Smart Network Design Methodologies

Novel Analysis Techniques at Low Voltage
Final Report

June 2019

Contents

0 Document control

0.1 Document history

Version	Status	Issue Date	Authors
1	Interim report	20/12/18	G. Edwards G. McFadzean
2	Final report	17/05/19	C. Higgins G. Edwards G. McFadzean

0.2 Document review

Name	Responsibility	Date
Francis Shillitoe	Project Manager (NPg)	17/05/2019
Alan Creighton	Technical Lead (NPg)	17/05/2019

0.3 Document sign-off

Name	Responsibility	Date
Mark Nicholson	Project Sponsor	11/06/2019

Executive Summary

This report introduces novel analysis techniques that have been developed in order to help a Distribution Network Operator (DNO) gain an understanding of the risk of unacceptable thermal loading and voltage excursions on their Low Voltage (LV) networks. The method accounts for both:

- The fundamental *variability* in a given customer's demand over a period of time, which is only partially predictable; as well as
- The significant *uncertainty* that DNOs are faced with when trying to model demand variability for a network supplying multiple customers, who may use their electricity in very different ways.

The method is designed to incorporate network-specific demand data for refining estimates of customer demand, as this becomes available, with particular attention given to smart meter data.

Motivation

With the exception of designing new connections, there has historically been little need to monitor or model LV networks, which means that existing models and the data they require can be reasonably simple. Like the other DNOs in Great Britain (GB), Northern Powergrid (NPg) tend to use one of two existing methods when estimating demands for LV networks:

- The ACE 49 method, developed in the 1970s and 1980s, which describes a simple statistical model for understanding the demand for group of customers types that could be supplied from an LV network. There are several strong assumptions made in order to derive this model, including the definition of a 90th percentile level of risk. In addition, the detail and intent of some important aspects of the method aren't completely clear.
- The After Diversity Maximum Demand (ADMD) method, which generally refers to empirically determined values of the per customer demand, for a group of N customers. By definition, ADMDs don't relate to a level of network risk.

Both methods have shortcomings, and these are discussed in more detail in Section 2.2. One weakness of both methods is that, in their typical use, they assume that all customers of a specific type exhibit the "average" behaviour for that type of customer. Trial data shows that this is not the case when dealing with small groups of customers. This is discussed in more detail in Section 2.2.3. Another is that power flows on the network are often not studied, or are studied in a simplified way. DNOs have used these methods for many years to provide customers with sufficient network capacity, which has helped to ensure that their quality of supply is not compromised.

There is the potential for rapid increase of demand on LV networks, particularly due to customers adopting Low Carbon Technologies (LCTs) in support of Government objectives for decarbonisation heat and transport. In the future, these methods may cease to be suitable for designing efficient and secure LV networks.

The roll-out of smart metering and LV monitoring could provide DNOs with a rich source of data to improve their LV design practices. Enhanced design processes and methods are required in order to incorporate this information and continue to design economically efficient networks.

Novel analysis techniques for LV design which utilise this richer data, like those presented here, are more suitable for managing uncertainty about customer demands, and represent an evolution towards more efficient risk-based network design methodology. Benefits would include:

- Improved business planning and more efficient investment programme for LV networks.
- Helping to inform the appropriate level of risk for networks.
- Improved understanding of the reduction in risk associated with network development.

Overview of method

The foundational principles of the ACE49 method are still very appropriate for LV network planning. In particular, the use of a statistical model to represent customer demands will continue to be appropriate, particularly as the adoption of LCTs could lead to far more uncertain patterns of demand on the LV network. Such a model allows the network to be thought about in terms of risk and uncertainty.

Our novel analysis techniques are therefore rooted in the use of a statistical model to reflect both the *variability* and *uncertainty* in demand on LV networks. More specifically, we have proposed the use of a *Bayesian statistical model* for representing demand. Bayesian statistics are described in more detail in Section 2.3, but the key points to note are:

- In Bayesian statistics, probabilities are viewed as representing subjective beliefs, rather than the long-run frequency of some measured phenomenon. This is important when dealing with problems for which there is not much data. Initial beliefs are formalised mathematically as *'prior probability distributions'*.
- Our method proposes that prior probability distributions should be formed based on existing data sets from projects such as NPg's Customer Led Network Revolution (CLNR) project. Bayesian statistics allows for the initial prior beliefs to reflect the uncertainty which exists when trying to understand the demand for customers at the LV level, without specific local data.
- When data becomes available from smart meters or LV monitoring, it can be used to update the 'prior', according to a procedure known as Bayesian updating, to form a *'posterior probability distribution'*. It is expected that this will reduce the uncertainty in the estimate.
- This procedure can be repeated indefinitely every time new data becomes available. Eventually, the initial prior belief will have very little influence on the modelled demand.

Section 3.2 sets out how the Bayesian approach could be layered on top of the probabilistic model using Gamma and Weibull distributions as described in the Smart Meter Data Analytics report.

Demonstration

Our method also captures the impacts which these demands will have on the network, in terms of thermal utilisation and voltage excursion, to inform network planning and new connection designs, based on detailed AC power flow modelling. However, rather than attempting to run 100,000s of AC power flow simulations using Monte-Carlo sampling, we have proposed a method which *decouples* the AC power flow modelling from the modelling of demand.

This is achieved by running large sets of customer demands through a network model, storing the outputs, and then regressing the outputs of the AC load flow against the demand inputs. This would require considerably fewer samples than a conventional Monte-Carlo AC load flow. This takes advantage of the fact that, even though AC load-flow is non-linear, by observing the results for 1,000s of combinations of demand, it should always be possible to estimate the results for all credible demands. For particularly complex network, machine learning methods may be used to explore different regression models, although these have not been demonstrated here.

These novel analysis techniques are demonstrated for two LV networks in Section 4. In the cases we have looked at, it has been possible to account for the majority of network behaviour using simple single-variable linear regressions. We have only done this for case studies that look at high demand, although this could be easily extended to also look at low demand cases where there is embedded generation. We have demonstrated how these techniques can be used to calculate risk-based "exceedance expectations" that inform a network planner of the long-term frequency of voltage limit violations or thermal overloads.

Key learning

Some of the key learning generated from the work is summarised below:

- Existing methods provide good foundations for LV network planning, and novel analysis techniques can evolve these so that new sources of data can be leveraged to address emerging challenges. In particular, the use of a statistical model, as within the ACE49 method, should be retained and incrementally enhanced.
- An appropriate statistical model for describing customer demand on an LV network, in a situation where there is limited data available to a network designer, has been developed and demonstrated. Approaches for integrating new sources of data within this model, as these become available, to help improve the precision of models have also been proposed.
- Even within a specifically defined customer type, there is still significant variability in the patterns of demand. Therefore, the benefits of reflecting certain customer type categorisation, such as MOSAIC, may not be justified, given the minor reduction in uncertainty.
- Variability and uncertainty should be propagated throughout the model, rather than just studying the impacts on the network of "average" customer demands. A method for avoiding the need for computationally expensive Monte-Carlo AC load flows has been demonstrated.
- The risk related to unacceptable thermal loading and voltage excursions is multi-faceted, particularly because DNOs do not have access to high quality data about how their LV networks are used. The method we proposed systematically accounts for multiple sources of risk, including the fundamental and inherent variability in customer demand as well as the uncertainty that the DNO is faced with due to lack of information.

Next steps

On the basis of the analysis presented in this report, it would be possible to construct a model which calculates the risk for LV network thermal loading and voltage violations due to load, including both generation and demand. Issues for near-term implementation will be considered in the project through development of a functional specification. Implementation issues could include:

- Determining how DNOs should think about load related risk in their LV network planning
- Practical use of smart-meter data, given challenges such as data privacy and partial penetration of smart-meters, although the fundamental options have already are been considered, and are presented and discussed here in Section 5.1.2 and Section 5.1.3.

Further developments of the novel analysis techniques would be beneficial, including:

- Capturing the behaviour of Low Carbon Technologies and their interactions with existing demand, when their penetration is much deeper than current levels.
- Improving predictions of "extreme" events, such as very infrequent demands, using a branch of statistics called extreme value theory (although the current model does have this ability).
- Assessing more complicated meshed networks, or networks with lots of imbalance across phases, which may require the use of multi-variate statistics.

We have described how the modelling methodology could address some of these implementation and modelling challenges in Section 5.

1 Introduction

With growing low carbon generation and demand and correspondingly more dynamic, controllable loads on the LV network, and with the improved ability to monitor behaviour through smart meters coming in the future, there is an increasing need to move towards a more sophisticated LV network modelling approach. Integration of smart meter data as well as other sources of typical network monitoring data e.g. Elexon Profile Class 5-8 customer data, with LV network modelling will provide significant value to LV network planning.

The project objective is to explore how smart metering data in combination with other sources of smart monitoring and existing network data can be used to improve the planning and design of the distribution network. This technical report forms part of the project and is focussed on development and demonstration of novel network analysis techniques at LV, based on use of smart meter data.

1.1 Risk-based network planning

The current security of supply methodology ER P2/6 is deterministically applied and based on a single peak demand rather than a statistical representation of peak demands. A capacity assessment then identifies any areas where reinforcement is required based on the forecasted peak demand. In the context of energy system changes, industry is now beginning to realise that implementing a more probabilistic methodology has the potential to utilise existing capacity more effectively through better understanding and characterisation of customer loading and how different types of distributed energy resources could contribute to providing demand security. This prospect has been examined extensively in the ongoing review of the ER P2 standard where the benefits and challenges of implementing more probabilistic supply security assessment methodologies have been considered.

In the draft ER P2/7 security of supply document, emphasis is on defining the minimum level of security of supply that should be achieved rather than how that level should be achieved.

1.1.1 Opportunities from new data sources

The increasing adoption of smart meters is providing a significant opportunity to collate and analyse data on customer load characteristics, enabling more accurate characterisation of networks, and facilitating the adoption of a robust probabilistic approach to network planning. There is also a wider application of monitoring across distribution networks, including at low voltage, to facilitate greater understanding and control of networks coupled with increased digitalisation of a wide range of geographic and demographic data. This includes monitoring to enable a number of smart solutions such as active network management, demand side response, voltage control and dynamic asset ratings. These data sources can all contribute to an improved understanding of customer and network behaviour within a statistical framework albeit within the constraints of customer privacy.

1.2 Summary of Approach

The approach that we have developed and present in further detail in this report is formulated based on advanced statistical techniques that apply to both customer demand and corresponding network states. It incorporates the following features:

- Sophisticated data-driven statistical modelling.
- Risk based – identifies demands that exceed circuit capacity, their impact and their frequency.
- Can extend to model patterns in simultaneous demands at multiple nodes, , for situations where the state of a network component cannot be determined with sufficient accuracy by a single aggregated demand e.g. because the network has a large non-domestic customer with an atypical pattern of demand connected at the end of a feeder.

- Takes particular care to ensure that the high and low extremes observed over multiple years are accurately modelled.
- Dynamic and flexible, can be updated to incorporate new data, such as increasing availability of smart meter data, and learning about new technologies and their uptake.

This approach will facilitate more efficient network planning in a changing energy system, maximising use of increasingly granular customer and network monitoring data.

1.3 Objectives

The objectives of the novel network analysis techniques at LV are as follows:

- Develop a statistical LV network customer load model capable of incorporating smart meter data, along with other data such as network monitoring and annual energy consumption values. It must also include consideration of the future uptake of low carbon technology (LCT).
- Develop a computationally efficient method for combining the customer demand model with a network model to obtain a probabilistic representation of network states.
- Develop a method that combines this probabilistic representation with rules[1] set by DNO managers to make automated decisions regarding network design or reinforcement requirements.
- Compare our method with existing methods, and demonstrate the core functionality of these novel analysis techniques.

[1] Although, determining what these rules should be is not in the scope of the project.

2 Background

2.1 Existing NPg approach to LV network planning

Currently, LV network planning and design (including for new connections) is carried out using several design tools across licence areas as summarised below:

- LV Design (DEBUT) (Northeast): LV Design (DEBUT) is an older version of WINDEBUT and is a graphical and analytical tool for LV network study. It performs thermal, voltage drop, fault current and earth fault loop impedance studies on the LV network, although in most use-cases these are based on average or typical relationships rather than power flow analysis of specific circuits. The user selects a customer type for each connection, along with either the customer's annual energy consumption – if available – or otherwise the average energy consumption for the customer type, and this is translated into a peak demand value.
  Whilst DEBUT includes a wide range of profiles for different customer types, in general, the DEBUT method is used to represent the demand of domestic customers. However, using CLNR data, the DEBUT customer related parameters have been extended to include EVs and heat pumps (but no type of generation).
- Design Demand Calculator (DDC) (Northeast and Yorkshire): This Excel based tool provides the Equivalent ADMD (kW) for any given LV feeder depending on the number of customers and the type of load connected. Apart from normal domestic load, it also takes account of heat pumps and electric vehicle loads. Based on the Equivalent ADMD (kW) values, the DDC calculates the total design demand (i.e. total load) in kVA that any network asset, e.g. a supplying transformer will have to supply. The DDC calculates the Equivalent ADMD (kW) values for four typical loads which are assumed to represent the majority of the network load i.e. General Domestic (GD) with no electric heating, General Domestic (GD) with electric heating, GD with Heat pumps and GD with Electric vehicles.
- Network Calculator Tool 3 ph (Excel Tool) (Northeast): This Excel based tool is primarily used for motors (starting studies), welders fusing & flicker study. It provides guidance on the size of fluctuating load that can be connected within the EREC P28 limits, loop impedance, voltage fluctuation, fault current and fuse rating.
- LV Volt Regulation & Fault Level Calculator (Yorkshire): This Excel based in-house tool calculates the voltage drop, earth loop impedance, fault current and required fuse size.

2.2 Existing statistical approaches to network modelling

In this section, we examine in further detail the existing statistical modelling approaches which NPg use when undertaking network planning, including:

- The model within DEBUT, which is based on the ACE 49 report, and
- The "After Diversity Maximum Demand" approach.

We then show some comparisons of the estimates produced by each approach against the underlying CLNR data.

2.2.1 ACE49

Existing network planning, using the ACE49 method implemented in DEBUT, already incorporates a statistical model, with the following features:

1. Type of distribution: During the central winter period (November to March), the demand (kW) on an LV circuit with N customers for any individual half-hour fits a 'normal' distribution. This distribution is characterised by its mean (G) and its standard deviation σ. Distributions such as these are described in more detail in Section 2.3.

The distribution parameters, namely the mean and standard deviation, are assumed to be unique for each of the 48 half-hour periods within the daily cycle, often referred to in this report as the time-of-day. For this reason, it is necessary to introduce a subscript to the notation, indicating the time-of-day i.e. $G_{t,}, \psi_t$.

2. Estimation of parameters: The parameters of this distribution (the mean and the standard deviation) are assumed to depend linearly on annual energy consumption (kWh).

$$G_t = N \times C \times \psi_{G,t}$$

$$\sigma_t = N \times C \times \psi_{\sigma,t} \times \sqrt{{\sigma_1}^2 + {\sigma_2}^2 + \frac{{\sigma_3}^2}{N}}$$

N is the number of customers.

C is average annual energy consumption.

$\psi_{G,t}$ and $\psi_{\sigma,t}$ are proportionality constants for each half-hour.

${\sigma_1}^2$, ${\sigma_2}^2$, and ${\sigma_3}^2$ are the variance components arising from different types of variability: ${\sigma_1}^2$and ${\sigma_2}^2$ are connected to variations that each customer experiences with 100% correlations, while ${\sigma_3}^2$ is associated with variability that is completely independent for each customer.

The standard describes how, from this model, the design demand for each half hour period can be expressed in terms of customers' annual energy consumption C, the number of customers N, a "mean demand factor" p, and an "enhancement demand factor" q.

3. Data sources: The parameters (and therefore, the distributions) that describe the demand are calculated very infrequently based on single data sets. Our understanding is that p and q values (see below) were initially produced in the late 1970s/early 1980s, based on winter demand data observed sometime in the 1970s, and were updated in the mid-2010s, during the Customer Led Network Revolution (CLNR) project.
4. Level of risk: The level of demand for which networks should be designed is calculated, based on the maximum demand across all 48 distributions. This is the 'design demand', described as the level of demand that has only a 10% chance of being exceeded, which is "taken as an acceptable risk". In other words, actual demand will be less than the design demand for at least 90% of the time, which for any normal distribution is given by:

$$D_t = G_t + 1.28 \times \sigma_t,$$

In terms of the parameters defined above, this is:

$$N \times C \times \left(\psi_{G,t} + 1.28 \times \psi_{\sigma,t} \times \sqrt{{\sigma_1}^2 + {\sigma_2}^2 + \frac{{\sigma_3}^2}{N}} \right)$$

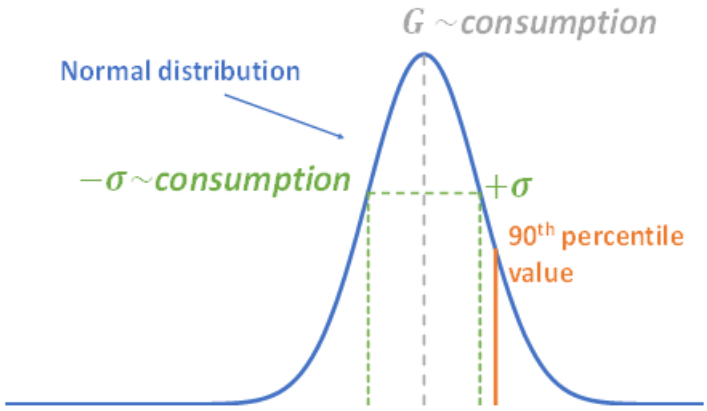
The design demand (for each half hour period) is then rephrased as:

$$D_t = N \times C \times \left(p_t + \frac{q_t}{\sqrt{N}} \right)$$

where p is the "mean demand factor" and q is the "enhancement demand factor". After calculating each D_t, the design demand D is set as the maximum value of D_t across all times of day. Our understanding of the intention of ACE49 is that p and q should be calculated empirically as parameters which provide the best fit to the data, rather than being expressed analytically. However, we are almost certain that the ACE49 approach simply takes p to be the mean demand per customer, while q is the mean of the total standard deviation per customer.

The main features of this model are represented graphically in Figure 2-1.

Figure 2-1 Probabilistic distribution of demand in a given half hour within ACE49



One important thing to note is that, although this is a statistical model, the statistical treatment is focused on the calculation of the design demand for the network via a deterministic process. Further, there is no sophisticated consideration of the relationship between the design demand and the actual condition of the network, namely power flow and voltage level. Statistical distributions for these network variables, specifically calculated for the unique features on a specific LV network, could in principle, be derived from a combination of the ACE49 demand model and a model of the network circuit. However, this is not a feature of the ACE49 approach, since the report methodology pre-selects a 90% risk level for demand, rather than some current or voltage level.

For a more detailed exposition of this statistical approach, please refer to the report "Review of the Distribution Network Planning and Design Standards for the Future Low Carbon Electricity System"[2] from CLNR.

2.2.1.1 Previous challenges to ACE49

The CLNR report referred to above raises some interesting challenges to the method, including one on "the basis of the statistical approach of the ACE49 standard". We have reproduced below a quote from this report, which will be explained in the subsequent paragraphs:

> "There are a several issues associated with the statistical modelling assumptions underpinning ACE49 which should be resolved in order to give the maximum degree of confidence in a planning approach for integrating LCTs. Most fundamentally, while ACE49 specifies design requirements in terms of a given percentile of a probability distribution, it does not specify clearly the definition of the variable of whose distribution this percentile is taken. There are also questions over the assumption of statistical independence between customers on a single feeder where LCTs are significant (i.e. supply or demand from some technologies such as solar PV may be highly correlated between properties). This report has concentrated on providing new datasets for use within the existing standards; however, as penetrations of LCTs become very high the issues raised in C6[3] should be addressed."

We have emphasised a key statement – this refers to the fact that the ACE49 method appears to take the 90th percentile value of the probability distribution for the demand on a given half-hour and on a given day

[2] http://www.networkrevolution.co.uk/wp-content/uploads/2014/12/ACE49-Report-1.1.pdf

[3] This relates to notation used within the CLNR report to identify different aspects of the model to be researched.

of the year as the design demand – i.e. the specific half hour that is used to establish the design demand. However, our interpretation of the ACE49 report is that it aims to design for a level of demand that would only have a 10% chance of being exceeded in a given winter. If so, it should be taking the 90th percentile value of the probability distribution for the maximum demand value that could occur at any time within the central winter period. This would be equivalent to a 1-in-10-year demand event.

It can be implied from the ACE 49 standard report that the method calculates the 90th percentile of 48 modelled distributions, which are the distributions for the demand on some given (but unspecified) weekday in the central winter period, during one of 48 time-of-day intervals of 30 minutes, e.g. 4pm – 4:30pm or 6:30pm – 7:00pm. It then takes the largest of these distribution percentiles to be the design demand. In reality, there are only 2 or 3 possible candidate half-hours (at most) that could plausibly be the time of day with the maximum 90th quantile of demand, but the largest among these generally changes for different customer numbers, N. This is significantly different from the 90th percentile of the maximum demand that could occur throughout the winter period.

To elaborate, although the ACE49 design demand will only be reached or exceeded 10% of the time during any single instance of the peak half-hour in winter, over the course of an entire winter, this level of demand is almost certain to be equalled or exceeded, most likely with a frequency of 1-in-10 days in the long-run. This is contradictory to the intention of the standard, which says that "a 90% probability of meeting the demand within the design voltage regulation was taken as an acceptable risk".

If, for example, a winter contains 90 working days, and we can assume for simplicity that the maximum demand could only occur between 4pm and 7pm[4], and - again temporarily, for convenience - that the demand distributions for those periods are identical, then this distribution is sampled 540 times in one winter. Calculating the maximum of these 540 samples is conceptually similar to tossing a coin multiple times and calculating the probability that at least one coin-toss comes up with a result of 'heads'[5]. After n coin tosses, the probability that at least one of the results in a 'head" is $(1 - 50\%)^n$. Even with a relatively small number of tosses, this becomes a near certainty – e.g. after 7coin tosses, the probability that none of them returned a result of 'heads' is only 0.8%. (In case the reader is concerned that the variability in demand around its mean values do not share the statistical independence of coin tosses, be assured that we are concerned with calculating the demand level that will be exceeded *on average* once every 10 years, and such correlations may be ignored in the calculation of averages)

This is illustrated in Figure 2-2, for individual events which only occur 10%, 1% and 0.5% of the time respectively. This shows that an event which will occur only 10% of the time is almost certain to occur (99.82% probability) if the observation is repeated only 60 times. For the statistical model, this means that the design demand (which is set at the 90th percentile level of the demand at a specific time) is almost certain to occur within any two-week period of winter. (This involves a 10% probability of the design demand being exceeded observed on a given time-step, repeated over 14 days of 6 measurements between 4pm and 7pm)

To get a design demand which is, as desired, only 90% certain to be equalled or exceeded over an entire winter, the percentile for the individual time-of-day distributions would have to be close to the 99.5th percentile (i.e. a level of demand which, in any individual half-hour only has a 0.5% chance of occurring).

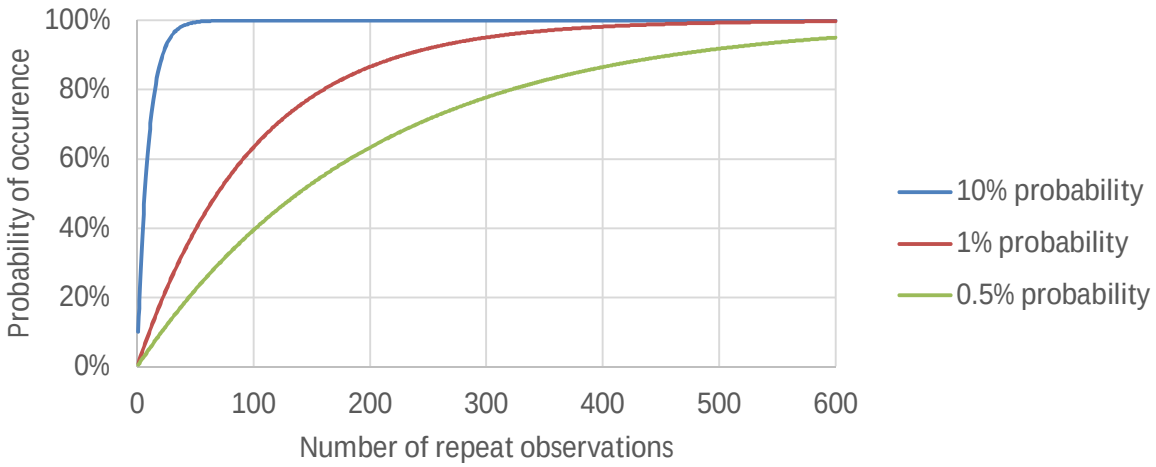
[4] We have assumed for simplicity that these events have the same distribution. It would be possible to repeat these calculations exactly using the underlying distributions.

[5] In practice, unlike concurrent tosses of a coin, where the outcomes are independent of each other, the demand in any give half-hour is likely to depend on the demand in the previous half-hour. We have ignored this effect for the purposes of this illustrative example. This can be justified by recalling that we are ultimately concerned with the mean frequency with which some demand level is exceeded, and correlations do not affect such mean values.

This analysis offers a compelling explanation why the ACE49 method systematically underestimates the observed customer demand for small numbers of customers (N), as highlighted in the NPg Code of Practice[6]. For larger N, the spread of the distribution become smaller relative to their mean values, and thus the extent of the error introduced by the misconception regarding distribution quantiles diminishes. This is because of diversity, i.e. fluctuations in the peak demand decreases with larger customer numbers, because customers switch appliances on and off at different times.

The design demands produced by ACE49 are compared to observed demand values from the CLNR datasets in Section 2.2.3.

Figure 2-2 Probability of at least 1 occurrence given repeated trials of a random variable



2.2.2 After Diversity Maximum Demand

The alternative popular approach to modelling aggregate demand is called the After-Diversity Maximum Demand (ADMD) approach. This is an entirely empirical yet deterministic approach, concerned with the maximum value of aggregated demand that is observed within a relatively large sample of a historical series – normalised by the number of customers in the aggregation group. More specifically, the method is concerned with fitting an analytical function to the rate at which the normalised peak demand diminishes as the number of customers in the group increases. The maxima can relate to a series of aggregated demands that were in fact witnessed on a real network, or an artificial aggregation of true and concurrent historical series that happened to occur on distinct and widely separated networks.

This approach and the associated metric are not truly risk-based, since they involve only a single observation of a derived random variable: i.e. the maximum demand experienced over 1 year, or some small number of years, for some combination of customers.

Some risk-based alternatives to this metric would be the *expected value* of the maximum observed out-turn over several years e.g. 1 in 2.5 years in the case of CLNR data, or e.g. the yearly maxima that have a 10% chance of being exceeded. Of course, ADMD values remain useful, e.g. ADMD values calculated from CLNR data serving as good proxies for the expected value of 1-in-2.5 years maxima for that dataset. However, the quality of the match is dependent on the quality of the deterministic modelling of the maximum observed demand for a given group size.

[6] IMP/001/911 Code of Practice for the Economic Development of the LV System, June 2018. https://www.northernpowergrid.com/asset/0/document/109.pdf

2.2.3 ACE49 and ADMD comparisons

This section examines the extent to which the ACE49 and ADMD approaches represent domestic customer aggregated peak demands, for the parameter values adopted by Northern Powergrid (in their application of ACE49 and ADMD), as established by studies conducted and published as part of the CNLR project, and naturally using the CLNR (TC1a) dataset. We have conducted an experiment whereby we simulated the aggregated demand series of 200 networks, or part of a network, with N domestic customers (i.e. General domestic customers with no electric heating). To achieve this, we randomly selected a group of customer profiles from those available in CLNR, and then calculated the total demand of this group. We have done this for groups of varying sizes, ranging from 1 to 120 customers, using smaller increments at the smaller end of the range.

Alongside this, we have calculated the design demand which would be provided for the same values of N by the ACE49 and ADMD methods as described in accordance with the NPg LV Design Code of Practice. The results of these trials are shown in Figure 2-3. The blue line indicates the mean result of the 200 "trials" for each value of N, with a range showing the 10th and 90th percentile values across those trials. The orange line shows the demand that would be produced using the ACE49 method, using the updated p and q values and the annual consumption for a URMC customer. The grey line shows the demand that would be produced by the ADMD approach, using the values from the LV Design Code of Practice.

The following observations are made based on these trials:

- ACE49 consistently underestimates demand for smaller values of N, whereas ADMD consistently overestimates it for all values of N.
- ACE49's underestimation is more pronounced for smaller group sizes. For larger N, the trial sample and the ACE49 design demand converge at approximately 1 kW per customer. ADMD, on the other, converges to a figure of approximately 2.1 kW per customer[7].
- There is significant variation in the peak demand produced by different groups of customers, especially for small values of N. Neither the ACE49 approach nor the ADMD approach account for this variation.

Figure 2-4 further demonstrates this variability, and how it diminishes for aggregated series involving increasing values of N. Specifically, it shows the decrease in the standard deviation of the 200 individual group maximum demand values per customer, normalised by their mean demand.

[7] Note that the LV Design Code of Practice quotes ADMDs in terms of the *marginal* ADMD of the nth customers. The additional ADMD imposed by the 100th customer is 1.7 kW, lower than the *average* ADMD of all 100 customers which is 2.1 kW.

Figure 2-3 CLNR Trial Demand Values vs ACE49 and ADMD representations, for 1 to 120 customers

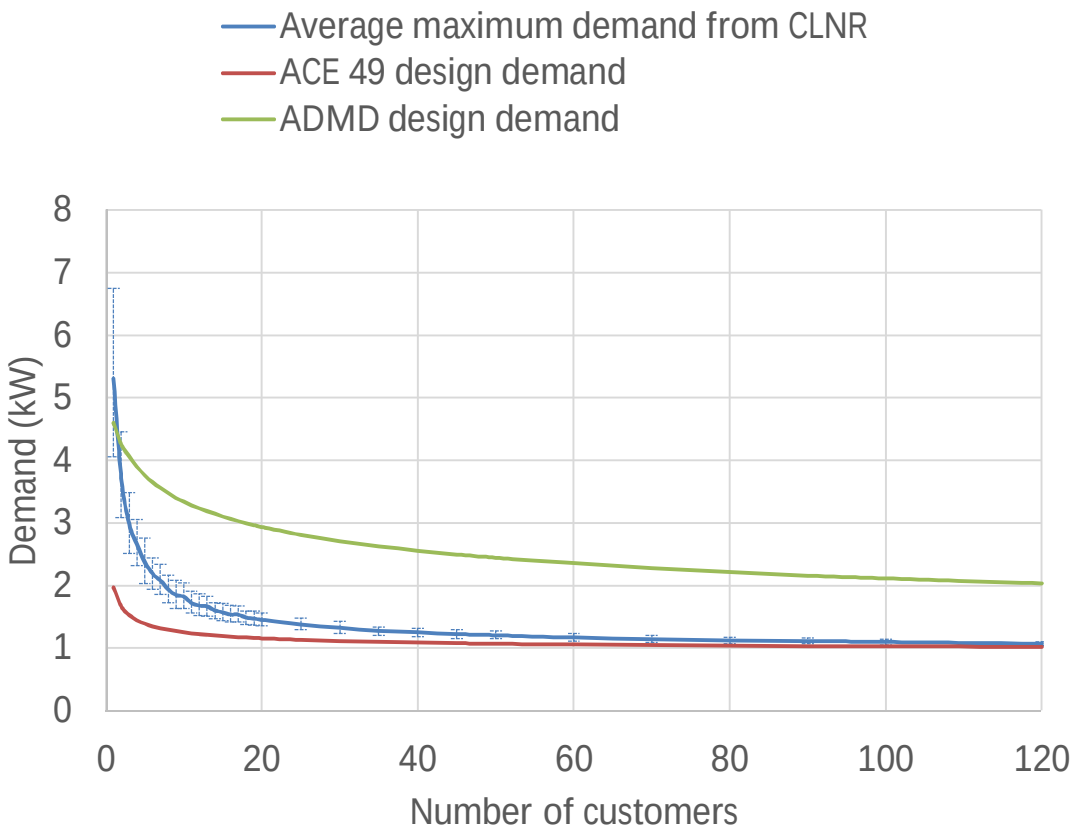
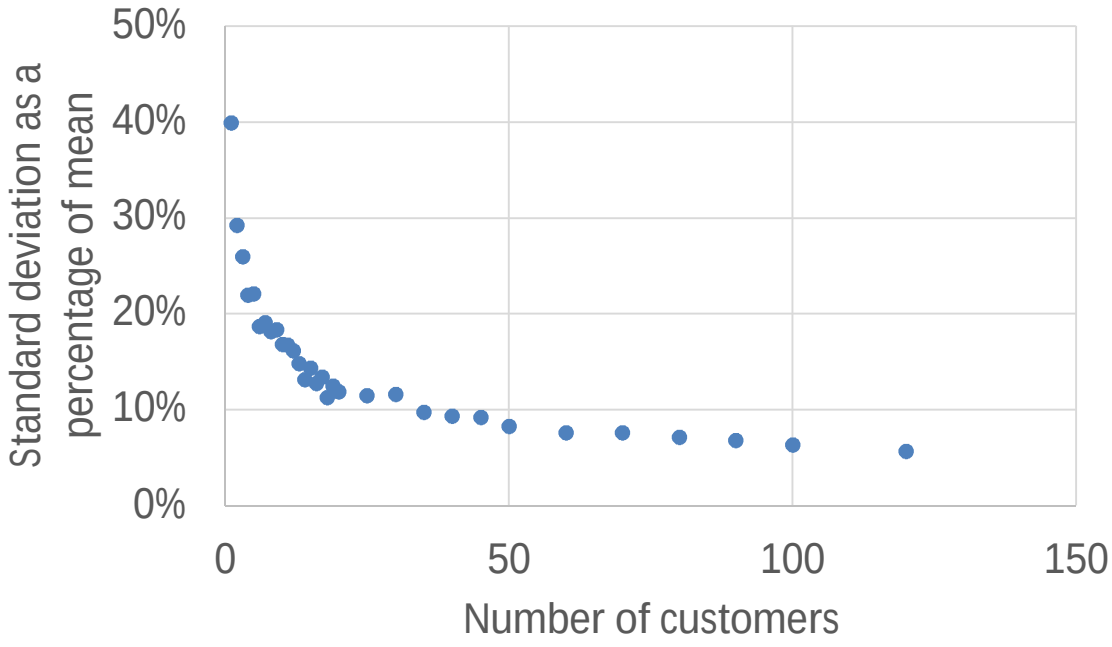


Figure 2-4 Standard deviation of series maxima as a proportion of the mean, for different customer numbers



2.3 Bayesian statistics

Much of our method is based on an approach to statistics known as 'Bayesian' statistics, and in particular 'Bayesian inference'. Although routinely used in a wide variety of real-world applications, its use may present the novice user with some challenges, due to the requirement of thinking about the fundamental nature of probabilities and samples in a slightly new way. To aid understanding of the method, this section provides an overview of these concepts, starting with a brief introduction of Bayes' Theorem.

2.3.1 Bayes' Theorem

Named after an 18th century statistician, Bayes' theorem describes the relationship between different conditional probabilities. The theorem is often expressed as follows:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

where A and B are events, $P(\,)$ is a probability of an event occurring, and the symbol | implies the probability of one event occurring, conditional on another event having occurred. For example, $P(B \mid A)$ is the probability of event B happening, given that A is true (i.e. 'has already happened').

Consider, as an example, that A is a randomly selected person in the world being Scottish, and B is a randomly selected person in the world having ginger hair. Since about 1% of the world's population has ginger hair, we may express this as $P(B) = 0.01$, and given Scotland's population of about 5.4 million compared to the world population of around 7.5 billion, we have $P(A) = 0.0007$ or 0.07%. As anyone who's visited Scotland knows, the proportion of red-heads is higher than 1%, with some estimates being 10%, which means in this case that $P(B \mid A) = 0.1$. An interesting question that arises is: if you know that someone has ginger hair, how does this change the probability that they're Scottish? Instinctively it is clear that the probability is somewhat increased, i.e. $P(A \mid B) > P(A)$, and Bayes' Theorem tells you how to calculate this using the other information available. In this case, the probability that a person is Scottish, given they have ginger hair, is

$$P(Scottish \mid Ginger\ hair) = \frac{P(Ginger\ hair \mid Scottish)\, P(Scottish)}{P(Ginger\ Hair)} = \frac{0.1 \times 0.0007}{0.01} = 0.007$$

or 0.7%, meaning that if someone has ginger hair, we should believe it is more likely that they are Scottish. As a further example, consider an application of medical diagnosis of a disease. Although possible to test patients for the disease, there is always a risk of "false positive" and "false negative" result, so a positive result does not necessarily mean someone has the disease or vice versa. Bayes Theorem' allows for a quantification of how likely the patient is to actually have a disease, dependent on their test results.

In this example, A is the event that the patient has the disease, and B is the event that the patient tests positive for the disease. For simplicity, assume that there is only a 1% rate of false positive and false negative results, and that the overall rate of occurrence of the disease in the population is 0.5%. That means:

- The probability of having the disease (before the test) is $P(A) = 0.005$, while the probability of not having the disease is $P(\,not\ A) = 0.995$.
- The probability of testing positive for the disease among those who truly have it is $P(B \mid A) = 0.99$.
- The probability of testing positive despite not having the disease is $P(B \mid not\ A) = 0.01$
- The total probability of the test result being positive, $P(B)$, is given by the (probabilistically) *weighted average* of the two conditional probabilities $P(B \mid A)$, and $P(B \mid not\ A)$, i.e.

$$P(B) = P(B \mid A) \times P(A) + P(B \mid not\ A) \times P(\,not\ A)$$

$$P(B) = (0.99 * 0.005) + (0.01 * 0.995) = 0.0149$$

- From Bayes' Theorem, we have that:

$$P(A|B) = \frac{P(B \mid A)\, P(A)}{P(B)} = 0.99 * \frac{0.005}{0.0149} = 0.3322$$

In other words, even though the test is quite accurate, the probability that the recipient of a single positive test result actually has the disease is only 33% - reflecting the fact that a false positive is more likely than a true positive, due to the prevalence of the disease. If we only expect 1 in 200 people to have the disease, and we know that the disease gives a false result 1 out of 100 times, then it is difficult to say with much certainty whether we have the disease on the basis of 1 test. The analysis used here has accurately accounted for our *prior* knowledge of the probability of occurrence of the disease.

