mitasova and Mitas, 1993; Šúri and Hofierka, 2004).

Whilst the atlas provided useful irradiance values for most purposes, the spatial resolution was not always as high as required for solar energy estimations, particularly where the local terrain is highly variable (Cros et al, 2004). Each pixel in the digital atlas represent a mean irradiance for a 10x10km area. To deliver a higher resolution, and to make the atlas more accessible to non-professionals, Suri et al (2007) created an open data platform called R.Sun. In particular this took into account a Digital Elevation Model to take incorporate local terrain variability down to a resolution of 1x1km. The R.Sun implementation provides global irradiance on horizontal and inclined surfaces, monthly averages of daily beam, diffuse and reflected irradiance.

CMSAF uses empirically derived algorithms to calculate the terrestrial diffuse and global irradiance derived from the albedo measured in the upper atmosphere (Rigollier et al., 2003). In Europe this

activity has been co-ordinated by the Climate Monitoring Satellite Application Facility (CMSAF) to produce a database of beam and diffuse irradiance maps using satellite imagery collected from 1998 to 2011. Typically resolutions of 10 km can also be obtained for horizontal irradiance. As with the R.Sun model, this is improved using the same digital elevation model (Huld, 2014).

The two irradiance sources and the resultant solar PV estimates have been compared (Huld et al, 2012). Whilst the ESRA/R.Sun has a higher spatial resolution than CMSAF, the former utilises interpolated data which is much further apart than this, Huld et al (2012) argue that CMSAF can be regarded as having a higher resolution especially in areas where the ground irradiance is not captured by representative ground station measurements. CMSAF delivers generally higher yields than the ESRA/R.Sun database except in mountainous areas. In the UK it is reported that CMSAF gives lower values in the West of the UK.

The input parameters required by PVGIS, which are entered on a web form are listed in Table 5-3. On submitting the data, a web page with the results is returned. An example output is shown in Figure 5-6.

Table 5-3 Parameters required by PVGIS

Parameter	**Options (units)**
Radiation Database	PVGIS-classic | PVGIS-CMSAF
Latitude	(Degrees)
Longitude	(Degrees)
Region	Europe | Africa
Nominal Power	(kWp)
Technology	Crystalline Silica | CIS | CdTe | Unknown
Mounting	Free | Building Integrated
System Losses	(Percent)
Inclination	Angle (degrees)
Aspect	Orientation angle (degrees)

Performance of Grid-connected PV

NOTE: before using these calculations for anything serious, you should read [this]

PVGIS estimates of solar electricity generation

Location: 50°12'44" North, 5°17'10" West, Elevation: 127 m a.s.l.,

Solar radiation database used: PVGIS-CMSAF

Nominal power of the PV system: 1.0 kW (crystalline silicon)
Estimated losses due to temperature and low irradiance: 7.9% (using local ambient temperature)
Estimated loss due to angular reflectance effects: 3.5%
Other losses (cables, inverter etc.): 14.0%
Combined PV system losses: 23.6%

Fixed system: inclination=34°, orientation=73°				
Month	E_d	E_m	H_d	H_m
Jan	0.82	25.4	1.05	32.7
Feb	1.43	40.0	1.81	50.6
Mar	2.44	75.6	3.08	95.6
Apr	3.67	110	4.75	142
May	3.97	123	5.23	162
Jun	4.21	126	5.61	168
Jul	3.76	117	5.04	156
Aug	3.44	107	4.60	143
Sep	2.79	83.8	3.68	110
Oct	1.64	51.0	2.14	66.3
Nov	1.00	30.1	1.30	39.0
Dec	0.69	21.5	0.90	28.0
Yearly average	**2.49**	**75.8**	**3.27**	**99.5**
Total for year		**910**		**1190**

E_d: Average daily electricity production from the given system (kWh)
E_m: Average monthly electricity production from the given system (kWh)
H_d: Average daily sum of global irradiation per square meter received by the modules of the given system (kWh/m^2)
H_m: Average sum of global irradiation per square meter received by the modules of the given system (kWh/m^2)

window.focus();

Figure 5-6 PVGIS HTML Result Page for a single roof showing the monthly global irradiance H_m and monthly yield E_m

Standard Assessment Procedure

The Standard Assessment Procedure (BRE, 2014) method for the estimation of Solar PV yield is an empirical model developed by the Building Research Establishment (BRE) in the UK. In order to be eligible to receive the FiT, Solar PV must be installed by Microgeneration Certification Scheme (MCS) accredited installers who are obligated to furnish adoptors with an estimate of the annual yield using SAP. This is calculated using an algorithm and data specified in the SAP guidelines. Because investment decisions and carbon savings are frequently based on this model this was selected for direct comparison with PVGIS. The electricity generated by a PV system with a rating R kW_p is given by Equation 5-9, where S_t is the annual insolation on the tilted plane in kWh and Z_{PV} is the shading factor (BRE, 2014, p96).

$$E = 0.8 \cdot R \cdot S_t \cdot Z_{PV} \qquad \text{Equation 5-9}$$

Z_{PV} is calculated heuristically by calculating the percentage of time the beam irradiance is obstructed using a sun-path diagram. S_t is obtained by summing each month's daily average insolation, S_{t,d_m}, multiplied by the number of days in the month n_m (Equation 5-10).

$$S_t = 0.024 \sum_{m=1}^{12} n_m \cdot S_{t,d_m} \qquad \text{Equation 5-10}$$

A month's S_{t,d_m} is calculated from the month's daily average horizontal irradiance, S_{h,d_m} and a factor, R_{ht_m} (Equation 5-11).

$$S_{t,d_m} = S_{h,d_m} \cdot R_{ht_m} \qquad \text{Equation 5-11}$$

S_{h,d_m} is taken from table U3 from the SAP manual using 1 of 24 UK regions for the system location, and the month. R_{ht_m} is a function of the factors which effect the angle of incidence of beam

irradiance on the plane of the PV array: the solar declination, δ, the latitude, ϕ, the aspect and the inclination, p. (Equation 5-12).

$$R_{ht_m} = A\cos^2(\phi - \delta) + B\cos(\phi - \delta) + C \qquad \text{Equation 5-12}$$

δ is taken from table U3 in the SAP manual, and is taken as the average solar declination for the month. The latitude, ϕ, is a representative value for the region. The constants A, B and C are determined from the inclination p and 9 constants k, which are functions of the orientation (Equations 5-13, 5-14 and 5-15).

$$A = k_1 \sin^3\left(\frac{p}{2}\right) \cdot k_2 \sin^2\left(\frac{p}{2}\right) \cdot k_3 \sin\left(\frac{p}{2}\right) \qquad \text{Equation 5-13}$$

$$B = k_4 \sin^3\left(\frac{p}{2}\right) \cdot k_4 \sin^2\left(\frac{p}{2}\right) \cdot k_6 \sin\left(\frac{p}{2}\right) \qquad \text{Equation 5-14}$$

$$C = k_7 \sin^3\left(\frac{p}{2}\right) \cdot k_8 \sin^2\left(\frac{p}{2}\right) \cdot k_9 \sin\left(\frac{p}{2}\right) + 1 \qquad \text{Equation 5-15}$$

Lookup tables for k values at 45° intervals are provided in table U5 in SAP version 9.2 (BRE, 2014). SAP guidance suggests that values for simulated solar PV systems with orientations within these intervals be interpolated.

SAP makes no use of information regarding the technology and BOS, and specifically loss factors in conversion of irradiance, particularly temperature. This is encapsulated in the hard-wired constant of 0.8 in Equation 5-9 which is equivalent to 20% system losses.

Table 5-4 Parameters required by SAP

Parameter	Options (units)
Region	1 of 24 values, used to look up horizontal irradiance and representative latitude
Nominal Power	(kWp)
Inclination	Angle (degrees)
Aspect	Orientation angle (degrees)

5.2.6 Summary

This section summarises the parameters discussed above, all of which are predictors of the PV system yield (Table 5-5). Additionally, through a thorough treatment of the theory behind solar PV yield, sources of uncertainty in its estimation have been elucidated.

The next section considers the data sources and assumptions utilised to furnish the model with these parameters for the geographic case study areas selected in Chapter 4.

Table 5-5 Predictor parameters for solar PV yield

Factor	Uncertainty
Aspect and Inclination	For a fixed array these parameters determine the time integrated beam irradiance and diffuse sky irradiance. However irradiance products will only ever give an estimate of actual insolation received by a system in a particular year.
Location	Since the latitude in particular influences the integrated irradiance of the tilted plane knowledge of the precise position of the PV system is important.
Insolation	Databases provide estimates of irradiance but these are modelled using empirical data. There are a number of uncertainties in these modelling algorithms in addition to the fact that these are historical values.
Semiconductor Material	Spectral responses and inherent efficiencies of different materials would make knowledge of this parameter useful for reducing uncertainty in yield estimates.
Power Rating	Different manufacturers will produce different ratings even if the material is the same due to different manufacturing processes. Knowledge of the module ratings could reduce uncertainty.
BOS	The performance of the system, particularly the inverter technology coupled with the high variability of irradiance ensure that even nominal power ratings of the BOS will only ever approximate the actual power throughput

5.3 Data Sources

There are no empirical data for solar PV generation on the domestic building stock located specifically in the four LSOAs discussed in chapter 4. This, therefore, needs to be estimated using the methods outlined in section 5.2. In order to evaluate the accuracy and precision of the simulated datasets, comparative empirical data have been acquired. The objective of this work is to be able to produce probabilistic datasets for the prediction of specific yield given available predictor parameters.

5.3.1 Conducting PVGIS Simulations

With many thousands of PV systems to simulate in this research the process was automated using a software script written in a Visual Basic for Applications (VBA) Excel Module. The script read the aspect, inclination, latitude and longitude data for each roof from a database. In all cases the PV module technology was assumed to be crystalline silicon, supported by evidence from empirical data. The nominal power was set to 1kW_p in order to return the specific yield in kWh/kW_p. System losses were set to 14% typical value for operational domestic UK systems. Simulations where conducted using both the ESRA, and the CMSAF climate models for estimating irradiance and for each of these both 'free' (rack mounted with free circulation of air underneath the modules) and 'building integrated' (no air circulation) settings for the mounting parameter were analysed. This gave four permutations for PVGIS estimations summarised Table 5-6 along with abbreviations used in the text.

Table 5-6 Four permutations of PVGIS estimation

Irradiance Model	Mounting	Abbreviation
ESRA	Building Integrated	ESRA-BI
ESRA	Free Standing	ESRA FS
CMSAF	Building Integrated	CMSAF-BI
CMSAF	Free Standing	CMSAF-FS

The VBA code functions as follows. An internet protocol HTTP request is constructed in the proprietary format (Huld, 2012) of a PVGIS web form request. This contains the name value pairs for all the parameters in Table 5-3. A response string is returned with contains monthly yield and irradiance data in the plane of array. This is parsed and summed to give the annual specific yield and insolation values and these are stored back in the database.

5.3.2 Conducting SAP Simulations

The equations and lookup tables discussed in section 6.2.6 where encoded as functions and look-up tables in a Microsoft Excel spreadsheet. The SAP method requires the values pairs in Table 5-4. The geographical co-ordinates given for the simulated PV systems were converted to the region code required by SAP. This was achieved automatically using the Postcode Address File (Ordnance Survey, 2014) to return the postcode of each location. This enabled the lookup of the correct region code using table U4 in the SAP documentation (BRE, 2014). The specific yield was stored in the database with the corresponding roof being simulated.

5.3.3 Variability of G_H Estimated by PVGIS for the Case Study Areas

PVGIS takes as parameters the absolute longitude and latitude of the PV array, and the claimed spatial resolution is of the order of 1km. This is similar to the dimension of an urban LSOA and therefore it was important to assess the spatial variation of the horizontal insolation to test for

discontinuities in the PVGIS model and determine whether, over the spatial scale of the LSOA, the irradiance could be assumed to be constant.

For each LSOA the global horizontal irradiance estimated by the CMSAF and ESRA irradiance models, was analysed over a one square kilometre grid at a spatial resolution of 20m in both West to East and South to North directions. The results are presented in Table 5-7. The average irradiance for the four areas reflects the expected correlation with latitude. Also observed is the higher insolation returned by the CMSAF database relative to ESRA as discussed above.

Table 5-7 Analysis of variation of horizontal insolation predicted by PVGIS

	Annual Insolation kWh/m²				Difference CMSAF ESRA/RSun (%)
	Average		**Coefficient of Variation (%)**		
LSOA	**PVGIS-classic**	**PVGIS-CMSAF**	**PVGIS-classic**	**PVGIS-CMSAF**	
Kerrier 008B	1093	1170	0.14	1.30	7.1
Charnwood 002D	948	1055	0.23	0.31	11.3
Kirklees 042B	942	976	0.05	0.05	3.7
Newcastle 008G	922	989	0.07	0.01	7.3
Measurements over 1km^2 grid with a longitudinal and lateral vertical spatial resolution of 20m					

The coefficient of variation (CV) is shown in Table 5-7 for both satellite and interpolation methods. For all the areas the CV is less than 0.25%, except for Kerrier 008B, using CMSAF, where the CV is 1.3%. Whilst small, this anomaly warranted further investigation and is represented graphically in Figure 5-7. A sharp step-change in irradiance is observed running horizontally (West to East) with approximately 30kWh/year difference in annual insolation, a difference of 2.6%.

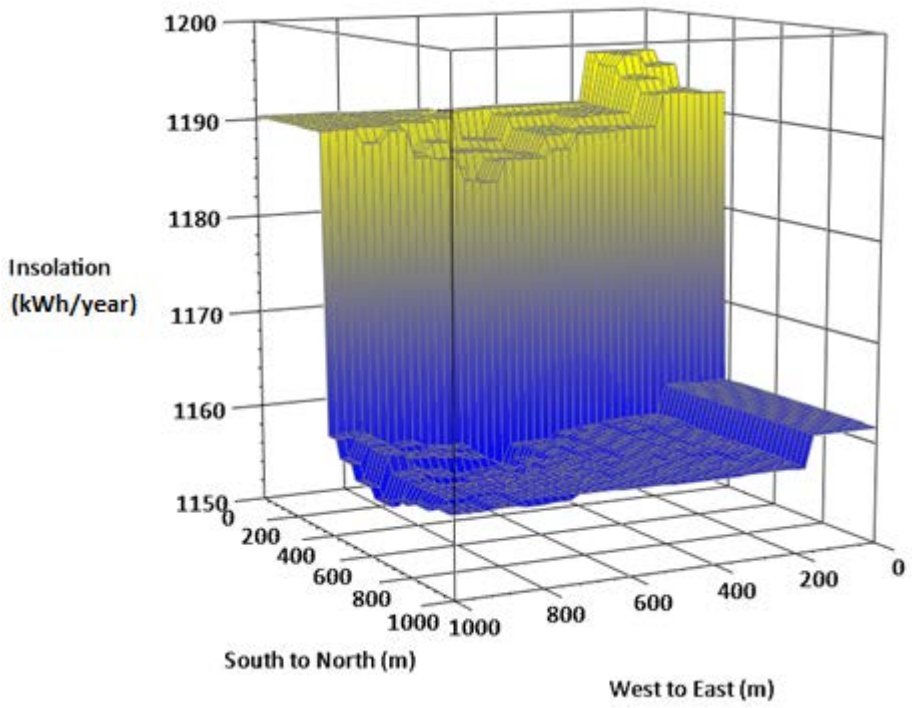


Figure 5-7 Variation in annual insolation predicted by PVGIS

This is perceived to be an artefact of the PVGIS CMSAF model which relies on the measured brightness of individual pixels in digital photography taken by satellites which have a spatial resolution of about 10km. Since this furnishes the database with an average irradiance for the whole pixel it is possible that pixel boundaries will deliver such sharp transitions. Huld (2014) has suggested this is likely in coastal areas where there can be large differences due to coastal mist impacting on the diffuse fraction.

It has been shown that over the spatial scale of the LSOA, the irradiance could be assumed to be constant. This supports the exclusion of longitude and latitude parameters in the BN submodel.

5.3.4 Comparison of G_H and Specific Yield estimation with SAP and PVGIS

The calculation of insolation using SAP employs a number of heuristic equations. The parameters are homogeneous within any of the 22 regions used by SAP (Figure 5-8). For this reason the analysis in the previous section is not relevant for SAP since all the values on LSOA 20m grid are estimated to be the same, and the CV is zero for all four LSOAs. It is worth noting this when considering the total uncertainty.

Figure 5-8 SAP regions for prediction of irradiance in the UK (BRE, 2014)

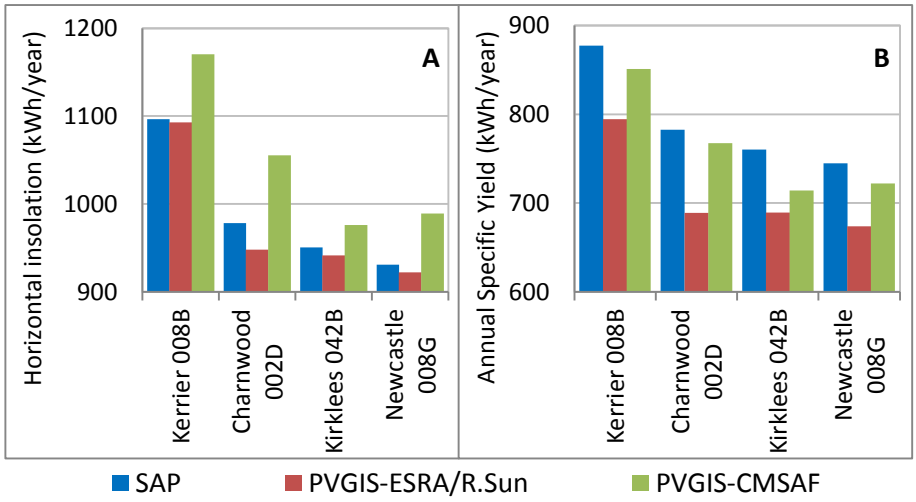


Figure 5-9 Comparison of Insolation estimated by PVGIS and SAP

Figure 5-9A shows a comparison of insolation estimates for SAP and the two irradiance databases used by PVGIS. SAP agrees more with the ESRA database than CMSAF, which is consistently higher. Given that SAP uses an average irradiance over larger geographical regions not much can be inferred from this except to say the more granular estimations by ESRA concur with SAP in three of the LSOAs and CMSAF in one other. In Figure 5-9B the specific yield for horizontally mounted arrays estimated by the same three irradiance models are compared for each LSOA. The yields follow the same pattern as the insolation and there appears to be no discernible climatic effects due to average ambient temperature variations.

5.3.5 Simulation of Yield as a Function of Pitch and Aspect and LSOA

The previous section has given confidence that irradiance and specific yield can be regarded as constant on the spatial scale of an LSOA. Comparison to SAP has been made which, by design, assumes a constant irradiance on a regional scale. Significant differences are noted for the average values for each LSOA which is due to the latitude as discussed above.

This constancy merits a simple model where only the tilt and aspect of the PV system, as well as the specific LSOA, are required to estimate the specific yield. For each LSOA, simulations were executed to estimate the predicted PV yield for each permutation of pitch, and aspect, at 2.5° and 5.0° intervals respectively. This creates a matrix of 2701 specific yield values. Such a matrix allows the estimation of specific yield for every conceivable orientation of a roof mounted PV array. In Chapter 7, the analysis of the building stock shows that not all orientations are of interest – only roofs facing towards the southern hemisphere (90° to 270°) and those which are either flat, or have a pitch from 20° to 50° were observed. Thus the size of matrix could be considerably curtailed. However, the model was furnished with a full complement of orientations and aspects in order to create a true object-oriented component which is reusable for other purposes.

A matrix can be generated for the SAP estimations and 4 permutations of PVGIS estimation in Table 5-6. A 3-D representation of such a matrix is shown in Figure 5-10, which has been generated for LSOA Kerrier 008B using the CMSAF-BI. PVGIS also allows the choice of Solar PV semiconductor technology (see Table 5-3). Since the majority of systems installed in the UK are either polycrystalline or monocrystalline silicon the option of Crystalline Silicon was chosen in all simulations.

In Figure 5-10 there is symmetry around the aspect of 180°, at which the irradiance peaks for all possible values of the pitch, except for a flat roof, which, logically, does not have an aspect. The variability is particularly stark for a vertical surface. As the tilt become less acute there is an optimal value around 38° at which the cosine law, for a particular latitude, maximises the plane of array irradiance.

The estimated specific yield data now provide a method of estimating the specific yield of a PV system installed on any roof in the LSOA, regardless of its orientation and aspect. The next section evaluates the estimation of specific yield in this way by making an analytical comparison with empirical data collected in the UK.

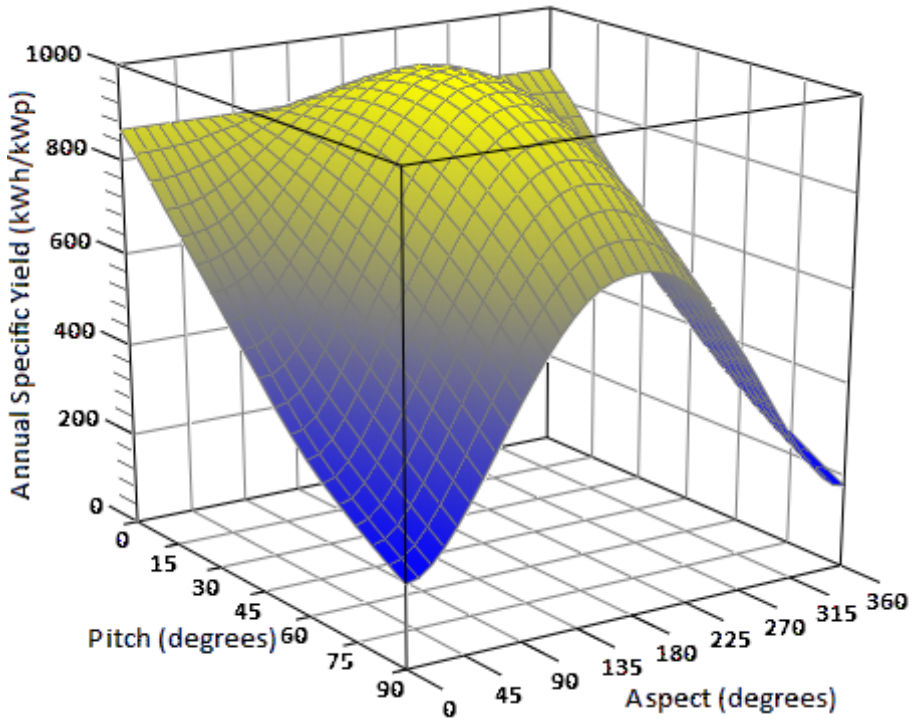


Figure 5-10 3-D representation of matrix of annual specific yield as a function of pitch and aspect

5.3.6 Empirical Solar PV Data in the UK

The evaluation of solar PV yields requires the collection of generation data at an appropriate temporal resolution for a number of systems defined in terms of their installation geometry, rating, technology and location. Over 400 systems where monitored for the UK photovoltaic domestic field trials (PDFT) at a temporal resolution of 5 minutes from 2001 to 2005 (BRE, 2006). Data was automatically collected using sensors and data loggers. The distribution of the annual specific yield is shown in Figure 5-13 (page 116) in comparison to the Sheffield Microgeneration Dataset discussed in Section 5.3.7. The expected value for the specific yield for this distribution is 670 kWh/kW_p but the field trials showed that there was a wide variability in performance in the UK. There are two issues which suggest the data may not be representative of systems contemporary with this research. Firstly, German studies have shown a marked improvement in performance of systems installed

towards the end of the first decade compared to those installed at the end of the last century (Reich et al., 2012). Secondly, the studies were part of an integrated programme of installation and associated monitoring. Thus, the conditions experienced now, under the UK feed-in tariff, the rapid growth of which has attracted many start-up companies and novice installers in to the industry, may not be reflective of those installed over 10 years ago. To address these potential issues contemporary data were sought for this research.

Dedicated field trials were prohibitively expensive and complex to set up. A number of commercial installers and inverter manufacturers offer real time automatic uploading of generation data to web portals using internet or GSM connectivity in order to offer value-added services to customers such as system monitoring, fault detection and reporting analytics. Such data however is rarely accessible to researchers due to confidentiality and data protection legislation. An alternative source is data donor projects (Leloux, et al., 2012A; Leloux, et al., 2012B). With the increasing adoption of solar PV by UK households, and the ubiquitous access to the internet, a number of projects have appealed to the PV user community to donate monthly yield data on a purpose built website. For this project, data has been obtained from such a project called the Sheffield Microgeneration Database (SMD) (Colantuono et al., 2014).

5.3.7 Sheffield Microgeneration Dataset

The data obtained from the SMD consists of over 6000 monthly generation readings donated by over 600 system owners collected from 2010 to 2013, with the majority of monthly readings collected during 2012 (Figure 5-11A). As well as the publically available generation data, the SMD project also provided the system data parameters shown in Table 5-8. To prevent disclosure of personal data the longitude and latitude where randomised to within 1km of the true co-ordinates (Everard, 2013).

The system rating is calculated from the peak power (W_p) of a single module, multiplied by the number of modules. The distribution of system rating (Figure 5-11B) shows a higher tendency for adopters to install systems with ratings of 2, 3 and particularly 4 kW_p. The distibution of the aspect (Figure 5-11C) shows a large peek at 180°. If the aspect is considered random i.e. all southerly facing roofs are equally likely to receive an installation. this peak is anomolous, or, if south facing roofs are more likely to receive an installation then this may be correct. Colantuono et al. (2014) have suggested that this is due to the datasheet, provided by the installation company for the client, having a propensity to state south facing. An exploration by them of a larger dataset indicated that there was a tendency for systems to be declared to be facing east, southeast, south, southwest or south, with a higher probability stating simply south, rather than a more precise angle. The distribution of roof pitch (Figure 5-11D) shows that the majority of roofs have a pitch between 25 and 45 degrees which is representative of the building stock data discussed in Chapter 7. The location of the PV systems in the dataset is shown in Figure 5-12. This demonstrates a widespread spatial distribution of the dataset, and in particular, shows data points in the Northeast, Yorkshire, the Midlands and Southwest, where the selected LSOAs, used in this study, are located.

In summary, the SMD dataset contains data for a significant number of PV systems with representative distributions of system rating, aspect, pitch and geographical location. There is some uncertainty in the owner provided data; in particular, there is a propensity for systems to be declared south facing. The dataset has enough parameters to execute simulations, in the same way as has been carried out for the four LSOAs using PVGIS and SAP, with which to make comparisons with the generation data donated by system owners. This is presented in the following sections.

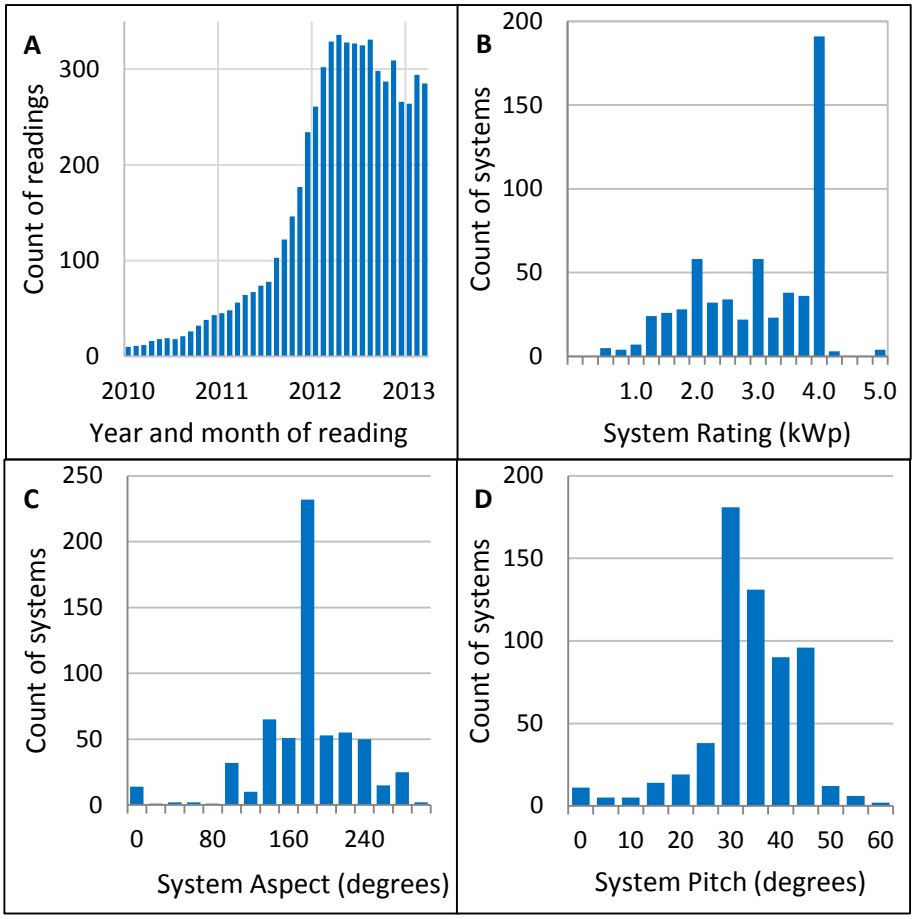


Figure 5-11 Characteristics of the Sheffield microgeneration dataset

Table 5-8 SMD System Data Parameters

Parameter	Comment
Latitude	Randomised to 1km
Longitude	Randomised to 1km
Inverter Make	Datum not always present
Inverter Model	Datum not always present
Pitch	Degrees
Aspect	Degrees from North
Height	Height of system above ground
Number of modules	Allows calculation of system rating
Manufacturer	Datum not always present
Model	Datum not always present
Wp	Peak power of a single module

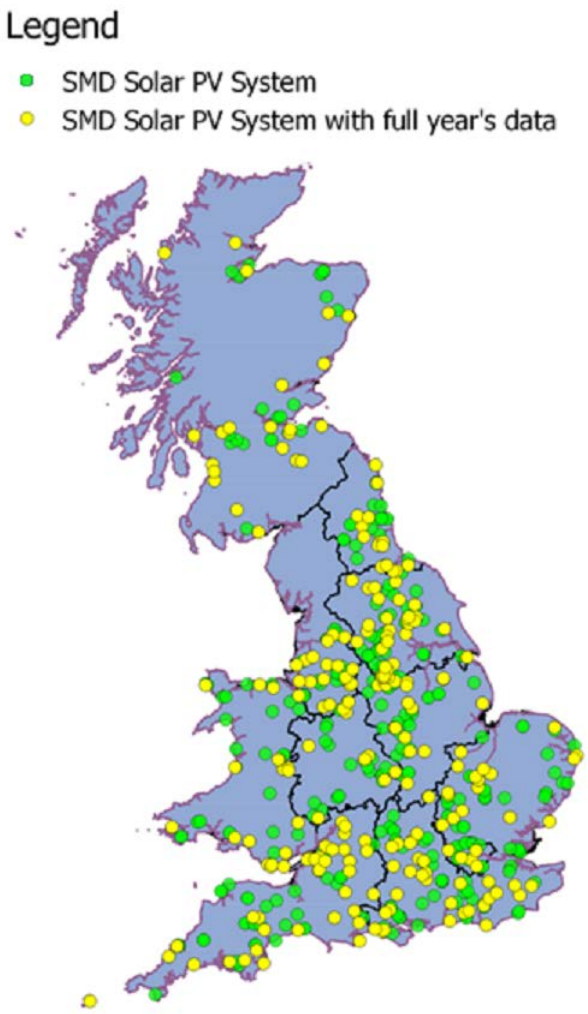


Figure 5-12 UK Locations of PV Systems in the SMD dataset

5.3.8 Annual Specific Yields of SMD Solar PV Systems

In order to calculate specific yields, the monthly generation data were aggregated by month and the average monthly value over 12 months was summed to give the annual yield. Only 245 systems had a full year's data and were selected for further analysis. These are differentiated by colour, on the map shown in Figure 5-12. This delivers the annual specific yields of 245 UK Solar PV systems. The probability distribution for the specific yield is shown in Figure 5-13, discretised in 100kWh bins. The yield distribution obtained from the PDFT data, discussed in Section 5.3.6, is also shown for comparison. The expected value for the SMD dataset is 855kWh/kW_p which is 184 kWh higher than

that for the PDFT. The comparison also demonstrates that there are less failing systems; 10.2% in the PDFT dataset and only 2.0% in the SMD dataset had annual specific yields less than 500 kWh.

This comparison with the earlier dataset indicates, similar to the German study (Reich et al., 2012), that contemporary systems are performing better than systems installed ten years ago, and supports the acquisition of this dataset.

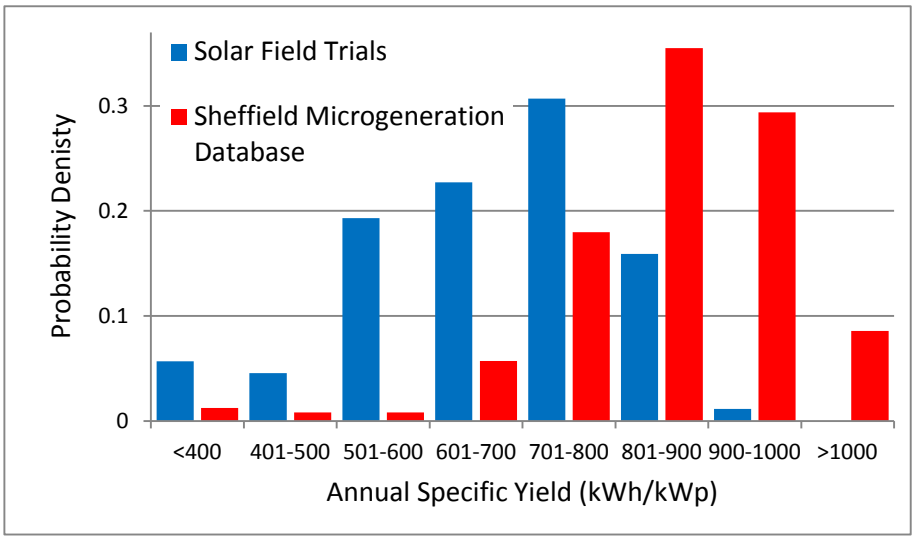


Figure 5-13 Comparison of specific yield distribution for PDFT and SMD PV systems

5.3.9 Correlation of Measured and Estimated Specific yields for SMD PV Systems

Using the aspect, pitch and geographic co-ordinates for each PV system in the SMD the same method employed for estimating solar yield in the 4 LSOAs was carried out. This allowed a direct comparison between measured yields and estimated yields using regression and statistical analytical methods. The four permutations (see Table 5-6) for yield estimation using PVGIS, as well as PVSAP, were compared.

Table 5-9 Comparison of correlations between measured and estimated specific yield for SMD Systems

Estimation Method	Average Measured	Average Estimated	MBE	%MBE	RMSE	%RMSE	SE_Y
SAP	867	867	0	0.0%	95	11.0%	101
ESRA BI	867	791	-76	-8.8%	124	14.3%	136
ESRA FS	867	829	-38	-4.3%	104	12.1%	116
CMSAF BI	867	860	-6	-0.7%	95	11.0%	101
CMSAF FS	867	902	36	4.1%	102	11.7%	95

Table 5-9 presents the results of this analysis. The mean value for measured specific yield for all 245 PV systems is 867 kWh. The mean bias error (MBE) and root mean square error (RMSE) were calculated using Equation 5-16 and 5-17 respectively. SAP delivered the most accurate estimation with an MBE of zero. The ESRA-BI and ESRA-FS systems underestimated on average by -76kWh -36kWh respectively, whilst CMSAF-BI underestimated by only -6 kWh, but overestimated by 36kWh for CMSAF-FS.

$$MBE = \frac{\sum(SY_{estimated} - SY_{observed})}{n} \qquad \text{Equation 5-16}$$

$$RMSE = \sqrt{\frac{\sum(SY_{estimated} - SY_{observed})^2}{n}} \qquad \text{Equation 5-17}$$

The MBE were all under 10% which compares favourably with other studies (Thevenard and Pelland, 2013). The precision, as furnished by the RMSE is quite consistent for all estimation methods, ranging from 11.0 to 14.3%. The RMSE is analogous to the standard deviation of the error and it can be estimated that approximately 66% of observations are within the estimate value ±100 kWh and 95% in ±200 kWh.

To test the efficacy of the estimated yield as a predictor of measured yield a regression analysis was undertaken. Figure 5-14 shows, for each simulation method, the line of best fit using linear least

squares regression analysis. The charts also show the line of equality where the estimated yield equals the measured yield, and the 95% (2σ) confidence intervals.

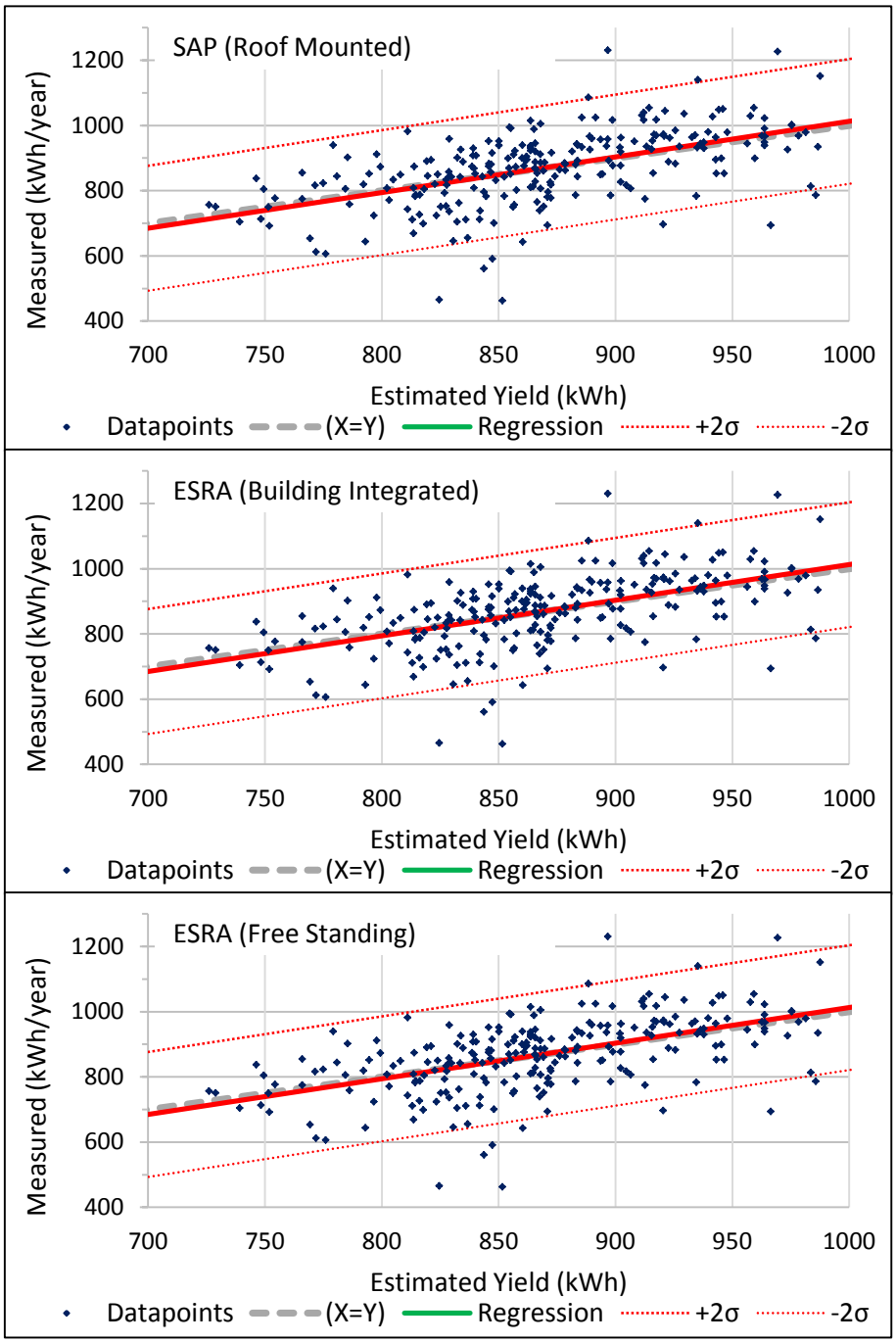


Figure 5-14 regression analysis of estimated against measured specific yield

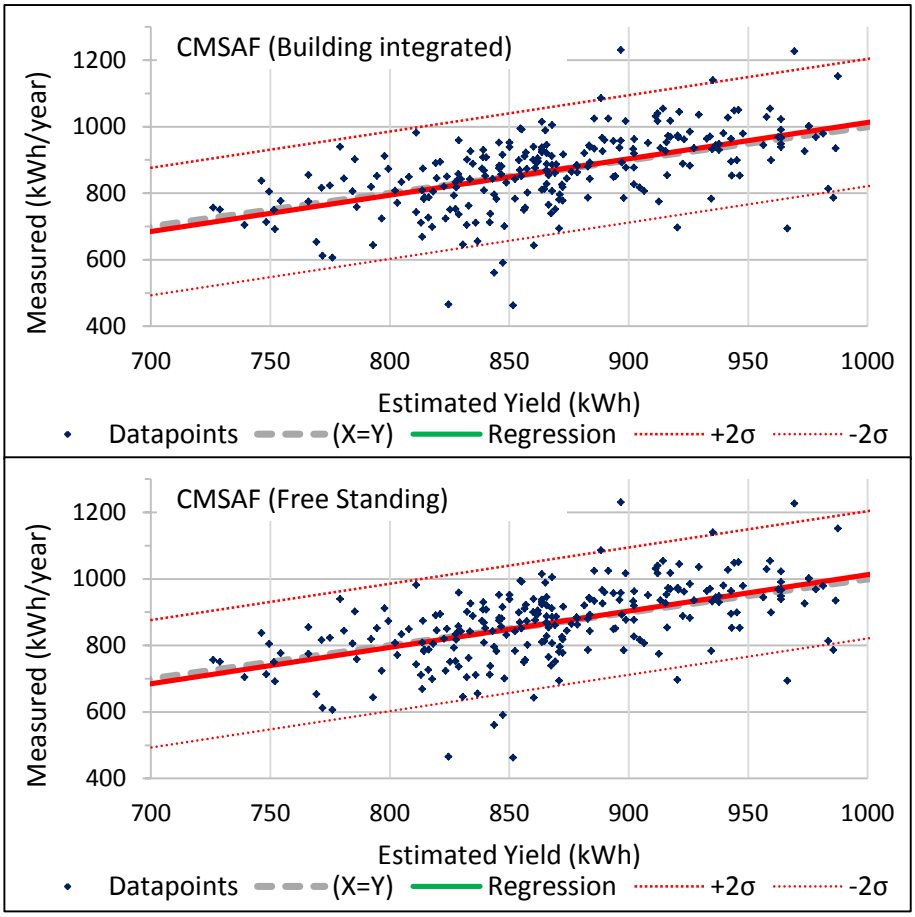


Figure 5-14 (continued) regression analysis of estimated against measured specific yield

The standard error of the estimation $\mathrm{SE_Y}$, given by Equation 5-18 (Lane, 2007), for these regression analyses are shown in Table 5-9.

$$SE_Y = \sqrt{\frac{\sum\left(SY_{predicted} - SY_{observed}\right)^2}{n-2}} \qquad \text{Equation 5-18}$$

There are three aspects to this analysis pertinent to producing a BN model for predicting specific yield. Firstly the analysis showed that SAP and CMSAF/building-integrated gave the most accurate results with the smallest MBE. Secondly following linear regression the CMSAF/Free-Standing gave the most precise results with an $\mathrm{RE_Y}$ of 95 kWh.

The third aspect concerns the standard assumptions of ordinary regression analysis namely that the variance is independent of the property being measured (homoscedasticity) and that the errors are normally distributed. Inspection of the distribution of residual errors for each of the regression analyses depicted in Figure 5-14 demonstrated that the distribution is not normally distributed but positively skewed. Figure 5-15 shows one such example using the CMSAF/Free-standing estimation model. This indicates a higher probability, when using these regression analyses, of over estimating the observed yields.

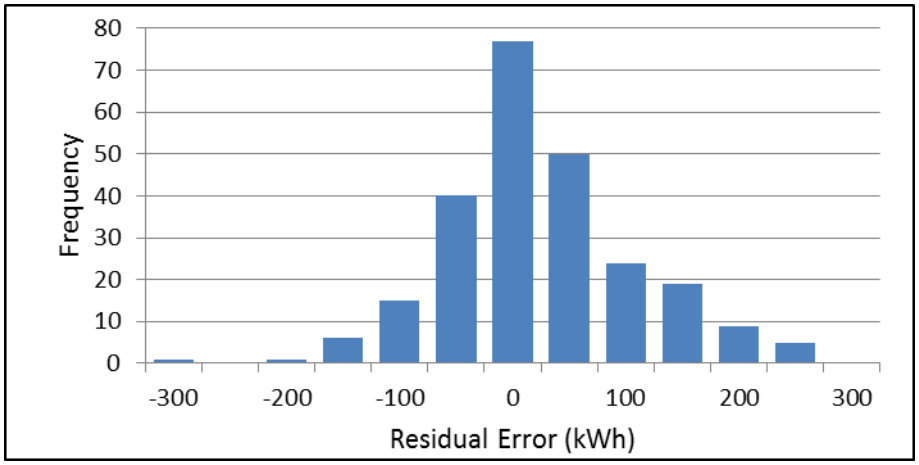


Figure 5-15 Residual errors for PVGIS estimation using CMSAF/Free-standing

Homoscedasticity was tested by taking the residual error of each data point as a percentage of the predicted value (percentage relative residual error), and plotting this against the predictor variable (in this case the estimated yield). Figure 5-16 shows this for the CMSAF/Free Standing estimation model. The skewed nature of the residual error distribution and a uniform degree of scatter across the range of estimated yields are apparent in this graph.

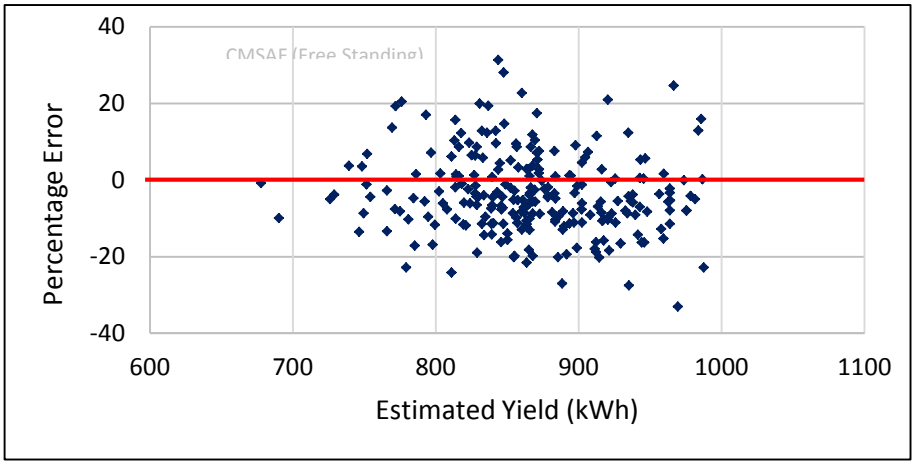


Figure 5-16 Residual error of each data point as a percentage of the predicted value for CMSAF/Free Standing estimation model

5.3.10 Prediction of Specific Yield and Uncertainty

The linear regression lines are valid as calibration curves for predicting real UK PV yield as a function of estimated yield; by definition the mean error of estimation is zero. The key factor to differentiate between the yield estimation models is the standard error of estimation, SE_Y, for which the lowest value of 95 kWh was obtained using the CMSAF-FS estimation model. This was therefore selected as the calibration curve for the prediction of specific yield; the calibration statistics are shown in Table 5-11 and the curve is represented by Equation 5-19.

Table 5-10 Specific Yield Estimation model calibration curve statistics

Statistic	**Value**
Slope m	0.828
Intercept c	119
Coefficient of determination r_2	0.33
Standard error of estimation SE_Y	95
F Statistic	117

$$SY_{observed} = 0.828 \times SY_{PVGIS_{CMSAF-FREE}} + 119$$ Equation 5-19

It has been shown that the model is homoscedastic, meaning that the variability in SE_Y is independent of the estimated yield. Furthermore, this variability is right-skewed independently of estimated yield. This shape of distribution resembles a number of well-known probability distributions including the Weibull, Lognormal, Beta and Gamma distributions. Best fit parameters for each of these was found using the Microsoft Excel Solver software add-in to minimise the sum of squared errors (John and Grosvenor, 2001).

The Gamma distribution gave the best statistical fit with a coefficient of determination of 0.999, although all the distributions were reasonable in reflecting the small degree of skew. Figure 5-17 shows the resultant parameterised fit, expressed as a cumulative gamma distribution with a cumulative normal distribution for comparison. Using this function the uncertainty in the standard error can be expressed as Equation 5-20.

$$SY = \frac{\beta^{\alpha} x^{(\alpha-1)} e^{-x\beta}}{\Gamma(\alpha)}$$ Equation 5-20

Where $\alpha = 19.25$, $\beta = 18.12$ and $\Gamma(\alpha) = (\alpha - 1)!$

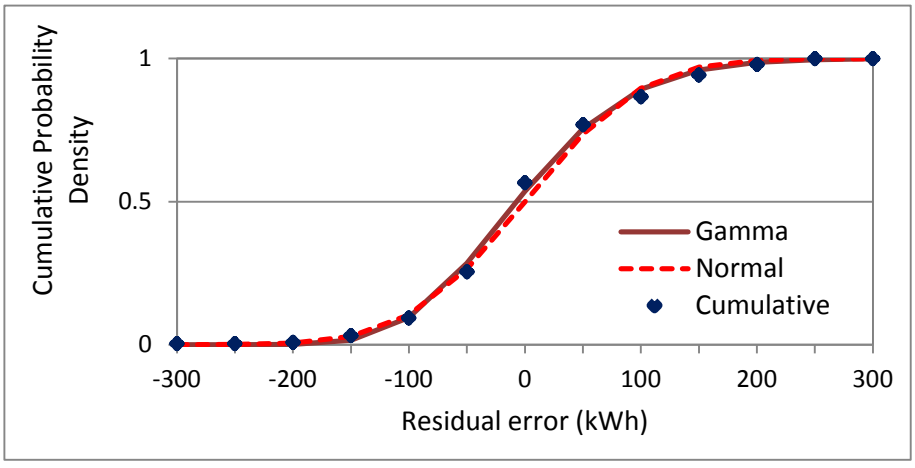
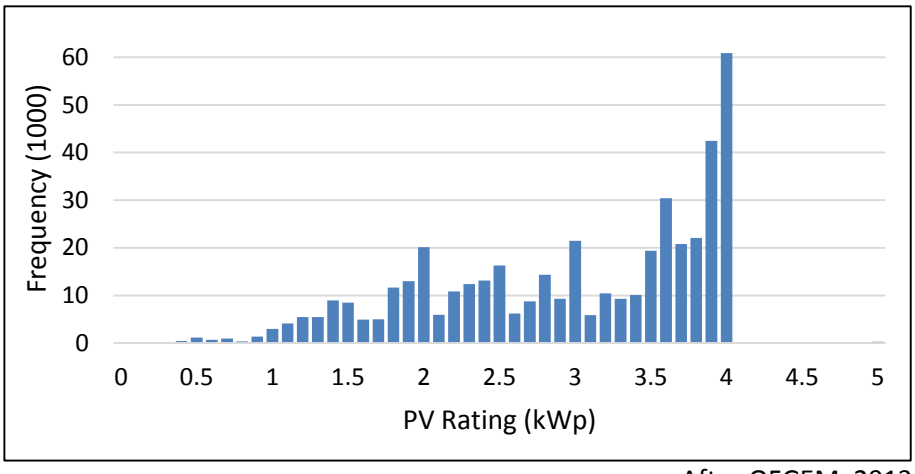


Figure 5-17 Cumulative distribution of residual error in PVGIS prediction

5.3.11 Estimation of System Yield

Once an estimation of specific yield is available the system yield of installed systems can be determined with knowledge of their system rating. Assuming that PV adopters maximise the use of available roof estate the rating can be estimated from the area of the roof and the size of commercially available modules. Evidence from the OFGEM FiTs register suggests, however that economic and regulatory factors are also a significant determinant of system ratings. Figure 5-18 shows the frequency of PV systems as a function of their ratings. The striking cut-off at 4kW$_P$ is the upper threshold of the first band for the FiT, above which there is an approximately 10% reduction in the tariff. There is also a regulatory barrier to systems over 4kWp which is due to a 16A maximum (equivalent to 3.68kW at 230V) which grid-connected generation equipment is permitted to feed in to a single phase of the low voltage network (Energy Networks Association, 2003). In the UK a 4kW$_p$ system can be safely coupled with a 16A inverter since the likelihood of exceeding 3.68 is very small indeed. Above this limit installers have to apply for permission from the network operator which may incur additional expense for grid reinforcement. The simplification of connection for systems 4kWp or less has reduced technical and economic barriers (Marsh, 2004), but the de facto upper limit shows that these remain for larger systems.



After OFGEM, 2013

Figure 5-18 Frequency of solar PV installations as a function of system rating on the FiT register

A further observation from Figure 5-18 is the propensity of systems to be clustered at 0.5 kW_p intervals. Solar PV system ratings are matched to inverters which system designers have available. Thus whilst the distribution of roof size may be continuous, the PV system sizes are affected by discontinuities due to available inverter sizes which are often rated to 0.5 kW intervals.

Indiscernible from Figure 5-18 is the role of economic factors in rating selection; an investor may have a roof large enough to accommodate a 3 kW_p system, but may only have the capital to realise the purchase of a smaller one.

All the factors above can contribute to roof under-utilisation. However, this, without an empirical study is difficult to quantify. In the absence of this, recourse is made to a simple heuristic relationship derived using the SMD (Section 5.3.7), the analysis of which allows the distribution of rating per square meter for these system to be plotted (Figure 5-19). The small peak centred at 0.19 kW_p/m^2, 19% of the total sample, is due to modules equipped heterojunction with intrinsic thin layer (HIT) technology which have a higher yield per unit area than standard silicon PV systems (Leloux et al., 2012). The distribution in Figure 5-19 can be used to estimate the probability of the system rating given the available roof area

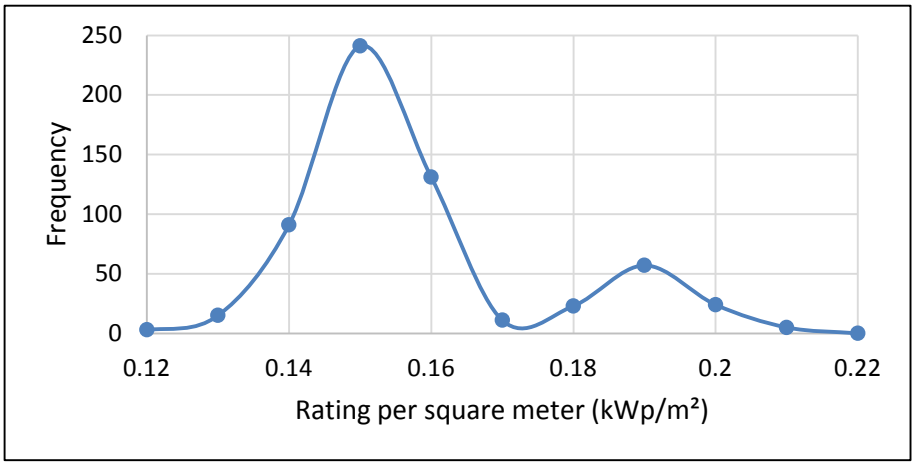


Figure 5-19 The rating density of solar PV modules deployed in the Sheffield microgeneration database sample

5.3.12 Summary

In this section three achievements have been presented. Firstly the utility of two estimation models has been evaluated and a dataset created using these models to estimate the specific yield as a function of aspect and pitch for specific regions corresponding to the case study areas. This created a data matrix for each LSOA and choice of estimation models and parameters.

Secondly this deterministic method of predicting yield has been augmented by using empirical data from the SMD. PVGIS using the CMSAF-FS configuration delivered the lowest variance when used to predict observed data. This calibration of PVGIS using the SMD data was quantified using a linear regression and the uncertainty quantified using a gamma distribution to account for the skewness of the residual errors which tends to over estimation.

Finally consideration has been to actual system ratings which are deployed in the UK contexts and a heuristic method of estimating probabilistically the system rating from a given roof area. The next step is to capture this analysis in the form of a BN model in Section 5.4

5.4 Bayesian Network Submodel for Solar PV Prediction

The domain ontology (Section 5.2) and the data sources acquired or simulated (Section 5.3) were used to guide the construction of a Bayesian Network sub-model. In Section 5.4.1 the DAG is presented. This is furnished with the quantitative data to construct the CPTs.