One way to address this uncertainly would be to complete a second test. Applying Bayes' theorem again, we have:

- $P(B|A)$ and $P(B|not\ A)$ must be the same as previously, as they are inherent to the test.
- The positive result *has updated our prior knowledge*, $P(A)$ from 0.005 to 0.33.
- This means that $P(B) = (0.99 * 0.33) + (0.01 * 0.67) = 0.3334$
- Applying Bayes' Theorem, we have that

$$P(A|B) = 0.99 * \frac{0.3322}{0.3334} = 0.9864$$

This result reflects the fact that the probability of a true positive result is considerably smaller than two false positives, and thus the test recipient almost certainly has the disease. This process can be repeated many times, with our prior knowledge being updated as more data becomes available.

2.3.2 Bayesian inference

Bayes Theorem has applications in many different areas of statistics, including more conventional "frequentist" statistical methods, in which probabilities are treated as the long-run frequencies of the occurrence of events. However, there is a separate branch of statistics known as Bayesian statistics which makes use of an approach known as Bayesian inference. Bayesian inference uses Bayes' theorem along with a slightly different view of probability in order to make predictions which account for uncertainty. Rather than viewing probabilities as long run frequencies of a phenomenon, the Bayesian approach to statistical inference holds that all probabilities are subjective, and are a means of quantifying a degree of belief about phenomena. As a result, any belief that is uncertain – which includes a state of knowledge about the material world - is treated as a random variable.

For example, an experimenter might be interested in statistically modelling the height of children within a school class. Both the Bayesian inference approach and the more 'regular' approach to inference (i.e. learning) would consider the height of an individual child as a random variable, drawn from some probability distribution, most likely a normal distribution.

The more common (frequentist) view is that the distribution of children's height has some fixed set of parameters that is initially unknown. The more children are available for measuring, the closer the statistician is able to get to the true value of the parameters. The mean height for a specific classroom is a random variable, but there is assumed to be a true population mean (essentially the mean for all children in the world) that is not random. The Bayesian view, however, is that both the children's height and the parameters of the associated distribution are random variables. Further, both the mean height for a single classroom and for 'all children in the world' are random.

A reasonable approach to Bayesian inference would be for the statistician to take as their prior belief a distribution of heights for children of the same age and from the same country reported in some reputable source. The prior distribution would however be updated by the evidence specific to that class, i.e. the set of measured children's heights, with probability density functions (PDFs) – i.e. mathematical functions providing the probability that an observation falls within a certain range - being subject to Bayesian updating in exactly the same way that the probabilities of discrete events A and B were updated in the previous section.

In equation form, we may substitute in Bayes' theorem the event probabilities $P()$ with PDFs $p()$, the event A with our *prior belief* about the distribution of heights (e.g. the mean and standard deviation of the normal distribution), and the event B with the evidence that has become available – i.e. the height observations gathered, to get:

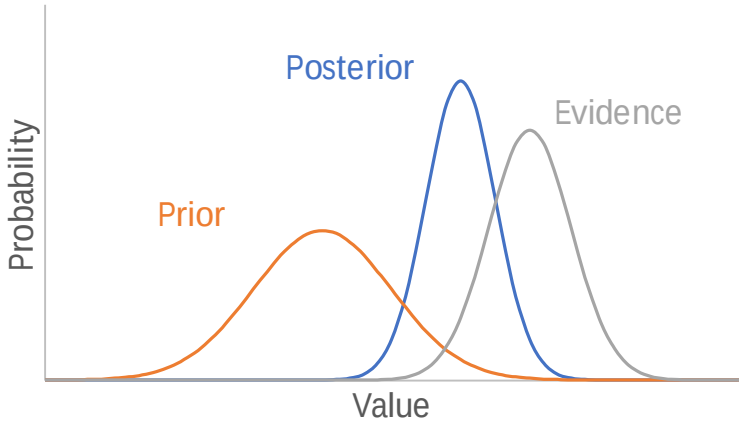
$$p(Belief \mid Evidence) = \frac{p(Evidence \mid Belief)\, p(Belief)}{p(Evidence)}$$

- The PDF of the belief, $p(Belief)$, is called the "prior" distribution
- The PDF of the belief after incorporating the evidence, $p(Belief \mid Evidence)$, is called the "posterior" distribution
- The function $p(Evidence \mid Belief)$ is the probability of observing the evidence, given some prior belief – e.g. some pair of mean and standard deviation values for our example. It describes how compatible the evidence is with any given prior belief. This is often referred to as the "likelihood" or "sample".
- The probability of the evidence taken as a weighted average over all possible parameter value combinations is the "marginal" distribution of evidence, $p(Evidence)$.

As in the example of the medical test, prior distributions can repeatedly be updated based on new sources of evidence – the previous "posterior" belief becomes the new "prior" belief with each new update.

Figure 2-5 shows how a prior belief and the likelihood function associated with some evidence can be combined to form a posterior distribution.

Figure 2-5 Bayesian inference example

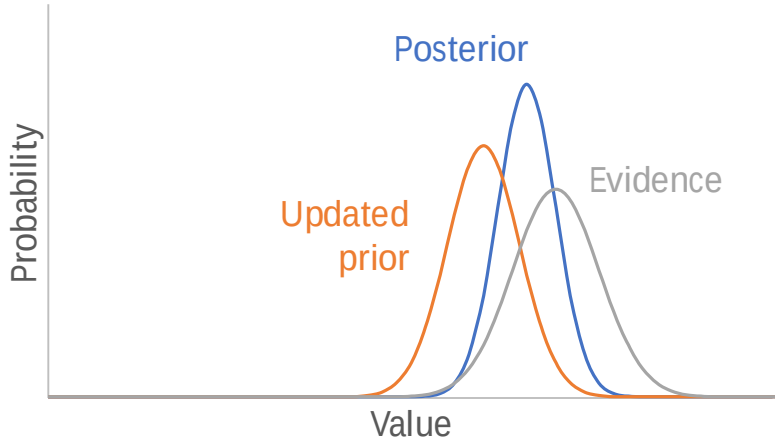


By adopting Bayesian inference, we essentially introduce a 'hierarchy' into our treatment of uncertainty. As previously stated, the probability distribution of the data is still dependent on some set of parameters (such as the mean and standard deviation) with the full set often denoted by the single mathematical symbol θ. However, we reflect the uncertainty in our belief by treating these parameters as a collection of random variables, themselves drawn from a distribution with its own set of 'hyper-parameters' represented by the single mathematical symbol α.

Indeed, the probability distributions in Figure 2-5 are actually distributions of the parameter values θ, (which we assume for convenience consists of only one random variable), rather than the directly observable random variables, i.e. customer demands in the present context. In the figure, we have a prior distribution for θ where the most likely values for the parameter are lower than the values most consistent with the evidence. Our initial level of certainty about the parameter value was low, and so the distribution has a relatively high variance. After observing some evidence with a relatively narrow likelihood function, a posterior distribution is formed with a peak that lies between the prior and the evidence likelihood peaks. As new evidence is incorporated, the posterior distribution will 'narrow', until we reach a point where we have a very thorough understanding of the phenomena.

This is illustrated in Figure 2-6 – the previous "posterior" forms an updated prior belief, and additional evidence causes the posterior to shift further.

Figure 2-6 Bayesian inference example with updated prior distribution



Recalling the example of the medical test, one of the quantities involved was $P(B)$, the marginal probability of the test outcome being positive, which was the probability-weighted sum of conditional outcomes over all possible outcomes of A – which in this case were two outcomes: the test recipient having the disease and not having the disease. In the case of customer demand for energy, if we are interested in e.g. the demand level that has a 1% chance of being exceeded, then the answer is generally determined by the values of the set θ. In order to get a marginal value, i.e. a value not dependent on a particular out-turn of the random θ, it is similarly necessary to take a weighted sum over all possible outcomes for θ. However, in the new example there are in fact an infinite number of possible values for θ, and so the probability-weighted sum becomes an integration (in the calculus sense).

By expressing prior beliefs probabilistically, uncertainty about this belief can be robustly quantified. This is particularly useful when dealing with small samples of data. It also formalises, within a mathematical framework, the processes by which people naturally incorporate their existing subjective views of the world. For example, if a network planner has a strong prior knowledge of a customer's demand which is incompatible with what some smart meter data says, Bayesian inference allows them to incorporate both pieces of information into their assessment, rather than just being forced to discard one of these pieces of information. For the same reasons, it is also very effective at dealing with outliers in data.

Our method, essentially, uses Bayesian inference as the means of estimating the parameters of the probability distributions which describe demand, in a manner which incorporates the (potentially significant) uncertainties associated with this. This is in contrast to ACE49, where parameters are estimated using a very simple deterministic assumed relationship between instantaneous demand and annual energy consumption. The model is explained in more detail in Section 3.2.

2.3.2.1 Bayesian inference example – tossing a coin

The usefulness of Bayesian inference is best further illustrated through an example. A common example is tossing a coin: if you only make a small number of coin tosses, what would you assume about the "fairness" of the coin?

For example, imagine picking up a random coin and tossing it five times, resulting in four 'heads' and one 'tail'. Traditional (also known as frequentist) statistics would estimate that the coin has an 80% chance of returning a value of heads, but that the sample size of the experiment is too small to state this with much confidence. However, this is not the conclusion that most people would be likely to draw - they would probably still assume that the coin was largely fair (i.e. that it has an approximately equal chance of resulting heads and tails). That's because people have a strong 'prior' belief about what the results of a coin toss is likely to be. As a result, a frequentist statistician might take the coin being fair as their null hypothesis (i.e. original view of the world), and calculate the extent to which the data forces them to reject that hypothesis.

In Bayesian statistics, such prior beliefs are incorporated in a more integral way, as prior distributions. The prior distribution of the coin's fairness (as defined by its hyper-parameters) will be closely centred around 50%, with a very small variation – this variation might be because people know that coins are not always perfectly fair due to minting imperfections, or that there are some trick coins in the world. Therefore, our posterior view of the fairness of the coin would not change much, at least, not until there were a much larger number of samples (evidence) suggesting the coin wasn't fair.

2.3.3 Bayesian prediction

It is important to emphasise again that the curves in figures 2-6 and 2-7 are the prior and posterior distributions of one of the *parameters* that define the probability distribution of the observable random variable. In our school class example, the observable quantity is the height of any child within the class, which we assume has a normal distribution. Therefore, the figures could represent the pdf of either the mean or the standard deviation of children's heights (both are random parameters and the same updating process occurs for both).

The natural question that arises is: how can I use these parameter distributions to make predictions about the height of an individual child – or whatever the observable variable may be. In Bayesian statistics, all statements and predictions we can make about the observable variable are encapsulated by a pdf known as the variable's 'predictive distribution'. More specifically, before the parameter distributions are updated due to the arrival of new data, the observable variable is represented by the *predictive prior distribution*, and after the arrival of data it is updated to become the *predictive posterior distribution*. The mathematics behind this are presented in the mathematical appendix.

The predictive distributions allow the modeller to make statements such as: "there is a 99% chance that a randomly selected child within the class is taller than x", or "there's a 50% chance that the height of a randomly selected child within the class is within the range a to b", or a simpler 'point forecast' such as "the predicted height of a randomly selected child within the class is h". The precise predictive distributions are obtained from the assumed distribution of the data, combined with a process of taking a probabilistically weighted mean across all possible values of the parameters – known as '*marginalising*' over those parameters. This process takes full and rigorous account of all model uncertainty, while very conveniently producing single-valued answers to questions such as 'what is the height that has only a 10% change of being exceeded by a randomly selected child in the class.

Considering the example of the coin, the frequentist predictive distribution for a new coin toss, given the results of the previous 5, would be a probability of 0.8 of getting a 'heads' (i.e. 80% chance) and a probability of 0.2 of getting a 'tail', by following the default method known as 'maximum likelihood

estimation'. However, the frequentist statistician could alternatively test whether the data is sufficient to reject their null hypothesis of a fair coin at a confidence level of 95%, would find that the answer is no, and their predictive distribution would be a probability of 0.5 for both outcomes. A Bayesian statistician, assuming that they choose their prior well, would derive a predictive distribution with e.g. a probability of 0.52 of getting a 'heads' and 0.48 of getting a 'tail'. After many 1000s of coin tosses, both the frequentist and Bayesian statisticians would derive predictive distributions that are extremely close to a probability of 0.5 for both outcomes.

Bringing consideration back to LV networks, the Bayesian methodology allows the following types of statements to be made: "for a randomly selected LV network feeder with 50 domestic customers, 3 of which have EVs and 3 of which have heat pumps, the level of aggregated demand that has a 10% chance of being exceeded between 5:30pm and 6:00pm on a randomly selected day in winter is d". Through further calculations, we can extend the statements that can be made to the following type: "for a randomly selected LV network feeder with 50 domestic customers, 3 of which have EVs and 3 of which have heat pumps, the level of aggregated demand that will be exceeded, on average, only once every 10 years is ".

The mathematics associated with producing posterior predictive distributions can be very challenging, or even impossible to solve numerically. Fortunately, sampling techniques can be used as computationally very cheap and intuitive alternatives, and the use of such methods will be presented in the case study section of this report. These have the additional benefit of illustrating the uncertainty in demand, and how this can change when new data is incorporated.

2.3.4 The opportunities for applying Bayesian inference to electricity demand modelling

Fundamentally, there are practical limits on the extent to which we can anticipate customer behaviour and their demands for electricity. For example, even if we had a very thorough understanding of all of the customers on a network, we would not be able to pinpoint the exact times of the day at which they might boil their kettles. That is, there is significant inherent randomness in how a customer uses electricity, which is still present even if we have a very full understanding of how that customer's characteristics drive that usage.

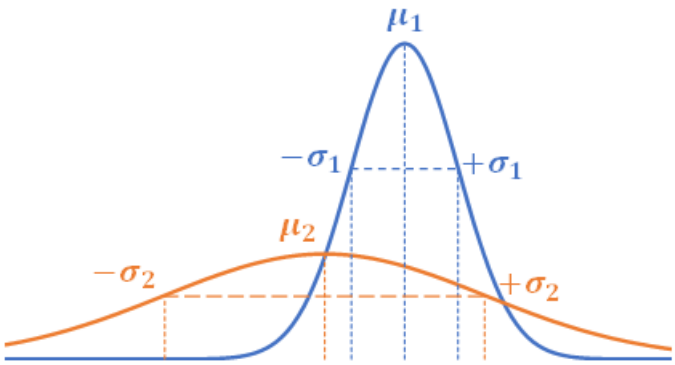
In the ACE 49 approach, that *inherent randomness* is captured by considering a probability distribution for how that specific customer will behave. This is illustrated in Figure 2-7 which show two examples of probability distributions – a 'normal' distribution. This distribution is characterised by:

- The expected value, μ, of the distribution[8]
- The standard deviation, σ, of the distribution (which describes the variance i.e. the extent to which values vary around the mean)

For example, there may be a category of customers whose demand, for a particular season/time-of-day, can be represented by a distribution with a mean of μ_1 and a standard deviation of σ_1, whereas another category of customer or the same customer at another season/time-of-day of would be represented by a distribution with a mean of μ_2 and a standard deviation of σ_2.

[8] Expected value and mean are used somewhat interchangeably – however, strictly speaking, a distribution has an expected value, whereas a sample has a mean.

Figure 2-7 Example of different normal distributions



The parameters which define a distribution are often referred to as the *parameter set*, and represented by θ. In the ACE49 approach, the parameter sets are 48 values of p and q, the annual energy consumption C, and the number of customers N, which are related analytically to the mean and standard deviation of a normal distribution.

If the characterisation of a group of customers were complete, the inherently random nature of demand means that the demands they place on a network at a specific time still cannot be completely predicted with 100% certainty – put precise probabilities can be associated with those demands exceeding some threshold or being within some range.

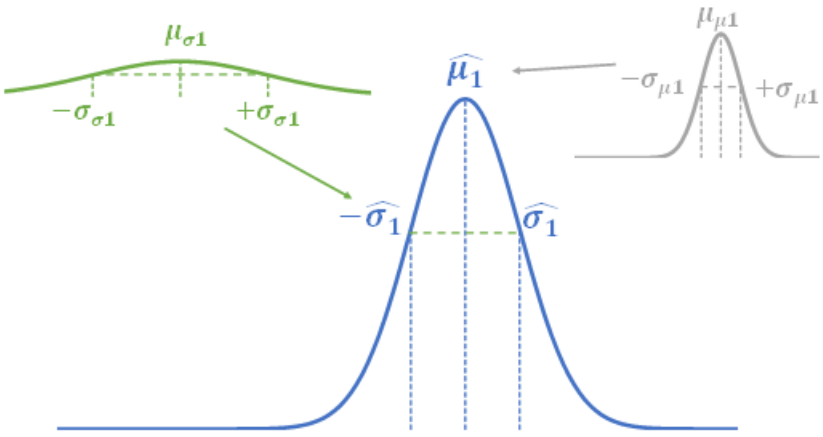
However, when the characterisation of the customers is incomplete, then even these probabilities may have considerable uncertainty associated with them[9]. There is always going to be some incompleteness in a network operator's characterisation of the customers on its network, due to the very partial information available.

Our analysis has demonstrated that this is certainly the case if characterising customers according to Mosaic categories, where the variability in statistics such as average demand observed within each category is roughly an order of magnitude greater than the variability between the average demand observed in different MOSAIC categories. There could also be uncertainty in the data which enables categorisation – for example, a DNO may be reasonably sure that a specific customer is of Category A, but depending on the data that is used to inform this, there may still be a chance that they are actually part of another category.

We believe that the only robust way to deal with this uncertainty is to explicitly acknowledge that, for any instance of a group of customers, there is significant uncertainty about the parameters that define the probability distribution of their combined demand. So, for a normal distribution, the mean μ_1 and standard deviation of σ_1 are themselves only one set of possible values from distributions. The mean μ_1 might actually be characterised by its own distribution, which has a mean $\mu_{\mu 1}$ and a standard deviation $\sigma_{\mu 1,}$ and the standard deviation σ_1 might actually be characterised by a distribution which has a mean $\mu_{\sigma 1}$ and a standard deviation $\sigma_{\sigma 1.}$ This is represented in Figure 2-8, which shows a distribution created from the distributions of each of the two values in the parameter set.

[9] This is illustrated in Fig 2.3 by the variations in the average demand derived from the CLNR data.

Figure 2-8 Example of generating parameters of a distribution from underlying distributions



This is a clearly an example which is well suited to the application of Bayesian Inference – this approach reflects that different types of uncertainty exist at different levels, and seeks to explicitly account for this. When using smart meter data sources (e.g. SMETS 2), this would be extended further so that the original understanding of these parameters (the 'priors') is periodically updated to produce new parameters, based on a regular review of smart meter data. Bayes Theorem is used to update these prior beliefs about probabilities based on new data.

The approach means we don't have to pretend that we know exactly what the demand distribution parameters are, when there is in fact considerable uncertainty, without compromising our ability to run accurate calculations.

2.4 Probabilistic network condition

In order to better understand the risk associated with different network states, it will be necessary to apply some sort of probabilistic approach to the modelling of the network condition, integrating the probabilistic model of demand. This is in contrast to the approaches generally taken now, which just look at a single snapshot of demand (and possibly generation), based on the 'design demand' or measured existing demand where it is available, and assess the network condition for that snapshot.

The aim of a probabilistic power flow is to give the network operator more information that they can use when planning their network. In particular, this enables a risk-based approach to network planning, as it gives the network operator information about both the severity of an unwanted network condition e.g. a network overload, as well as the likelihood of that condition occurring.

In the approach we have developed, this is achieved by calculating the exceedance expectation – i.e. the average number of times in a year that the utilisation of a circuit (or voltage) will exceed a specified threshold (such as the circuit's thermal rating or the defined voltage limits).

There are two main elements that contribute to a probabilistic power flow:

- A distribution network topology: the conditions, e.g. thermal and voltage of the distribution network depends in a deterministic way on the demands of customers connected to it. Therefore, the determination of power flows and voltage in the network is deterministic[10].
- Customer demands: a customer's demand is inherently random.

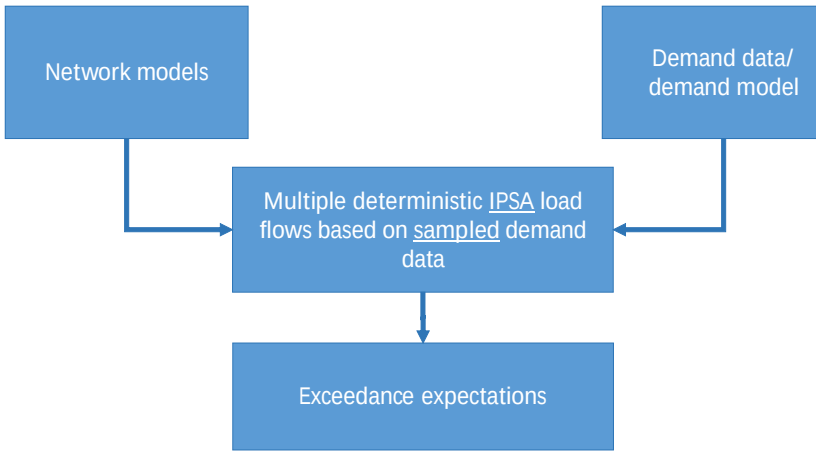
[10] In practice, simulations may exhibit some variation due to aspects of the implementation of load-flow algorithms e.g. whether or not a network is flat-started.

Combining probabilistic inputs with a deterministic process is well suited to Monte-Carlo analysis. The basic process is shown below, with an indication of how this process would apply to probabilistic power flow modelling:

1. Randomly generating inputs from a probability distribution
   This would involve sampling customer demands at each node from the relevant probability distribution.
2. Performing a deterministic computation on the inputs
   This would be the calculation of the AC power flow, for example using IPSA.
3. Aggregating the results
   This would be combining all of the results in order to calculate exceedance expectations.

One option for how this model could look is illustrated in Figure 2-9.

Figure 2-9 Option for Monte-Carlo load flow analysis



This would involve sampling sets of demand inputs from probability distributions, and then running each of these sets through an IPSA model.

However, it is likely that this could be very computationally inefficient, since there is a requirement of the Monte-Carlo analysis called convergence. That means that new samples need to be created and included in the calculation of desired output, up until there have been sufficient relevant outcomes to ensure that the 'wobble' in results as more samples are included falls within a suitable limit.

As a rule of thumb, it typically takes around 100 times the expected frequency of the event for results to converge. Therefore, in order to robustly identify a once per year event, 40,000 load flows might be required. For a 1-in-10-year event, 400,000 load flows might be required. These requirements are impractical for any regular network planning activity. For example, we expect that with the LV IPSA models we have created in this project it might take 1-2months to complete 1,000 load flows.

Therefore, a straightforward Monte-Carlo load flow approach is unlikely to be the right approach. As well as being time consuming to run, it is something that has been explored regularly in previous innovation projects on network design and may not be considered innovative enough for this project.

2.5 Implications for our novel analysis technique

There are four key conclusions that should be drawn from the information set out in this section.

1. NPg (and other DNOs) already use a range of tools for calculating the demand on their networks, and some of these tools are based on statistical models of electricity demand. Therefore, developing an updated statistical model for electricity demand is an evolution of the existing approach.

2. However, both the approaches taken currently to estimating demand (ACE49 and ADMD) have limitations, and there is potential to develop a model which improves on these in several areas. One key observation from the CLNR data is that, particularly for small groups of customers, the variation in demand patterns across groups of customers (and even across groups of similar customers) is very significant (as shown in fig 2.3).
3. The opportunity to apply Bayesian statistical inference techniques to the modelling of electricity demand is potentially very attractive. These techniques allow for explicit quantification of uncertainty within a statistical model, and also enable a range of different data sources to be integrated in a mathematically robust way. In addition, these techniques are well suited to situations where there is limited data available. By adopting a Bayesian approach, it would be possible to update the demand model in a consistent way as new sources of data become available including, but not limited to, smart meter data.
4. When assessing probabilities associated with undesirable conditions on the network, a full Monte-Carlo AC load flow is unlikely to be an appropriate approach, given the relatively high computational cost of completing AC load flows. For the purposes of LV design studies, which need to be completed relatively quickly due to the high volumes, an alternative approach will be required.

3 Novel analysis techniques methodology

3.1 Overview

The novel analysis techniques that we have developed and tested are presented at a high level in Figure 3-1. Each component is described in further detail below.

One of the key features of the method we have developed is that variability and uncertainty in customer demands are preserved and propagated throughout the entire model, to the greatest degree possible. That is, we want to know certain properties (such as the mean, or a certain percentile) of the functions of the random customer demands, rather than the functions of those properties (mean, percentiles) of customer demands.

As an illustrative example, imagine that a year consisted of only 3 time-steps (rather than 17,520), and that the set of demands, in kW, for a particular year was 3, 5, 4. Imagine also that a voltage, v, in which we are interested is related to the demand, d, according to $v = 0.4 \cdot d + 0.1 \cdot d^2$. The fundamental approach adopted by most existing methods is to calculate the mean demand as 4 kW, and state that the corresponding voltage is $(0.4 \cdot 4) + (0.1 \cdot 16) = 3.2V$. However, our approach propagates the variability by calculating that the 3 demand values correspond to voltages of 2.1V, 3.2V and 4.5V, so that the mean voltage is 3.27V. This can be expressed more formally as in the following example, where we adopt the common notation that random quantities are uppercase, while deterministic quantities are lowercase.

The ACE49 method involves determination of the 90th percentile value of demand, which is then set as the design demand $\hat{d}$:

$$\hat{d} = q_{90}(D)$$

where D is assumed to be normally distributed with mean μ and variance σ^2:

$$D \sim N(\mu, {\sigma_D}^2)$$

Then, network impacts such as utilisation caused by this design demand, $\hat{u}$, are evaluated as functions of $\widehat{Dd}$.

$$\hat{u} = g(\hat{d}) = g(q_{90}(D))$$

where $u = g(d)$ is the solution of the power flow problem for a given network with demand d.

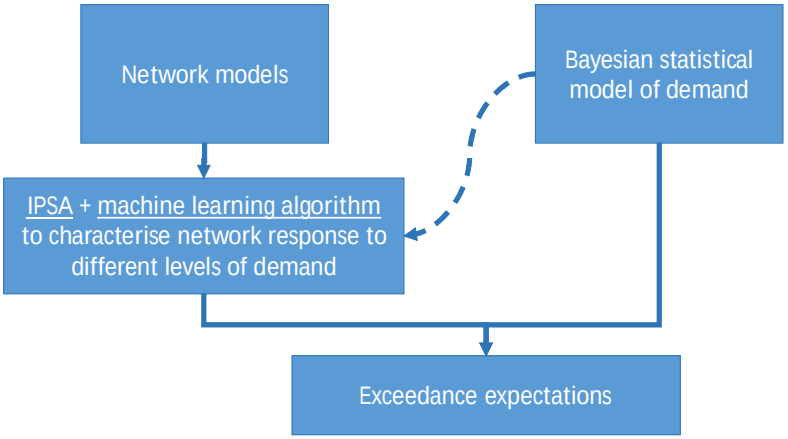
Our method enables a similar, but more accurate calculation to be completed, involving the direct calculation of the 90th percentile value of utilisation:

$$u_{90} = q_{90}(g(D))$$

The method enables translation of the distribution of demand, $D \sim N(G, {\sigma_D}^2)$, into a distribution of utilisations U, from which any quantity of interest (e.g. the mean, or a given quantile such as 90th percentile) can be determined. To achieve this, our method decouples network modelling from the statistical customer demand model, which has the additional benefit of reducing the time it takes to complete a study, compared to the typical probabilistic approach of Monte Carlo simulation. As previously stated, the Monte Carlo approach is not desirable due to the very large number of power flow simulations required, combined with the relatively large computational resource required to carry out AC power flow modelling.

A high-level overview of our approach is provided below, and each block will be explained in turn in the following report sections. It should be noted that the output is represented here as 'exceedance expectations', which means a function for any desired network variable e.g. current or voltage that shows the expected number of half hours per year that any threshold value for that variable is exceeded. However, it also possible to express the model in terms of the value of current or voltage that has some predetermined exceedance expectation (expressed as a number of hours per year).

Figure 3-1 Novel analysis technique components



In developing this technique, we have been mindful of the following high-level functional requirements:

- The analysis technique should be as automated as possible and should not require the LV designer to understand the underlying theory and methodology in order to make design decisions;
- The inputs to the analysis should be limited due to the high volume of LV networks that are assessed;
- The application of automated design policies should be considered;
- There should be an option for the LV designer to use some autonomy in defining the LV network demand based on their expert knowledge of the network e.g. specify the demand of commercial customers, if desired, without risking a loss of consistency within the approach; and
- The results should be presented in an easy to understand and apply manner.

3.2 Demand Model

3.2.1 Customer demand distributions

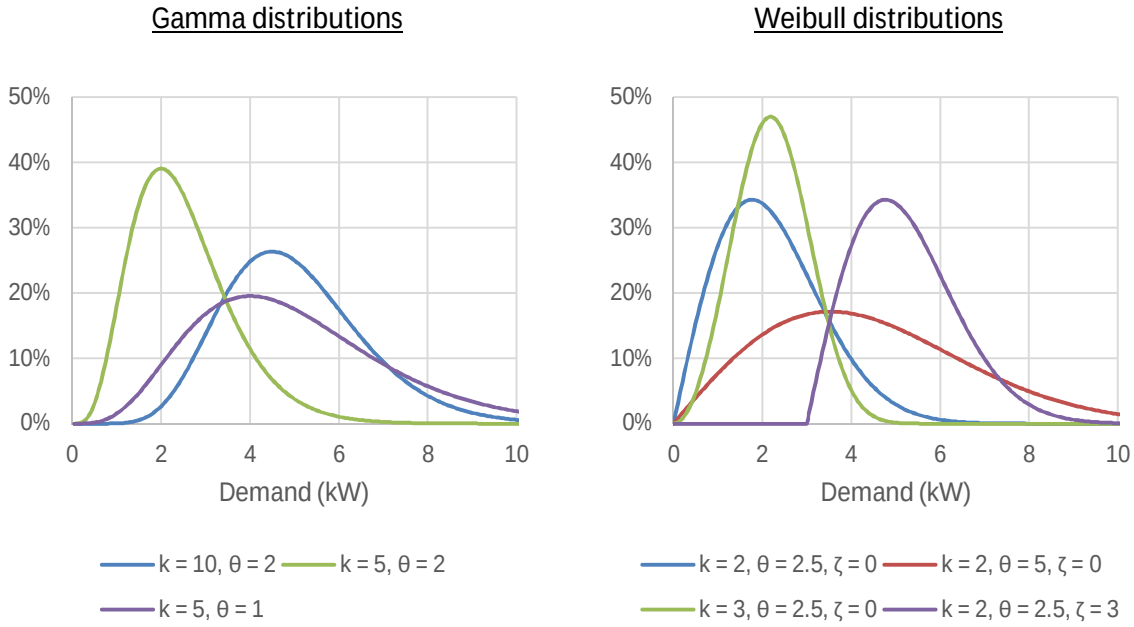
Customer demand is modelled as random variable with a probability distribution based on a set of representative demand data. For this study, we have used the TC1a dataset from the CLNR project, comprising half-hourly load data for 8000 customers over two and a half years. It is envisaged that in future, smart meter data would increasingly supplement the dataset, enabling a more accurate representation of the demand and demand uncertainty of customers both with and without a smart meter.

We have explored segmenting domestic customers into 7 socio-economic "MOSAIC" categories, along with commercial and industrial categories derived from the CLNR project. These represent a significantly reduced number of types compared to those adopted within the CLNR project, for both domestic customers and SMEs. It is clear that the adoption of such types is useful in that they reduce the uncertainty around both the scale and patterns of demand from customer groups – though typically not very significantly. Further, it is not clear that it is practical to expect LV design engineers to accurately identify the types of customers supplied from different parts of LV network, and as such the specification of a new, novel, methodology should strive to be agnostic to customer segmentation. It may well be the case that as the proposed novel methodology evolves in the future, possibly along with an increasing ability to automatically analyse the type of customers supplied from an LV network external sources, categorisation could bring additional benefits. Some new system of categorisation may be introduced to account for new sources of data, rather than socio-economic categorisation (like MOSAIC categories). That might account for different heating sources, the type of the house, new types of customer, and/or new low carbon technology, but it is very difficult to currently predict exactly how that system of categorisation might work.

As will be discussed in greater depth below, one of the most significant differences in our approach compared to ACE49 is our adoption of Gamma and Weibull distributions, rather than normal distributions, as the former are much more flexible than the latter. Indeed, Gamma and Weibull distributions can accommodate positive-valued variables with a very wide variety of distribution shapes. In more formal statistical terms, we can say that while the normal distribution allows specification of mean and variance only, the Gamma and Weibull distributions allow these to be specified along with skewness - which defines how lopsided a distribution is, and the kurtosis - which relates to the likelihood of moderately rare events[11], compared to central and extreme events. This is the model which has been assessed in the smart meter data analytics workstream.

Examples of these distributions for different parameter sets are shown in Figure 3-2.

Figure 3-2 Example Gamma and Weibull distributions



For each distribution, the Gamma and 3-parameter Weibull distributions have two or three parameters each to define their shape. The detail of this is described in Appendix A. By fitting these distributions to the observed customer demand, we end up with an exact equation which describes the probability of the aggregated customer demand at a specific time being above or below some threshold, or within some interval.

3.2.2 Time of day and seasonal distributions

For a single customer, there could, in principle, be as many as 1,000[12] parameters to completely define their demand. This is largely due to distributions being genuinely different for each season, and each of the 48 half-hour slots in the day, combined with the additional hyper-parameters the Bayesian inference approach introduces for each distribution. This is clearly overly complex and therefore, our approach

[11] These four characteristics – expectation, variance, skewness and kurtosis are technically referred to as *moments*.

[12] This assumes that there are 48 distributions for each of the four seasons, and that each distribution has 2 or 3 parameters, each of which is defined by 2 hyper-parameters. This gives a range of 768 to 1152 parameters, depending on whether Gamma or Weibull distributions are used.

introduces some well-justified simplifications. Specifically, our model considers sequences of times-of-day[13] within which the differences in distribution means are relatively small, and where either:

- the differences in distribution means and standard deviations are roughly proportional, or;
- differences in standard deviation can effectively be neglected.

These consecutive periods can be temporarily represented as having the same distribution parameters, through the use of either constant multiplicative or additive factors, so that the parameters are fitted to a sufficiently large set of observations. The same factors can then be used to restore the uniqueness of the individual distributions. The details of this process are presented in Appendix A.

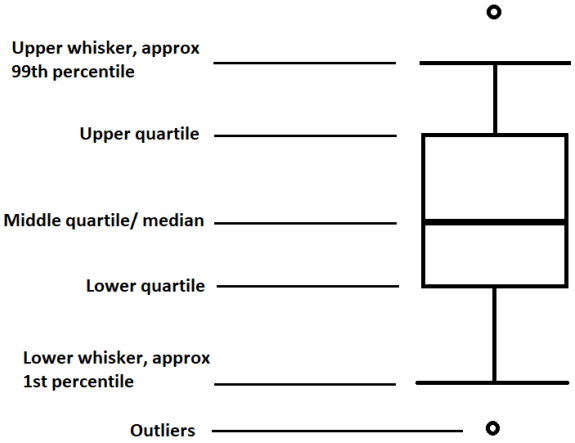
3.2.3 Customer categorisation

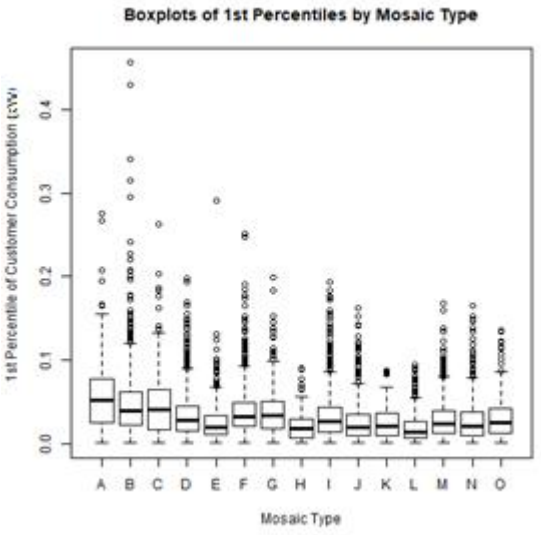
As stated above, customers can be categorised into various types to improve the demand model e.g. based on socio-economic data. Each customer type's Gamma or Weibull distribution will have a k and θ, and a ζ for Weibull. In principle, the more detailed the customer-type model, the narrower the parameter distributions become and our knowledge of the network demand and the associated state is improved. However, increases in the number of customer types make the model of how the parameters depend on the number of customers of each type more complex, and therefore likely to be represented more imperfectly. There will therefore exist an optimum number of customer types where these considerations are balanced.

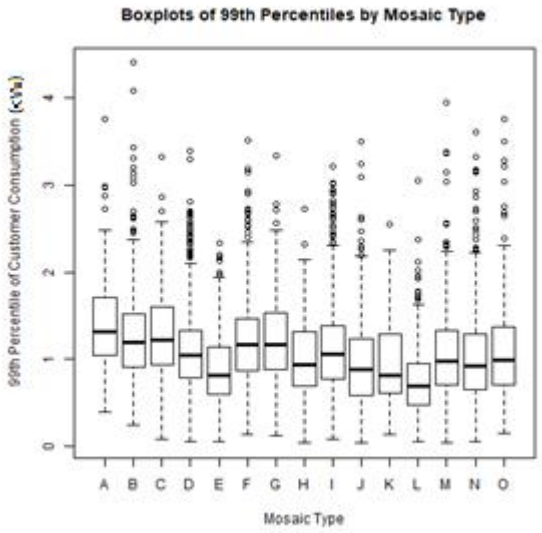
For example, a demand distribution can be defined for a set of customers categorised as "Elderly Needs" (from the Experian MOSAIC system, used to categorise customers in the TC1a dataset). To explore this issue, we temporarily suspend the treatment of demand as a set of distributions for each unique combination of time-of-day and season, and consider a single distribution over all such combinations. The demand distribution of all Elderly Needs customers can be represented by an "average" category distribution defined by k_{EN} and θ_{EN}. An individual Elderly Needs customer will, however, also have their own unique demand distribution defined by k_i and θ_i. Demand modelling approaches typically assume that as long as there is a sufficient number of customer types (the 15 MOSAIC types, for example, being sufficient), the deviations of the individual k_i and θ_i from k_{EN} and θ_{EN} are essentially negligible, and all customers of that type can be modelled as identically distributions. However, our examination of the CLNR data shows that there is in fact very large variability between the distributions of demand observations, within the MOSAIC types. This is illustrated in Figure 3-3, which presents box plots for the 1st, 99th and 99.99th percentile of demand for each customer within a category.

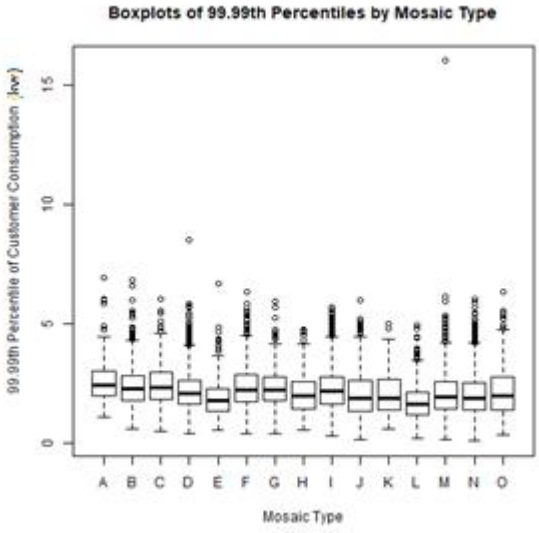
Figure 3-3 Boxplots of 1st, 99th, and 99th percentile values of demand across MOSAIC categories type

[13] For example, 15:30 to 22:00 in winter is one such sequence.









Indeed, the figures show that the variability within the customer groups tend to be greater that the variation across the groups, although this variability is somewhat reduced for the most extreme values (i.e. the 99.99th percentiles). This suggests that the use of domestic customer categorisation using a system such as MOSAIC is not strictly necessary and potentially not particularly worthwhile. However, preliminary analysis on the CLNR project's TC1b dataset on SME customers, not reported here, demonstrated that the identification and characterisation of non-domestic customers is important, and that the system adopted by the CLNR project is not sufficient.

3.2.4 Bayesian Updating

The application of Bayesian inference to network demand modelling with different sources of data is illustrated in Figure 3-4, with five illustrative steps identified. This diagram differentiates between the data used, the methods applied, and the models produced at different points in time and for different purposes. It also differentiates between demand models which are determined "generically" and applied to all parts of the network across the entire licence area, and those which are determined for a specific part of the network. The mathematical foundations of this approach are presented in Appendix A.

1. The first step is to establish an initial set of parameters (or "prior" belief) for the demand model. This could be based on the data within the CLNR project, where the prior belief might be based on extracting a large number of different customer profiles at random and fitting statistical models to each of them[14]. Such a generic CLNR based demand model could be refined using new smart meter data to produce a refined generic model. Other datasets may also be used, particularly for Low Carbon Technologies (LCTs), where the CLNR project's datasets are small. In the smart meter data analytics workstream[15] it is illustrated how creating demand models that include LCTs might be achieved by continuing to fit Gamma and Weibull distributions to the observations of net demand for different groups of customers. It is also possible that other approaches could be useful, particularly for LCTs where there is much less data available. For example, a "generic" statistical model for the "marginal distribution"[16] of LCT demands and PV generation output could be constructed. This step would only need to be completed once, and could actually be developed without *any* smart meter datasets. However, a challenge for LV modelling in general is capturing the temporal statistical relationship between LCT demand, PV generation and traditional domestic demand. These challenges are described in more detail in Section 5.2.1.

2. For a specific area of the network where smart meter data is available, this refined generic model would be combined with that specific smart meter data (subject to rules around aggregation), and any other network specific data, to produce a refined network-specific demand model. The creation of a bespoke network demand model would be done each time that specific network was to be studied, and hence this process would need to be automated. This process would treat the new smart meter data as a sample in the Bayesian inference process, and use it to update the "priors" from the previous generic or refined generic demand model, producing a network bespoke demand. A detailed presentation on such Bayesian inferencing is available in Appendix A.A.1.1.

3. The Bayesian updating process would not be limited to direct observations of customer demands. Other informative data sources that could be used[17] include network monitoring, annual or seasonal energy consumption values, or the model of LV network total demand developed for NPg by Element Energy. The latter is a model that includes very finely-grained spatial information about the socio-economic characteristics of the customers at a given location, along with the presence of specific buildings such as hospitals and schools. This spatial data is combined with Elexon settlement data on energy consumption at higher levels of aggregation, with the latter acting as a set of constraints on individual modelled demand at LV network level.

4. In the case of such indirect demand observations, additional statistical models will have to be developed to represent the relationship between the indirect observations and demand distribution parameters. These will allow the indirect observations to be translated into likelihood functions for the demand distribution parameters. Exactly the same principle applies where only some of the customers on a feeder have smart meters (which will almost always be the case). That is, we have two quantities of interest: (i) a random variable of principal interest that we wish to update, which is either the total demand (net generation) on a feeder or main, or downstream of some node within the circuit; (ii) a 2nd variable that is statistically related to the 1st, and for which we have direct observations. The 2nd variable could be values recorded by monitoring devices at the beginning of a feeder, or aggregated smart meter data from a subset of customers on a feeder.

[14] For a full implementation of the model, this might require 1,000s or even 10,000s of different samples.

[15] Smart Meter Data Analytics Report, March 2019

[16] The term marginal distribution is used to describe the probability distribution of a single variable (e.g. PV output) without reference to the values of other variables (e.g. domestic demand), in a situation where the simultaneous values of multiple correlated variables is also of interest (e.g. in order to understand total demand net generation).

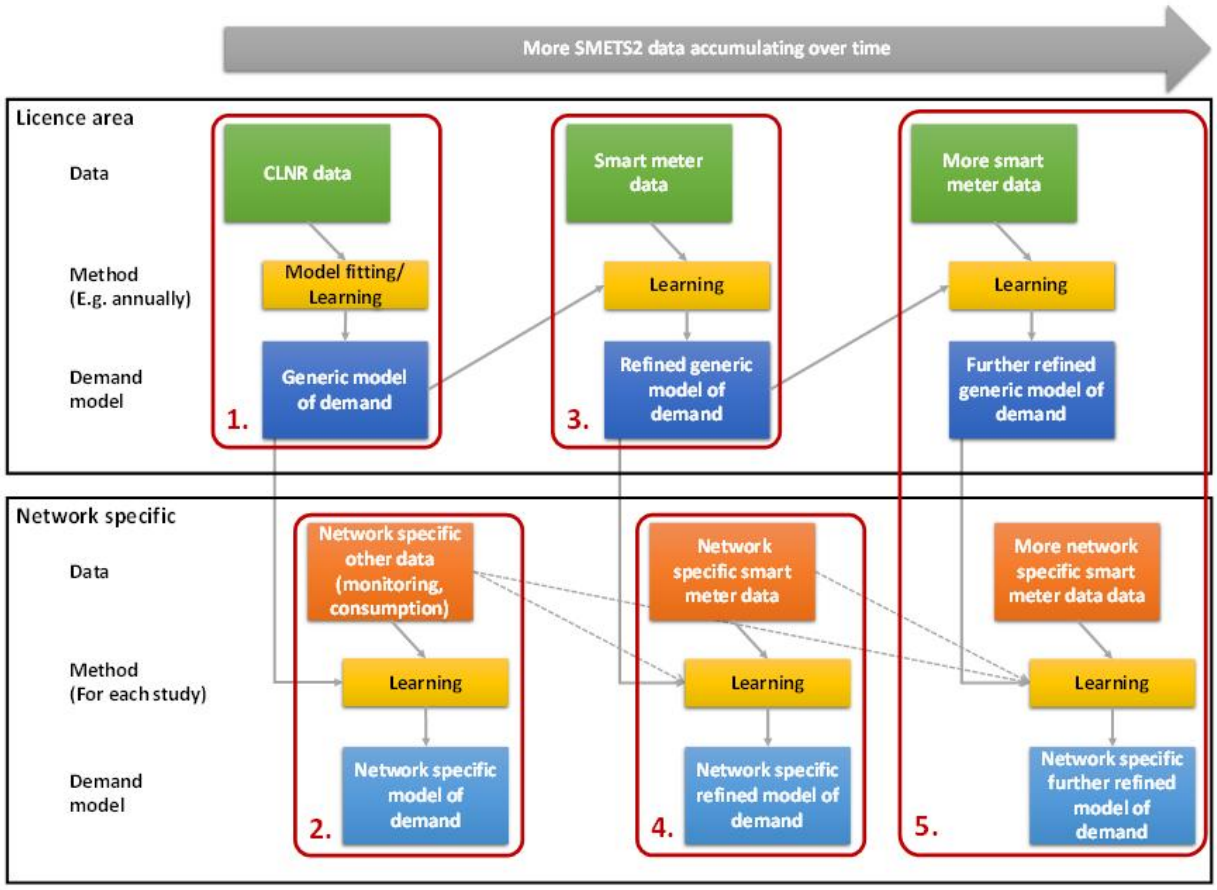
[17] We have not explored the detail of how these data sources would be integrated into the method.

Regardless of the nature of the observations, the essence of the challenge is to mathematically model the relationship between the observed variable and the variable of most interest, as explained in the presentation on Bayesian updating using indirect measurement is available in Appendix A.A.1.2. It is likely that artificial intelligence, or other machine learning method, would conduct automatic characterisation and prediction of these relationships, as discussed in Section 5.

5. The case of Element Energy forecasts is unique among additional data sources, in that they are point forecasts, rather than observations. Indeed, they can only become a useful source of data for Bayesian updating of the model presented here once compared to observations, so that a statistical model of their errors can be built. It is unclear whether annual or seasonal energy consumption values would ultimately be used directly to update the Bayesian models of demand, or used to update the error distribution associated with Element Energy forecasts, which in turn would update the demand model, or – most likely – both.
6. While even a small amount of monitoring at the LV level will be useful immediately as a form of model validation, it will take several years of monitoring at many locations before sufficiently accurate models can be built for this type of data to significantly update the Bayesian demand models.
7. We envisage that the generic demand models would be updated (using these learning approaches) on a periodic basis, e.g. annually, incorporating new smart meter data sets along with all other data sources . Over time, it is likely that the number of updating data sources would increase, such as the introduction of more granular categorisation of customers. As before, whenever a specific part of the network was to be assessed, the generic demand model would be updated using information from that specific part of the network to produce a bespoke refined demand model for that particular network.

We have explored the first step of this process within the case studies which are described in Section 4 although we have used only 100 randomly selected net demand series for the purposes of the illustration, rather than 1,000s, which would be required to implement this methodology in practice.

Figure 3-4 Process for refining demand models as more data becomes available

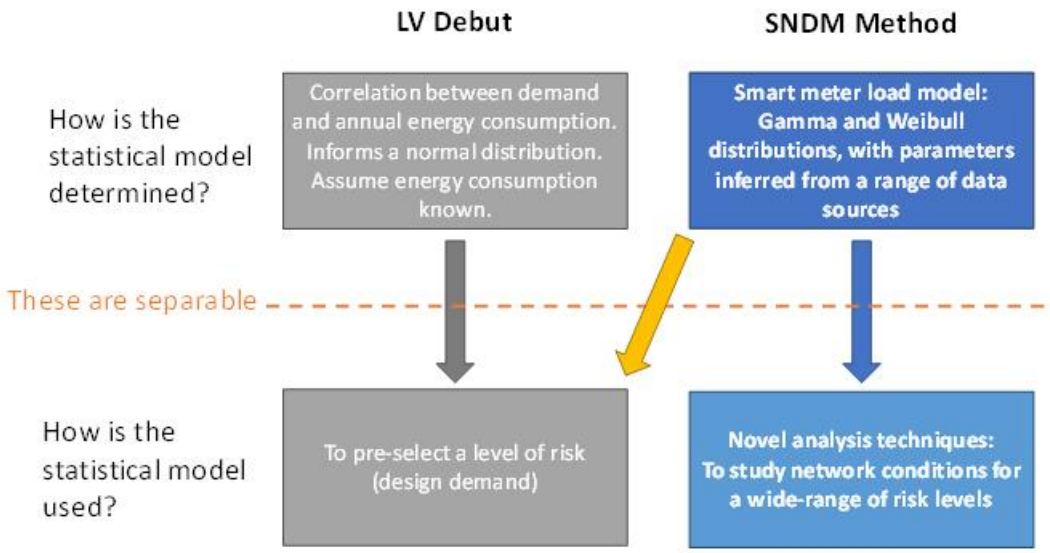


3.2.5 Summary and Conclusion

The proposed approach for creating and refining the demand model is presented above. Examples, explored through a case study, are presented in Chapter 4. The approach is relatively flexible as to how the estimated demand patterns are used to explore impacts on the network. It can be used to carry out a simple assessment of network impacts e.g. by computing a design demand and then comparing this to thermal limits of the network components. However, it can also be applied alongside more sophisticated statistical techniques that analyse network impact across a range of risk levels. Further detail on this is given in Section 3.3.

Figure 3-5 illustrates that the determination of the statistical demand model is largely decoupled from its application.

Figure 3-5 Components of SNDM method



3.3 Network Response Model

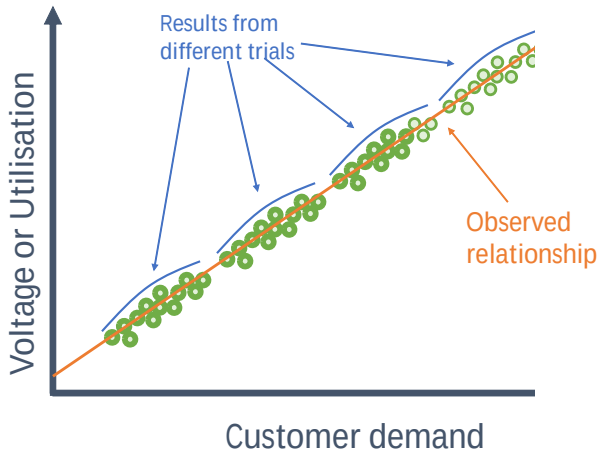
We can characterise how a network model responds to varying levels of customer demand using a power systems model (e.g. IPSA) and machine learning techniques. The power systems model calculates how the network will respond to different combinations of individual customer demands. This is analogous to modelling how individual customer demand and thus network conditions might vary across the course of any random day. Parametric equations can then be derived that describe:

- Thermal loading of any feeder section or transformer, based on downstream feeder demand;
- Voltage behaviour of any node based on the customer demand on the feeder.

Then, for any defined set of individual customer demands, it is possible to estimate the resulting power flow and node voltages from the parametric equations. This can be expressed as how the utilisation of a given branch or the voltage at a specific node varies based on all of the demands in the network e.g. $F_i(d_1, d_2, \dots, d_N)$ for branch/node i. Ideally, the network variable would depend on only one aggregate demand. This is likely to be the case for thermal utilisation in un-tapered radial networks. However, for more complex networks, such as tapered networks, networks with fewer customers dispersed along long feeders, or networks with significant unbalance across phases, it may be necessary to consider multiple aggregate demands, which will then require the use of multi-variate statistics (described in Section 5.2.2).

Characterisation of network response for each feeder section and node in our approach involves running 100s or 1000s of "trials" through IPSA based on plausible levels of concurrent customer demands. This is illustrated in Figure 3-6.

Figure 3-6 Illustration of trials and fitting a relationship for voltage or utilisation



Note that the values of demand which are used to carry out these trials do not need to be sampled from any particular distribution, i.e. their relative frequency of occurrence does not attempt to reflect reality, and any source providing *credible* input demand values could be used. In testing this technique, we have used the CLNR time series which are readily available. As it is important to understand the behaviour of the network under low probability, high impact peak loading, customer demands can be scaled up to reproduce this.

Based on our initial assessment, 1,000 IPSA runs takes in the region of 5-10 minutes for a representative LV network. However, in principle, these trials could be run at any time e.g. in the final step of the LV model creation, and not necessarily when the LV designer is using the model. Therefore, when the LV designer uses the network response model (the fitted relationship between demand and network utilisation or voltage) as part of the novel analysis technique, applying this should be much faster.

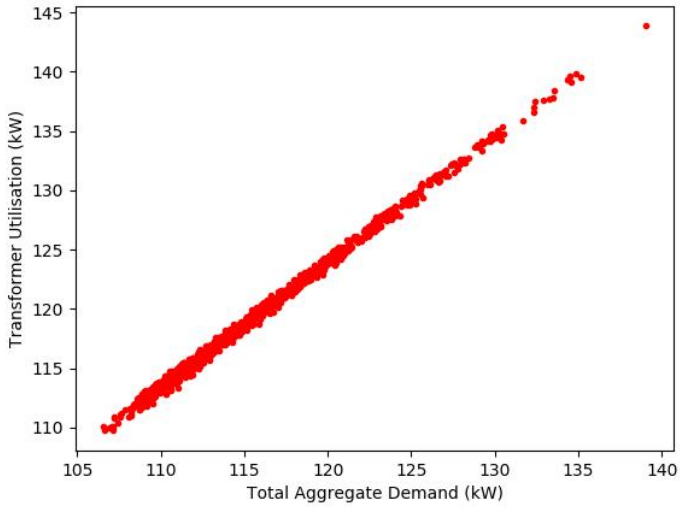
It is worth noting that different trials would need to be run to establish a revised network response model when assessing any changes to the network configuration (such as changing normally open points), new solutions or reinforcements, contingencies, or changes in voltage control set points. For example, Figure 3-7 shows how this might be approached for a range of running arrangements: each running arrangements leads to a different set of trial results, and (in this illustrative example) a different linear relationship. The overall utilisation would then be the most onerous result for each possible value of demand (the orange line).

Figure 3-7 Illustration of trials and fitting a relationship for utilisation for different contingencies

3.3.1 Network response model example

An example of a network response model is shown in Figure 3-8 which compares the relationship between the total aggregate demand on an LV network and the utilisation of the secondary transformer. In this case, it is easy to determine a simple linear relationship to explain how the transformer utilisation will vary for any combination of down-stream demands. With such a relationship, the majority of the variation in the transformer utilisation can be explained based on knowing only the behaviour of the aggregate downstream demand.

Figure 3-8 Example of response of transformer utilisation to downstream demand



The presence of a broadly linear relationship is no surprise – it is a well-known result that when per-unit voltages are close to unity the power flow equations can be linearised (this is often known as DC load flow). Our expectation is that a small amount of inaccuracy in the results (due to not re-running the power flow simulations) will be acceptable if it significantly boosts the time taken to run the models, and in principle this error, and the increase in uncertainty that it would lead to, could even be included in the statistical model directly.

It may not always be possible to find a relationship which explains close to 100% of the variation based on only a single variable. For example, a large non-domestic load on the end of a feeder with an unusual consumption pattern could affect voltages along the length of the feeder. However, our expectation based

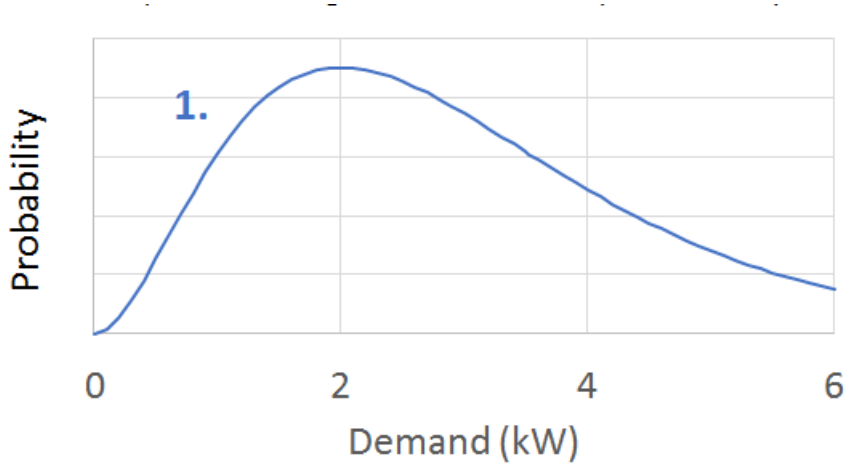
on tests carried out on the Sinderby network is that it should be possible to explain the majority of the variation in most cases based on aggregating downstream demand into one or two groups. This helps simplify the statistical modelling, as multivariate statistics can be very complicated, particularly where there are more than 3 or 4 variables.

However, it may also be beneficial in terms of aggregation when assessing the customer demands for specific customers on a particular part of the network. In such cases, there are concerns about customer privacy, if DNOs (or others) are allowed access to completely disaggregated customer smart meter data. This is discussed in more detail in Section 5.1.2

3.4 Stylised example of determining probabilistic network condition

In this section, we present a stylised example of how the novel analysis techniques can be applied to the design of an LV network. Note that this uses purely illustrative values. We consider first the case where demands are modelled in a frequentist manner, i.e. their probability distributions have fixed and (approximately) known parameter values[18]. As a result, the term 'predictive distribution' has no specific meaning other than probability density functions (PDFs), as illustrated in Figure 3-9[19].

Figure 3-9 Step 1: Demand probability density function



Step 1: Produce PDF's of the aggregate customer demand for each hour-of-day and season, using the methods described in Section 3.2.

Step 2:Use this set of PDFs to generate a set of "exceedance functions", i.e. the probability that demands exceeds the input value, as shown in Figure 3-10. This is one minus the cumulative probability of any given level of demand.

Exceedance probabilities can be converted to exceedance expectations by multiplying by the number of periods with the same time-of-day, for each season. This is done for each combination of time-of-day and season, and then summed. For example, for a customer whose demand is always positive, their exceedance expectation will be 17,520 half-hours for 0 kW (24h/day x 365 days x 2).

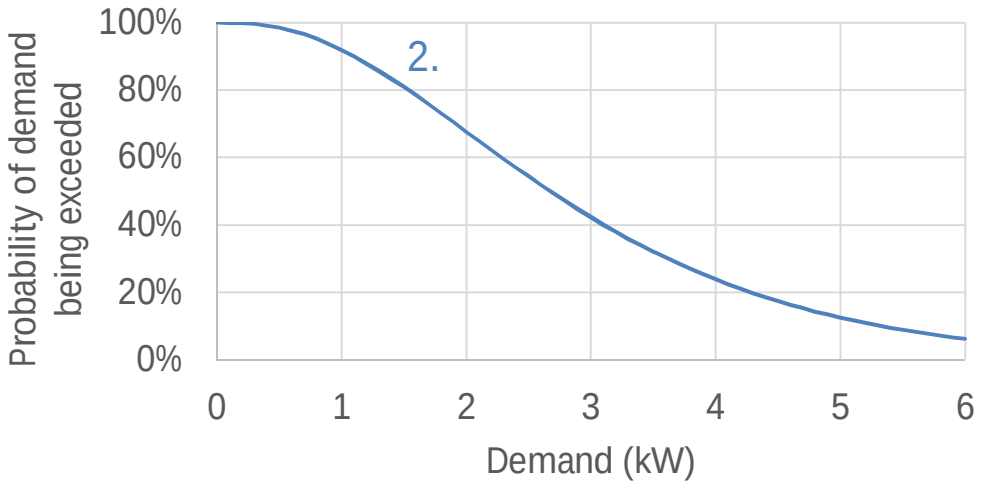
[18] This could alternatively correspond to a Bayesian posterior predictive distribution, where the uncertainty has been collapsed out and reflected within the distribution.

[19] We have deliberately omitted the scale from this graph, as it is difficult to meaningfully define this for a continuous PDF.

A 1-in-10-year event therefore has an exceedance expectation of 1/10, or a probability of 1/175,200 (which is equivalent to a 0.001% probability), assuming this is as defined as the half-hourly average demand level which occurs once in a ten-year period.

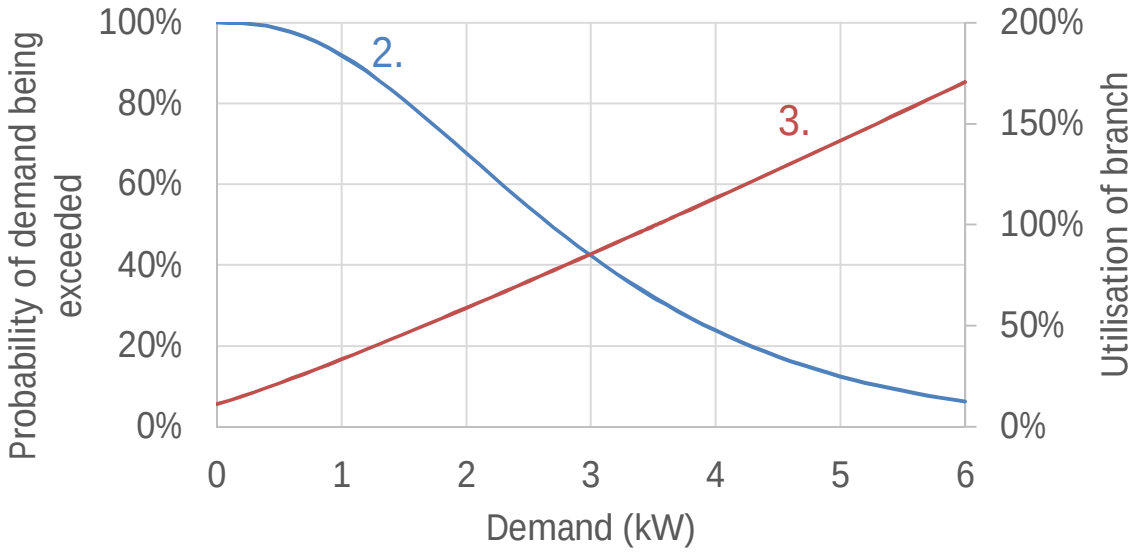
For simplicity when discussing these concepts diagrammatically, we will continue to work with exceedance probability curves (in units of % rather than half-hours) and look at larger probabilities (e.g. 20%) than a network designer would be interested in practice (which would be closer to 0.001%).

Figure 3-10 Step 2: Probability of demand being exceeded



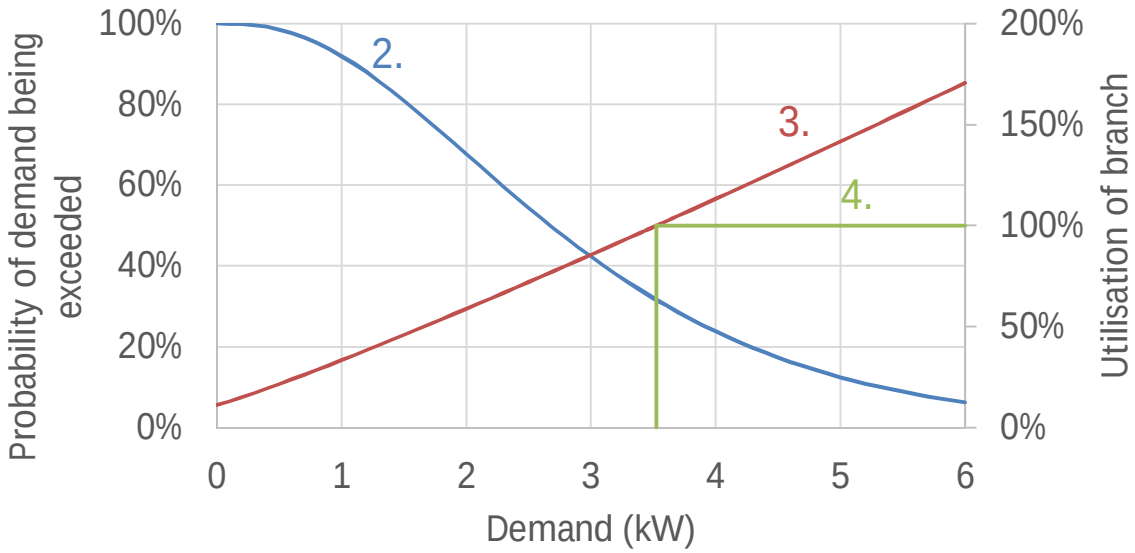
Step 3: Based on the network response trials completed in IPSA, an equation can be found which describes the utilisation (or voltage) of a feeder section as a function of the relevant aggregated demand. In this illustrative case, this is represented reasonably well by a linear function, which depends on only one variable (e.g. aggregate downstream demand) as shown in Figure 3-11. Multi-variate applications are discussed at a high level in Section 5.2.2.

Figure 3-11 Step 3: Utilisation of network varying with demand



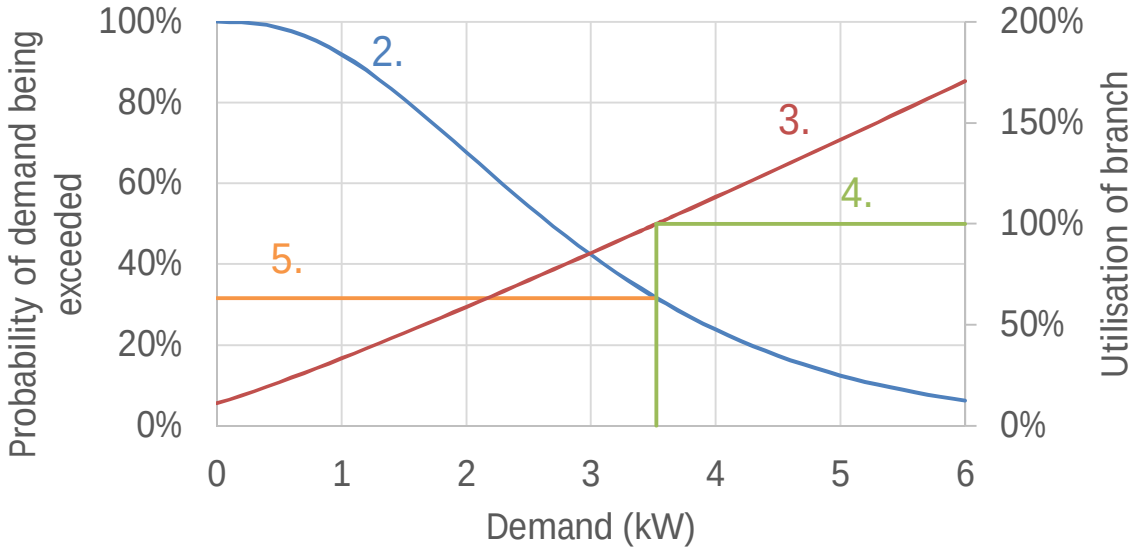
Step 4: The network planner can determine what level of aggregate customer demand leads to a utilisation of 100% from Step 3. In this case, a demand of approximately 3.5 kW leads to a utilisation of 100%.

Figure 3-12 Step 4: Demand associated with a given utilisation



Step 5: By comparing this to the demand model exceedance function, the probability that demand will be high enough such that the feeder section has a utilisation of 100% or higher can be estimated. In this stylised example, there is a 32% chance of the feeder section utilisation being exceeded. This would be mean that for a given year, there are 5,606 half-hour periods (0.32 x 24h/day x 365 days x 2) during which half-hourly aggregate customer demand is likely to cause the feeder section utilisation to exceed 100%.

Figure 3-13 Step 5: Probability of demand level



3.4.1.1 Extension to Bayesian inference

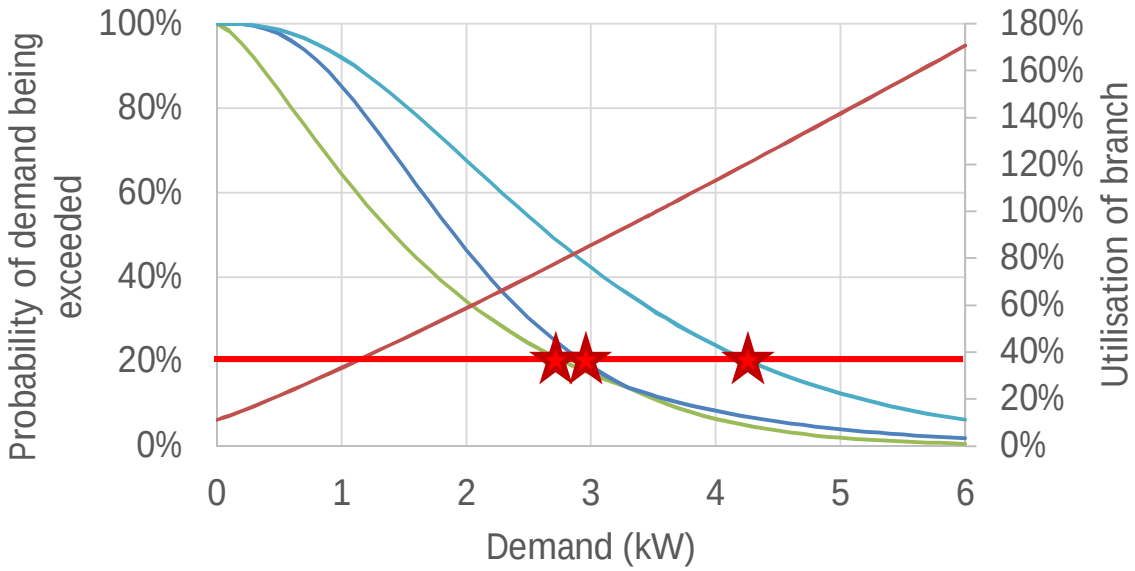
The stylized example above involved a demand model with a single probability distribution with known and fixed parameter values. However, in our “Bayesian” inference model, these parameters are also uncertain and have a distribution, and in this sub-section we present how the process above can be adapted to

account for this. The completely correct adaptation is to replace the generic demand PDFs with Bayesian posterior predictive distributions, calculated as shown in appendix A.1[20].

However, it is not always straightforward or even possible to calculate these quantities precisely, and an approximate, sampling-based approach can be adopted instead[21]. The essence of such an approach is to randomly sample sets of PDF parameter values from their distributions of possible values, in order to produce multiple PDFs.

Figure 3-14 shows three demand models characterised by PDFs with different sampled parameter values.

Figure 3-14 Example of multiple different distributions



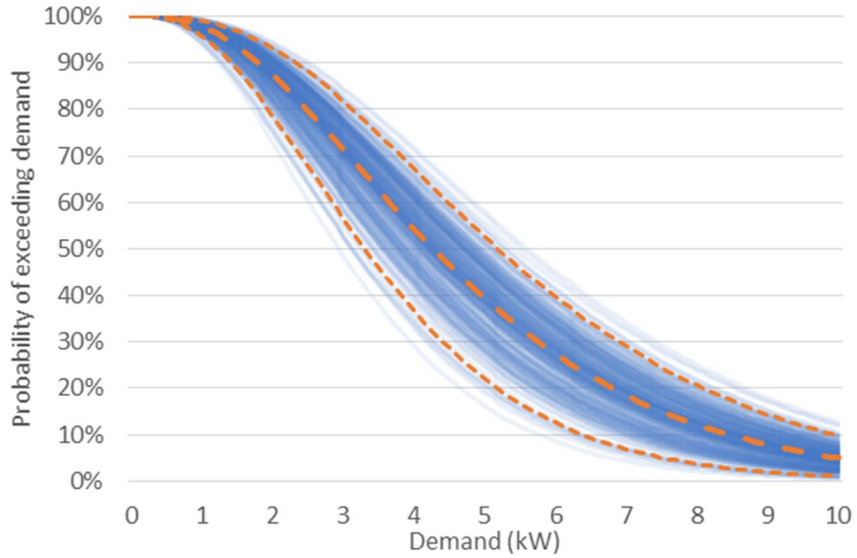
Each PDF has a different level of demand for which the probability of exceedance is 20% - this is because the demand for the specific group of customers represented by a particular PDF is fundamentally variable and only partially predictable. However, if a designer does not know which of these three groups is actually connected to their particular network, then there is also *uncertainty* about the distribution of demands. This means, that for a given utilisation, there will be a range of possible exceedance expectations – one for each sampled set of parameters. For example, with three samples, we get three possible demand values associated with any probability of demand being exceeded.

We can visualize the impact of 100s of repeated samples from the possible values of the distribution's parameters, as in Figure 3-15.

[20] The prior and posterior distributions should ideally be enhanced by the calculation of accompanying 'credible intervals', that reflect the range of values the distribution parameters might credibly take, and are therefore roughly the Bayesian equivalent of confidence intervals.

[21] We have adopted the sampling-based approach in our case study, but a tool that implements this technique could use either approach.

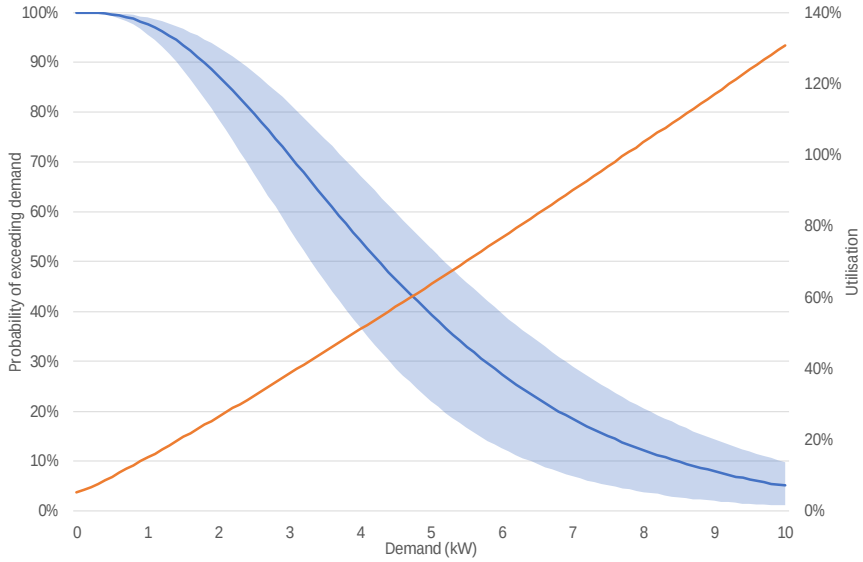
Figure 3-15 Example of hundreds of different distributions



In this case, there are 100s of possible distributions all overlaid, all based on subtly different sets of parameters. The orange lines show the average probability associated with each level of demand, as well as the range of values which include 95% of the possible probabilities. One option would be to collapse down this range at this point and only consider the average value – this would be a completely valid approach, but it does mean that some of the useful information about the uncertainty in this estimate would be lost.