5.4.1 The Directed Acyclic Graph

The outcome of the review of the literature is that orientation, pitch and geographic location are the important predictors of PV yield. It was shown that the longitude and latitude for PVGIS estimations can be represented by a single value for the entire LSOA. The location can be represented by a

regional parameter corresponding to the LSOA. The PV technology and BOS are also important predictors of Yield. The semiconductor technology for the PV systems in the SMD is known to be 98% crystalline silicon PV modules (Taylor and Buckley, 2014). It was assumed that any hypothetical penetration of PV into the LSOAs would have the same technology distribution as the SMD therefore no node to represent the technology was included. The DAG to summarise the available parameters and their dependencies is shown in Figure 5-20.

The Simulated Yield node is dependent on orientation, pitch and region. In order to endogenise the uncertainty represented by the error of estimation, the measurement idiom (Chapter 3) was employed. The PVGIS estimation model delivers, using Fenton and Neil's (2013, p178) terminology, namely the "*actual value of the attribute"*, albeit in this case simulated. The measurement idiom requires a second parameter which delivers the *assessment accuracy* of the actual value. In the DAG the assessment accuracy is represented by the Yield Uncertainty node. These two nodes are parent nodes of the *assessed value of the attribute*, which gives the actual value augmented by the uncertainty in the assessment accuracy. In the DAG this is represented by the Specific Yield node. The System Yield is predicted by the Specific Yield and the System Rating which in turn is predicted by the Roof Area and Rating Density.

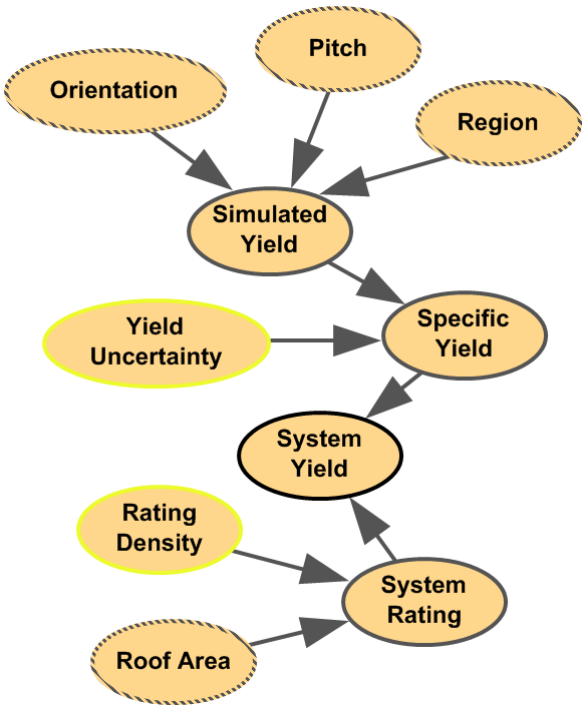


Figure 5-20 DAG for the Specific Yield BN Submodel

5.4.2 Node Probability Tables

The data sources discussed in section 5.3 were used to construct the NPTs using the methods discussed in Chapter 3 for Netica. The approach for each node is summarised in Table 5-11. Only the region is a discrete node; all others are discretised to an appropriate degree for performance efficiency and fidelity to the continuous distributions. The values listed were found not to compromise either.

Table 5-11 Summary of approach for PV yield model nodes

Node	NPT	Type	Discretisation	Units
Orientation	Case File/Learning	Continuous	10	Degrees
Pitch	Case File/Learning	Continuous	5	Degrees
Region	Case File/Learning	Discrete	n/a	n/a
Simulated Yield	Case File/Learning	Continuous	20	kWh
Yield Uncertainty	Equation to table	Continuous	10	kWh
Specific Yield	Equation to table	Continuous	20	kWh
System Yield	Equation to table	Continuous	200	kWh
System Rating	Equation to table	Continuous	250	W_p
Rating Density	Case File/Learning	Discrete	0.01	kW/m^2
Roof Area	Case File Learning	Continuous	5	m^2

The matrix for simulated yield as a function of LSOA, orientation and pitch, cast as a table, provides the case file for NPT learning. Table 5-12 shows the first 5 of 2812 rows. There is only one case per combination of variable.

Table 5-12. Top 5 rows of the case file for learning NPTs for region, orientation, pitch and yield

Region	Orientation	Pitch	Simulated Yield
Camborne	0	0	901.10
Camborne	0	5	855.80
Camborne	0	10	805.96
Camborne	0	15	756.63
Camborne	0	20	707.77

The NPT for the yield uncertainty node is created by configuring the node with Equation 5-20 and using Netica's equation to table feature (Chapter 3). The generated discrete probability distribution is shown in Figure 5-21. This has been modelled in Excel and Netica in order to verify the BN software's interpretation of the gamma function.

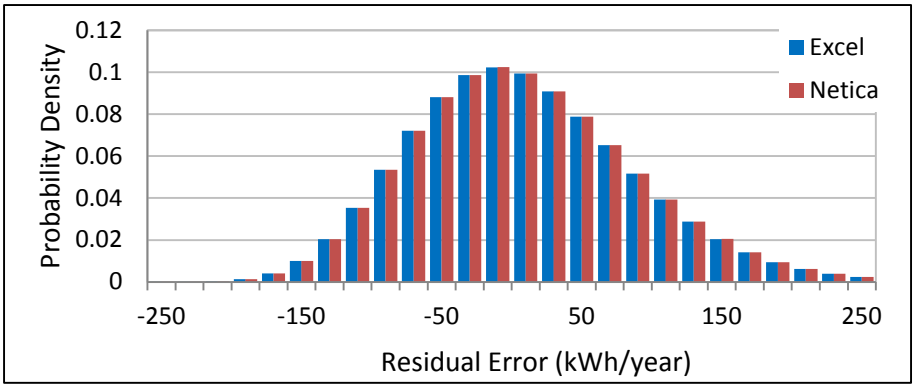


Figure 5-21 Yield uncertainty modelled by the gamma distribution discretised into 20kWh intervals.

The specific yield, SY, probability distribution was generated using the equation to table method, taking the simulated yield, SY_{SIM}, and the yield uncertainty, YU, as inputs. The simulated yield is adjusted using the calibration curve and (Equation 5-19) and is added to the uncertainty variable. This is shown in Equation 5-21.

$$SY(SY_{SIM}, YU) = 0.828 \times SY_{SIM} + 119 + YU \qquad \text{Equation 5-21}$$

Netica automatically calculates the joint probability distribution, $P(SY|SY_{SIM}, YU)$ as discussed in chapter 3.

The prior distribution for the Rating Density is taken from the analysis in Figure 5-19. From this, and the roof area, which is an input node, the System Rating is calculated as a product of the Roof Area and Rating Density with conditionality such that if the rating is greater than 5 it is set to 5, and if the Roof Area is less than 10, the System Rating is set to zero.

5.4.3 Netica Specific Yield BN Sub-model

The resultant BN submodel in Netica is shown in Figure 5-22. The model was verified in a number of ways to ensure its behaviour reflected the source data.

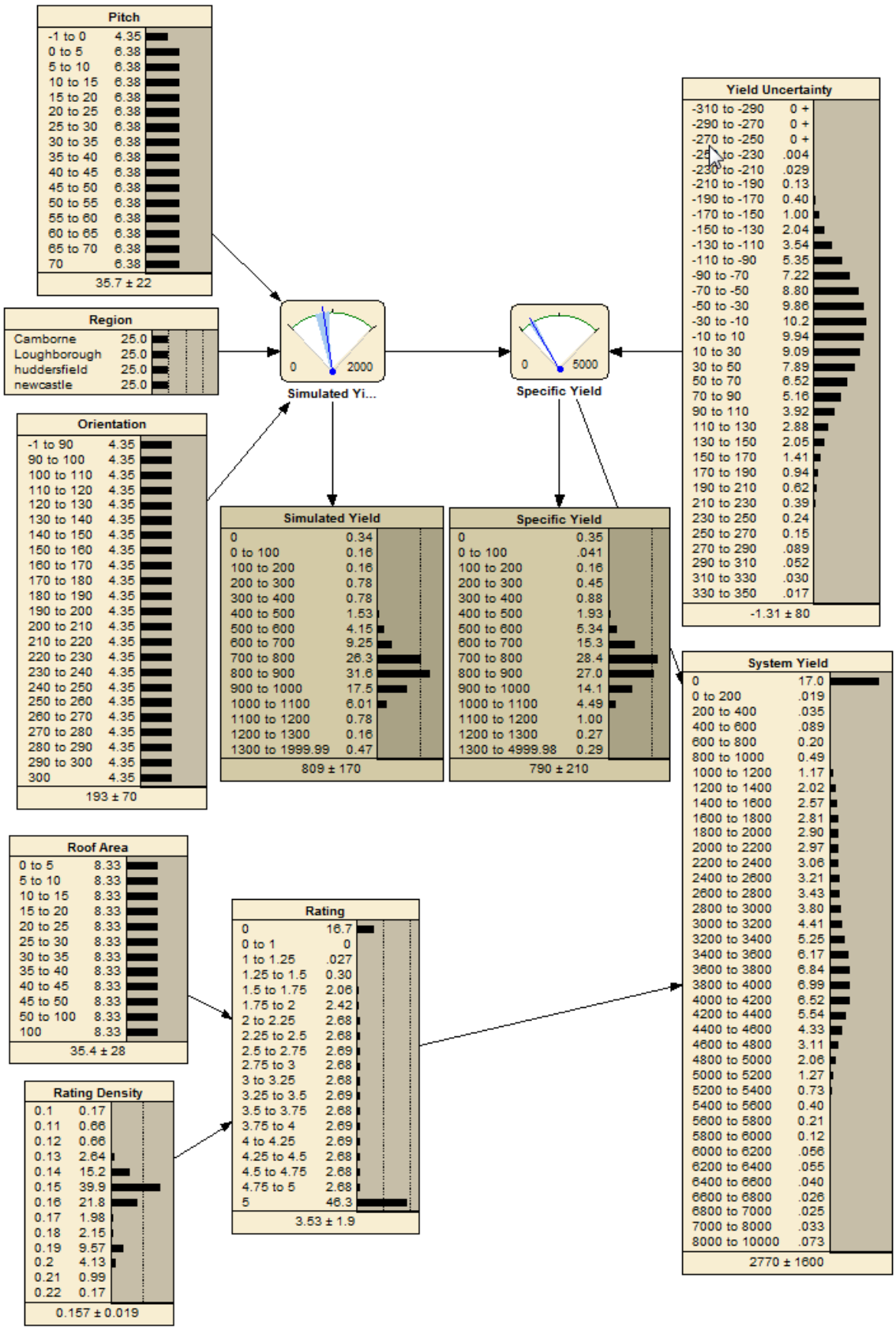


Figure 5-22 Bayesian Network Submodel in Netica

The orientation, region and pitch are uniform distributions. This is expected since the case file has been generated with regular intervals of orientation and aspect which gives one estimated yield for each combination of the two intervals and region.

When selecting hard evidence for the pitch, orientation and region the posterior state for the simulated yield should correctly predict the actual measured value using PVGIS. This was verified for several pitch and orientations for each region against the value in the case file and found to be correct.

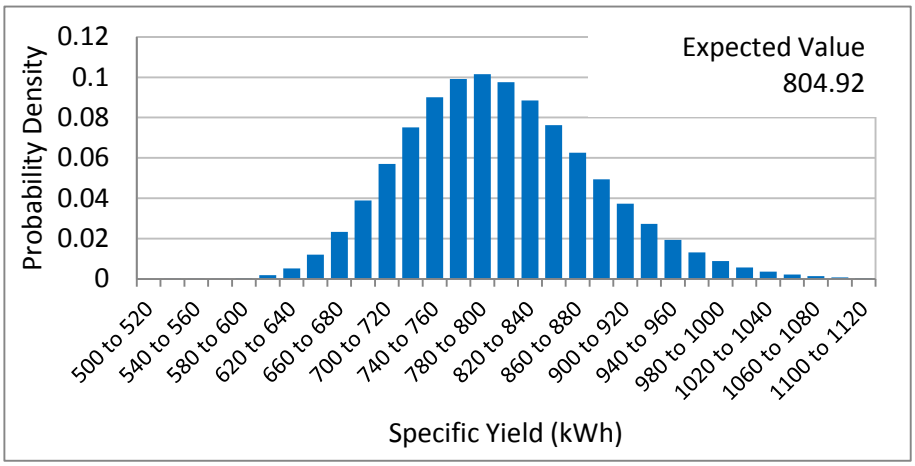


Figure 5-23 Posterior distribution for Specific Yield with hard evidence for simulated yield

A typical posterior distribution for the specific yield is shown in Figure 5-23, resulting from hard evidence for the state 820-840. This has an expected value of 805, which compares to the value of 806 for the predicted yield when an estimated value of 830 (the average of the selected state) is entered in to the calibration curve. This was verified for the range 710 to 1090 kWh for the estimated yield (Figure 5-24).

It is not possible to verify the probability distribution for specific yield directly against empirical data. However the behaviour of the node should reflect the behaviour of the calibration curve augmented by the yield uncertainty. Thus when hard evidence for the simulated yield node is selected, the

probability distribution of the specific yield node should deliver an expected value equal to the specific yield predicted by the calibration curve.

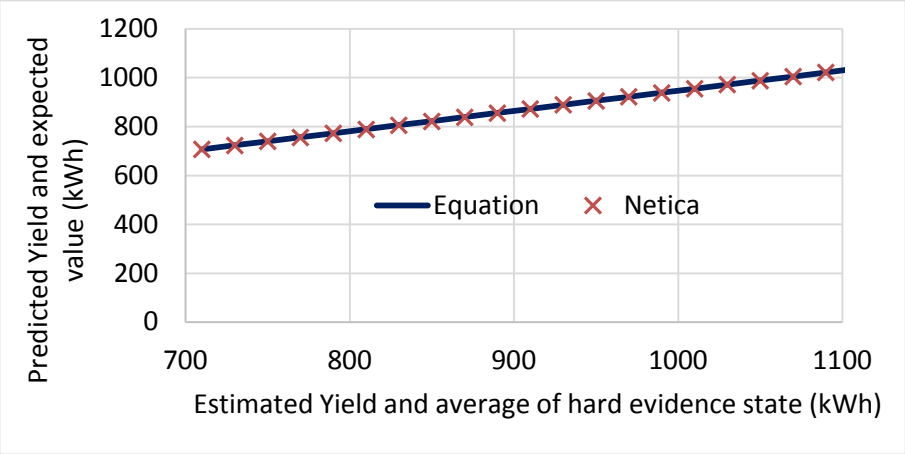


Figure 5-24 Verifying the specific state PD return the correct expected value when entering hard evidence for the simulate yield

5.5 Discussion and Conclusion

Theory and data have been combined to produce a BN model for predicting the yield of PV systems in the UK. These have been developed into a BN sub-model which models the specific yield of PV in 4 distinct LSOAs. PVGIS has been used as an estimation tool which has been calibrated using empirical data from the SMD. This corrects the inherent inaccuracy when using PVGIS to model system yields.

The uncertainty of this method was rigorously assessed and found to be homoscedastic and positively skewed. This has been has been quantified and endogenised in the BN model. A heuristic method has been developed to predict system yields by taking the roof area and empirical rating densities to calculate the distribution of system rating. This, when multiplied by the specific yield, furnished the model with a distribution of the system yield.

Because of the significant data analysis and processing techniques employed to furnish the model with quantitative data with which to derive NPTs, the behaviour of the model in Netica was

extensively verified to ensure it concurred with the selected calibration curved and degree of uncertainty.

Whilst the model has been developed as a component for integrating with the wider system model, it can also be used independently. Certain or uncertain evidence can be applied to the main interface nodes, orientation, pitch and region in order to probabilistically predict the specific yield for any PV system orientation in the 4 LSOAs. The main purpose of the model, however, is to integrate it with the wider model. The interface nodes will receive probabilistic evidence about orientation, pitch and region from the Building Stock node discussed in Chapter 7. The specific yield, as an output from this node will be passed as probabilistic evidence to calculate the total energy yield for PV systems.

This completes the main energy generation component of the integrated model; the next chapter looks at the energy demand component and explores the modelling of domestic energy consumption.

6 Building Energy Consumption

6.1 Introduction

In Chapter 5 the result of the development of a BN component to model the energy generation by a domestic PV system was presented. This chapter considers the theory, data and analysis to support the development of the Bayesian network component for domestic energy consumption. The aim of the chapter is to determine dwelling and occupants attributes which can serve as predictors of domestic gas and electricity consumption, and show how this can be used with available data sources to construct a 'building energy consumption' (BEC) component of the integrated model.

Using the same format as the previous chapter, supporting evidence from the literature is reviewed to elicit the domain ontology to help derive a structure of the BN component (section 6.2). In section 6.3, the datasets which have been sourced for this component are critically reviewed and analysed to support the model structure. Together the review and the available data determine the design of the BEM BN component (section 6.4). This section presents the DAG structure, the data required to populate the NPTs, and the resultant Netica BN model. Finally section 6.5 presents a discussion and conclusion.

6.2 The Domain Ontology

The conceptual map exposited in Chapter 4 proposed a set of direct dependency relationships for the renewable technology installation site/stakeholder subsystem. For the domestic adopter vector this subsystem can be understood as a domestic property and its occupants (Figure 6-1). This heuristic model needs to be converted to a DAG using the using the procedure outlined in Chapter 4.

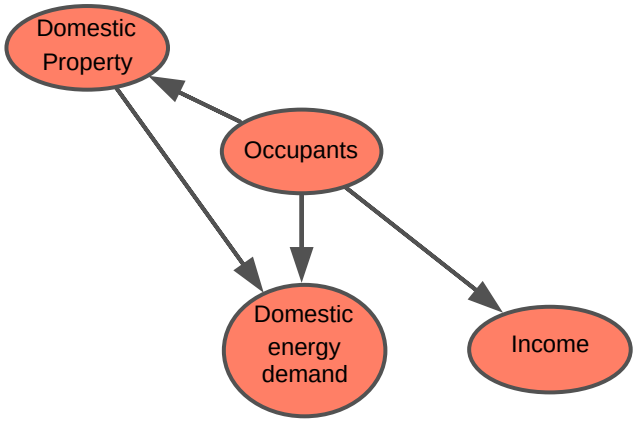


Figure 6-1 Sociotechnical system as the location of the solar PV generation system

Building physics methods and statistical models (see Chapter 2) are reviewed below, with the specific purpose of elucidating the salient parameters, dependencies and uncertainties.

6.2.1 Building Energy Model Parameters

Building physics models require a large number of data points to represent the building attributes with which to model the energy balance at the single building level. Three domains of energy use are commonly considered:

1. Space heating and cooling.
2. The heating of domestic hot water.
3. Appliances and lighting.

Knowledge of the performance of heating and cooling technologies, the loads and efficiencies of appliances, and the modelling of heat losses and gains through the building envelope due to conduction, radiation and convection, and air mass exchange through infiltration and ventilation, are all required to determine the energy balance (Kavgic, opt. cit.).

The UK's industry standard building model for estimating domestic energy consumption is the Building Research Establishment Domestic Energy Model (BREDEM). First developed in the early

1980s, (Anderson 1985; Shorrock and Anderson, 1985) the model has been updated to accommodate new construction methods, renewable technologies and energy efficiency measures (Anderson and Chapman, 2010). BREDEM is the underlying methodology in the Standard Assessment Procedure (SAP) (BRE, 2014), which is the UK Government's preferred method for the energy performance assessment of domestic buildings. The purpose of SAP is "to provide accurate and reliable assessments of dwelling energy performances that are needed to underpin energy and environmental policy initiatives". SAP is the UK's method of choice for compliance with the EU's directive on building energy performance for the production of energy performance certificates (EPC) (EP, 2010).

In order to estimate the energy demand of a domestic dwelling numerous physical data points are required. The BREDEM-8 model requires approximately 80 data points per dwelling (Anderson et al., 2002). These are, for example, the areas of dwelling elements (walls, floor, roof, windows and doors), their thermal conductivity (U-Values), ventilation and infiltration rates, the efficiency of and type of fuel used by space and water heating technology, internal and external temperatures, and the potential for solar gains. A number of factors to estimate occupant dependent energy use are included, for example, domestic hot water consumption and heating patterns may be specified (BRE, 2014).

A SAP assessment assumes a two zone heating regime of 21°C in the main living space and 18°C in the remainder of the occupied rooms. The sum of the fabric heat losses through building elements is calculated on a monthly basis taking into account typical outdoor temperatures. Incidental internal heat gains through metabolic rates, lighting and appliances are taken into consideration. Whilst

criticised as an inferior steady state model (Hens 2007)[17], BREDEM and SAP have been widely accepted following initial tests which showed the models worked well when compared to more detailed simulation models and empirical consumption data (Dickson et al., 1996, Shorrock et al, 1994).

This validation of BREDEM and relative computational simplicity led to its wide adoption in national bottom-up stock models developed to model the UK housing stock. These deterministic models attempt to achieve the requisite variety by due consideration of a large number of representative housing archetypes which is a promising method of modelling the residential sector (Famuyibo et al, 2012). Its general applicability, however, has been brought into question. In particular empirical data suggest that heating demand in real dwellings is lower and is more variable than the simplified two-zone heating regime in SAP (Huebner et al., 2013; Kelly et al, 2012). The modelling of renewables using SAP did not agree with more detailed simulations, particularly when non-standard solar PV modules were used (Murphy et al., 2011), and heating demand was underestimated for a low energy houses when compared to the Passive House Planning Package (Reason and Clarke, 2008).

Table 6-1 Data requirements for steady state energy estimation model

Main data requirements	**Applicable parameters**
Building Element Areas	Floors, walls, roofs, windows and doors
Materials and construction	U-Values, cavity walls, double glazing
Exposed Elements	External walls, ground floors, roofs
Ventilation and Infiltration	Open chimneys, vents, drafts
Weather	Temperature, wind speed and irradiance
Hot Water Demand	Fuel type and efficiency
Space Heating	Indoor temperature and heating pattern
Lighting	Incandescent and low energy
Insulation measures	Loft insulation, cavity walls
Renewable technologies	Solar hot water, Solar PV, dimensions
Occupancy profile	Number of active residents

[17] Steady state here assumes a constant heat transfer through the building envelop with constant maintained internal temperature. More advanced dynamic models do not assume a steady state and use differential heat transfer equations.

Nevertheless, BREDEM and SAP offer a valuable broad-brush ontology for predicting domestic building energy demand. This is summarised in Table 6-1.

6.2.2 Reduced Datasets

The large number of data points presents a practical problem with the BREDEM/SAP family of engineering models for estimating the energy demand of an existing building. For a new build typically most of the structural parameters will be known. For existing properties, however, without extensive, and potentially intrusive, site surveys, only a limited number of data points can be obtained. Therefore, reduced datasets are used where variables such as building age and built form serve as a proxy for a number data points. SAP, for example, has a co-defined reduced dataset SAP method (rdSAP), which estimates construction materials, their U-values, floor heights, external wall dimensions etc. from categorical building age data (Table 6-2) by accounting for building practices and building regulations[18] in operation at the time of construction (BRE, 2014, Appendix S). The rational is that UK dwellings are built to a minimum energy efficiency standard enforced at the time (Rylatt et al., 2003). Similarly, built form is used to predict the number of exposed external walls for heat loss calculations.

[18] The first building regulations are recorded as far back as 1189 but formal national regulations came into force in 1965 in England and Wales, a year later than Scotland (Manco, 2014).

Table 6-2 Categorical dwelling built form and age bands for the purposes of assigning U-values and other data used in rdSAP

SAP Dwellings Type 1
House
Bungalow
Flat
Maisonette
Park home

SAP Dwellings Type 2
Detached
Semi-detached
Mid-terrace
End-terrace
Enclosed mid-terrace
Enclosed end-terrace

Age Band
before 1900
1900-1929
1930-1949
1950-1966
1967-1975
1976-1982
1983-1990
1991-1995
1996-2002
2003-2006
2007-2011
2012 onwards

Whilst reduced, there are still over 100 parameters in the data to be collected (ibid, table S19) for rdSAP, more if multiple building components are present. Thus data collection still requires site surveys (Rylett et al., 2003), or a good knowledge of the building stock such as might be available to a social housing provider (Mhalas, 2013).

6.2.3 Parameters in UK Bottom-up Building Physics Models

The still significant requirements of a reduced dataset model like rdSAP is problematic for estimating the baseline energy demand of a large quantity of domestic building stock, such as for a local authority, region or even nation. The family of bottom-up building physics models address this problem by using a more limited and therefore predictable set of parameters. The purpose of this section is to explore these limited datasets as an expert solicitation of the parameters required for a BN submodel.

Table 6-3 Bottom-up Building Physics Models, Key Parameters and Dwelling Type

Model	Year	Variety of Dwelling Types	Key Attributes	Reference
BREHOMES	1997	Over 1000	Age group, building form, tenure and central heating ownership.	Shorrock and Dunster, 1997
NHER			Age and built form	NHER, 2004
EMERALD	2003		Age, built form and Building Footprint.	Rylett et al., 2003
UKDCM	2005	12,000 (20,000 in 2050)	Geographical Area, Age, Build form, Tenure, Number of Floors, Construction.	Boardman et al. 2005
DECarb	2007	8064 43384	Age, built form, insulation characteristics (7 variables in total)	Natarajan and Levermore 2007
DECORUM	2009	N/A	50 general parameters, 18 derived from Age, 5 derived from built form, 22 "from walk by survey"	Gupta, 2005
CDEM	2010	47	Build form and age. EHS used for parameter prediction.	Firth et al., 2010
DECM	2011	4 16000+	Scaling up of EHS dataset	Cheng and Steemers, 2011
CHM	2012	16000+	Scaling up of EHS dataset	Hughes et al., 2013
Mahlas et al	2013	16000+	EHS dataset	Mhalas et al., 2013

Table 6-3 lists some of the energy stock models develop over the last 18 years in the UK, all of them based on the BREDEM family of reduced dataset energy models. As discussed above, building age and built form are invoked in reduced dataset models to infer a large number of parameters. Each model uses these and a number of additional parameters, each of which may be used to predict other values for the 80 to over 100 parameters in the reduced dataset BREDEM model.

As an example, the Community Domestic Energy Model (CDEM) uses 6 build form classifications and 7 building age bands to deliver 54 combinations. In practice some combinations are omitted because

their occurrence in the building stock is rare[19] resulting in 47 "building archetypes" (Firth et al., 2010). With the inclusion of further parameters, the number of potential archetypes rapidly increases. Thus the UK Domestic Carbon Model (UKDCM) created approximately 12,000 archetypes resulting from the variability derived from combinations of the geographical area, age category, built form, tenure, number of floors, and construction method[20] (Boardman et al. 2005). This rose to 20,000 as further age categories were created for each decade leading up to 2050 in order to estimate the energy demand for defined future energy scenarios.

The energy demand of each archetype is calculated using the building physics model of choice; unknown parameters are inferred by reference to empirical building survey data such as the English Housing Survey (EHS) or its predecessor, the English Housing Condition Survey (EHCS). In order to calculate an aggregated baseline domestic energy demand for all the dwellings in spatial area of interest the number of dwellings of each archetype must be determined. This can be estimated for national and regional areas. For each archetype the number of dwellings is multiplied by its calculated energy demand and the sum over all archetypes yields the desired aggregated demand (Shorrock and Dunster, 1997).

A more direct method than using archetypes is to calculate the energy demand for a representative sample of survey data. Thus the Cambridge Housing Model (CHM) uses EHS data and rdSAP to calculate the energy demand for each member of a sample of 15000 dwellings, each with a known weight in the UK housing stock (Hughes et al., 2013). The model is used to estimate the national domestic energy demand and a range of statistics on energy demand segmented by many variables (Palmer and Cooper, 2012).

---

[19] For example converted flats in post 1945 dwellings and purpose-built flats pre-1900.

[20] Calculated as the product of the cardinality (number of elements) of each discrete variable. For example:

$$\text{Built Form } (10) \text{ x Age } (7) \text{ x Region } (9) \text{ x Tenure } (2) \text{ x Floors } (3) \text{ x Construction } (3) = 11340$$

The 115 parameters in rdSAP give a potential variety of 4.6×10^{42} unique combinations or states[21]. A large number of these states will be of very low probability (e.g. district heating in terraced houses), or zero (e.g. high-rise flats with PV systems) so will not be prevalent in the building stock population of interest. Furthermore, sensitivity analysis of building physics models shows that aggregated building energy consumption estimates are insensitive to many of the variables (Hughes et al., 2013). It is pertinent, however, to enquire whether using all 15000 members of the EHS sample, or a lower number of archetypes delivers a model with the requisite variety to represent the uncertainty in empirical energy demand.

The parameters included in the restricted datasets used by the bottom-up building physics models in Table 6-3, are summarised in Table 6-4. One of the models (DECORUM) uses 20 additional data points collected by inference on "walk by surveys". This may include reference to GIS and other sources. EMERALD also makes use of GIS mapping tools using a combination of automatic and user controlled software routines to determine building dimensions from built form and age category using a geometric model (Chapman, 1994).

[21] Estimated by taking the product of the cardinality (number of elements) of each discrete variable in the rdSAP model. If all the continuous variables were included, this figure would be infinitely higher.

Table 6-4 Parameters used in bottom-up building physics models

Parameter	Model in which Parameter is used
Geographical Area	UKDCM
Building Age	BREHOMES, EMERALD, NHER (Level 0); UKDCM;DECORUM;CDEM
Built Form	BREHOMES; EMERALD;NHER (Level 0); UKDCM;DECORUM;CDEM
Area	EMERALD
Perimeter	EMERALD
Central Heating Ownership	BREHOMES
Footprint	EMERALD
Orientation	EMERALD
Storeys	EMERALD; UKDCM
Tenure	BREHOMES;UKDCM
Number of Floors	UKDCM
Construction Method	UKDCM
Number of Rooms	NHER Level 0
20 Parameters + Walk by Survey	DECORUM
80 Parameters EHS Survey	CHM;DECM
Occupancy Pattern	DECM

Whilst Building Age and Built Form are present in each model, developers have each introduced additional parameters, apart from Firth et al. (2010) who have used average values for constants such as the floor area for each archetype. The data used in such studies is often an outcome of the data that was available to the researchers at the time (Swan and Ugursal, 2009).

The inclusion of occupancy as an integral parameter in a building stock dataset is difficult to attain. Thus most models estimate the energy demand based on a standard or average occupancy pattern. Models which do consider occupancy are the DECM model, NHER level 2 and 3, and modern variants of SAP which have been developed to consider occupancy factors for Green Deal assessments. These latter models are aimed at single dwelling assessments rather than building stock estimates. Models which use the weightings in the EHS to scale up to national or regional building stock do account for occupancy since the parameter is included in the survey. However, these models cannot account for the variability of occupant influence on the total domestic energy demand due to different energy behaviours and practices.

To summarise, a number of parameters with a predictive power for energy demand has been solicited from the literature. To take this beyond a pure taxonomic description of the problem domain it is necessary to consider the dependency graph which expresses the conditional relationships between the parameters. Thus the literature which provides an insight into appropriate PGM structures is considered in the next section.

6.2.4 Towards a Dependency Graph for Building Parameters

Many studies provide insight into the direct dependencies between parameters and/or the strengths of statistical relationships which predict building energy demand. Three types are presented here: sensitivity analysis, statistical techniques and hierarchical statistical techniques.

Sensitivity Analysis

It is instructive, a priori, to consider the sensitivity of building energy demand to candidate parameters. Sensitivity analysis is considered by modellers from a variety of disciplines as an essential prerequisite to the building of robust models (Firth et al. 2010). Whilst critiqued as "perfunctory" by Saltelli and Annoni (2010), a number of building energy modellers use a one-at-a-time local sensitivity analysis (OATSA). This varies one variable locally around its mean value to observe the impact on a target variable whilst holding all other variables at their mean values.

Table 6-5 presents the normalised sensitivity coefficients[22] for the impact of the most influential parameters on carbon emissions from a number of recent studies using OATSA. These parameters

[22] A normalised sensitivity coefficient represent the percentage change in target parameter given a 1% change in the input parameters (Firth et al. 2010). Generally the change is assumed to be a linear

can be categorised as environmental, building, building services, and occupant attributes. The most influential environmental parameter is the external air temperature; in building models this is often represented by the geographic region. Floor area is the most influential building attribute. Boiler efficiency is very influential, and of the two behavioural characteristics, the 'heating demand temperature' is the most influential parameter of all, at an average 1.55.

Table 6-5 Normalised sensitivity coefficients reported for three BREDEM based models

Parameter	**Parameter Type**	**Normalised Sensitivity coefficient**		
		CDEM	**DECM**	**CHM**
External Air Temp (°C)	Environment	-0.58	-0.61	-0.59
Storey Height (m)	Building Attribute	0.48		0.46
Floor Area (m^2)	Building Attribute	0.34	0.77	0.53
Wall U Value (W/m2K)	Building Attribute	0.27	0.21	0.18
Window U-value (W/m2K)	Building Attribute	0.19	0.12	0.11
Boiler Efficiency	Building Service	-0.45	-0.48	-0.66
Heating demand temperature (°C)	Occupant behaviour	1.55	1.55	1.54
Length of daily heating period (hrs)	Occupant behaviour	0.62		
CDEM (Firth et al., 2010); DECM (Cheng and Steemers, 2011); CHM (Hughes et al 2012)				

Statistical Methods

In contrast to the use of a modelled energy demand in the above BREDEM models, a number of statistical approaches have used empirical datasets. Through the use of concurrent building survey, occupant interview and measured energy consumption data, the influence of building attributes in conjunction with occupant characteristics on energy consumption can be investigated. As alluded to in Chapter 4, the two behaviours, that of the dwelling (driven by its physical properties) and that of the occupants (driven by their "demographic, biophysical and psychological attributes"), can be

---

function of the input parameter change with a limit of, typically ± 1%. This limitation means only a small area of the parameter space is covered (Tian, 2013).

described as a complex interaction of two components of a sociotechnical system (Hitchcock, 1993). An understanding of this aspect is essential if the objective of exploring socio-economic parameters using a Bayesian Network is to be fulfilled.

Regression techniques can be used to model the impact of a dependent variable given an array of predictor variables. This technique has been applied to estimate metered energy demand given a selection of variables such as those introduced above (Nielson, 1993; McLoughlin et al. 2012) and to predict heating demand (Catalina et al., 2008). Kelly (2011) used multiple linear regression (MLR) on a dataset of the English Housing and Condition Survey from 1996 which incorporated empirical metered energy data from a follow-up energy study on over 2531 homes. The results are shown in Table 6-6, Column 1. Floor area was the second most influential variable, followed by the number of occupants.

Steemers and Yun (2009) used a Generalised Linear Model[23] on a dataset of over 4000 US dwellings and occupants to show the measure of variability in space heating consumption explained by the variability of a range of predictor variables (essentially a correlation study). The number of 'heating degree days' was the most significant influence. However, this did not sufficiently explain all the energy demand. Of the building attributes, 'floor area' was the second most correlating variable with 'building age' and 'built form' also partially explaining the energy demand. The third most influential variable was the 'type of heating', and of the behavioural parameters it was number of 'heated rooms'. Of the demographic parameters, 'income' and 'number of occupants' were influential but not significantly so. The relative strength of these influences is shown in Table 6-6, Column 2.

A number of generalised models have been created which show the predictive influence of variables using segmented averages. The Local Area Resource Analysis (LARA) model has shown that energy

[23] GLM is related to MLR but has a vector of dependent variables.

and carbon emissions are strongly influenced by household income, but other factors such as built form, tenure, household composition, rurality, and socio-economic characteristics are also very important (Druckman and Jackson, 2008). The Distributional Impacts Model for Policy Scenario Analysis (DIMPSA) using aggregated data shows that domestic emissions are strongly correlated with income with the richest income decile emitting three times as much as the poorest decile (Preston et al., 2013).

Table 6-6 Studies using statistical models and parameters which influence energy consumption

Parameter	**Parameter Type**	**Study**		
		[1]	**[2]**	**[3]**
Degree Days	Environment		0.306	
Storey Height	Building Attribute			
Floor Area	Building Attribute	0.262	0.216	
Building Age	Building Attribute		0.125	
Built Form	Building Attribute		0.129	✓
SAP Rating	Building Attribute	(0.053)		
Type of Heating	Building Service		0.202	
Heating demand temperature	Occupant behaviour	0.042	0.099	
Number of Rooms Heated	Occupant behaviour		0.181	
Heating Pattern	Occupant behaviour	0.112		
Number of Occupants	Occupant demographics	0.297	0.076	✓
Household Income	Occupant demographics	0.140	0.083	✓
Tenure	Occupant demographics			✓

Study:
[1] Measure standardised MLR coefficients predicting SAP rating (Kelly, 2011)
[2] Measure of R^2 for variables predicting space heating energy consumption (Steemers and Yun, 2009)
[3] Influence on energy expenditure in LARA model (Druckman and Jackson, 2008).

Hierarchical Methods

Evidence in the literature suggests that both direct and indirect influences on energy demand should be considered. Steemers and Yun (opt cit.) observed that household income has a weak direct influence on energy consumption, but, they report, it has a strong influence on floor area and the

number of heated rooms. These in turn both strongly influence energy consumption. They investigated the same dataset using Path Analysis[24] which reveals the strength of direct and indirect influences. A summary of their result is shown in Figure 6-2 which shows that occupant parameters (income, size of household and the age of household reference person) have strong direct influences on the building parameters (built form, building age, floor area), equipment and behavioural parameters (number of heated rooms), and these in turn have a direct influence on heating energy consumption. There is also a weak direct influence (the dashed line in Figure 6-2) of occupant parameters on energy use.

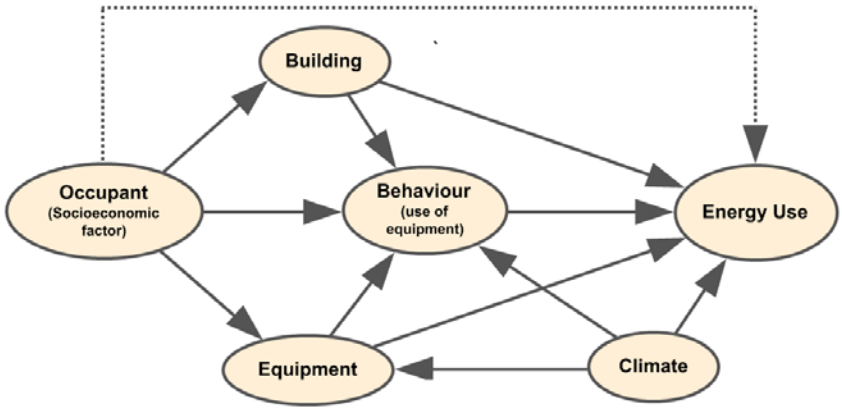


Figure 6-2 Path diagram using path analysis after Steemers and Yun (2009)

Kelly (opt. cit.) applied structural equation modelling to the same data as analysed using MLR. This technique relies on substantive prior research to represent the expected causal relationships between several manifest variables which are likely to explain domestic energy consumption. The results are shown in Figure 6-3. The numbers on the connecting arrows indicate the strength of influence. This hierarchical model again positions household income as an important indirect influence on energy expenditure via the strong direct influence of floor area. In contrast to Steemers

[24] Path analysis uses hierarchical causal maps and are a precursor technique to Bayesian networks (Geiger and Pearl, 1988)

and Yun (opt. cit.), Kelly also shows that the number of occupants influences household income strongly, but also influences the energy expenditure directly.

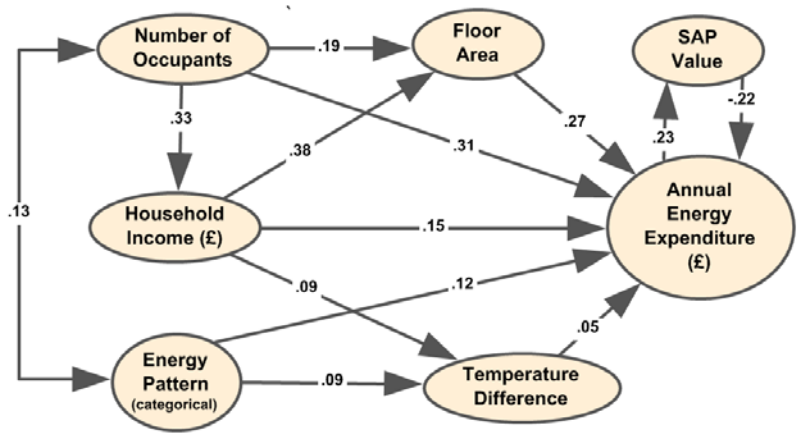


Figure 6-3 Path diagram of structural equation model showing influences on energy expenditure after Kelly (2011)

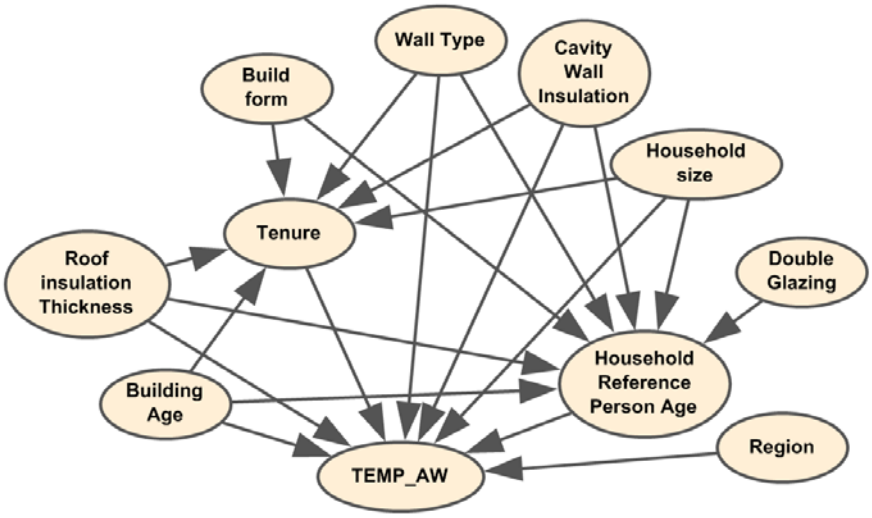


Figure 6-4 Bayesian Network model for predicting internal temperature after Olivier (2008)

A Bayesian Network modelling approach was adopted to specifically explore occupant parameters and their influence on internal temperature to which energy demand is sensitive (Shipworth 2013). A model discovery algorithm was used to create a BN which is theory agnostic, but delivers the model

which is a 'best fit' for the data. The results of this work are unpublished but have been cited by Telenko and Seepersad (2014) and the DAG for the model has been produced by Olivier (2008). This shows (Figure 6-4) that household size, tenure and the age of the household reference person all directly influence the temperature at which the dwelling is kept. However, a large number of other influences are observed which are difficult to theorise but which can facilitate inter-disciplinary, deliberative discussion (Shipworth, opt cit.).

6.2.5 Summary

A strong literature base in the domestic building energy domain has been purposefully reviewed with the intention of discerning important parameters and their dependencies (i.e. an ontology) to support the design of a Bayesian Network. Limited statistical datasets are available to populate such a model. Some parameters may not be important and should be omitted in pursuance of model parsimony. Those to which the energy demand is insensitive can be assumed to be independent of all other parameters in the Bayesian Network. In terms of a JPD they will not contribute significantly to the variability and can be safely marginalised.

Statistical methods and some hierarchical statistical techniques have been explored. The former, such as MLR or GLM models, can fail to take into account the major causal influences, simplifying the model to one where the dependent variable is linearly dependent on a mutually independent collection of predictor variables (Fenton et al., 2002). As such, they are equivalent to the modelling of a naïve Bayesian classifier. The latter (hierarchical techniques) include, and are more formally related to, Bayesian Networks, since they discern a nodal structure expositing both direct and indirect dependency relationships. For example occupant parameters, it is observed, such as income, have a weak direct influence on energy consumption, but also a strong indirect influence via building attributes such as floor area.

There is enough extant knowledge to construct a BN for this knowledge domain. However, any DAG constructed must have the necessary data with which to populate the NPTs, and so the resultant DAG is a synthesis of the optimal model structure and the data available in practice. The next section considers the available datasets for the construction of the building energy model.

6.3 Data Sources

In this section a number of publicly available empirical datasets are discussed which enable the partial quantification of the statistical relationships between the key parameters identified in the previous section. The analysis of these datasets is performed to a) support independence relationships represented by the DAG, and b) to quantify the NPTs. The principle datasets discussed are the national energy efficiency data (NEED) framework, the English Housing Survey, the Cambridge Housing Model and the Living Costs and Food Survey.

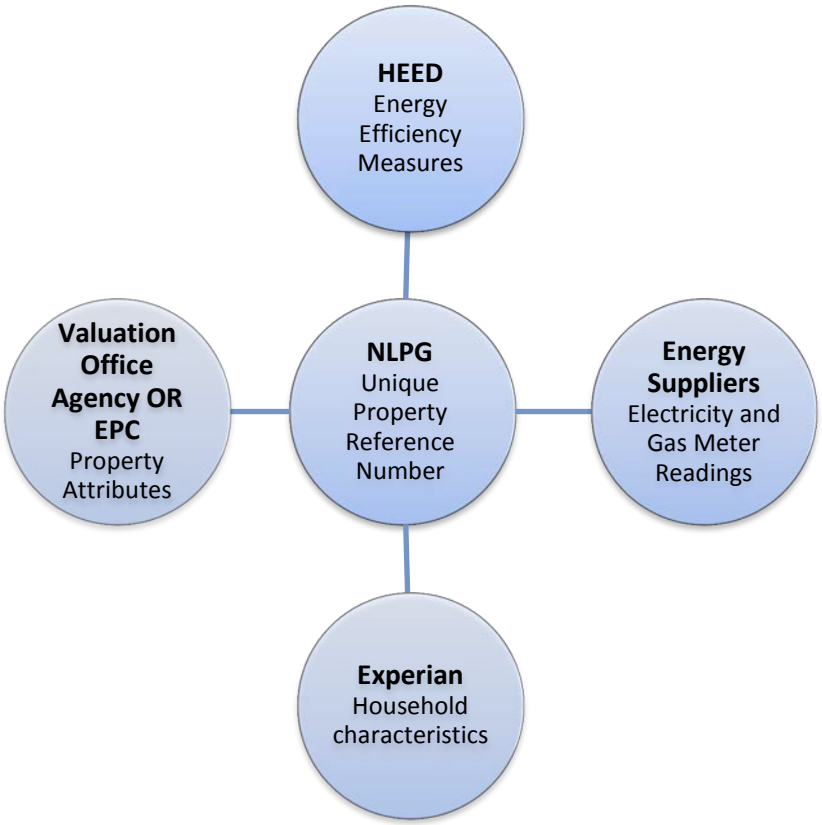
6.3.1 National Energy Efficiency Data (NEED) Framework

The NEED framework is a multi-agency data sharing framework established by DECC in partnership with the energy suppliers, the Energy Saving Trust and the Valuation Office Agency. Its purpose is to provide Government and delivery partners with empirical data with which to inform evidence based energy efficiency policy and programmes (Foulds and Powell, 2014).

Provenance of the data

At the core of the framework is a unique property reference number (UPRN) assigned to each address by the National Land and Property Gazetteer (NLPG). This UPRN is assigned by matching

postal addresses to four other datasets thereby allowing records in one dataset to be linked to the corresponding records in another, as shown in Figure 6-5 (DECC, 2012).



After DECC, 2012

Figure 6-5 provenance of the data integrated in the NEED Framework

Energy suppliers have provided the NEED framework with metered gas and electricity data obtained from billing data since 2004. The data are annualised and weather corrected to convert the actual demand for a non-annual period to an estimate of seasonal normal demand for a 365-day year. This has the effect that where the weather has been colder (or warmer) than the seasonal normal, the consumption is adjusted downwards (or upwards). The magnitude of these adjustments are calibrated using daily consumption data from 12,000 meter points (DECC, 2014).

Properties with an annual consumption of less than 100kWh for gas and electricity, or over 50,000kWh per annum for gas and 25000kWh for electricity, were rejected. This ensured that empty

properties and properties where the meters are misclassified as domestic but are in fact in commercial properties, do not distort the dataset.

The Home Energy Efficiency Database (HEED) was established to record data on energy efficiency measures carried out on the UK housing stock by national government programmes since 1995. This includes the Energy Efficiency Commitment, Carbon Emissions Reduction Target, and the Community Energy Savings Programme, as well as specific activity reported by trade associations (Hamilton et al., 2013). Fifty percent of UK homes have one or more records[25] on the HEED database. The database is managed by the Energy Saving Trust for the NEED Framework. In addition to energy efficiency measures HEED also contains records of properties which have been retrofitted with renewable energy technologies, particular solar thermal and solar PV systems. Other than this the HEED database has very little relevance to this study.

The Valuation Office Agency (VOA) maintains a national register of all properties in the UK for the purposes of setting appropriate local government taxation based on property values. The register holds property age, dwelling type, number of bedrooms and floor area for each property address.

Experian Plc. is a commercial data company specialising in modelling socio-economic data at the address level. Statistical algorithms which use publically available demographic data and proprietary data such as credit references, are used to model the income group of household occupants and derive a modelled household income for a target property. Data were purchased by DECC for 3.5 Million dwellings.

An analytical dataset was created linking the above four sources of data, though not all properties could be matched across all four databases. Using the VOA property attributes as a baseline the percentage of records matched to the other datasets is presented in Table 6-7 (DECC, 2012).

[25] If more than one programme measure has been registered for a property then a property will have more than one record on HEED.

Table 6-7 Datasets used in the NEED framework and success of address matching

Dataset	Percentage Match
VOA	100
HEED	99
Gas	97
Electricity	94
Experian	82
After DECC, 2012	

The reasons for the less than perfect matching are not given by DECC. However it is known that the electricity and gas meters are often attributed to the wrong address since many dwellings of multiple occupation and 'flats over shops' do not have on premise meters. The 82% match for Experian data reflects the incomplete coverage of UK households for credit data thus it is not possible to derive modelled income for all properties.

Analysis of the NEED data

The resultant analytical dataset has been analysed by DECC (2012). This showed that only a third of the variation in electricity and gas consumption could be explained by property attributes and household characteristics in the NEED dataset. Floor area had the largest influence with both gas and electricity consumption increasing with floor area. The effect of building physics is discernible. Thus the average energy demand decreases in the order detached, semi-detached, end terraced, terraced and flat, reflecting the number of external walls for each built form. Similarly the improvement in energy efficiency standards is observed with a decrease in gas consumption, largely used for space heating, with newer buildings. In contrast electricity consumption was little influenced by building age.

Considering occupant characteristics the average consumption for dwellings of different tenure showed social housing occupants at lower energy consumption than owner occupied properties,

with privately rented dwellings in between. Gas and electricity consumption increased with household income, an effect which was more marked at high incomes above £40,000 per year.

DECC also release statistics for the aggregated data on which the above analyses were based, including mean, median and quartile data. These enable an appreciation the wide variability of domestic energy consumption, shown in Table 6-8 for the whole dataset.

Table 6-8 Summary of annual consumption (kWh) statistics for 2010

	Mean	Standard Deviation	Lower Quartile	Median	Upper quartile
Gas	15,100	8,000	9,700	14,000	19,200
Electricity	4,200	3,100	2,200	3,500	5,300

However the primary tabular data were not released thus rendering impossible a more detailed multi-parametric analysis. Using an ad-hoc data release of a broad range of percentiles points for each permutation of region, built form, floor area and building age, parametric probability distributions were fitted to the cumulative distributions for each row. This enabled the synthesis of tabular data over a broad range of gas consumption values. Examples of such plots, created using the Weibull distribution function, are shown in Figure 6-6. These parametric fits suffered from systematic errors at low and high gas consumptions. Furthermore, income was not included in this data and no analogous data for electricity consumption were released by DECC[26]. Nevertheless this synthesised tabular data does make a strong case for analyses which endogenise the variability of gas and electricity consumption given building attributes. When used to furnish a naïve BN similar findings to the analysis by DECC (2012) were reproduced. But these shortcomings thwarted the

[26] This was finally released in March 2015.

development of an integrated model to include both gas and electricity. This is addressed in the next section.

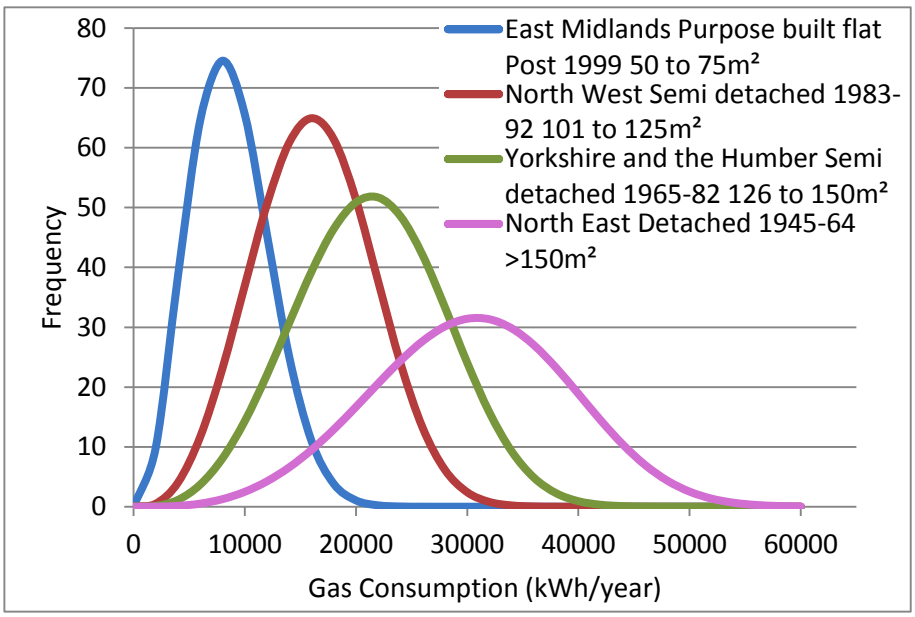


Figure 6-6 Gas consumption distributions for various parameter combinations generated using a Weibull probability distribution fitting to percentile data points.

The Anonymised NEED Dataset

In 2014 DECC published a repeated cross sectional study of just over 4 million records representing households with annualised and weather adjusted electricity and gas consumption data from 2005 to 2012 (DECC, 2014B). The data were released with building attributes along with energy efficiency measures undertaken; household income data were not included. Thus this tabular data partially addresses the shortcomings of the synthesised data set discussed previously.

The data were anonymised by constraining the geographic precision to the region[27] for each record and banding the main columns such as the property age, floor area and metered fuel consumption.

[27] Based on the now defunct government office regions (GOR)

Experian plc did not permit the release of estimated household income data, and legal requirements prevented the inclusion of VOA property attributes. The latter were replaced by property attributes from the energy performance certificate (EPC) database. This precludes properties without an EPC, making the sample less representative[28], but does allow the EPC rating to be included in the dataset.

The data dictionary for the dataset is shown in Table 6-11 which includes the bandings. The discretisation of the gas and electricity consumption is non-uniform and is rounded to a nearest ceiling value based on its absolute value. These bands are shown in Table 6-9 for gas and 6-10 for electricity.

Table 6-9 Banding ranges for annual gas consumption

Range (kWh)	Rounding
100 – 7,999	500 kWh
8,000- 15,999	100 kWh
16,000 – 24,999	500 kWh
25,000 – 34,999	1,000 kWh
35,000 – 50,000	5,000 kWh

Table 6-10 Banding ranges for annual electricity consumption

Range (kWh)	Rounding
100 - 9,999	50 kWh
10,000 - 11,999	100 kWh
12,000 - 14,999	500 kWh
15,000 - 19,999	1,000 kWh
20,000 - 25,000	5,000 kWh

DECC have shown that the EUL dataset is representative of the entire NEED dataset. Weightings are also provided to allow the dataset to be scaled up to national and regional frequencies. Therefore the dataset represents the most up to date empirical tabular data on energy consumption with corresponding property attributes. Despite the exclusion of income data an opportunity to develop a BN model was presented.

[28] Energy Performance Certificates (EPCs) were first introduced in England and Wales in 2007 and only new builds and properties which have changed ownership since then will have one.

Table 6-11 Data dictionary for the NEED framework 'EUL' anonymised dataset.

Column	Description
HH_ID	Household identifier. Created specifically for these datasets.
REGION	Former Government Office Region: North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales
IMD_ENG	Index of multiple deprivation (IMD) 2010 for England, quintiles. Households are allocated based on the IMD rank of the Lower Layer Super Output Area (LSOA) they are located in.
IMD_WALES	Welsh Index of multiple deprivation 2011. Households are allocated to one of five bands based on the deprivation rank of the LSOA they are located in.
FP_ENG	EUL only. Fuel Poverty Indicator. Households are allocated to five bands based on the estimate of the proportion of household in fuel poverty in the LSOA they are located in; low income high cost definition.
EPC_INS_DATE	EUL only. Provides information on the date of the EPC inspection (grouped by pre-2010 and 2010 or later).
GconsYEAR*	Annual gas consumption in kWh.
GconsYEARValid*	Flag indicating records with valid gas consumption and off gas households.
EconsYEAR*	Annual electricity consumption in kWh.
EconsYEARValid*	Flag indicating record with valid electricity consumption.
E7Flag2012	Flag showing households with Economy 7 electricity meters.
MAIN_HEAT_FUEL	Description of main heating fuel (gas or other).
PROP_AGE	Age of construction of the property (six bands): before 1930, 1930-1949, 1950-1966, 1967-1982, 1983-1995,1996 onwards
PROP_TYPE	Type of property: Detached house, Semi-detached house, End terrace house, Mid terrace house, Bungalow, Flat (including maisonette)
FLOOR_AREA_BAND	Floor area band: 1 to 50, 51-100, 101-150, Over 151
EE_BAND	Energy Efficiency Band: Band A or B, Band C, Band D, Band E, Band F, Band G
LOFT_DEPTH	Depth of loft insulation (150mm or more, or less than 150 mm).
WALL_CONS	Wall construction (cavity wall or other).
CWI	Cavity wall insulation installed through a Government scheme.
CWI_YEAR	Year cavity wall insulation installed.
LI	Loft insulation installed through a Government scheme.
LI_YEAR	Year of loft insulation installed.
BOILER	Boiler installed in property.
BOILER_YEAR	Year of boiler installation.
WEIGHT	EUL only. Weighting based on Region, property age, property try and floor area band.
**The columns in which YEAR appears are repeated for each year, 2005 to 2012*	

TAN Analysis of the Anonymised NEED Dataset

The first step to evaluate the anonymised NEED dataset for BN modelling was to use the TAN learning algorithm (Chapter 3). Setting, in turn, the electricity and gas consumption as the classifier variable allowed the sensitivity to a finding at other nodes to be evaluated. Tree augmentation also suggests dependencies between other parameters. The resultant BNs are shown in Figure 6-8 and Figure 6-9 with electricity and gas as the classifier variable respectively. Remembering that the classifier node is automatically connected to all other nodes in the network – a naïve model – it is significant in that consistent dependencies are determined between region and build form, built form and floor area and built form and building age though the direction is not consistent. This pattern was repeated setting other parameters as the classifier variable. The undirected graph representing this finding is shown in Figure 6-7 and strongly suggests that arcs should be drawn between build form and both building age and floor area.

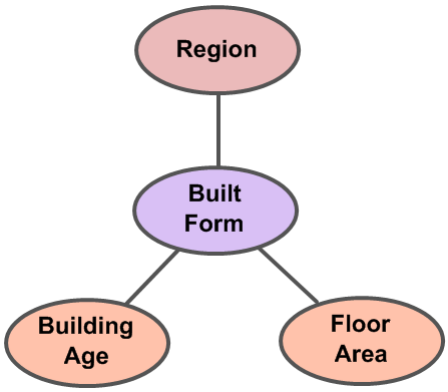


Figure 6-7 dependencies between building attributes and region discovered using TAN learning of anonymous NEED data

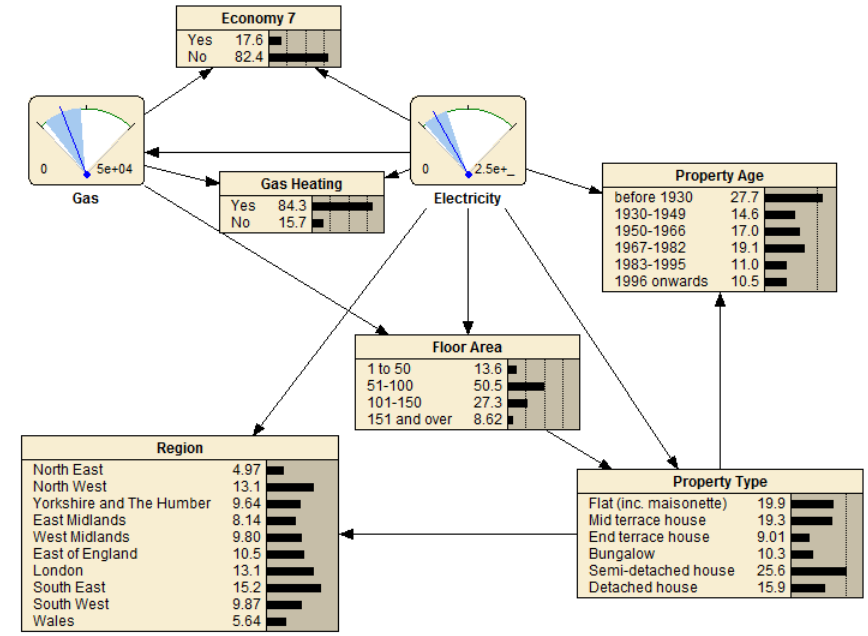


Figure 6-8 TAN BN Structure in Netica with electricity consumption as the classifier using the anonymised NEED dataset.

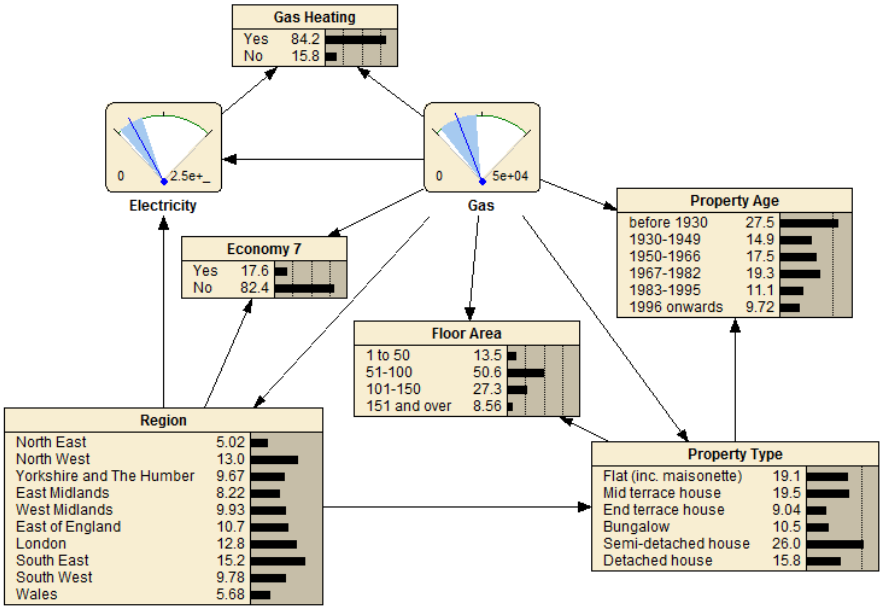


Figure 6-9 TAN BN Structure in Netica with gas consumption as the classifier using the anonymised NEED dataset.

The sensitivities to findings at other nodes, in the TAN BN models above, for gas and electricity are shown in Table 6-12 and 6-13 respectively. A noteworthy result is that the electricity consumption

has greatest sensitivity to gas consumption, at 18.4%. And of the three building attributes, floor area is more sensitive predictor of gas consumption at 20.1% compared to electricity at 7.5%. Built form is also a stronger predictor of gas consumption at 15.8%, compared to 3.5% for electricity. Generally electricity consumption is not as sensitive to building attribute parameters as is gas consumption. Indeed electricity consumption is most sensitive to variations in gas consumption.

Table 6-12 Sensitivity of electricity consumption to findings at other nodes

Variable	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
Electricity	1.03E+07	100	7.47061	100	0.9869058
Gas	1.90E+06	18.4	0.2372	3.18	0.0000176
Gas Heating	1.10E+06	10.7	0.07061	0.945	0.0000026
Floor Area	7.67E+05	7.46	0.09567	1.28	0.0000055
Economy 7	4.70E+05	4.57	0.03156	0.422	0.000001
Property Type	3.50E+05	3.4	0.06079	0.814	0.0000036
Region	6.53E+04	0.635	0.00726	0.0971	0.0000002
Property Age	4.46E+04	0.434	0.00666	0.0891	0.0000002

Table 6-13 Sensitivity of gas consumption to findings at other nodes

Variable	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
Gas	9.36E+07	100	6.37519	100	0.9432639
Gas Heating	2.67E+07	28.5	0.48523	7.61	0.080842
Floor Area	1.88E+07	20.1	0.27389	4.3	0.003818
Property Type	1.48E+07	15.8	0.21283	3.34	0.0034978
Electricity	5.88E+06	6.28	0.23378	3.67	0.0051153
Economy 7	4.82E+06	5.15	0.09907	1.55	0.008829
Property Age	3.16E+06	3.37	0.04547	0.713	0.0008371
Region	1.19E+06	1.27	0.02202	0.345	0.0005131

In summary the NEED dataset was found suitable for the construction BN models based on empirical data. Relationships between key building attributes can be specified supported by a sensitivity analysis and TAN learning algorithm. The strength of these dependencies are consistent with other studies, thus floor area is a strong predictor of domestic energy consumption.

A key weakness with respect to research objectives is the lack of income data in this dataset and the ability to integrate this in to the model. This was pursued in the next two sections with an analysis of two public datasets which combine income metrics with energy consumption.

6.3.2 Living Costs and Food Survey

The Living Costs and Food Survey (LCF) has been produced by the ONS on behalf of the Department for Environment, Food and Rural Affairs (DEFRA) since 2008, replacing predecessor surveys stretching back to 1957 (ONS, DEFRA, 2010). Used primarily for information on the retail prices index and trends in nutrition it provides multi-purpose data on all household consumer purchases including domestic energy, accompanied by meta-data on household characteristics. The underlying methodology is largely through computer assisted personal interviews (CAPI) augmented by diary keeping methods. Most variables need to be processed to yield estimates of weekly or annual expenditures (ONS, DEFRA, opt cit.). The 2010 survey has a sample size of 5,116 UK households.

The LCF is one of the few public sources of data which estimates domestic energy expenditure and corresponding data points for household metrics such as income, dwelling type (built form) and main fuel types. It has been used by DECC to augment domestic energy use statistics and by academic researchers (Druckman and Jackson, 2008). For this study the fields shown in table 6-14 were of interest for domestic energy demand studies.

Table 6-14 ECF Data used in this study

ECF variable code	Description
case	Case Number
weighta	Annual weight
A049	Household size
A116	Category of dwelling
A121	Tenure - type
A150	Central heating by electricity
A151	Central heating by gas
A152	Central heating by oil
A153	Central heating by solid fuel
A154	Central heating by solid fuel and oil
A155	Central heating by calor gas
A156	Other gas central heating
B170	Gas amount paid in last account
B175	Electricity - amount paid in last account
EqIncOp	Equivalised income (OECD Scale) - top-coded
a114p	Rooms in accommodation - anonymised

The annual weight field enables each row of survey data to be scaled to represent all similar UK households. Category of dwelling and tenure type are categorical variables which take one of several values as in Tables 6-15 and 6-16. The household income is equivalised using the OECD scale[29]. Several of the fields are anonymised to minimise the risk of personal data disclosure. The household size is top-coded[30] to six, and the weekly income to £1859/week.

[29] Equivalisation is an adjustment to actual income to account for, and make comparable, different household size and composition. The UK has traditionally used the McClements equivalence scale but since 2009 the 'OECD' method is used to enable international comparisons (Horsfield, 2011).

[30] Top-coding means putting an upper limit on the published data so as not to risk disclosure of less frequent values.

Table 6-15 Built form categories in the LCFS.

Code	Category
0	Not Recorded
1	Detached
2	Semi-detached
3	Terraced
4	Purpose-built flat
5	Converted flat
6	Other

Table 6-16 Tenure categories in the LCFS

Code	Category
0	Not Recorded
1	Social Rented
2	Private Rented
3	Owner Occupied

The data were processed to yield a dataset with categories substituted for coded values, and the central heating fields (A150-A156) converted into a single central heating type field.

Analysis of the LCF

The LCF, by using the annual weighting parameter can provide a JPD for a number of key parameters. A meta-analysis of the LCF serves to test its suitability for building a BN. Since this is the only dataset which delivers both household income and energy expenditure consumption this was the primary purpose of the investigation.

Figure 6-10 shows the equivalised household income for the LCFS sample (alongside that for the English Housing Survey discussed in the next section), and Figure 6-11 shows the frequency distribution of gas and electricity expenditure. To observe the dependency of the expenditure distributions on income the mean value for each was calculated for each income decile. This is shown in Figure 6-12 and shows consumption increase monotonically with income as previously reported for an earlier LCFS data set (Druckman and Jackson, 2008.). This trend concurs with studies discussed above and the NEED framework data. In order to explore the strength of this relationship, and tease out other dependencies, a TAN BN was constructed using household income as the classifier variable.

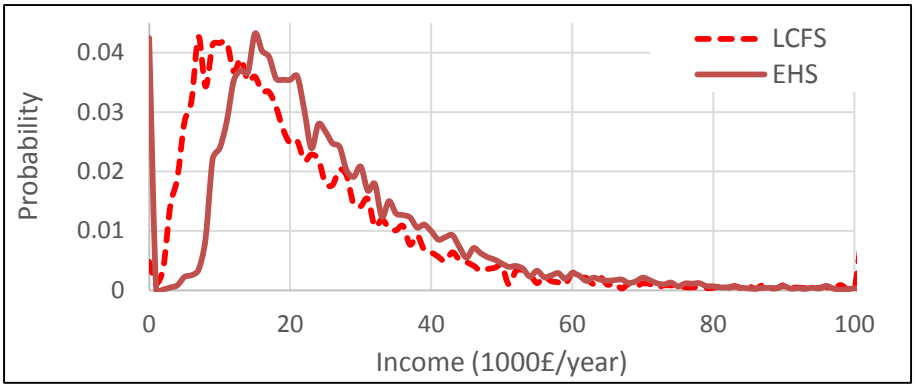


Figure 6-10 Frequency distribution of equivalised household (OECD) income from LCF survey 2010 and EHS 2010-11

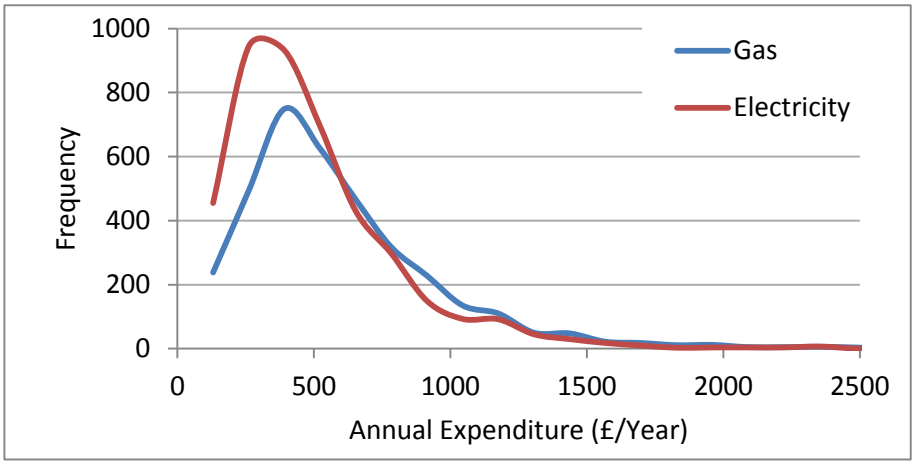


Figure 6-11 Frequency distribution of domestic gas and electricity expenditure from LCF 2010

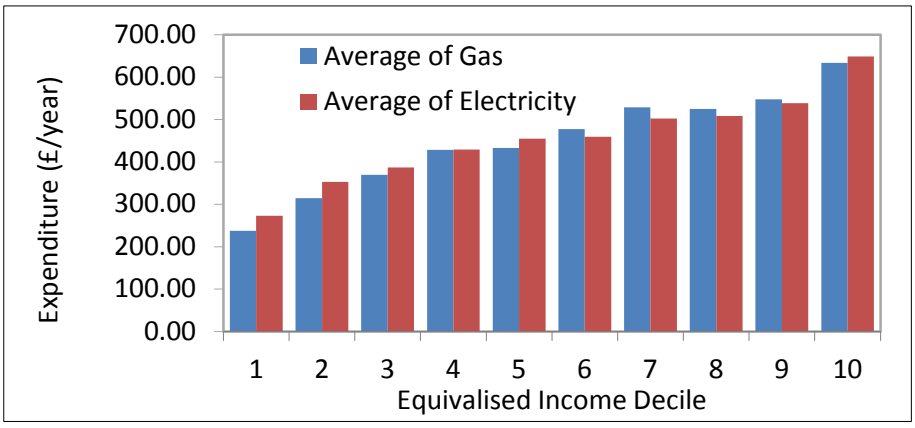


Figure 6-12 Annual average household expenditure on gas and electricity as a function of equivalised household income (OECD) decile

TAN analysis of the living cost and food survey dataset

All the pertinent variables in the dataset where further explored using a TAN BN learnt using Netica followed by NPT learning using the count method. Income was chosen as the classifier variable and the resultant BN is shown in Figure 6-13. A variance reduction sensitivity analysis was carried out with household income, gas consumption, and electricity consumption as the target nodes (Tables 6-17, 6-18 and 6-19).

This analysis indicates a highest correlation of income to tenure type, and significant correlation with the remaining variables apart from central heating type. Gas and electricity have a high correlation with each other, suggesting that high consumers of one fuel are also high consumers of the other as also noted in the NEED analysis. The BN clearly showed that when evidence was entered for the state with the highest electricity demand there was still the highest probability of the highest gas demand (39%) but also showed the second highest probability (26%) of zero gas demand (Figure 6-14).

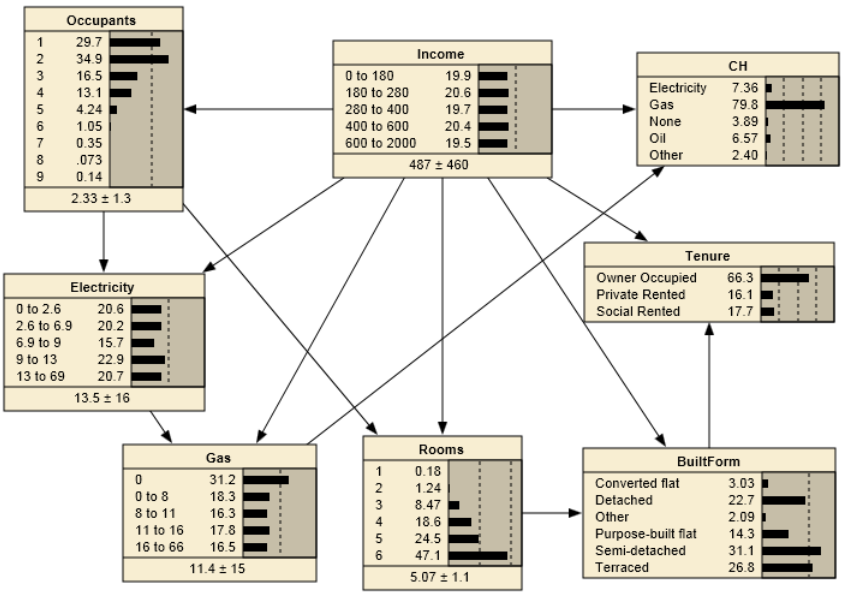


Figure 6-13 Tree Augmented Naïve Bayesian Network Classifier for Income Using the LCF.

Gas and electricity exhibit a sensitivity to household income; significantly electricity is more sensitive to the number of occupants than is gas, supporting Steemers and Yun's (2009) suggestion that buildings are heated regardless of occupancy, but that electricity demand is driven by consumer appliances and hence active occupancy as supported by Richardson et al. (2010).

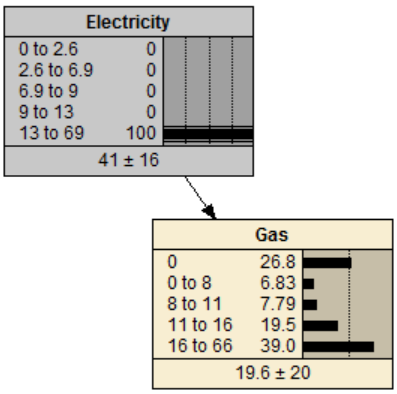


Figure 6-14 Selecting the state for the highest electricity consumption shows both highest and lowest electricity consumers with high probability

Table 6-17 Sensitivity of Income to findings at other nodes

Node	Variance Reduction	Percent
Tenure	1.84E+04	8.7
Electricity	1.12E+04	5.28
Rooms	8359	3.96
Built Form	7881	3.73
Gas	7510	3.56
Occupants	6513	3.08
CH	1540	0.729

Table 6-18 Sensitivity of Gas to findings at other nodes

Node	Variance Reduction	Percent
Electricity	32.37	13.9
CH	25.04	10.8
Income	8.581	3.69
Occupants	2.091	0.9
Tenure	1.619	0.697
Rooms	1.183	0.509
Built Form	0.7665	0.33

Table 6-19 Sensitivity of Electricity to findings at other nodes

Node	Variance Reduction	Percent
Gas	28.09	10.7
Occupants	13.75	5.21
Income	9.026	3.42
Rooms	3.699	1.4
Tenure	2.099	0.796
Built Form	1.975	0.749
CH	1.215	0.461

In summary the LCF provides a JPD for income and energy expenditure which can also include several other variables such as tenure. It is evidenced from the variance reduction technique that the TAN BN classifier for income demonstrates significant direct dependencies on energy expenditure. Furthermore electricity shows a strong dependency on gas though this is not a monotonic relationship; further detail can be extracted with a probabilistic analysis to represent off-gas or low gas consumers.

In theory the dataset could be used to augment the NEED energy consumption model with an income parameter. This would require the conversion of energy expenditure to energy demand using known energy costs. This introduces additional uncertainty since tariffs are a latent variable in this model. A conversion was carried out by Druckman et al (2008), but Preston et al. have found the method to be unreliable and adopted alternative approaches in their DIMPSA model (2010). A final attempt to source tabular data with which to link income data and energy demand is discussed next.

6.3.3 English Housing Survey and the Cambridge housing Model

The English Housing Survey (EHS) for 2010-11 has been produced by the ONS on behalf of the Department of Communities and Local Government (DCLG, 2012). The survey has been carried out in its current form since 2008 though predecessor surveys have been conducted since 1965. It is the UK Government's primary tool to categorise and assess the condition of the English housing stock.

The survey consist of CAPI data of around 17,000 households and a follow up physical building survey of half this number over two years. This has provided 14,000 records of building attributes linked to occupant data such as number, age, and household income. The dataset is furnished with annual weights to allow scaling up to the national housing stock.

Energy consumption data is not generally an integral part of the survey though this has been carried out in 1997 and a small follow-up energy survey in 2011 which cannot be analysed in conjunction with the EHS. As discussed in section 6.2, EHS data is used to furnish bottom up building stock models with building attributes for the estimation of energy consumption. In this analysis estimated energy consumption data generated by the Cambridge Housing Model is included for comparison to the LCF data.

Analysis of the EHS

As with the LCF, the purpose of the analysis of the EHS was to assess the dataset as a JPD for the building energy sub model and is of particular interest since it provides tabular data which includes household income, each of the building attributes included in the NEED dataset, along with gas and electricity consumption, albeit estimated using a BREDEM model. This is in contrast to the LCF with its very limited building attributes (only built form), and energy consumption which can only be estimated using a derived annual fuel expenditure.

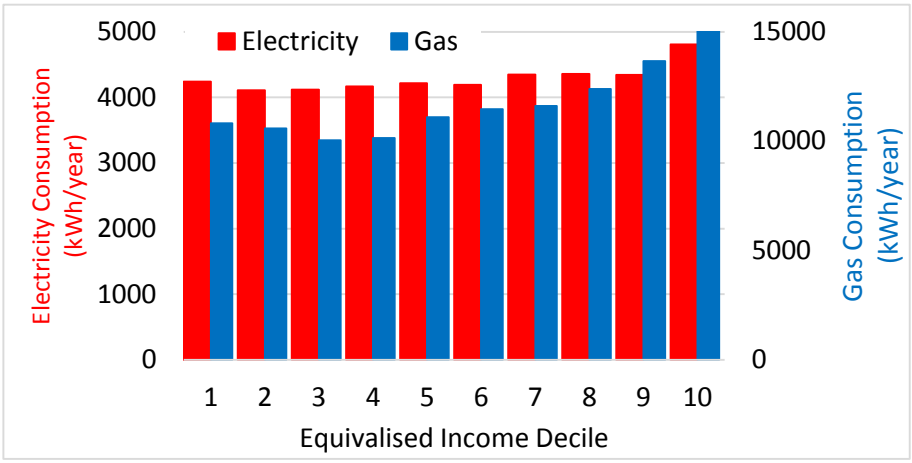


Figure 6-15 Gas and electricity consumption estimated using the Cambridge housing model for the English Housing Survey stock and interview data.

Figure 6-15 shows the average gas and electricity consumption as a function of household income decile. This should be compared with the similar LCF analysis (Figure 6-12), which, although it shows energy expenditure as a function of income, is indicative of a more stark dependence of energy consumption on household income than the CHM modelled data.

The variability of modelled gas and electricity consumption was compared with the empirical data in NEED dataset (Figure 6-16). The distributions occupy the same region of gas consumption and the mode values are similar, at about 10GWh per year. It is clear, however, that the CHM fails to model very low and very high gas consumption well. Modelled electricity consumption delivers a poor representation of the empirical distribution exhibited by the NEED dataset. The preponderance of low consumers in the empirical data is not present in the modelled data and the mode consumption is significantly higher at 3.5GWh per year compared to 2.5GWh per year for the empirical data.

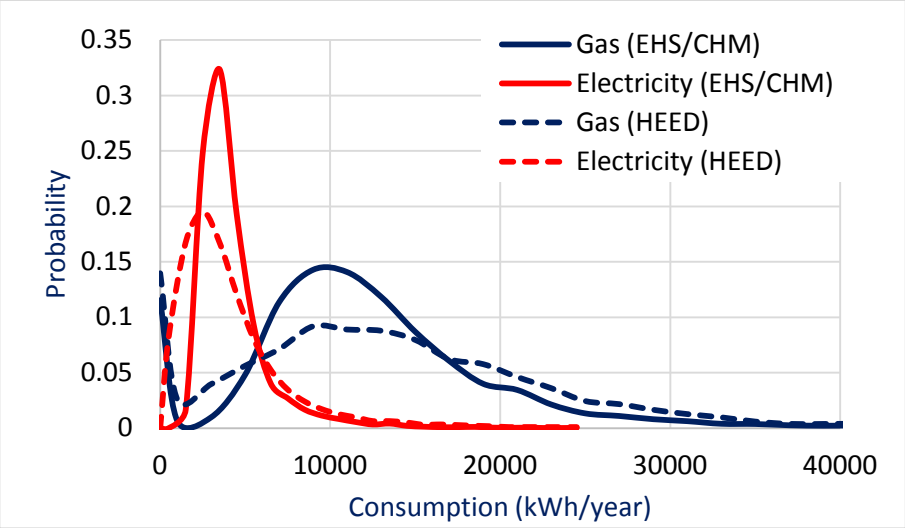


Figure 6-16 Distribution of estimated Gas and electricity consumption using the Cambridge housing model for dwellings in the 2010-11 EHS

Finally, when comparing the OECD equivalised income distribution for the EHS dataset with the LCF, it is observed that the former has significantly higher income values than the latter. The reason for this is unclear since both surveys claim to include income from earnings, benefits and other sources and use the OECD method to equivalise household income.

These shortcoming cast doubt on the utility of the EHS/CHM modelled energy consumption data set for delivering a tabular dataset which incorporates both building attributes and household income.

6.3.4 Summary

Three empirical datasets have been explored to supporting the structure of a BN model and furnishing a BN model with the data required to create NPDs. The salient variables in these are summarised in Table 6-20 along with the tabular data sources which include them.

Table 6-20 Parameters for a building energy model with sources of tabular data

Variable	Abbreviation	Sources
Annual Electricity Consumption	E	NEED;EHS[a];LCF[b]
Annual Gas Consumption	G	NEED;EHS[a];LCF[b]
Built Form or Building Type	T	NEED;EHS;LCF
Total Floor Area	F	NEED;EHS
Building Age	A	NEED;EHS
Region	R	NEED;EHS;LCF
Household Income	I	EHS[c];LCF[c]

Notes
a. Modelled data using BREDEM model
b. Derived data from expenditure
c. Incomes between two source inconsistent

For the purpose of learning BN structures and dependencies (conditional probabilities) a summary of the JPDs which can be derived from the tabular data using the given data sources are as follows:

NEED $P(E, G, T, F, A, R)$

EHS $P(T, F, A, R, I)$

LCF $P(T, R, I)$

The LCF has been useful for reinforcing belief in a relationship between income and energy consumption, but does not deliver comprehensive empirical data for a probabilistic model. The ability of the EHS to furnish models a JPD which include energy consumption is rejected due to the poor comparison between the modelled and empirical data. The EHS does however permit the integration of income with building attributes. In contrast the NEED data delivers a JPD which does include building attributes and energy consumption, but does not integrate income.

Thus it is possible to model the building attribute variables which have a direct influence on energy consumption using the NEED dataset, but not the weak direct influences of household income identified by Kelly, and Yun and Steemers. But by using the EHS dataset it is possible to model the dependencies between building attributes and household income, and therefore the stronger indirect influences on energy consumption.

Evidence has also been found for a dependency relationships between built form and building age, floor area and region which might be represented by a conditional relationship, $P(T|R, A, F)$.

The next section combines the potential of the datasets discussed in this section, and the dependencies generally identified in section 3.2 to propose a BN subcomponent.

6.4 Bayesian Network Submodel for Building energy Demand Prediction

6.4.1 The Directed Acyclic Graph (DAG)

The outcome of the review to create an ontology of variables, coupled with a critical review of available data has yielded a number of potential predictor variables for domestic gas and electricity demand together with their dependency relationships. The direct influence of household income cannot be modelled with the available data. Dependencies between building attributes have also been highlighted.

The building stock model, in which these same dependencies can also be modelled, is discussed in chapter 7. This presents what has been termed in this thesis 'the dependency ownership dilemma' of the object oriented model. In this context the problem is represented by Figure 6-17. Here the dependencies between building attributes can be represented in the building stock model (represented by the red nodes and arcs) or in the building energy model (represented by the blue nodes and arcs).

Since the two objects are joined by interface variables which are common to both objects and which therefore have a one-to-one relationship, it might seem unimportant where these dependencies are represented. However, in this thesis the view is taken that the intra-building attribute dependencies should be modelled in the building stock model, not in the building energy model founded upon the

NEED dataset. This is because the building stock model encapsulates more spatially specific knowledge of the building stock, at the LSOA level, rather than the less spatially specific relationships in the NEED dataset which can only be spatially resolved to the regional scale.

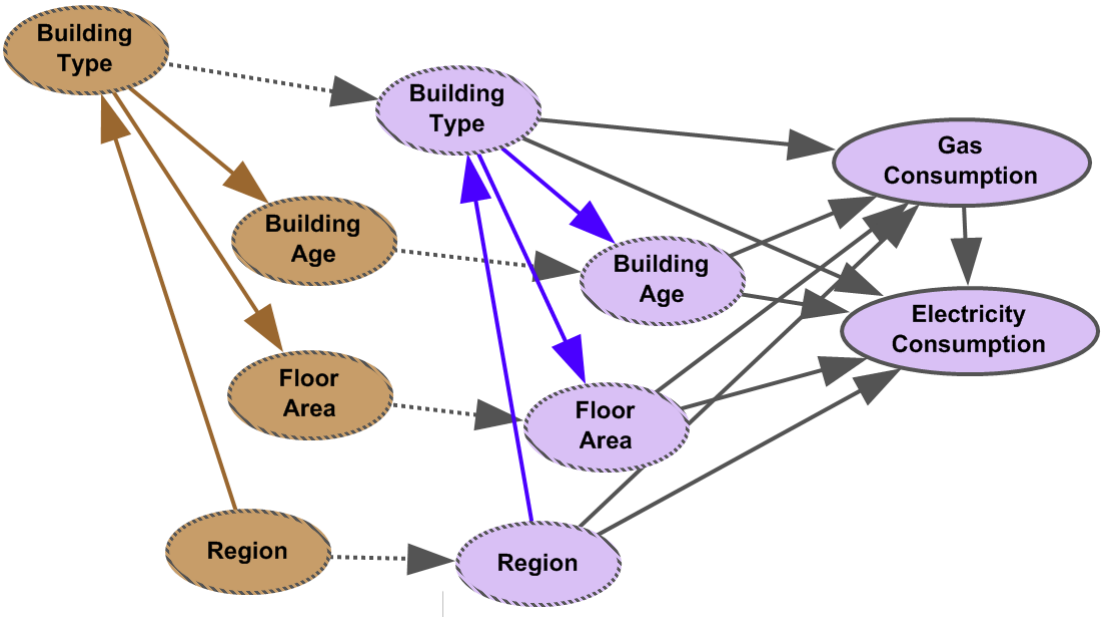


Figure 6-17 the dependency ownership dilemma between the building stock model (red nodes and arcs) and the building energy model (blue nodes and arcs).

In the exploratory TAN BNs developed using the NEED dataset the findings on one energy consumption vector was found to be sensitive to the other. Thus it is deemed appropriate to draw an arc between gas and electricity. The direction of this is immaterial since there is no causal assumption represented by this arc; it merely represents the observed probability of gas consumption conditional on electricity consumption or vice versa.

Taking into account the delegation of building attribute dependencies to the building stock model, and the relationship between gas and electricity consumption, the final submodel for building energy demand is shown in Figure 6-18. Here gas consumption has been made a child of electricity

consumption. The next section presents the data with which to construct the node probability tables for each of the nodes in this model.

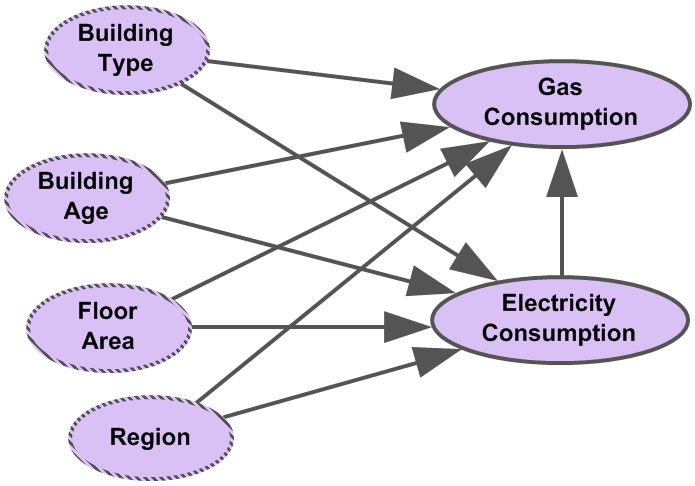


Figure 6-18 DAG for the Bayesian network submodel for building energy demand

6.4.2 Building the Netica Model

The DAG in Figure 6-18 was built in Netica as a standalone BN (Figure 6-19). Region, Building Age, Floor Area and Built Form nodes all represent discrete variables. Gas and electricity, whilst not continuous were present in 129 bands; these were reduced to only 25. This is particularly important for performance of the model since the energy nodes have four or five parents resulting in an unusually large CPT.

The anonymised NEED data set was used to furnish the BN submodel in Figure 6-19 with NPTs. These were learnt using the counting method using a Microsoft Access database as a case file consisting of over four million records. Some preparation of the original source file was required. All nominal categories in the NEED dataset are numeric codes which had to be converted to text values to be acceptable for Netica. This was achieved using standard SQL update queries to precede each numeric

code with a letter. In order for Netica to recognise the weighting column this was renamed to 'NumCases' which Netica uses to identify the column as a multiplier in the counting algorithm.

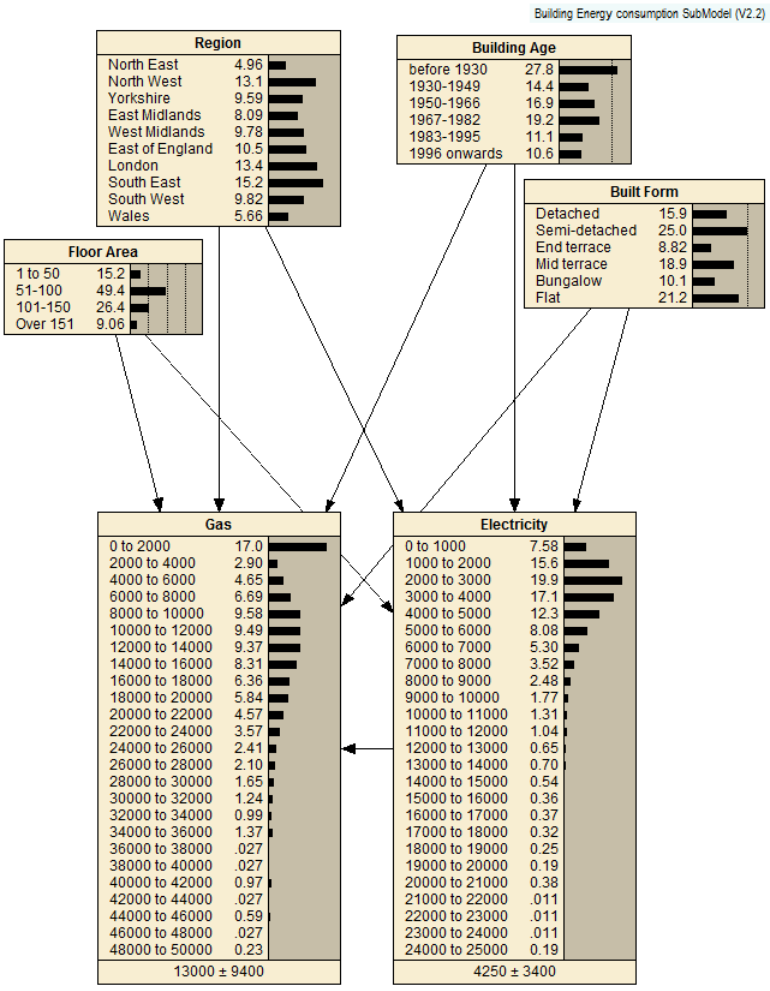


Figure 6-19 Netica Bayesian network sub-model for building energy The Units for gas and electricity are kWh/year, and the floor area is m^2.

After learning the case file the BN was verified against the results in the NEED analysis report to ensure similar trends were observed for gas and electricity consumption, as region, property type, floor area and property age were varied by selecting hard evidence for the respective nodes. The BN was found to concur with the published results. As an example of the outputs of the BN submodel, Figure 6-20 and Figure 6-21 show the distributions of gas and electricity consumption produced by the model in Netica as different hard evidence for floor area is selected.

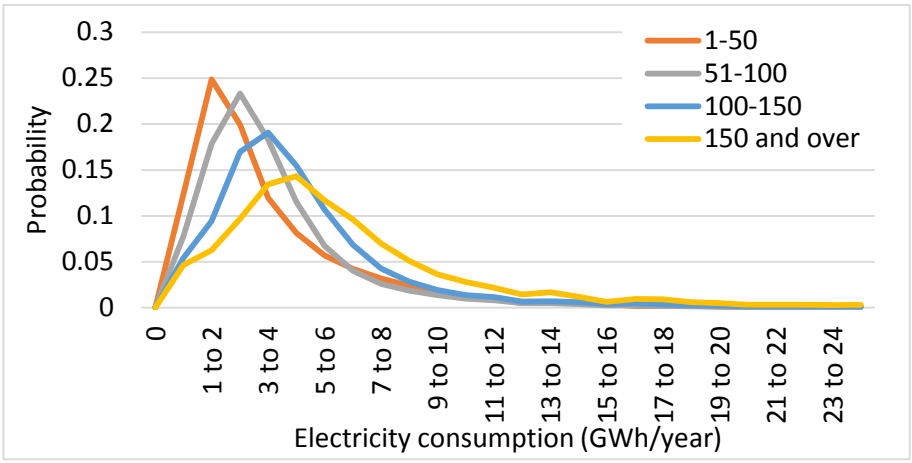


Figure 6-20 Data extracted from Netica sub-model showing probability distribution of electricity consumption with hard evidence for floor area

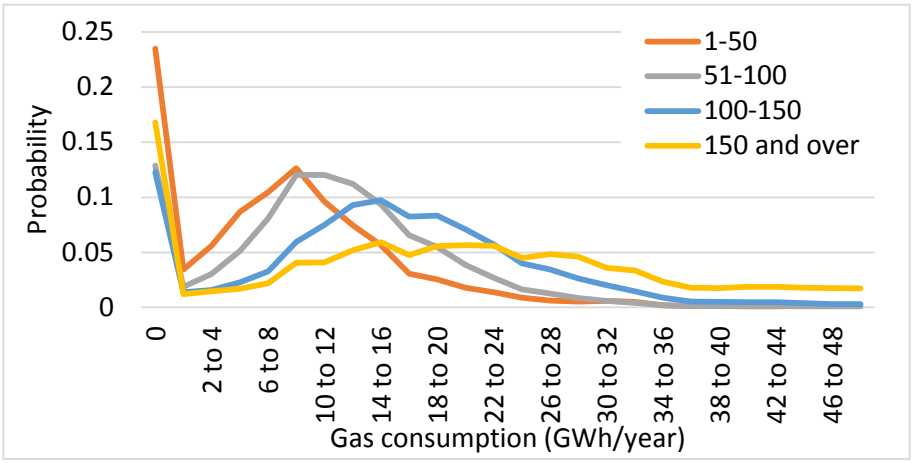


Figure 6-21 Data extracted from Netica sub-model showing probability distribution of gas consumption with hard evidence for floor area

In addition, statistical analysis of the source tables was used to verify expected values given by the BN for specific hard evidence. The result of this for hard evidence for floor area is shown in Table 6-21.

Table 6-21 Expected value for gas and electricity consumption for different floor areas compared with source data.

Floor Area (m^2)	*Expected Value kWh/year*	
	Electricity	**Gas**
1-50	3668	8415
51-100	3800	11523
100-150	4744	15595
150 and over	6159	19209

6.5 Discussion and Conclusions

Bottom-up building energy models traditionally use building physics models to deterministically predict gas and electricity consumption given a number of key building attributes. Statistical models which have been extensively used to research the relationships between parameters which influence energy consumption have been reviewed. This is to support the construction of a probabilistic graphical model which can predict gas and electricity consumption given hard or probabilistic evidence for key building attributes and UK region.

The NEED framework dataset has been used to provide the marginal probability distributions in the model. The critical elements in the model are the CPTs used to predict, probabilistically, annual electricity and gas consumption given the region, and key building attributes, floor area, building age and built form, and in the gas of gas, the electricity consumption (Equations 6-1 and 6-2).

$$P(G|E,T,F,A,R) \qquad \text{Equation 6-1}$$

$$P(E|T,F,A,R) \qquad \text{Equation 6-2}$$

The model has been verified against existing published NEED data, exhibiting the trends shown by this and other research. The model can be used on its own. The marginal distributions are representative of the national building stock thanks to the weighting column. However, it is essentially a naïve BN – there are no dependencies between building attributes and the region.

By furnishing one or more nodes with hard evidence, a probability distribution of gas and electricity consumption is presented to the user. This is a very different result than a deterministic building energy model which will deliver a deterministic answer with a margin of error. In this model uncertainties have been endogenised and the answer has to be interpreted as a probability for each possible value in the discretised distribution, or one could determine the probability of the answer being over or below a specific value for the purposes of risk assessment. For example, there is a 20% probability of a flat in the East Midlands with a floor area of 50m^2 or less having a gas consumption of 5000 kWh/year or more. This probability rises to 32% if the flat is known to be built before 1930.

The NEED analysis discussed above, which had full access to the income data from Experian, reported that only 30% of the variability of total energy consumption can be explained by building attributes alone. Thus all things being equal there is still a large variability in energy consumption. This in itself suggests that if risk evaluation is an objective, a probabilistic approach must be entertained. A key outstanding question is: what are the latent variables which would explain some of the remaining variability? One known unknown in this model are occupant influences; a large number of studies and datasets explored above shown that household income is a predictor, both directly and indirectly of energy demand. There are several theses to support these dependency relationships. Firstly, it is known that dwelling size and built form influence energy type and it is can be normatively suggested that higher earning households are more probably found in larger dwellings and detached houses – the indirect influence. Secondly, it can also be reasoned that higher income households will have a higher direct energy demand since there may be more occupants with commensurate more energy consuming practices or behaviours. It might also be normatively assumed that even with an equal and similar amount of occupants, greater incomes might lead to more profligate behaviours, though this would need to be supported by empirical evidence.

It is, therefore, disappointing that household income could not be incorporated directly in to the building energy demand model to explain the direct influences of occupant parameters. Thus, this

variable remains latent in this model. However, the uncertainty attributed to it is endogenised. It should be compared with the CHM model which badly models low and high consumers possibly due to occupant modelling which does not reflect real life behaviours. This is certainly true of heating temperature settings.

The question remains as to whether the indirect occupant influences can be modelled – namely the relationships between household income, building attributes, and region. In the next chapter, on the building stock model, it is proposed that it can be. Indeed in the discussion above on 'the dependency ownership dilemma', it was concluded that the building attribute inter-dependencies are best determined using the actual building stock being modelled – and this argument applies equally for the income.

Whilst the model can be used on its own, and is, in itself, useful for exploring the NEED data, its purpose is to be a constituent component of a larger OOBN. As inputs it receives probabilistic evidence for the region, property age, built form, and floor area. These are to be provided by a building stock model which reflects the spatial area of interest. The next chapter considers a building stock model for the chosen spatial scale, the LSOA, and the four case study areas.

7 Building Stock

7.1 Introduction

A key element in the conceptual model (Chapter 4) is the 'installation site'. This has direct and indirect influences on both energy generation and demand. This chapter presents the development of the installation site element as a BN sub-model. Since the installation sites are buildings, this is designated the *building stock sub-model*. The purpose of this sub-model is to furnish the generation and demand components of the OOBN with probabilistic evidence for the requisite building stock attributes whilst encapsulating their dependencies.

Section 7.2 presents the required parameters and explores the dependency relationships between them to provide an ontology for the domain. Section (7.3) considers the data sources which can be used to empirically quantify the parameters and their dependencies, as quantified by conditional probabilities. These data sources are critically reviewed and their processing and analysis described in detail.

Section 7.4, in a similar manner to early chapters, integrates the theoretical requirements of the ontology, and the practical requirements of the available data sources, and proposes the DAG for the BN sub-model. The required NPTs and data sources are presented which determine the final BN sub-model. Finally a discussion and conclusions are presented in Section 7.5.

7.2 Ontology

The requisite parameters for the building stock sub-model are those required as inputs by the solar PV yield and the building energy consumption components. These constitute the interfaces between the sub-models, as shown in the UML diagram in Figure 7-1. Required as inputs to the building

energy consumption sub-model are the floor area, building age, built form, and region; the PV system yield sub-model requires region, orientation, pitch and roof area (Table 7-1).

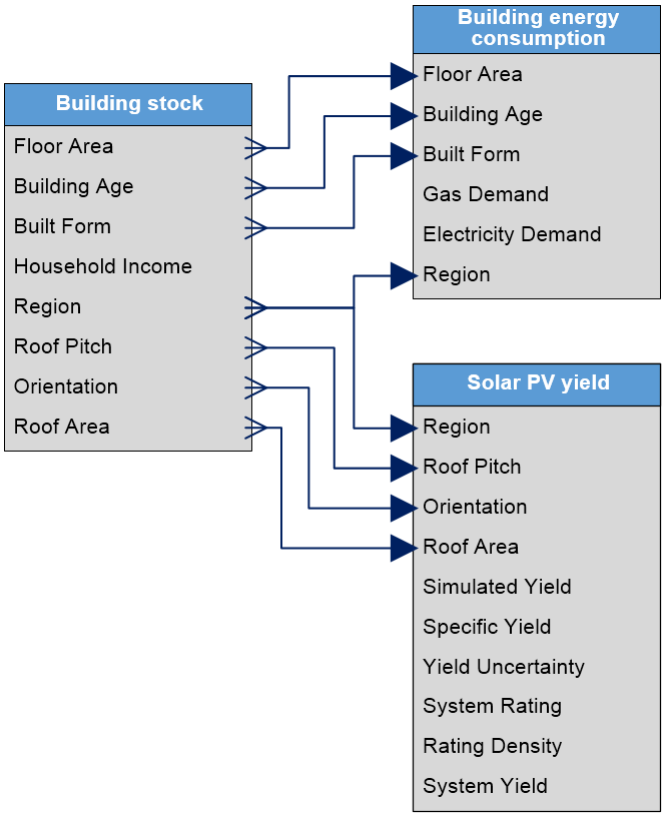


Figure 7-1 UML diagram showing the interfaces between the building stock, building energy demand and energy yield sub-models

Table 7-1 Parameters required in the building stock sub-model alongside the sub-model to which they interface. Abbreviations are used in equations.

Building attribute	**Abbreviation**	**Sub-model**
Income	I	
Floor Area	FA	Building energy consumption
Building Age	BA	Building energy consumption
Built Form	BF	Building energy consumption
Region	R	Building energy consumption PV Yield
Orientation	O	PV Yield
Pitch	P	PV Yield
Roof Area	RA	PV Yield

It is not proposed to discover more parameters - and there is a debate to be had whether the models so far developed are too parsimonious - but to learn the dependency relationships between those already included. The building stock sub-model could be assumed to be a naïve Bayesian classifier (Friedman et al. 1997) for the particular building stock dataset under consideration, i.e. one where all the parameters are assumed to be strongly independent of each other (Figure 7-2). However, this idea was rejected due to the presence of several manifest dependency relationships described below. A key objective of the building stock analysis described in this section was, therefore, to characterise any such dependencies as may exist.

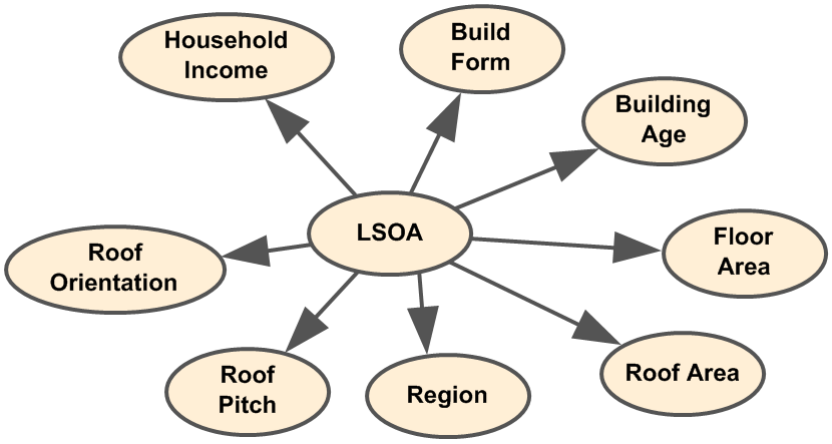


Figure 7-2 Naïve Bayesian network classifier for the building stock model where all parameters are dependent on the LSOA dataset but mutually independent of each other

Furthermore, some of these dependencies have already been elicited in chapter 6 but were not encoded in the DAG. There are two reasons for this. Firstly, there are dependencies between parameters which appear in separate models. For example, the roof area, which appears in the PV yield model, has a manifest dependency on floor area, which appears in the building energy model. The second reason is due to what in Chapter 3 was termed the 'dependency ownership dilemma' and was discussed in this context in section 6.4.1. The encoding of dependencies with CPTs in the building stock model is more appropriate than a generic encoding in the building energy

consumption or PV yield sub-models. This is in concurrence with the object-oriented paradigm where the sub-models are agnostic with respect to actual housing stock inputs. This section, therefore, will 'bring forward' the dependencies learnt in chapters 5 and 6.

7.2.1 Dependencies between building stock parameters

Evidence from analysis of the EHS (Chapter 6) suggested that the four parameters required for building energy consumption exhibit dependencies as depicted in the undirected graph in Figure 6-7. Graphical representations of these dependencies are shown as conditional probabilities $P(BA|BF)$ in Figure 7-3 and $P(FA|BF)$ in Figure 7-4. It is clear from these that converted flats have a very high probability of being built before 1918, whereas detached dwellings are most probably post 1945. Flats have the highest probability of having 50m^2 or less floor area, whereas detached dwellings are more likely to be 150m^2 or over. Generally, therefore, building age and floor area exhibit a dependency on built form and Equations 7-1 and 7-2 apply.

$$P(BA|BF) \neq P(BA) \qquad \text{Equation 7-1}$$

$$P(FA|BF) \neq P(FA) \qquad \text{Equation 7-2}$$

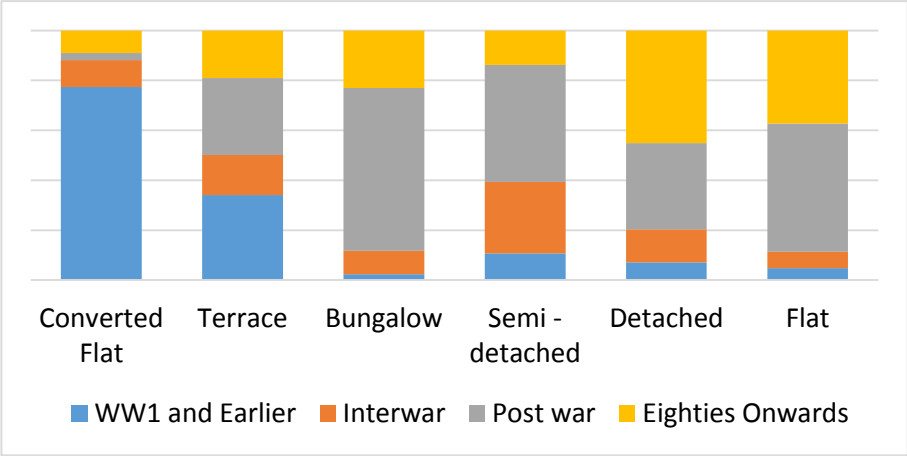


Figure 7-3 Distribution of age for housing stock of each built form

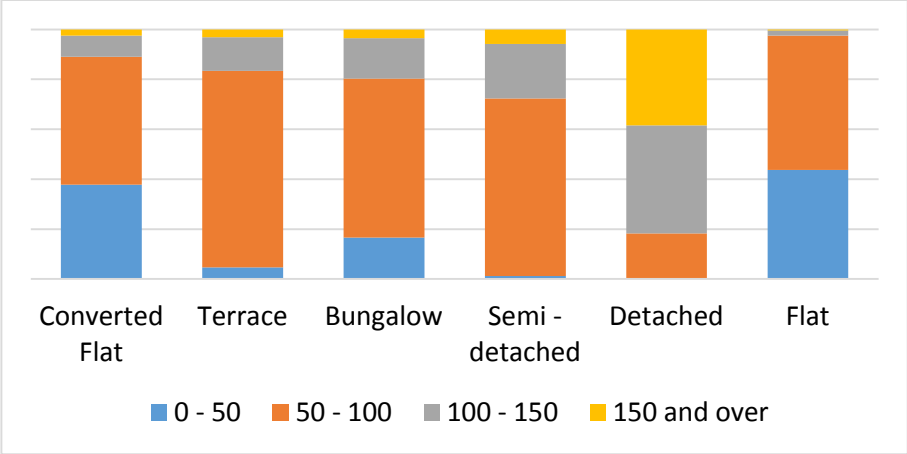


Figure 7-4 Distribution of floor area for housing stock of each built form

The dependency of built form on the region for the EHS dataset is shown in Figure 7-5. London is exceptional with a large probability of purpose built and converted flats. The remaining regions are similar with a slightly greater propensity for northern regions to have terraced housing and slightly less detached dwellings.

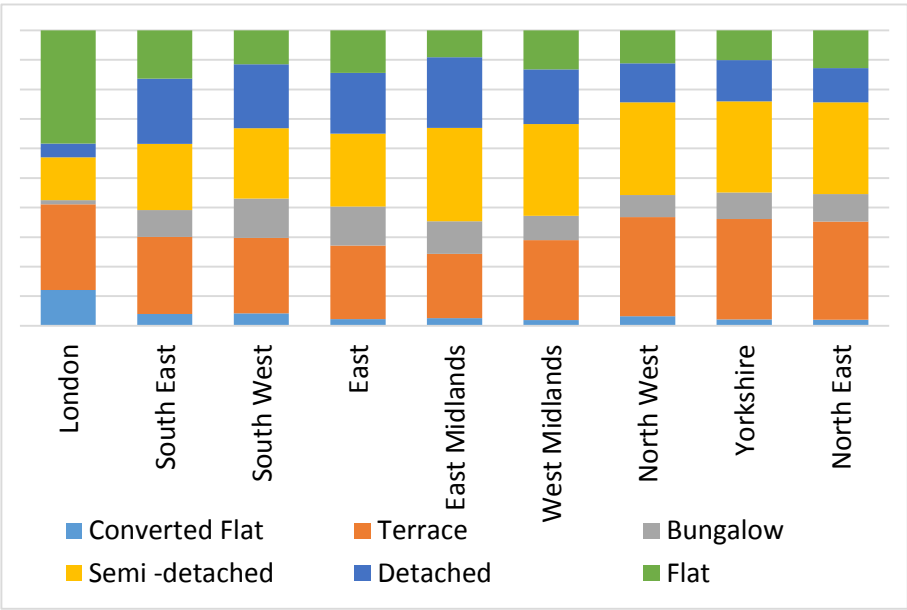


Figure 7-5 Distribution of built form for each region in the EHS dataset

Whilst the national and regional pictures are academically interesting, these are not likely to reflect the specific spatial scale of the LSOA. This suggests a graphical structure where the building attributes are all dependent on the LSOA node, but building age and floor area depend on the built form (Figure 7-6). This reinforces the idea that these specific LSOA dependencies should be encoded in the building stock model and learnt using the actual building stock data for the LSOA.