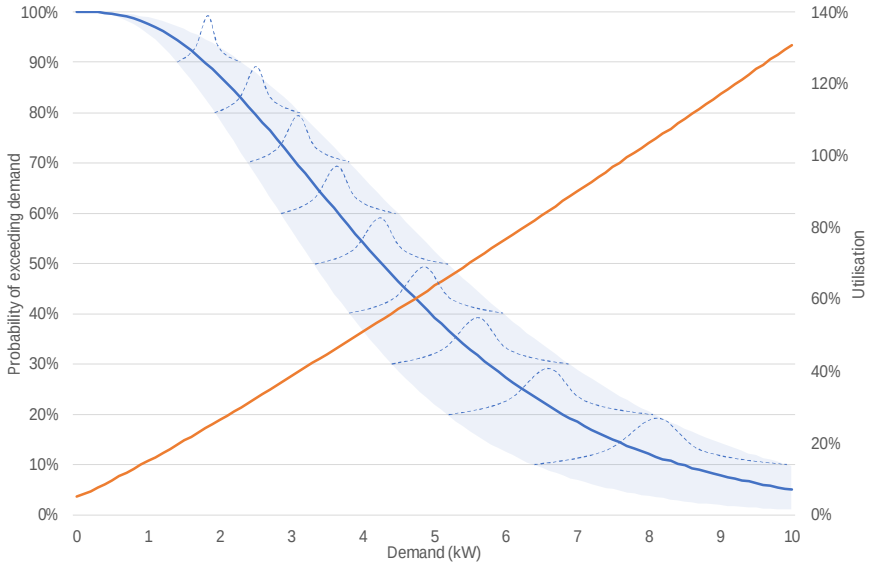
This is simplified slightly in Figure 3-16 – for any given exceedance probability, there is a 95% chance that the demand is within the blue shaded range. For example, the demand that is exceeded 20% of the time is 95% sure to be within the range of 5.5 kW and 8 kW.

Figure 3-16 Expected and 95% possible range of distributions



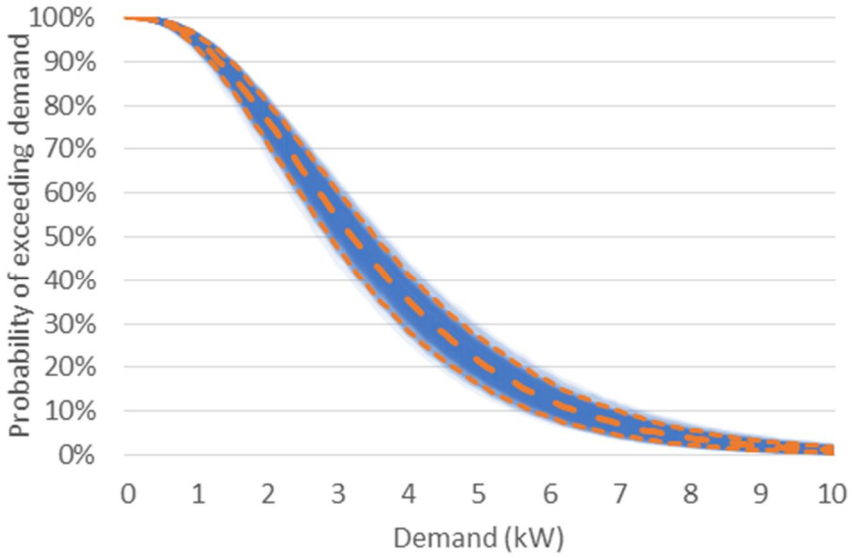
Essentially, for any level of risk (probability of exceeding demand) there is actually a distribution of possible values of demand. This distribution is defined by the uncertainty in the parameter set which defined the original demand distribution. This is illustrated in Figure 3-17.

Figure 3-17 Expected and 95% possible range of distributions



Note that, over time, as more new data is incorporated within the model and the classification of customers improves, the uncertainty around the parameters decreases and the range of possible values "tightens" around the average[22] as illustrated in Figure 3-18.

Figure 3-18 95% possible range 'tightening' as new data is incorporated



3.5 Usage and outputs of the method

In this section, we build on the stylised example of Section 3.4 in order to describe the different ways in which an LV designer could use the methods, the sort of outputs they might produce and the role of different options within the network planning process.

Full model: The "full" model combines the Bayesian representation of demand with knowledge of the network response (from the decoupled IPSA model) to produce a full probabilistic model of the utilisation

[22] This 'tightening' is not always 100% guaranteed – it is possible, in principle, for consumption patterns to change sufficiently quickly (e.g. due to a sudden but uncertain adoption of electric vehicles) that the process of learning about these changes couldn't keep up, and uncertainties would temporarily increase.

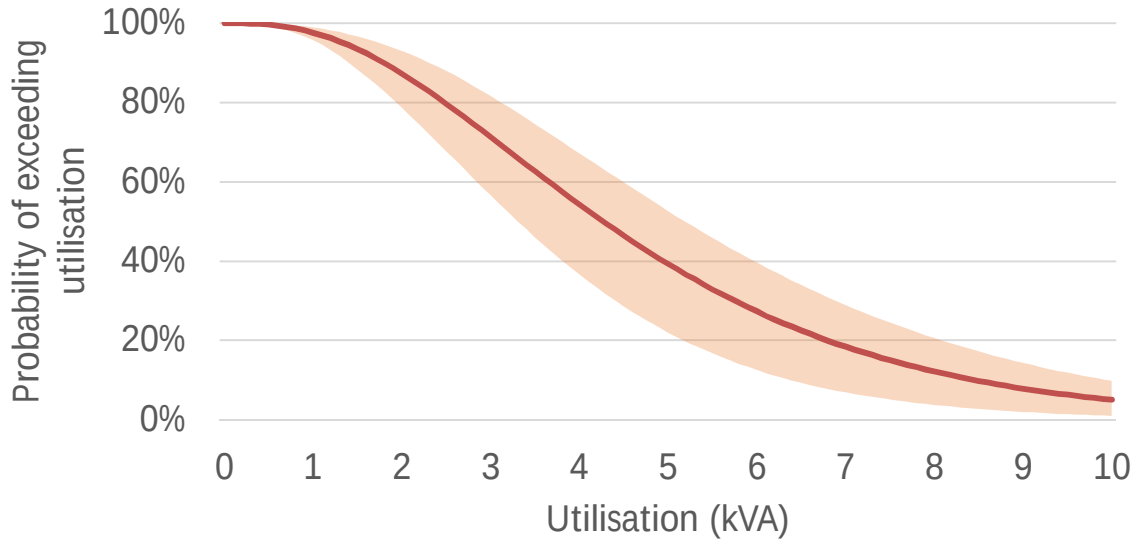
and voltage of each branch node. This is a full risk-based assessment involving the calculation of ranges for exceedance probability functions where the function arguments are branch utilisations and busbar voltages, rather than demands.

To reiterate and expand commentary in the previous section, the Bayesian approach, if followed correctly, provides a single PDF that fully accounts for the fundamental variability in customer demand, and the uncertainty that exists about which customers are connected to the network, i.e. the posterior predictive distribution. This is the thick orange line shown in Figure 3-19.

However, due to the mathematical complexity of calculating this distribution exactly, it can be approximated as the average of many curves resulting from randomly sampled parameter sets. We also believe that it is illustrative to include the 'envelope' that contains most of the sampled curves, as a measure of the model uncertainty, although this would not need to be provided as an output of a modelling tool that implements this methodology. This is illustrated as the transparent orange area in Figure 3-19.

An example of the output of this "full" model for a single branch is shown in Figure 3-19.

Figure 3-19 Output of the "full" model



In this example, our approximated predictive distribution states that a utilisation of 7.7 kVA is expected to be exceeded 15% of the time. We can supplement this prediction by stating that in 95% of our sampled curves, the demand exceeds 7.7 kW between 5% and 25% of the time. In addition, we have a complete understanding of similar outputs for *all* other levels of utilisation.

Pre-defined risk or utilisation level: The first optional simplification of this model is to consider a "fixed" level of risk (or equally simply, a fixed level of utilisation). The assessment would then be based on a set risk level or set utilisation/voltage, the corresponding demand according to the estimated predictive distribution, and the range of demands obtained from the sampled exceedance probability functions. For example, the designer may want to know what is the modelled utilisation, and accompanying uncertainty, corresponding to the demand value expected to occur only once in every 10 years.

For example, our predicted circuit utilisation is 85% for a demand level we expect to be exceeded 20% of the time. We also note that for 95% of the sampled parameter sets, the utilisation level defined in this way lies between 65% and 105%. Remember that this is only calculated for one level of demand i.e. the one which is expected to be exceeded 20% of the time (like ACE49). This is illustrated in Figure 3-20.

Conversely, we could examine the risk associated with a circuit utilisation of 100%, and find that the model predicts this would be exceeded 15% of the time, on average. However, we also note that for 95% of

sampled parameter sets, this utilisation's exceedance probability ranged between 5% and 25%. This is illustrated in Figure 3-21.

The only difference between this and the "full" model is that these calculations are only completed for a pre-set level of either risk or utilisation.

Figure 3-20 Calculating utilisation for a specified level of risk

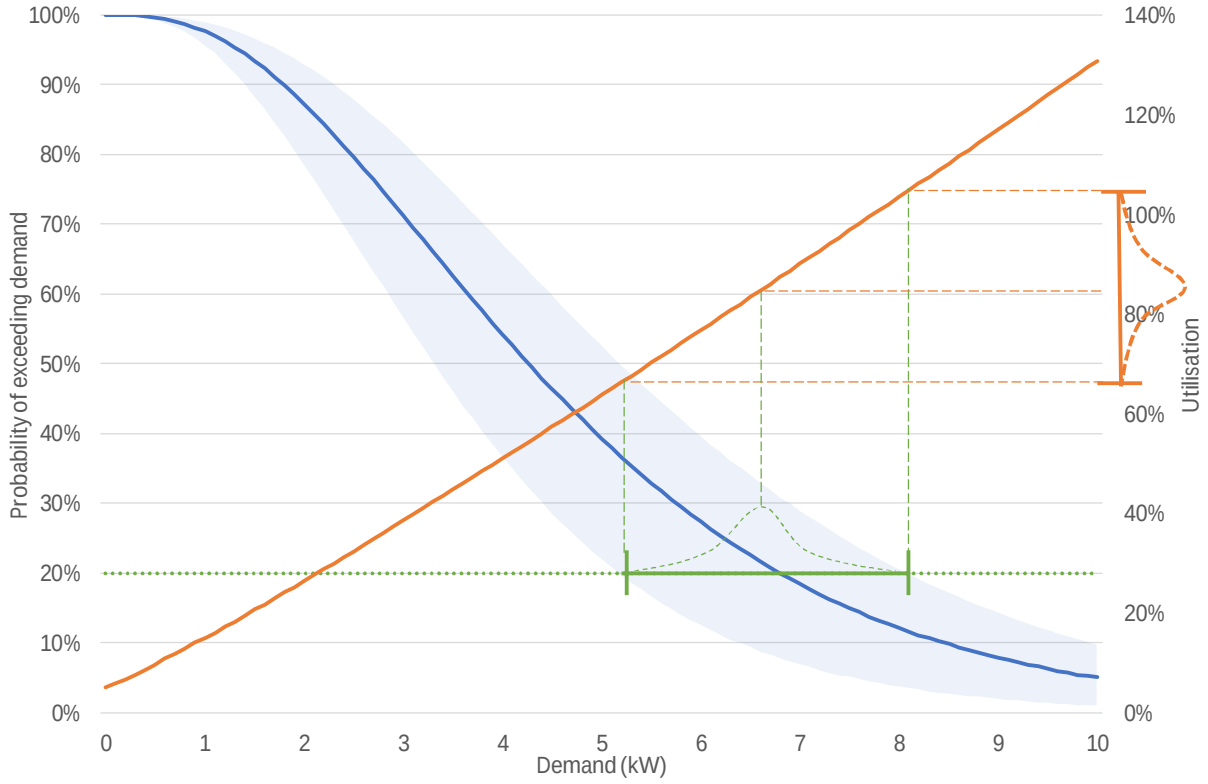
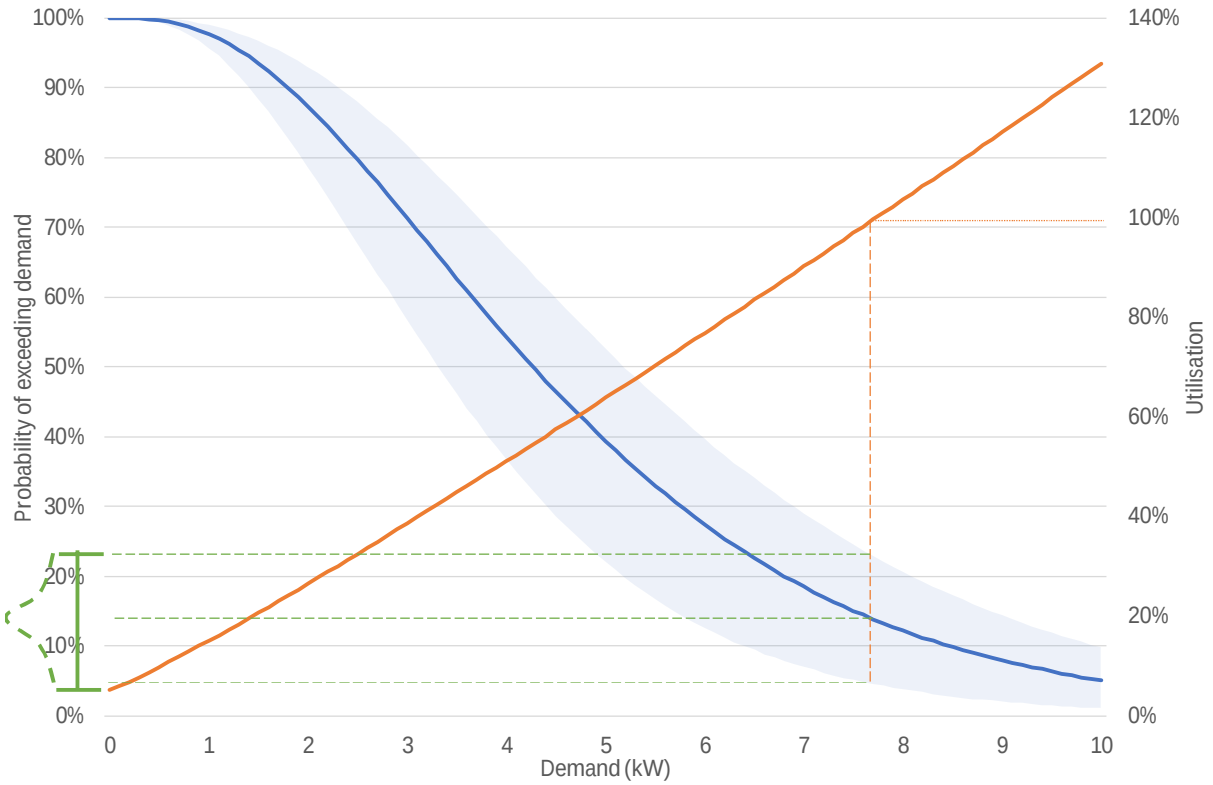


Figure 3-21 Calculating exceedance expectation for a specified utilisation



Expected results: As suggested earlier, a further simplification would be to only take notice of the approximated predictive distribution, and ignore the ranges found in the sampling. This would just tell us that the utilisation corresponding to the level of demand that is exceeded 20% of the time is predicted to be 85% (Figure 3-20), or that 100% utilisation is predicted to occur 15% of the time (Figure 3-21).

Load aggregation: It would be possible to simplify the model further by assuming a simplified model of the network, or potentially not considering any model of the network whatsoever (i.e. only considering the aggregation of load and comparing this to, for example, feeder and transformer ratings). This is *very* similar to what ACE49 does currently. It is not clear whether this saves much time or effort compared to a single IPSA run, and additionally this would not account for losses and reactive power. This would present results in a very similar manner, but by simplifying or removing the modelling of the network, they would inevitably be less accurate.

Outputs: We believe that the LV Designer should be shown results for the whole LV network even if, behind the scenes, the algorithm is looking at each branch and voltage one-by-one. As well as presenting tabulated results (like expected values and ranges), results could be displayed as a "heat map" of network thermal and voltage conditions.

Role of different options: This is not an exhaustive list of options – these could be broken up into more variants with added or reduced complexity. Our view is that all (or most) of these options could be part of a (mostly) automated modelling process. This process would start by applying the simplest method and, if the result was at all ambiguous, escalate to a higher level of complexity before the result is clearer.

For example, if it is found that the 1-in-10-year demand is expected to correspond to a 50% asset utilisation, then there is little need to use a more complex method. However, if the expected utilisation is found to be 90%, then it might be necessary to rigorously incorporate the uncertainty associated with the model and produce an approximate Bayesian predictive distribution.

How exactly this process all fits together will depend on the time and the resource available, and the importance of precision in the calculations. This might depend, for example, on the nature of the customers connected to the network or the voltage level which is being studied.

3.6 Comparison with existing statistical approaches

It is useful to compare the approach we are proposing with the existing statistical approach documented within ACE49. This illustrates that our approach is an evolution of existing practice, rather than being something completely new and unprecedented.

- Type of distribution: Our approach uses *Gamma* and *Weibull* distributions, rather than the simpler *Normal* distribution. Although more complicated, these have been found to be a better bit for the data observed from CLNR.
- Estimation of parameters & data sources: Our approach does not impose any strong assumptions on how the parameters of these distributions are estimated, certainly none as strong as assuming a hard-link between half hourly demand and annual energy consumption. Instead, parameters would be calculated automatically based on statistical inference from smart meter data, on a circuit-by-circuit basis.

Uncertainty about distributions will be reflected using the Bayesian approach, in which parameters themselves are subject to uncertainty. Rather than having parameters calculated initially and only recalculated them infrequently, our approach is designed to periodically draw on new data (which will be provided periodically from smart meters and elsewhere) to update and refine these parameters.

- Level of risk: Our approach does not necessarily involve selecting a risk-level in advance (e.g. calculating a design demand based on 90th percentile demand values). Instead, it could allow a much greater range of demand conditions to be studied within a power flow model. This allows quantification of both the *likelihood* and *magnitude* of any network conditions, rather than just a calculation of the magnitude of a single demand condition with a pre-specified likelihood.

This means that the method can consider quantiles of risk of network conditions (e.g. circuit utilisations and voltages), rather than just quantiles of risk associated with demand. Because power flow is a highly non-linear problem, the risk quantiles associated with network conditions and demand will not necessarily align.

As discussed in 2.1, the preceding statement is also true for the statistical model which underpins the ACE49 approach, however, this is not done in practice as the 90% level of risk is 'baked-in'. However, it would still be possible to reduce the proposed novel approach statistical model down to a single pre-defined level of risk, as is currently done with the ACE49 approach, and to only assess the power flow for this single demand condition. There would still be considerable benefits from the other aspects or our proposed approach.

4 LV Case Studies

4.1 Case study overview

To demonstrate the novel analysis techniques, we have carried out case studies on two of the LV network IPSA models which have been built within this project: Cranwood, and Sinderby.

After the network has been built in the power systems software, there are four steps to be followed in order to complete the case study as per Figure 3-1, as described in Section 3:

1. Fit a probabilistic model for the network's customer demands (Section 3.2)
2. Characterise the response of the network model (Section 3.3)
3. Produce predictive distributions for the network conditions (Section 3.4)
4. When network-specific data becomes available – essentially smart meter data, network monitoring or annual energy consumption data, use it to (i) update the demand model and (ii) use the updated demand model to update the predictive distributions of network thermal utilisation and voltage.

The extent to which we could carry out the final step was limited by not having any such data specific to Sinderby and Cranwood customers. Therefore, we have simulated the effect of obtaining partial smart meter data, for illustrative effect, using 'local' SMETS2[23] data synthesised from the CLNR dataset, and have not simulated the acquisition of monitoring data or annual energy consumption.

Each of the steps above is described in detail below, first for the Cranwood network and then Sinderby.

4.1.1 CLNR data

In this case study, we have used extensively the customer consumption time series produced as part of the Customer Led Network Revolution (CLNR) project. We have used this in the absence of SMETS2 data, although in practice, we expect the CLNR data will continue to be useful for many years even as SMETS2 data starts to accumulate. This data consists of multiple "test cells" e.g. TC1a which provides the consumption profiles for domestic customers.

These case studies have exclusively used the TC1a data. This includes approximately 8,000 half-hourly consumption time series, covering two and a half years and including two winters, collected between January 2011 and December 2013. After filtering out monitored customers with poor data quality, the number of series/ customers fell to around 5,000. As a result of the trial start and end dates, for each customer we essentially have data for two winters and three summers.

For creating seasonal demand models for an individual or specific group of customers, the number of data points available is given by: $[number\ of\ days\ in\ the\ season] \times [number\ of\ repetitions\ of\ the\ season]$ x 48, and the number of data points for specific times of day is given by: $[number\ of\ days\ in\ the\ season] \times [number\ of\ repetitions\ of\ the\ season]$. This latter quantity evaluated for each season is summarised in Table 4-1 below. In practice, for a given customer profile, this might be lower due to individual data quality issues.

[23] Note that this could be any smart meter data, including enrolled SMETS1 data.

Table 4-1: Data available from CLNR trials

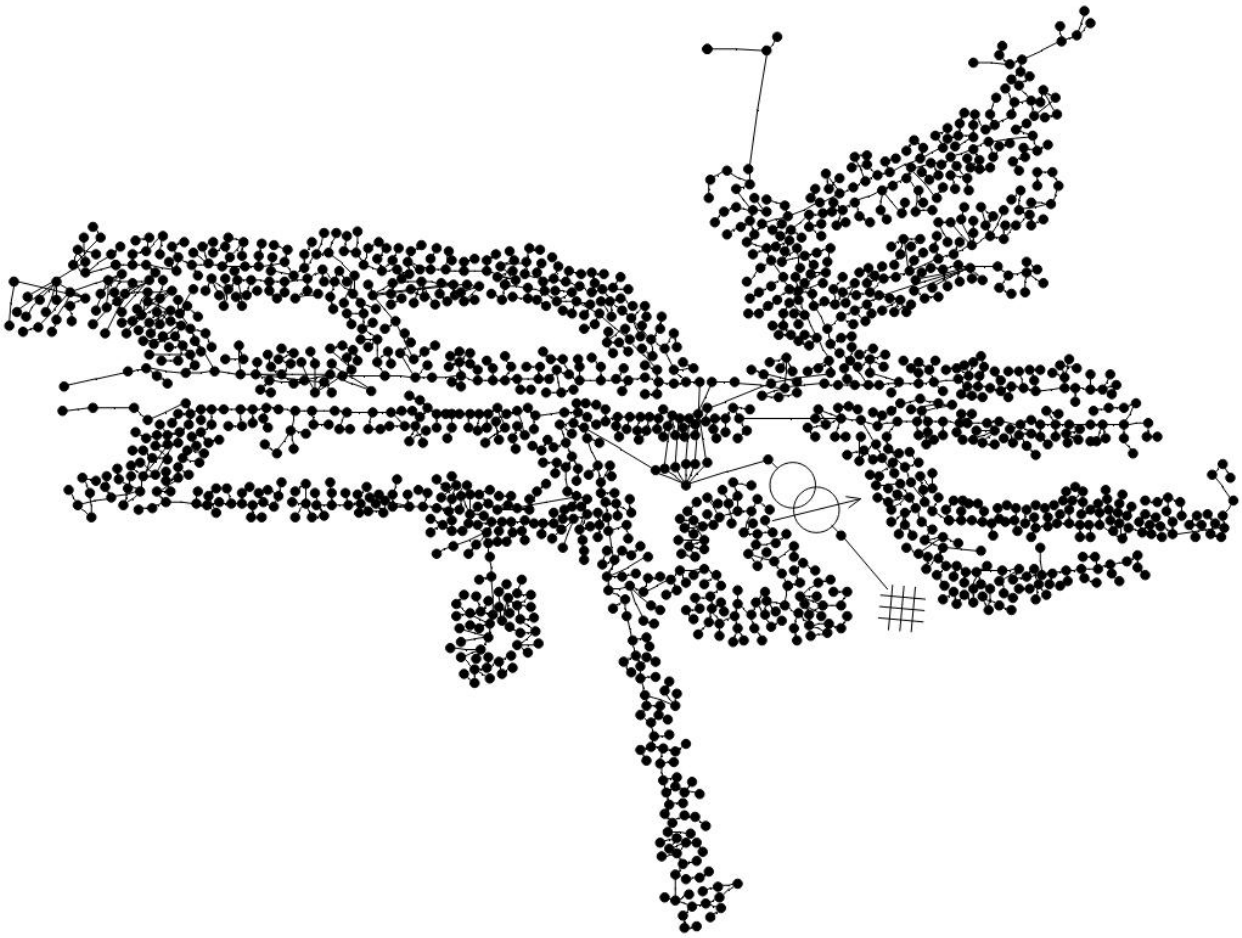
Season	Maximum possible number of half-hours for each time of day	Average number of half-hours for each time of day per complete year	Number of years of complete data
Spring (Mar – May)	217	92	2.36
Summer (Jun – Aug)	254	92	2.76
Autumn (Sep – Nov)	211	91	2.32
Winter (Dec – Feb)	191	90.25	2.12

These are time series of total kWh consumption in each half-hour. To use them in our demand modelling, we multiply each measurement by two, to convert from kWh to kW, giving the average half-hourly power demand.

4.2 Cranwood case study

The Cranwood network has been built in IPSA based on data contained in eAM Spatial. The network is shown in Figure 4-1.

Figure 4-1: Skeleton of Cranwood LV Network



In this case study, we have considered the risk of each of the six circuits becoming thermally overloaded, and of the nodes at the end of feeders experience unacceptable levels of voltage drop.

Cranwood supplies a total of 611 domestic and non-domestic customers as well as 120 unmetered supplies, on six feeder circuits. Data is not available on the exact split of domestic and non-domestic customers. Five of these circuits have a rating of 239 kVA, whereas the sixth has a rating of 308 kVA. All cables are four core copper PILC. The first five have cross sectional areas of 0.2 sq inch, and Feeder 6 has a cross sectional area of 0.3 sq inch.

A description of the six feeders is provided in Table 4-2.

Table 4-2: Description of six Cranwood Feeders

Feeder	Customers	Unmetered supplies	Rating	Busbar Ref
Feeder 1	92	18	239 kVA	100171386
Feeder 2	168	28	239 kVA	100152251
Feeder 3	99	26	239 kVA	100173740
Feeder 4	42	7	239 kVA	100149279
Feeder 5	127	22	239 kVA	100152853

Feeder	Customers	Unmetered supplies	Rating	Busbar Ref
Feeder 6	83	20	308 kVA	100153721

The 400V network is supplied by an 11kV to 433V transformer. This means that the nominal voltage, assuming the transformer is on the nominal tap) at the secondary busbar is typically at around 1.0825 per unit.

4.2.1 Probabilistic model for demand

The first stage in the case study is to produce separate probabilistic representations of the demand of all the customers on each of the six Cranwood feeders – this is necessary as we are interested in the thermal and voltage characteristics of each of these six feeders, and this is caused by the demands on each of these feeders[24]. This was achieved by:

1. Sampling from the CLNR datasets to produce time series of the aggregate demand of the customers supplied from each feeder.
2. Fitting Gamma and Weibull distributions to these time series, for each period in the day (half-hours) and season, as described in the Smart Meter Data Analytics Report.
3. Calculating demand exceedance expectations across the whole year from these distributions.

In order to account for the variability between different customers of the same group (as discussed in 3.2), we repeat this entire process 100 times for each feeder using 100 randomly selected groups of customers each time. Each of these groups is defined the same way, i.e. the same customer numbers, the same number of unmetered supplies, and the same customer types during initial investigations – but nonetheless they exhibit quite distinct consumption patterns. Considering 100 different possible combinations therefore allows us to account for this variability, and quantify the extent to which it introduces uncertainty into the network planning process. As previously discussed, the average of the 100 fitted distributions is taken as an approximation of the prior predictive distribution.

We refer to each of these 100 random samplings from the CLNR data as a "draw". We chose 100 draws as this is a large enough number to allow us to illustrate and understand the variability, but not so large that it becomes computationally intensive to carry out the analysis.

4.2.1.1 Sampling from CLNR

For each of the feeders in the network, customer profiles were randomly drawn from the CLNR datasets. This process was repeated 100 times resulting in a table with 100 columns and N rows, where N is the number of customers connected to that LV feeder. Each "draw" is defined the same way but, as can be seen in Figure 4-2, the value of each row is different for every draw, and so we are clearly accounting for the variability between customers when carrying out the case study. To be entirely accurate, the table would extend to 100 columns (one for each draw) and as many customers as are in the group whose demand is being modelled.

Each profile listed in this table (e.g. TC1a_1139) is a time series of half-hourly consumption kWh values, spanning two and a half years, for a total of approximately 42,000 data points. Again, we use 100 draws for the purposes of illustrating the variability, but in practice the model could use more.

[24] If we had included the transformer in the case study, then it would have been necessary to also fit a model (i.e. probability distribution parameters) for the aggregate demand of all of the customers supplied by the transformer.

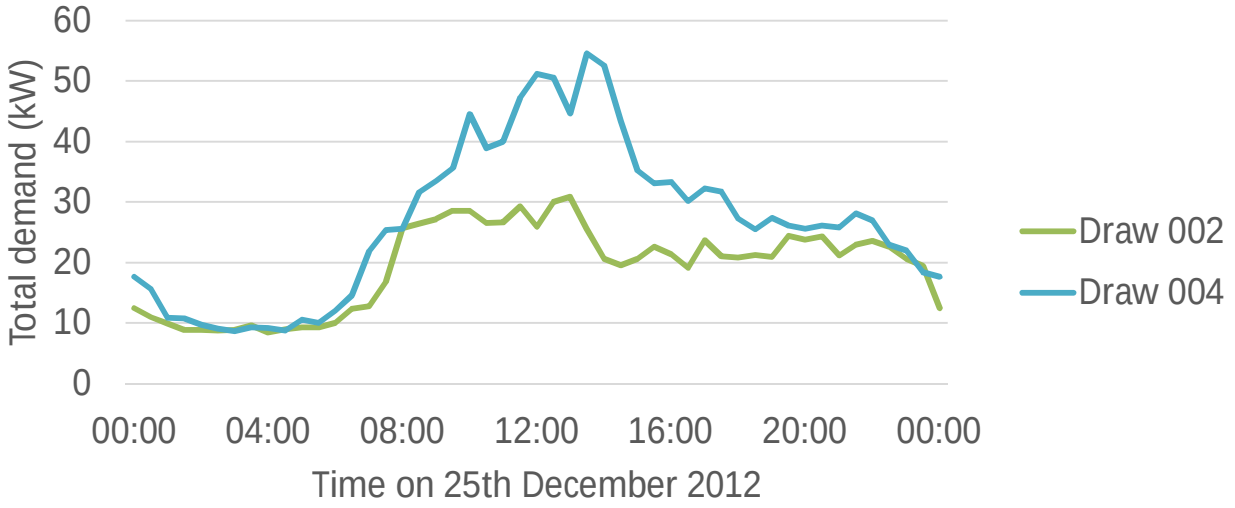
Figure 4-2: Example of randomised CLNR profile selection for illustrative feeder

	Draw 000	Draw 001	Draw 002	Draw 003	Draw 004	Draw 005	Draw 006	Draw 007	Draw 008	Draw 009
Customer 1	TC1a_1139	TC1a_3516	TC1a_3215	TC1a_5346	TC1a_7737	TC1a_4098	TC1a_631	TC1a_5988	TC1a_5474	TC1a_1699
Customer 2	TC1a_5346	TC1a_7528	TC1a_570	TC1a_8495	TC1a_315	TC1a_2815	TC1a_1599	TC1a_8615	TC1a_8137	TC1a_8255
Customer 3	TC1a_8844	TC1a_5879	TC1a_1078	TC1a_2335	TC1a_7581	TC1a_126	TC1a_2129	TC1a_2023	TC1a_725	TC1a_4773
Customer 4	TC1a_6233	TC1a_9016	TC1a_7704	TC1a_4355	TC1a_1424	TC1a_2816	TC1a_4051	TC1a_5151	TC1a_431	TC1a_7353
Customer 5	TC1a_5036	TC1a_3934	TC1a_3135	TC1a_5165	TC1a_6243	TC1a_5957	TC1a_5389	TC1a_1478	TC1a_4977	TC1a_865
Customer 6	TC1a_8709	TC1a_6419	TC1a_1650	TC1a_5492	TC1a_5215	TC1a_5796	TC1a_73	TC1a_2179	TC1a_2526	TC1a_546
Customer 7	TC1a_1827	TC1a_1277	TC1a_1441	TC1a_1050	TC1a_2815	TC1a_7848	TC1a_859	TC1a_6186	TC1a_7388	TC1a_2108
Customer 8	TC1a_6750	TC1a_2015	TC1a_4007	TC1a_191	TC1a_3921	TC1a_4535	TC1a_2866	TC1a_4190	TC1a_6109	TC1a_917
Customer 9	TC1a_7294	TC1a_1344	TC1a_3181	TC1a_3850	TC1a_3185	TC1a_7801	TC1a_1843	TC1a_7766	TC1a_3456	TC1a_6601
Customer 10	TC1a_8968	TC1a_4820	TC1a_4882	TC1a_3106	TC1a_4798	TC1a_5672	TC1a_4911	TC1a_1685	TC1a_2183	TC1a_5869
Customer 11	TC1a_807	TC1a_297	TC1a_3385	TC1a_4037	TC1a_3961	TC1a_7412	TC1a_8704	TC1a_6877	TC1a_5111	TC1a_6752
Customer 12	TC1a_4757	TC1a_8349	TC1a_1900	TC1a_1994	TC1a_3017	TC1a_8882	TC1a_2803	TC1a_2178	TC1a_5315	TC1a_6954
Customer 13	TC1a_513	TC1a_1174	TC1a_5682	TC1a_9088	TC1a_7968	TC1a_6981	TC1a_4525	TC1a_249	TC1a_4531	TC1a_3854
Customer 14	TC1a_6894	TC1a_6917	TC1a_6198	TC1a_9074	TC1a_6738	TC1a_5252	TC1a_3492	TC1a_947	TC1a_9015	TC1a_1747
Customer 15	TC1a_9006	TC1a_5799	TC1a_9163	TC1a_6360	TC1a_473	TC1a_4066	TC1a_3831	TC1a_660	TC1a_1692	TC1a_2689
Customer 16	TC1a_8868	TC1a_2442	TC1a_121	TC1a_5711	TC1a_2129	TC1a_4334	TC1a_8337	TC1a_3987	TC1a_1216	TC1a_6339
Customer 17	TC1a_2548	TC1a_6803	TC1a_3985	TC1a_7445	TC1a_1918	TC1a_6719	TC1a_4355	TC1a_231	TC1a_3908	TC1a_5439
Customer 18	TC1a_7410	TC1a_6561	TC1a_6534	TC1a_1882	TC1a_7716	TC1a_8567	TC1a_455	TC1a_5752	TC1a_7878	TC1a_5413

For simplicity, we have represented the non-domestic customers within the network using domestic profiles. This is because data is not available for Cranwood to define the split between domestic and non-domestic at a granular level. A brief inspection on Google maps suggests that any non-domestic customers will have reasonably small demands, as they are small shops, hair salons etc. There is significant variability in non-domestic customer demands within the CLNR data so we believe that this is appropriate given the limited number of non-domestic customers within the network and the reduced credible range for their peaks, compared to the CLNR SME dataset.

For each draw (e.g. down each column in Figure 4-2), the total demand in each half-hour was determined by aggregating together the demands from each customer and adding 100W for each unmetered supply i.e. considering an unmetered supply to be represented by a constant load. This resulted in 100 sets of time series – i.e. a table with 100 columns (representing each draw) and around 42,000 rows (representing all of the half-hours monitored in the CLNR trials). An example is shown in Figure 4-3, which presents the total aggregate demand of the 42 customers on Feeder 4 for one day (48 periods) for two of the 100 draws. The overall shape of the demand profile is similar but with some important differences (e.g. the peak demand is much higher for Draw 004).

Figure 4-3: Example of randomly selected total demand profiles



To represent the presence of SMETS2 meters, we fixed the selection of a subset of profiles when sampling from CLNR. This had the effect of reducing the variability in demand observed between different draws. This represents the fact that, for customers with SMETS2 meters, we have no (or at least significantly reduced) uncertainty about their patterns of demand, but for those without SMETS2 meters, there is no reduction in uncertainty about their patterns of demand (which, in the Bayesian context, means their parameter values are still uncertain). The overall impact on the group of customers is that the uncertainty reduces somewhat– the greater the assumed penetration of smart meters, the greater the reduction in uncertainty.

For example, if we wanted to simulate the presence of ten SMETS2 meters on a feeder with 18 customers, we would select the same CLNR profiles for customers 1-10, and then randomly select the profiles to represent customers 11-18. This is illustrated in Figure 4-4, where the first ten rows are identical across all columns, and rows 11 to 18 change.

For this case study, we have therefore repeated all of the demand modelling for each of the feeders for a scenario where around 2/3 of customers have a SMETS2 meter – i.e. by "fixing" 2/3 of the randomly selected profiles.

Figure 4-4: Example of randomised CLNR profile selection, with ten profiles fixed

	Draw 000	Draw 001	Draw 002	Draw 003	Draw 004	Draw 005	Draw 006	Draw 007	Draw 008	Draw 009
Customer 1	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808	TC1a_7808
Customer 2	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021	TC1a_7021
Customer 3	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821	TC1a_8821
Customer 4	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423	TC1a_6423
Customer 5	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601	TC1a_1601
Customer 6	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627	TC1a_6627
Customer 7	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750	TC1a_8750
Customer 8	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429	TC1a_429
Customer 9	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987	TC1a_6987
Customer 10	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247	TC1a_1247
Customer 11	TC1a_807	TC1a_297	TC1a_3385	TC1a_4037	TC1a_3961	TC1a_7412	TC1a_8704	TC1a_6877	TC1a_5111	TC1a_6752
Customer 12	TC1a_4757	TC1a_8349	TC1a_1900	TC1a_1994	TC1a_3017	TC1a_8882	TC1a_2803	TC1a_2178	TC1a_5315	TC1a_6954
Customer 13	TC1a_513	TC1a_1174	TC1a_5682	TC1a_9088	TC1a_7968	TC1a_6981	TC1a_4525	TC1a_249	TC1a_4531	TC1a_3854
Customer 14	TC1a_6894	TC1a_6917	TC1a_6198	TC1a_9074	TC1a_6738	TC1a_5252	TC1a_3492	TC1a_947	TC1a_9015	TC1a_1747
Customer 15	TC1a_9006	TC1a_5799	TC1a_9163	TC1a_6360	TC1a_473	TC1a_4066	TC1a_3831	TC1a_660	TC1a_1692	TC1a_2689
Customer 16	TC1a_8868	TC1a_2442	TC1a_121	TC1a_5711	TC1a_2129	TC1a_4334	TC1a_8337	TC1a_3987	TC1a_1216	TC1a_6339
Customer 17	TC1a_2548	TC1a_6803	TC1a_3985	TC1a_7445	TC1a_1918	TC1a_6719	TC1a_4355	TC1a_231	TC1a_3908	TC1a_5439
Customer 18	TC1a_7410	TC1a_6561	TC1a_6534	TC1a_1882	TC1a_7716	TC1a_8567	TC1a_455	TC1a_5752	TC1a_7878	TC1a_5413

4.2.1.2 Gamma and Weibull model

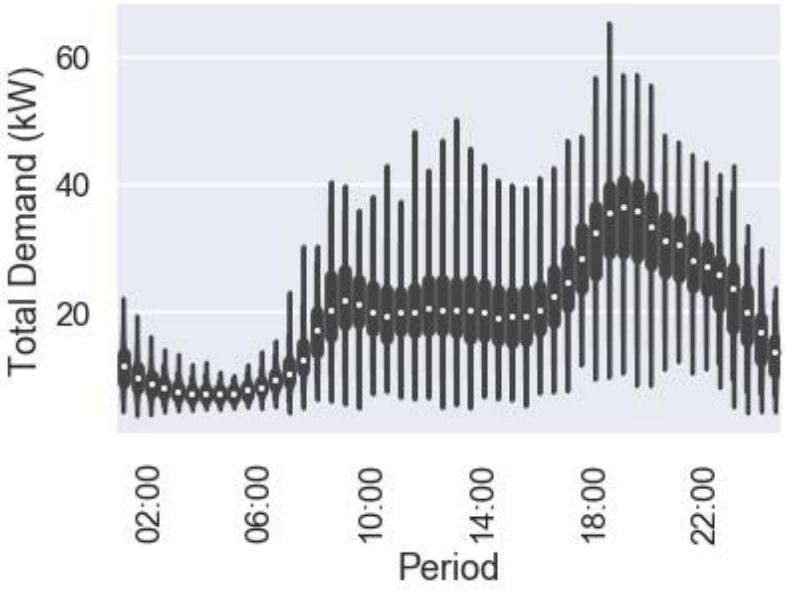
The next stage was to fit a statistical model to the sampled values of total aggregated demand for the group, where a group represents the total demand on one of the six feeders.

Throughout most of the rest of this section, up until Figure 4-16, we show the example of the model being fitted for the same specific draw of demand data for Cranwood Feeder 4, which has 42 customers and 8 unmetered supplies. However, when applying this to the novel analysis techniques method, it is important to recognise that the model for a single draw is incomplete, and it is necessary to consider the variability and uncertainty introduced by having separate models for each of the 100 draws.

The first step is to split the year into four seasons and 48 times of day – this is because demand is known to be heavily season and time-of-day dependent. Figure 4-5 shows the range of demand values observed for a specific group of customers across all of the half-hours in winter. The thick and thin black bars are actually a compressed box plot[25], showing the four quantiles of the data (i.e. the 0th to 25th percentile, 25th to 50th percentile, the 50th to 75th percentile and the 75th to 100th percentile). These show the wide range of possible values. The white dot indicates the average demand in the winter in each time period.

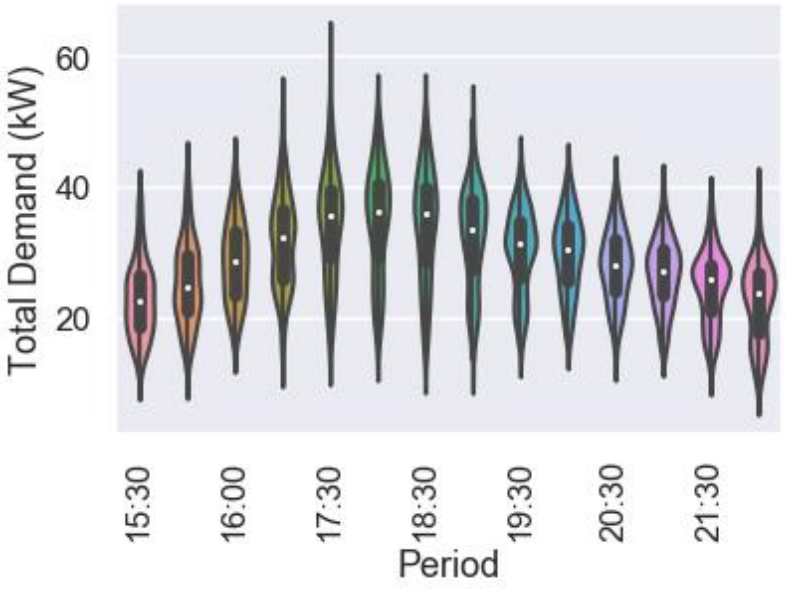
[25] The white dot shows the average, the thin black line at the bottom shows the 0th to 25th percentile, the thick black line below the white dot shows the 25th to 50th percentile, the thick black box above the white dot shows the 50th to 75th percentile, and the thin black line at the top shows the 75th to 100th percentile.

Figure 4-5: Distribution of total demand across half-hour periods in winter, for a single draw of Cranwood Feeder 4



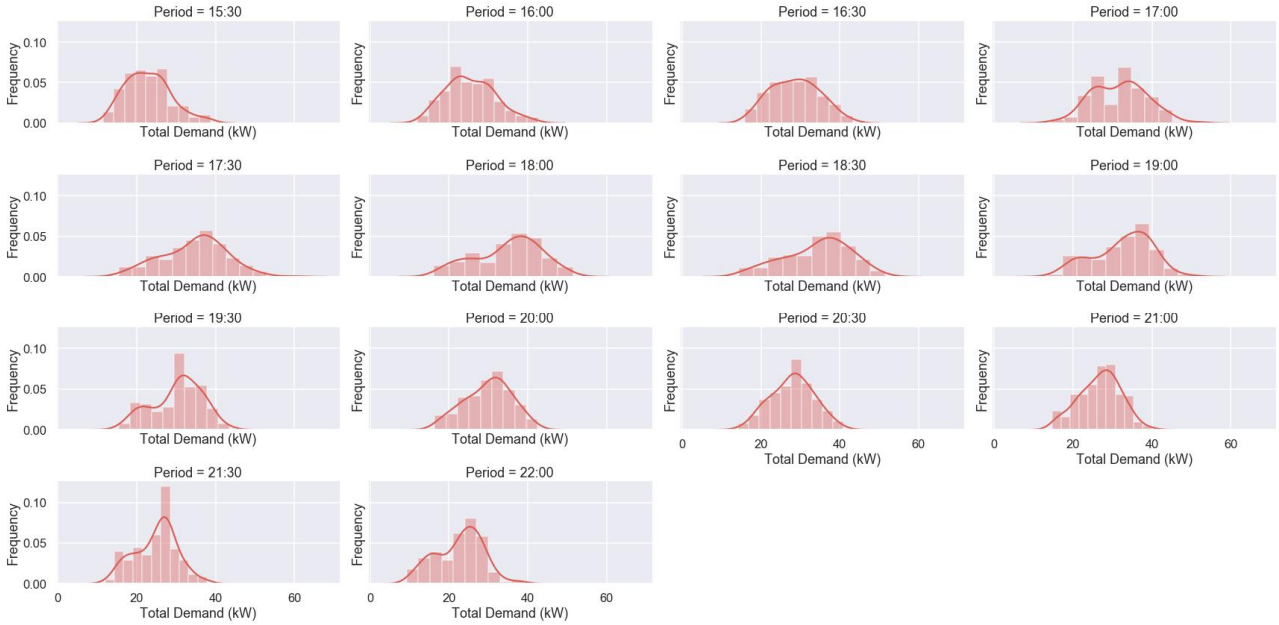
As described in Appendix A, periods are grouped into sequences based on observed features of the data (e.g. relationships between the mean and standard deviation). For example, one sequence is winter between 15:30 and 22:00, as shown below in Figure 4-6. The total demand in each of the periods in this sequence are similar, but have different means, and a standard deviation which varies approximately in proportion to this mean.

Figure 4-6: Winter afternoon and evening sequence of modelled periods, for a single draw for Cranwood Feeder 4



This “violin plot” is essentially showing the histogram of the data for each period, rotated on its side. These twelve histograms are shown explicitly in Figure 4-7. These histograms show the range and frequency of possible values of total demand on Cranwood Feeder 4 for this draw, for each of the half-hour periods in this sequence.

Figure 4-7: Histograms of data for winter afternoon and evening sequence, for a particular draw for Cranwood Feeder 4

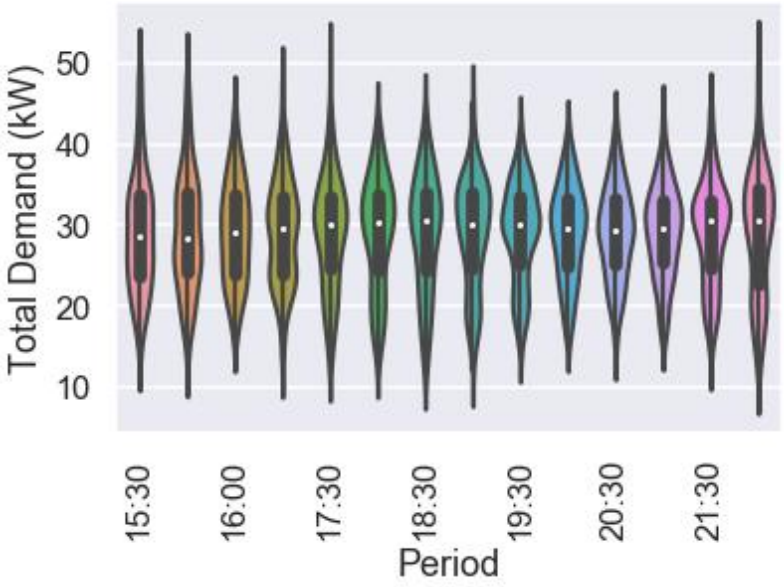


There are limited data points (fewer than 200) for each half-hour period in this winter sequence, since there are only ~90 days in each year in winter from which to obtain measurements, and CLNR includes two winters, as described in Section 4.1.1. Fitting a model to such a small amount of data – e.g. a distinct distribution for each time of day and season - could be challenging to do robustly, and could result in "overfitting"[26]. In order to overcome this, and to reduce the number of parameters needed for the model, all of the data in this sequence is pooled together. This is done by normalising the data with respect to the mean across the whole sequence, so that each period has the same mean (and, for distributions that are modelled with Gamma, approximately the same standard deviation).

This particular period is best modelled using Gamma distributions, therefore the normalisation is achieved by $Normalised\ demand = Demand \times \frac{Sequence\ Mean}{Period\ Mean}$. This is illustrated in Figure 4-8. However, similar steps would be followed for periods which are modelled using Weibull distributions.

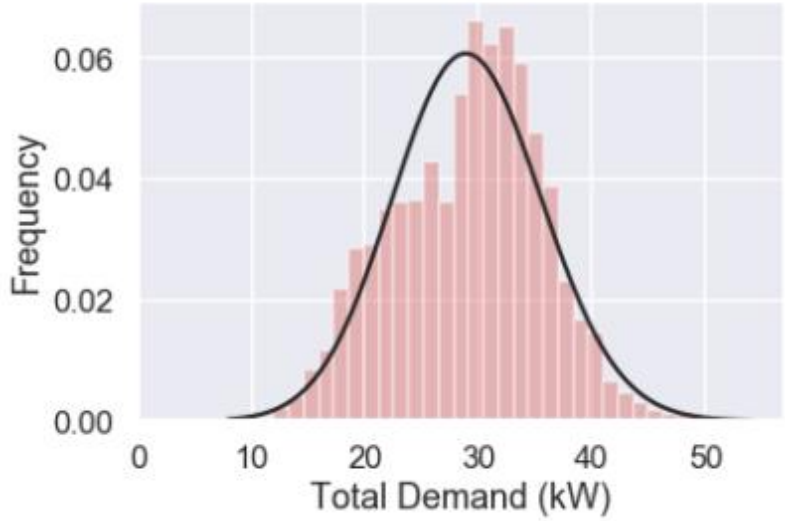
[26] Overfitting occurs when, in fitting a model to a small data set, the resultant model describes that particular small data set very well, but would provide a poor fit for additional data gathered in the future, suggestion a level of precision that is not justified by the amount of data being used.

Figure 4-8: Winter afternoon and evening sequence of modelled periods following mean normalisation, for a particular draw for Cranwood Feeder 4



The combined data set, following normalisation across all periods, is shown in Figure 4-9, with a Gamma distribution overlaid in black[27]. This Gamma distribution is defined by a shape parameter k and a scale parameter θ.

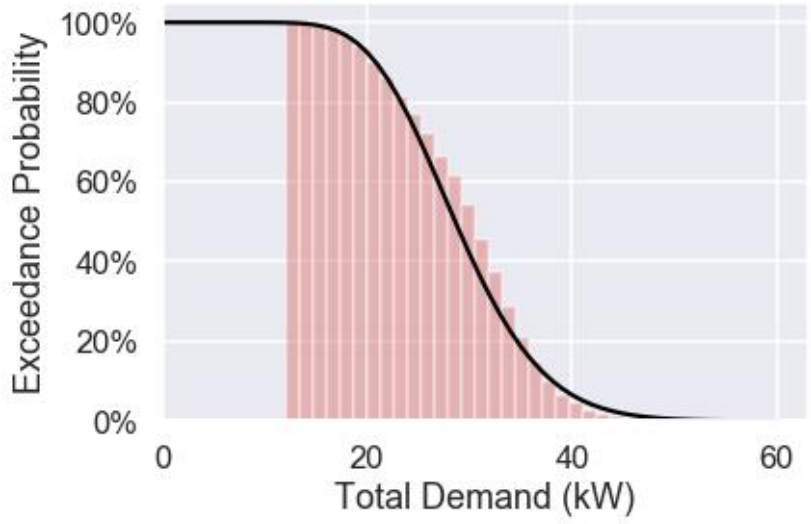
Figure 4-9: Pooled, mean-normalised histogram and PDF for winter afternoon and evening sequence, for a particular draw for Cranwood Feeder 4



The histogram and fitted model of the exceedance probability is shown in Figure 4-10. The red histogram counts the number of data points that are equal to our larger than each value of demand, while the black line shows how this exceedance probability would be described by the gamma distribution

[27] In this case, it appears that this sequence might have been possible to characterise with a normal distribution. However, the gamma and Weibull distributions are much more flexible in terms of the shapes that they can accommodate, and can fit data which is heavily "skewed", which a normal distribution cannot do.

Figure 4-10: Pooled, mean-normalised reverse cumulative histogram and exceedance probability for winter afternoon and evening sequence, for a particular draw for Cranwood Feeder 4



To recover the distributions for each individual period, we rely on the special property of Gamma distributions, that both their mean μ and their standard deviation σ vary linearly with the scale parameter θ:

$$\mu = k \times \theta \qquad\qquad \sigma = \sqrt{k} \times \theta$$

Therefore, we can recover the means (and standard deviations) for each individual period by transforming the scale parameter, essentially undoing the normalisation step taken previously.

$$\theta_{Period} = \theta_{Sequence} \times \frac{Period\ Mean}{Sequence\ Mean}$$

By repeating this for all sequences of periods in all four seasons, 192 probability distributions are found which describe the demand of this group of customers during all 48 time-periods in each of the four seasons. These distributions are a mix of Gamma and Weibull, depending on how the sequences are defined.

This model is described in more detail in [Appendix A]. In the appendix, we define a new parameter $\Lambda = \frac{Period\ Mean}{Sequence\ Mean}$ for Gamma sequences, and an equivalent Γ for Weibull sequences. We have continued to use the statistical convention of representing uncertain variables with capitalised characters as in the Bayesian model these multiplicative (and for the Weibull sequence, additive) parameters are also uncertain.

4.2.1.3 Exceedance expectation

The final stage of the demand modelling is to calculate exceedance expectations for every level of demand. The exceedance expectation is the average number of half-hours in a year for which the total demand from a group of customers is expected to be equal to or larger than any value of demand. This is similar conceptually to a "load duration curve".

The demand exceedance expectations are found by:

1. Evaluate survival functions:

For each probability distribution, the "survival function" of the distribution is evaluated at every level of demand. This gives the probability of each level of demand being exceeded on each half-hour within one day. Figure 4-11 shows an example survival function, for period 36 (17:30 – 18:00) in winter. This shows

that, for example, there is around a 20% chance that demand will be equal to or higher than 40 kW in winter during period 36 (the half hour between 17:30 and 18:00).

Figure 4-11: Survival function example

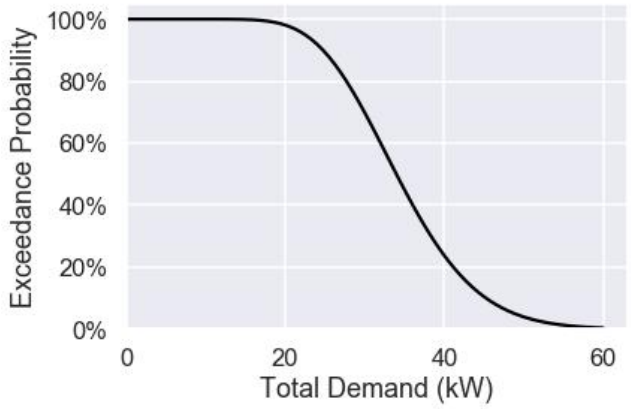
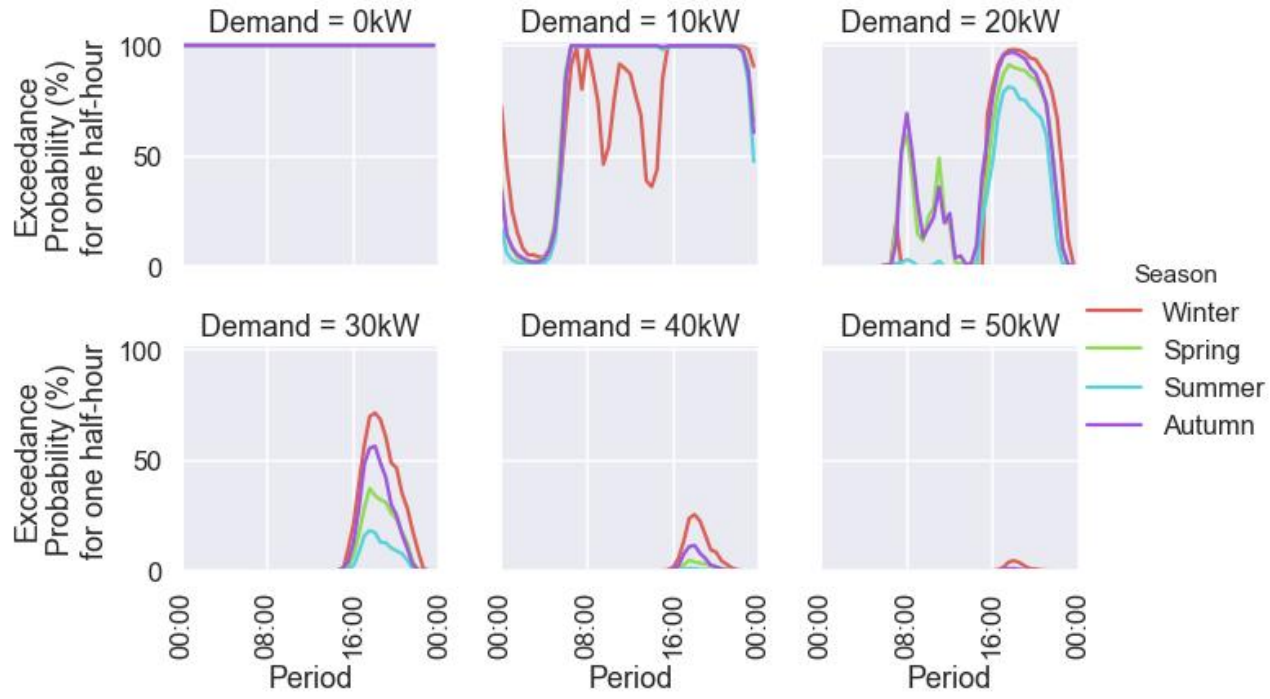


Figure 4-12 shows the result of evaluating the survival functions for each period (for the same group of 42 customers) at various levels of demand.

This shows that, for example, the probability of exceeding 0 kW is always 100%, irrespective of time and day and season, whereas the probability of exceeding 40 kW is 0% for most seasons and times of day, except for a handful of periods in the evenings, where it varies between 0% and 25% depending on season and time of day. For higher demands, probabilities are higher during the winter season (season 1), as would be expected, and lower during the summer (season 3).

For 10kW, there is a clear day/night pattern, although, interestingly, there are some periods in winter afternoons where the chance of having a demand of 10 kW or greater is relatively low compared to the other seasons. This is illustrated by the red line taking lower values between 08:00 and 16:00 for a demand of 10 kW.

Figure 4-12: Evaluation of survival function at different values of demand, across time-of-day and season, for a particular draw for Cranwood Feeder 4

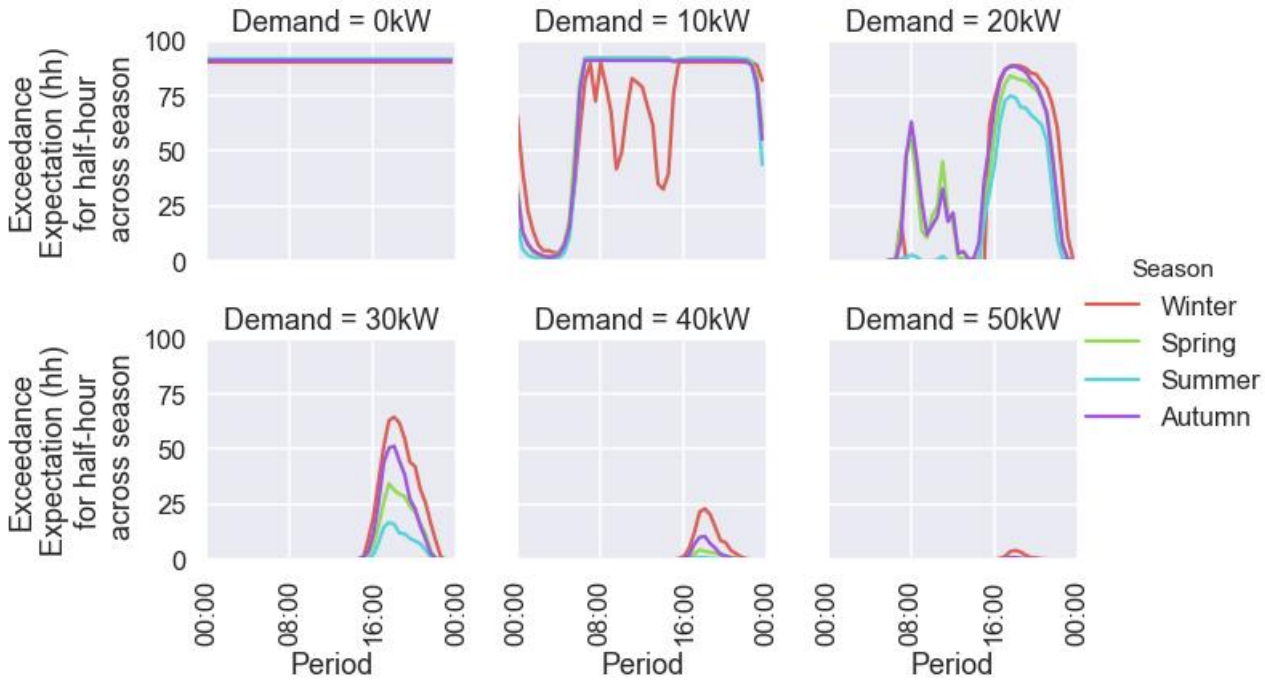


2. Weighting

Each probability is multiplied by its weighting (i.e. the number of days in each particular season). This gives the expected probability in terms of half-hours per season per time-of-day of each level of demand being exceeded during each year during that specific season/time-of-day.

For example, the plot for demand of 30kW shows that there are expected to be around 60 half-hours within a typical year during which the demand on Cranwood Feeder 4 during 18:00 in winter exceeds 30kW. For the same time during the summer season, this is only expected to occur for around 15 half-hours per year.

Figure 4-13: Weighted exceedance expectations for each season/time-of-day in half-hours per year, for a particular draw for Cranwood Feeder 4



3. Aggregating

For each season, the 48 exceedance expectations are added together, and then the values for the four seasons are added together, for each value of demand, to give the probability of each level of demand being exceeded across an entire year. The aggregation across the seasons for different levels of demand is shown in Figure 4-14, with the seasons coloured as in the preceding figures (red – winter, green– spring, blue– summer, purple– autumn). This shows that, for example, over all seasons and times of day in a year, a demand of 20 kW is expected to be exceeded in around 5,000 half hours each year.

Figure 4-15 shows the same information as Figure 4-14 except with a reduced scale for the y-axis. This shows that there is still a non-zero exceedance expectation for higher values of demand such as 40 kW and 50 kW, and that this is highest for winter, then spring, then autumn, and almost zero for summer for 40 kW or higher.

Figure 4-14: Aggregating exceedance expectations across seasons, for a particular draw for Cranwood Feeder 4

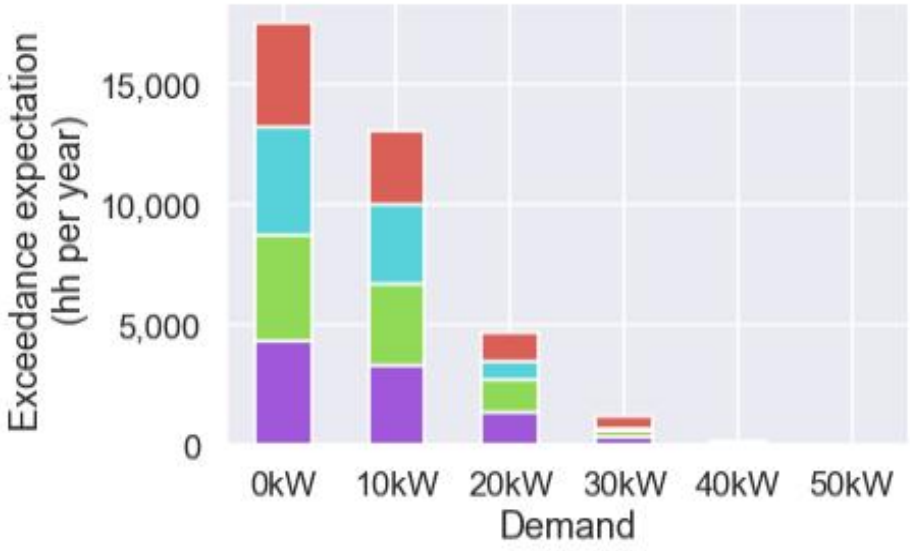


Figure 4-15: Aggregating exceedance expectations across seasons, for a particular draw for Cranwood Feeder 4, with reduced y-axis scale

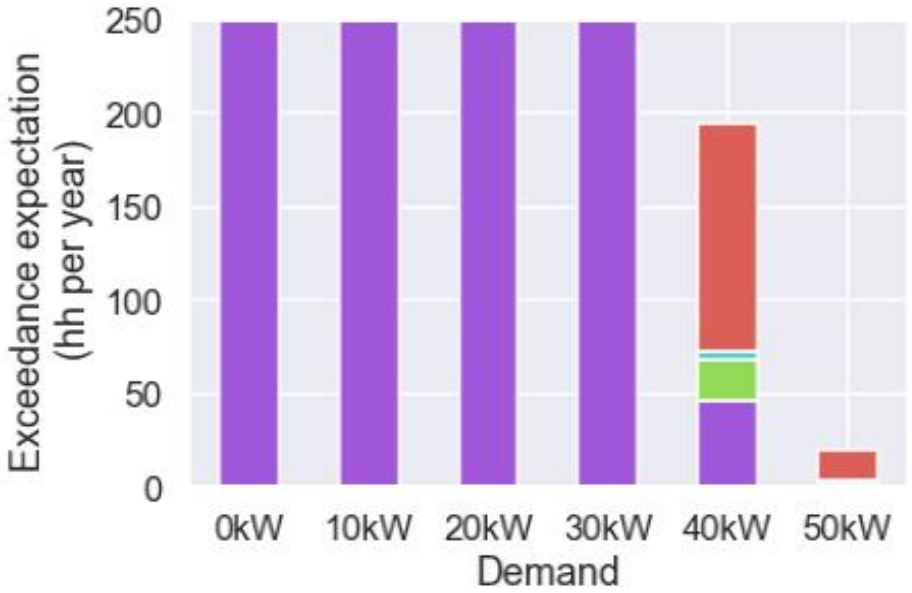


Figure 4-16 shows the exceedance expectation curve for Cranwood Feeder 4. This is equivalent to Figure 4-14 except:

(i) it is plotted over a continuous range of demand values, rather than just a pre-defined set of demand values (0 kW, 10 kW etc), and

(ii) it shows the uncertainty associated with fitting the model to 100 different "draws" from the CLNR database, as described at the start of Section 4.2.1.

(iii) It compares the values from the raw CLNR data with the statistical model that we have developed. The statistical model is shown in red, while the values from the original sampled CLNR data are shown in blue.

The shaded range around each line is the standard deviation of the exceedance expectation for each level of demand, across all 100 draws from the CLNR data. This accounts for the fact that each of these 100 draws will results in a different model being fitted. This shows that, for a given level of demand, there is uncertainty about exactly how frequently that level of demand will be exceeded[28]. These are produced due to the sampling-based approach we have taken to approximating the Bayesian posterior predictive distribution. They would not be a necessary output of the actual modelling, but we have retained them in the discussion of the results as they further illustrate the uncertainty that our proposed method is accounting for.

Plotting these curves on a logarithmic scale gives a clearer view as to what the model is predicting with respect to peak demands. This is shown in Figure 4-17. The model predicts that a demand of 60 kW is expected to be exceeded on average around once every ten years (because the value of the red line at 60 kW is 0.1).

The curves are "flat" up until about 5kW. This is because 5 kW is the minimum demand that will ever be observed from this group of customers.

The model provides a reasonably good fit to the CLNR data, over the whole range of values of demand. The fit is not quite so good for intermediate values of demand (10 to 20 kW), which could be because of the Weibull sequences (which assume that standard deviation is constant across the sequence). This shouldn't affect the analysis of peak values.

In the "tail" (e.g. for very high values of demand), it is difficult to say much about the accuracy of the model, as there are very few data points for the model to fit to and CLNR data points to compare to. Values of 50 kW to 60 kW only occur in some of the draws, and even then, only once or twice across the whole trial period. With more data, it may be possible to better ascertain whether the model fit needs to be improved or whether this is a more modelling limitation.

[28] For the purposes of the case study, we have further simplified the modelling process summarised in Section 3.2. We have not gone as far as fitting a second statistical model to the range of distribution parameters (e.g. a multivariate normal distribution with up to 576 variables – 192 periods with 3 parameters each). Instead, we have directly used the 100 "empirical" fitted distribution parameters in all the subsequent steps of the case study. This significantly increases the clarity of the analysis without any real loss of insight.

Figure 4-16: Demand exceedance expectation for 42 customers and 7 unmetered supplies (Feeder 4)

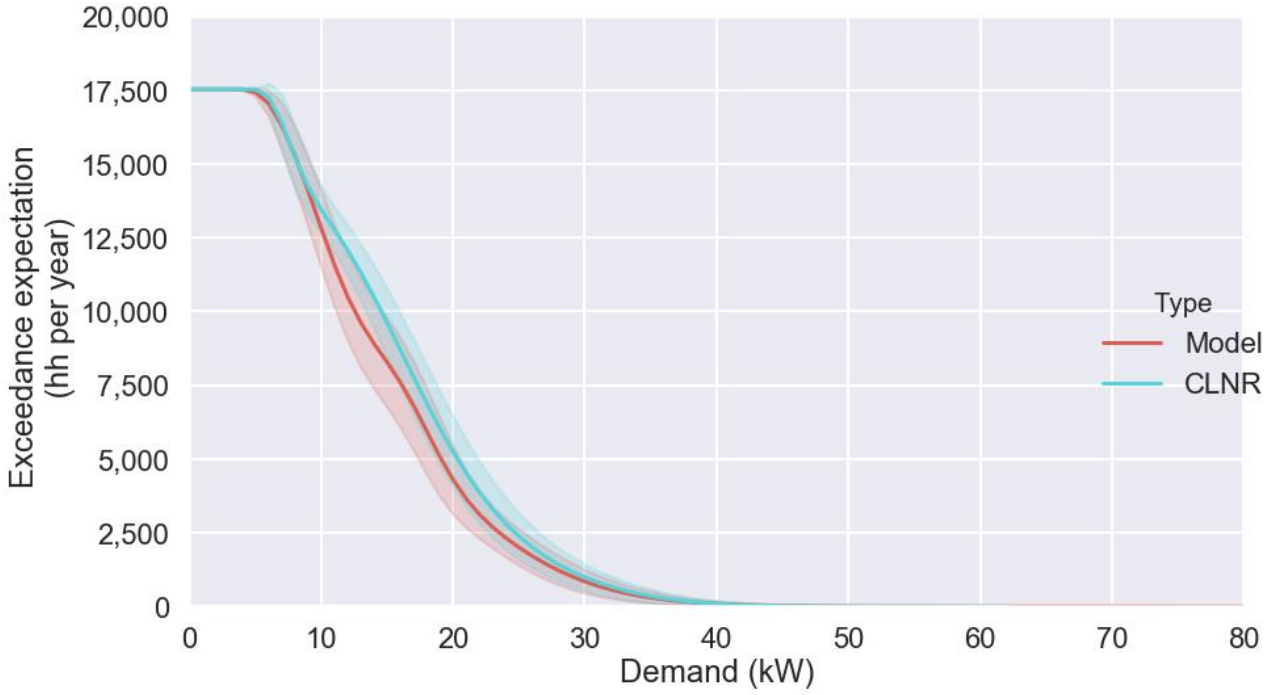
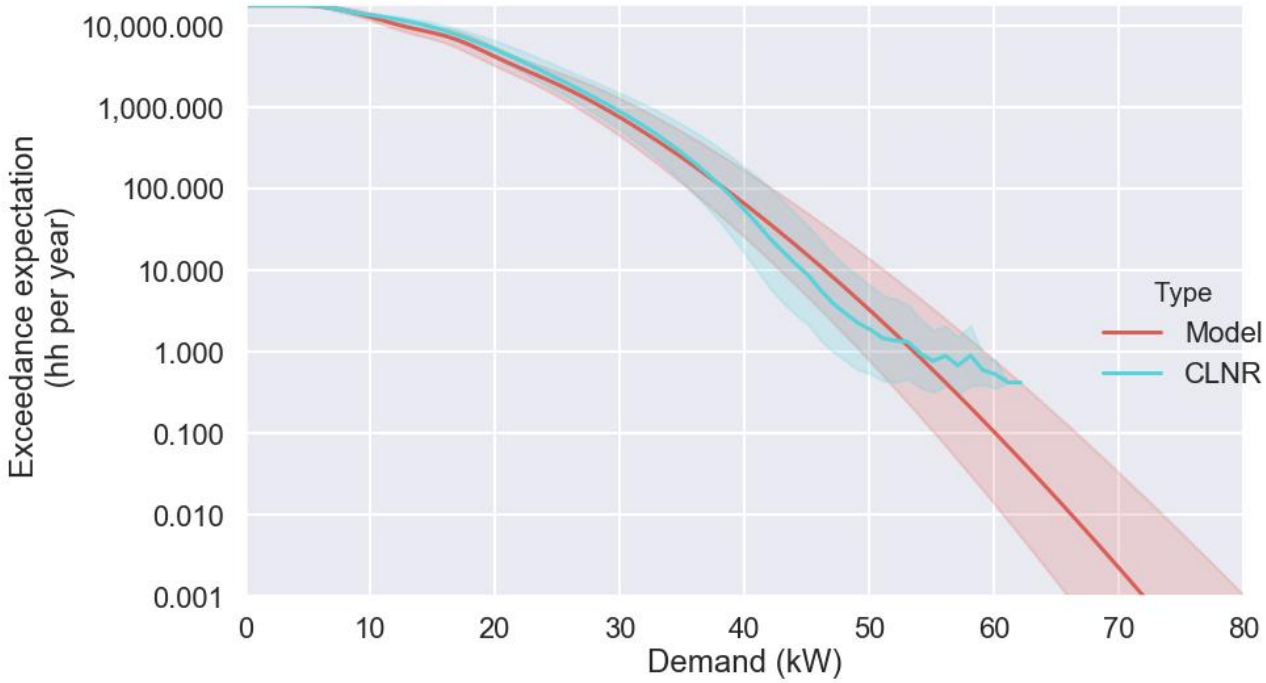


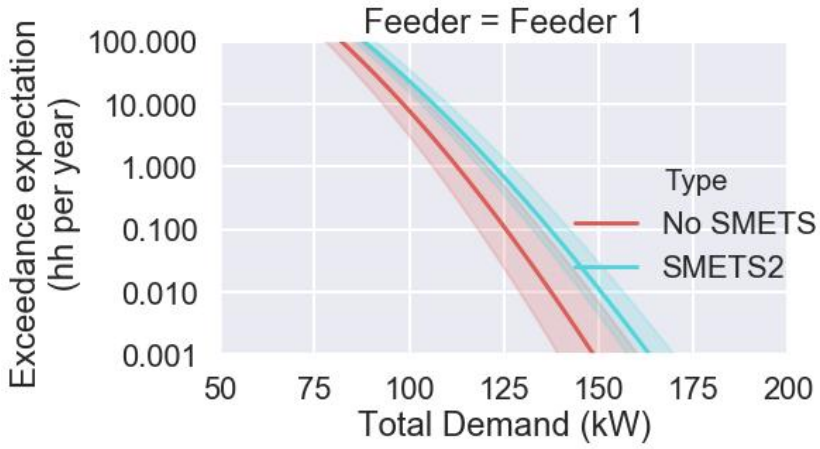
Figure 4-17: Demand exceedance expectation for Feeder 4, logarithmic scale



The modelling of the more extreme values of the "tail" could in principle be further improved using "extreme value theory", which can enable more robust modelling of the extreme values of a dataset. Such approaches are taken in other sectors such as insurance, finance, and hydrology. However, this might not be possible to integrate with the time-of-day/seasonal approach model studied in this report – for example, the time/season dependency might be used for demands up to 40kW, beyond which extreme value theory would be used to fit a "tail" model to describe the higher demands across the whole year. Any application of extreme value theory would probably require data sets covering a longer period of time to be available.