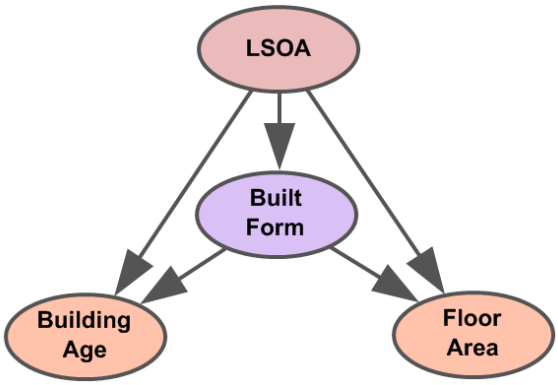


Figure 7-6 Suggested dependencies between building attributes and the LSOA

The remaining parameters in Table 7-1, the roof pitch, area and orientation influence the PV yield. There is a large body of knowledge on solar energy potential of building roofs discussed briefly in the next section.

7.2.2 Domestic Roofs

The diffusion of solar PV in to the urban environment is driven by the ready availability of roofs for their mounting. As discussed in Chapter 5, the geometry and spatial orientation of urban roofs is a determinant of the solar potential of the built environment. It is pertinent to discuss whether there are dependencies between these parameters.

The roof area is manifestly dependent on the overall building foot print. If both these parameters can be sourced for the building stock model, the dependency between them can be encoded using CPT learning in a straightforward manner. The building, and therefore roof orientation, is most often dependent on the street orientation rather than any purposeful orientation by architects to the axes of the compass. Such practices are a relatively modern phenomena of low carbon building designed to maximise passive solar gain.

The third parameter is the roof slope, which is a feature of the roof style. The latter may be dependent on building attributes, particularly built form, but, without a large set of tabular data the statistical relationships between slope and other parameters cannot be determined at present, although work in this area is in progress by Palmer et al., (2015).

Roof slope is furthermore problematic, since it is not, generally, a feature of any building stock dataset. To obviate the need for expensive or time consuming site surveys the development of models to estimate solar energy potential from geospatial data sources is an attractive proposition for research and urban planning purposes (Rylatt et al., 2003). Such techniques are often augmented by statistical knowledge of roof architecture to, for example, infer pitch, roof styles and the likelihood of shading (Ordonez et al., 2010), or, to extrapolate to larger geographies (Wiginton et al., 2010).

The availability of contemporary high resolution digital photography has made the characterisation of roofs for solar potential readily accessible. Small areas of heterogeneous housing stock have been characterised in this way to yield probability distributions of annual insolation (Araya-Munoz et al., 2014) whilst multi-story buildings have had their roof geometries estimated for the socio-economic assessment of PV (Orioli and Di Gangi, 2014). A disadvantage of this approach was the labour intensity, making the quantification of large areas impractical. Automatic roof characterisation was executed using an analysis of vector maps in GIS software (Rylatt et al., 2003), and using

sophisticated feature recognition algorithms to auto-detect roof features such as chimneys and dormer windows in digital raster maps (Ordonez et al., 2010).

The most significant development in recent years is the use of lidar[31], a distance measuring technique using laser light, which can be used to make high resolution topographic (3D) maps (Melius et al., 2013). This overcomes the 2-dimensionality of digital photographic methods which require the employment of statistical approaches for the assessment of solar potential. Lidar has found increasing application in landscape surveying since it has a high enough 3D spatial resolution to impart detailed dimensional information about buildings in both rural and urban contexts. A vertical and horizontal accuracy of 50cm, this has found application in the analysis of the urban roof-scape, with the ability to determine pitch and identify larger obstructions and discontinuities automatically. Collection of lidar data requires numerous overflights by aeroplanes and is expensive for general use but a number of research projects have evaluated this technique for solar potential assessment (Nguyen et al., 2012). Low (2m) resolution lidar data had to be supplemented by a applying a roof profile from a common catalogue within the building footprint (Jacques et al., 2014). In section 7.3 the sourcing of lidar assessment of roof parameters to supplement the building stock is for the LSOA case study areas is presented.

7.2.3 Household income

The acquisition of household income data alongside associated building attributes for a housing stock of interest is not a straightforward proposition. In Chapter 6, the EHS and the LCFS were assessed for their utility for using income to predict energy consumption. The EHS provides tabular data which incorporates income with the main building attributes, but not empirical energy

[31] Lidar is a portmanteau if 'light' and 'radar'.

consumption. Whilst there it was found infeasible to determine direct dependencies between income and energy consumption, the dependencies between income and building attributes are theoretically determinable (Chapter 6). However, the EHS, and other government surveys, are not spatially disaggregated to a spatial scale smaller than the region and there is a paucity of empirical socio-economic data for small area geographies (Anderson, 2011). The UK census has a range of questions to elicit household incomes from user responses. However, this is not released as 'microdata' for specific localities, but as part of an anonymised data set, or presented as aggregated statistics for small areas such as the index of multiple deprivation.

A number of commercial data providers have resorted to modelling income distributions using credit ratings and other consumer intelligence to estimate household incomes for small areas. Experian plc was selected by DECC to furnish the NEED framework with household income data though this, due to commercial license agreements, was not released as part of the anonymised dataset (Chapter 6). CACI ltd provide income data in £5000 bands, mean and median and mode for every UK postcode derived from market research data and UK Census returns (CACI, 2014). Both products are widely used for commercial marketing purposes. Data could not be sourced for this research, however, as costs were prohibitive. A further disadvantage is that their methodologies are not in the public domain (Whitehead et al., 2009).

In contrast, a fully documented micro-simulation method has been developed to estimate household incomes at the LSOA level (Anderson 2011, Anderson 2013). This technique uses iterative proportional fitting (IPF) to simulate microdata using "exogenous data constrained by known endogenous parameters" (Lovelace and Ballas, 2013). In this case employment status, the number of earners, the tenure and gender of the household reference person from census data were used to constrain a regional dataset of the Family Resources Survey, which contains those same variables alongside household income data. The sourcing of simulated household income data using this method, and the use of IPF to fit this to LSOA building stock data are described in section 7.3.

7.2.4 Dependencies between Building Attributes and Income

Evidence is required to support the relationships between household income and building attributes. Household income has a manifest influence on the types of houses people purchase or rent. Whilst it may be normatively assumed that higher income households are more likely to live in larger, detached dwellings, little statistical evidence is available. This section seeks to gain insight into the quantitative dependency relationship between each parameter and household income using the EHS dataset. Using income decile as a proxy for household income the following figures show an analysis as a function of floor area, built form and building age. Note that care was taken to consider the weights in the EHS data so that these relationships represent that observed for England and Wales.

Figure 7-7 shows the average and standard deviation of floor area for each income decile, demonstrating an increase in floor area as income rises albeit with a large variability. This supports the observation by Kelly (2011) of the indirect effect of income on energy expenditure due to its direct influence on floor area (Figure 6-3).

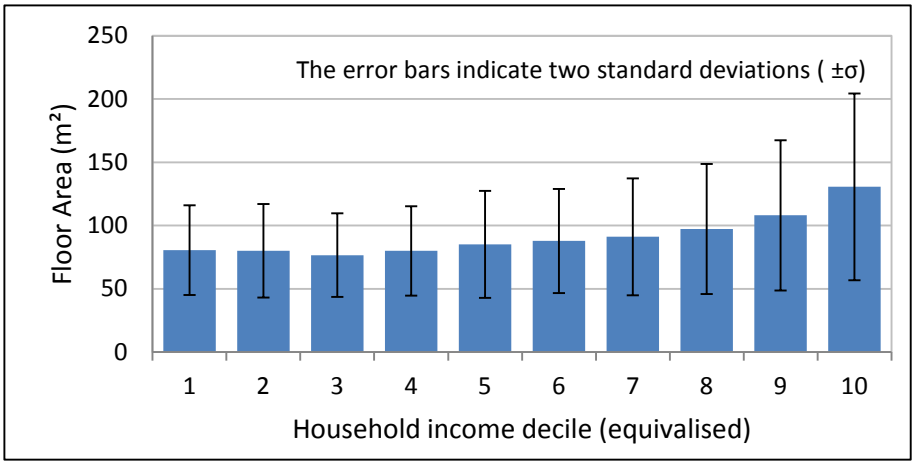


Figure 7-7 Floor area in the EHS dataset as a function of income decile

Figure 7-8 shows the proportions of built form for each income decile showing a marked dependency between them. This is highlighted by comparing the first (lowest income) decile with the tenth highest) income decile. The former are much less likely to live in a detached and more likely to live in

flats and terraced housing whereas for the latter the probability is reversed. Semi-detached are the most prevalent UK built form with an almost constant likelihood of this property type for all income groups except for the highest deciles.

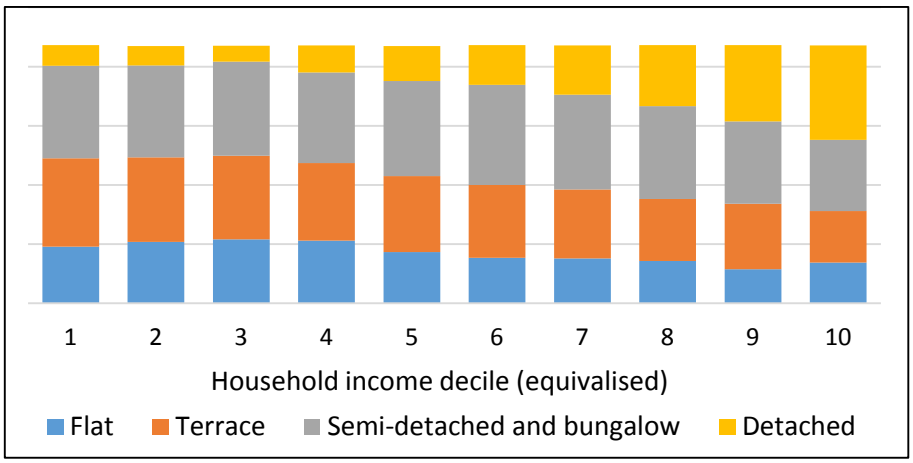


Figure 7-8 Proportions of built form by income decile

The dependency between building age category is observed in Figure 7-9. The very newest and oldest properties are more likely to be inhabited by higher income decile groups whereas there is a greater probability of post-war properties being inhabited by those on lower incomes.

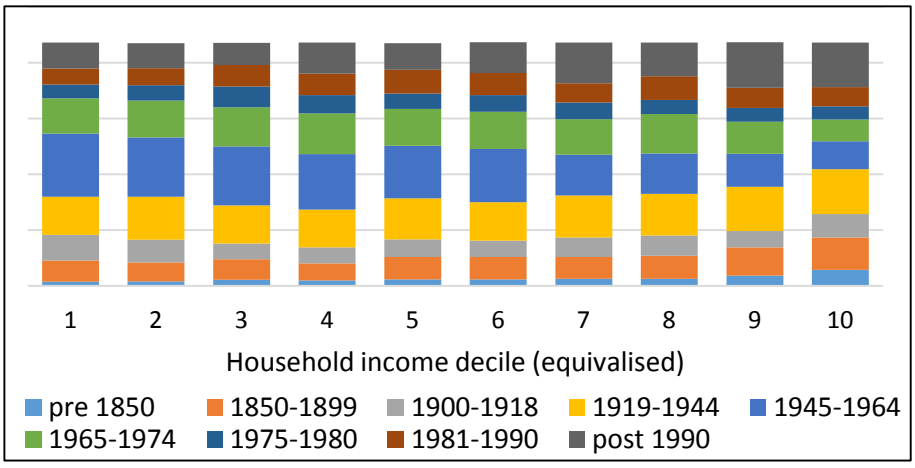


Figure 7-9 Proportions of building age categories by income decile

It would be appropriate to quantify these relationships with regression techniques but these graphical analyses are convincing enough to suggest the dependencies between these broad building attributes and income need to be encapsulated in the BN model. Generally we can state with confidence that Equation 7-3, Equation 7-4 and Equation 7-5 apply.

$$P(I|BF) \neq P(I) \qquad \text{Equation 7-3}$$

$$P(I|FA) \neq P(I) \qquad \text{Equation 7-4}$$

$$P(I|BA) \neq P(I) \qquad \text{Equation 7-5}$$

7.2.5 Summary

The purpose of this section was to characterise the dependencies between parameters required for the building stock model using literature and other sources. However, there was little extant literature which relates dependencies between household and building attributes. Thus recourse was made to national building stock data and income data within the EHS. A semi-quantitative analysis of this revealed distinct dependency relationships between attributes which can help create a more representative model than a naïve Bayesian network classifier based on the LSOA. It was shown that quantifiable dependencies exist between the building attributes (Figure 7.6), and that all the building attributes were dependent upon income. Section 7.3 presents specific data at the spatial scale of the LSOA with which these dependencies can be modelled.

7.3 Data Sources

This section presents the acquisition and processing of data for the creation of LSOA building stock datasets. Firstly, the acquisition of building attributes is discussed (section 7.3.1), followed by the

roof attributes (section 7.3.2). The challenge of combining these two data sources and further processing to provide the quantitative data for the BN submodel is described in section 7.3.3. Following this, the acquisition of household income data is presented (section 7.3.4) and its integration with the building stock data to yield a tabular data combining all variables at the LSOA scale (7.3.5).

7.3.1 The Geoinformation Group 'National Building Class'

Building stock data is required for the four case study LSOAs selected in Chapter 4. The building stock in English urban areas with a population above 10,000 has been classified by the Geoinformation Group, a commercial company which specialises in the photointerpretation of high resolution digital aerial photography (GIG, 2013). This process uses trained image interpreters with experience of period building architectures, and supporting evidence such as chimney styles, roof tile types and colours (GIG, 2012). The Geoinformation Group is clear the data is not 100% accurate; buildings are allocated to age groups and types based on best available evidence. Occasionally field visits are made and other supporting evidence such as historical maps are utilised.

Despite its subjective component in building classification, datasets released by Geoinformation have been frequently used by Local Authorities to augment their local Gazetteer housing stock data (Keirstead and Calderon, 2012). Taylor et al. (2013) used the age and type data to create a hygrothermal model for urban areas to simulate the post-flood drying of dwellings. Notably they used a GIS spatial join to link the building attributes to building outlines from mapping datasets. Building age and archetypes were used to create 21 building categories in the calibration of bottom-up building energy models for the aggregated energy demand at the LSOA scale using Bayesian regression (Booth and Choudhary, 2012).

There is a lack of validation for Geoinformation Group data, however, In Newcastle, inconsistencies where found on comparison to the city council's gazetteer based on OS Mastermaps©. "A significant amount of misclassification of buildings" was subsequently improved by the supplier (Calderon et al., 2012). The suppliers acknowledge there is no formal measure of accuracy and suggest that local knowledge and context should be used. In particular, building age is more difficult to judge than built form since architectural designs span a range of age boundaries (GIG, 2012).

Properties are classified into five age bands and 15 different building structural types. An individual dwelling may, therefore, be assigned to one of 57 categories. According to the company literature this classification utilised "*the company's considerable expertise in classifying a whole range of structural, regional and other property characteristics from its imagery archive using photo interpretation skills" (GIG, 2013)*.

The product was purchased for each of the four LSOAs. In addition to building attributes, the geographic co-ordinates, post office address file data for each dwelling and the building footprint (the area of ground covered by the building) were provided. A purchased dataset had the structure shown in Table 7-2.

Table 7-2 Geoinformation group building stock data file columns

Attribute	**Permitted Values**
Latitude	Degrees (°)
Longitude	Degrees (°)
Address	House number and street
Postcode	UK Post office address codes
Age Category	1870-1914; 1914-1945; 1945-1964; 1964-1979; 1979-1999; Recent Unknown
Type	See table 7.3
Area	Continuous (m^2)

A preliminary analysis of this dataset was carried out to prepare frequency distributions for the building area, the age and built form classifications, as subsequently employed in the BN sub-models.

The area of the building footprint for dwellings was presented in the dataset to the nearest square meter. The measurement of the building footprint area was achieved by a GIS analysis of OS mapping data, specifically the MasterMap 1:1000 raster layer which provides building outline geometries. A frequency distribution was created using discrete bins with an interval of 5m² to create a vector with 13 elements, ranging from 40m² and 100 m². Figure 7-10 shows the building footprint area frequency distributions for each LSOA.

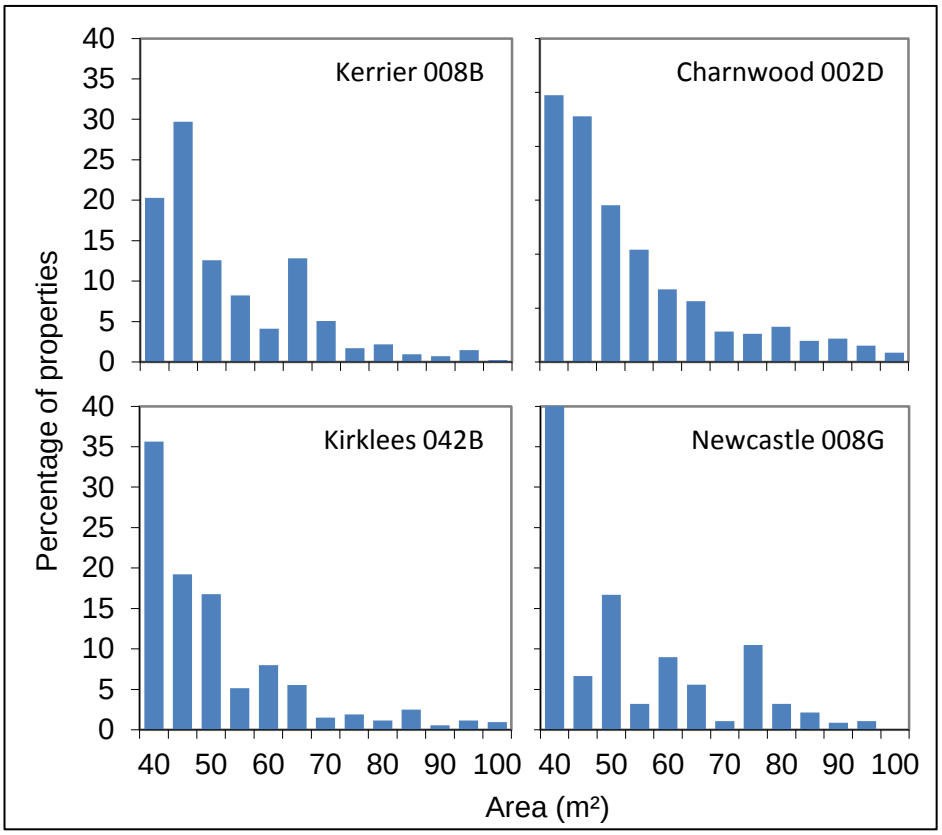


Figure 7-10 Building footprint distribution for each LSOA

The four charts show that for each locality the mode value is 40m² except for LSOA Kerrier 008B where it is 45m^2. The Northern localities (LSOA Kirklees 042B and LSOA Newcastle 008G) have slightly smaller domestic properties with the median value (0.5) of 45 and 44 m^2 respectively with LSOA Charnwood 002D at 55m^2 and LSOA Kerrier 008B 49m^2.

The age bands into which the properties fall for each of the four areas are given in Figure 7-11. LSOA Kerrier 008B is a largely new settlement with 75% of properties built after 1979 whereas LSOA Charnwood 002D consists of largely Edwardian or earlier dwellings building with 50% built before 1914. LSOA Kirklees 042B is a mixed area with three building periods: before the 1914, the war and inter-war years and immediate post-war. LSOA Newcastle 008G, in contrast, consists of largely inter-war dwellings and a small number of late sixties and seventies properties.

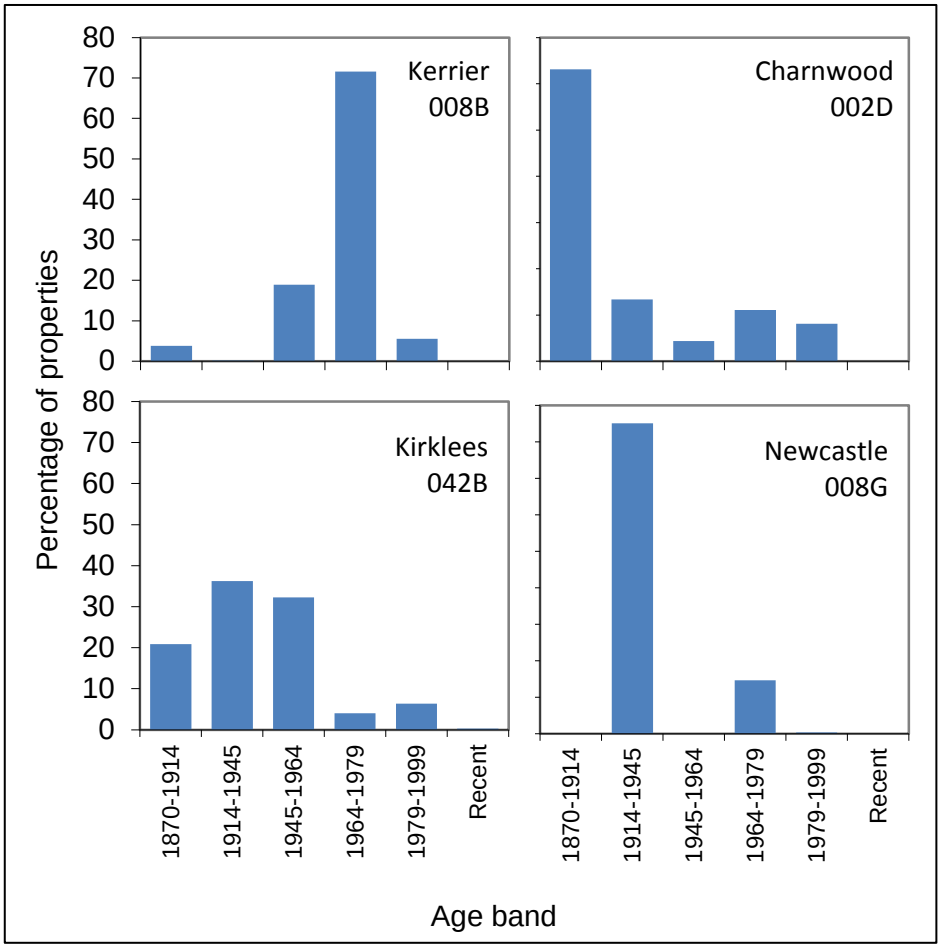


Figure 7-11 Age band distribution for each LSOA

The Geoinformation Group has developed its own proprietary building type classification with 16 archetypes used in their products. The principle types are shown in Table 7-3 along with an

identification code. The frequency distributions for the building type for each of the four LSOAs are shown in Figure 7-12. The frequency of built form within each area reflects the respective age of the dwellings. Thus the majority of inter-war and post war dwellings are of the common British semi-detached residence archetype. This applies to LSOA Kerrier 008B, LSOA Kirklees 042B and LSOA Newcastle 008G. In contrast, LSOA Charnwood 002D, with its large number of pre-1914 dwellings has a large density of late Victorian terraced housing.

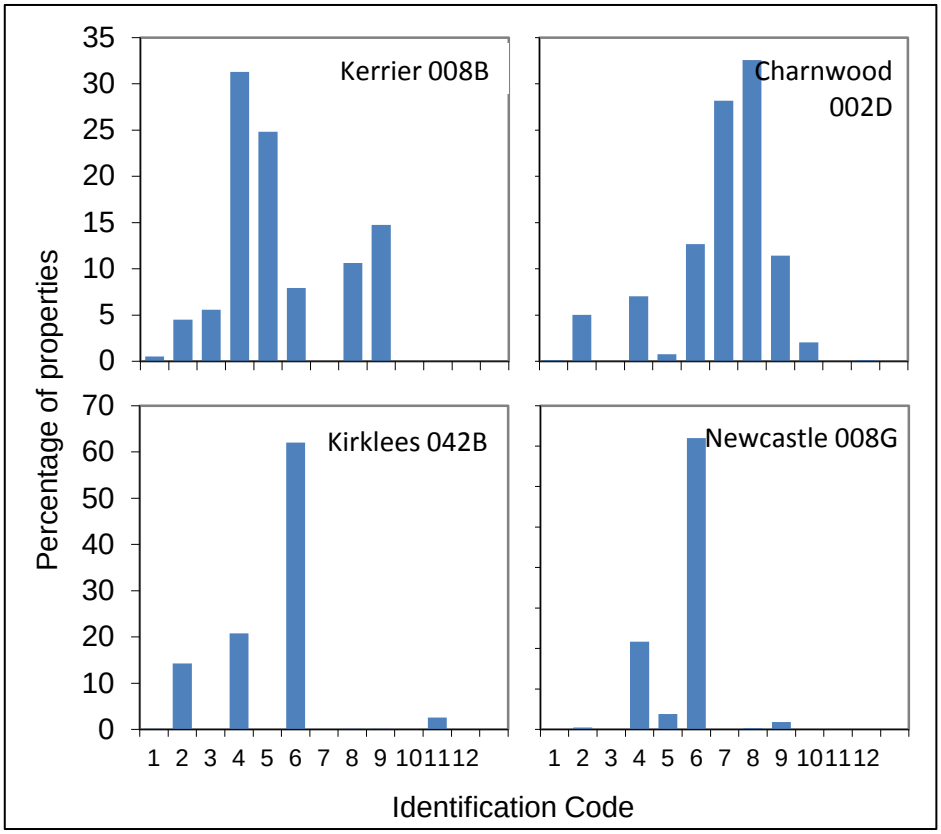


Figure 7-12 Building archetype distribution for each LSOA

Table 7-3 Built form archetypes used in Geoinformation Group products

Code	Archetype
1	Large detached
2	Smaller detached
3	Large property semis
4	Standard size semis
5	Linked and step linked houses 2-3 mixed and 3 stories
6	Semi type house in multiples of 4,6,8 etc.
7	Low terraces 2 stories with large T extension
8	Low terraces, small
9	Lower 3-4 storey Flats
10	Medium Height Flats 5-6 Story flats
11	Probably residential, unknown classification
12	Tall terraces 3-4 stories

7.3.2 Roof Geometry

Lidar data (see Section 7.2.2) has been obtained to establish roof parameters for the building stock sub-model. Processing of aerial scans has been undertaken by the data providers, BlueSky LTD, who used 3-D imaging software to identify roof elements and determine, for each element the geographical co-ordinates, area, pitch, and aspect. A subjective shading attribute was estimated for each element by the operator at the time of analysis from the 3-D imagery (by comparing the height of nearby trees to the house, for example). One of four discrete values was assigned to the attribute for each roof (Table 7-4).

Table 7-4 Shading factor for roofs prepared by BlueSky using lidar data

Shading factor	
0	none (or very little)
1	up to 30%
2	up to 60%
3	very heavy

Roof data for each LSOA was received in the GIS ESRI shape file standard which is compatible with GIS software systems. The files contained polygon data for the ‘most favourable roof’ which was

assigned by BlueSky LTD to be most South facing. Attribute data were stored in an associated database file and contained the fields shown in Table 7-5. In addition to the supplied roof attributes data, BlueSky also supplied a list of roofs which were rejected, since a valid area could not be identified, or was ambiguous. Each LSOA has a known number of domestic properties as described earlier. Some properties, with more complex roof geometries have two or more roofs deemed suitable for solar PV. The number of roofs rejected by BlueSky is also given. Inspection of these in a GIS, overlaid on MasterMap, showed that the reason for rejection was due to the software selecting inappropriate polygons as roofs, for example shed or garage roofs, or ambiguous geometries on the ground. A numerical overview of the supplied data is presented in Table 7-6.

Table 7-5 Lidar dataset attributes

Attribute	**Description**
UPRN	Unique Property Reference Number
USRN	Unique Street Reference Number
BSPG_ID	Blue Sky Assigned unique reference number
GEOX	x co-ordinate (Easting)
GEOY	y co-ordinate (Northing)
ROOFID	Identity number assigned to roof
AREA	Area of roof (m^2)
PITCH	Inclination of roof from horizontal (°)
ASPECT	Orientation ($0\,° \equiv Due\ North$)
SHADE	Shading Factor (0, 1, 2, or 3)
FLAT	Indicates if roof is flat

Table 7-6 Number of properties and roofs in each LSOA in the BlueSky dataset

LSOA	**Number of Properties**	**Number of Roofs**	**Rejected**
LSOA Kerrier 008B	556	451	29
LSOA Charnwood 002D	747	740	330
LSOA Kirklees 042B	774	747	87
LSOA Newcastle 008G	693	660	39

A preliminary analysis follows to explore the distributions for the key parameters used for the calculation of the specific yield of the solar PV systems.

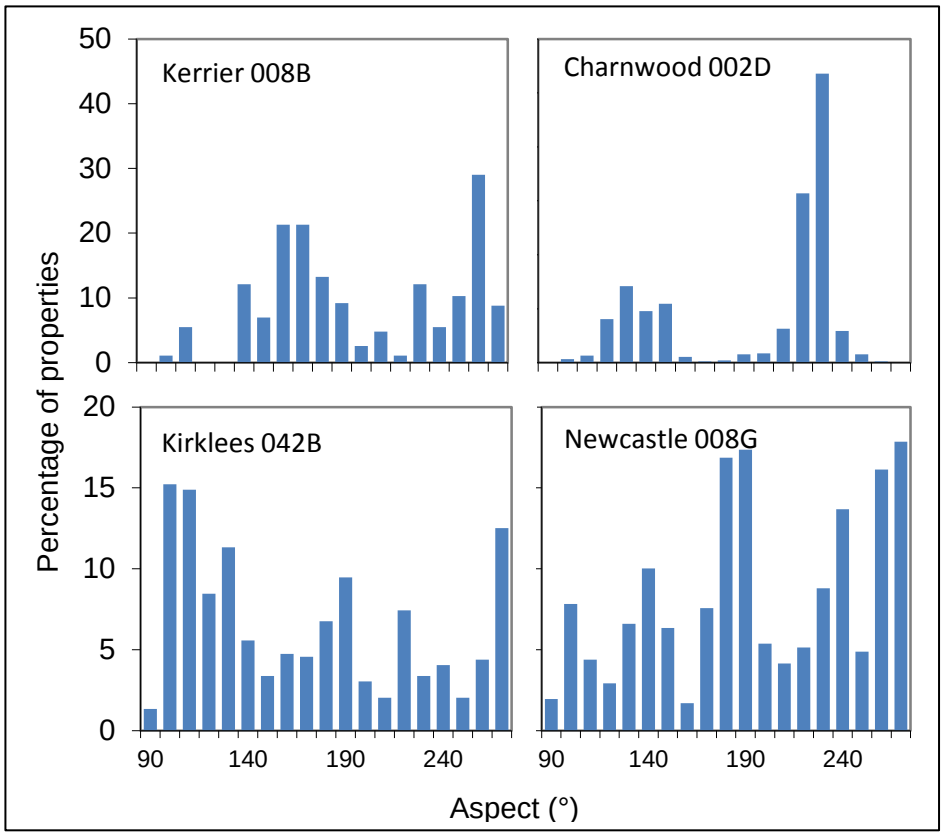


Figure 7-13 Roof aspect distribution for each LSOA

The distribution of the roof aspect for each LSOA is shown Figure 7-13. Properties tend to be orientated along an axis parallel to the road or street on which they are built. Given the relatively small number of streets and properties in the LSOAs, the distribution of orientations is not a uniform one which might be expected in a large random dataset. A small local network of streets will have a greater probability of being parallel to each other, or orthogonally intersecting. This is revealed in the fine structure of the orientation distribution: in LSOA Kerrier 008B a peak in the orientation distribution at 165° is commensurate with a second peak, approximately 90° apart at 260°. In LSOA Charnwood 002D the two orthogonal street peaks are at 135° and 225°. The orthogonal streets

layouts resulting in this structure in the distributions can be observed on the OS maps. In contrast LSOA Kirklees 042B has a more random distribution, reflected by a more variable orientation of streets; LSOA Newcastle 008G again is more variable but the fine structure resulting from orthogonal streets can be discerned at 190° and 270°. Every local LSOA in the country will have a unique signature of building orientations which the probabilistic model needs to account for, since as will be discussed below, the fine structure in the orientation distributions resulting from the street architecture will appear in the distribution of yields. It is useful that the reader note this physical trait at this juncture.

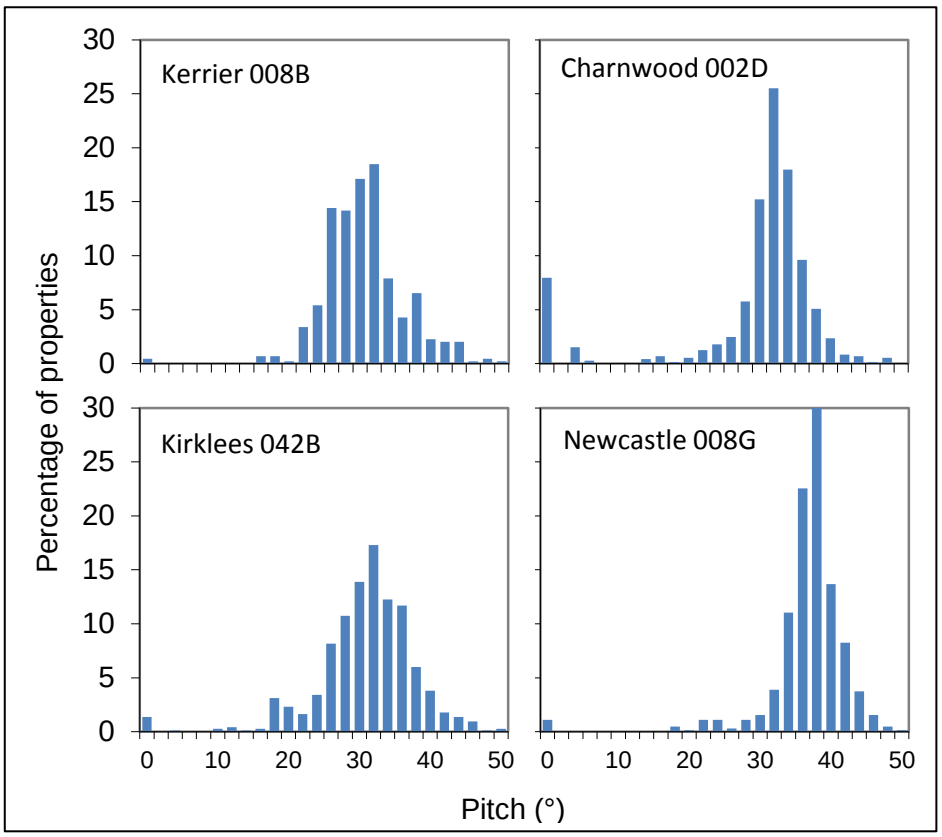


Figure 7-14 Roof pitch distribution for each LSOA

Figure 7-14 show the distribution of roof pitch for each LSOA, obtained from the Lidar data. It is assumed that a hypothetical Solar PV system will be co-planar with the plane of the roof, the usual

practice on inclined roofs. The distribution of roof pitch is noteworthy for its breadth, indicative of a large variety of building styles.

Finally, the third predictor variable for PV yield is the area of the roof most suitable for PV installation. The distribution of roof area for each LSOA is shown in Figure 7-15. To be noted is the degree of variability of roof sizes within the four areas, and the differences between the LSOA areas.

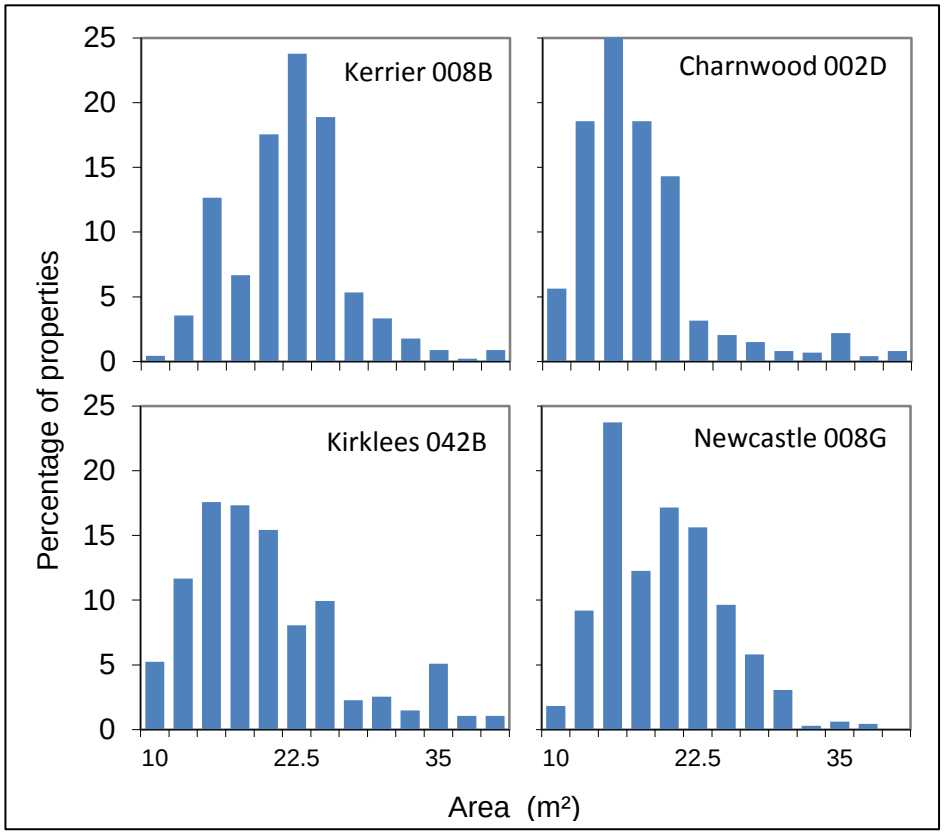


Figure 7-15 Roof area distribution for each LSOA

7.3.3 Combined Building Attribute Dataset

The UPRN number provided with the Lidar dataset enabled the address data for each roof to be accessed from the Postcode Address File (Ordnance Survey, 2014). The Geoinformation group data set was already furnished with the address data. This allowed the pairwise matching of the

properties in the two independent datasets to create a unified dataset for each LSOA. This, however, was a less than perfect matching process. On loading of data into the GIS software it was discovered that both the GeoInformation Group, and BlueSky, had done imprecise matching of building elements to post office addresses. Using an iterative procedure of matching GIS polygon, vector and point data with the two datasets loaded, the matching of roofs to building stock was carefully executed by moving GIS features and executing spatial queries. A large number of buildings did not have a matching roof, or occasionally had more than one. The online aerial photography mapping tool, Google Earth, was employed to visually cross check all the roof-building matches and missing roofs in the GIS system (Figure 7-16).

Figure 7-16 Use of Google Earth™ and QGIS™ to visually cross check roof and building data

Each building was visually inspected using Google Earth and the attribute database was updated to record features which might prevent the installation of a PV system. Typical impediments were dormer windows, skylights and hipped or intersecting roofs presenting small surfaces (Figure 7-17).

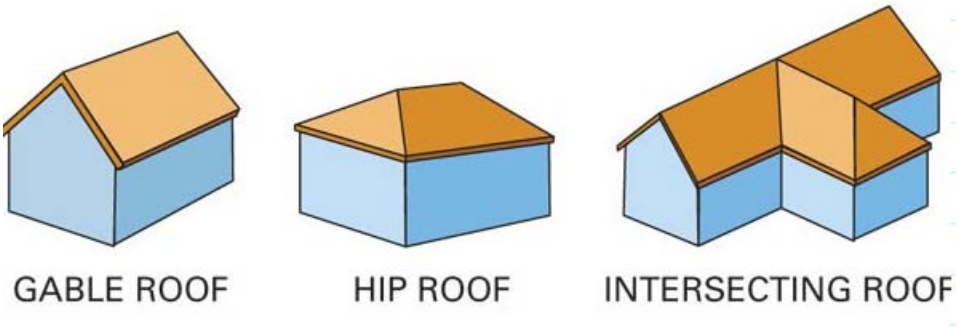


Figure 7-17 Common roof types identified using Google Earth aerial photography

A significant number of property addresses in the GeoInformation Group dataset were found to be allocated to the same building polygon identified in the MasterMap data. The use of QGIS spatial queries allowed each building polygon to be given a sharing factor – the number of addresses allocated to the building. This was over 20 for some large flats. The data did not include the bungalow built form archetype and a missing feature in the provided datasets was the number of floors. Whilst the majority were typical 2 story dwellings, there were a number of bungalows, three or more story properties. Visual inspection using Google Earth in Street View, which allows the user to explore the front (street-facing) elevations of buildings, was used to inspect and record the number of floors in the attribute database.

In addition to a virtual street walk and flyover of the four LSOAs using Google Earth, in two, LSOA Charnwood 002D and LSOA Kerrier 008B, the streets were physically walked. A key objective was to verify the building age categorisation. In LSOA Charnwood 002D, modern post war in-fill properties nestled amidst Victorian dwellings, and old cottages amidst modern buildings in LSOA Kerrier 008B, had all been appropriately aged. A finding of note in LSOA Kerrier 008B was a large number of the properties had been recently retrofitted with external wall insulation. The results of the detailed building stock analysis are shown in Table 7-7.

Table 7-7 Summary of roof assessment for the building stock in each LSOA

Assessment	Kerrier 008B	Kirklees 042B	Charnwood 002D	Newcastle 008G
Suitable	395	517	427	508
Suitable 20% Shaded	33	66	49	34
Suitable 40% Shaded	8	16	19	10
Suitable 60% Shaded	7	2	1	2
Shaded	6	2	1	11
Apartment	78	86	197	106
Dormer	12	1	12	1
Hip Roof	2	0	14	3
Intersecting Roofs	2	72	16	15
Missing Roof	0	11	7	2
North Facing	4	1	4	1
Skylight	9	0	0	0
TOTAL	**556**	**774**	**747**	**693**

The shading assessments have been provided by Blue Sky operators and the remaining roof issues using the software tools outline above. The final result shows significant numbers of dwellings which either do not have a roof suitable for a PV system due to structural constraints, or, they are apartments or flats that do not have roof elevation. Table 7-8 presents the same data summarised into four categories – suitable, affected by shading, apartment dwelling (has no dedicated roof) and structural constraints. Overall, only two thirds of dwellings have a suitable roof. LSOA Charnwood 002D, with a proportionately higher number of flats and apartments has only 57% of dwellings with a suitable roof. LSOA Kirklees 042B has the highest percentage of structural constraints at 11%. This is due to a common post-war architectural design featuring intersecting roofs which presents many small faces unsuitable for PV modules.

Table 7-8 Summary of broad category roof assessments in each LSOA (% suitable)

Assessment	Kerrier 008B	Kirklees 042B	Charnwood 002D	Newcastle 008G	TOTAL
Suitable	71.0	66.8	57.2	73.3	66.7
Affected by shading	9.7	11.1	9.4	8.2	9.6
Apartment (No roof)	14.0	11.1	26.4	15.3	16.9
Structural Constraints	5.2	11.0	7.1	3.2	6.8

The percentage of dwellings with suitable roofs will influence the potential socio-economic impact of PV in these areas. A further key factor in the assessment of this impact is the actual level of income in the various dwelling types. This is considered in the next section.

7.3.4 Household Income

Using the methods discussed in section 7.2.3 gross income distributions for the four case study LSOAs were generated for the year 2009/10 by Anderson using his published method (Anderson 2013). The probability density distributions (Figure 7-18) for each LSOA, alongside the distribution for the whole of England and Wales, show that, despite the normative assumption that LSOAs are designed to be relatively socio-economically homogeneous, the variability of household income is large. The coefficient of variation for each area is above 50%, and as high as 68% in LSOA Kirklees 042B. These statistics are given in Table 7-9, along with the expected value for the area. LSOA Newcastle 008G has the lowest mean income of £20,870 per annum and LSOA Kirklees 042B the highest at £29,680. For comparison, Table 7-9 shows the rank of the income score used in the index of multiple deprivation. Whilst LSOA Newcastle 008G is clearly identified as a low income, LSOA Charnwood 002D has the third highest mean income but has the highest ranking income score. However, the expected values are very sensitive to small changes in the probabilities of the higher income intervals which could mean that the distributions inaccurately represent higher than average incomes.

These distributions provide estimations of the marginal distributions of gross income for the LSOA. There is no information regarding the dependencies on the housing stock data discussed in the previous section (7.3.3). In the next section it is shown how a joint probability distribution for housing stock parameters and household income can be simulated.

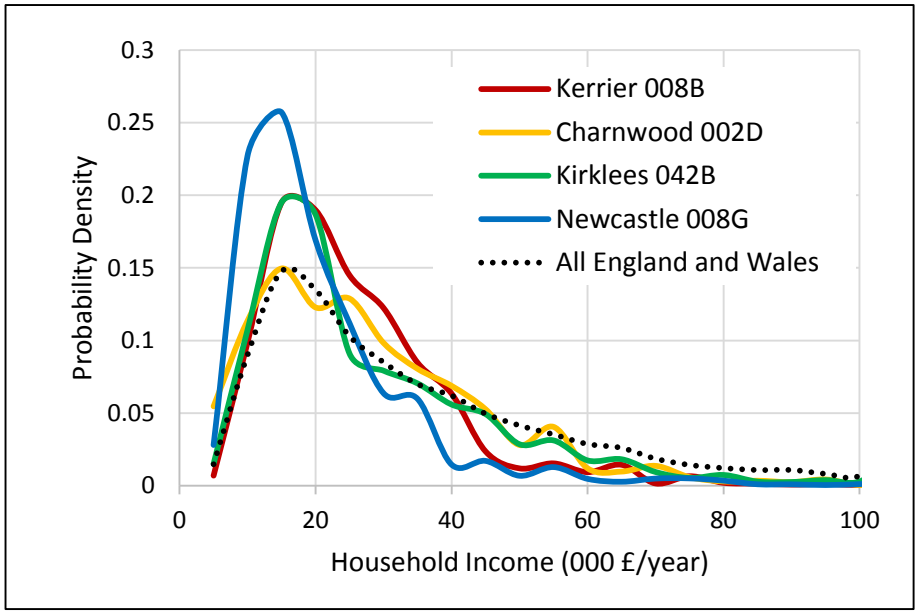


Figure 7-18 Probability distribution of household income for each LSOA

Table 7-9 Expect value (mean), standard deviation and coefficient of variation (CV) of annual household income for each LSOA

LSOA	Income Expected Value (£)	Standard Deviation (£)	CV (%)	Rank Income Score
Newcastle 008G	20870	13515	65	19
Kerrier 008B	26095	14420	55	336
Charnwood 002D	28185	17507	62	9184
Kirklees 042B	29680	20289	68	1198

7.3.5 Simulating a Joint Probability Distribution for Housing Stock Including Income

In the preceding two sections, for each LSOA, a dataset of building attributes has been created to yield a joint probability distribution of the salient building attributes (Equation 7-6), and a marginal distribution of household income has been sourced and assessed (Equation 7-7).

$$\boldsymbol{P}(\boldsymbol{BF}, \boldsymbol{BA}, \boldsymbol{FA}, \boldsymbol{RA}, \boldsymbol{O}, \boldsymbol{S}) \qquad \textit{Equation 7-6}$$

$$\boldsymbol{P}(\boldsymbol{HI}) \qquad \text{Equation 7-7}$$

If the not unreasonable assumption is made, that income is independent of the building orientation and roof attributes[32], then, as suggested in Section 7.2.4, only the dependencies on core building attributes (built form, building area and floor area) are required. Thus the desired JPD for each LSOA is as shown in Equation 7-8.

$$\boldsymbol{P(BF, BA, FA, HI)} \qquad \text{Equation 7-8}$$

In this section it is shown how iterative proportional fitting (IPF) can be used to simulate such a JPD. Firstly, in the remainder of this section, the theory of IPF is outlined, followed by its practical application to simulate a JPD as in Equation 7-8.

Theory of IPF

A common problem in small-area data analysis is a lack of contingency tables or JPDs for small geographic areas, whereas for large areas, microdata are often available, from which one can construct contingency tables or JPDs (Kurban et al., 2011). The iterative proportional fitting (IPF) procedure was proposed as a solution to adjust the values in a target contingency table, when the expected marginals are known, based on knowledge derived from a reference contingency table (Deming and Stephan, 1940). IPF has been extensively adopted to solve the problem of the lack of microdata for small area geographies by simulating contingency tables using less spatially specific reference datasets, a technique known as microsimulation (Ballas et al, 2013). A thorough treatment of the method and its uses has been detailed in a widely cited working paper by Norman (1999).

Crucially the "interaction structure" of the reference contingency table is preserved in the adjustments made to the target values (Mosteller, 1969), which is tantamount to the preservation of

[32] This is assuming property purchase decisions are not generally influenced by roof orientation or roof slopes, but rather more likely determined by their locality.

the dependencies between the parameters in the contingency table. Fienberg (1970) has shown that the adjustments show a convergence to final values.

Using IPF a marginal distribution of a target parameter can be used to simulate a multidimensional contingency table if the contingency table of all the remaining parameters, referred to as constraint parameters, is known. The fitting is executed such that the proportion of the fitted parameter to the constraint parameters matches the proportions in a reference dataset which contains those same parameters. Table 7-10 shows the required data sources and the resultant target for three hypothetical variables A, B and C. Variable C is to be fitted to a table with a known contingency table containing parameters B and C (the constraints), to produce a simulated target table with all three parameters A, B and C. A reference table with these same parameters is available to provide information on the proportions for C in the new target table.

Table 7-10 Components for performing iterative proportional fitting

Joint or Marginal Probability Distribution	**Description**
$\mathrm{P}_{reference}(\mathrm{A}, \mathrm{B}, \mathrm{C})$	Reference table which has information about three variables A, B and C
$\mathrm{P}_{target}\ (\mathrm{A}, \mathrm{B})$	Constraints table with known information about A and B in the target dataset
$\mathrm{P}_{target}\ (\mathrm{C})$	Marginal distribution of C for the target dataset
$\mathrm{P}_{target}(\mathrm{A}, \mathrm{B}, \mathrm{C})$	Simulated target table

As well as a requirement that the reference dataset contains the same parameters as the target, it is important that these parameters are effective predictors of the variable which is to be fitted (Anderson, 2013). It is also suggested that the best fitting is achieved if the reference data is as spatially proximate, or characteristically similar, to the constraint dataset as possible.

Executing IPF

It was demonstrated (Section 7.2.4) that the EHS dataset shows dependencies between income and building attributes and so can serve as a suitable reference dataset for performing IPF. However, as suggested above, the best fitting is achieved if the reference data is as spatially proximate, or characteristically similar, to the constraint dataset as possible. For this reason, London and the Southeast were excluded when constructing the reference dataset since these areas have a larger proportion of flats and maisonettes, and generally higher income households.

Table 7-11 Components for performing iterative proportional fitting to simulate an LSOA level building stock dataset with integrated household income

IPF table	**Joint or Marginal Probability Distribution**	**Description**
Reference	$\mathrm{P}_{EHS}\ (\mathrm{HI}, \mathrm{BA}, \mathrm{BF}, \mathrm{FA})$	English housing survey reference table
Constraint	$\mathrm{P}_{LSOA}\ (\mathrm{BA}, \mathrm{BF}, \mathrm{FA})$	Constraint building stock table for the LSOA
Marginal	$\mathrm{P}_{LSOA}\ (\mathrm{HI})$	Marginal household income table for the LSOA
Target	$\mathrm{P}_{LSOA}\ (\mathrm{HI}, \mathrm{BA}, \mathrm{BF}, \mathrm{FA})$	Target simulated table for the LSOA
Key: HI: household income; BA: building age; BF: built form; FA: floor area		

Table 7-11 summarises the three required datasets and the target simulated dataset created using IPF. Floor area, built form and building age were used as constraints - evidence above suggests that floor area is the most influential. Unfortunately the constraint data and the reference data each use quite different built form and building age classifications. This was solved by mapping their respective classifications to those used in the NEED framework, since this is later required when interfacing the building stock model with the building energy consumption model. Table 7-12 and Table 7-13 show how this was done for the building age parameters which are mapped

probabilistically in proportion to the number of years which overlap with each category[33]. Consider, for example, the EHS building age category '1945-1964'. This is mapped to both categories '1930-1949' and '1950-1966' in the NEED framework, in the proportion of 4:14.

Table 7-12 Mapping building age in the EHS to the NEED parameter

	Before 1930	**1930-1949**	**1950-1966**	**1967-1982**	**1983-1995**	**1996 onwards**
EHS	NEED Framework					
pre 1850	1.00	0.00	0.00	0.00	0.00	0.00
1850-1899	1.00	0.00	0.00	0.00	0.00	0.00
1900-1918	1.00	0.00	0.00	0.00	0.00	0.00
1919-1944	0.44	0.56	0.00	0.00	0.00	0.00
1945-1964	0.00	0.21	0.79	0.00	0.00	0.00
1965-1974	0.00	0.00	0.11	0.89	0.00	0.00
1975-1980	0.00	0.00	0.00	1.00	0.00	0.00
1981-1990	0.00	0.00	0.00	0.11	0.89	0.00
post 1990	0.00	0.00	0.00	0.00	0.33	0.67

Table 7-13 Mapping building age in the Geoinformation Group dataset to the NEED parameter

	before 1930	**1930-1949**	**1950-1966**	**1967-1982**	**1983-1995**	**1996 onwards**
Geoinformation	NEED Framework					
1870-1914	1.00	0.00	0.00	0.00	0.00	0.00
1914-1945	0.52	0.48	0.00	0.00	0.00	0.00
1945-1964	0.00	0.22	0.78	0.00	0.00	0.00
1964-1979	0.00	0.00	0.13	0.87	0.00	0.00
1979-1999	0.00	0.00	0.00	0.15	0.65	0.20
Recent	0.00	0.00	0.00	0.00	0.00	1.00

The built form parameters where simply mapped to the most appropriate built form in the NEED framework. Since build form parameter in the EHS dataset did not distinguish between mid-terraced

[33] This method was suggested by Delcroix (2013).

and end-terraced building types, this were mapped in a ratio 2:1 respectively which is the ratio found in the NEED framework.

In this fashion, both the EHS reference dataset and the LSOA building stock dataset had their built form and building age parameters converted to be compatible with the NEED framework. Using these converted sources IPF was executed using the *mipfp* software package written in the R software programming language (Barthélemy et al, 2015). This is an implementation of several methods for updating an initial N-dimensional array with respect to given marginal distributions, which may also be multi-dimensional.

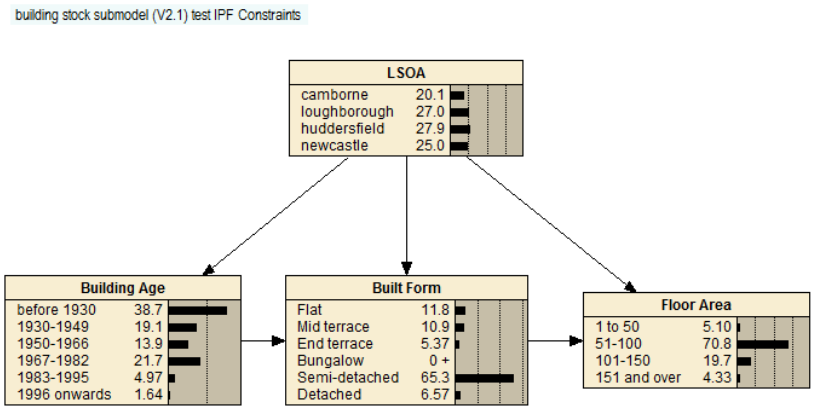


Figure 7-19 Bayesian Network in Netica constructed using the reference dataset

The IPF procedure was conducted for each of the four LSOAs; the resultant target JPDs were verified to ensure that the marginal distributions of each variable matched the source distributions; this ensured that nothing untoward had occurred with the software[34]. This was achieved by importing the data into a purposefully created BN. Figure 7-19 shows the network constructed using the constraint data, $\mathrm{P}_{LSOA}\,(\mathrm{BA}, \mathrm{BF}, \mathrm{FA})$, and Figure 7-20 the network constructed using the target data,

[34] The mentioned software, R-mipfp, is open source and whilst the R-project is well supported by the academic community, it routines are not always independently verified.

P_{LSOA} (BA, BF, FA, HI). Comparison of these two networks shows that the prior distributions for the building attributes were not altered by the IPF procedure.

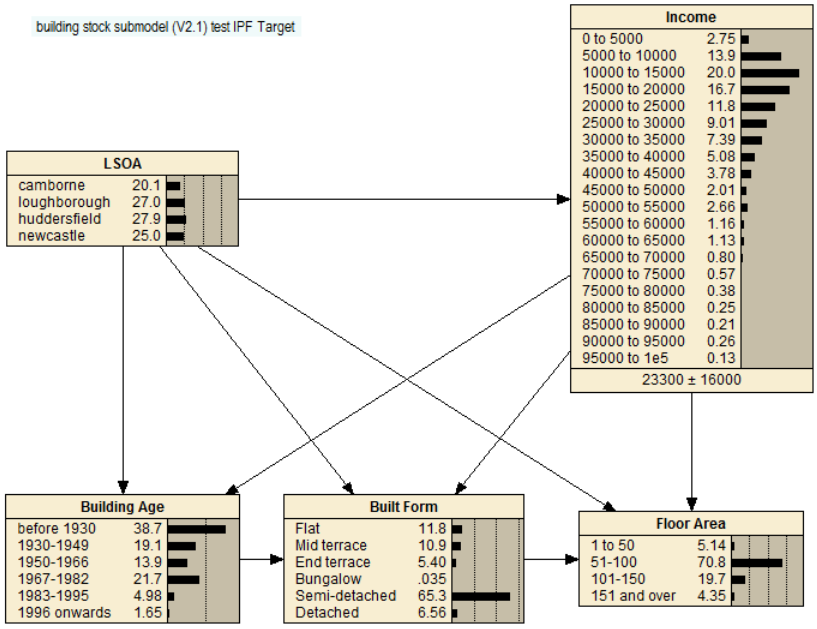


Figure 7-20 Bayesian Network in Netica constructed using the dataset from the IPF procedure.

A second verification step undertaken was to check that the dependencies observed in the resulting dataset, particularly those pertaining to income and building attributes, bore some resemblance to the observations already made with the EHS dataset. In Figure 7-21 a high income category was selected to compare with a low income category in Figure 7-22. The results show that low income households are more likely to occupy flats and smaller dwellings than those on high incomes who are more likely to occupy larger dwellings. This is summarised in Table 7-14.

Table 7-14 Comparing percentage of building attributes for low and high income households

Income (£/year)	Detached	Semi-detached	Flat	Floor Area	
				1-50	151 and over
0-5000	9.75	32.7	46.1	55.1	9.77
50-55000	4.64	53.3	0	0	16.4

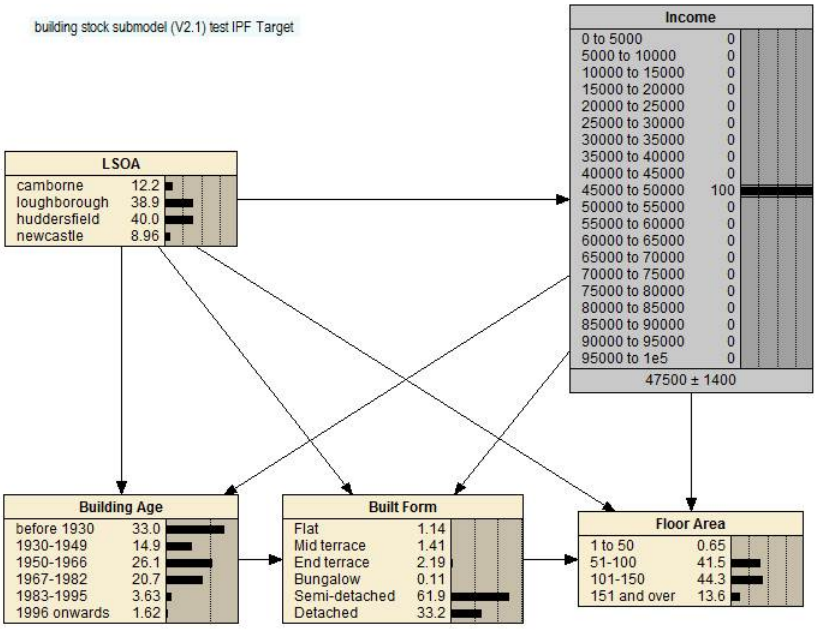


Figure 7-21 Posterior distributions for building attributes after selecting a high income category.

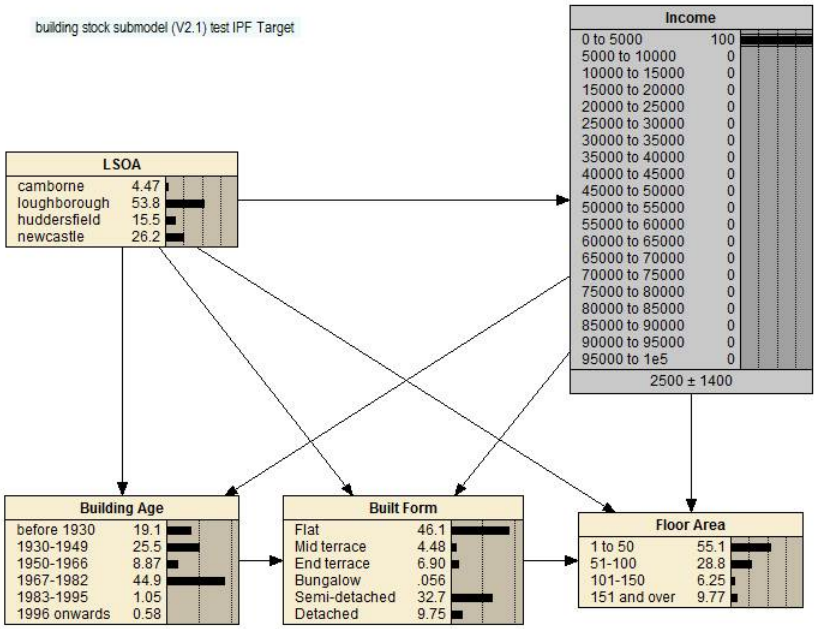


Figure 7-22 Posterior distributions for building attributes after selecting a low income category.

The IPF procedure has delivered an integrated income parameter which shows dependencies on building attributes which are consistent with those observed in the EHS. The resultant dataset, albeit

simulated, contains dependences between empirical income data and the building stock attributes in each of four LSOAs. Such detailed spatially resolved microdata would be difficult and expensive to garner through empirical means.

This concludes the discussion of data sources for the building stock model. The next section presents the construction of a BN consistent with the literature studies in Section 7.2 and the data presented in Section 7.3.

7.4 Bayesian Network Submodel for the Building Stock

7.4.1 The Directed Acyclic Graph (DAG)

The structure of the DAG, which captures the dependencies discussed above on the domain ontology and the empirical data sources, is shown in Figure 7-23. The DAG is essentially a naïve classifier network with the LSOA of interest predicting building, roof and the household income attributes. In this way the LSOA serves as a method of selecting the appropriate marginal distributions for all these parameters.

However, additional dependencies between building attributes, as discussed in Sections 7.2.1 and 7.24, have been incorporated into the DAG structure. Thus the built form predicts the floor area and building age. The roof area has been assumed to be dependent on floor area and built form. Whilst the former dependency is manifest, the dependency on built form allows for the possibility that the type of building influences the roof area too. This also holds for the roof pitch which is also influenced by the LSOA – steeper roofs were observed, for example, in LSOA Newcastle 008G.

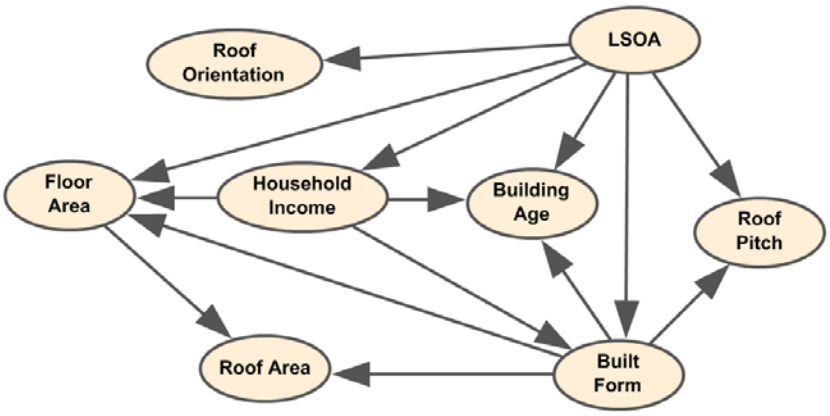


Figure 7-23 DAG for the LSOA building stock model

Roof orientation is also dependent on the LSOA since this is determined by the orientation of the streets in the given area. However it is apparent that this is dependent on the building orientation which in turn is only dependent on the street layout.

With the accomplishment of the IPF procedure carried out in preceding section, the relationships between household income and key building attributes have also been integrated. The EHS showed that income influences the floor area and built form, but is less predictive of building age. For completeness, however, all three building attributes where made dependent on the income parameter.

7.4.2 Node Probability Tables (NPTs)

Table 7-15 summarises all the NPTs required by the model alongside the data source used to furnish each one. The counting method was used for learning the NPTs. The model is designed to furnish the building energy consumption and the PV yield sub-models with inputs. Therefore, the discretisation of the continuous variables used as interfaces to these models was carefully matched to them. In particular, the categories used for the building attributes derived from the Geoinformation data source were converted to match the NEED attributes for the purposes of

executing IPF (Section 7.3.5), as well as maintaining compatibility with the building energy consumption submodel (Chapter 6).

Table 7-15 Summary of approach learning NPTs for the building stock mode

NPT	Data Source	Discretisation	Units
P(LSOA)	1	n/a	n/a
P(Orientation|LSOA)	1	10	Degrees
P(Pitch|LSOA,BF)	1	5	Degrees
P(RA|FA^{GI},BF)	1	5	m^2
P(BA|LSOA,HI,BF)	2	As NEED*	n/a
P(FA|LSOA,HI,BF)	2	As NEED*	n/a
P(BF|LSOA,BF)	2	As NEED*	n/a
P(HI|LSOA	3	5000	£
Notes 1. Combined building attribute dataset (Section 7.3.3) 2. Simulated target dataset using IPF (Section 7.3.5) 3. Household income (Section 7.3.4) * See categories above for floor area, built form and building age			

7.4.3 Netica Building Stock Sub-model

The resultant BN sub-model in Netica is shown in Figure 7-24. This shows, in comparison to the theoretical DAG in Figure 7-23 two nodes for the floor area; in addition to the broad interval discretisation used in the NEED framework, a floor area node to represent the continuous variable was retained, discretised in intervals of 10 m^2. This allows the model to more accurately reflect the roof areas empirically measured for each LSOA using lidar data. This will permit a granular calculation of system ratings in the PV yield sub-model; the broad floor area intervals used in the NEED framework loses this granularity and distorts the range of system ratings.

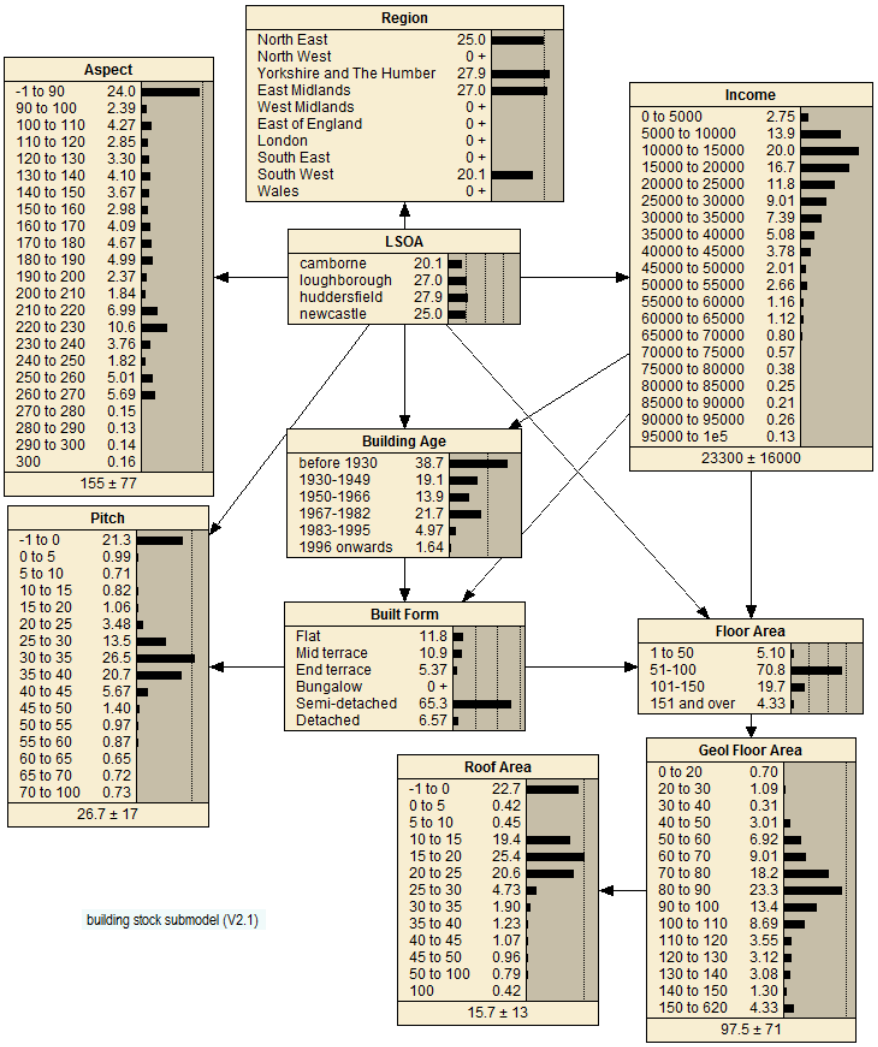


Figure 7-24 Building stock sub-model in Netica

7.5 Discussion and Conclusion

In this chapter, empirical data for building and roof attributes have been combined using GIS tools for four LSOAs. Socio-economic parameters in the form of household income distributions have been derived for these same areas using IPF and census data constraints. The IPF technique has been employed to fit this household income data to the building stock data using building attributes as constraints and the EHS as a reference dataset.

Using the study of publically available datasets and the acquired data sources, a BN sub-model was created which models probabilistic relationships between building stock attributes for each the four lower super output areas. This was designed to interface with the BN sub-models for PV yield and building energy consumption discussed in Chapters 5 and 6 respectively.

The model has been verified to ensure that the marginal distributions concur with the data sources used to furnish the NPTs with quantitative data. Verification also extended to observing posterior distributions of parameters on the selection of hard evidence. For example, it was verified that the variability of floor area for each built form selected in turn as hard evidence, matched those in the source distributions.

The integration of household income into the model was achieved using IPF simulation; this in effect presents the most likely distribution of income given the available building attributes, assuming the LSOA level dependencies match those in EHS reference source. The result is not easy to verify but Table 7-14 shows expected trends of higher income households being more probably allocated to larger and/or detached properties, and lower income households allocated to smaller flats.

The income floor area relationship can be studied using the BN sub-model. Figure 7-25 shows the result of varying hard evidence of the 'Income' node and observing the expected value (mean) of the floor area node. The scatter of the results shows that the IPF fitting is beset with some random noise due, probably, to small sample size for each income interval in the reference dataset. However, the trend, as indicated by the 2nd order polynomial fit is definitively one of increasing floor area as household income increases. This result is gratifying, not least because it concurs with observations made by other researchers (Kelly, 2011), and the analysis of the EHS (Figure 7-7), but also because floor area is a strong predictor of building energy consumption. So, whilst the direct influence of income on energy consumption could not be modelled in Chapter 6 due to the inadequacy of available data sources, at least the strong indirect influence on energy consumption is realised by the

influence of income on floor area in the building stock model. This achievement allows, in the final integrated OOBN discussed in Chapter 9, the inclusion of a socio-economic dimension.

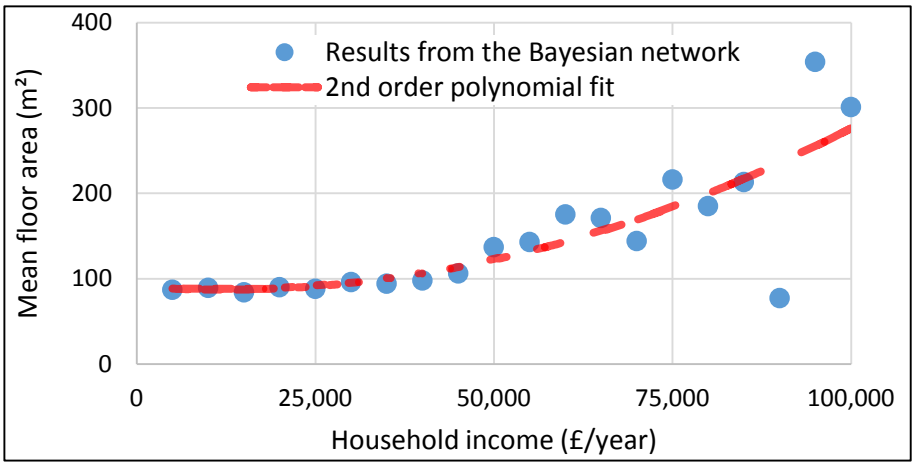


Figure 7-25 Mean floor area as a function of household income obtained from the building stock sub-model.

At the outset of this chapter it was stated that no new parameters were currently sought for the building stock model, the aim was to keep the model parsimonious and restrict its purpose to the interfacing with the building energy consumption and PV yield sub-models. This introduces some weaknesses. For example, building tenure is known to influence energy consumption. Marginal distributions of tenure could not be obtained for the LSOAs; if it had, it could have been fitted in a similar way to household income, using the EHS as a reference. A further simplification was made with regard to building orientation; this was not made dependent on the building type. Thus, any building in the data set is deemed to have the same orientation probability distribution which renders the model not a 100% authentic representation of the LSOA.

In conclusion, however, the building stock model provides the necessary inputs for the building energy consumption and PV yield sub-models discussed in Chapters 5 and 6. These two components provide inputs in to the self-consumption model which is the subject of the next chapter.

8 Self-Consumption of Domestic Solar PV Generated Electricity

8.1 Introduction

The previous chapter has presented a model where a local building stock model provides marginal distributions as inputs to the solar PV yield sub-model (chapter 5) which predicts electricity generation (energy yield), and the building energy sub-model (chapter 6), which predicts consumption (energy demand). This chapter makes the all-important link between generation and consumption with an examination of self-consumption and how this can be incorporated in to a probabilistic model.

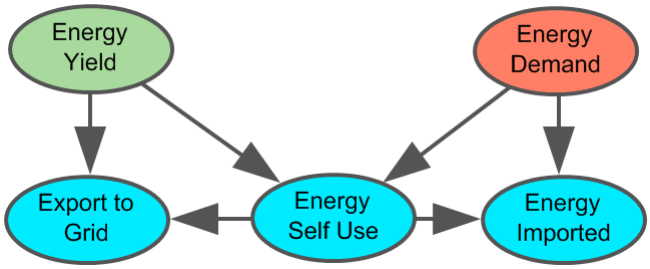


Figure 8-1 Energy self-consumption predicted by demand and yield

The relationship between the energy yield and energy demand has already been alluded to in Chapter 4; Figure 8-1 shows this portion of the conceptual model abstracted into a submodel where both yield and demand are depicted as predictors of energy self-consumption.

The objective in this chapter is to determine a representative the CPT which will model the conditional probabilities to allow the prediction of self-consumption (SU) given demand, D, and yield, Y : $P\ (SU|D,Y)$.

In the next section (8.2) the concept of self-consumption is defined and its required high resolution time-frame discussed. Section 8.3 presents analysis of empirical data from the UK Solar PV domestic field trials. These are shown not to deliver an adequate CPT for the model and therefore section 8.4 presents the solution to this using a simulation of annual self-consumption for UK dwellings. The results of this are presented as a BN model in Netica (Section 8.5) followed by a discussion and conclusion in Section 8.6.

8.2 The Self-consumption Factor

The self-consumption factor (SUF) means the fraction (or percentage) of energy generated which is used to do meaningful on a site work, rather than be dumped or exported (Cao and Sirén, 2014). Self-consumption is favourable to the household economy since solar PV generated electricity is produced at zero marginal cost[35] and avoids the cost of imported electricity supplied at domestic tariffs.

Under the current UK FiT subsidy regime there is an extra payment to the PV system owner for electricity exported to the grid (the export tariff) but this is significantly less than avoided costs. This incentivises self-consumption rather than export. This is a deliberate act of policy in order to reduce potential grid impacts of wide-scale PV penetration, a lesson learnt from the German FiT experience. The economic incentive for self-consumption is further peculiarly reinforced by the fact that the grid exported electricity is generally not metered for domestic PV systems. Instead 50% of the total generation is deemed to have been exported and attracts the export tariff.

[35] A marginal cost is the increase in costs due to production of a good. Solar and wind energy are unique in that the energy resource is free and therefore generation of unit energy does not incur any extra costs.

Subsidies aside, there are householders who are not the owners of the PV system on their home's roof. These may be tenants, and homeowners who have had Solar PV installed under a so-called rent-a-roof scheme. Such users only benefit financially due to the avoided grid electricity costs.

It is important therefore to get a clear estimate of the magnitude of self-consumption since this has a significant impact on the economics of PV under both with and without subsidy. The latter is pertinent to post-subsidy scenarios. Knowledge of self-consumption also has relevance on potential requirements to mitigate low-voltage grid impacts resulting from widespread penetration of solar PV (van der Welle and de Joode, 2011). Theoretical and empirical evaluations of the impact of load-shifting (McKenna, 2013), electric vehicle charging and electrical and thermal energy storage are examples of potential mitigating technologies and behavioural change.

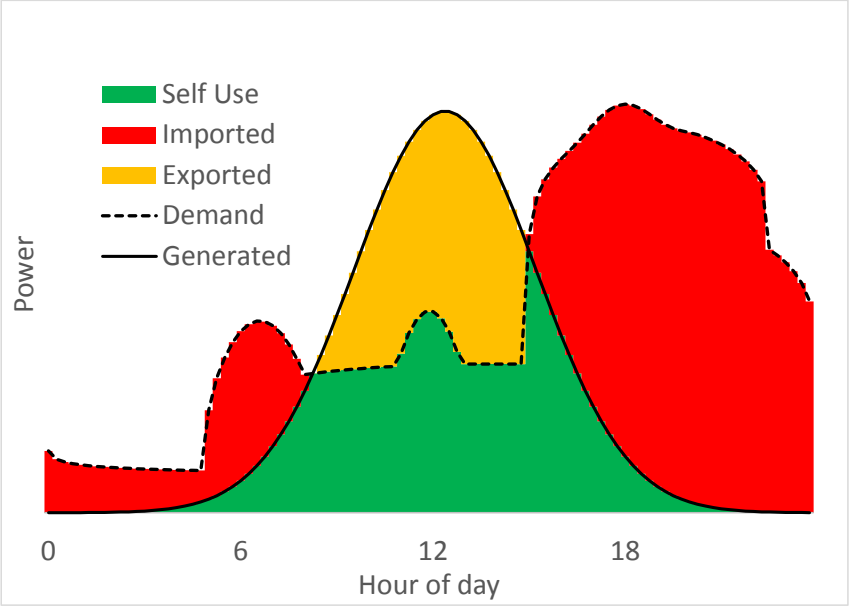


Figure 8-2 Idealised demand and generation profiles demonstrating self-consumption, export of excess generation and import

Self-consumption occurs when generation temporally matches or exceeds the load (the demand on a building's electricity supply due to the use of appliances). For this reason it is sometimes known as

the load match index (Voss et al, 2010) or cover factor. The temporal load profile is dependent on the unpredictable use of appliances, lighting, cooling and heating within the building (Richardson et al., 2009; Richardson et al., 2010). Similarly, the temporal profile of solar PV generation is subject to the unpredictability of the weather as discussed in Chapter 6. This is illustrated in Figure 8-2 which shows an idealised demand profile (black dashed line) and generation profile over 24 hours.

It is intuitive that self-consumption will increase as both generation and demand increase since there will be more overlap between the two profiles[36]. This will be empirically confirmed below. In practice, even with very high demand, 100% self-consumption is rarely attained. This is due to rapid fluctuations in both electricity demand and generation at the domestic level (Richardson and Thompson, 2012). This is exemplified by Figure 8-3 which shows representative 1-minute resolution generation and demand profiles simulated by Richardson and Thompson's model (opt. cit.). The 'spikey' behaviour of both generation, caused by cloud transients, and demand, caused by short bursts of high load, such as might be caused by an electric shower, result in lower than expected match between generation and demand due to the less probable temporal coincidence of narrow sharp peeks compared to broader flatter profiles. This is demonstrated in Figure 8-4, which shows the same data as that in Figure 8-3, aggregated over 1 hour time periods. The sharp spikes on both the generation and the demand are smoothed out which results in an apparent greater overlap between the load and generation profiles. Comparing simulations with 1 hour and 5 minute temporal resolutions has been shown to give errors as large as 80% (Cao and Sirén, 2014).

[36] Consider the increase in area under either demand, generation or both temporal profiles, then the green area representing self-consumption will increase.

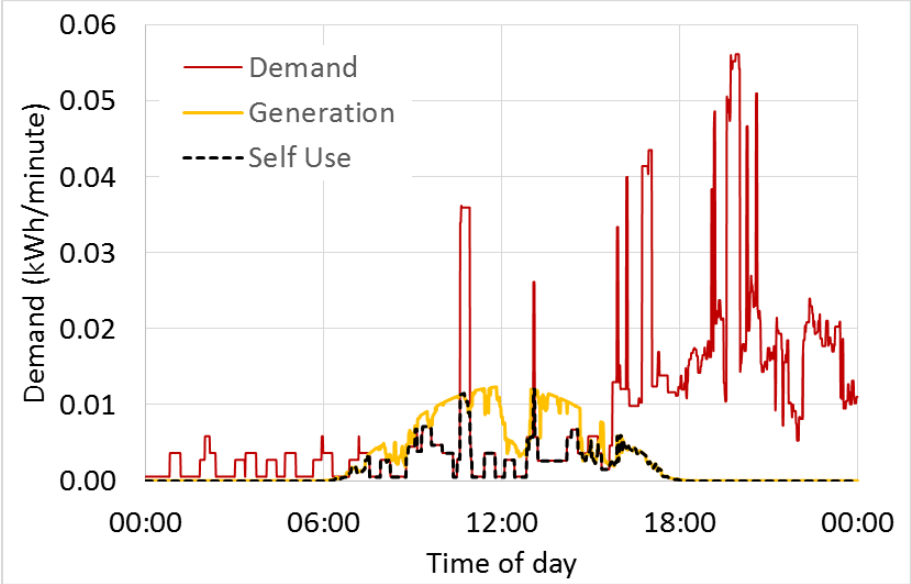


Figure 8-3 Domestic electricity demand and PV generation profile at 1 minute resolution

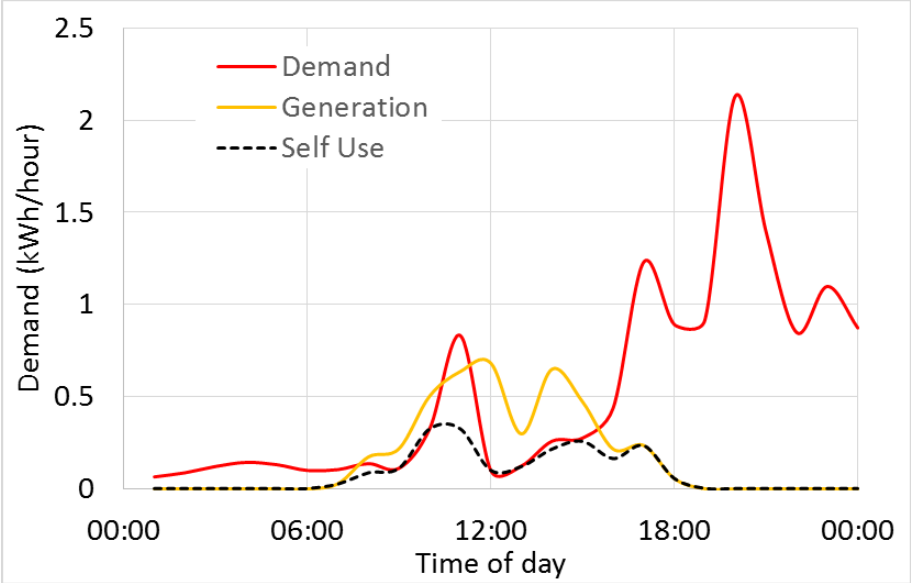


Figure 8-4 Domestic electricity demand and PV generation profile at 1 hour resolution

The over estimation of overlap between demand and generation profiles exhibited by data with a course temporal resolution is well known, and, for this reason one-minute resolution data was used

by Thompson and Murray (2012). In summary there are several distinct features to consider when constructing a probabilistic model of self-consumption:

- The overlap between load and generation is determined by the magnitude of both the energy generation and the energy demand. The greater the magnitude of either, the greater is the probability of overlap and therefore the greater the SUF.
- The stochastic nature of demand and yield suggests that the relationship is not an easily modelled deterministic one. The problem lends itself to probabilistic modelling using a BN.
- The temporal frame for the stochastic events occurs on the minute time-frame but the socio-economic impacts are required over a much larger time-frame, typically one year.

The challenges of garnering data for this model are considerable. Firstly the range of annual electricity consumption needs to match empirical data as shown by the NEED framework data (Chapter 6). And for each annual demand a range of solar PV generation needs to be sampled. The next section evaluates the insights yielded by the UK's photovoltaic domestic field trials (PDFT).

8.3 Self-consumption and the UK Photovoltaic Domestic Field Trials

Domestic field trials for solar PV technology were undertaken comprising over 300 domestic PV installations on 17 separate sites in the UK (Munzinger et al, 2006). Amongst environmental variables such as ambient temperature and irradiance the study collected at 5 minute resolution the AC output of the PV system, the electricity imported from the grid and electricity exported to the grid. For this study, a previously cleaned dataset was used from which erroneous data had been removed and data from several sites were not used at all due to the malfunction of sensors (McKenna, 2013, p 134). This yielded 135 systems with 23 months of 5 minute data. The ratings of systems included in the analysis are shown in Table 8-1. These system ratings are considerably lower than the systems studied in the Sheffield dataset (Chapter 5).

Table 8-1 Rating of systems used in analysis of the PDFT

Rating (kW_p)	Number of Systems
1 - 1.5	86
1.5 - 2	30
2 - 2.5	14
3 - 3.5	4
4 - 4.5	1

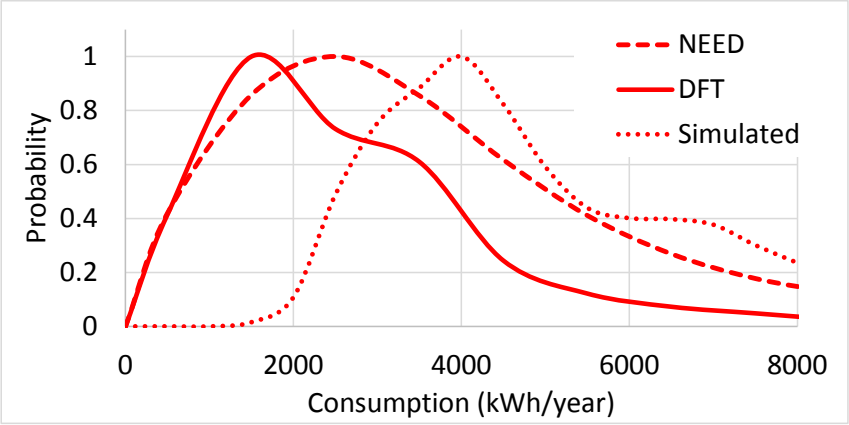


Figure 8-5 Comparison of annual electricity consumption in the NEED dataset for 2010, the PDFT sample, and simulated data

For each system the annual specific yield, the total household annual electricity demand and the self-consumption were determined. The specific yields have already been contrasted with the Sheffield Microgeneration Database yields in Chapter 5, where it was noted that contemporary systems are performing significantly better than those used in the PDFT. A comparison of the annual electricity consumption with empirical data from the NEED dataset (Chapter 6) is shown in Figure 8-5. This illustrates that the PDFT sample exhibits somewhat lower consumption than that of the national population. An exact correspondence of the marginal annual electricity demand profile in the PDFT and NEED datasets is not critical for the construction of a valid CPT; of importance is the availability of enough cases in the sample for each interval of annual consumption for which to derive the conditional probabilities. Note that this figure also shows the annual consumption in the simulated data discussed in the next section.

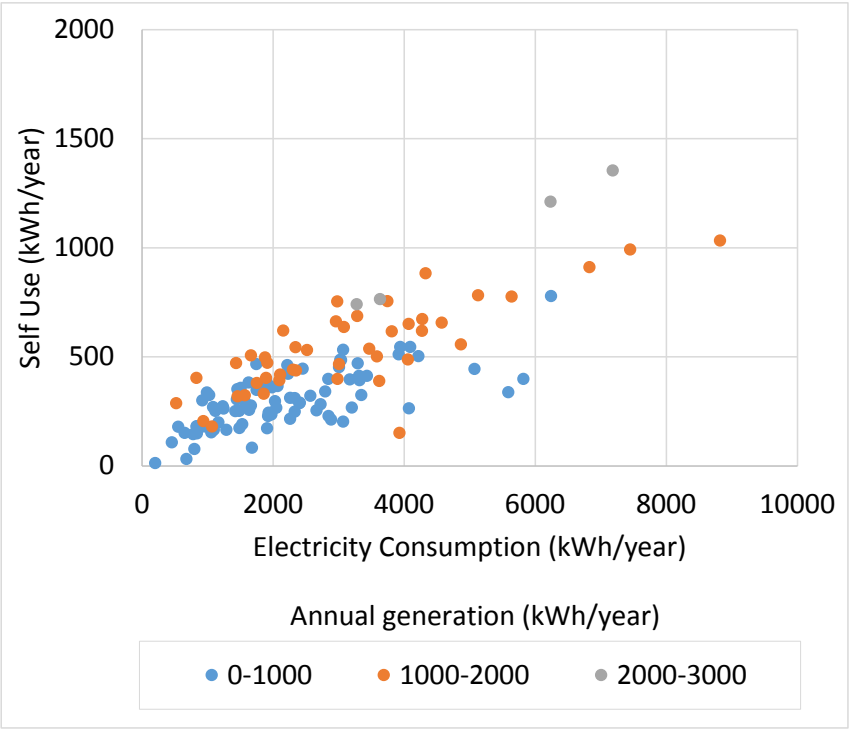


Figure 8-6 Annual self-consumption as a function annual electricity consumption segmented by annual system yield (generation) from the PDFT data.

Nevertheless the PDFT self-consumption data do demonstrate the hypothesised trends represented by the BN model. Figure 8-6 shows an increase in self-consumption as both the annual household electricity demand and solar PV generation increase. The scatter of the data also supports the notion of a highly stochastic model predicated upon a wide variety of occupant behaviours.

The resultant PDFT data were used to generate a CPT using counting learning algorithms in Netica (see Chapter 3) but this had two fundamental problems. Firstly the sample size was too small to deliver a noise-free CPT with some states having a number of samples in single figures. Secondly the range of system ratings, dominated by the small one kW_p systems employed in the trials, did not produce a satisfactorily wide range of generation. The PV Generation sub-model in chapter 6 estimates a much higher range of generation and no probabilistic evidence is available in this region. The comparisons of specific yield, rating, annual electricity consumption suggest that the PDFT data

is a biased sample, not generally representative of the national population. Indeed the PDFT did not use a random sample of PV adopters but a purposeful selection of new build social housing (Munzinger et al, 2006).

Two solutions to overcome the lack of time resolved empirical data with adequate temporal resolution where employed. The first was to resort to a simplified self-consumption model based on the UKs policy of deeming 50% self-consumption, but adding an appropriate degree of uncertainty to this. Figure 8-7 shows the distribution of the self-consumption fraction for the PDFT data. 50% does indeed appear to be the median value, but with a broad (almost normal) distribution covering 0 to 100%. However this model only allows the magnitude of generation to predict the self-consumption and would render it statistically independent of electricity consumption, which, as demonstrated by the PDFT data, is counterfactual.

The second solution to this problem was to employ simulated data using the Richardson and Thompson model. This is presented in the next section.

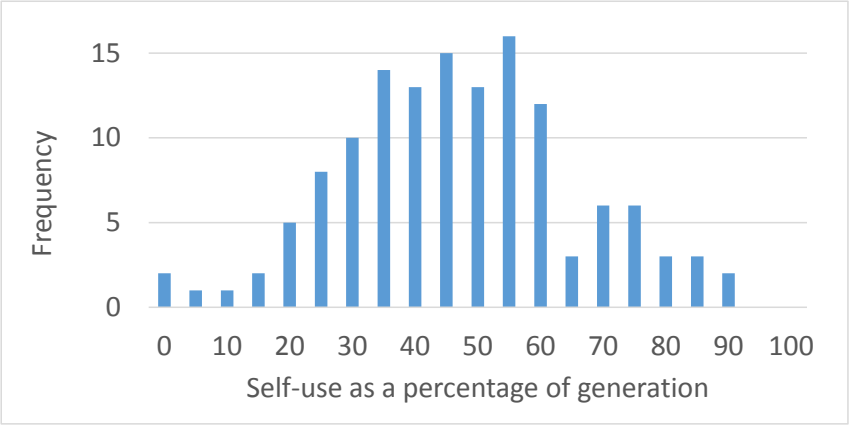


Figure 8-7 Self-consumption as a percentage of total annual generation for the PDFT data

8.4 Simulation of Self-consumption

This section presents a construction of a self-consumption dataset analogous to that presented above using simulated domestic demand and PV generation data. The simulation used is wholly based on that published by Richardson and Thomson (2012) with a few modifications. This is available as an open source Excel Spreadsheet application written in visual Basic for Applications.

The electricity load profile generator is constructed from the minute aggregated demand from a set of household appliances which are randomly assigned to the dwelling based on published statistics of appliance ownership and ratings. Appliances are categorised into several groups: those that run all the time, consuming a base load such as refrigerators and freezers, and those that require an active occupant (a person who is at home and not in bed) performing a particular activity to operate them, such as a television or an electric shower. The active occupancy and activity profiles within the dwelling are simulated stochastically using temporal probabilities derived from the UK's time-use survey (TUS) for between 1 and 5 residents. Separate probability data are used for weekdays and weekends. The model also features a seasonally linked lighting simulation module (Richardson et al., 2009) and the operation of primary and secondary electric heating sources.

The PV generation profile is simulated by first calculating the clear-sky irradiance for every minute of the day using published sun path algorithms as discussed in chapter 5. A clearness index is used to model the attenuation of the clear-sky irradiance due to clouds (Skartveit and Olseth, 1997). Richardson and Thompson (opt. cit.), using a one-minute time-series of empirical horizontal irradiance data recorded in Loughborough, England over a whole year (Betts and Gottschalg, 2007), created a transition probability matrix (TPM) which allows the stochastic prediction of the clearness index at time t_{n+1}, given the clearness index at time t_n. In this way the TPM is used to generate a one-minute time series for clearness index for a whole day. By multiplying each corresponding minute's clearness index by the clear sky irradiance a realistic time series of horizontal irradiance is simulated.

The tilt and azimuth of the plane of array are taken into account to calculate the irradiance in the plane of array (Dusabe et al, 2009) and a simple system efficiency method is used to convert the irradiance into an estimation of the minute by minute AC electrical output of the system.

The application allows the user to specify the location, size, azimuth and tilt of the PV panel. The day of the year, whether it is a weekday or weekend, and the number of residents between one and five must be entered. The appliances present in the dwelling can be user-determined, or randomly allocated prior to running the simulation.

Once the start parameters are entered the simulation takes a few seconds to run once the clear sky irradiance has been generated for the location and day of interest, which takes about 30 seconds. Graphical outputs for the PV generation, occupancy and load are presented along with the aggregated values for the whole day.

Modifications to the Software

The requirement is to generate an aggregated SUF over a whole year for appropriate combinations of annual consumption, generation. If the maximum range for these were taken to be 10,000 and 5,000 kWh/year respectively, self-consumption were assumed to maximise at 100% (i.e. also 5,000 kWh/year) and 1000 kWh sampling intervals are required, then an average sampling rate of 100 simulations per bin would require 25,000 simulations. With automated start parameter entry (for example changing the day number) this would require ten years of CPU time. A modification to the application architecture brought about a 50 fold increase in the speed of calculating the clear sky irradiance, thus making the attainment of a reasonable number of simulations in a short time feasible. This was further extended by running the modified application on several PCs and combining the results for analysis.

The application was further modified to automatically cycle through every day of a whole year with a fixed set of start parameters. The automatic entry of these is carried out by selecting a random value between an upper and lower limit for each parameter as tabulated (Table 8-2)

Table 8-2 Start parameters for automated annual simulation

Parameter	Lower Value	Upper Value
Number of residents	1	5
PV System Rating	1	6
Azimuth	-90	90
Slope	35	35
Demand Calibration	0.25	4

```
Start Annual Simulation
      Set Start parameters
            Random number of residents
            Random PV System Rating
            Random Azimuth
            Random Slope
            Random Demand calibration
            Allocate appliances
      For each day in the year
            Set day number
                  Updates clear-sky irradiance for the location
            Simulate occupancy profile
            Simulate clearness index time series
            Simulate lighting use
            Simulate appliance use

            Aggregate demand for the day
            Aggregate generation for the day
            Aggregate export for the day

            Add day's aggregated demand to year's running total
            Add day's aggregated generation to year's running total
            Add day's aggregated export to year's running total
      Next day in the year

      Save total demand for the year
      Save total generation for the year
      Save total export for the year

Go to Start Annual Simulation
```

Box 8-1 Descriptive software code to automate simulation for multiple years

Following this the appliances are allocated to the dwelling and the simulation is primed to run for a whole year. The aggregated demand, generation and export are determined for each day and added to a running total for the year. After the last day of the year the running annual totals are saved along with the start parameters and the simulation repeats for another whole year with new random start parameters. Box 8-1 shows this in a descriptive software code.

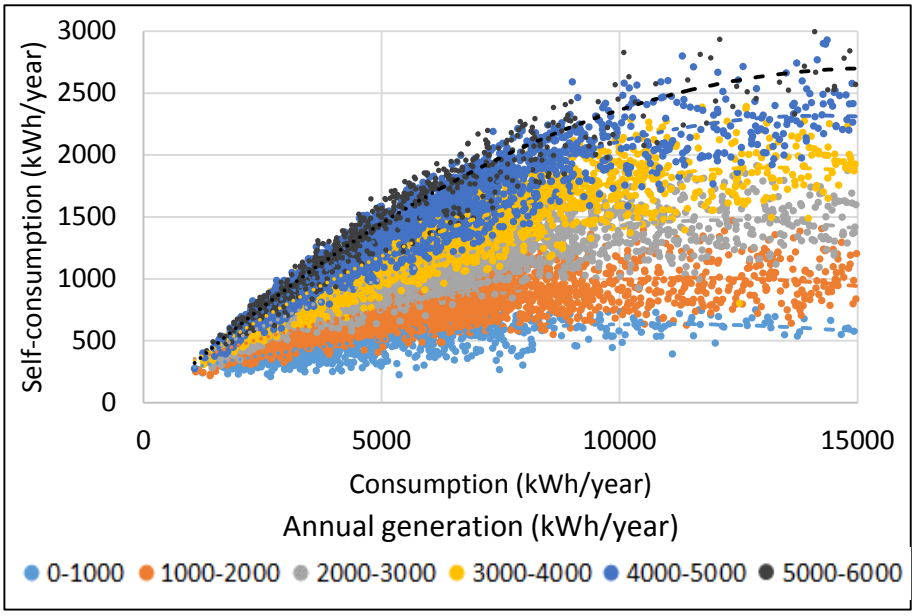


Figure 8-8 Annual self-consumption as a function annual electricity consumption segmented by annual generation for simulated data.

Result of Simulation

Figure 8-8 shows the magnitude of self-consumption as a function of annual electricity consumption segmented by annual generation, obtained from approximately 25,000 annual simulations. The simulated dataset exhibits a suitably high range of annual electricity demand and generation but does not have any data points at very low consumptions. This is shown in Figure 8-5 (see Page 228) alongside the empirical NEED and PDFT data. It is apparent from this that the stochastic simulation model under-represents low electricity consumers prevalent in the general population. It is not fully

clear why this should be so; it could be due to over estimation of occupancy or activities in the TUS, or the overestimation of appliance ownership; all of which militate against the observation of low electricity energy consumption. At the high consumption end specific appliance signatures are observed. For example a kink in the distribution between 6 and 8 thousand kWh is due to simulations where domestic electricity storage water heating (DESWH) was present (Figure 8-9). Similarly night storage heaters contribute strongly to the cases with an energy consumption above 10,000 kWh; night-time demand does not of course contribute to self-consumption.

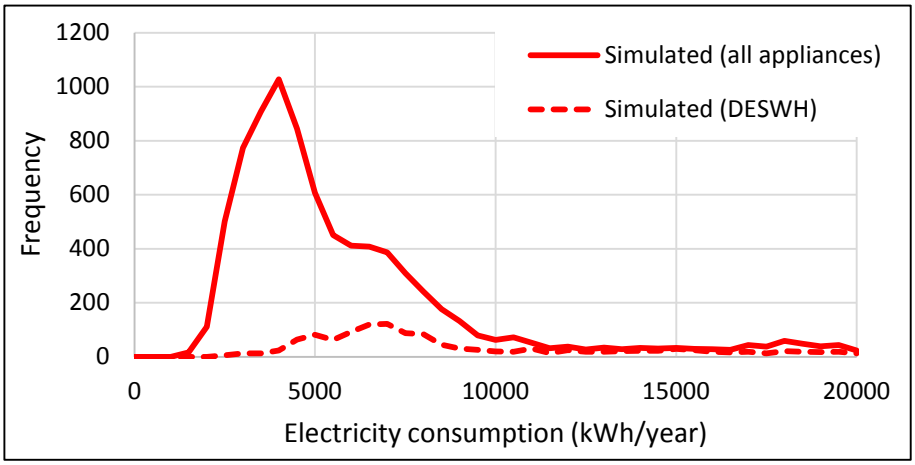


Figure 8-9 Simulated electricity demand showing 'signature' of water heating appliance at high electricity consumption values

It is instructive to construct a frequency distribution of self-consumption modelled by the simulated data in Figure 8-8 to compare with the empirical data in Figure 8-7. This shows (Figure 8-10) that the mode self-consumption lies towards 35%, and not 50% as the PDFT data suggested.

Despite these features of the simulation, the methods purpose is to generate realistic demand and generation temporal profiles as opposed to delivering accurate aggregated demand, or a distribution thereof which reflects the general population. It can be suggested, however, that faith in the temporal profiles would be boosted if annual demand profiles were consistent with empirical data. Nevertheless, there is good agreement with actual empirical data observed at low consumption and

generation, and the general trends concur with the hypothesis that both generation and consumption are strong predictors of self-consumption with an expected large variability. Thus, it is concluded that a useful joint probability distribution has been created which can be used to furnish a BN with a CPT. This is discussed in the next section.

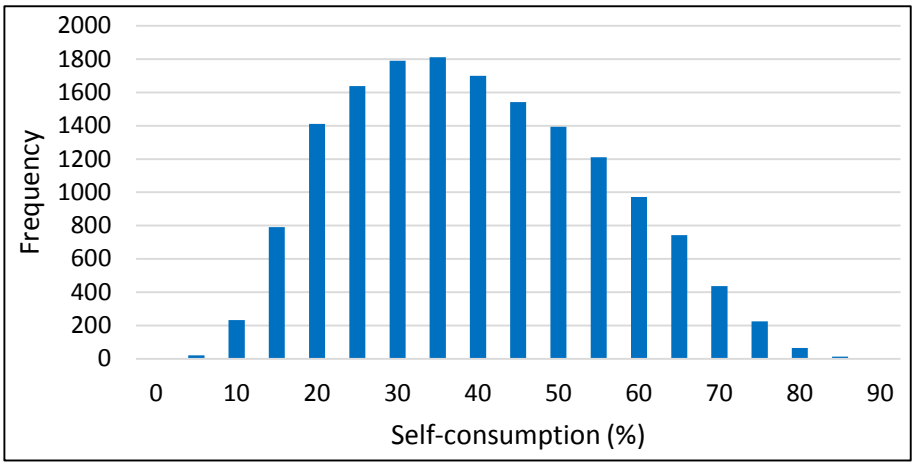


Figure 8-10 Self-consumption as a percentage of total annual generation for the simulated data

8.5 Bayesian Network Submodel for Self-consumption

A Bayesian network model was constructed in Netica with both consumption and generation as parents of the self-consumption node (Figure 8-11). An extra percent self-consumption node was added; this is a deterministic node where the CPT was calculated from the ratio of self-consumption to generation, expressed as a percentage. The discretisation interval for consumption and generation was set to 500 kWh and for self-consumption 200 kWh was used.

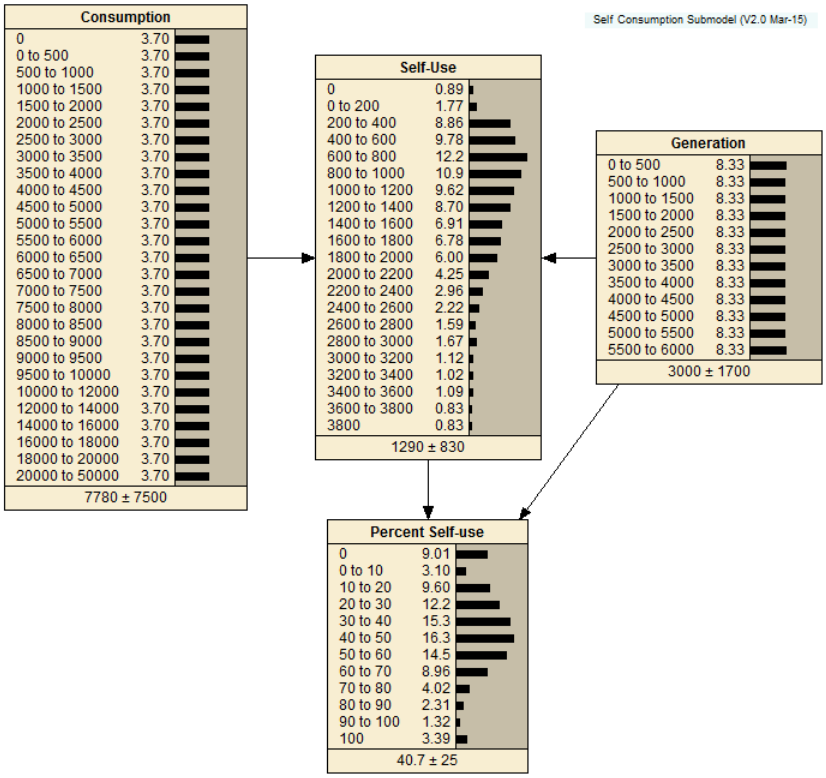


Figure 8-11 Bayesian network model for self-consumption derived from simulated and empirical data.

Both empirical data from the PDFT and the simulated data were used as case files for model learning using the counting method (see Chapter 3). Because the marginal distributions of consumption and generation are not pertinent to the final use of the model these were set to a uniform distribution following the learning operation. Note that this does not change the underlying CPT, but renders the model ambivalent about the marginal distributions of the parent nodes in the case files. This is entirely appropriate since the empirical or modelled probabilistic evidence furnished by building energy model and PV yield sub-models in the integrated model is more important than the marginal data from the simulations.

8.6 Discussion and Conclusion

A potential flaw in the generation profiles is due to the reliance on a single TPM for the clear sky index which is based on a whole year's data collected at a single location. The probabilities in the matrix are therefore deemed independent of season and location, which may result in seasonally and spatially unrepresentative profiles, a fact which the authors themselves have highlighted (Richardson and Thomson 2012). Since the greater contribution to self-consumption will occur in the summer months when generation is higher the use of this TPM may have an impact on the self-consumption factor as winter clear sky transition probabilities influence the summer generation profiles.

The model shows that very high percentage self-consumption does not occur with high probability even as consumption rises. This is because high consumption is probably due to electric winter heating and appliance and lighting loads occurring in the evening which is when occupancy is the highest. This is demonstrated by Figure 8-12 which shows the average occupancy on a weekday, generated for 500,000 simulations, 100,000 for each resident count (1 to 5).

Figure 8-12 hints at a flaw in the demand simulations since each day in the annual simulation generates a new occupancy profile. Implicit in this is that a domestic unit, consisting of 1 to 5 residents, has a different daily behaviour which is unlikely since the weekday occupancy pattern, if not also the weekend pattern, for most households, are likely to have a high degree of consistency, particularly if there is regular employment and school attendance. Thus the simulations run here are more randomised than they should be. Yao and Steemers (2005) have proposed five domestic load archetypes which relate to active occupancy (Table 8-3). A brief exploration of the simulated occupancy profiles using a BN analysis suggests that these archetypes are only weekly present but further work is required to apply pattern recognition techniques to time-series data as carried out by Aerts et al. (2014)

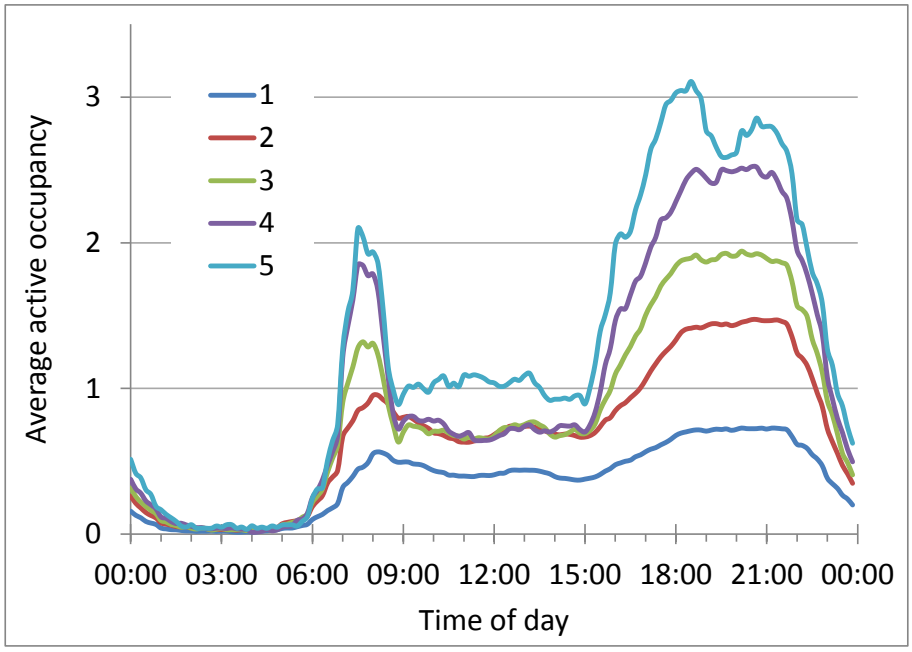


Figure 8-12 Average weekday active occupancy for dwellings with one to five residents. Each curve has been generated by averaging 100,000 simulations using the 2-state occupancy model.

The absence of occupancy patterns used in these simulations are likely to have some impact on the determination of the self-consumption factor. The random allocation of daily occupancy will have a tendency to deliver a more average value of self-consumption since consistent daily patterns with high and low occupancy with commensurate high a low electricity demand will not be present in the dataset.

Table 8-3 Typical appliance load profiles for average domestic household related to occupancy archetypes

Load Pattern Archetype
Unoccupied 9:00 – 13:00
Unoccupied 9:00 – 16:00
Unoccupied 9:00 – 18:00
Unoccupied 13:00 – 18:00
All Day Occupied

Finally a further parameter thus far ignored in these simulations is the influence of aspect of the PV system and the relationship to occupancy patterns. Figure 8-13 shows the average household

occupancy superimposed on clear sky irradiance profiles for East, South and West facing panels for both summer and winter. The aspect is seen to have a significant influence on the overlap between demand and generation with South facing panels benefitting from day time occupancy, West facing benefit from the increasing number of home-comings between four and six in the afternoon and East facing panels generate the most for the early morning 'breakfast' surge. Due to the randomised daily simulation none of these effects could be observed and further work is required to test the sensitivity of self-consumption to the PV array's aspect.