As described in 4.2.1.1, the presence of SMETS2 can be simulated by "fixing" some of the CLNR profiles when sampling. Figure 4-18 shows the results of this for Feeder 1, which has 92 customers[29]. The blue line has 61 of the profiles fixed, to represent having 61 SMETS2 meters (i.e. around 2/3 penetration).

Figure 4-18: Feeder 1 exceedance expectation, showing change with SMETS2 data

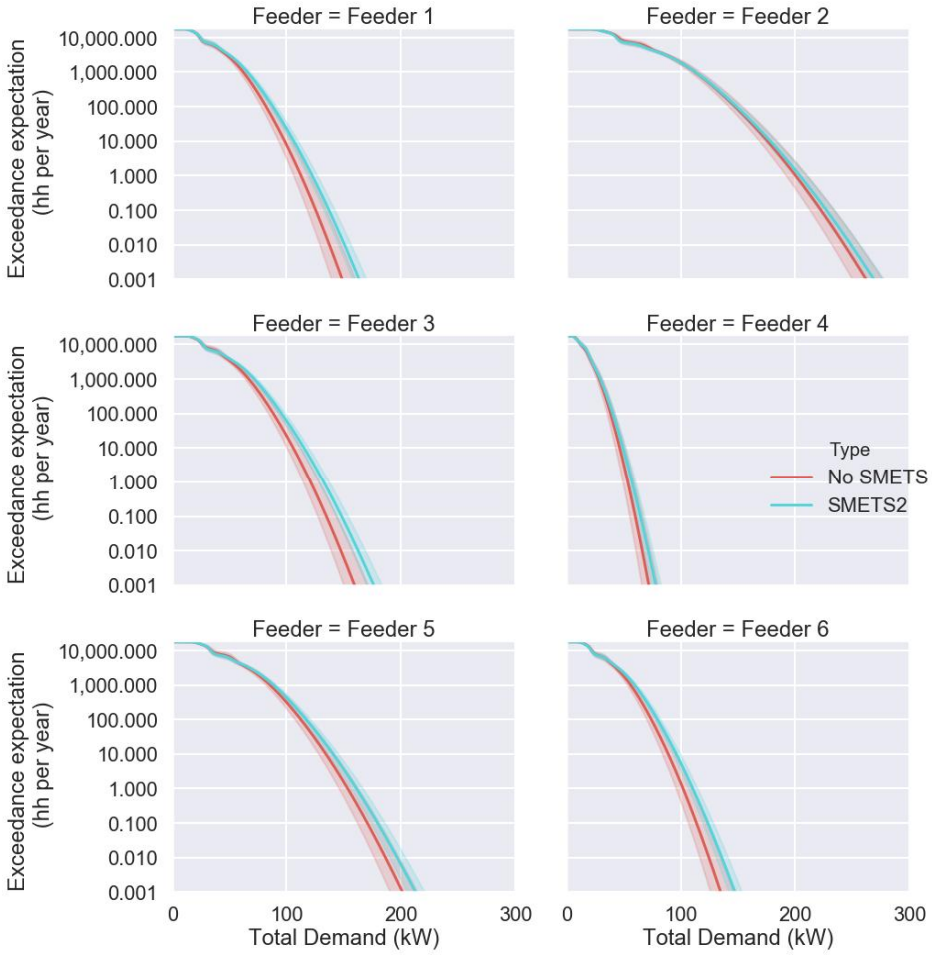


When there are known demand profiles as part of the group, then the uncertainty in the modelled demand reduces. In a full Bayesian implementation, this would correspond to a reduction in the variance of the posterior distributions of the parameters after incorporating smart meter data. This is why the area around the line (which is illustrating the impact of the parameter uncertainty) "tightens" for the blue line compared to the red. In addition, the estimate of the exceedance expectation for each level of demand shifts – for example, in this illustrative case, for a one-in-ten-year event, without SMETS2 data it is ~125 kW, and with it, it is ~137.5kW - although it is still reasonably close to the original estimate (approximately 1 standard deviation away from this central estimate).

Similar curves for all six feeders, without and "with" SMETS2 data, are shown in Figure 4-19.

[29] The figure is "zoomed" in on the range 50 – 200 kW, to show the curves shifting.

Figure 4-19: Demand exceedance expectation curves for all six Cranwood feeders



4.2.2 Network response

The second stage in the case study is to determine (approximately) how the network will respond to different patterns of customer demand[30]. There are three stages to this:

1. Datasets are pulled from the CLNR datasets to represent customer demand for a range of conditions. These are referred to as "trial" datasets. There is no statistical significance placed on this demand data – all that matter is that it is "credible" i.e. it represents a combination of possible values of demand that could feasibly occur on the network. For the example, we have used 2,000 data points – 1,000 with very high demand and 1,000 with very low demand. This is a large enough number to provide some confidence in the output of the regression testing, but isn't overly onerous for modelling in IPSA.
2. The customer demand trial datasets are run through IPSA to analyse voltage and thermal impacts in each case.
3. Regression tests are run on the trial data and IPSA outputs in order to find simple equations which describe the response of the network.

[30] A full implementation of the model would also consider the response to embedded generation.

4.2.2.1 Trial data

Demand "trial data" is passed into the IPSA model in order to enable us to estimate how the model will respond to other sets of customer demand, without having to actually run these through a power systems model. There is therefore no statistical significance attached to the trial data – however it is important that the thermal and voltage conditions which will arise because of these inputs are credible, and that the trial data covers a wide enough range of credible data.

This data is generated in a similar manner as described in Section 4.2.1.1, except with three differences:

1. Since we place no statistical significance on the demands, there is no need to repeat the sampling for multiple draws.
2. We extract 1,000 samples with high aggregate demand, and 1,000 samples with low aggregate demand from the previous sampling exercise – we are only interested in the values of the demands and the impact they have on the network, and not the season or time of day they are associated with.
3. We are now interested in the "disaggregated" profiles of individual customers so that these can be modelled at individual customer nodes within IPSA – rather than just the aggregated demand.

Since Cranwood has a total of 611 customers, the form of the trial data is a table with 2,000 rows (1,000 high demand and 1,000 with low demand) and 611 columns (one for each customer).

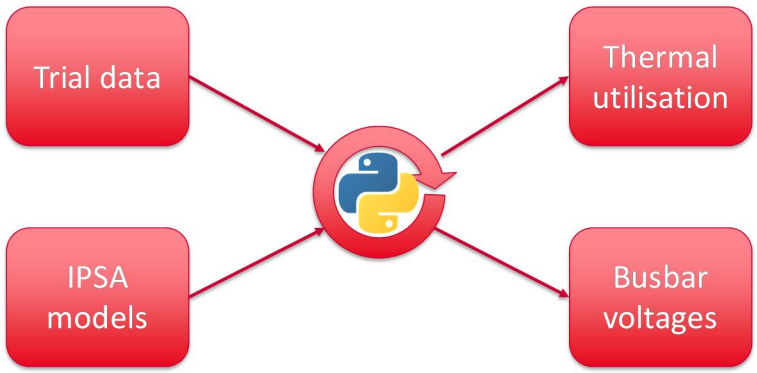
As before, 100W is used for unmetered supplies.

4.2.2.2 Running IPSA trials

The trial data customer demand profiles are automatically loaded into the Cranwood unbalanced IPSA model using a script, and then an unbalanced load flow is executed for each of the 2,000 samples. The model also defines which phase each customer is connected to.

The script records the total kVA utilisation of each branch in the model, as well as the individual red, blue and yellow phase voltages at every busbar. This is illustrated in Figure 4-20.

Figure 4-20: Summary of process for fitting IPSA regression models.



4.2.2.3 Regressions

The aim of the regression testing is to determine simple equations which can approximate the outputs of a full, iterative load flow. In this project, it has been sufficient to use single variable linear equations – although, in practice, more complicated and sophisticated methods might be required. These equations use demand as the "independent variable" and either thermal utilisation or voltage as the "dependent variable".

The total demand on each feeder is aggregated together in order to return single variable regressions – this simplifies the method in that it allows us to avoid using much more complex multi-variate statistics. This

means that, for example, thermal loading will take the form of $y = m \times x + b$ (where y is thermal loading and x is the sum of aggregate demand) rather than $y = b + m_1 \times x_1 + m_2 \times x_2 + m_3 \times x_3 \ldots$ (where x_i is the sum of demands for the nodes within group i).

We envisage that a future modelling tool would "scan" the network as part of the model-build phase, in order to inform how to produce these regressions, and to highlight the branches and the nodes that are the most at risk. This might involve identifying the branches of each feeder which are the most heavily loaded or the nodes which have the highest and/or lowest voltages.

For the purposes of this case study, we have restricted the analysis to consider the first branch of each feeder at the secondary transformer, and a single node near the end of each feeder. These branches and nodes are listed in Table 4-2

The regressions for thermal utilisation (i.e. the kVA or kA circuit loading in per unit of rating) are shown in Figure 4-21. Utilisation accounts for the losses and reactive power associated with the demand (kW) on the network.

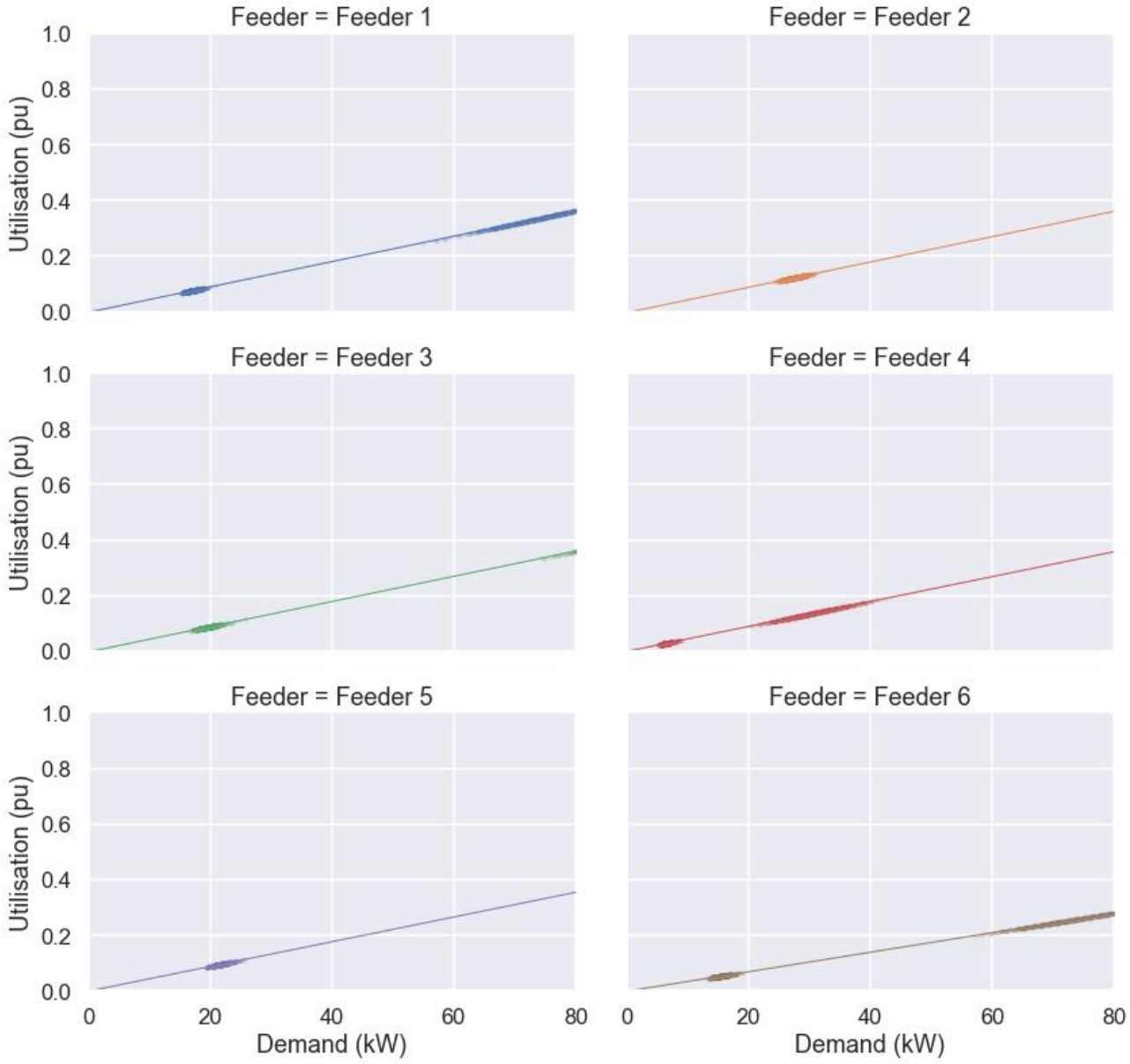
Table 4-3 lists the slope and intercept of these linear equations, as well as their R^2 scores (where R is the coefficient of correlation), which measures how well the linear equation models the results for utilisation[31]. As the R^2 values show, these equations explain almost *all* of the variation in utilisation, based on only the sum of downstream demand on the feeder.

Table 4-3: Regression results for six Cranwood feeders, thermal

Feeder	Slope	Intercept	R^2
Feeder 1	0.0046	-0.0025	1.0000
Feeder 2	0.0046	-0.0040	1.0000
Feeder 3	0.0045	-0.0024	1.0000
Feeder 4	0.0045	-0.0006	1.0000
Feeder 5	0.0044	-0.0012	1.0000
Feeder 6	0.0035	-0.0014	1.0000

[31] To be more precise, the R^2 score describes the proportion of the variance in utilisations which can be explained by the aggregate demand on the feeder

Figure 4-21: Regressions for thermal utilisations of first sections of Cranwood feeders



The regressions for red, blue and yellow busbar voltages are shown in Figure 4-22. For each node assessed, there are three separate sets of data points, and lines, representing each of the voltage phases.

Figure 4-22: Regressions for phase voltages of the ends of Cranwood feeders

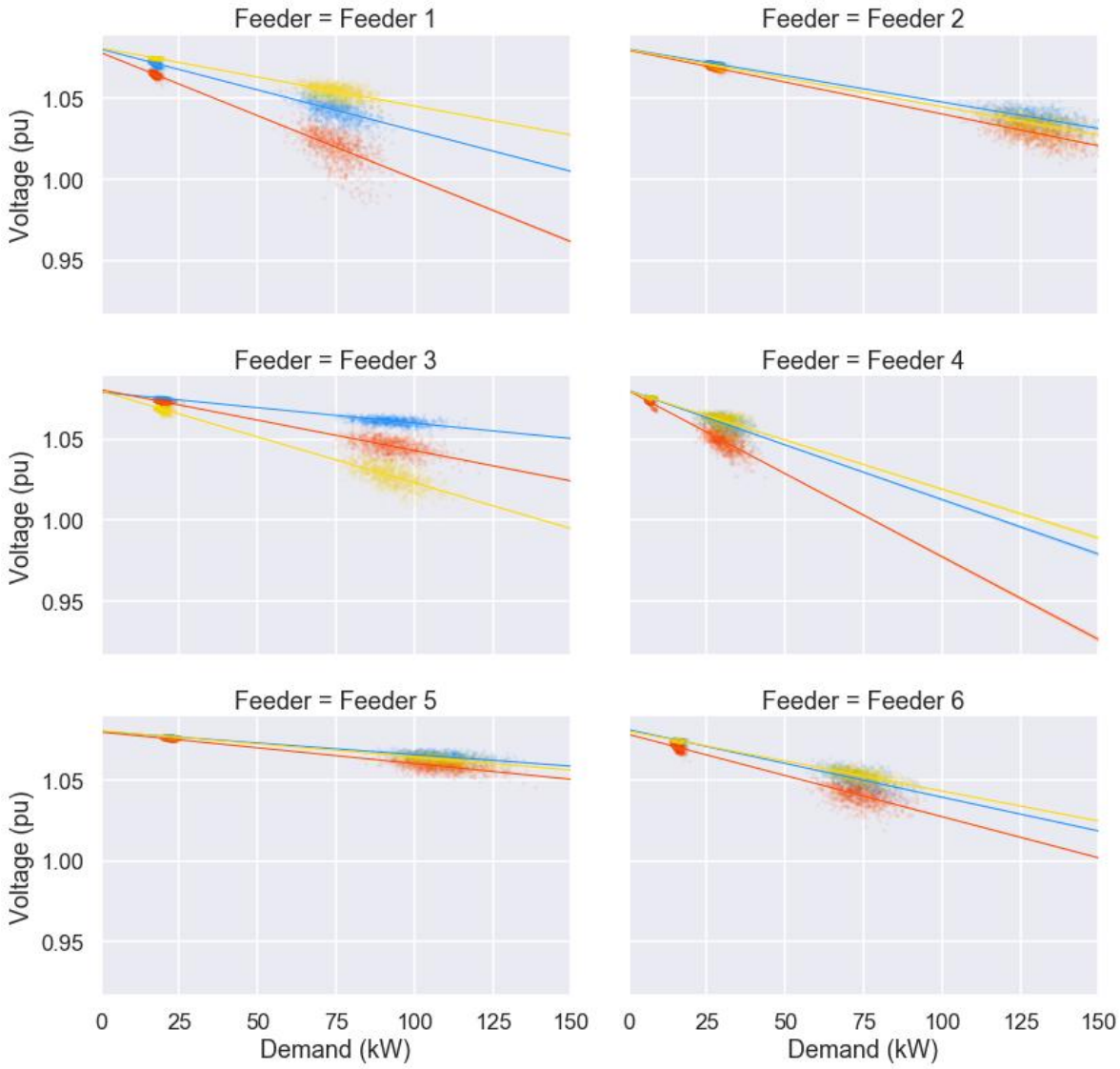


Table 4-4 lists the slope and intercept of these linear equations for voltage (of the form of $y = m \times x + b$ where y is nodal voltage and x is the sum of aggregate demand), as well as their R^2 scores. The "goodness of fit", in terms of R^2, is qualitatively assessed using red, amber and green colours in this table, to highlight which have better and poorer fits. Both the figure and the table show that feeder demand explains a significant amount of the variation in voltage, however, it is clear that there are variations that are not explained purely by the total demand. This is because voltage at the end of the feeder depends not only on the demand on the feeder, but also on how that demand is distributed along the feeder.

The secondary transformer was modelled with a locked tap changer, with the high voltage busbar set to have a constant voltage of approximately 1.08pu, although in practice the transformer tap changer would be set to give a -2.5% buck.

Table 4-4: Regression results for six Cranwood feeders, voltage

Feeder	Node Ref	Slope	Intercept	R^2
Feeder 1	100171386	-0.0008	1.0784	0.9402
		-0.0004	1.0812	0.9411
		-0.0005	1.0805	0.9401
Feeder 2	100152251	-0.0004	1.0798	0.9716
		-0.0004	1.0802	0.9687
		-0.0003	1.0805	0.9692
Feeder 3	100173740	-0.0004	1.0806	0.9536
		-0.0006	1.0797	0.9691
		-0.0002	1.0791	0.9618
Feeder 4	100149279	-0.0010	1.0800	0.9368
		-0.0006	1.0798	0.8655
		-0.0007	1.0803	0.9058
Feeder 5	100152853	-0.0002	1.0797	0.9596
		-0.0002	1.0804	0.9541
		-0.0001	1.0801	0.9235
Feeder 6	100153721	-0.0005	1.0782	0.9473
		-0.0004	1.0803	0.9641
		-0.0004	1.0812	0.9427

For example, for "samples" where more demand is clustered at the end of the feeder, this will result in even lower voltages, but where more of the demand is clustered near the distribution substation, voltage drop will be less. In the future, a multi-variable regression could be used to further improve the accuracy of these regressions (and therefore increase the R^2 value) e.g. to represent varying voltage on the LV side of the distribution transformer busbar as the aggregate demand increases.

4.2.3 Probabilistic network condition

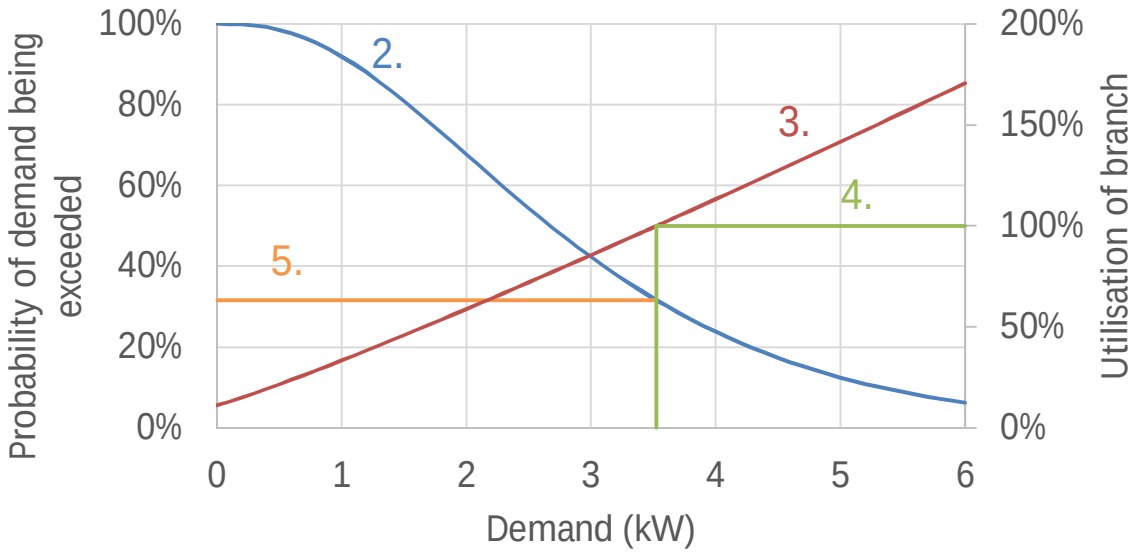
The final stage of the case study is to combine the probabilistic demand model with the characterisation of the network response, in order to determine the probabilistic network condition.

In these results, we have presented far more information that we expect would typically be provided for a LV designer in their day-to-day activities. We have highlighted the results that we expect would typically be provided.

4.2.3.1 Thermal utilisation

Applying the regression equations in Table 4-3 to the exceedance expectation functions in Figure 4-19, returns the thermal utilisation exceedance expectations shown below in Figure 4-24. This is the same as the final step set out in Figure 3-13, which is reproduced below for ease of reading. With the demand curve (2) and the regression line (3) known, we can read off for every level of utilisation (4) the demand associated with this level of utilisation, and therefore, the probability associated with it (5).

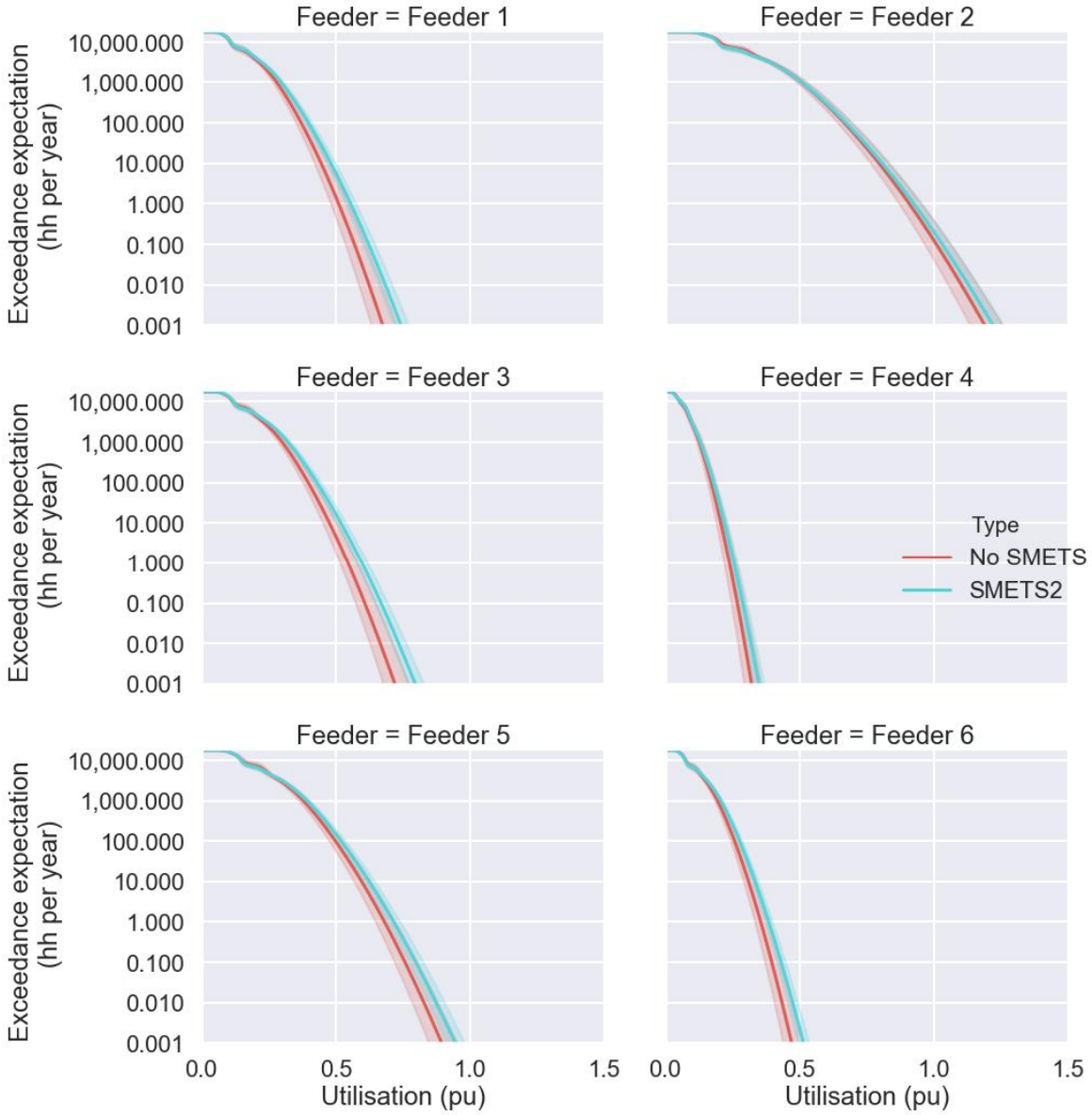
Figure 4-23 Step 5: Probability of demand level



These show the number of half hours in each year that the thermal utilisation of the first branch of each feeder will exceed a certain level (expressed in terms of per unit on kVA rating). Again, the average of all 100 samples is shown, with a shaded area to illustrate the variability around this average.

For most of these feeders, the 1.0 per unit utilisation is reached at well below an exceedance expectation of 0.001 – e.g. 1.0 per unit utilisation is expected to happen less frequently than once in every ten thousand years. However, for Feeder 2, the exceedance expectation of the feeder rating is around 0.1 – this circuit is therefore at risk of overload. This is to be expected, given this feeder has the most customers connected.

Figure 4-24: Thermal utilisation exceedance expectations for feeders

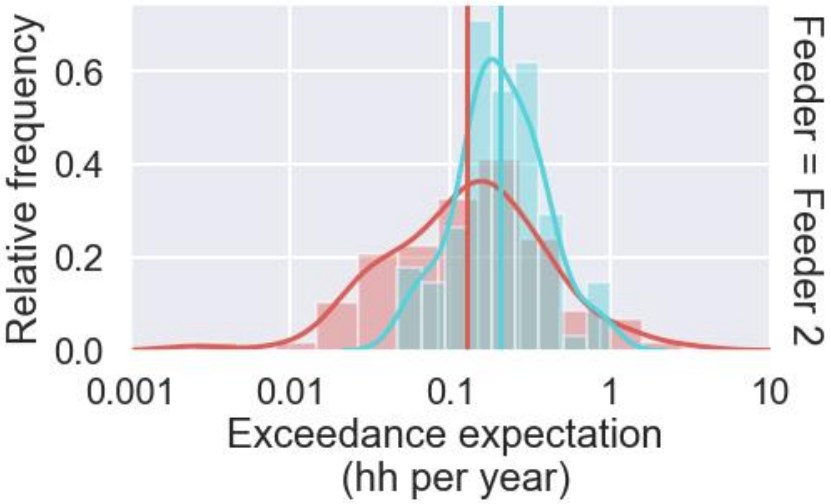


The risk for this feeder can be better understood by examining the histogram of exceedance expectations for the 1.0 per unit utilisation, which further illustrates the impact of the parameter uncertainty. This can be thought of as taking a vertical slice of the curves in Figure 4-24, along the point where utilisation is equal to 1.0pu. These histograms are presented in Figure 4-25.

It can be seen that the exceedance expectations are, approximately, log-normally distributed (i.e. the natural logarithm of the exceedance expectation is normally distributed). This means that the distribution of exceedance expectations is heavily skewed, with a long right tail – i.e. the median is lower than the mean. Plotting this skewed data with a logarithmic scale returns data which is approximately normally distributed.

In addition, the effect of the SMETS2 data is clear – the distribution of sampled values shifts (in this case, to the right) and it also tightens. This means that, the increased information about the customers on this feeder has (i) informed us that they actually have slightly higher than typical demands and (ii) reduced the uncertainty in the parameters that inform this estimate.

Figure 4-25: 100% of rating exceedance expectation for Feeder 2



We can double check our interpretation of the results by plotting a violin plot of the results for each feeder. Figure 4-26 plots the utilisations for each feeder, with and without simulated SMETS2 data, for an exceedance expectation of 0.1 (i.e. once in every ten years). As expected, the one-in-ten-year utilisation for Feeder 2 is around 1.0pu, whereas for all the other feeders, it is comfortably lower.

Figure 4-26: One-in-ten-year utilisation

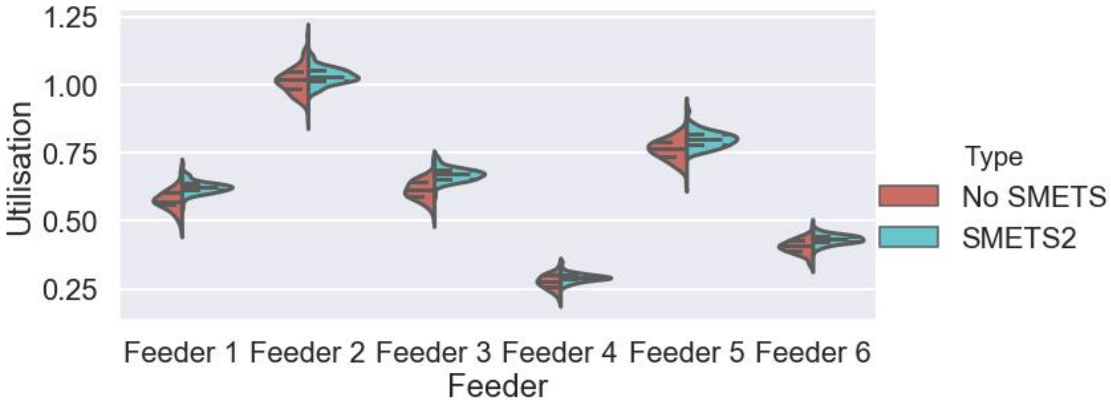


Table 4-5 presents some summary statistics of the 1.0pu exceedance expectation for Feeder 2. It is notable that when the SMETS2 data is simulated, the coefficient of variation (which is the standard deviation divided by the mean) is almost halved, and that the 90% range (between the 5% and 95% value) narrows. To remind the reader, the mean values here are approximations of the Bayesian prior and posterior predictive distributions. The 5% value simply means the 5th largest sampled value (as there were 100 samples), while the 95% value is the 5th smallest sampled value, with these measures of variation only provided to illustrate the impact of the parameter uncertainty

Table 4-5: Summary of 100% thermal utilisation exceedance expectation for Cranwood Feeder 2

Variant	Mean (hh-per-year)	Standard Deviation (hh-per-year)	Coefficient of Variation	5% Value (hh-per-year)	95% Value (hh-per-year)
No SMETS2	0.233	0.350	1.501	0.024	0.748

Variant	Mean (hh-per-year)	Standard Deviation (hh-per-year)	Coefficient of Variation	5% Value (hh-per-year)	95% Value (hh-per-year)
data					
"With" SMETS2 data	0.249	0.175	0.702	0.065	0.593

We expect that the information in Table 4-5 is the sort of information that could be provided for a LV designer. In practice, they would probably only be provided with the mean, which is the approximate predicted value for the exceedance expectation. The value of 0.233 half-hours per year means that exceedance of this rating is a 1-in-4-year event. The range represented around this reflects the uncertainty in the parameters, in a manner we believe is more intuitive than the Bayesian convention. Such values could identify to a planner circuits where there is the most significant uncertainty about the variation in customer demand. This may be helpful when making decisions about the deployment of monitoring, for example. However, this variability is already accounted for in the calculation of the mean of 0.233, as described in Section 2.3.3.

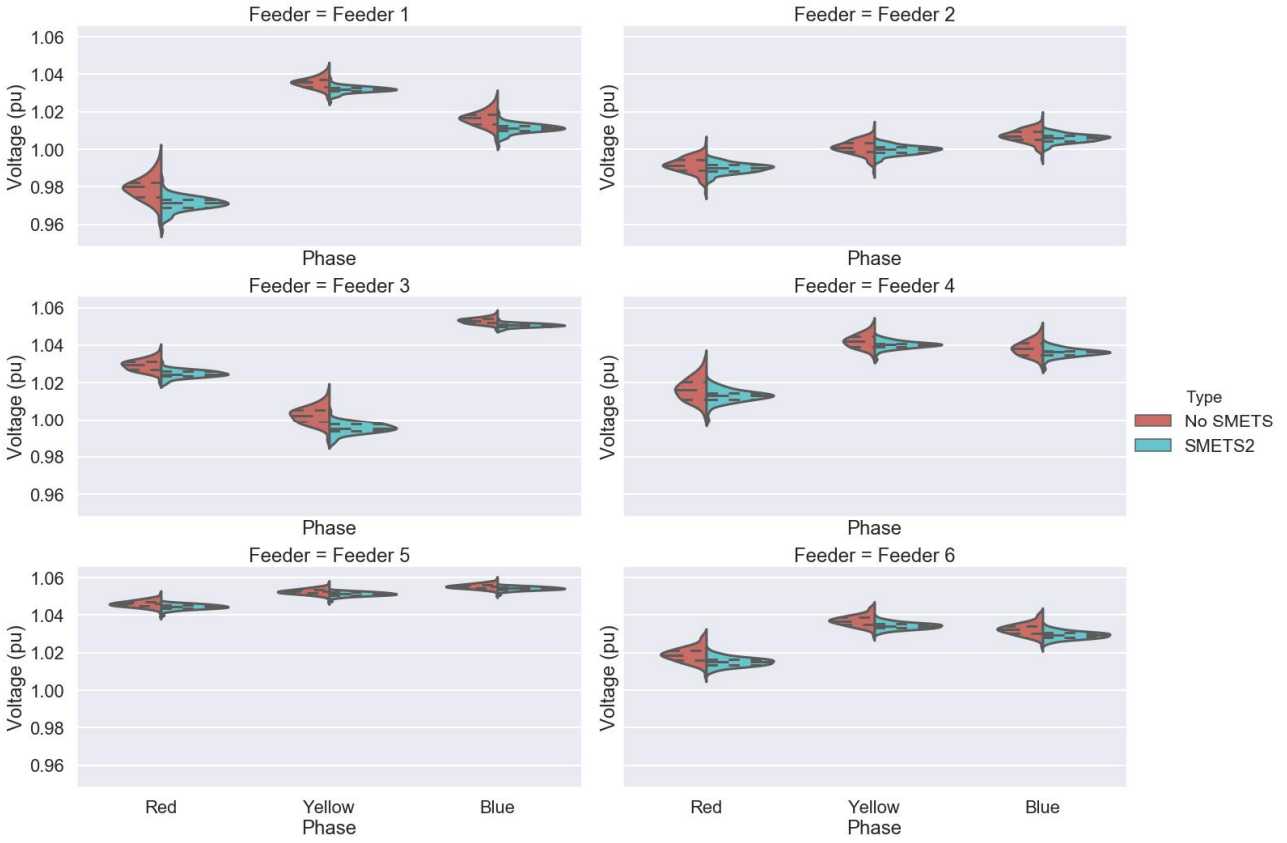
We expect that NPg would need a policy to dictate what actions should be taken with different levels of risk for example, a mean exceedance expectation of greater than 0.1 might require some type of intervention, whereas a mean of less than 0.1 with a 95th percentile value of greater than 0.5 might mean that monitoring is required.