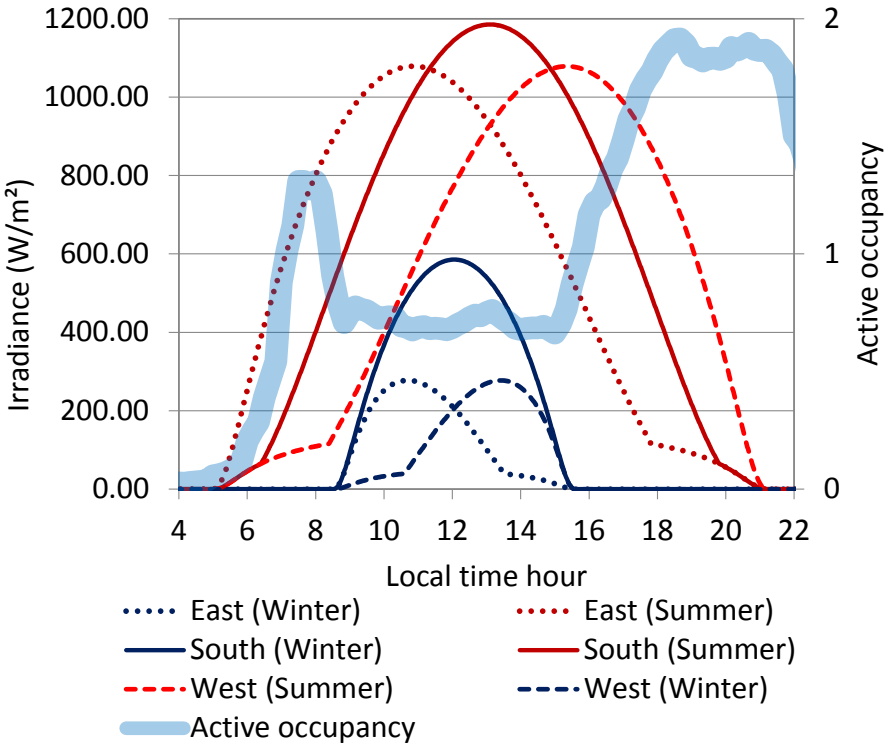


Figure 8-13 Average weekday occupancy superimposed on clear-sky irradiance profiles for different aspects and seasons.

In conclusion self-consumption is intuitively dependent on the magnitude of electricity consumption, and generation due to increased probability of overlap between the temporal profiles of each. This has been verified with both empirical and simulated data, the latter for over 16,000 years' worth of minute resolved temporal profiles of demand and generation. The aggregated

magnitudes of demand and generation in these simulations are less important than achieving the same ‘spikiness’ in generation due to rapid variations in cloud cover, and in demand due to sudden demand surges resulting from the cycling of appliances.

9 The Integrated Bayesian Network

9.1 Introduction

The conceptual model presented in Chapter 4 has, at its apex, the spatial context for a renewable energy technology which influences the renewable energy yield and the total energy consumption. These in turn influence the degree of self-consumption on the site. Reified for the case of PV in domestic urban contexts, these four 'cornerstones' of a balanced energy system have been developed as four separate Bayesian network models in Chapters 5 to 8 (Figure 9-1).

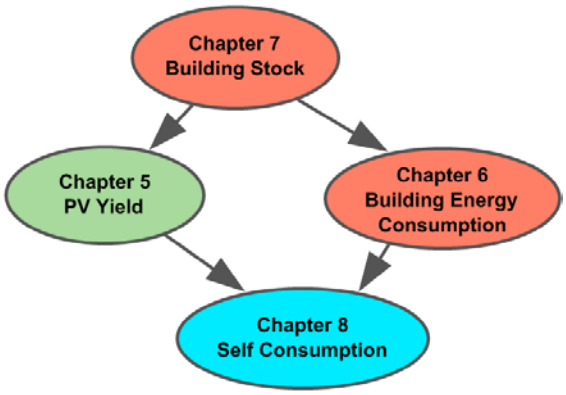


Figure 9-1 The four 'cornerstones' of the integrated model for PV

The purpose of this chapter is two-fold. Firstly, the integration of these four components to create an OOBN model is described (Section 9-2) and some initial findings are presented in Section 9-3. Secondly, the three representative SEE indicators selected in Chapter 4 are introduced as components to the network. Section 9.4 explains the general treatment of these indicators, and Section 9-5, 9-6 and 9-7 present the indicator for carbon savings, techno-economics and fuel affordability respectively. Section 9-8 presents and discusses findings pertaining to these indicators in the context of the integrated model.

9.2 Creating the Integrated Object Oriented Bayesian Network (OOBN)

An OOBN allows the connection of sub-models 'objects' using interface nodes. The four 'cornerstone' models have been purposefully designed to represent autonomous knowledge domains which share probabilistic data through these interface nodes. The OOBN can be represented by an entity relationship (ER) diagram (Figure 9-2). The blocks represent sub-models and the interfaces are represented by connectors between the parameters, with the arrows' tails indicating outputs, and the arrowheads, inputs.

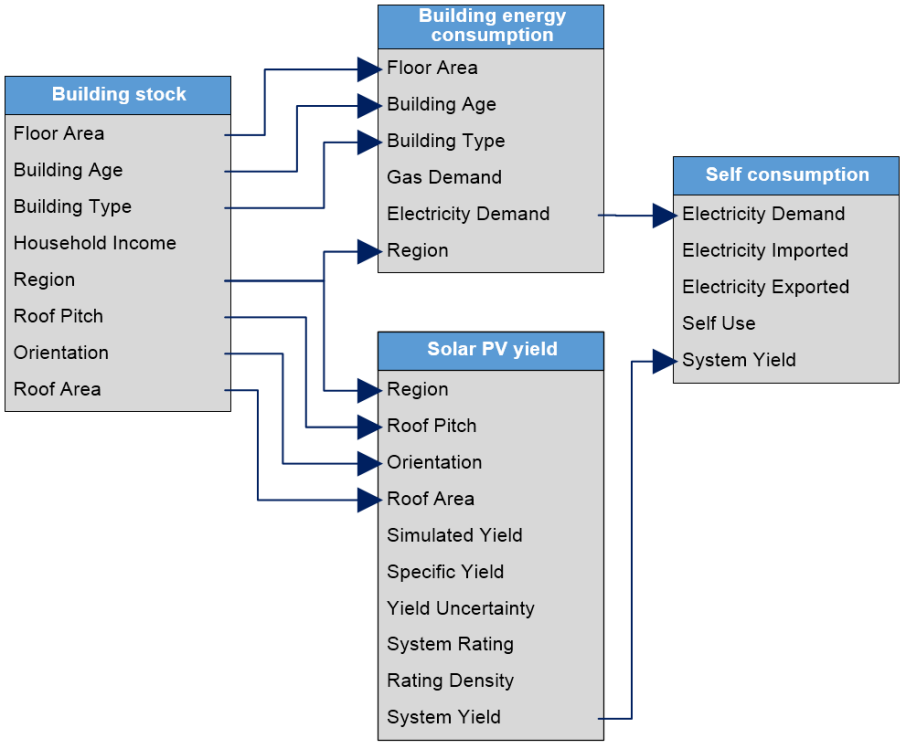


Figure 9-2 Entity relationship (ER) diagram representation of the integrated PV model showing the interfaces between the objects

To create the OOBN in the Netica software, each functioning sub-model is copied to a new network. Connections are made between Interface nodes, which were purposefully designed to have the

same number, and values, of discrete states[37]. This entails converting the input node to a deterministic node and equating its distribution to the corresponding output node. This ensures that evidence received at, or applied to, any side of the interface is faithfully reproduced at the other, since the evidence at each side of the interface cannot differ. This is illustrated in Figure 9-3, where, the prior distribution for the 'Property Age' node at the output side of the interface is faithfully reproduced at the input side (A). Similarly, if evidence is applied to the variable at either side of the interface (B) this is also faithfully reproduced at both sides of the interface.

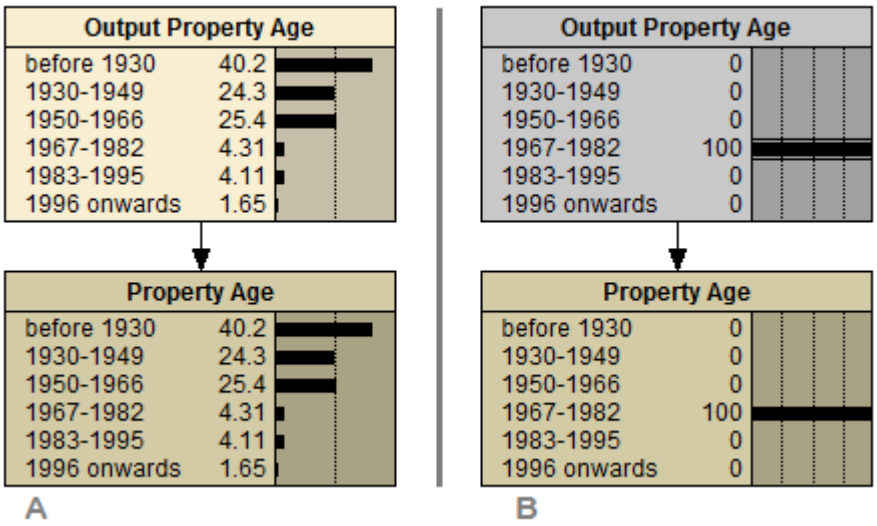


Figure 9-3 Example of an interface node with (A) a prior distribution for the variable and (B) hard-evidence applied to either input or output side of the interface

An image of the final OOBN with the four sub-models in Netica is shown in Figure 9-4, albeit at a low resolution – the purpose here is to give the reader an insight into the size, and complexity, of the network. To understand the architecture of the whole model in detail, the use of the **ER** diagram is advocated which shows the connections more clearly, and the individual BN sub-models are

[37] In Netica, the ordinal position of states must also be the same at both sides of the interface.

discussed in depth in their dedicated chapters. In the next section observations on output nodes in the dependent sub-models, as a result of connection to the building stock sub-model, are presented.

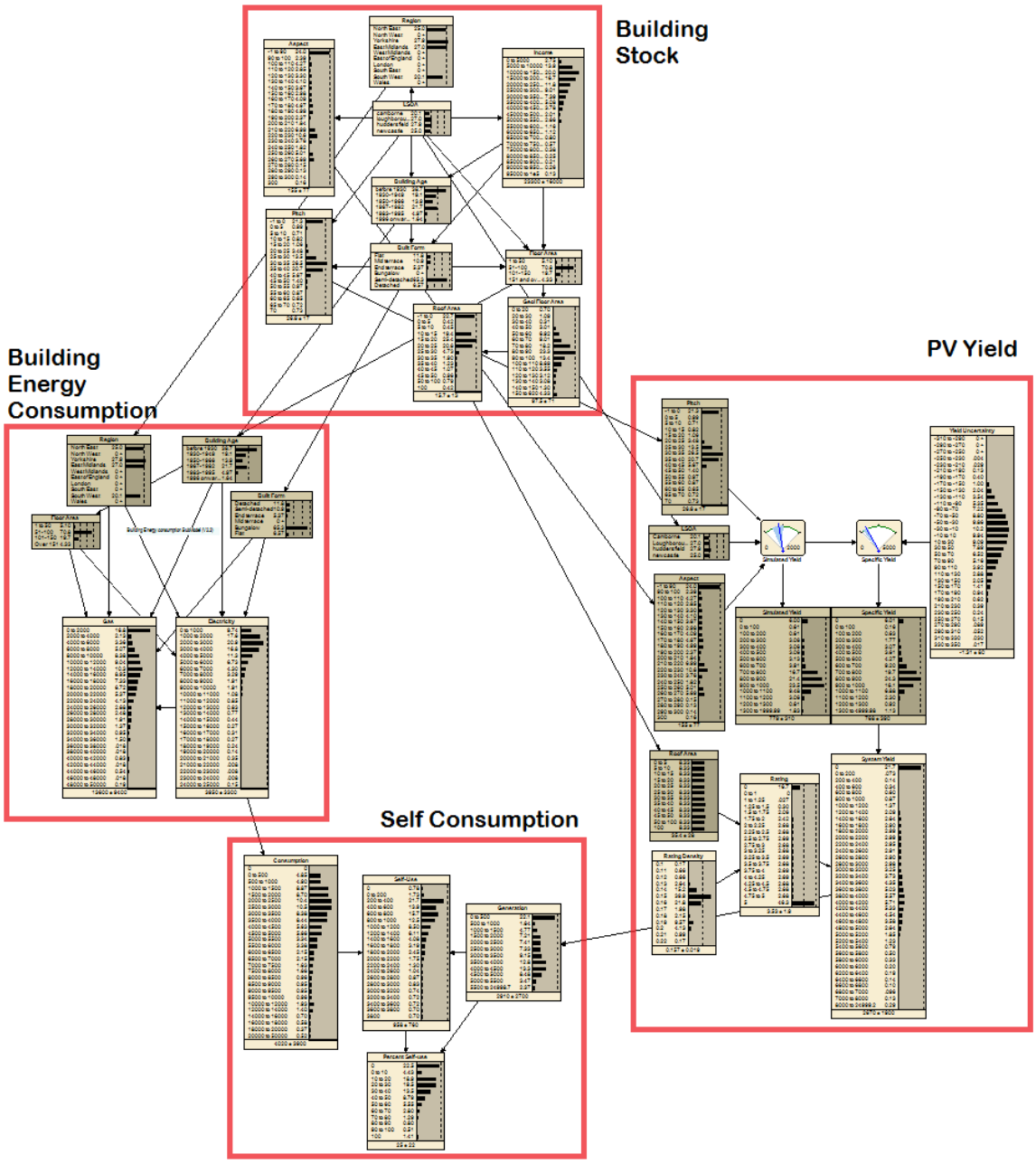


Figure 9-4 The OOBN, consisting of the 'four cornerstones' sub-models connected together in Netica

9.3 Preliminary Observations for the OOBN

The properties of the BN objects which constitute the OOBN have previously been explored in the sub-model specific chapters (Chapters 5 to 8). Here, for the first time, the posterior probabilities of nodes belonging to the PV yield, building energy consumption, and self-consumption sub-models, can be explored in response to the updating of building stock attributes with the prior distributions for each LSOA as encoded in the building stock sub-model. This is achieved by selecting the LSOA node of the building stock model as hard evidence, which then sets all the building attributes to the empirical distributions for the selected LSOA. This is tantamount to assigning probabilistic evidence to these attributes, which, by virtue of the interfaces between sub-models, propagates new probabilistic evidence to the dependent sub-models. Moreover, this new probabilistic evidence is spatially specific, emanating from the selected LSOA.

To examine the properties of the OOBN it is instructive to observe gas consumption, electricity consumption, PV yield and the self-consumption – four key output nodes of interest which appear in the dependent sub-models. The histograms for the four parameters, extracted from the OOBN, are displayed in Figure 9-5 for each LSOA. The distributions appear similar, and in order to facilitate comparison statistics, expected value (EV), standard deviation (SD) and coefficient of variation (CV), are also displayed for each chart.

The charts in Figure 9-5 include buildings, which, for reasons discussed in Chapter 7, have no viable roof to host a PV system. These manifest as the large probability of low electricity generation, representing approximately 22 to 26% of households in the LSOA without a viable roof. The model also shows a significant proportion of properties with zero or low gas consumption, as learnt from the NEED framework dataset (Chapter 6) and encoded in the CPT for this sub-model. LSOA Kerrier 008B and LSOA Charnwood 002D, in particular, have higher probability for low gas consumption.

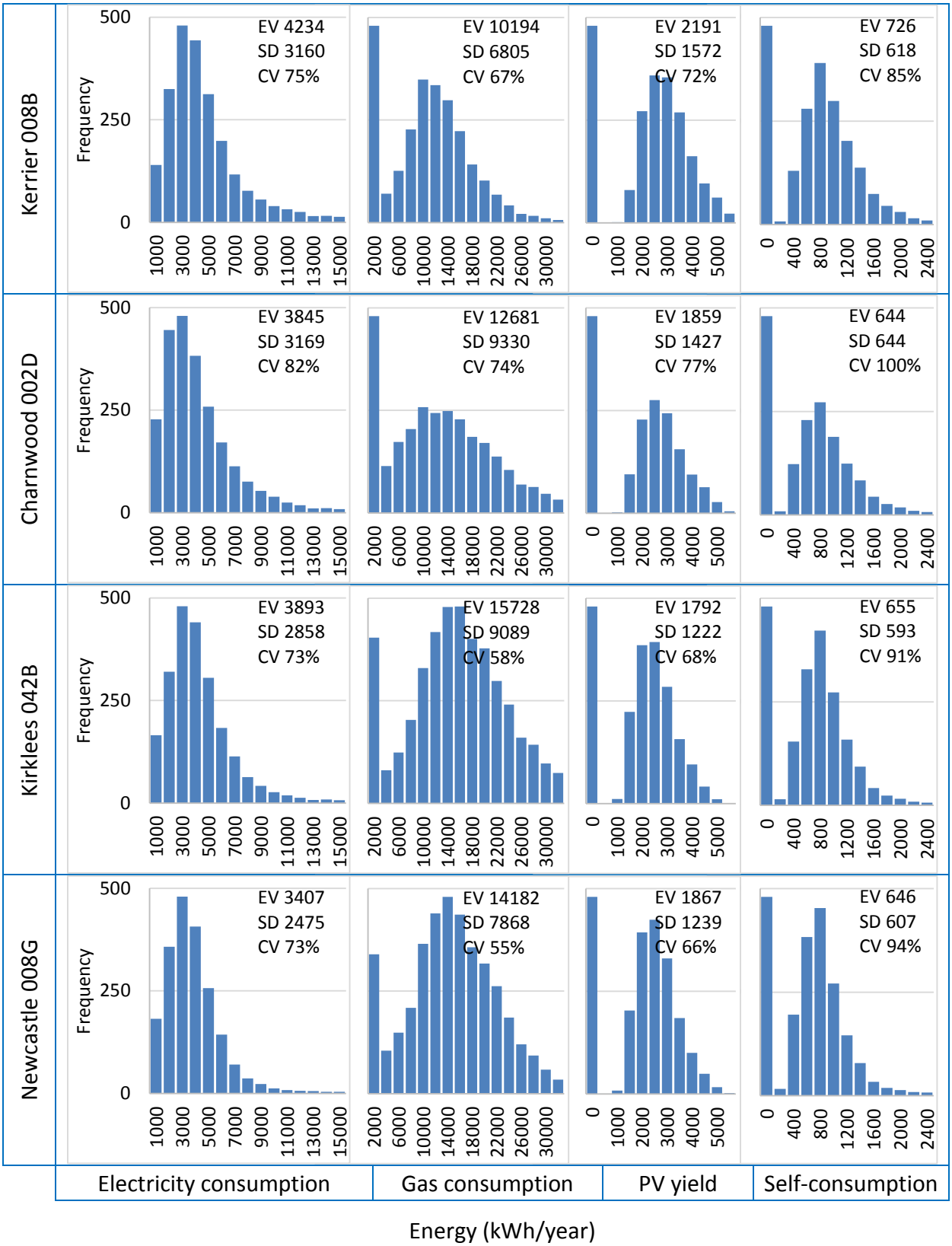


Figure 9-5 Electricity consumption, PV yield and self-consumption distributions with expected value (EV), standard deviation (SD) and coefficient of variation (CV).

This reflects the greater probability of off-gas properties in the South West region, and the greater propensity for the larger number of flats in LSOA Charnwood 002D to be without gas.

A different perspective on output nodes of interest is obtained by selecting only properties with a viable roof and those which consume gas. Figure 9-6 shows the same charts as in Figure 9-5 but the 'zero and low' gas consumption state has been given a zero likelihood, as has the state for zero 'roof area', whilst maintaining all other category likelihoods as unity (Equation 9-1).

$$P(Observation \mid roofarea = 0) = 0$$

$$P(Observation \mid gas = '0\ to\ 2000') = 0$$

Equation 9-1

With this hard evidence applied, the EV of gas consumption rises by 9 to 23% and the CV drops by a factor of 15 to 30% when compared to the entire building stock. Similarly the EV for PV yield is 40% higher for LSOA Charnwood 002D, and 28-29% for the other LSOAs. The coefficient of variation approximately halves. The increase in yield and decrease in variability is not unexpected when buildings with no viable roof are omitted. The larger increase for LSOA Charnwood 002D is due to the larger percentage of flats for this LSOA which have no viable roof.

This demonstrates how the BN allows output parameters attributed to locality's entire building stock can be compared to those with specific observations, as exemplified by Equation 9-1. This presents a dilemma; should the output (posterior) distributions and their statistics, be discussed in terms of properties which do not consume gas and/or have no viable roof to host a PV system, or the opposite? The answer, of course, depends on the questions asked; the achievement here is to have justifiable probabilities for both scenarios. For the present purposes it is worth examining the OOBN for expected trends with the zero states for gas consumption and roof area set to zero likelihood in order to verify if the model is delivering expected results.

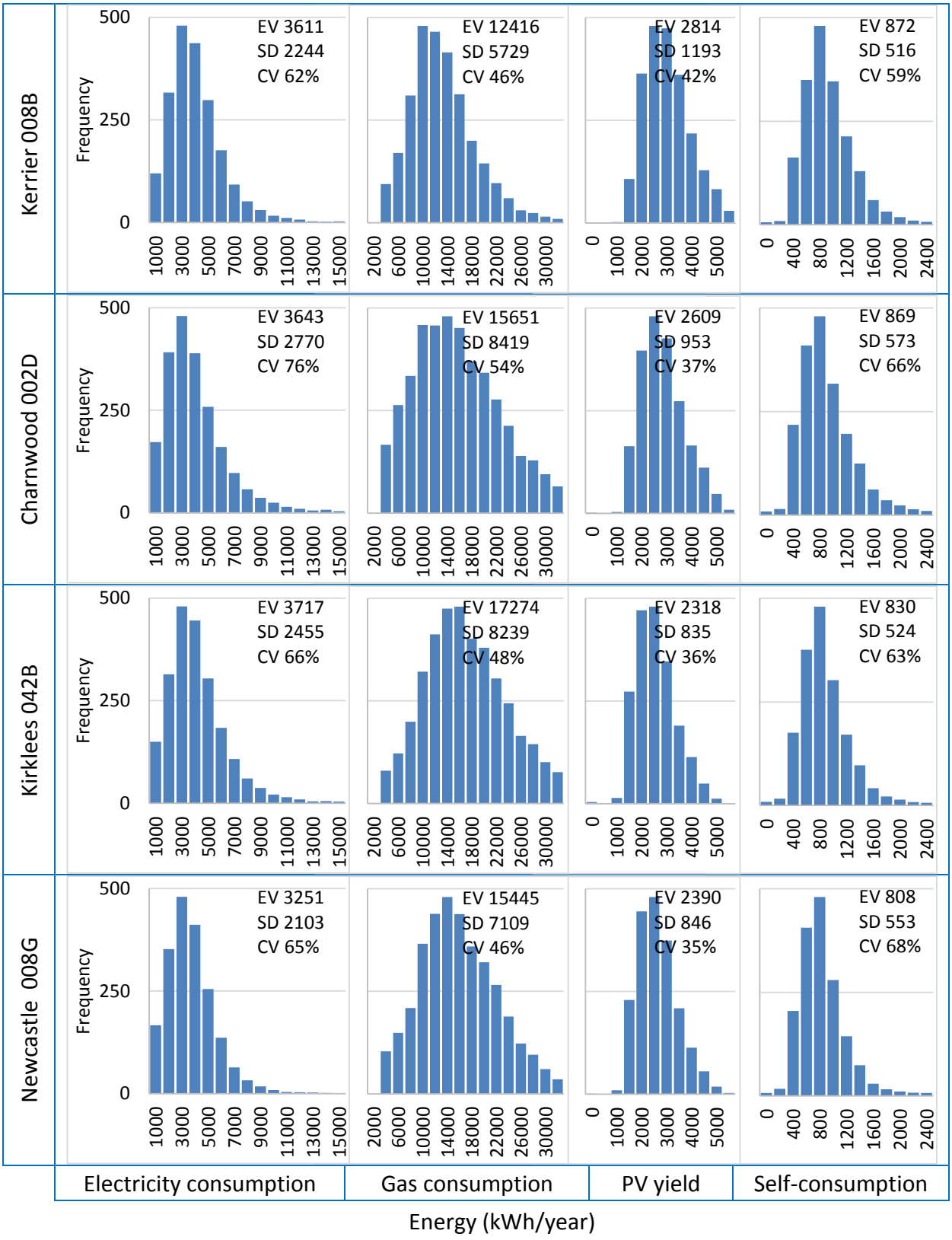


Figure 9-6 Electricity consumption, PV yield and self-consumption distributions with expected value (EV), standard deviation (SD) and coefficient of variation (CV) with zero-area roofs and low gas consumption excluded.

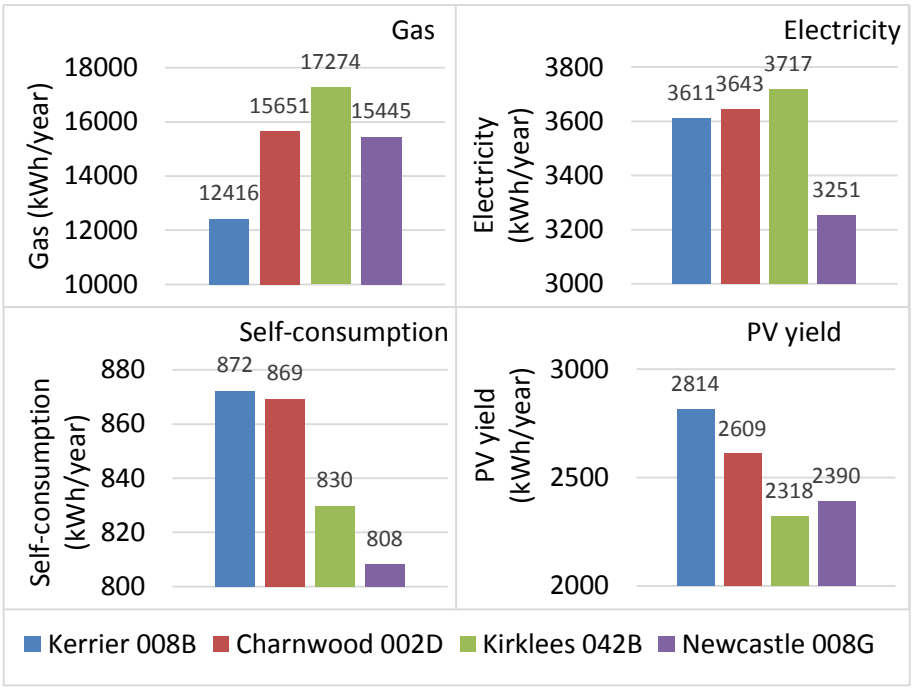


Figure 9-7 Expected value for the gas consumption, electricity consumption, PV yield and self-consumption distributions.

Firstly, comparisons of the expected value (mean) for the four key output variables in each LSOA are shown in Figure 9-7. The analyses were performed excluding zero and low gas consumers, and zero-roof area dwellings. The EV of gas consumption shows a steady increase in consumption going North (LSOA Kerrier 008B, LSOA Charnwood 002D, LSOA Kirklees 042B), with LSOA Newcastle 008G showing an exception to this trend, having a value similar to LSOA Charnwood 002D. The EV for electricity consumption shows a similar, but less dramatic rise, and again LSOA Newcastle 008G is the exception, having the lowest electricity consumption of all. Sense of these trends can be made by comparing them with the EV of building energy demand predictors in the building stock model. Figure 9-8 compares the EV for floor area, income, and building age[38] for each LSOA. Thus LSOA

[38] The building age has been artificially calculated by taking the median value of the building age category ranges used in the NEED framework. Whilst not a rigorous method of calculating the ages of buildings, it serves as a method for comparing the average age of buildings in each LSOA.

Kerrier 008B has the lowest gas consumption, possibly because the building stock is newer and incomes are lower; it is also milder in the South West. LSOA Newcastle 008G, in the (colder) North East might be expected to have the highest gas consumption; this, may be mitigated by the lowest incomes and floor areas of the LSOAs, rendering the consumption similar to LSOA Charnwood 002D.

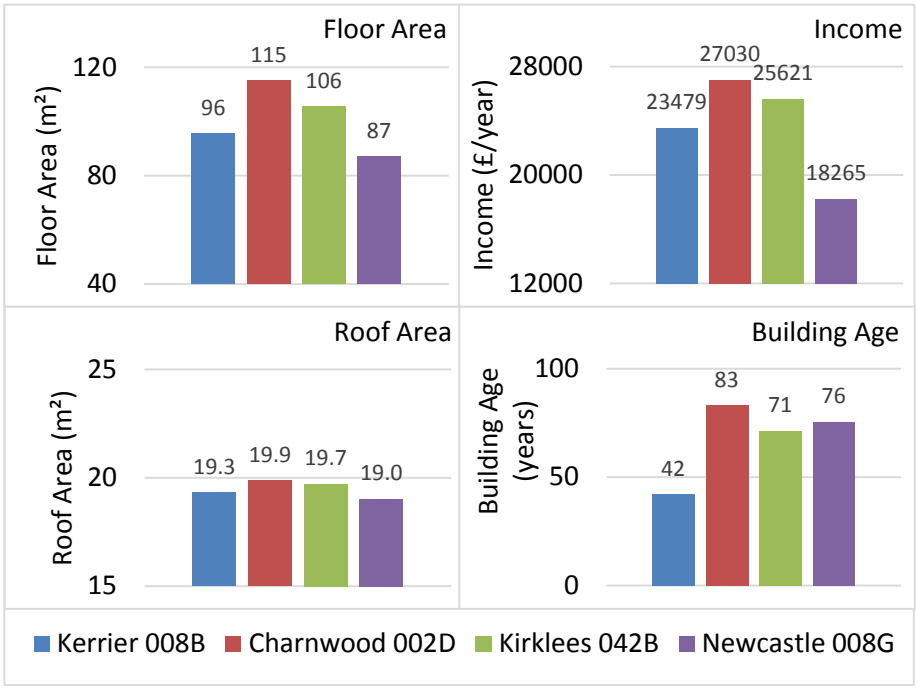


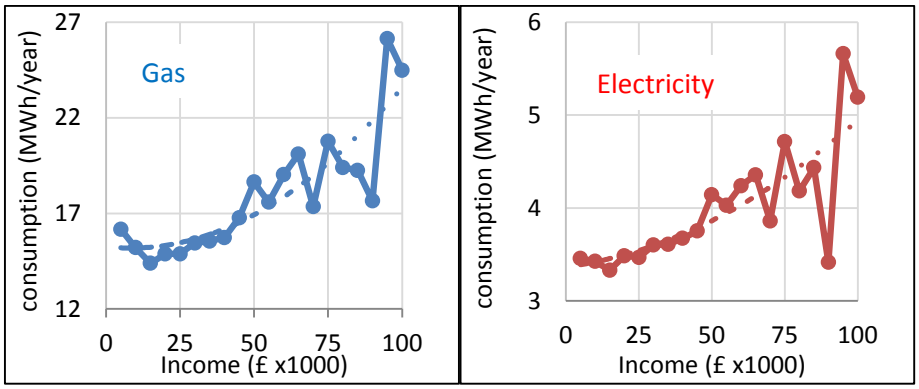
Figure 9-8 Comparison of the expected value for key predictor variables in the building stock model

Electricity consumption is less sensitive to building stock and climate parameters and more sensitive to occupant factors, as discussed in Chapter 6. This may explain the small difference in EV for electricity consumption between LSOA Kerrier 008B, LSOA Charnwood 002D and LSOA Kirklees 042B, and the lower EV for in LSOA Newcastle 008G since this LSOA has a significantly lower income than the other three.

LSOA Kerrier 008B has the highest PV yield, followed by LSOA Charnwood 002D, LSOA Newcastle 008G and then LSOA Kirklees 042B have the lowest; this follows the same trend as the autonomous PV Yield model demonstrated for specific yield (Chapter 5). There is no significant effect of the

building stock, a fact not surprising when the roof areas are taken into account (Figure 9-8), which are very similar, indicating similar sized PV systems can be installed in each LSOA.

The expected value of self-consumption is slightly higher for LSOA Kerrier 008B than LSOA Charnwood 002D, with a drop for LSOA Kirklees 042B and a further reduction for LSOA Newcastle 008G. This follows the expected trend given the relative electricity consumption and PV yield values. However, the difference in EV for self-consumption is not very great between the four LSOAs.



The dashed lines are second order polynomial regression fits to observe the trend of the data

Figure 9-9 Expected value of annual gas and electricity consumption, as a function of hard evidence for household income states, aggregated for all four LSOAs.

As well as observing the trends in expected values for key output and input parameters it is also pertinent to test the impact of hard evidence on the model. Figure 9-9 shows the effect of setting hard evidence for household income to each successive state in turn, and observing the EV for gas and electricity consumption. Both of these show an increase in energy demand with income. This shows the efficacy of the IPF, discussed in Chapter 7, applied to the integrated model. The trends in EV for gas and electricity consumption follow very closely the trend in floor area (Figure 7-25, page 220) which is a key predictor for building energy consumption (Chapter 6).

The preliminary observations of this integrated model suggest that expected trends are observed by observing statistical metrics for the posterior probability distributions. In the following sections the

OOBN is further enhanced by the inclusion of output indicators as discussed in Chapter 4 (Table 4-7, page 63).

9.4 General treatment of output indicators

As discussed in Chapter 3, Netica, in common with many BN software applications, allows the creation of deterministic nodes. In the following three sections, BN sub-models, which augment the OOBN with impact indicators, are presented and described. These use deterministic nodes which may have as inputs any of the probabilistic variables included in the OOBN (e.g. PV yield, gas consumption, electricity consumption, income, or floor area). Additional parameters may be required, represented by nodes which have an empirical or theoretical probability distribution, or are simply furnished with a uniform distribution.

The following sections give a brief, though rigorous, treatment of the indicators; the scope of this work does not permit a more thorough treatment of the respective knowledge domains, but assumptions will be highlighted. The key purpose is to provide an insight into how the OOBN model can be enhanced with these deterministic models to deliver decision support features as well as deliver some interesting outputs of the model thus far.

9.5 Carbon Savings

Solar PV generated electricity is either self-consumed or exported to the grid (see Chapter 8). The total PV yield displaces grid electricity (disregarding transmission and distribution losses incurred by the exported component). Thus the PV yield lessens the load on generators, reducing their consumption of fuel. Since a significant proportion of these fuels are derived from fossil sources, this results in carbon emission savings. For a PV system the carbon savings, C, are equal to the product of

the carbon intensity, I, defined as the mass of carbon released per unit grid electricity generated, and PV yield Y(Equation 9-2).

$$C = I \cdot Y \qquad \textit{Equation 9-2}$$

However, the determination of carbon intensity is complex, since, a basket of fuels, each with different carbon intensities, constitute the UK electricity supply. Furthermore, the proportion of each fuel varies according to the instantaneous load, which results in a variation of carbon intensity on both an hourly and seasonal timescale (Figure 9-10).

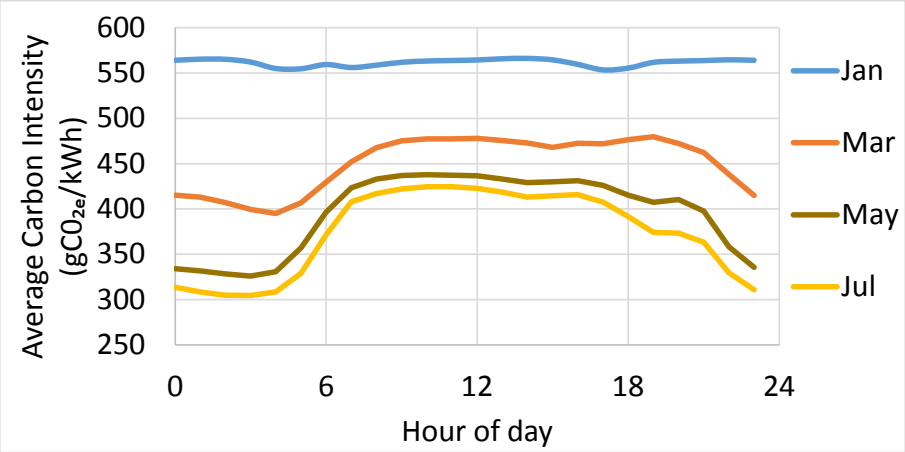


Figure 9-10. The average hourly carbon intensity of the UK electricity supply for 24 hours for January, March, May and July (after Hart-Davis, 2013).

For the assessment of the UK building stock an average annual carbon intensity is used. SAP recommends a standard five year average of 517 gCO_{2e}/kWh for electricity, and, for comparison, 198 gCO_{2e}/kWh for natural gas. In the 2012 edition of SAP (BRE, 2014) this was revised down to 502 and 401 gCO_{2e}/kWh for 5 five year and 15 year electricity averages respectively in. This downward trend is due to the decreasing carbon intensity of the electricity supply, which, over the period 1990 to 2010, has reduced from 770 to 490 gCO_{2e}/kWh (Figure 9-11). This is due to a shift from carbon intensive coal to less carbon intensive gas (Utley and Shorrock, 2009). As climate change targets are

fulfilled, carbon intensity is set to decrease further as more renewable and other low carbon energy generators are introduced into the energy basket. The Committee on Climate Change (CCC, 2012) is recommending a drop to 50 $gC0_{2e}$/kWh by 2030 (CCC, 2012).

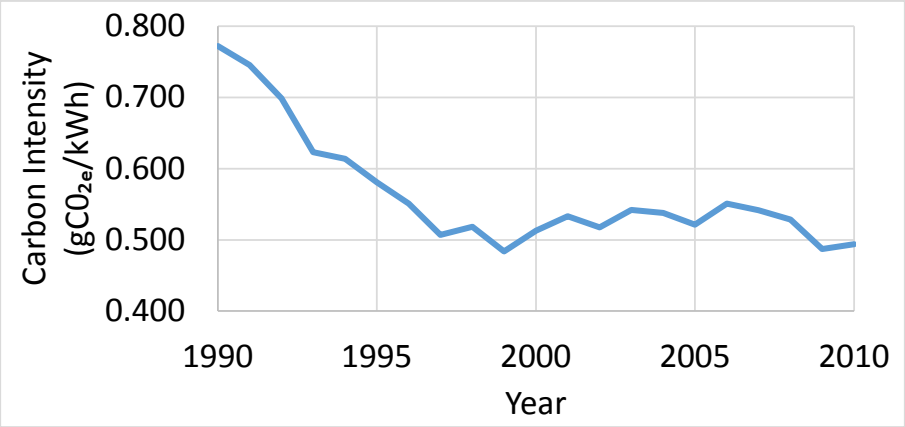


Figure 9-11 Electricity generation emission factor from 1990 to 2010, including imported electricity and transmission and distribution losses (after DEFRA, 2012)

9.5.1 Carbon savings BN sub-model

The DAG for the deterministic BN sub-model is shown in Figure 9-12. 'Carbon Savings' is deterministic node, and uses Equation 9-2. The PV Yield is taken as an input from the PV Yield sub-model.

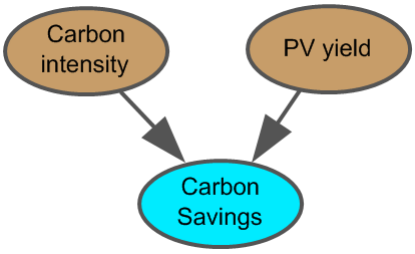


Figure 9-12 Deterministic BN Sub-model for carbon savings

The carbon intensity is represented by a probability node with a uniform distribution between 400 and 600 gCO_2/kWh. This encompasses the value ranges discussed above and thus allows a variety of scenarios to be modelled. In the SAP model, carbon intensity represents the mean value for grid electricity for the whole year. However, the daily and seasonal variations shown in Figure 9-10 will impact the carbon intensity of PV displaced electricity since, most generation is during seasons when carbon intensity is below average, but also at times of the day when it is above the daily average. This is demonstrated by Figure 9-13 which shows a typical PV yield average monthly generation profile, alongside carbon intensity values calculated between the hours of 09:00 and 17:00 hours, normalised for an annual average intensity of 500g/kWh. Aggregating the monthly carbon savings using monthly '09:00 to 17:00' carbon intensities, and average monthly PV yields, the annual carbon saving delivers a value 523 g/kWh, 4.3% higher than the annual average intensity.

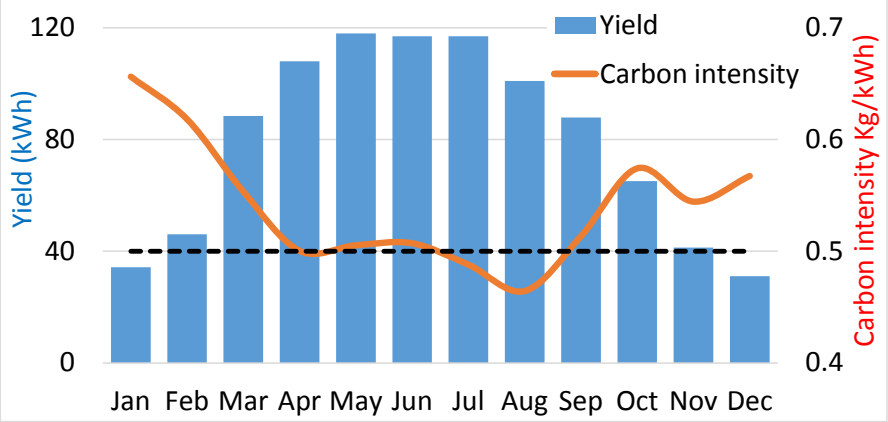


Figure 9-13 Typical monthly specific yield and average monthly carbon intensity between 9:00 and 17:00 hours, normalised to an average annual carbon intensity of 500 g/kWh

To account for this, a correction factor was introduced into the equation to calculate the carbon savings (Equation 9.3).

$$C = 1.043 \cdot I \cdot Y$$ *Equation 9-3*

The resultant component for carbon emission reductions was constructed in Netica and connected to the PV yield node (Figure 9-14). Note that a carbon intensity must be selected for the simulation – it is meaningless to retain a uniform distribution for all available intensities.

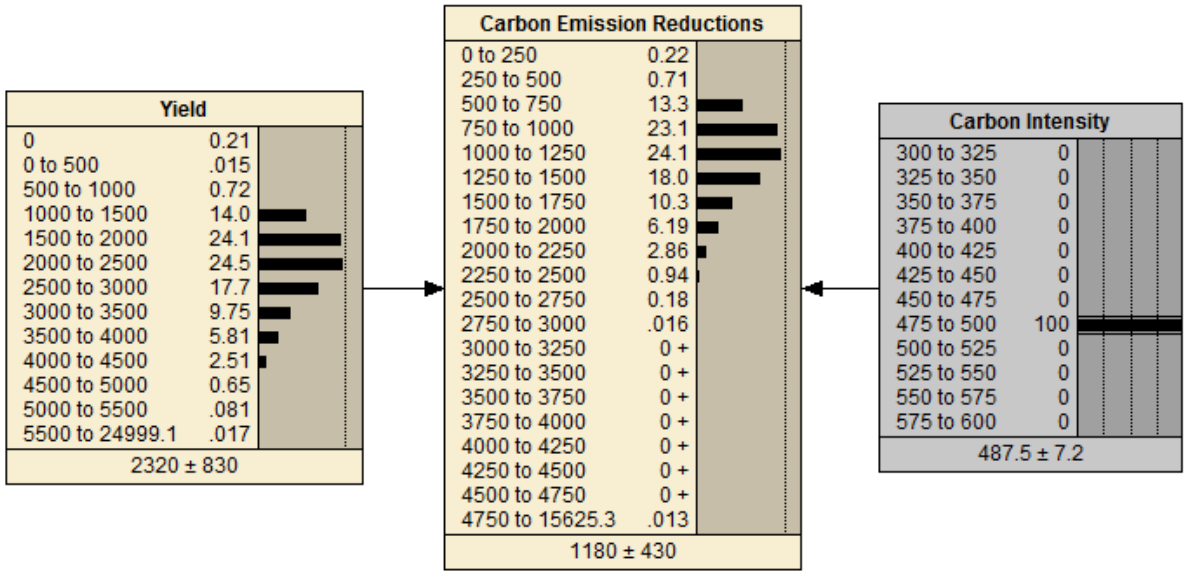


Figure 9-14 Deterministic BN model to predict carbon emission savings, influenced by the carbon intensity of the UK electricity grid, and the PV system yield.

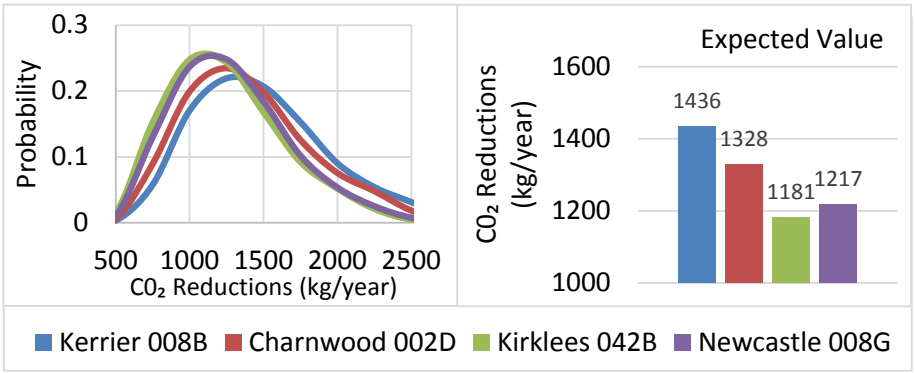


Figure 9-15 Comparison of carbon emission reductions for each LSOA

Figure 9-15 shows a comparison of the expected value for CO_2 reduction for the LSOAs and the posterior distributions with a carbon intensity of 475 to 500 g/kWh/year. Expected values range from 1181 to 1436 kg/year.

9.5.2 Summary of carbon savings

It has been demonstrated how a deterministic sub-model can be added to the OOBN for the purpose of creating a carbon reduction indicator. This takes the PV yield as a probabilistic input variable, and a fixed value for the carbon intensity for grid electricity. Industry standard grid intensity should be corrected for the temporal variability, both daily and seasonal. This analysis suggests that carbon savings with the current energy basket are 4.3% higher than otherwise predicted. Since carbon savings are proportional to the yield, it exhibits the same variability and dependencies as discussed in section 9-3.

9.6 Techno-economics

The second indicator integrated with the model is discounted cash flow analysis (DCFA). This method of asset valuation was documented by Fisher (1930) and has been used to inform investments in renewable energy deployment (Short et al, 2005), including the techno-economic assessment of microgeneration projects (Wood and Rowley, 2011), and as a decision support criterion for renewable technologies (Azzopardi et al, 2013). DCFA therefore, has been chosen to demonstrate the creation of a probabilistic techno-economic indicator. This section presents the derivation of the algorithm used and shows how the probabilistic parameters required are introduced from the OOBN.

9.6.1 Discounted Cash Flow Analysis (DCFA)

Future cash flows are discounted to deliver their sum (net value) in the present day, termed the net present value (NPV), using Equation 9-4.

$$NPV = \sum_{n=0}^{\lambda} \frac{V_n}{(1+i)^n} \qquad \textit{Equation 9-4}$$

V_n is the net cash flow at time interval n and i is the discount rate which is the interest rate at which an alternative method of investing the initial sum, V_0 could accrue value, and λ is the lifetime of the investment project.

Negative cash flows include the large initial ($n = 0$) capital expenditure, C_0, and subsequent expenditure, during interval, n, for maintenance, M_n, and repair, R_n over its estimated lifetime. A final decommissioning expenditure, D_λ may be incurred. Positive cash flows arise from the monetisation of generated energy, E_n, during each interval. Thus the net cash flow in interval n is given by Equation 9-5.

$$V_n = E_n - M_n - R_n \qquad \textit{Equation 9-5}$$

Under the FiT E_n is the sum of the value of generated electricity G_n, exported electricity X_n, and avoided imported electricity, A_n (Equation 9-6).

$$E_n = G_n + X_n + A_n \qquad \textit{Equation 9-6}$$

Under the UK's subsidy regime their values, in year n, are given in Equations 9-7 to 9-9 where Y_n is the annual yield, F_{G_n} is the generation tariff, F_{X_n} is the export tariff, and T_{E_n} is the electricity tariff during interval n. S is the self-consumption fraction.

$$G_n = Y_n \cdot F_{G_n} \qquad \text{Equation 9-7}$$

$$X_n = \frac{Y_n}{2} F_{X_n}$$ Equation 9-8

$$A_n = Y_n \cdot S \cdot T_{E_n}$$ Equation 9-9

To account for inflation, FiT rates are incremented commensurate with the annual Retail Price Index (RPI); this measures the percentage annual increase (inflation) in the price of consumer goods. The cost of domestic energy changes at a different (frequently faster) rate than other consumer goods, so economists use a distinct energy inflation rate (EIR) (Cucchiella et al., 2012). Equations 9-10 to 9-12 calculate the tariffs in year n, relative to year 0, assuming average inflation rates.

$$F_{G_n} = F_{G_0}(1 + RPI)^n$$ Equation 9-10

$$F_{X_n} = F_{X_0}(1 + RPI)^n$$ Equation 9-11

$$T_{E_n} = T_{E_0}(1 + EIR)^n$$ Equation 9-12

A further factor to consider is the degradation in performance of a PV system, d, over its lifetime (Jordan and Kurtz, 2011). Assuming a yield Y_0 in the first year of operation, the yield in year n is given by Equation 9-13.

$$Y_n = Y_0(1 - d)^n$$ Equation 9-13

Substituting Equations 9-7 to 9-12 into Equation 9-6 yields Equation 9-14, which is the income from monetised electricity generation in year n.

$$E_n = Y_0(1 - d)^n \left(F_{G_0} + \frac{F_{X_0}}{2}\right)(1 + RPI)^n + Y_0(1 - d)^n T_{E_0} S(1 + EIR)^n$$ Equation 9-14

Assuming no additional expenditures for repair not covered by the warranty, and neglecting decommissioning costs, the total NPV is given by Equation 9-15.

$$NPV = \sum_{n=0}^{\lambda} \left(Y_0 \left(F_{G_0} + \frac{F_{X_0}}{2} \right) \alpha^n + Y_0 T_{E_0} S \beta^n \right) - C_0 \qquad \text{Equation 9-15}$$

Where

$$\alpha = \frac{(1-d)(1+RPI)}{(1+i)} \qquad \text{Equation 9-16}$$

$$\beta = \frac{(1-d)(1+EIR)}{(1+i)} \qquad \text{Equation 9-17}$$

If α and β are assumed to be constant over the lifetime of the technology then Equation 9-15 is the sum of two geometric progressions. These can be simplified using Equation 9-18 (Riley et al., 2006).

$$\sum_{n=0}^{\lambda-1} cr^n = c\left(\frac{1-r^{\lambda}}{1-r}\right) \qquad \text{Equation 9-18}$$

The initial capital expenditure,C_0, is replaced by the system rating R multiplied by the installation cost per unit rating C_u. Thus replacing C_0 and substituting the sums of the geometric progressions of α and β i to Equation 9-15 gives Equation 9-19.

$$NPV = Y_0 \left(F_{G_0} + \frac{F_{X_0}}{2} \right) \left(\frac{1-\alpha^{\lambda}}{1-\alpha} \right) + Y_0 T_{E_0} S \left(\frac{1-\beta^{\lambda}}{1-\beta} \right) - R \cdot C_u \qquad \text{Equation 9-19}$$

Equation 9-19 has 12 parameters which are summarised in Table 9-1. The first three, self-consumption fraction, the initial system yield, and the system rating, are taken as probabilistic inputs from nodes in the OOBN. In the following sections values for the other parameters will be extracted from the literature and public sources.

Table 9-1 Parameters for NPV calculation in Equation 9-19

Variable	Name	Units	Type
S	Self-consumption fraction	Fraction (per year)	Probabilistic input
Y_0	Initial system yield	kWh/year	
R	System rating	kW_p	
d	Annual degradation rate	Fraction (per year)	Probabilistic marginal
C_u	System cost	£/ kW_p	Constant
RPI	Retail price index	Fraction (per year)	
EIR	Energy inflation rate	Fraction (per year)	
i	Discount rate	Fraction (per year)	
F_{G_0}	Initial fit generation tariff	£	
F_{X_0}	Initial fit export tariff	£	
T_{E_0}	Initial electricity tariff	£	
$\lambda - 1$	Life time of technology	Years	

9.6.2 Annual Degradation

To reflect the impairment of performance resulting in a diminished yield over time from the initial value Y_0 it is important to include the PV module degradation rate in the techno-economic analysis of PV (Darling et al, 2011). Jordan and Kurtz (2013) have conducted a review in which almost 2000 long-term degradation rates for modules or entire systems were assessed to produce the frequency distribution shown in Figure 9-16. The average and median values for this analysis were 0.8 and 0.5 %/year respectively. Darling et al (opt cit.) used a qualitatively similar distribution to carry out levelised cost of energy calculations for solar PV using a Monte Carlo approach.

Whilst an initial rapid light-induced deterioration of yield over the first few days of exposure is documented in the literature (Dunlop, 2003; Kroposki and Hansen, 1997), the assumption that the annual degradation rate is a gradual process is supported by observations of year-on-year degradation, as opposed to catastrophic failures. This long-term gradual decrease in efficiency occurs due to a number of degradation processes caused by thermal and mechanical shocks, and physico-chemical changes, which result in physical damage to module components, and corrosion following humidity ingress (Kaplanis and Kaplani, 2011).

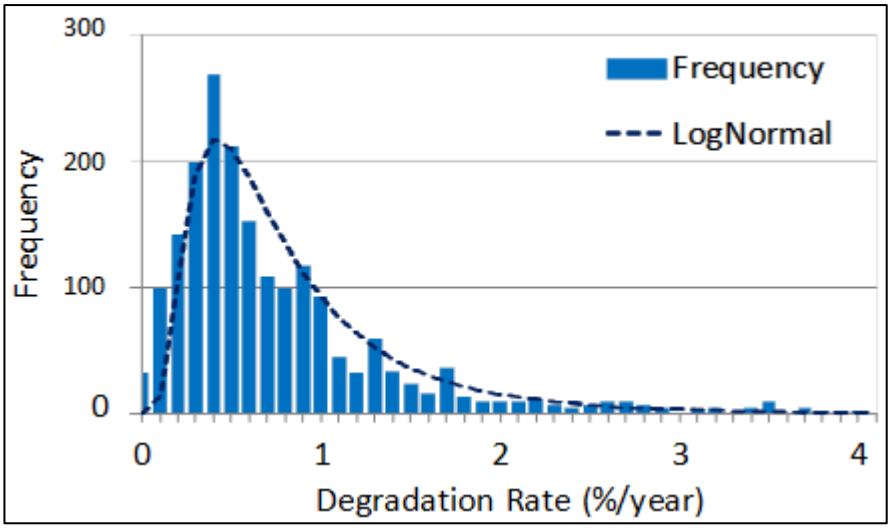


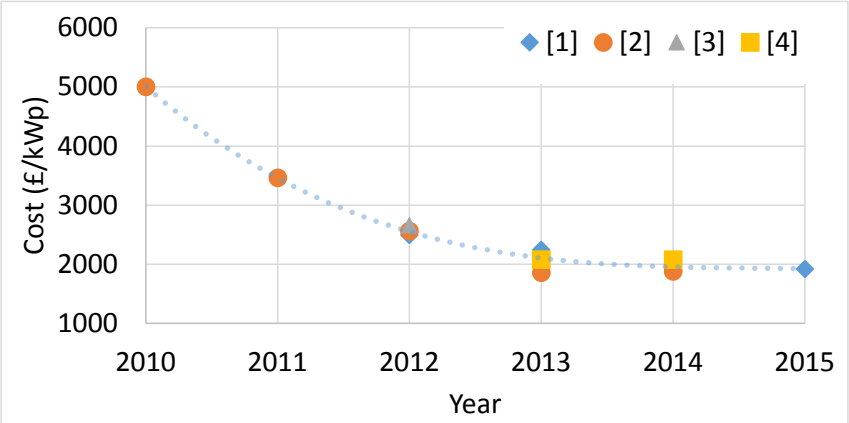
Figure 9-16 Frequency distribution of degradation rates after Jordan and Kurtz (2013)

Equation 9-19 assumes a geometric degradation rate, whereas cited rates are usually assumed to be linear. This allows a simpler formula to be used (based upon the sum of a geometric progression) rather than that based upon a more complex arithmetico-geometric series. However, the discrepancy between a geometric and an arithmetic (linear) degradation is only 2.2% after 20 years, at an annual degradation rate of 1%. The majority of reported degradation rates are less than this, clustered around a value of 0.5%, at which this discrepancy falls to only 0.5%. Thus, given the intrinsic uncertainty in degradation rates, a geometric degradation rate was assumed and this has been incorporated into the composite discount factors, as represented by Equations 9-16 and 9-17.

9.6.3 System Costs

The cost of domestic solar PV reduced rapidly from 2010, when the UK FiT scheme commenced, from typically £5000, to less than £2000 per kWp in 2015 (Figure 9-17). The inherent variability of prices by supplier/installers results in a natural distribution of installed cost which has not been quantified in this work due to a lack of available data. However, variability is also introduced by the

way prices are determined which includes some costs that are independent of system rating, (for example scaffolding costs), thus leading to the notion of fixed and marginal costs (Parsons Brinkenhoff, 2012). The effect of this is to render the price per kWp for smaller systems higher than that for larger systems.



[1] Parsons Brinckerhoff
[2] Green business watch
[3] Parsons Brinckerhoff (2)
[4] DECC

Figure 9-17 Average cost of capital expenditure costs of UK solar PV system between 2010 and 2015 from public sources

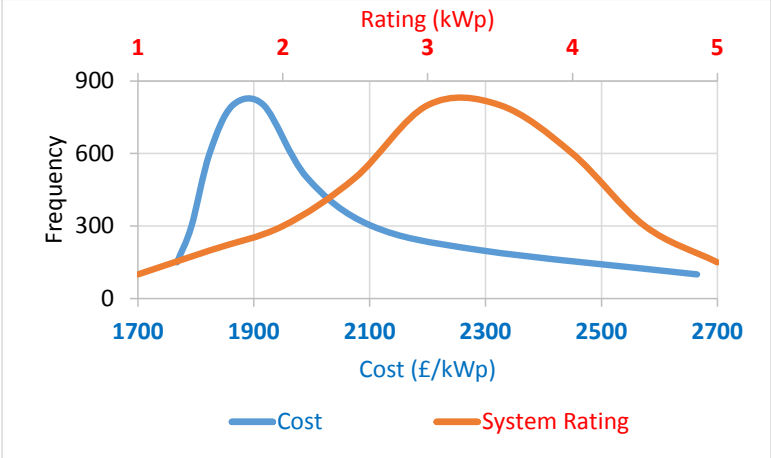


Figure 9-18 The distribution of cost per kWp for an empirical distribution of UK PV ratings based on a fixed cost of £1122 and a marginal cost of £1543 for 2014/15 (After Parsons and Brinckerhoff, 2012)

With the observed distribution of system ratings skewed towards the larger systems in the band, this results in a positively skewed distribution for cost per kWp (Figure 9-18). Thus it is necessary to include both fixed and marginal costs for domestic PV systems in order to account for the higher fixed costs per kWp for smaller systems.

9.6.4 Retail Price Index (RPI) and Energy Inflation Rate (EIR)

Both the RPI and the EIR are integral components of the discount factor used in in Equation 9-19. The former is the UK government's preferred method of incrementing FiT tariffs each year to account for inflation and thus maintain the value of the incentive over its 20 year duration (Cherrington et al., 2013). The RPI consists of a composite index which measures price variations for a wide range of consumer items. Since 2010, the RPI applied to increment FiT tariff has ranged from 4.8% to 1.6%, whilst over the same period the electricity component of the RPI, which measures the percentage change in the price of domestic electricity, has fluctuated from 10.6% to 4.7% (ONS, 2015). It has been extremely volatile over the past 25 years, subject to rapid reductions and negative values corresponding to price reductions during the 1990s (Helm, 2002), and a rapid increase between 2003 and 2008 (Figure 9-19). Equation 9-19 utilises an average figure for the period of interest, an assumption also made by other researchers (Cucchiella et al., 2012). The average RPI between 2014 and 1988 is 3.5% with a standard deviation of 2.0%, whilst the average EIR is 4.3% with a standard deviation of 6.1%.

The use of an average value, whilst common practice in DCFA, is problematic; ideally the actual year-on-year inflationary value (RPI or EIR) in each year, rather than an average value over the lifetime of the calculation, should be used. The error that this introduces into Equation 9-19 can be deduced over the range of values for the RPI and EIR exhibited in Figure 9-19. This is given by the difference between the result obtained using year-on-year values and calculating the sum of the series

(Equation 9-20), and that obtained using an average value as a geometric factor and the formula for the sum of a geometric progression (Equation 9-21). In these equations I_n is the RPI, or EIR, in year n, and $\bar{I}$ is the average of the index over the period of λ years. A Monte Carlo simulation using a randomised sequence of values for the period 1988 to 2014 gave a mean standard error of 30% for the EIR (standard deviation 15%), the positive value signifying that the exact method using Equation 9-20 is higher. For the RPI the standard error was only 4% (standard deviation 4%).

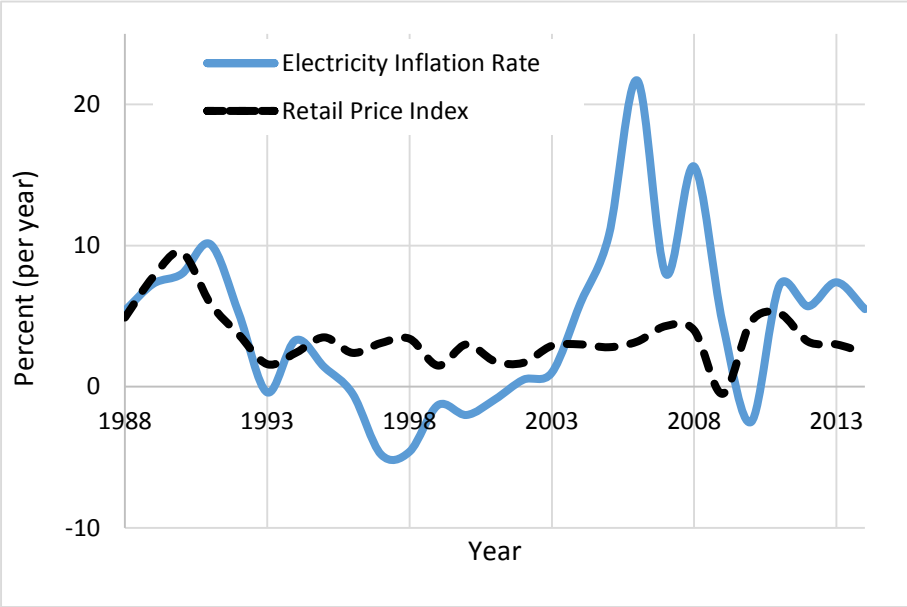


Figure 9-19 ONS data on RPI and Electricity inflation rate (EIR) between 1988 and 2014)

In practice, using Equation 9-19 requires the use of an estimated average value as other researchers have done, however it has been shown here that the factor due to the inflation may be significantly out by a factor of 30% in periods of high price volatility.

$$\sum_{n=0}^{\lambda}(A(1-I_n)^n) \qquad \text{Equation 9-20}$$

$$\frac{1-(1+\bar{I})^{\lambda}}{\bar{I}}$$ Equation 9-21

9.6.5 Generation, Export and Electricity Tariffs

There are three tariffs used in Equation 9-19, namely the initial generation, the export, and the retail electricity tariff respectively. The FiT tariffs have undergone significant reductions commensurate with the significant PV system cost reductions which have occurred since 2010 (Figure 9-20) (OFGEM, 2015). For systems up to 4kWp, the generation tariff reduced from 43.3p/kWh in 2010 to 21p/kWh in March 2012, with another reduction to 16.0p/kWh only 5 months later. At this juncture, a more responsive approach to degression - a systematised quarterly reduction in tariffs – was introduced, which allows accurate prediction of tariff reductions as long as deployment targets have been met. If deployment levels are low then the tariff reductions may be skipped for up to two quarters. The current generation and export tariffs are 13.4p/kWh and 4.85p/kWh respectively for systems installed on or after 1 April 2015.

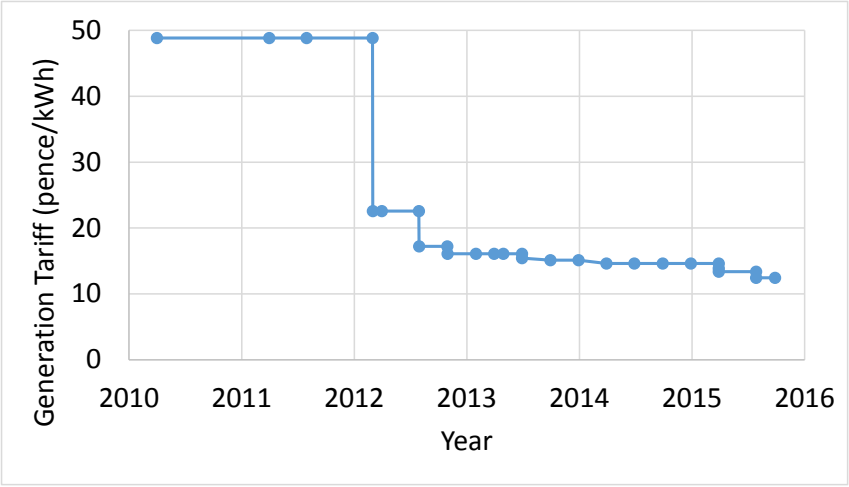


Figure 9-20 PV FiT rate for <4kWp system for EPC Grade D retrofit at the 2015/16 values i.e. RPI corrected.

Since the implementation of the FiT scheme, average retail electricity prices have increased from 12.6p/kWh in 2010 to 17.5p in 2014. As described in the previous section, fuel prices are volatile and therefore likely to be highly uncertain going forward. Furthermore, unlike the FiT rates, electricity tariffs are subject to significant market uncertainties. For example, in terms of available tariffs, these might be lower cost long-term contracts, or high cost card meter consumers. Thus the self-consumption contribution to the financial impact is subject to further uncertainty, and therefore the OOBN model assumes a constant value, whilst allowing this to be varied to explore a range of electricity cost scenarios.

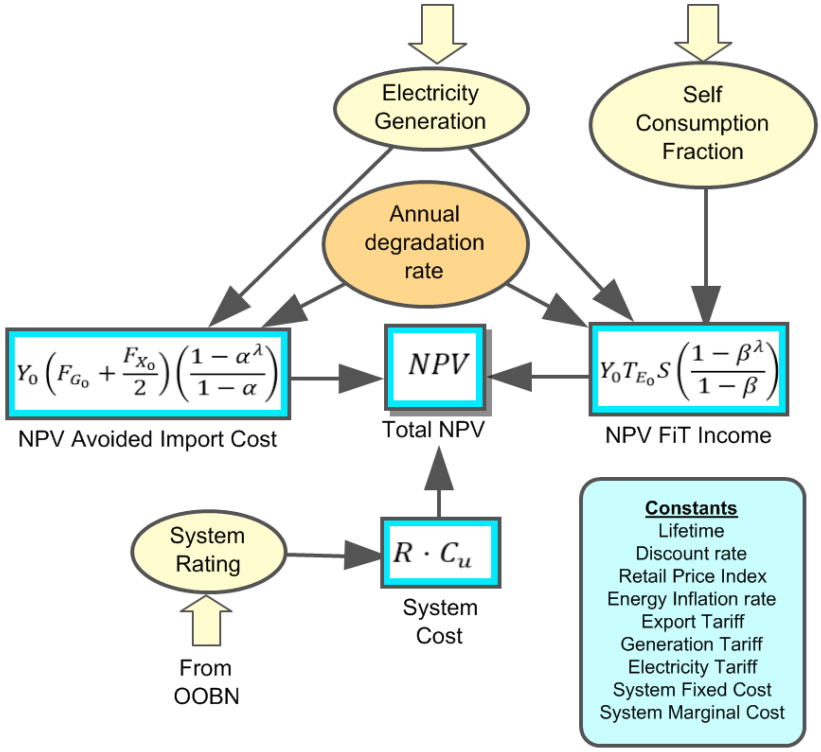


Figure 9-21 BN sub-model to calculate net present value showing the deterministic nodes with their defined equations and the interface nodes which connect to the rest of the model.

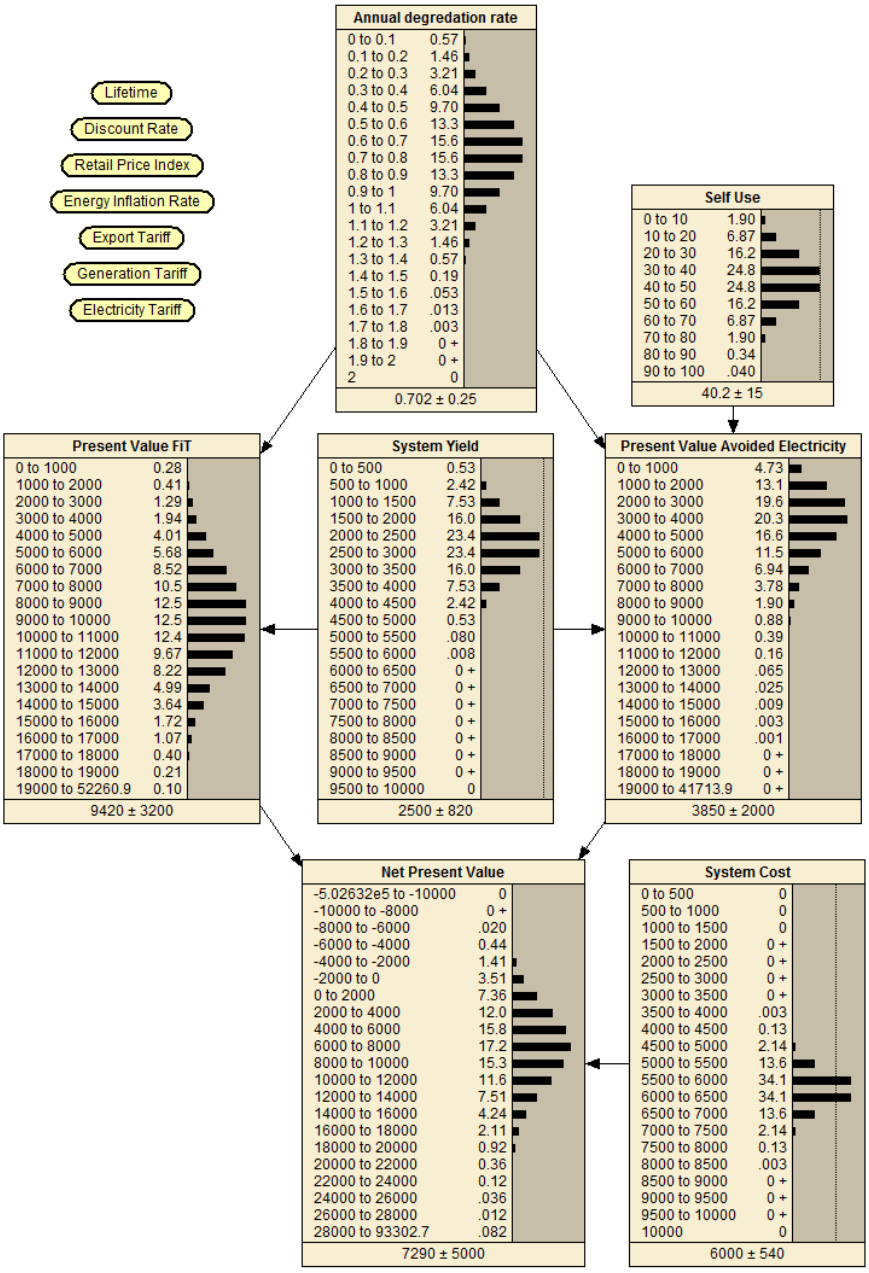


Figure 9-22 Bayesian network sub-model for net present value calculations

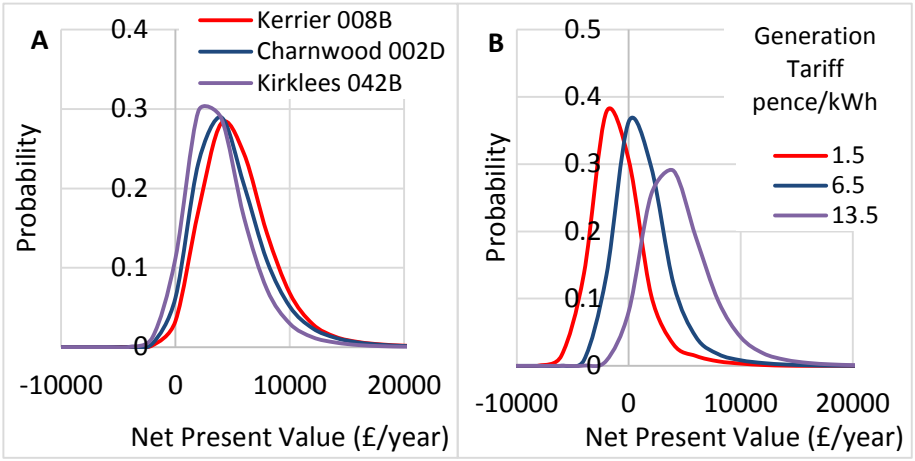
9.6.6 Net Present Value BN Sub-model

Equation 9-19 can be rationalised into three components as shown in Figure 9-21. This shows three key parameters as probabilistic inputs delivered from the OOBN model (self-consumption fraction, electricity generation and system rating). The annual degradation rate is represented as a marginal

distribution and the remaining values are fixed constants. This representation shows the contributions of the initial capital outlay, the FiT subsidy and the avoided import costs to the overall NPV. The BN representation of this model in Netica is shown in Figure 9-22.

9.6.7 Summary of Techno-Economics

Figure 9-23A shows the posterior NPV distributions for three census areas used in this study. The constant parameters (see Table 9-1) were given the values shown in Table 9-2.



A: Using marginal distributions for LSOA building stock parameters and constants as in Table 9-2. B: For all census areas but using the given scenarios for the FiT PV Generation Tariff

Figure 9-23 Net Present Value distributions

The uncertainties inherent in the system rating, PV yield, electricity consumption and self-consumption, as endogenised in the various BN sub-models which constitute the OOBN, are propagated into the NPV BN sub-model to deliver a realistic uncertainty in the value of total NPV. The median NPV value for LSOA Kerrier 008B is circa £4100, with an interquartile range from £2300 to £6250. The top decile of systems would attain an NPV greater than £8700 whilst the lowest decile less than £750. It is observed that LSOA Kerrier 008B, as the more southerly area modelled, delivers

the most favourable NPV due to higher irradiance, but it can be seen that here, as in the other areas, the risk of a low return is significant. The NPV is seen to be sensitive to building stock parameters. Thus, the influence of orientation shows a difference in NPV of £2000 between systems facing East or West compared to the optimal azimuth (due South). Notably, the system degradation rate has a significant influence, with a 10% reduction in NPV when comparing a relatively conservative 0.1% annual degradation rate to a degradation rate of 1%.

Table 9-2 Value for the constant parameters used to generate NPV distributions in Figure 9-23A

Constant	**Value**
Discount Rate	3.5%/year
Electricity Inflation Rate	10%/year
Electricity Tariff	£0.18 /kWh
Generation Tariff	£0.135 /kWh
Export Tariff	£0.05 /kWh
Fixed Cost	£1122
Marginal Cost	£1543 /kWp
Lifetime	20 years
Retail Price Index	3%/year

The model also allows the variation of constants built into the model[39]. Thus Figure 9-23B shows the impact on NPV of different generation tariffs for all the census areas modelled in the study. This shows that under less generous generation tariffs, whilst maintaining current system costs and levels of self-consumption, tariffs any lower than the current £0.135 would subject a high percentage of PV adopters to a severe risk of having no economic return at the social discount rate of 3.5%. The low

[39] This sounds like an oxymoron; ideally constants which were desired to vary could be established as nodes (variables) in their own right and thus simply allowing the selection of a state. However the downside is that this rapidly increases the computer memory requirement for the whole model.

value of £0.015 was recently proposed by the UK Government's review of its FiT subsidy regime for domestic PV systems less than 10kWp.

It has been demonstrated that using a probabilistic model for calculation, a wide variation in NPV can in theory be realised. This distribution can be explored under a variety of different cases for input parameters. The large uncertainty in investment has been demonstrated.

9.7 Fuel Affordability

The third indicator, for which a sub-model adjunct to the OOBN has been constructed, predicts the impact of domestic solar PV on domestic energy expenditure. By comparing energy expenditure with household income, both with and without PV, its impact on the domestic economy can be ascertained. As discussed in Chapter 1, fuel poverty has been routinely estimated using the quotient of energy spending required for the maintenance of adequate thermal comfort, to household income, though this measure has been superseded by the high cost low income indicator (Hill, 2012).

Both gas and electricity consumption have been probabilistically predicted from the building stock model, and using IPF, equivalised household income has been interlocated into the appropriate CPT. Thus the OOBN encapsulates dependency relationships between income, building attributes and energy consumption. Using published energy and feed-in tariffs as constants, and the posterior distributions for gas and electricity consumption, PV system yield, self-consumption and household income from the OOBN as inputs, a deterministic BN model can be constructed to predict the fuel spend, and the ratio of spend to income – an energy affordability impact indicator.

Table 9-3 Parameters for Fuel spend (Equation 9-22) and fuel affordability (Equation 9-23)

Variable	Name	Unit	Type
F_S	Fuel Spend	£/Year	Probabilistic Output
G	Gas consumption	kWh/year	Probabilistic input
T_G	Gas tariff	£/kWh	Constant
E	Electricity consumption	kWh/year	Probabilistic input
T_E	Electricity tariff	£/kWh	Constant
Y	System Yield	kWh/year	Probabilistic input
F_G	FiT generation tariff	£/kWh	Constant
F_X	FiT export tariff	£/kWh	Constant
S	Self-consumption	kWh/year	Probabilistic input
F_A	Fuel Affordability	*(ratio)*	Probabilistic Output
I	Household income	£/year	Probabilistic input

Using the terms defined in Table 9-3, the household fuel spend can be aggregated from three components (similar to the discounted cash flow analysis in Section 9.6, though here, gas consumption is included). Firstly, the total spending on imported energy is given by Equation 9-22.

$$Spending\ on\ Imported\ Energy = (G \times T_G) + (E \times T_E) \qquad \textit{Equation 9-22}$$

Secondly, there is a subsidy from the FiT, from both the total generation and the deemed export (Equation 9-23), and finally there is the avoided cost of electricity imports due to direct self-consumption given by Equation 9-24.

$$\text{Feed in Tariff Income} = \left(Y \times \left(F_G + \frac{F_X}{2}\right)\right) \qquad \textit{Equation 9-23}$$

$$\text{Avoided electricity cost} = (S \times T_E) \qquad \textit{Equation 9-24}$$

Combining these three equations yields the total impact on domestic energy spending (Equation 9-25). The energy affordability is given by the ratio in Equation 9-26.

$$F_S = ((G \times T_G) + (E \times T_E)) - \left(Y \times \left(F_G + \frac{F_X}{2}\right)\right) - (S \times T_E) \qquad \textit{Equation 9-25}$$

$$F_A = \frac{F_S}{I}$$ *Equation 9-26*

9.7.1 Fuel Affordability Netica Sub-model

Equations 9-25 and 9-26 were represented as a BN in Netica (Figure 9-24). Each of the three components: imported energy spend, FiT income, and avoided electricity costs, were represented by a node. This enables comparisons to be made between the income streams for PV. The three components are combined in the total final fuel spend node. The total fuel spend is divided by the household income to deliver an energy affordability indicator; the model also displays a node to display the probability of a household spending more than 10% of its income on energy.

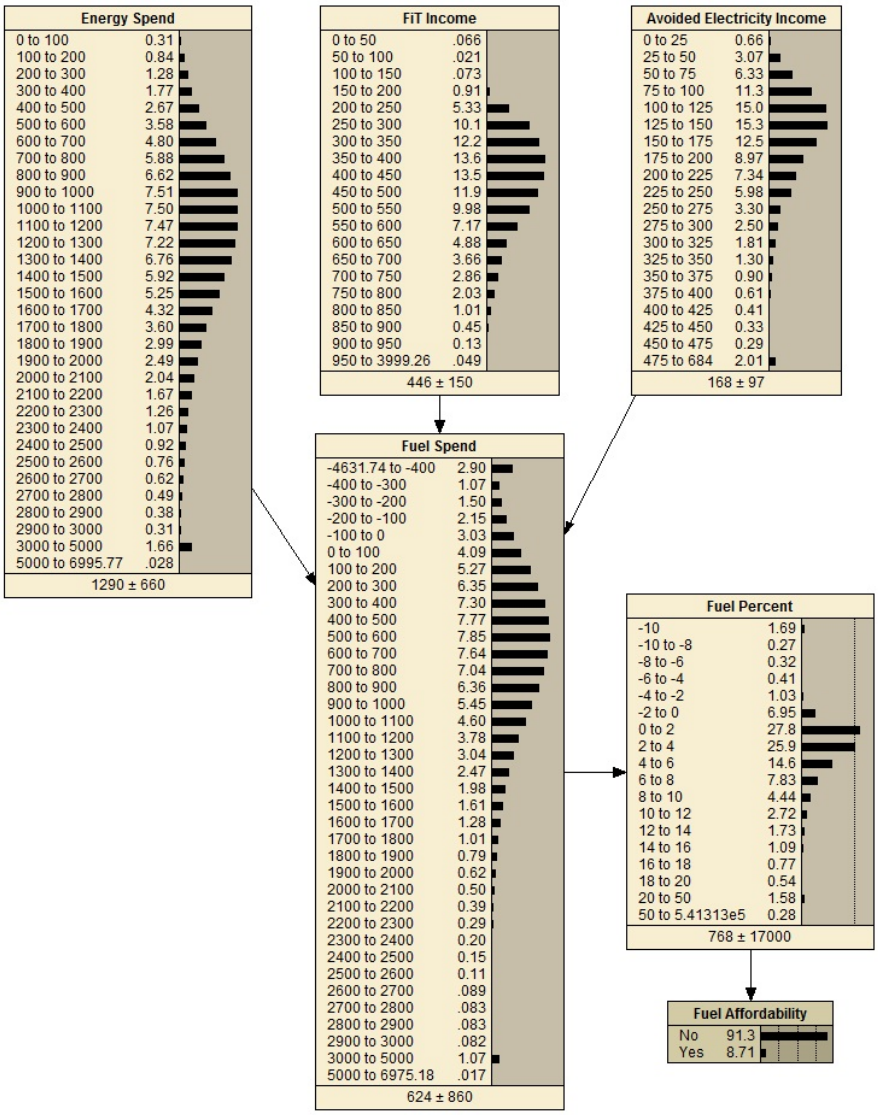


Figure 9-24 Bayesian network sub-model for fuel affordability calculations showing the actual spending on fuel (Fuel Spend) after the benefit of FiT income and avoided electricity costs have been subtracted. The percentage of income spent on fuel (Fuel percent) is presented.

9.7.2 Results from the Energy Affordability Netica Sub-model

Figure 9-24 gives an indication of the posterior distributions for each of the variables used to calculate the spending on gas and electricity, the returns from the FiT payments, and savings due to avoided electricity costs. The resultant distribution of household fuel spend on gas and electricity, both before and after the financial benefits of PV have been subtracted, are shown in Figure 9-25.

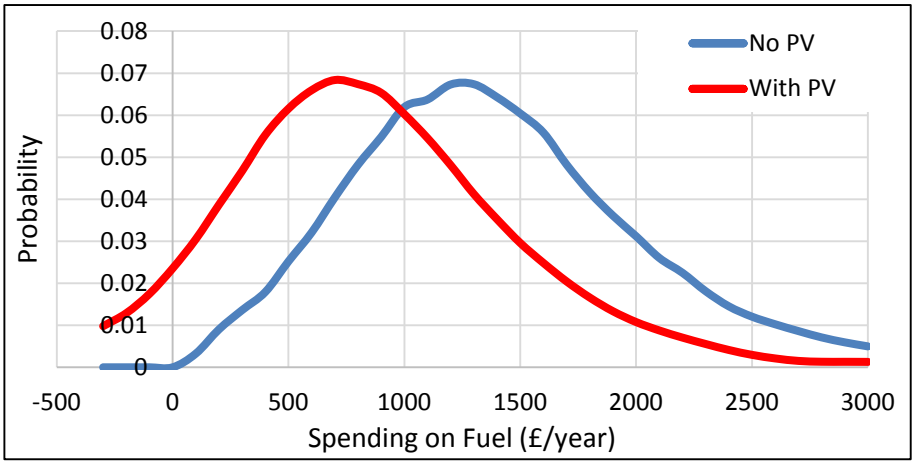


Figure 9-25 Prior distribution of aggregated household fuel spending (F_S) on gas and electricity per year for all four census areas before the financial returns of PV are subtracted (No PV), and after the financial returns have been subtracted (With PV)

Note that thus for this ignores the cost of the investment and assumes that the householder in in receipt of all the FiT payments. Figure 9-26 shows the expected values for these distributions for each census area. The required energy spend is what households would spend on grid electricity and gas as predicted by the energy demand model. This shows LSOA Kirklees 042B housing stock with the highest energy costs, and LSOA Kerrier 008B the lowest. Fit income varies between £372 and £446 per year. The avoided electricity saving is highest in LSOA Kerrier 008B, ranging from £168, to £149 in LSOA Newcastle 008G. The FiT income and avoided electricity saving, when subtracted from the required energy spend, gives the actual energy spend. The saving on the household bills is of the order of 52% for LSOA Kerrier 008B, and 36% for LSOA Kirklees 042B.

Each of these parameters has its own distribution for each LSOA; there is no correlation since each distribution is predicted by the diverse building stock and geographic factors in each area. Thus energy spending is high in the more northerly LSOA Kirklees 042B, where properties are relatively large, compared to Cornwall, which has more modern smaller dwellings. Cornwall in contrast benefits from higher irradiance and commensurate FiT returns.

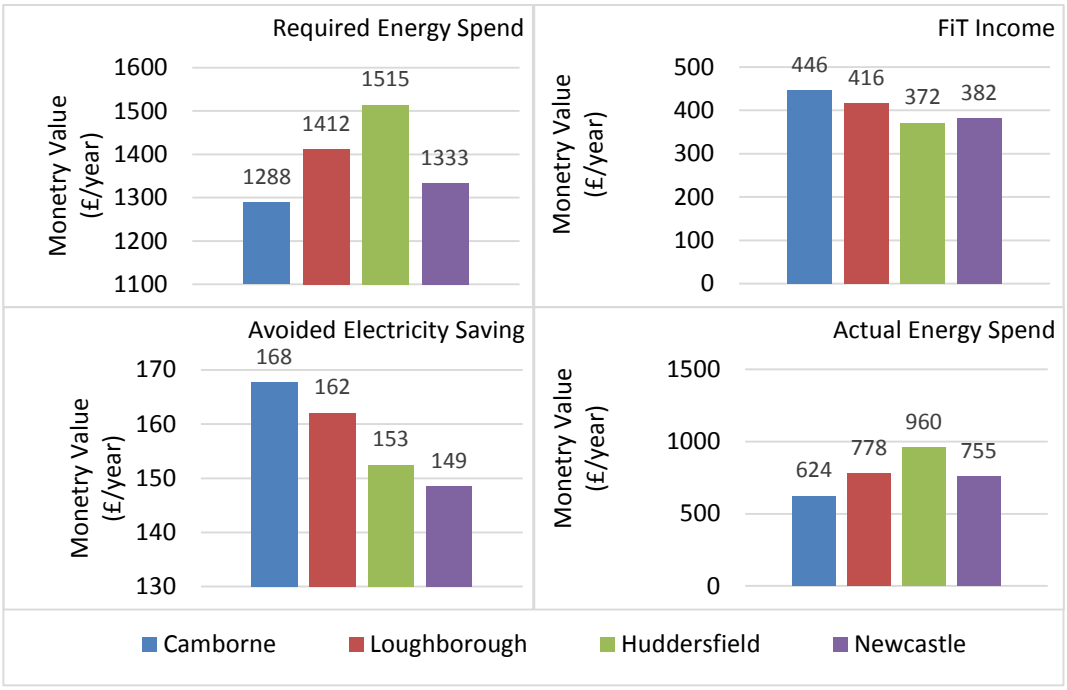


Figure 9-26 Expected value for the monetary value (£/year) for the required energy spend, FiT income, avoided electricity saving and actual energy spend for each census area

The fuel affordability (F_S) has been defined here as the ratio of spending on energy services to the household income. This delivers the fuel percent ratio. Figure 9-27 compares the expected values for this ratio. Thus, in Cornwall the average fuel spend to household income ratio is 8.1% without PV, dropping to 4.1% with PV installed. LSOA Newcastle 008G, having the lowest incomes (see Chapter 7) has the highest ratio at 11.9%, dropping to 7.1% with PV.

A fuel affordability benchmark has been defined here as the percentage of households having a fuel affordability ration of more than 10%. The upper two charts in Figure 9-27 show how this index is impacted by the installation of PV. In LSOA Newcastle 008G, there is a 41% probability of exceeding a fuel affordability of 10%, which halves to 21% if a property has PV. Similar 50% reductions occur for the other LSOAs in the study though the initial fuel affordability is not as high due to higher incomes in these areas.

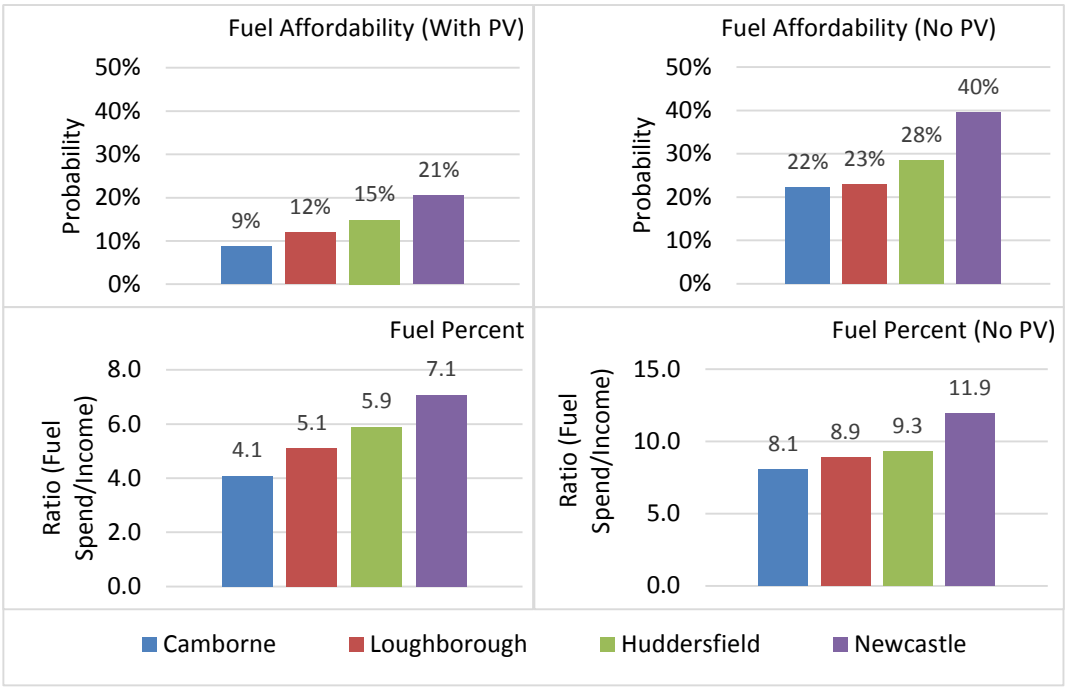


Figure 9-27 Fuel affordability index with and without PV, and the expected value for 'fuel percent', with and without PV for each census area

9.7.3 Summary of the Fuel Affordability

Spatially disaggregated empirical energy demand and household income have been modelled by the OOBN to provide probabilistic indicators which give the absolute spending, and the percentage of income spent on fuel, both with and without PV. Official fuel poverty indicators use a modelled energy demand base calculated using a normative heating regime. Since UK households are generally not heated to the same intensity (Shipworth et al., 2010), the official fuel poverty is generally higher than our proxy indicator might suggests. Nevertheless, our probabilistic approach provides a useful spatially disaggregated proxy indictor which can help with the targeting of mitigation interventions.

9.8 Discussion and Conclusion

This chapter describes the construction of an integrated OOBN from four purposefully designed BN sub-models. This resultant OOBN model (Figure 9-4) exposes a variety of nodes which can be used as probabilistic inputs to deterministic models which can be automatically converted into probabilistic BN models. Three such indicator models have been demonstrated: domestic fuel affordability, carbon emission savings, and discount cash flow analysis (DCFA), each as representative social, environmental and economic indicators, respectively. All the dependencies between the variables are represented in the OOBN, thus conditional probability tables reflect the influence of variables on each other in a manner consistent with empirical observations.

A key utility of the OOBN is the facility to enter hard or soft evidence (observations) for one or more parameters in order to constrain the model, and to support detailed analysis, such as that for a more localised assessment. The uncertainty of the remaining parameters is propagated to, and reflected in, the target variables. Thus, a variety of scenarios may be rapidly evaluated in terms of impact upon energy affordability, carbon emissions and discounted cash flow output.

As well as the application of evidence to a single input node, observations can be applied to multiple nodes. For example, a low self-consumption parameter, combined with an easterly facing array and a 1% degradation rate, resulted in an NPV of approximately 30% less than that for a system with more optimal characteristics.

In this analysis the marginal distributions of nodes of key energy parameters have been described and displayed in charts for each LSOA. This is a key paradigm shift in modelling; results are not presented as singular values with an uncertainty range. Rather, as in Figures 9-23 and 9-25, a parameter is presented as posterior distribution given some initial observations or a prior distribution when all other variables are in their initial states. This distribution endogenises the uncertainty of a target variable given all the initial uncertainties of the input variables.

As well as the individual reported results there are a myriad of scenarios which could be established in the model. For example, one might be interested in a use case where only terraced houses are to be considered, or only South facing dwellings. The model itself therefore, is a key result which permits all of these scenario permutations to be explored. As such it is a holistic knowledge representation of domestic PV in the four LSOA case study areas embedded within the OOBN.

There are of course limitations, as discussed in the individual chapters and sections in which the sub-models are described. The OOBN itself warrants further philosophical discussion as to its ontological accuracy and the resulting epistemic utility. Thus, just as Chapter 2 set out a research gap for probabilistic whole system modelling for renewable energy deployment, the Chapter 10 reflects upon the ideas therein, and whether this research has answered some of the key questions.

10 Conclusions and Further Work

10.1 Introduction

Discussion and conclusions have been presented in specific chapters of this thesis for the various knowledge domains which have been encapsulated into the constituent components of the OOBN and the integrated model (Chapter 9). In this Chapter, further concluding discussion is presented in Section 10.2 in the context of the research aims and objectives introduced in Chapter 1. Section 10.3 presents the overall conclusions of the thesis and summarises the contribution of the research. Finally, suggestions for further work are discussed in Section 10.4.

10.2 Concluding Discussion

The aim of this work was to develop, apply and evaluate a whole system modelling approach which endogenises uncertainties for key performance indicators in the deployment of solar PV, in the context of UK communities. In Chapter 2 it was argued that there was a gap in current knowledge and literature, in that most studies were deterministic, and did not consider the variability of parameters. Something more than a simplistic sensitivity analysis was required, in order to fully quantify the risk and uncertainty. Before arguing that this aim has been achieved by this work, the delivery of the specific objectives (Section 1.5) which contribute towards it are discussed below.

Objective 1 required a number of KPIs to be integrated into the model. This has been realised by the development of an innovative method of taking what are essentially deterministic relationships for carbon intensity, discounted cash flow and fuel affordability, and representing these by probabilistic BN models. These are interfaced with the core OOBN to furnish the user with probabilistic outputs for the KPIs. The statistics of these output distributions: expected value, median, deciles, quartiles

and interquartile ranges, for example, are easily realisable from the model outputs to deliver objective indicators which can be compared, or, as demonstrated with fuel affordability, a binary indicator can be created, in this case using the 10% fuel spend to income ratio benchmark.

The second objective was to characterise the uncertainty in solar PV yield, self-consumption and electricity exports. For theoretical installations the method was to model the solar potential using building stock parameters and use one of many predictive models. However, this would not deliver the uncertainty observed in the real world. The solution to this, discussed in Chapter 6, and integration with the building stock model (Chapter 9), delivered the end result of deterministically predicted yields enhanced with realistic empirically derived uncertainty.

The quantification of self-consumption required a novel new approach due to the need to construct a BN which operated with annualised data but which also endogenised the uncertainties in self-consumption experienced on the one-minute timescale. The amassing of 30,000 years of minute resolved load and generation profiles for a wide range of aggregate consumption and generation values allowed the creation of a three way contingency table for annual consumption, generation and self-consumption with which to generate the CPT for this sub-model. This novel approach demonstrates how a BN model, required to work with annual data, can be populated by aggregating time series data of a much shorter temporal resolution in order to maintain a fidelity to empirical studies. An interesting facet of this method is that the simulations do not have to yield the empirical aggregated marginal distributions of generation and consumption, but only need to furnish the model with the conditional probabilities. The required empirical distributions of electricity consumption and generation are then furnished by the PV yield and building energy demand sub-models which have been constructed using annualised generation and consumption data respectively. Thus, armed with annual data only, the model is able to predict the distribution of annual self-consumption despite a theoretical requirement to derive this at a very low temporal resolution.

The contribution of solar PV to the total domestic energy consumption required other energy vectors to be considered. In this study the focus has been on urban areas with predominantly dual fuel (gas and electricity) or electricity-only supplied dwellings. The potential presence of solid fuel, heat pumps and biomass heating has been neglected. Furthermore, the empirical distributions of dual fuel and electricity only households encapsulated in the OOBN model uses NEED data which is spatially resolved at the regional scale, whereas this analysis focuses on distinct LSOA census areas. Thus the marginal distributions for gas consumption and electricity at the regional level might not be representative of the actual LSOAs.

A comparison of the simulated Cambridge Housing Model's gas and electricity consumption with the empirical consumption in the NEED dataset showed a significant difference in the probability distributions (Section 6.3.3); low and high demand households evident in the empirical data are unrepresented in the simulated demand data. Thus the integration of the NEED framework dataset into the model introduces empirical consumption patterns into the model. This makes the evaluation of the contribution of PV to the domestic energy demand more representative than if simulated annual demands had been used. This fulfils Objective 3.

In Chapter 9 the utility of the OOBN in assessing likely probabilistic impacts of PV on carbon reduction, low carbon technology investment, and domestic economics – the fourth objective - has been successfully demonstrated. It should be apparent to the reader that there are numerous potential results to chart or tabulate, based on scenarios for hard or probabilistic evidence, for any combination of key input variables. It is this notion that demonstrates that the substantive aim of the research, to create a whole system model which endogenises uncertainties for KPIs in order to evaluate the impact of PV, has been achieved. There are several underlying features of this approach which are pertinent.

Firstly, the model encapsulates a JPD which contains the knowledge about the probability of every possible permutation of all the discretised variables in the domain. Using the Chain Rule, the BN has

allowed the construction of this JPD component by component; the need for a singular large data set has been obviated – instead the relationships between clusters of variables have been modelled and these have been joined together through common parameters to construct a factorised JPD.

This joining together of components has delivered the whole-system model. Of course the whole-system was not defined at the outset with a clear system boundary; this was an open and moving boundary and other components of the system could have been added, for example to model other SEE indicators such as employment or levelised cost of energy. For those components included within the system boundary, a more expansive ontology could have been developed. For example, the simple building stock model (Chapter 6), built on three main predictor parameters (building age, built form and floor area), could have included other parameters such as heating system type and controls. However, this burdens the research with the need for even more detailed data and potential data fitting requirements. The positive aspect is that the uncertainty due to hidden or missing variables in the models is endogenised in the OOBN. Thus, given the known attributes of building age, built form and floor area, the model yields a distribution of fuel consumption. Heating system type, controls, and occupancy behaviours, are hidden variables but the uncertainty they represent is present in the probabilistic outputs. This applies to all the models; thus the PV yield model could contain system components such as module or inverter type. Again, the use of empirical data ensures that these hidden variables are reflected in the probabilistic outputs. Thus the research output is indeed a whole system model – within a certain system boundary, where epistemic and aleatory uncertainties are endogenised. The inclusion of representative SEE KPIs for PV deployment has also been demonstrated.

10.3 Conclusions

A whole system object oriented Bayesian network has been developed to model the energy balance of domestic dwellings in a UK context. This provides inputs into probabilistic models which furnish posterior distributions for carbon savings, discounted cash flow analysis, and fuel affordability as three representative social economic and environmental (SEE) indicators. The integration of knowledge across domains has been accomplished to create, in essence, a transdisciplinary knowledge representation consisting of building energy demand, applied solar PV, and building stock modelling. A probabilistic OOBN model has been developed which can predict the energy flows between the electricity grid and dwelling using empirical and simulated self-consumption probabilities. The model has been developed with data for four UK LSOA case study census areas, but is scalable to larger and smaller areas, and other geographical areas; all that is required are hard or probabilistic evidence for the building stock, as well as further enhancement of the solar PV yield sub-model to incorporate the irradiance in other geographies.

The object oriented design facilitates a better understanding of the distinct components of the model. Furthermore, in true object oriented fashion, it enables components to be altered without changing the whole model; it is envisaged, for example, that other renewable technologies, or even domestic energy efficiency interventions, could be integrated in to the model. Opportunities therefore exist for other researchers to expand and build upon this methodology.

Uncertainties have been endogenised in the model. Aleatory uncertainty, pertaining to, for example, weather, or occupancy behaviours, as well as epistemic uncertainties pertaining to, for example, occupancy level, or PV module and balance of system components, have been endogenised through the use of field data in which variability exists but is not known due to a lack of data. Methods have been devised of producing objective outputs such as decile, expected value, and median for the posterior distributions on key target variables. This presents a valuable research direction for the

potential construction of a multi-criteria decision support tool, which since uncertainty is endogenised, is better able to accommodate decision making involving risk analysis.

This methodology has its limitations, the main one being that the model outputs and inputs are a snapshot of time; there is no dynamic element since BNs are not easily able to model cyclic relationships which feedback loops which a dynamic system would require. However, that aside, the key benefits of this approach are as follows:

Data from multiple disciplines can be integrated using probabilistic relationships derived using empirical data and extant research. This is not possible with a deterministic system since linear relationships between key parameters are unknown and often cannot be derived. To illustrate this, consider, for example, a deterministic relationship between a household income and the built form of the dwelling in which the household live. This is not possible to model using linear equations without losing the uncertainty inherent in the empirical data; however a probabilistic relationship can be set out and modelled using conditional probabilities based upon such data. Thus, the OOBN presents a significantly different analysis paradigm; whereas the scientist/engineer is used to presenting deterministic relationships between parameters, and presenting average values, the new paradigm requires the presentation of outputs as probability distributions, and to beware the 'flaw of averages' (Savage et al., 2012). BNs, as other PGMs (Koller and Friedman, 2009), are part of this transition, which is facilitated by the increasing availability of large datasets – five years ago for example, it would have been unimaginable to have access to 5 million records for domestic fuel consumption. With smart meter implementation the amount of metered energy data will multiply inexorably. The new calculus, based on distributions requires new mathematic and algorithmic methods and tools which are in continuous development (e.g. Bessiere et al., 2013).

10.3.1 Contribution to Knowledge

The position taken with this research is that it has been very much a necessary early step to attempt to apply a probabilistic calculus to the economically, socially and environmentally important field of renewable energy, and in particular solar PV, deployment. It has been suggested, to the author, that the statistical reorganisation of extant data does not represent scientific endeavour (Romanos, 2014). However, the *thesis* presented here is that a fundamental scientific endeavour has been achieved; namely several disciplinary ontologies have, for the first time, been re-engineered into an ordered pattern to create a larger ontology. This has delivered a new epistemological tool – enabling the acquisition of new knowledge - from the resultant transdisciplinary domain. The user can interrogate the model by applying hard or probabilistic evidence representing observations or proposed scenarios, to one or more nodes to in order to deliver previously unknown results from this extant data delivering new insights and learning.

This is the first time that such a model has been developed, and applied to the deployment context for solar PV. Due to the object oriented design, it can potentially find wider application in other spatial contexts, such as non-domestic buildings, or different micro-generation technologies, such as solar thermal systems, heat pumps or micro-CHP etc. As such, the OOBN here represents a valuable contribution to knowledge both as a methodological discovery, and for unique tangible outputs, which are embedded in the model as a whole system representation of the knowledge domain.

10.4 Further Work

The following sections introduce several important areas of potential further research and development which build upon the work presented here.

10.4.1 Software Development

Work on the standardisation of evidence for BNs (Mrad et al., 2015) needs to be taken further by the BN software community. The weakness in current BN applications is that they do not all recognise the different types of evidence one might want to apply to a node, and in particular they do not easily facilitate the application of probabilistic evidence such as might be delivered from the application of this model to a new LSOA for example.

10.4.2 Low Carbon Interventions

Currently, the model only applies to with solar PV. However, the object oriented design was deliberately implemented to facilitate the integration of other technologies. That is not to say that such an endeavour would require the swapping of the generation component; significant new engineering would be required and the acquisition or simulation of new data. In particular, electricity technologies would have a different temporal generation profile than solar PV, rendering the CPT for self-consumption inapplicable. A further object of study could be to integrate energy efficiency interventions into a similar model, monetising the energy saved and hence enabling financial impact comparisons to be made between various low carbon interventions.

10.4.3 Geographic Information System Integration

The model alludes to Geographic Information System (GIS) integration by virtue of the geographic nature of the building stock model which is derived from LSOA census area data, and has, in this work, already been manipulated using GIS. A GIS system could spatially reference the distributions for the required parameters and, upon selection of a geographic area, apply new prior distributions as probabilistic evidence to the appropriate nodes. Thus, for any selected geographic area new

posterior distributions could be produced for key output nodes, such as the KPIs developed in this work for example. This is already a growing area of academic study, as mentioned in Chapter 2.

10.4.4 Decision support

An important research question (Section 1.5) asked whether insights in to a decision and policy support tool might be obtained using probabilistic methods. A fully fledged decision support tool this is not; there are no decision or utility nodes incorporated, as theorised by Smith (2010), and implemented in a BN for multi-criteria decision support as carried out by Delcroix et al (2013). Questions remain unanswered as to how results, presented as probability distributions, can be utilised since such outputs are typically not immediately accessible to decision makers (Buys et al., 2014).

The subjective, or objective, interpretation of probability distributions in decision making is an area of continuing academic study (McCloy, 2013); decision makers frequently desire binary values, whereas a distribution presents the cognitive challenge of a vector of results. Nevertheless, useful insights have been gained here, and with decision support expertise the model could be developed further in this direction. An additional aspect to this would be the implementation of a wider range of KPIs from multiple disciplines in order to facilitate multi-criteria decision making from a variety of stakeholder perspectives. This could build on the SEE approach developed here, perhaps widening it to other conceptual models for sustainability.

11 Bibliography

Aerts, D., Minnen, J., Glorieux, I., Wouters, I. and Descamps, F. (2014). A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. Building and Environment, 75, pp.67-78.

Authors B.R. Anderson, P. F. Chapman (2010). Bredem-12. Building Research Establishment, Volume 439, Watford, UK.

Anderson, B., Chapman, P., Cutland, N., Dickson, C., Doran, S., Henderson, G., Henderson, J., Iles, P., Kosmina, L. and Shorrock, L. (2002). BREDEM-8: Model Description 2001 Update. Building Research Establishment, Watford, UK.

Anderson, B. (1985). BREDEM: BRE Domestic Energy Model: Background, Philosophy and Description. Building Research Establishment, Watford, UK.

Anderson, B. (2011). Estimating Small Area Income Deprivation: An Iterative Proportional Fitting Approach Cresi Working Paper Number: 2011-02

Anderson, B., (2013). Estimating small-area income deprivation: an iterative proportional fitting approach. In: Tanton, R., Edwards, K. (Eds.), Spatial Microsimulation: A Reference Guide for Users, Understanding Population Trends and Processes, vol. 6. Springer, Netherlands, pp. 49–67, Chapter 4.

Armstrong, D. (2006). The quarks of object-oriented development. Communications of the ACM, 49(2), pp.123-128.

Arrhenius, S. A., (1896). On the influence of carbonic acid in the air upon the temperature of the ground, Philos. Mag., 41, 237.

Azzopardi, B., Mutale, J. and Martínez-Ceseña, E. (2013). Decision support system for ranking photovoltaic technologies. IET Renewable Power Generation, 7(6), pp.669 – 679.

Ballas, D., O'Donoghue, C., Clarke, G., Hynes, S., & Morrissey, K. (2013). A review of microsimulation for policy analysis. Springer. Chapter 3, p. 264.

Barthélemy, J., Suesse, T., Namazi-Rad, M., (2015) mipfp: Multidimensional Iterative Proportional Fitting and Alternative Models, Software [online] Available from: http://cran.r-project.org/. CRAN. [Accessed 23-Feb-2015].

Bartholomew, D. (1965). A Comparison of Some Bayesian and Frequentist Inferences. Biometrika, 52(1/2), p.19.

Bessiere, P., Mazer, E., Ahuactzin, J. and Mekhnacha, K. (2013). Bayesian Programming. Chapman & Hall/CRC.

Betts, T.R. and Gottschalg, R. (2007). Irradiance data: solar irradiation data for Loughborough. Centre for Renewable Energy Systems Technology (CREST), Loughborough University, UK.

Betts, T.R. and Gottschalg, R., (2013). Expected Variability of Monthly and Annual Energy Yield over PV System Lifetime, in: Proceedings of PVSAT-9. Swansea, pp. 75–78.

Bibby, P.R., Shepherd, J.W. (2004). Developing a New Classification of Urban and Rural Areas for Policy Purposes – the Methodology [online] available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/137655/rural-urban-definition-methodology-technical.pdf [Accessed 5-Jul-2013]

Boardman, B., Darby, S., Killip, G., Hinnells, M., Jardine, C., Palmer, S., Sinden, G., et al. (2005). 40% house. Oxford: ECI, University of Oxford.

Boardman, B. (2012). Fuel poverty synthesis: Lessons learnt, actions needed. Energy Policy, 49, pp.143-148.

Boland, J., Huang, J., Ridley, B. (2013). Decomposing global solar radiation into its direct and diffuse components. Renewable and Sustainable Energy Reviews, 28, pp749-756.

Booth, A. and Choudhary, R. (2012). Calibrating micro-level models with macro-level data using Bayesian regression analysis. Proceedings of the 12th IBPSA Conference, November 14-16, Sydney, Australia, pp. 641-648.

Booth, A. and Choudhary, R. (2013). Decision making under uncertainty in the retrofit analysis of the UK housing stock: Implications for the Green Deal. Energy and Buildings, 64, pp.292-308.

Borsuk, M., Stow, C. and Reckhow, K. (2004). A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling, 173(2-3), pp.219-239.

Box, G. (1976). Science and Statistics. Journal of the American Statistical Association, 71(356), pp.791-799.

BRE (Building Research Establishment), (2006). Domestic Photovoltaic Field Trials – Final Technical Report. Department of Trade and Industry Contractor Report, Building Research Establishment, Watford, UK.

BRE (Building research Establishment) (2014). The Government's Standard Assessment Procedure for Energy Rating of Dwellings 2012 edition, Published on behalf of Department of Energy and Climate Change by BRE, Garston, Watford, UK.

Bridge, G., Bouzarovski, S., Bradshaw, M. and Eyre, N. (2013). Geographies of energy transition: Space, place and the low-carbon economy. Energy Policy, 53, pp.331-340.

Burgess, J., Stirling, A., Clark, J., Davies, G., Eames, M., Staley, K. and Williamson, S. (2007). Deliberative mapping: a novel analytic-deliberative methodology to support contested science-policy decisions. Public Understanding of Science, 16(3), pp.299-322.

Burgess, P. (2014). Analysis of Historical Solar Irradiance Data and Development of a Virtual Pyranometer Method for Real-time PV Monitoring. PhD Thesis, University of Reading.

Buys, L., Mengersen, K., Johnson, S., van Buuren, N. and Chauvin, A. (2014). Creating a Sustainability Scorecard as a predictive tool for measuring the complex social, economic and environmental

impacts of industries, a case study: Assessing the viability and sustainability of the dairy industry. Journal of Environmental Management, 133, pp.184-192.

CACC (2015). One Million Climate Jobs [online] Available from: http://www.climate-change-jobs.org/ [Accessed 13/01/2015].

CACI (2014). Paycheck: Gross household income estimates at postcode level. [online] Available at: http://www.caci.co.uk/products/product/paycheck [Accessed 19 Apr. 2014].

Calderon, C. and Keirstead, J. (2012). Modelling frameworks for delivering low-carbon cities: advocating a normalized practice. Building Research & Information, 40(4), pp.504-517.

Calderon, C., James, P., Alderson, D., McLoughlin, A. and Wagener, T. (2012). Data availability and repeatability for urban carbon modelling: a carbon route map for Newcastle upon Tyne, presented at Retrofit 2012, Salford, England.

Campoccia, A., Dusonchet, L., Telaretti, E. and Zizzo, G. (2014). An analysis of feed'in tariffs for solar PV in six representative countries of the European Union. *Solar Energy*, 107, pp.530-542.

Cao, S. and Sirén, K. (2014). Impact of simulation time-resolution on the matching of PV production and household electric demand. Applied Energy, 128, pp.192-208.

Carmona, G., Varela-Ortega, C. and Bromley, J. (2013). Supporting decision making under uncertainty: Development of a participatory integrated model for water management in the middle Guadiana river basin. Environmental Modelling & Software, 50, pp.144-157.

Carta, J., Velázquez, S. and Matías, J. (2011). Use of Bayesian networks classifiers for long-term mean wind turbine energy output estimation at a potential wind energy conversion site. Energy Conversion and Management, 52(2), pp.1137-1149.

Catalina, T., Virgone, J. and Blanco, E. (2008). Development and validation of regression models to predict monthly heating demand for residential buildings. Energy and Buildings, 40(10), pp.1825-1832.

CCC (2012). Letter: The need for a carbon intensity target in the power sector. Lord Deben, Committee on Climate Change [Online] Available from: https://www.theccc.org.uk/publication/letter-the-need-for-a-carbon-intensity-target-in-the-power-sector/ [Accessed: 18-Dec-2012].