4.2.3.2 Voltage

Similar results are produced for the network voltages, by applying the equations in Table 4-4 to the x-axis of the curves in Figure 4-19. The voltage on each phase of each feeder would be lower than these voltages once in every ten years on average in the long run, for a given sample. The violin plots show the distribution of results across all 100 samples.

However, for Cranwood, the transformer turns ratio mean that none of the feeders are at any significant risk of an undervoltage. The one-in-ten-year low voltages are presented in Figure 4-27, showing that all of them are significantly higher than 0.94 per unit (which is the lower voltage statutory limit). These are the voltages that arise due to periods of high demand, and will be closely linked to the one-in-ten-year high demand.

Figure 4-27: One-in-ten-year low voltage



4.2.4 Adjusted secondary voltage

In order to more comprehensively demonstrate the possible outputs of the method for voltage, we have examined a scenario where the secondary voltage of the transformer at Cranwood is changed from 433V to 416V[32]. This might be implemented as, for example, part of a general LV policy for modifying networks to accommodate more small-scale solar PV, which could otherwise lead to voltage rise on LV networks.

The deceedance[33] expectation curves for each phase on all six feeders, following the reduction in the tap setting, are shown in Figure 4-28. Visual inspection suggests that there is a risk for the node on red phase of feeder 1, given the position of the exceedance expectation curve at 0.94pu voltage is close to 1 half-hour per year. The other feeder nodes appear to be still safe from the risk of undervoltages.

Summary statistics for the 0.94pu deceedance expectation of the studied node on the red phase voltage of Feeder 1 are presented in Table 4-6. As before (Table 4-5), the values are log-normally distributed, such that the distributions are skewed and the median value is less than the mean. The coefficient of variation decreases and the 90% range narrows with the presence of the SMETS 2 data.

Table 4-6: Summary of 0.94 pu voltage deceedance expectation for Cranwood Feeder 1, with 416V no-load secondary voltage setting

Variant	Mean	Standard	Coefficient of	5% Value	95% Value

[32] For the purposes of this illustrative case study, we haven't specified exactly how this is achieved but it could be achieved by setting the transformer tap changer to reduce the target voltage.

[33] We use term "deceedance" to describe under voltage conditions and exceedance to describe over voltage conditions.

	(hh-per-year)	Deviation (hh-per-year)	Variation	(hh-per-year)	(hh-per-year)
No SMETS2 demand data	0.487	0.757	1.553	0.019	1.516
"With" SMETS2 demand data	1.465	1.047	0.714	0.547	3.589

This means that for Cranwood Feeder 1, voltages lower than 0.94 pu are expected in the long-run once every two years (corresponding to the value of 0.487 half-hours per year)

Figure 4-29 and Figure 4-30 show violin plots of, respectively:

(i) the deceedance expectation of 0.94 per unit, and

(ii) the one-in-ten-year low voltage.

These confirm that the only feeder with any pronounced risk is Feeder 1, specifically on the red phase. This is due to the very significant phase imbalance on Feeder 1, as can be observed in the position of the regression lines in Figure 4-22.

Figure 4-28: Low voltage utilisation deceedance expectations, with reduced secondary voltage

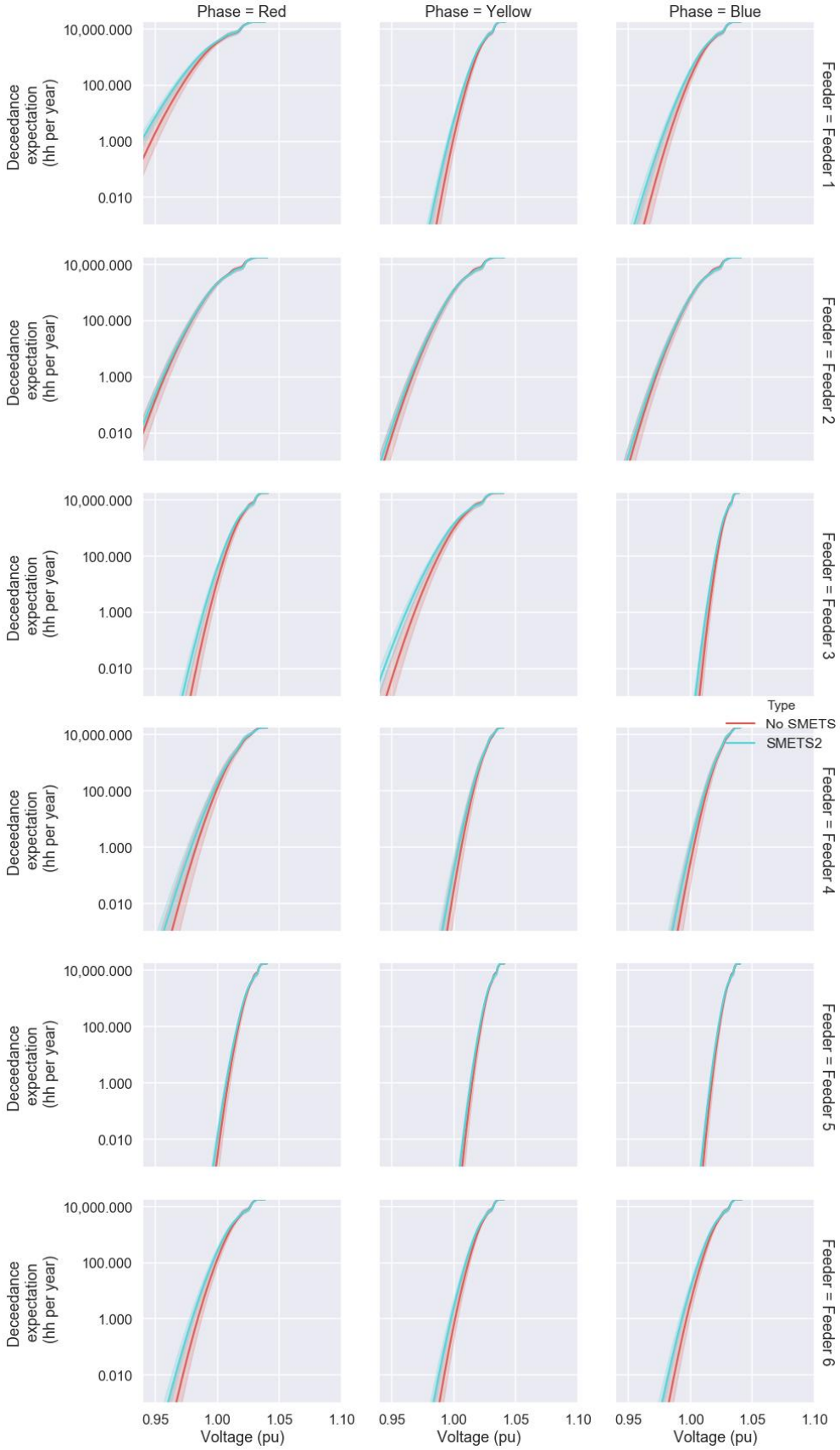


Figure 4-29: 0.94 lower voltage deceedance expectation, with reduced secondary voltage

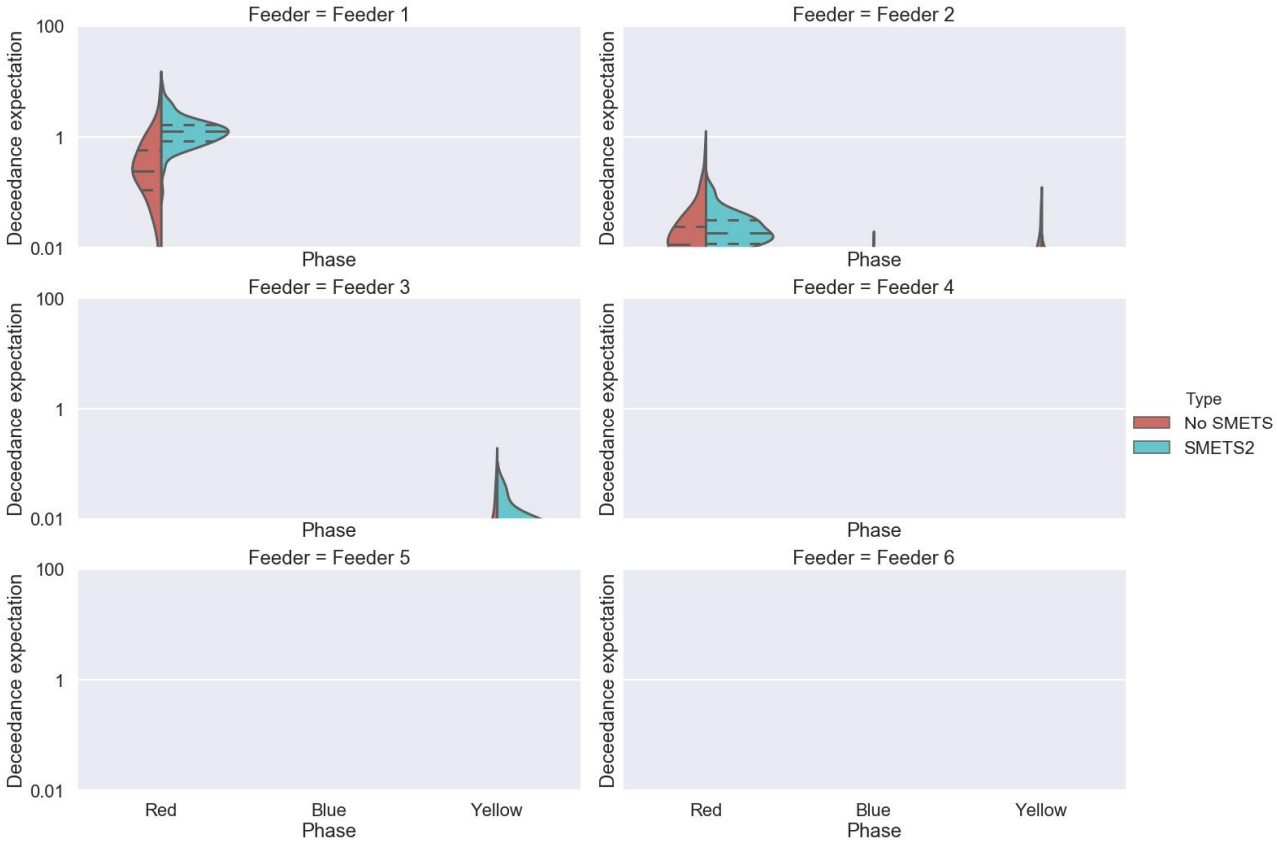
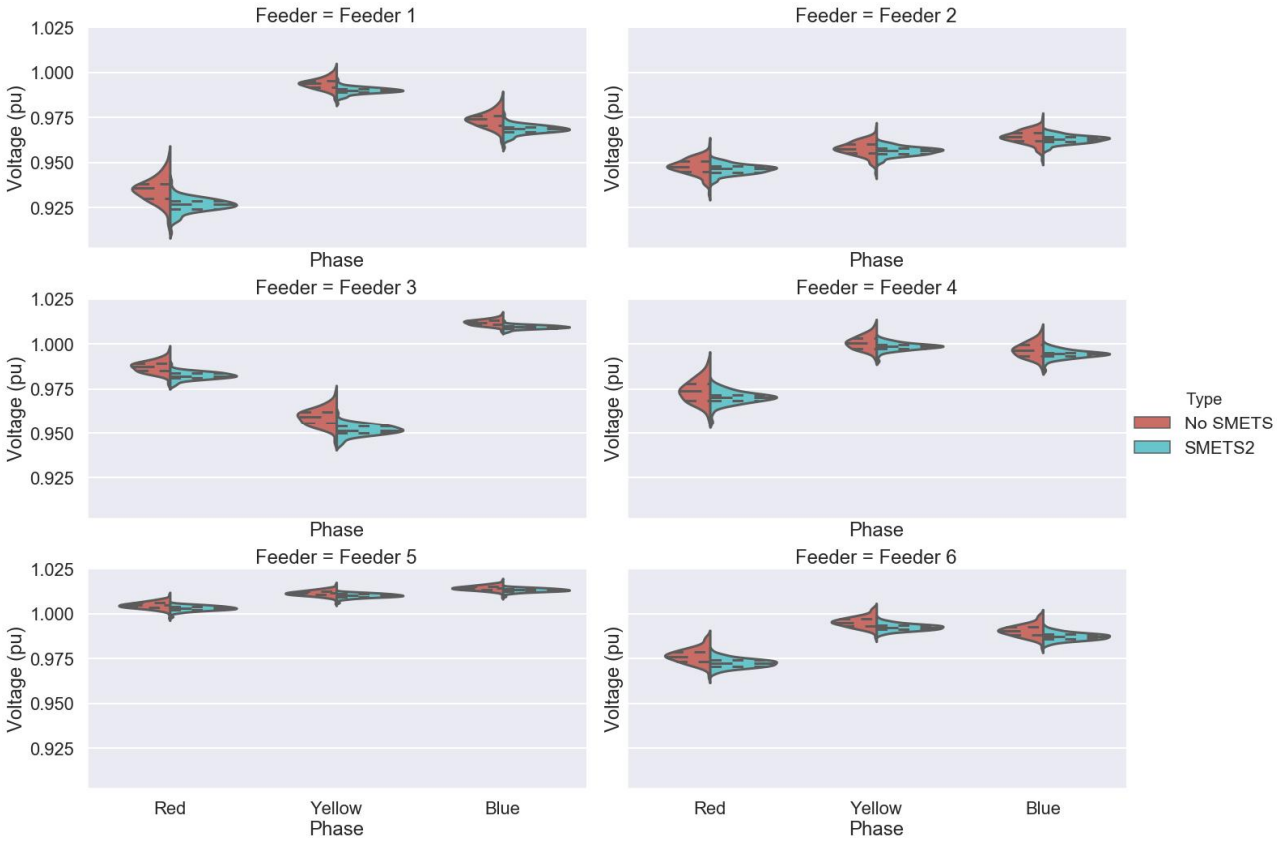


Figure 4-30: One-in-ten-year low voltage, with reduced secondary voltage



4.3 Sinderby case study

The steps described in Section 4.2 were also repeated for the Sinderby LV network model. The Sinderby network has been built in IPSA based on data contained in eAM Spatial.

Sinderby supplies a total of 55 domestic and non-domestic customers as well as 10 unmetered supplies, on two feeder circuits. As with Cranwood, data is not available on the exact split of domestic and non-domestic customers. The North Feeder is rated at 232 kVA and the South Feeder at 301 kVA.

A description of the two feeders is provided in Table 4-2, including a breakdown of customer connections by phase.

Table 4-7: Description of two Sinderby Feeders

Feeder	Customers	Unmetered supplies	Rating	Branch	Busbar Ref
North Feeder	18	4	232 kVA	141107475_1	137956470
Red Phase	5	4			
Yellow Phase	6	4			
Blue Phase	11	4			
South Feeder	37	6	301 kVA	141107463_1	138026433
Red Phase	23	6			
Yellow Phase	7	6			
Blue Phase	13	6			

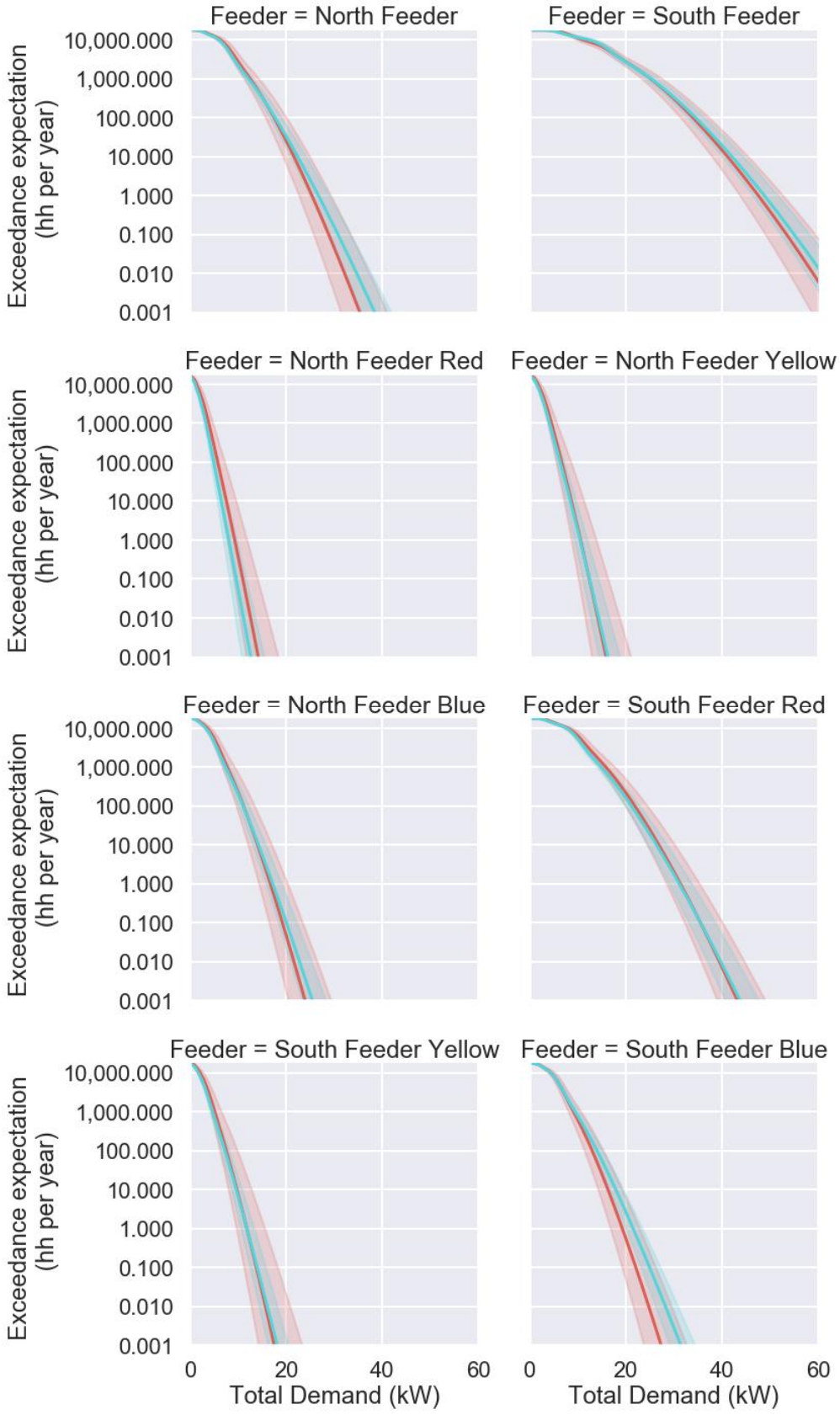
The 400V network is supplied by an 11 kV to 433 V transformer. This means that the nominal no load voltage at the secondary busbar is typically at around 1.0825 per unit.

4.3.1 Probabilistic model for demand

The demand model is determined using the same steps as set out in Section 3.2 and demonstrated for Cranwood in Section 4.2.1. Although the numbers of customers are different, the steps taken are exactly the same, and are therefore not presented in full again for Sinderby.

The exceedance expectations for demand from this model are shown in Figure 4-31. For Sinderby, it is necessary to model the demand individually for each phase, as described in the next subsection.

Figure 4-31: Demand exceedance expectation curves for all Sinderby feeders and phases



4.3.2 Network response

The response of the branches and busbars in Table 4-7 is determined in the manner described in Section 3.3 and demonstrated for Cranwood in Section 4.2.2.

All of the regression equations, and their R^2 values are listed in Table 4-7, with the regressions plotted in Figure 4-32 and Figure 4-33 for thermal and voltage respectively. The regression equations provide a less good fit (in terms of R^2) for Sinderby than for Cranwood for voltage, suggesting that this might be a network where a multi-variate regression is required.

Note that we are not forcing the regression lines to take any specific value at demands of 0 kW. This means that, for demands of 0 kW, the results will differ slightly from the actual voltages in the network – which we would expect to all be equal to ~1.08pu on all three phases. However, doing this would result in a poorer fit for the model at the demand levels in which we are interested – therefore, it is considered acceptable to have a slightly less accurate model for 0 kW since it results in greater accuracy for higher levels of demand[34].

Figure 4-32: Regressions for thermal utilisations Sinderby feeders

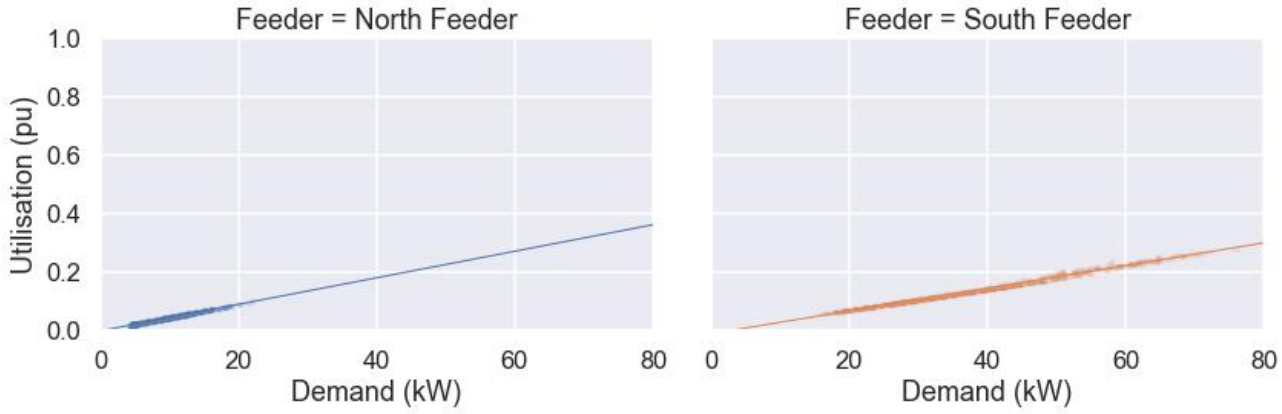
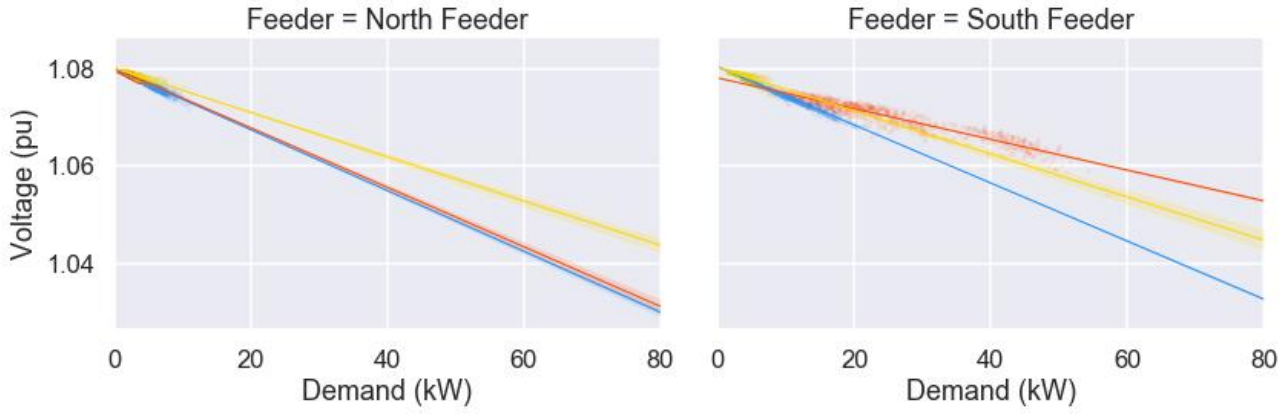


Figure 4-33: Regressions for phase voltages of notes at the end of Sinderby feeders



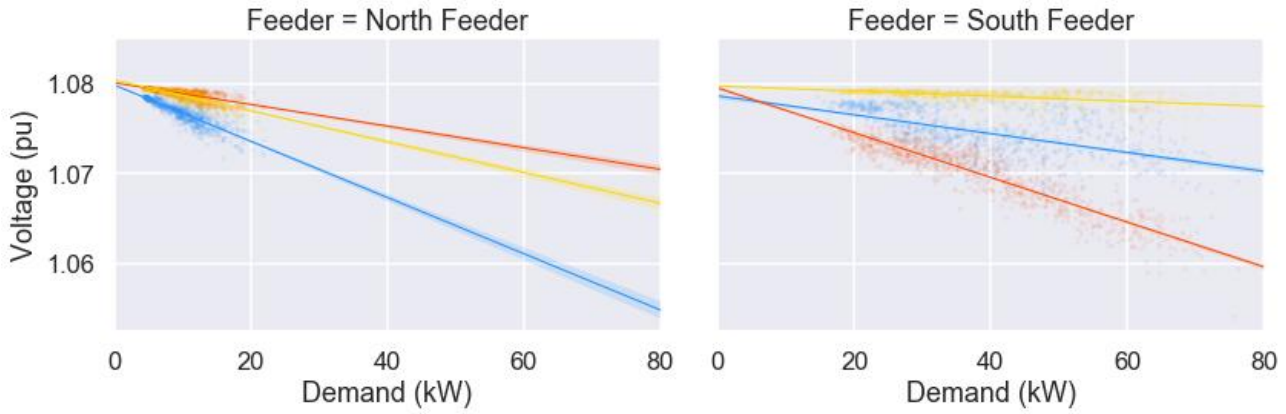
[34] When assessing cases where there is embedded generation, it would also be necessary to consider low demand and high generation trials as part of this analysis.

Table 4-8: Regression results for two Sinderby feeders

Feeder	Constraint	Phase	Branch/node	Slope	Intercept	R^2
North	Thermal	Three-phase	141107475_1	0.0045	-0.0001	1.0000
North	Voltage	Red	137956470	-0.0006	1.0800	0.8463
North	Voltage	Yellow		-0.0005	1.0801	0.8189
North	Voltage	Blue		-0.0006	1.0799	0.9432
South	Thermal	Three-phase	141107463_1	0.0039	-0.0098	0.9947
South	Voltage	Red	138026433	-0.0003	1.0780	0.9104
South	Voltage	Yellow		-0.0004	1.0801	0.7397
South	Voltage	Blue		-0.0006	1.0803	0.9839

For Cranwood, it was possible to find equations to describe the single-phase voltages which depended on the total demand across all of the phases on the feeder. However, for Sinderby, this resulted in a very low R^2 scores. We have not presented all of these here, but as an example, for the yellow phase voltage on the south feeder the R^2 was only 0.23, whereas it is 0.74 when using separate demand models as shown in the table. This is likely due to the significantly lower number of customers connected to the LV feeders in the Sinderby network, reducing diversity on each phase. This informs the future approach and how it might be best applied to a wide range of network types and topologies. This is illustrated in Figure 4-34.

Figure 4-34: Regressions for phase voltages of Sinderby feeders, against total feeder demand



As a result, Sinderby needs to have separate models for the demand on each phase, rather than a single model describing the total aggregate demand across all phases.

4.3.3 Probabilistic network condition

The probabilistic network condition was determined as it was for Cranwood in Section 4.2.3. However, because customer demand is so low relative to the ratings of the assets and the impedances of the branches, the thermal exceedance and voltage deceedance expectations are effectively zero in all cases, meaning there is no real risk of either voltage excursions or thermal overloads.

The 1-in-10-year utilisations and voltages are plotted in Figure 4-35 and Figure 4-36. These show that even for relatively extreme values of demand, utilisation is still well below 100% and voltages are still much

higher than 0.94 per unit. Nevertheless, this case study further demonstrates application of the novel analysis approach to assess LV network thermal loading and voltage.

Figure 4-35: One-in-ten-year utilisation

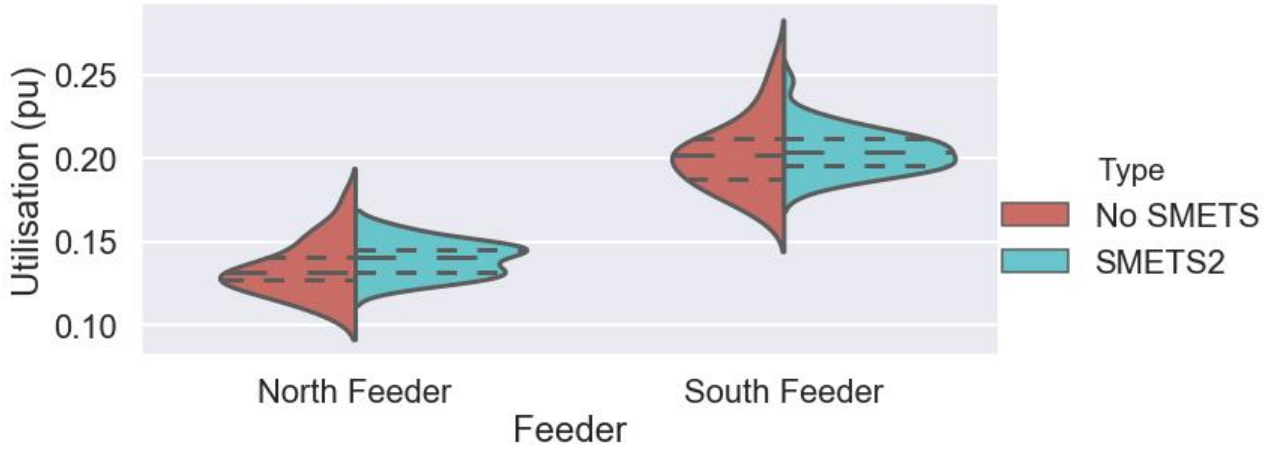
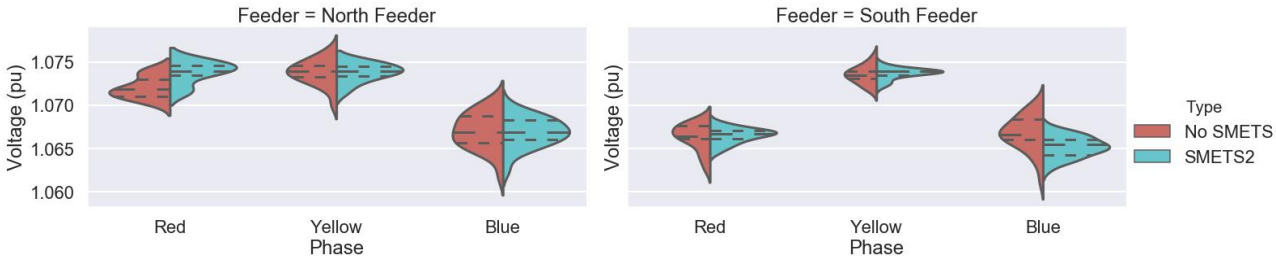


Figure 4-36: One-in-ten-year low voltage



5 Next steps

In this report, we have described the context for our novel analysis techniques for probabilistic assessment of low voltage networks, and developed and demonstrated the foundations of a methodology to achieve this. This method could now be adopted for implementation when analysing simple LV networks in the absence of significant volumes of LCTs. However, there are some areas which need to be considered in more detail for implementation, and further developments that would be required in the future to ensure that this method is robust enough to work with more complex networks and can be used to model LCTs. These potential future developments are described in this section.

5.1 Implementation

5.1.1 Acceptable levels of risk for low voltage networks

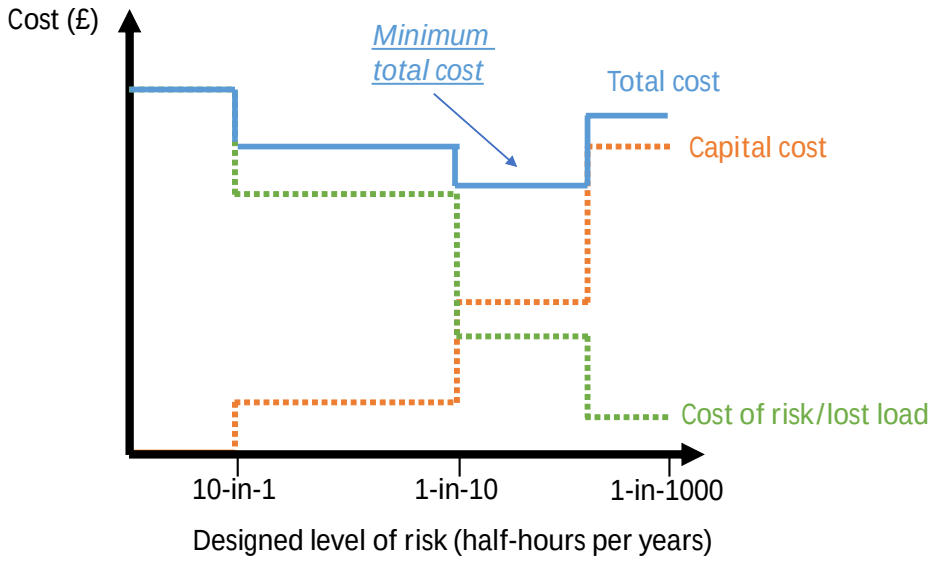
This report has focused on how to calculate probabilistic representations of network demand and how to use these to assess the risk of certain network conditions. We have calculated "exceedance expectations" in terms of the numbers of half-hours per year (or perhaps more meaningfully, expected rate or occurrence in years) for which network assets have voltages outside of their limits or cables and transformers exceed their thermal ratings.

However, we have not made any recommendation as to what sort of levels of exceedance expectation might prompt a DNO to invest in the network. However, we have, in places, put more emphasis on the 1-in-10-year demand, on the basis of ACE49 considering the 90th percentile of demand. For reference, planning of the gas transportation system, the licence requires that the system is designed for a 1-in-20-year demand, derived based on at least 50 years' data.

In principle, acceptable levels of risk should be determined on the basis of the trade-off in cost and benefit between network investment, and outage risk. Essentially, this corresponds to trading off the cost of investment with the cost of customer interruptions and customer minutes lost.

This is illustrated in Figure 5-1 – as the network is designed for more extreme levels of risk, the capital cost of meeting these requirements will be greater, as this will require higher rated assets or more extensive works. However, at the same time, the risk of exceeding these asset ratings will decrease, which means the cost of this risk (in terms of lost load) will decrease. The total minimum cost is therefore achieved by designing for a level-of-risk which minimises the total of these two costs. In this illustrative example, this is the 1-in-10-year demand, which strikes the optimal balance between capital cost and cost of risk.

Figure 5-1: Illustration of trade-off between capital cost and cost of risk



In principle, this could be determined on a case by case basis, accounting for specific local variation in the value of lost load. However, this is unlikely to be possible in practice, and it is more likely that generically appropriate levels of risk would have to be determined and defined within LV network planning policies.

5.1.2 Aggregation of smart-meter data

Aggregation of individual customers' smart meter data could have implications for the future implementation of the approach set out in the report. This could require some additional development beyond the scope of this project, depending on the final functional requirements of the modelling tool.

In the event that data is required at a lower level of aggregation than is allowed by Ofgem, e.g. if there are 2 isolated customers at the end of a long feeder cable, but the smallest level of aggregation for which DNOs are allowed access is e.g. 3, then the problem can be resolved by only calculating and extracting histograms of the aggregated demand of the two customers, as the raw data is not actually required to estimate parameter values to a reasonable approximation. Instead, a synthetic series could be sampled from the histogram, by assuming that all values within the range of each histogram box is equally likely.

Another potential option would be for the algorithms which produce the demand models to operate on the disaggregated data within the secure data store, so that NPg never has access to this data directly, but can still realise the possible benefits of disaggregated data when fitting the demand models. This could potentially have implications for the existing IT systems that will be used to store the smart meter data.

5.1.3 Partial penetration of smart meters

The process for Bayesian updating set out in this report assumes that the variable which is being statistically modelled – that is, demand on a feeder – is the same as the variable for which we have measurements. However, even once the smart-meter rollout is complete, this is unlikely to be the case for the vast majority of the networks, with an overall penetration of smart meters expected to be materiality less than 100%.

Therefore, it is necessary to consider how smart meter data can be used for updating the Bayesian model in this situation, where the measured variable and the modelled variable are not the same. A detailed mathematical exposition of this is set out in Appendix A.1.2. Essentially, this sets out the necessity of forming a statistical relationship between the parameters that characterise the distribution of the modelled aggregate demand variable of interest, and the parameters that characterise the distribution of the

aggregate demand variable for which measurements happen to be available. This could be done by fitting models for a group of N customers, then fitting the model for a group of $N - M$ customers (where the $N - M$ are part of the original group of N), and then finding the statistical relationships between these two sets of parameters.

5.2 Further developments

5.2.1 Accounting for low carbon technologies

In this report, we have described a model which finds probability distribution parameters for different numbers of customers, N. We have demonstrated this generically, without differentiating between different customer categories, although in practice this could be achieved using existing data.

In order to use this method to forecast future scenario demands, it would be necessary to also account for the future adoption of LCTs such as EVs, heat pumps and solar PV. We envisage that LCTs will just extend this to find distributions for groups of customers and LCTs, N_i, where i is an index over different types of technology (EVs, HPs, PV etc) and customers. For example, a group of domestic customers on a feeder with 40 customers, and a 20% penetration of EVs, would use the distribution hyperparameters for $N_{Domestic} = 40, N_{EV} = 8$. These distribution hyperparameters would have to be determined in a manner which robustly accounts for the interactions between different LCTs

It is unlikely to be efficient to determine unique parameters sets for all possible combinations of customer categories and LCTs, before those combinations happen to occur in an LV circuit. For example, for groups of up to 100 customers with up to 100 EVs and up to 100 HPs, there would be 100^3 = 1,000,000 possible sets of hyperparameters covering all of these different combinations.