Celik, A. and Muneer, T. (2013). Neural network based method for conversion of solar radiation data. Energy Conversion and Management, 67, pp.117-124.

Cetiner, I. and Edis, E. (2014). An environmental and economic sustainability assessment method for the retrofitting of residential buildings. Energy and Buildings, 74, pp.132-140.

Cha, Y. and Stow, C. (2014). A Bayesian network incorporating observation error to predict phosphorus and chlorophyll a in Saginaw Bay. Environmental Modelling & Software, 57, pp.90-100.

Chapman, P. (1994). A geometrical model of dwellings for use in simple energy calculations. Energy and Buildings, 21(2), pp.83-91.

Checkland, P., (2000), Soft Systems Methodology: A Thirty Year Retrospective, Systems Research and Behavioural Science Syst. Res. 17, S11–S58.

Cheng, V. and Steemers, K. (2011). Modelling domestic energy consumption at district scale: A tool to support national and local energy policies. Environmental Modelling & Software, 26(10), pp.1186-1198.

Cherrington, R., Goodship, V., Longfield, A. and Kirwan, K. (2013). The feed-in tariff in the UK: A case study focus on domestic photovoltaic systems. Renewable Energy, 50, pp.421-426.

Chmutina, K., Sherriff, G. and Goodier, C. (2013). Success in international decentralised urban energy initiatives: a matter of understanding?. Local Environment, 19(5), pp.479-496.

Chmutina, K., Wiersma, B., Goodier, C. and Devine-Wright, P. (2014). Concern or compliance? Drivers of urban decentralised energy initiatives. Sustainable Cities and Society, 10, pp.122-129.

Cinar, D. and Kayakutlu, G. (2010). Scenario analysis using Bayesian networks: A case study in energy sector. Knowledge-Based Systems, 23(3), pp.267-276.

Colantuono, G., Everard, A., Hall, L. and Buckley, A. (2014). Monitoring nationwide ensembles of PV generators: Limitations and uncertainties. The case of the UK. Solar Energy, 108, pp.252--263.

Connor, P., Xie, L., Lowes, R., Britton, J. and Richardson, T. (2015). The development of renewable heating policy in the United Kingdom. Renewable Energy, 75, pp.733-744.

Conrady, S., and Jouffe, L. (2011), Introduction to Bayesian Networks, Conrady Applied Science, LLC, Franklin, USA

Cook, J. and Cowtan, K. (2015). Reply to Comment on 'Quantifying the consensus on anthropogenic global warming in the scientific literature'. Environ. Res. Lett., 10(3), p.039002.

Cook, J., Nuccitelli, D., Green, S., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P. and Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. Environ. Res. Lett., 8(2), p.024024.

COP (2015). For a universal climate agreement. 21st Session of the Conference of the Parties to the United Nations Framework Convention on Climate Change [online] Available from: http://www.cop21.gouv.fr/en [Accessed 30-Jul-2015].

Cros S., Mayer D., Wald L., (2004) The availability of irradiation data. International Energy Agency, Report IEA-PVPS T 2, Vienna, Austria.

Cucchiella, F., D'Adamo, I., Gastaldi, M. and Koh, S. (2012). Renewable energy options for buildings: Performance evaluations of integrated photovoltaic systems. Energy and Buildings, 55, pp.208-217.

Dagum, P. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artificial Intelligence, 60(1), pp.141-153.

Daly, R., Shen, Q. and Aitken, S. (2011). Learning Bayesian networks: approaches and issues. The Knowledge Engineering Review, 26(02), pp.99-157.

Darling, S., You, F., Veselka, T. and Velosa, A. (2011). Assumptions and the levelized cost of energy for photovoltaics. Energy Environ. Sci., 4(9), p.3133.

DCLG (Department for Communities and Local Government). (2010). English Housing Survey, 2010-2011: Household Data [computer file]. 2nd Edition. Colchester, Essex: UK Data Archive [distributor], January 2013. SN: 7040

DCLG (Department for Communities and Local Government), (2011). The English Indices of Deprivation 2010, ISBN: 978-1-4098-2922-5.

De Finetti, B. (1974). Theory of probability. London: Wiley.

de Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, 41, pp.40-49.

Deakin, M. (2012). The case for socially inclusive visioning in the community-based approach to sustainable urban regeneration. Sustainable Cities and Society, 3, pp.13-23.

DECC (2012). Summary of Analysis National Energy Efficiency Data-Framework, London, URN 12D/405.

DECC (2012A). Identifying trends in the deployment of domestic solar PV under the Feed-in Tariff scheme, London, URN 12D/247

DECC, (2014). Overview of Weather Correction of Gas Industry Consumption Data. [online] Available from: https://www.gov.uk/government/statistics/overview-of-weather-correction-of-gas-industry-consumption-data [Accessed 3-Nov-2014].

DECC (2014A). National Energy Efficiency Data-Framework, 2014 [computer file]. Colchester, Essex: UK Data Archive [distributor], July 2014. SN: 7518.

DECC (2015). Solar photovoltaics deployment': Feed-in Tariff Statistics, June, available from https://www.gov.uk/government/statistics/solar-photovoltaics-deployment

DECC (2015A), Monthly feed-in tariff commissioned installations by month, Excel Spreadsheet, Downloaded from https://www.gov.uk/government/statistics/monthly-small-scale-renewable-deployment

Decker B., Grochowski J. and Jahn U. (1993). Results and experience from the German 1000-Roof-Photovoltaic Programme-140 grid connected PV systems in Lower Saxony. Proc. ISES Solar World Congress, Budapest, pp. 95-100.

Delcroix, V., (2013), personal communication on the method of relating disparate data categories.

Delcroix, V., Sedki, K. and Lepoutre, F. (2013). A Bayesian network for recurrent multi-criteria and multi-attribute decision problems: Choosing a manual wheelchair. Expert Systems with Applications, 40(7), pp.2541-2551.

Deming, W. and Stephan, F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. Ann. Math. Statist., 11(4), pp.427-444.

Devine-Wright, P. (2008). Reconsidering public acceptance of renewable energy technologies: a critical review. In Grubb M, Jamasb T, Pollitt M (eds.) Delivering a Low Carbon Electricity System: Technologies, Economics and Policy, Cambridge University Press, 443-461.

Devine-Wright, P. (2010). Renewable energy and the public. London: Earthscan.

Dickson, C., Dunster, J., Lafferty, S. and Shorrock, L. (1996). BREDEM: Testing monthly and seasonal versions against measurements and against detailed simulation models. Building Services Engineering Research and Technology, 17(3), pp.135-140.

Doylend, N. (2015). Evaluating building energy performance: A lifecycle risk management methodology. EngD. Thesis, Loughborough University.

Druckman, A. and Jackson, T. (2008). Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. Energy Policy, 36(8), pp.3177-3192.

Duespohl, M., Frank, S. and Doell, P. (2012). A Review of Bayesian Networks as a Participatory Modeling Approach in Support of Sustainable Environmental Management. Journal of Sustainable Development, 5(12).

Dunlop, E.D. (2003)'Lifetime performance of crystalline silicon PV modules In: Proceedings of the 3rd World Conference on Photovoltaic Energy Conversion. Osaka, Japan, pp. 2927–2930, 2003.

Dusabe, D., Munda, J. and Jimoh, A., (2009). Modelling of cloudless solar radiation for PV module performance analysis. J. Electr Eng, 60(4), pp 192–197.

ECJRC (European Commission Joint Research Centre), (2013). Photovoltaic Geographical Information System (PVGIS) Version 4. [online computer program] Available from: http://re.jrc.ec.europa.eu/pvgis/ [Accessed 12-Jan-2013].

Edenhofer, O., Hirth, L., Knopf, B., Pahle, M., Schlömer, S., Schmid, E. and Ueckerdt, F. (2013). On the economics of renewable energy sources. Energy Economics, 40, pp.S12-S23.

EERA. (2013). Economic, environmental and social impacts (JP e3s) [Online] Available from: http://www.eera-set.eu/eera-joint-programmes-jps/economic-environmental-and-social-impacts-jp-e3s/ [Accessed 30/07/2014]

Elkington, J. (1998). Cannibals with forks Cannibals with forks: the triple bottom line of 21st century business. Gabriola Island, BC: New Society Publishers.

Energy Networks Association, (2003). Engineering Recommendation G83/1 Recommendations For The Connection Of Small-scale Embedded Generators (up To 16 A Per Phase) In Parallel With Public Low-voltage Distribution Networks. Published by Engineering Directorate, Energy Networks Association, London.

EP (European Parliament), 2010). Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings. Bruxelles, Belgium.

ESRI (1998). Shapefile Technical Description - An ESRI White Paper. [online] Available from: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf [Accessed 5-Oct-2013].

EU (European Union), (2009). Directive 2009/28/EC of the European Paliament on the promotion of the use of energy from renewable sources.

Everard, A., (2013) Personal communication regarding anoymisation of Sheffield Microgeneration Data.

Famuyibo, A., Duffy, A. and Strachan, P. (2012). Developing archetypes for domestic dwellings—An Irish case study. Energy and Buildings, 50, pp.150-157.

Fenton, N. and Neil, M. (2012). Risk assessment and decision analysis with Bayesian networks. Boca Raton: Taylor & Francis.

Fenton, N., Krause, P. and Neil, M. (2002). Software measurement: uncertainty and causal modeling. IEEE Softw., 19(4), pp.116-122.

Fienberg, S. (1970). An Iterative Procedure for Estimation in Contingency Tables. Ann. Math. Statist., 41(3), pp.907-917.

de Finetti, B., (1937). La prévision: See Lois Logiques, ses Sources Subjectives," Annalesde l'Institut Henri Poincaré, Vol. 7,1937.

de Finetti, B. (1937). La prévision: See Lois Logiques, ses Sources Subjectives. Annales de l'Institut Henri Poincaré, 7.

Firth, S., Lomas, K. and Wright, A. (2010). Targeting household energy-efficiency measures using sensitivity analysis. Building Research & Information, 38(1), pp.25-41.

Fisher, I. (1930). The theory of interest. New York: Macmillan Co.

Foulds, C. and Powell, J. (2014). Using the Homes Energy Efficiency Database as a research resource for residential insulation improvements. Energy Policy, 69, pp.57-72.

Fouquet, R. and Pearson, P. (1998). A Thousand Years of Energy Use in the United Kingdom. EJ, 19(4).

Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29, pp131-163.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, Vol 29, 131-163.

Gadsden, S., Rylatt, M. and Lomas, K. (2003). Putting solar energy on the urban map: a new GIS-based approach for dwellings. Solar Energy, 74(5), pp.397-407.

Geiger, D. and Pearl, J.(1988). On The Logic of Causal Models. Proc. of the 4th Workshop on Uncertainty in Artificial Intelligence, St. Paul, Minnesota, USA, pp. 136-147.

GIG, (2012). Cities Revealed Building Class Datasets Notes to Users 2012. The Geoinformation Group UK [online] Available from: http://www.geoinformationgroup.co.uk/products/ukbuildings [Accessed 3-Jun-2014].

GIG, (2013). UKBuildings. The Geoinformation Group UK [online] Available from: http://www.geoinformationgroup.co.uk/products/ukbuildings [Accessed 3-Jun-2014].

Goodier, C., Austin, S., Soetanto, R. and Dainty, A. (2010). Causal mapping and scenario building with multiple organisations. Futures, 42(3), pp.219-229.

Goss, B., Gottschalg, R. and Betts, T.R., (2012), Uncertainty Analysis of Photovoltaic Energy Yield Prediction. 8th Photovoltaic Science Application and Technology (PVSAT-8) Conference and Exhibition, Newcastle, England, 2nd-4th April 2012, pp.157-160.

Goss, B., Gottschalg, R. and Betts, T. (2012). Uncertainty Analysis of Photovoltaic Energy Yield Prediction. In: 8th Photovoltaic Science Application and Technology (PVSAT-8) Conference and Exhibition. Newcastle: UK National Section of the International Solar Energy Society, pp.157-160.

Goss, B. (2010) Choosing Solar Electricity. 1st edn. Machynlleth: CAT

Green, M. A., Emery, K., Hishikawa, Y., Warta, W. and Dunlop, E. D. (2013). Solar cell efficiency tables (version 42). Prog. Photovolt: Res. Appl., 21, pp827–837.

Green, M.A. (2003). General temperature dependence of solar cell performance and implications for device modelling. Progress in Photovoltaics: Research and Applications, 11, 5 pp333-40.

Gupta, R. (2005). Investigating the potential for local carbon djoxjde emission 1·eductions: developing a CIS-based domestic energy, carbon-counting and carbon-reduction model. Refereed Technical Paper, in 2005 Solar World Congress, Orlando, Florida, USA.

Hamakawa, Y. (2002). Solar PV energy conversion and the 21st century's civilization. Solar Energy Materials and Solar Cells, 74(1-4), pp.13-23.

Hamilton, I., Steadman, P., Bruhns, H., Summerfield, A. and Lowe, R. (2013). Energy efficiency in the British housing stock: Energy demand and the Homes Energy Efficiency Database. Energy Policy, 60, pp.462-480.

Hammond, G. and Pearson, P. (2013). Challenges of the transition to a low carbon, more electric future: From here to 2050. Energy Policy, 52, pp.1-9.

Hart-Davis, D., (2013). Analysis intensities 2009 (spreadsheet) downloaded from http://www.earth.org.uk/note-on-UK-grid-CO2-intensity-variations.html#fullyear2009 [Accessed 19-May-2014].

Heath, M., Walshe, J. and Watson, S. (2007). Estimating the potential yield of small building-mounted wind turbines. Wind Energy, 10(3), pp.271-287.

Heckerman, D., Geiger, D. and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Mach Learn, 20(3), pp.197-243.

Heeren, N., Jakob, M., Martius, G., Gross, N. and Wallbaum, H. (2013). A component based bottom-up building stock model for comprehensive environmental impact assessment and target control. Renewable and Sustainable Energy Reviews, 20, pp.45-56.

Helm, D. (2002). Energy policy: security of supply, sustainability and competition. Energy Policy, 30(3), pp.173-184.

Helm, D. (2003). Energy, the state, and the market. Oxford: Oxford University Press.

Hens, H. (2007). Building physics – heat, air and moisture. Fundamentals and Engineering Methods with Examples and Exercises. Ernst & Sohn.

Heo, Y., Choudhary, R. and Augenbroe, G. (2012). Calibration of building energy models for retrofit analysis under uncertainty. Energy and Buildings, 47, pp.550-560.

Hills, J. (2012). Getting the measure of fuel poverty: Final Report of the Fuel Poverty Review. CASE report 72 ISSN 1465-3001, DECC, London.

Hinrichs-Rahlwes, R. (2013). Renewable energy: Paving the way towards sustainable energy security. Renewable Energy, 49, pp.10-14.

Hirsch, D. (2015) A Minimum Income Standard for the UK in 2015 York: Joseph Rowntree Foundation

Hitchcock, G. (1993). An integrated framework for energy use and behaviour in the domestic sector. Energy and Buildings, 20(2), pp.151-157.

Horsfield, G. (2011). Family spending: A report on the 2010 living costs and food survey, Edited G. Horsfield, Published by Office for National Statistics, Newport, Wales, ISSN 2040-1647.

Howard, C. and Stumptner, M. (2009). Automated compilation of Object-Oriented Probabilistic Relational Models. International Journal of Approximate Reasoning, 50(9), pp.1369-1398.

Huebner, G., McMichael, M., Shipworth, D., Shipworth, M., Durand-Daubin, M. and Summerfield, A. (2013). Heating patterns in English homes: Comparing results from a national survey against common model assumptions. Building and Environment, 70, pp.298-305.

Hughes, M., Palmer, J., Cheng, V. and Shipworth, D. (2013). Sensitivity and uncertainty analysis of England's housing energy model. Building Research & Information, 41(2), pp.156-167.

Huld T., Müller R. and Gambardella A. (2012). A new solar radiation database for estimating PV performance in Europe and Africa. Sol Energy, 86 pp1803–15.

Huld, T. (2012). Personal communication regarding automatic interrogation of PVGIS using HTTP requests

Huld, T. (2014). Personal communication concernning the use of R.Sun

Hulme, M., Conway, D., Jones, P., Jiang, T., Barrow, E. and Turney, C. (1995). Construction of a 1961–1990 European climatology for climate change modelling and impact applications. International Journal of Climatology, 15(12), pp.1333-1363.

Hutchinson, M., Booth, T., McMahon, J. and Nix, H. (1984). Estimating monthly mean valuesof daily total solar radiation for Australia. Solar Energy, 32(2), pp.277-290.

IEA (International Energy Agency) − (2012), Renewable Energy Medium-Term Market Report 2012, OECD/IEA, Paris.

IPCC (Intergovernmental Panel on Climate Change) (2013). Climate Change 2013: The Physical Science Basis, Working Group I Report, World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP).

Iqbal, M. (1983), An Introduction to Solar Radiation, Academic Press, New York.

Iversen, G. (1984). Bayesian statistical inference. 1st ed. Newbury Park, Calif: Sage Publications

Jackson, T. (2012). Prosperity without Growth. Hoboken: Earthscan.

Jacques, D., Gooding, J., Giesekam, J., Tomlin, A. and Crook, R. (2014). Methodology for the assessment of PV capacity over a city region using low-resolution LiDAR data and application to the City of Leeds (UK). Applied Energy, 124, pp.28--34.

Jenkins, A. (2013). The Sun's position in the sky. Eur. J. Phys., [online] 34(3), pp.633-652.

Jensen, F. and Nielsen, T. (2007). Bayesian networks and decision graphs. New York: Springer.

Jensen, F.V., Lauritzen, S.L. and Olesen, K.G. (1990). Bayesian Updating in Causal Probabilistic Networks by Local Computation. Computational Statistical Quarterly, 4.

Jiang, Z., Li, W., Apley, D.W., and Chen, W. (2012), A System Uncertainty Propagation Approach With Model Uncertainty Quantification in Multidisciplinary Design in Proceedings of the ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Buffalo, USA, August 17–20, 2014, ISBN: 978-0-7918-4632-2

John, E.G. and Grosvenor, R.I. (2001). Failure distribution curve fitting using spreadsheet add-ins. Int.J.Elec.Eng.Educ., 38(2), pp165-172.

Johnson, S. and Mengersen, K. (2011). Integrated Bayesian network framework for modeling complex ecological issues. Integr Environ Assess Manag, 8(3), pp.480-490.

Johnson, S., Mengersen, K., de Waal, A., Marnewick, K., Cilliers, D., Houser, A. and Boast, L. (2010). Modelling cheetah relocation success in southern Africa using an Iterative Bayesian Network Development Cycle. Ecological Modelling, 221(4), pp.641-651.

Johnson, S. (2009), Integrated Bayesian Network frameworks for modelling complex ecological issues, PhD Thesis, School of Mathematical Sciences, Queensland University of Technology, Queensland, Australia.

Jordan, D. C. and Kurtz, S. R. (2013). Photovoltaic Degradation Rates - an Analytical Review. Prog. Photovolt: Res. Appl., 21, pp12–29.

Kaplanis, S. and E. Kaplani (2011). Energy performance and degradation over 20 years performance of BP c-Si PV modules', Simulation Modelling Practice and Theory, 19(4), pp.1201-1211.

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z. and Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. Building and Environment, 45(7), pp.1683-1697.

Keirstead, J. and Calderon, C. (2012). Capturing spatial effects, technology interactions, and uncertainty in urban energy and carbon models: Retrofitting newcastle as a case-study. Energy Policy, 46, pp.253-267.

Keirstead, J. and Calderon, C. (2014). Corrigendum to “Capturing spatial effects, technology interactions, and uncertainty in urban energy and carbon models: Retrofitting Newcastle as a case-study” [Energy Policy 46 (2012) 253–267]. Energy Policy, 74, p.714.

Keirstead, J. and Schulz, N. (2010). London and beyond: Taking a closer look at urban energy policy. Energy Policy, 38(9), pp.4870-4879.

Keirstead, J., Jennings, M. and Sivakumar, A. (2012). A review of urban energy system models: Approaches, challenges and opportunities. Renewable and Sustainable Energy Reviews, 16(6), pp.3847-3866.

Kelly (Letcher), R., Jakeman, A., Barreteau, O., Borsuk, M., ElSawah, S., Hamilton, S., Henriksen, H., Kuikka, S., Maier, H., Rizzoli, A., van Delden, H. and Voinov, A. (2013). Selecting among five common modelling approaches for integrated environmental assessment and management. Environmental Modelling & Software, 47, pp.159-181.

Kelly, S., Shipworth, M., Shipworth, D., Gentry, M., Wright, A., Pollitt, M., Crawford-Brown, D. and Lomas, K. (2013). Predicting the diversity of internal temperatures from the English residential sector using panel methods. Applied Energy, 102, pp.601-621.

Kelly, S. (2011). Do homes that are more energy efficient consume less energy?: A structural equation model of the English residential sector. Energy, 36(9), pp.5610-5620.

Kindermann, R. and Snell, J.L. (1980). Markov Random Fields and Their Applications (PDF). American Mathematical Society. ISBN 0-8218-5001-6.

Kindermann, R. and Snell, J. (1980). On the relation between Markov random fields and social networks*. The Journal of Mathematical Sociology, 7(1), pp.1-13.

Koch, K. (2007). Introduction to Bayesian statistics. Berlin: Springer.

Koller, D. and Friedman, N. (2009). Probabilistic graphical models. Cambridge, MA: MIT Press.

Koller, D., and Pfeffer, A. (1997). Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97), pages 302-313, Providence, Rhode Island, August 1-3, 1997

Kolmogorov, (1933). Kolmogorov, A.N., Foundations of the Theory of Probability, Chelsea, New York, 1950 (originally published in 1933 as Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin).

Kolmogorov, A.N. (1950) Foundations of the Theory of Probability. Chelsea, New York (originally published in 1933 as Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin)

Krawczynski M., Strobel, M.B., Betts, T.R., Gottschalg, R. (2010). Spectral influences on estimations of useful irradiance for different PV technologies. IN: Sixth Photovoltaic Science Application and Technology Conference (PVSAT-6), 24-26th Mar, Southampton, 5pp.

Kroposki, B. and Hansen, R., Technical evaluation of four amorphous silicon systems at NREL, Proceedings of the 26th PV Specialists Conference, Anaheim, CA, 1357–1360, 1997 DOI: 10.1109/PVSC.1997.654342

Kurban, H., Gallagher, R., Kurban, G.A., and Persky, J. (2011). A Beginner's Guide to Creating Small-Area Cross-Tabulations. Cityscape: A Journal of Policy Development and Research 13: 225–235.

Lane, D. (2007). Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

Laniak, G., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R. and Hughes, A. (2013). Integrated environmental modeling: A vision and roadmap for the future. Environmental Modelling & Software, 39, pp.3-23.

Lauritzen, S.L., and D.J. Spiegelhalter (1988). Local Computation with Probabilities in Graphical Structures and Their Applications to Expert Systems. Journal of the Royal Statistical Society B, 50(2).

Lecklin, T., Ryömä, R. and Kuikka, S. (2011). A Bayesian network for analyzing biological acute and long-term impacts of an oil spill in the Gulf of Finland. Marine Pollution Bulletin, 62(12), pp.2822-2835.

Lefèvre, M., Cros, S., Albuisson, M., and Wald, L. (2004), Developing a database using METEOSAT data for the delivery of solar radiation assessments at ground level.

Lehikoinen, A., Luoma, E., Mäntyniemi, S. and Kuikka, S. (2013). Optimizing the Recovery Efficiency of Finnish Oil Combating Vessels in the Gulf of Finland Using Bayesian Networks. Environmental Science & Technology, 47(4), pp.1792-1799.

Leicester, P., Goodier, C.I. and Rowley, P. (2011). Evaluating the impacts of Community Renewable Energy Initiatives, ISES Solar World Congress, Kassel, Germany, 28 Aug.- 2 Sept. 2011.

Leloux, J., Narvarte, L. and Trebosc, D. (2012). Review of the performance of residential PV systems in Belgium. Renewable and Sustainable Energy Reviews, 16(1), pp.178-184.

Leloux, J., Narvarte, L. and Trebosc, D. (2012). Review of the performance of residential PV systems in France. Renewable and Sustainable Energy Reviews, 16(2), pp.1369-1376.

Liu, B.Y.H., Jordan, R.C., (1960). The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. Sol. Energy 4, pp1–19.

Lovelace, R. and Ballas, D. (2013). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. Computers, Environment and Urban Systems, 41, pp.1-11.

Lyytimäki, J., Tapio, P., Varho, V. and Söderman, T. (2013). The use, non-use and misuse of indicators in sustainability assessment and communication. International Journal of Sustainable Development & World Ecology, 20(5), pp.385-393.

Mallaburn, P. and Eyre, N. (2013). Lessons from energy efficiency policy and programmesin the UK from 1973 to 2013. Energy Efficiency, 7(1), pp.23-41.

Manco, J. (2014). History of Building Regulations. [online] Available from: http://www.buildinghistory.org/regulations.shtml [Accessed 4-Aug-2014].

Marcot, B., Steventon, J., Sutherland, G. and McCann, R. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. Canadian Journal of Forest Research, 36(12), pp.3063-3074.

Marsh, G. (2004). Lowering the barriers to RE. Refocus, 5(6), pp.45-47.

Martin, D., Nolan, A., Tranmer, M., 2001. The application of zone-design methodology in the 2001 UK Census. Environment and Planning A 33, 1949–1962

Mauthner F, Weiss W. (2014), Solar Heat Worldwide - Markets and contribution to the energy supply 2012. International Energy Agency, Edition 2014.

McCloy, R. (2013). PURE Research Blog: Understanding probability and uncertainty [Online] Available from: https://connect.innovateuk.org/web/pure-research-programme/article-view/-/blogs/pure-research-blog-understanding-probability-and-uncertainty-dr-rachel-mccloy-university-of-reading [Accessed 15/07/2015].

McCormick, P. G. and H. Suehrcke (1991). Diffuse fraction correlations. Solar Energy, 47, pp311–312.

McKenna, E. (2013). Demand response of domestic consumers to dynamic electricity pricing in low-carbon power systems. Thesis, Loughborough University.

McLoughlin, F., Duffy, A. and Conlon, M. (2012). Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. Energy and Buildings, 48, pp.240-248.

Meadows, D. (1972). The Limits to growth. 1st ed. New York: Universe Books.

Melius, J., Margolis, R., and Ong, S. (2013). Estimating Rooftop Suitability for PV: A Review of Methods, Patents, and Validation Techniques, National Renewable Energy Laboratory, Technical Report, NREL/TP-6A20-60593.

Mendonça, M., Jacobs, D. and Sovacool, B. (2009). Powering the green economy. Sterling, VA: Earthscan.

Menezes, A., Cripps, A., Bouchlaghem, D. and Buswell, R. (2012). Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. Applied Energy, 97, pp.355-364.

Mhalas, A., Kassem, M., Crosbie, T. and Dawood, N. (2013). A visual energy performance assessment and decision support tool for dwellings. Vis Eng, 1(1), p.7.

Mingers, J., & Rosenhead, J. 2004. Problem Structuring Methods in Action. European Journal of Operational Research, 152(3): 530-554.

Mitášová, H. and Mitáš, L. (1993). Interpolation by regularized spline with tension: I. Theory and implementation. Mathematical Geology, 25(6), pp.641-655.

Mitchell, C., Watson, J. and Whiting, J. (2013). New Challenges in Energy Security: The UK in a Multipolar World. Palgrave Macmillan.

Molina, J., Bromley, J., García-Aróstegui, J., Sullivan, C. and Benavente, J. (2010). Integrated water resources management of overexploited hydrogeological systems using Object-Oriented Bayesian Networks. Environmental Modelling & Software, 25(4), pp.383-397.

Molina, J., Pulido-Velázquez, D., García-Aróstegui, J. and Pulido-Velázquez, M. (2013). Dynamic Bayesian Networks as a Decision Support tool for assessing Climate Change impacts on highly stressed groundwater systems. Journal of Hydrology, 479, pp.113-129.

Mosteller, F., (1968). Association and estimation in contingency tables. J. Amer. Statist.Assoc. 63 1-28.

Mrad, A., Delcroix, V., Piechowiak, S., Leicester, P. and Abid, M. (2015). An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. Applied Intelligence, [online] pp.1-23.

Muneer, T. (1990). Solar radiation model for Europe. Building Services Engineering Research and Technology, 11(4), pp.153-163.

Munzinger, M., Crick, F., Dayan, E., Pearsall, N. and Martin, C. (2006). PV Domestic Field Trial: Final Technical Report [online] Available from: www.bis.gov.uk/files/file36660.pdf [Accessed 10-Dec- 2012].

Murphy, G., Kummert, M., Anderson, B. and Counsell, J. (2011). A comparison of the UK Standard Assessment Procedure and detailed simulation of solar energy systems for dwellings. Journal of Building Performance Simulation, 4(1), pp.75-90.

Nadkarni, S. and Shenoy, P.P. (1999). A Bayesian network approach to making inferences in causal maps,

Nadkarni, S. and Shenoy, P. (2004). A causal mapping approach to constructing Bayesian networks. Decision Support Systems, 38(2), pp.259-281

Nadkarni, S. and Shenoy, P.P. (2004). A causal mapping approach to constructing Bayesian networks. Decision Support Systems, 38(2), pp.259–281.

Navigant Consulting (2006). A review of PV inverter technology cost and performance projections. NREL. [online] Available from: http://www.nrel.gov/docs/fy06osti/38771.pdf [Accessed 19-March-2014].

Neapolitan, R. (2004). Learning Bayesian networks. Upper Saddle River, N.J.: Pearson Prentice Hall.

Nguyen, H. and Pearce, J. (2010). Estimating potential photovoltaic yield with r.sun and the open source Geographical Resources Analysis Support System. Solar Energy, 84(5), pp.831-843.

Nguyen, H., Pearce, J., Harrap, R. and Barber, G. (2012). The application of LiDAR to assessment of rooftop solar photovoltaic deployment potential in a municipal district unit. Sensors, 12(4), pp.4534--4558.

NHER, (2004). *NHER Levels of Analysis*. [online] National Home Energy Rating Scheme. Available at: http://www.nesltd.co.uk/ [Accessed 3 Feb. 2015].

Nielsen, L. (1993). How to get the birds in the bush into your hand. Energy Policy, 21(11), pp.1133-1144.

Norman, P., (1999), Putting Iterative Proportional Fitting on the Researcher's Desk, Working paper 99/03, School of Geography University of Leeds, UK.

Norsys, 1995. Netica Application. Available at: https://www.norsys.com/netica.html.

Notton, G., Lazarov, V. and Stoyanov, L. (2010). Optimal sizing of a grid-connected PV system for various PV module technologies and inclinations, inverter efficiency characteristics and locations. Renewable Energy, 35(2), pp.541--554.

Nugent, D. and Sovacool, B. (2014). Assessing the lifecycle greenhouse gas emissions from solar PV and wind energy: A critical meta-survey. Energy Policy, 65, pp.229-244.

OFGEM (2013). Central Feed-in Tariff Register (CFR) User Guide [online] Available from:https://www.ofgem.gov.uk/publications-and-updates/central-feed-tariff-register-cfr-user-guide [Accessed 13/05/2014].

OFGEM (2015). Feed-in Tariff Rates [online] Available from: https://www.ofgem.gov.uk/ [Accessed 18-May-2015].

Olivier, C. (2008). Modelling UK Home Energy Use Using Bayesian Networks Documentation and Experiments. [online] Available from: www.ofrancois.tuxfamily.org/carb/ [Accessed Jun-2014].

ONS (2011) Census: population and household estimates for Wards and Output Areas in England and Wales. [online] Available from: http://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuspopulationandhouseholdestimatesforsmallareasinenglandandwales/2012-11-23. (Accessed August 2014).

ONS (2015), Retail Price Index data, 1986 to 2014, Office for National Statistics, 2015, http://ons.gov.uk/

Office for National Statistics and Department for Environment, Food and Rural Affairs, Living Costs and Food Survey, 2010 [computer file]. 2nd Edition. Colchester, Essex: UK Data Archive [distributor], July 2012. SN: 6945 , http://dx.doi.org/10.5255/UKDA-SN-6945-2v

Ordnance Survey, (2014). PAF Public Sector Licence | Business and government | Ordnance Survey. [online] Available at: https://www.ordnancesurvey.co.uk/business-and-government/public-sector/mapping-agreements/paf-psl-update.html [Accessed 4 Jan. 2014].

Ordonez, J., Jadraque, E., Alegre, J. and Martinez, G. (2010). Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain). Renewable and Sustainable Energy Reviews, 14(7), pp.2122--2130.

Orioli, A. and Di Gangi, A. (2014). Review of the energy and economic parameters involved in the effectiveness of grid-connected PV systems installed in multi-storey buildings. Applied Energy, 113, pp.955-969.

Page, J. (2005). First conference on measurement and modeling of solar radiation and daylight "Challenges for the 21st Century" Napier University, Edinburgh, 15–16 September 2003. Energy, 30(9), pp.1501-1515.

Palmer, J. and Cooper, I. (2012). United kingdom housing energy fact file, Tech. rep., Department of Energy and Climate Change.

Palmer, D., Betts, T., Gottschalg, R., (2015), Assessment of Potential for Photovoltaic Roof Installations by Extraction of Roof Slope from Lidar Data and Aggregation to Census Geography, Proceedings of Conference C97 of the SOLAR ENERGY Society, PVSAT-11, 15-17/04/15, Leeds, UK, Eds: Michael Hutchins and Alex Cole, ISBN 0 904963 81 0.

Parsons Brinkenhoff, (2012). Solar PV Cost Update Prepared by for the Department of Energy and Climate Change, Parsons Brinkenhoff, London.

Pearl, J. (1985). Bayesian networks. [Los Angeles, Calif.]: UCLA, Computer Science Dept.

Pearl, J. (1986). Fusion, Propagation, and Structuring in Belief Networks. Artificial Intelligence, 29.

Perez R., Seals R., Ineichen P., Stewart R., Menicucci D. (1987). A new simplified version of the Perez diffuse irradiance model for tilted surfaces. Solar Energy, 39(3), pp221–31.

Perez, R., Seals, R., Zelenka, A., (1997). Comparing satellite remote sensing and ground network measurements for the production of site/time specific irradiance data. Solar Energy, 60, pp89-96.

Pérez-Miñana, E., Krause, P. and Thornton, J. (2012). Bayesian Networks for the management of greenhouse gas emissions in the British agricultural sector. Environmental Modelling & Software, 35, pp.132-148.

Piercy, E., Granger, R. and Goodier, C. (2010). Planning for peak oil: learning from Cuba's 'special period'. Proceedings of the ICE - Urban Design and Planning, 163(4), pp.169-176.

Pourret, O., Naïm, P. and Marcot, B. (2008). Bayesian networks: Applications. Chichester, West Sussex, Eng.: John Wiley.

Preston, I., White, V., Guertler, P. (2010). Distributional impacts of UK Climate Change Policies Final Final report to eaga Charitable Trust. Centre for Sustainable Energy and Association for the Conservation of Energy, Bristol, UK.

Preston, I., White, V., Thumim, J. and Bridgeman, T. (2013). Distribution of Carbon Emissions in the UK: Implications for Domestic Energy Policy. Joseph Rowntree Foundation, York.

Probert, L., (2014). Energy supplier involvement in English fuel poverty alleviation: a critical analysis of emergent approaches and implications for policy success, PhD Thesis, Loughborough University.

Probst, O. (2002). The apparent motion of the Sun revisited. European journal of physics, 23(3), p.315.

Przytula, K.W. and Thompson, D. (2000). Construction of Bayesian Networks for Diagnostics. Proceedings of 2000 IEEE Aerospace Conference, March 18-24.

Reason, L. and Alan Clarke, A. (2008). Projecting Energy Use And CO2 Emissions From Low Energy Buildings: A Comparison Of The Passivhaus Planning Package (PHPP) And SAP. The Association for Environment Conscious Building, Llandysul, UK, pp43.

Rees, S. and Curtis, R. (2014). National Deployment of Domestic Geothermal Heat Pump Technology: Observations on the UK Experience 1995–2013. *Energies*, 7(8), pp.5460-5499.

Rees, S. and Curtis, R. (2014). Correction: Rees, S. and Curtis, R. National Deployment of Domestic Geothermal Heat Pump Technology: Observations on the UK Experience 1995–2013. Energies 2014, 7, 5460–5499. Energies, 7(9), pp.6224-6224.

Reich, N., Mueller, B., Armbruster, A., van Sark, W., Kiefer, K. and Reise, C. (2012). Performance ratio revisited: is PR > 90% realistic?. Prog. Photovolt: Res. Appl., 20(6), pp.717-726.

Reindl, D.T., Beckman, W.A., Duffie, J.A., (1990). Diffuse fraction correlations. Sol. Energy 45, pp1–7.

Richardson, I. and Thomson, M. (2012). Integrated simulation of photovoltaic micro-generation and domestic electricity demand: a one-minute resolution open-source model. Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 227(1), pp.73-81.

Richardson, I., Thomson, M., Infield, D. and Delahunty, A. (2009). Domestic lighting: A high-resolution energy demand model. Energy and Buildings, 41(7), pp.781-789.

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

Rigollier C., Lefèvre M., Cros S., Wald L., (2003). Heliosat 2: an improved method for the mapping of the solar radiation from Meteosat imagery. In Proceedings of the 2002 EUMETSAT Meteorological Satellite Conference, Dublin, Ireland, 1-6 September 2002. Published by EUMETSAT, Darmstadt, Germany, pp585-592.

Riley, K., Hobson, M. and Bence, S. (2006). Mathematical methods for physics and engineering. Cambridge: Cambridge University Press.

Rogers, J., Simmons, E., Convery, I. and Weatherall, A. (2012). Social impacts of community renewable energy projects: findings from a woodfuel case study. Energy Policy, 42, pp.239-247.

Romanos, P., (2014), personal communication to the author at Loughborough University.

Rosenow, J., Platt, R. and Flanagan, B. (2013). Fuel poverty and energy efficiency obligations – A critical assessment of the supplier obligation in the UK. Energy Policy, 62, pp.1194-1203.

Rowley, P., Leicester, P., Palmer, D., Westacott, P., Candelise, C., Betts, T., Gottschalg, R., (2015), Multi-domain Analysis of Photovoltaic Impacts via Integrated Spatial & Probabilistic Modelling, IET Renewable Power Generation, pp.1-8, ISSN 1752-1416

Rylatt, R., Gadsden, S. and Lomas, K. (2003). Methods of predicting urban domestic energy demand with reduced datasets: a review and a new GIS-based approach. build serv eng res technol, 24(2), pp.93-102.

Şahin, Ş., Ülengin, F. and Ülengin, B. (2006). A Bayesian causal map for inflation analysis: The case of Turkey. European Journal of Operational Research, 175(2), pp.1268-1284.

Saltelli, A. and Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. Environmental Modelling & Software, 25(12), pp.1508-1517.

Saltelli, A. (2008). Global sensitivity analysis. 1st ed. Chichester, England: John Wiley.

Saluja, G. and Muneer, T. (1987). An Anisotropic Model for Inclined Surface Solar Irradiation. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 201(1), pp.11-20.

Savage, S., Danziger, J. and Markowitz, H. (2012). The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty.

Scharmer, K. and Greif, J. (2000). The European solar radiation atlas. 1st ed. Paris: Les Presses de l'Ecole des Mines.

Seyfang, G., Park, J. and Smith, A. (2013). A thousand flowers blooming? An examination of community energy in the UK. Energy Policy. 61. pp. 977-989.

Seyfang, G., Hielscher, S., Hargreaves, T., Martiskainen, M. and Smith, A. (2014). A grassroots sustainable energy niche? Reflections on community energy in the UK. Environmental Innovation and Societal Transitions, 13, pp.21-44.

(Shenoy, P.P. and Shafer, G. (1990). Axioms for Probability and Belief-Function Propagation, Classic Works of the Dempster-Shafer Theory of Belief Functions Volume 219 of the series Studies in Fuzziness and Soft Computing pp 499-528.

Shipworth, M., Firth, S., Gentry, M., Wright, A., Shipworth, D. and Lomas, K. (2010). Central heating thermostat settings and timing: building demographics. Building Research & Information, 38(1), pp.50-69.

Shipworth, D. (2013). The Vernacular Architecture of Household Energy Models. Perspectives on Science, 21(2), pp.250-266.

Shorrock, L.D. and Anderson, B.A. (1995). A guide to the development of BREDEM, BRE Information Paper IP 4/95, Building Research Establishment, Watford, UK.

Shorrock, L. and Dunster, J. (1997). The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock. Energy Policy, 25(12), pp.1027-1037.

Shorrock, L.D., Dunster, J.E, Seale, C.F., Eppel, H. and Lomas, K.J. (1994). Testing BREDEM-8 against measured consumption data and against simulation models. Proc. BEPAC Conf. Building Environmental Performance, 1994.

Short, W., Packey, D. and Holt, T. (2005). A manual for the economic evaluation of energy efficiency and renewable energy technologies. Honolulu, Hawaii: University Press of the Pacific.

Singh, H., Muetze, A. and Eames, P. (2010). Factors influencing the uptake of heat pump technology by the UK domestic sector. Renewable Energy, 35(4), pp.873-878.

Skartveit A. and Olseth J.A., (1992). The probability density and autocorrelation of short-term global and beam irradiance, Sol Energy, 49(6), pp 477–487.

Skoplaki, E. and Palyvos, J. (2009). On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. Solar Energy, 83(5), pp.614-624.

Smith, J. (2010). Bayesian decision analysis. Cambridge: Cambridge University Press.

Solarbuzz (2013). Multicrystalline Silicon Modules to Dominate Solar PV Industry in 2014, According to NPD Solarbuzz. [online] Available from: http://www.solarbuzz.com/news/recent-findings/multicrystalline-silicon-modules-dominate-solar-pv-industry-2014 [Accessed 13-Mar-2014].

Steemers, K. and Yun, G. (2009). Household energy consumption: a study of the role of occupants. Building Research & Information, 37(5-6), pp.625-637.

Stern, N., 2005. Stern Review on the Economics of Climate Change. London

Strunz, S. (2014). The German energy transition as a regime shift. Ecological Economics, 100, pp.150-158.

Sudtharalingam, S., Leach, M., Brett, D., Staffell, I., Bergman, N., Barton, J., Kelly, N., Brandon, N., Infield, D., Peacock, A., Baker, P., Woodman, B., Hawkes, A., Blanchard, R., Jardine, C. and Matian, M. (2010). UK microgeneration. Part II: technology overviews. Proceedings of the ICE - Energy, 163(4), pp.143-165.

Suri, M. and Hofierka, J. (2004). A New GIS-based Solar Radiation Model and Its Application to Photovoltaic Assessments. Transactions in GIS, 8(2), pp.175-190.

Šúri, M., Huld, T., Dunlop, E. and Ossenbrink, H. (2007). Potential of solar electricity generation in the European Union member states and candidate countries. Solar Energy, 81(10), pp.1295-1305.

Šúri,M., Remund, J., Cebecauer, T., Dumortier, D., Wald, L., Huld, T. and Blanc, P. (2008). First Steps in the Cross-Comparison of Solar Resource Spatial Products in Europe. Proceeding of the EUROSUN 2008, 1st International Conference on Solar Heating, Cooling and Buildings, Lisbon, Portugal.

Šúri, M. (2007). Solar resource data and tools for an assessment of photovoltaic systems. In: J. Arnulf, ed., Scientific Technical Reference System on Renewable Energy & Use Efficiency - Status Report 2006, 1st ed. Luxemburg: JRC Publications Repository of the European Communities.

Swan, L. and Ugursal, V. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews, 13(8), pp.1819-1835.

Taylor, J. and Buckley, A. (2014). Personal communication on the distribution of system module technologies held on the Sheffield Microgeneration database.

Taylor, J., Davies, M. and Lai, K. (2010). The Simulation of the Post Flood Drying of Dwellings in London. In: *International Conference on Sustainable Built Environment (ICSBE)*.

Telenko, C. and Seepersad, C. (2014). Probabilistic Graphical Modeling of Use Stage Energy Consumption: A Lightweight Vehicle Example 1. Journal of Mechanical Design, 136(10), p.101403.

Thevenard, D. and Pelland, S. (2013). Estimating the uncertainty in long-term photovoltaic yield predictions. Solar Energy, 91, pp.432-445.

Thirkhill, A. (2015). Evaluating the Uncertainty in the Performance of Small Scale Renewables. Thesis, Loughborough University.

Tian, W. (2013). A review of sensitivity analysis methods in building energy analysis. Renewable and Sustainable Energy Reviews, 20, pp.411-419.

Ticehurst, J., Newham, L., Rissik, D., Letcher, R. and Jakeman, A. (2007). A Bayesian network approach for assessing the sustainability of coastal lakes in New South Wales, Australia. Environmental Modelling & Software, 22(8), pp.1129-1139.

Tronchin, L. and Fabbri, K. (2012). Energy Performance Certificate of building and confidence interval in assessment: An Italian case study. Energy Policy, 48, pp.176-184.

Tweed, C. (2013). Socio-technical issues in dwelling retrofit. Building Research & Information, 41(5), pp.551-562.

UN (United Nations) (1997). Kyoto Protocol to the United Nations Framework Convention on Climate Change,

UNCED (United Nations Conference on Environment and Development), (1992). Rio Conference and Earth Summit, Rio de Janeiro, June 1992

Unruh, G. (2000). Understanding carbon lock-in. Energy Policy, 28(12), pp.817-830.

Unruh, G. (2002). Escaping carbon lock-in. Energy Policy, 30(4), pp.317-325.

UOSRML (University of Oregon Solar Radiation Monitoring Laboratory) (2014). Sun Path Chart Program [online] Available from: http://solardat.uoregon.edu/SunChartProgram.html [Accessed 10-Jan-2014].

Summerfield, A., Lowe, R. and Oreszczyn, T. (2010). Two models for benchmarking UK domestic delivered energy. *Building Research & Information*, 38(1), pp.12-24.

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling, 203(3-4), pp.312-318.

Vallverdú, J. (2003). The False Dilemma: Bayesian vs. Frequentist. XIIth International Congress of Logic, Methodology and Philosophy of Science, held in Oviedo.

van der Schoor, T. & Scholtens, B. (2015). Power to the people: Local community initiatives and the transition to sustainable energy. Renewable and Sustainable Energy Reviews. 43. pp. 666-675.

van der Welle, A. and de Joode, J. (2011). Regulatory road maps for the integration of intermittent electricity generation: Methodology development and the case of The Netherlands. Energy Policy, 39(10), pp.5829-5839.

Varis, O. and Kuikka, S. (1999). Learning Bayesian decision analysis by doing: lessons from environmental and natural resources management. Ecological Modelling, 119, pp.177–195.

Verbong, G. and Geels, F. (2010). Exploring sustainability transitions in the electricity sector with socio-technical pathways. Technological Forecasting and Social Change, 77(8), pp.1214-1221.

Voss, K., Sartori, I., Napolitano, A., Geier, S., Goncalves, H., Hall, M., et al. (2010). Load matching and Grid Interaction of Net Zero Energy Buildings. EuroSun, Graz Austria, September 29th-October 1st.

Weber, P. and Jouffe, L. (2006). Complex system reliability modelling with Dynamic Object Oriented Bayesian Networks (DOOBN). Reliability Engineering & System Safety, 91(2), pp.149-162.

Weidl, G., Madsen, A. and Israelson, S. (2005). Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes. Computers & Chemical Engineering, 29(9), pp.1996-2009.

Wenham, S. (2011). Applied photovoltaics. London: Earthscan.

Whitehead, C., Monk, S., Clarke, A., Holmans, A. and Markkanen S. (2009). Measuring Housing Affordability: A Review of Data Sources. Cambridge Centre for Housing and Planning Research.

Wieder, S. (1982). An introduction to solar energy for scientists and engineers. 1st ed. New York: Wiley.

Wiginton, L., Nguyen, H. and Pearce, J. (2010). Quantifying rooftop solar photovoltaic potential for regional renewable energy policy. Computers, Environment and Urban Systems, 34(4), pp.345--357.

Wild, M., 2009. Global dimming and brightening: A review. J. Geophys. Res. 114, D00D16.

Wood, S. and Rowley, P. (2011). A techno-economic analysis of small-scale, biomass-fuelled combined heat and power for community housing. Biomass and Bioenergy, 35(9), pp.3849-3858.

Woodman, B. and Mitchell, C. (2011). Learning from experience? The development of the Renewables Obligation in England and Wales 2002–2010. Energy Policy, 39(7), pp.3914-3921.

Yao, R. and Steemers, K. (2005). A method of formulating energy load profile for domestic buildings in the UK. Energy and Buildings, 37(6), pp.663-671.

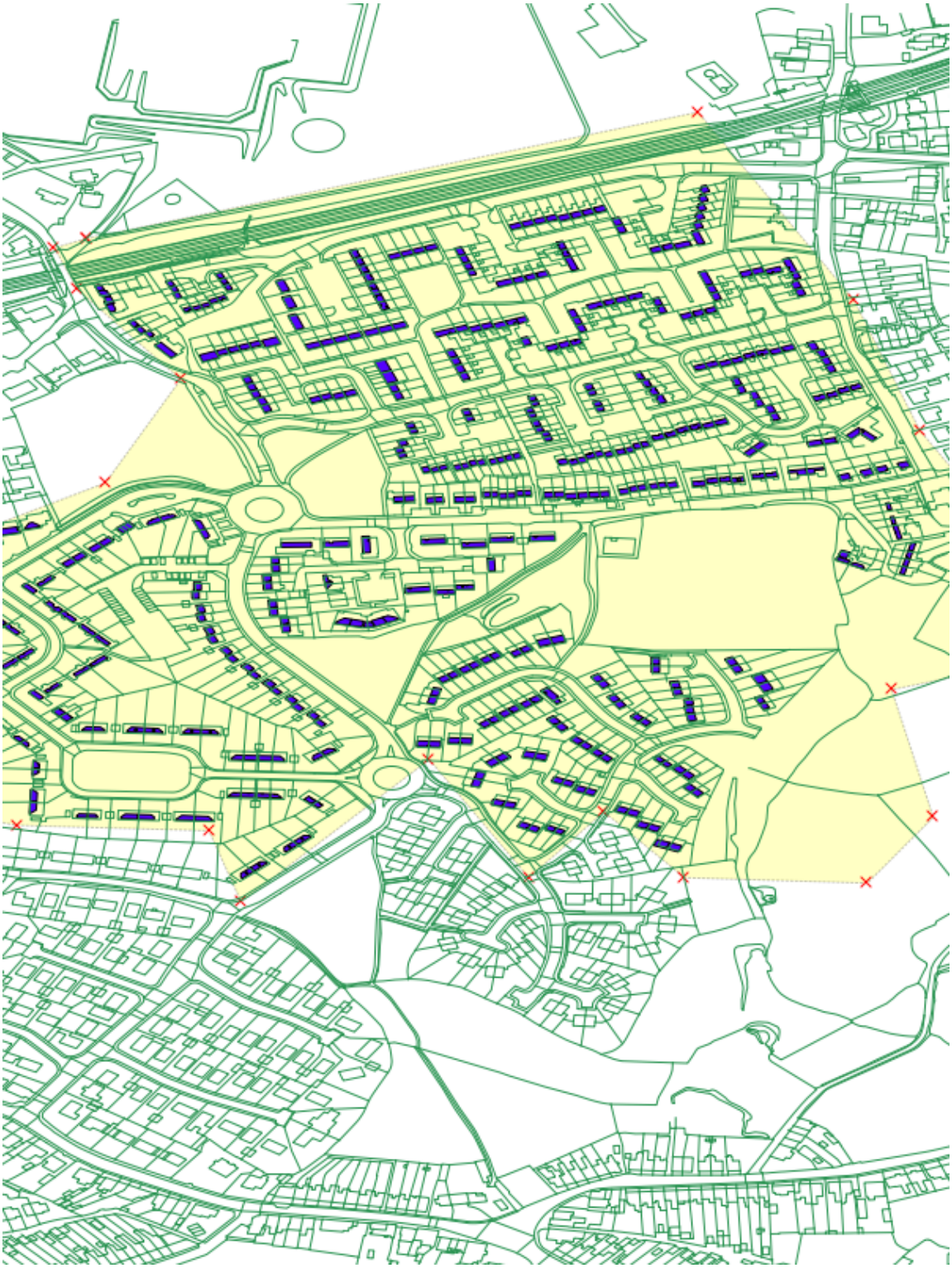
Yonder Haar, T. H., and J. S. Ellis, (1978) Determination of the solar energy microclimate of the United States using satellite data. Final Report, NASA Grant NA5-22372, Colorado State University.

Zelenka, A., Czeplak, G., D'Agostino, V., Josefson, W., Maxwell, E. and Perez, R. (1992) Techniques for Supplementing Solar Radiation Network Data. Geneva, International Energy Agency Technical Report No IEA-SHCP-9D-1

Zhou, Y., Fenton, N. and Neil, M. (2014). Bayesian network approach to multinomial parameter learning using data and expert judgments. International Journal of Approximate Reasoning, 55(5), pp.1252-1268.

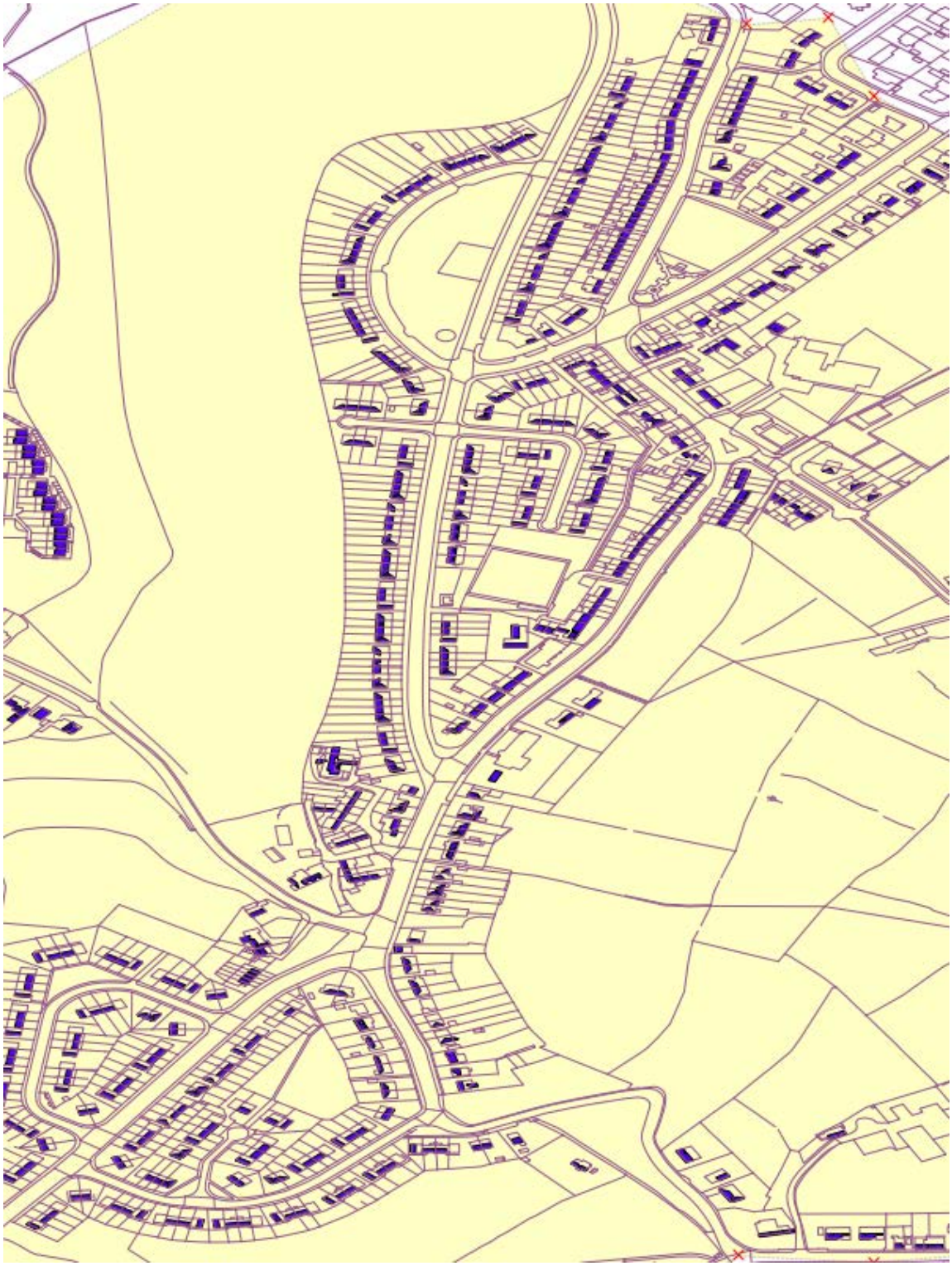
Appendix 1. GIS images of Census Area Building Stock

11.1 LSOA Kerrier 008B



11.2 LSOA Charnwood 002D

11.3 LSOA Kirklees 042B



11.4 LSOA Newcastle 008G

This Page Intentionally Blank

Loughborough University 2015