Machine learning methods are likely to offer the best type of solutions to deal with this problem.. Hyperparameters could be determined by fitting statistical models for 1,000s of different combinations of customers and LCTs. This could be used as training data for a machine learning regression, which would enable a DNO to accurately predict the hyperparameters for combinations which have not yet been fit. For example, after fitting a model for 10 customers with 5 EVs, and 10 customers with 7 EVs, it is not unreasonable that we could accurately predict the model for 10 customers with 6 EVs.

There are some outstanding challenges associated with this approach which need to be considered, and are not yet entirely resolved. These would not prevent the development of initial versions of the tool, although it might lead to LCTs being more simply represented to begin with. These challenges are described below.

5.2.1.1 Limited LCT data

In many cases, there is surprisingly little information available in the public domain to provide a thorough understanding of LCT demand. From a brief review, LCT datasets often possess one or more undesirable features, including:

- Monitoring of a relatively short time period: In the case of the CLNR EV datasets, these were monitored for less than whole year and only included one winter.
- Limited customer numbers: CLNR includes around 100 EV profiles, which is far less than the 8,000 domestic customer profiles provided.
- Interventions: Many datasets for LCTs have been gathered in innovation projects, which have also been concerned with studying the impacts of various smart interventions on managing the impacts of these LCTs. These profiles will not give a good indication of the impact of LCTs, unmitigated, on electricity networks.
- Early adopters: There is a risk that existing LCT profiles only reflect the consumption patterns of early adopters, and that these could be very different to the patterns of other types of consumer.

In general, these factors provide compelling reasons for treating LCT demand with a Bayesian approach, due to the general limitation in data. However, the model might need to make assumptions in order to achieve this. For example, it may be necessary to assume that, for any given season and time-of-day, the probability distributions for underlying domestic demand and LCTs are completely independent. Some of the most promising machine learning methods currently being utilised in other sectors involve combining neural networks with Bayesian inferencing to form 'Bayesian deep learning', and such methods may well be the best choice for the current application.

5.2.1.2 Locational dependency

There is potentially a strong locational dependency for certain types of LCT, in particular solar PV or other types of small scale weather-based generation. Demand and solar PV output are both strongly dependent on weather, and therefore there is likely to be a correlation between them. If this correlation isn't accounted for, then it is likely that the outputs of any modelling of PV would lead to inaccurate results.

It is also possible that, due to similarities in customer circumstances and behaviour, there would be strong correlations in the demand for other LCTs (such as EVs) for customers in a specific small local area, e.g. that might be served by a single LV network. To gain an understanding this would require datasets which measure EV demand for such a specific local area e.g. professional families with EVs served by a single LV feeder.

5.2.1.3 Uncertainty in LCT numbers and ratings

It is very likely that, particularly during earlier stages of adoption, DNOs will not know with certainty how many LCTs of certain types are present on a specific LV network. They certainly won't be able to say exactly how many LCTs will be on a specific network in the future. Therefore, it is possible that the model will have to reflect uncertainty in the numbers and types of LCTs, as well as uncertainty around the patterns of consumption for those LCTs. It might be possible to address this using a "mixture model", which is essentially a weighted sum of probability distributions. For uncertain numbers of LCTs, this might take a form similar to:

$$F(Demand) = \sum_i \phi_i \times F_i(Demand\,|\Theta)$$

In this formulation, ϕ_i is the probability that there are i LCTs, and $F_i(Demand\,|\Theta)$ is the probability distribution of demand for i LCTs, with uncertain parameter set Θ.

5.2.2 The multi-variable case

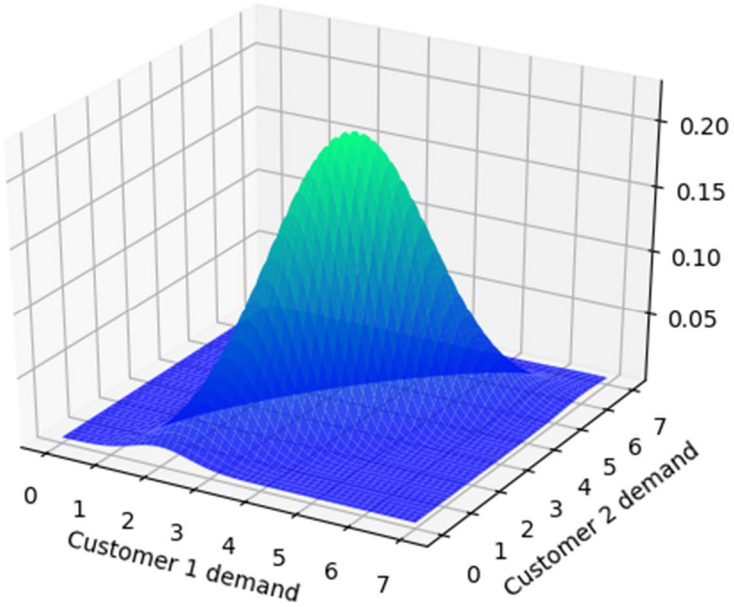
As described throughout this report, there are cases where there isn't a single aggregation demand variable which explains the utilisation of an asset, or the voltage at a node, to an acceptable degree of accuracy. For example, when modelling the voltage at the end of a feeder, it may be that this is best understood by splitting demand into two groups: demand closest to the transformer and demand furthest from the transformer. In most cases, simple multivariate linear or quadratic equations should be sufficient to understand this, for example, for the case described above:

$$V = \beta_0 + \beta_1 \cdot D_1 + \beta_2 \cdot D_2$$

It would be relatively easy to extend this to more than two groups of aggregated demand.

Multiple groups of aggregate demand present no new challenges for the modelling of the network condition. However, it then becomes necessary to consider statistical dependence between demand D_1 for customer group 1 and the demand D_2 for customer group 2. This requires the use of multi-variate statistics, which significantly increases the complexity of the probabilistic modelling. Instead of single-variable probability distributions, it is necessary to work with multi-variate distributions. The simplest possible multivariate distribution – the joint-Normal distribution, is illustrated in Figure 5-2.

Figure 5-2: Example of multi-variate probability distribution



Fitting and assessing multivariate probability distributions is complex. One common approach is to fit separate "marginal" or univariate distributions for the two variables of interest (e.g. demand D_1 for customer group 1 and the demand D_2 for customer group 2), and then join then using a "copula" function, which are functions that define the dependence between random variables. A common and reasonably simple option is the Gaussian copula, which requires the probability distributions for the two variables to be determined separately, as well as their linear correlation. Other, more complex, copula functions could also be used. Further work would be needed to develop the multi-variate evolution of the model.

Appendix A – Mathematical formulation

This section sets out a detailed mathematical exposition of the statistical model proposed by TNEI for the aggregated electrical demand of a group of customers on an LV network at a given time, starting with a general introduction to the theory of Bayesian inference. The document will also cover how the model for individual short periods translate into the statistics of extremes over long periods.

A.1 Mathematical presentation of Bayesian inference

In this section, we set out the principles of Bayesian inferencing in a more precise and detailed mathematical way than in the main body of the report.

A.1.1 Bayesian updating based on direct observations

In this subsection, we present the process of updating our parameter distributions, and ultimately predictions about observable quantities, when direct observations of those variables become available. This contrasts with the case presented in the next subsection, where the observations are of a different, but statistically related quantity.

We begin with the following definitions:

- X, an observable random variable, taking values x. This may in fact be a vector of values, e.g. the aggregate demand at multiple nodes on a circuit.
- Θ, the parameter set of the data point's probability distribution, i.e. $X \sim p(x|\Theta)$. This is typically a vector of parameters, e.g. mean and variance for normal distributions. The probability distribution here is very general, and could be discrete, continuous or a combination of both. As this is a Bayesian model, Θ is itself random, taking values θ.
- α, the (deterministic) hyper-parameter set of the uncertain parameter's distribution, i.e. $\Theta \sim p(\theta|\alpha)$. Again, this is likely to be a vector of hyper-parameters.
- $\underline{\mathbf{X}}$ is a sample of observations of X, i.e. a set of n observed data points, i.e., $x_1, \dots, x_n$.

Bayesian updating works as follows:

- The prior distribution is the distribution of the parameter(s) before any data is observed, i.e. $p(\theta|\alpha)$. This tends to represent our view of the world in the general case, e.g. that coins are likely to be essentially fair, prior to a coin tossing experiment.
- The sampling distribution is the distribution of the observed data conditional on its parameters, i.e. $p(\underline{\mathbf{X}}|\theta)$. This is the probability of the sample $\underline{\mathbf{X}}$ occurring, given that their distribution is characterised by the parameter set Θ taking the particular values θ. It is convenient to view this as a function of θ rather than $\underline{\mathbf{X}}$, since the latter is fixed and known, in which case the quantity is termed the likelihood $L(\theta|\underline{\mathbf{X}}) = p(\underline{\mathbf{X}}|\theta)$.
- The marginal likelihood (sometimes also termed the *evidence*) is the distribution of the observed data marginalized over the parameter(s), i.e. $p(\underline{\mathbf{X}}|\alpha) = \int p(\underline{\mathbf{X}}|\theta)\, p(\theta|\alpha) \mathrm{d}\theta$
- The posterior distribution is the distribution of the parameter(s) after considering the observed data. This is determined by Bayes' rule, which forms the heart of Bayesian inference:

$$p(\theta|\underline{\boldsymbol{X}}, \alpha) = \frac{p(\underline{\mathbf{X}}|\theta) \cdot p(\theta|\alpha)}{p(\underline{\mathbf{X}}|\alpha)}$$

- The posterior distribution could also be defined in terms of a new set of hyper-parameters:

$$p(\theta|\underline{\mathbf{X}}, \alpha) = p(\theta|\alpha_{New})$$

  Essentially, by incorporating new evidence through Bayesian inference, the hyper-parameters of the prior distribution change, resulting in reduced uncertainty.

Note these are the prior and posterior distributions of the *parameters* which define the probability distribution of the observable random variable, x, rather than the observable variable itself. The probability distributions for a realised value x, or more generally a vector of observations $\bar{x}$ (that are yet to occur) are known as predictive prior and predictive posterior distributions. These allow statements of the form "there is a 60% probability that $\bar{x}$ will be between 1 and 2", or "the expected values of my prediction for $\bar{x}$ is [1.7, 2.3]". Forecasts that take the form of probability distributions rather than single points are in fact necessary to make mathematically optimal decisions under uncertainty. The prior predictive distribution is the best prediction for $\bar{x}$ that can be made before the sample $\boldsymbol{X}$ has become available, given by:

$$p(\bar{x}|\,\alpha) = \int p(\bar{x}|\theta)\, p(\theta|\alpha)\mathrm{d}\theta$$

The posterior predictive distribution is the best forecast that can be made after $\underline{\mathbf{X}}$ becomes available:

$$p(\bar{x}|\,\underline{\mathbf{X}},\alpha) = \int p(\bar{x}|\theta)\, p(\theta|\underline{\mathbf{X}},\alpha)\mathrm{d}\theta = \int p(\bar{x}|\theta)\, p(\theta|\alpha_{\mathrm{New}})\mathrm{d}\theta$$

These predictive distributions are optimal since they are *marginalised* over the prior and posterior parameter distribution, respectively. That is, averaging across all of the uncertainty associated with the parameter distributions is conducted correctly.

For convenience, these distributions may be expressed as a series of quantiles, e.g. the 10th, 20th, ..., 90th percentiles, or as the expected value plus an indication of uncertainty, e.g. the standard deviation, or the 5th and 95th percentiles, giving a 90% interval. Note that these are known as prediction intervals, rather than confidence intervals.

A.1.2 Bayesian updating based on indirect observations

In this subsection, we present the full mathematical formulation of Bayesian updating for a random variable Y, when observations become available for some other variable X that has a statistical relationship to Y. In this extended situation, we have the following definitions:

- y, a data point for the (theoretically) observable random variable Y, the variable in which we are ultimately interested. This may in fact be a vector of values, e.g. the aggregate demand at multiple nodes on a circuit.
- φ, a realised value of random variable Φ, which is the uncertain parameter set of the distribution of y, i.e. $Y \sim p(y|\varphi)$. This is typically a vector of parameters, e.g. mean and variance for normal distributions. The probability distribution here is very general, and could be discrete, continuous or a combination of both.
- α_y, the (deterministic) hyper-parameter set of Φ's distribution, i.e. $\Phi \sim p(\varphi|\alpha_y)$. it is likely to be a vector of hyper-parameters.
- x, a data point for the observable random variable X. This is not the variable in which we're ultimately interested, but is relevant because we have a set of observations for it, to use in updating. Again, this may be a vector of values, e.g. the aggregated smart meter consumption record for a number of customer subsets from the same network (where not all customers have smart meters).
- $\underline{\boldsymbol{X}}$, the sample of observations available, i.e. a set of n observed data points, i.e., $x_1, \ldots, x_n$.
- θ, a realised value of the random variable Θ, which is the uncertain parameter set of the probability distribution of x, i.e. $X \sim p(x|\theta)$. This is typically a vector of parameters, and again the probability distribution here is very general.
- α_x, the hyper-parameter set of the uncertain parameter Θ's distribution, i.e. $\Theta \sim p(\theta|\alpha_x)$. Again, this is likely to be a vector of hyper-parameters.

Bayesian inference with indirect updating works as follows:

- We have prior distributions for the parameter sets φ and θ: $\Phi \sim p(\varphi|\alpha_y)$, $\Theta \sim p(\theta|\alpha_x)$, representing our view of the world in the absence of specific data.
- The sampling distribution for the observations $\underline{\boldsymbol{X}}$, i.e. the distribution of the observed data conditional on the parameter set θ, i.e. $p(\underline{\mathbf{X}}|\theta)$. This is the probability of the sample $\underline{\mathbf{X}}$ occurring, given that their distribution is characterised by θ. It is convenient to view this as a function of θ rather than $\underline{\mathbf{X}}$, in which case it is termed the likelihood $\mathrm{L}(\theta|\underline{\mathbf{X}}) = p(\underline{\mathbf{X}}|\theta)$.
- The marginal likelihood of $\underline{\mathbf{X}}$ is the distribution of the observed data marginalized over the parameter set, i.e. $p(\underline{\boldsymbol{X}}|\alpha_x) = \int L(\theta|\underline{\boldsymbol{X}}) \cdot p(\theta|\alpha_x)\mathrm{d}\theta$.
- The posterior distribution of θ is the distribution of the parameter(s) characterising x after considering the observed data. This is determined by Bayes' rule as:

$$p(\theta|\underline{\mathbf{X}}, \alpha_x) = (L(\theta|\underline{\boldsymbol{X}}) \cdot p(\theta|\alpha_x))/(p(\underline{\boldsymbol{X}}|\alpha_x))$$

- In order to understand what $\underline{\boldsymbol{X}}$ tells us about the distribution of Φ and ultimately predictive distributions for $\bar{y}$ – a set of future or unknown observations of Y, we need to understand the statistical relationship between X and Y. We may represent this relationship through the conditioning effect of Φ on Θ, i.e. $\Phi \sim p(\varphi \mid \theta, \alpha_y)$.
- Given this relationship, the updating effect of $\underline{\mathbf{X}}$ on Φ may be expressed as:

$$p(\varphi \mid \underline{\boldsymbol{X}}, \alpha_x, \alpha_y) = \int p(\varphi \mid \theta, \alpha_y) \cdot p(\theta \mid \underline{\mathbf{X}}, \alpha_x)\, \mathrm{d}\theta,$$

$$p(\varphi \mid \underline{\mathbf{X}}, \alpha_x, \alpha_y) = \int p(\varphi|\theta, \alpha_y) \cdot L(\theta|\underline{\boldsymbol{X}}) \cdot p(\theta|\alpha_x)\, \mathrm{d}\theta \Big/ \int L(\theta|\underline{\mathbf{X}}) \cdot p(\theta|\alpha_x)\, \mathrm{d}\theta$$

- The predictive posterior distribution for $\bar{y}$ (is the best forecast that can be made after $\underline{\mathbf{X}}$ becomes available) is given by:

$$p(\bar{y}|\, \underline{\mathbf{X}}, \alpha_x, \alpha_y) = \int p(\bar{y}\,|\varphi) \cdot p(\varphi \mid \underline{\mathbf{X}}, \alpha_x, \alpha_y)\, \mathrm{d}\varphi$$

$$p(\bar{y}|\underline{\mathbf{X}}, \alpha_x, \alpha_y) = \iint p(\bar{y}\,|\varphi) \cdot p(\varphi|\theta, \alpha_y) \cdot L(\theta\,|\underline{\boldsymbol{X}}) \cdot p(\theta|\alpha_x)\, \mathrm{d}\theta \mathrm{d}\varphi \Big/ \int L(\theta|\underline{\mathbf{X}}) \cdot p(\theta|\alpha_x)\, \mathrm{d}\theta$$

A.2 Statistical model of aggregated demand

A.2.2.1 Demand as a discrete-time random process

We adopt a discrete-time framework of 30-minute intervals, which means that we treat the power demand during these periods as a single value. We initially index the 30-minute intervals by t, and model the demand at time t as the continuous random variable D_t. The sequence of demands over consecutive intervals form the random process $\{\dots, D_{t-1}, D_t, D_{t+1}, \dots\}$ and, initially, we do not assume stationarity on this process. That is, we initially allow the probability distribution for each time step to be distinct, and characterised by a probability density function (PDF) written as $f_t(d)$. This very general case will later be simplified by the introduction of identical PDFs for all time steps with the same combination of time-of-day and season.

Unless otherwise stated, an uppercase letter indicates a random variable, while lowercase letters indicate either realised values of those variables, or other deterministic quantities such as probability distribution parameters. Times of day are expressed as e.g. 00:00 or 00:30, and these examples refer to the periods 00:00 – 00:29 and 00:30 – 00:59, respectively.

A.2.2 Parametric distribution selection

It has been established through exploration of the TC1a dataset of domestic demand collected within the CLNR project that Gamma and 3-parameter-Weibull distributions are suitable parametric families for

capturing distributions of demand. This was found to be true where the number of customers contributing to the aggregate demand ranges from 1 to 100s, which is the essential range for LV networks. Each D_t in our model is assigned either a Gamma or a 3-parameter-Weibull distribution (henceforth referred to as Weibull, for convenience), with the choice between them depending on the time-of-day and season of t, as is explained in the next section.

The form of the Gamma distribution's PDF for the demand at time t is:

$$f_t(d; k_t, \theta_t) = \frac{d^{k_t - 1} \cdot e^{-d/\theta_t}}{\theta^{k_t} \cdot \Gamma(k_t)}$$

where d is a level of demand, k_t is the shape parameter for time t, θ_t is the scale parameter for time t and $\Gamma()$ is the Gamma function – an extension of the factorial function.

The form of the Weibull distribution's PDF for the demand at time t is:

$$f_t(d; k_t, \theta_t, \zeta_t) = \frac{k_t}{\theta_t} \cdot \left(\frac{d - \zeta_t}{\theta_t}\right)^{k_t - 1} \cdot e^{-\left(\frac{d-\zeta_t}{\theta_t}\right)^{k_t}}$$

where d, k_t and θ_t have the same meaning as above, and where ζ_t is an additional location or 'shift' parameter for the Weibull distribution at time t. We use lower case symbols for parameters here as we are not explicitly considering the Bayesian formulation of this model.

A.2.3 Seasonalities and a time-collapsed model

Although customer demand is random, it is also clearly periodic – both across the hours of the day, and across seasons. For this reason, our model assumes a distinct probability distribution for each unique combination of time-of-day and season, and assumes that each time step t with these combinations are identically distributed. This is a simplification of the general case stated above that each time step t within the random process could potentially have its own distinct distribution, characterised by the PDF $f_t(d)$. For clarity we adopt a new notation where i indexes the time-of-day, starting with 00:00, and s indexes the season, starting with winter. So, the full set of distinct probability distributions is characterised by the PDFs $f_{i,s}(d)$, where $i = 1, 2, \dots, 48$ and $s = 1, \dots, 4$, and e.g. $f_{1,1}(d)$ is the PDF for demands for 00:00 in winter.

We define the 4 seasons of the year as winter – Dec to Feb, spring – Mar to May and so on, so that there is a total of 192 unique distributions. There are, therefore, 90 repetitions per year for each of the 48 winter distributions (ignoring leap years), rising to 92 repetitions per year for each spring and summer distribution, and 91 time-steps for autumn distributions. For convenience, we use the letter r to represent the number of distinct distributions, i.e. $r = 192$. We also adopt $q_{i,s}$ to represent the number of time-steps with the PDF $f_{i,s}(d)$, and as stated above these range from 90 – 92, depending on s.

We adopt a 'time-collapsed' model, where we are not concerned with the statistical relationship between consecutive time-steps (except for the process of simplifying parameter estimation, as will be covered in another section below). This does not compromise our modelling in any way, as long as we limit ourselves to the expected values of any derived random variables. We also use τ to index the repetitions for each unique distribution, so that our original concept of demands forming the non-stationary random process $\{\dots, D_{t-1}, D_t, D_{t+1}, \dots\}$ having PDFs $\{\dots, f_{t-1}(d), f_t(d), f_{t+1}(d), \dots\}$ has changed to thinking about a set of r stationary random processes $\{\dots, D_{i,s,\tau-1}, D_{i,s,\tau}, D_{i,s,\tau+1}, \dots\}$ with PDFs $f_{i,s}(d)$. Further, one year of observed demand values are seen, in this framework, as a set of samples from each $f_{i,s}(d)$, with $q_{i,s}$ trials in each sample.

Given this reformulation, the parameters of our Gamma and Weibull distributions become $k_{i,s}, \theta_{i,s}$ for both, and $\zeta_{i,s}$ for Weibull.

A.2.4 Exceedance Expectations

We are generally interested in the probability that the random demand, D_t, as some time-step t exceeds some level, d. In order to calculate this, we must know the relevant probability distribution, and therefore must express t as the set of indices i, s, τ. The probability of the demand $D_{i,s,\tau}$ – i.e. the demand on instance τ of the time-of-day and season combination i, s – exceeding the value d can be written as $\mathbf{P}_{i,s}(D_\tau \geq d)$, and is given by $\mathbf{P}_{i,s}(D_\tau \geq d) = 1 - \mathrm{F}_{i,s}(d)$, where $\mathrm{F}_{i,s}(d)$ is the cumulative distribution function (CDF) associated with $f_{i,s}(d)$. These are easily evaluated for the chosen Gamma and Weibull distributions described above.

We are interested in the exceedance expectation for this level of demand, i.e. the expected number of time-steps in a year where this level of demand will be exceeded. It must be noted that the number of occurrences within individual years may deviate considerably from this average. Taking advantage of the linearity of the expectation operator, the result is simply given by:

$$Annual\ Exceedance\ Expectation\ of\ d = \sum_{s=1}^{4} \sum_{i=1}^{48} q_{i,s} \cdot \left(1 - \mathrm{F}_{i,s}(d)\right).$$

For the peak demands of interest to network design, only a small subset of the i, s combinations will make a significant contribution to this summation.

For LV circuits with significant presence of distributed generation, the net demand can take negative values. Similar principles can be applied to calculate the number of times these negative extremes will be exceeded on average, by simply replacing $(1 - \mathrm{F}_{i,s}(d))$ with $\mathrm{F}_{i,s}(d)$.

A.2.5 Bayesian Formulation of the Demand Model

In this section, we extend our demand model of demand to a Bayesian version. The main difference is that the distribution parameters become random variables, i.e. $K_{i,s}$ taking realised values $k_{i,s}$, $\Theta_{i,s}$ taking realised values $\theta_{i,s}$, and $Z_{i,s}$ taking realised values $\zeta_{i,s}$.

The model is obviously considerably more complicated, as each of these random parameters have their own distributions, characterised by sets of hyper-parameters, which we write as $\alpha_{i,s}^{K}$, $\alpha_{i,s}^{\Theta}$, and $\alpha_{i,s}^{Z}$ respectively. Model building now involves choosing suitable parametric families for these distributions, as well as hyper-parameter values for the prior distributions. One obvious and sensible approach is to sample aggregate series from the TC1a dataset and fit optimal parameters to each one, using a standard method from frequentist (i.e. non-Bayesian) statistics. The standard method is to find maximum likelihood estimates (MLEs), i.e. the parameter set that maximises the likelihood function of the data. The prior parameter distributions can then be fitted, again as MLEs, to the distribution of parameter values obtained from the sampled series.

A further complication might be a need to consider the statistical relationship between the various random parameters. A common way of representing such relationships is with Gaussian copulas – i.e. the assumption that if the individual (marginal) distributions are transformed to be Normal (also known as Gaussian), then the together they will form a joint-Normal distribution. This has the very convenient quality that their statistical relationship is entirely captured by linear correlations. In this case, the set of hyper-parameters would be extended from vectors to covariance matrices, capturing the correlations between each pair of parameters, along with their individual variances.

When considering any instance of a group of customers, the prior distributions can be updated as outlined in Section A.1 whenever data specific to those customers becomes available – almost certainly indirect observations in our case.

A.2.6 A Method for Reducing the Number of Parameters

One major challenge with the method presented thus far is the very large number of parameters involved: there are 192 distributions, each with either 2 or 3 parameters, and each of those parameters has at least a set of 2 or 3 hyper-parameters, if not rather large covariance matrices. Another problem is that when fitting the model to a finite dataset, the number of observations associated with any i, s pair may be relatively small. For the TC1a dataset, spanning 2 winters and 3 summers, the number of observations associated with the i, s pairs vary between 180 and 270. It is therefore wise to impose some additional conditions on the distributions in order to reduce the parameter numbers, and temporarily reduce the number of distinct distributions. This is achieved by first dividing the sequence of 48 times-of-day, for each season, into 3 or 4 subsets of consecutive i – labelled $S_{j,s}$. The principles behind the optimal segmentation choices are that:

1. they should cluster similar mean values as much as possible.
2. one of the following relationships should be approximately true within the sequence:
   i. the variability in the mean and the standard deviation of demands across the sequence are roughly proportional.
   ii. there is significant variability in mean across the sequence, but the variability in standard deviation is negligible.
   iii. there is significant variability in both mean and standard deviation across the sequence, but there is no simple relationship between their patterns.

For simplicity, when condition (iii) holds, we act as though condition (ii) is in fact true, so the 1st simplifying assumption we make is that either condition (i) or (ii) is true within sequences. This means sacrificing the quality of standard deviation modelling, but is seen as a necessary compromise to overcome the stated problems.

The next step of the parameter reduction process makes use of the useful characteristics of the mean and standard deviation of Gamma and 3-parameter-Weibull distributions, i.e. what happens when you transform Gamma-distributed variables with a constant multiplicative factor, and transform Weibull-distributed variables by adding a constant.

The mean of a Gamma distribution with parameters k_t, θ_t is given by $\mu_t = k_t \cdot \theta_t$ and standard deviation by $\sigma_t = \sqrt{k_t} \cdot \theta_t$, which means that the multiplication or division of all demand values in a historic series by some constant factor will change θ_t by that same amount, without affecting k_t. Consequently, the 2nd simplifying assumption we make is that within a sequence where condition (i) is true, each distribution is Gamma, and they share a common shape parameter, i.e. $K_{i,s} = K_{j,s}$ for all $i \in S_{j,s}$. As a result, we can find a set of values $\Lambda_{i,s}$ for $i \in S_{j,s}$ such that multiplication of the random demands by the relevant value makes all mean and standard deviation values within the sequence equal. The upper case Λ is used, due to being in a Bayesian framework. This means that all transformed demand values within the sequence are identically distributed, with parameters $K_{j,s}, \Theta_{j,s}$, and we can fit these parameters to a dataset that's typically 10 – 20 times bigger than the original, separate ones. The original demand series can be easily restored, as they have the parameters $K_{j,s}, \Theta_{j,s}/\Lambda_{i,s}$.

In the case of Weibull distributions with parameters k_t, θ_t, ζ_t the mean and standard deviation in this case are given by:

$$\mu_t = \zeta_t + \theta_t \cdot \Gamma\left(\frac{1}{k_t} + 1\right), \qquad \sigma_t = \theta_t \cdot \sqrt{\Gamma\left(\frac{2}{k_t} + 1\right) - \Gamma\left(\frac{1}{k_t} + 1\right)^2}.$$

This means that adding or subtracting some fixed constant to all the data in a historical series will change ζ_t, and therefore the mean, by that amount – while having no effect on k_t and θ_t, and therefore the standard deviation. Therefore, the 3rd simplifying assumption we make, following a similar process to that

above, is that for sequences where condition (ii) is true, each distribution is 3-parameter-Weibull and share both shape and scale parameters i.e. $K_{i,s} = K_{j,s}$ and $\Theta_{i,s} = \Theta_{j,s}$ for all $i \epsilon S_{j,s}$. As a result, we can find a set of values $\Upsilon_{i,s}$ for $i \epsilon S_{j,s}$ such that addition of the relevant values to the random demands make both the mean and standard deviation values within the sequence equal. This means that all transformed demand values within the sequence are identically distributed, with parameters $K_{j,s}, \Theta_{j,s}, Z_{j,s}$. The original demand series can again be easily restored, as they have the parameters $K_{j,s}, \Theta_{j,s}, Z_{j,s} - \Upsilon_{i,s}$.

Naturally, all random parameters have an associated set of hyper-parameters, and in the case of $\Lambda_{i,s}$ and $\Upsilon_{i,s}$, covariance matrices will almost certainly be required for accurate modelling. These will all require parametric families to be chosen, as well as prior values for the hyper-parameters. This can again be achieved through sampling aggregate series from the TC1a dataset and MLE parameter fitting.

Capgemini has tested this model, including these simplifications on CLNR data in their SMA3 report[35].

They found that winter demand could be decomposed into three sequences:

- Between 07:30 and 15:00, a 3-parameter Weibull distribution is used
- Between 15:30 and 22:00, a Gamma distribution is used
- Between 22:30 and 07:00, a Gamma distribution is used.

For winter, i.e. $s = 1$, this model would therefore have 55 parameters:

- 2 values each of the Gamma shape and scale factors (corresponding to 2 j-values)
- 1 value each of the Weibull shape, scale and shift parameters (corresponding to 1 j-value)
- 32 values for the multiplicative factors, combining the number of i-values in the two Gamma sequences
- 16 values for the additive factors, representing the number of i-values in the Weibull sequence

If the parameters are all assumed to be normally distributed, as Capgemini's analysis suggests, then the model has 110 hyper-parameters – a mean, and a standard deviation for the normal distributions of each of these 55 parameters. This is comparable to the number of parameters currently used within the ACE49 model, which takes 48 values of p and 48 values of q. However, more hyper-parameters may be necessary if is found that the multiplicative and additive factors display significant correlations, as discussed above.

A.3 Customer numbers, customer types, LCT demand and generation

One major omission to the model presented so far is that it does not make explicit the dependency on the number of customers, n. Of course, such a dependency must exist even for the simplest possible model, and is best expressed by making the hyper-parameters vary with n, e.g. $\alpha^{K}_{i,s,}(n)$. We consider first the case where n is known (i.e. deterministic).

The next level of complexity is, if customers are divided into types, e.g. types A, B and C. The types could be the MOSAIC categories used in the CLNR project as explored in this project, or any other system based on the type of property, or the socio-economic features of the property's area. In this case, the hyper-parameters would change with the number of customers of each type - n_A, n_B, n_C, to that the hyper-parameters would be expressed as e.g. $\alpha^{K}_{i,s}(n_A, n_B, n_C)$.

The further level of complexity is to introduce different numbers of LCT demands and generation capacities. In this case, the hyper-parameters would change with the numbers of customers of each type and the numbers of LCTs present, e.g. $\alpha^{K}_{i,s}(n_A, n_B, n_C, n_{EV}, n_{HP}, n_{PV})$, where n_{EV} is the number of electric vehicles present, n_{HP} is the number of heat pumps and n_{PV} is the number of photovoltaic installations.

[35] Workstream 4 Output – Smart Meter Data Analytics Final Report

We assume here that the Gaussian and 3-parameter Weibull distributions are, between them, sufficiently flexible to accommodate at least a modest amount of LCTs simply with a change of parameters – the only caveat being that when PV is present in significant amounts, the distribution must have a shift-parameter that allows for the minimum supported values to be negative. Perhaps a point might eventually be reached in the future where the nature of demand has changed so significantly that Gamma and Weibull distributions cannot adapt to provide a good fit, but this is not a current concern.

For any combination of domestic demands of various types along with LCTs, the model can be constructed 'from scratch', by establishing prior values for hyper-parameters by sampling from TC1a along with other datasets – and pertinent details of those other datasets are discussed in the following sections. However, it should be possible to avoid this full process by 'learning' the relationship between the set of inputs $n_A, n_B, n_C, n_{EV}, n_{HP}, n_{PV}$ and the set of prior hyper-parameters, through neural network methods. Indeed, the relevance of neural network methods to this problem has recently increased, with surge of interest and available tools that combine Bayesian statistics with neural networks to form the new field of deep Bayesian networks.

Appendix B – Scripts

Python scripts for network modelling:

1. 'Spatial_Importer.py' – this is the main script which calls all the other modules. The script has a Boolean flag 'bRunLF', if set to TRUE the script calls the module 'ProfileBasedLF.py' to run the load flows. The user also has the option of specifying if the load flow runs will be on a balanced or an unbalanced network.
2. 'ProfileBasedLF.py' – this script reads in the demand values for all the customers in a network from a csv file 'load_matrix.csv'. It then loops through each demand scenario or sample from CLNR data, runs a load flow and stores the line flow results (kW), the nodal voltages (pu) and the demand values (kW) in a variable (dictionary). At the end of all the load flow runs, the results variable is exported to three different csv files 'BusVolt_Scale_1.csv', 'LinekW_Scale_1.csv' and 'LoadkW_Scale_1.csv'
3. 'network_parameterisation.py' – this script runs the simple regression analysis on the power flow modelling, aggregating network demands and then regressing voltage and utilisation against these.

Python and R scripts for demand modelling

A total of 89 R-scripts were written in the process of processing the CLNR datasets, conducting data analysis and demonstrating the model. With the majority of these reliant on multiple data files, it is not practical to present all of them, rather a selection of the most essential. Further, most of the data files are large, with several being multiple gigabytes in size. For this reason, the data files are available on request, rather than automatically shared.

1. 'tc1a_rearrange.R' – transforms the raw TC1a dataset of domestic customers from the CLNR project into a more compact and logical form, adhering to the 'tidy data' principle that data that's ready to be analysed using a tool such as R should comprised of a table where each column is a unique variable and each row is a unique observation. Requires the (very large) data file 'TrialMonitoringData_6.csv'.
2. 'tc1b_rearrange.R' – does the same as the file above, but for the TC1b dataset of SME customers, and requiring the data file 'TrialMonitoringData.csv'. Similar scripts were developed for each 'test cell' dataset from the CLNR project.
3. 'main.py', 'scenario_creator.py', 'definitions.py', 'data_assembler.py' – together, these four scripts are used to sample data from the CLNR TC1a and TC1b datasets, allowing different combinations of customers (scenarios) to be flexibly defined.
4. 'mle_fit.py' – this script finds the best fitting Gamma and Weibull distributions, in line with model described in this report, for CLNR data sampled from the previous four scripts. These are fit on the basis of "maximum likelihood estimation", hence the name "mle_fit".
5. 'exceedance_expectations.py' – this script evaluates and aggregates the multiple seasonal and time-of-day probability distributions in order to determine the exceedance expectations for demand.
6. 'graphs_cranwood.py' – this script was used to calculate the exceedance expectations for thermal utilisation and voltage, based on the network parameterisation and demand modelling. A similar script was also prepared for the Sinderby case study.

Appendix C – CLNR Data Quality

The following analysis was initially conducted on the TC1a data, to examine its quality, and to establish some of the most salient aspects of domestic demand distributions:

Analysis of how much data is present in the series, i.e. not 'NA' values, as a function of time, and as a distribution across customers.

The analysis of availability across time is presented in figures A.C.1 and A.C.2, which are time-series plots of the proportion of customers reporting non-NA data for each 30-minute interval in the trial period. The series are presented separately for customers using two different smart meter brands: Logica (figure A.C.1) – which includes roughly 2/3 the customers, and Trilliant (A.C.2). Figure A.C.1 shows that the availability of data is generally quite high for customers with Logica devices, albeit with near constant customer attrition after an initial stable period, and which several very brief periods where either all or most of the data is NA, for unknown reasons. Figure A.C.2 shows that the proportion of non-NA values reported by customers with Trilliant devices was generally lower, with a reverse trend of general improvement over time, and a brief period of good quality toward the end of the period.

Figure A.C.1 A time-series plot of the proportion of TC1a customers with Logica Meters returning non-NA values

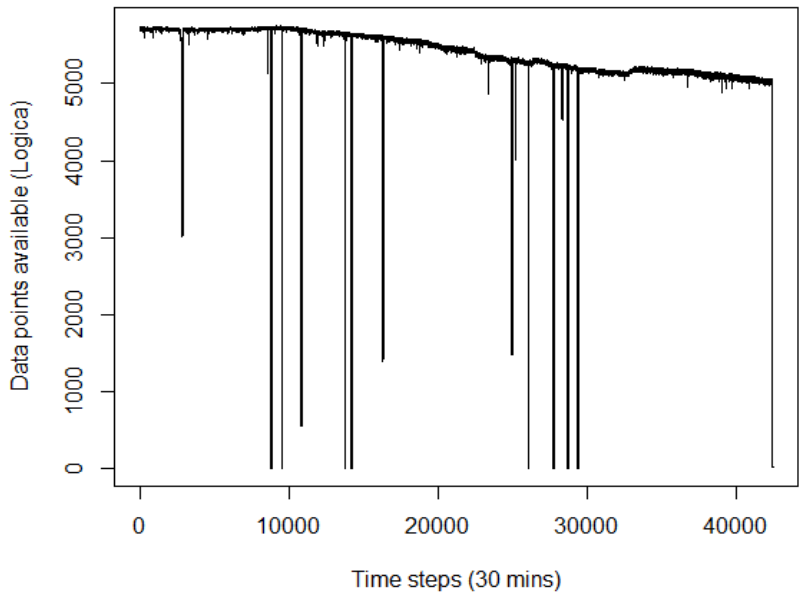


Figure A.C.2 A time-series plot of the proportion of TC1a customers with Trilliant meters returning non-NA values

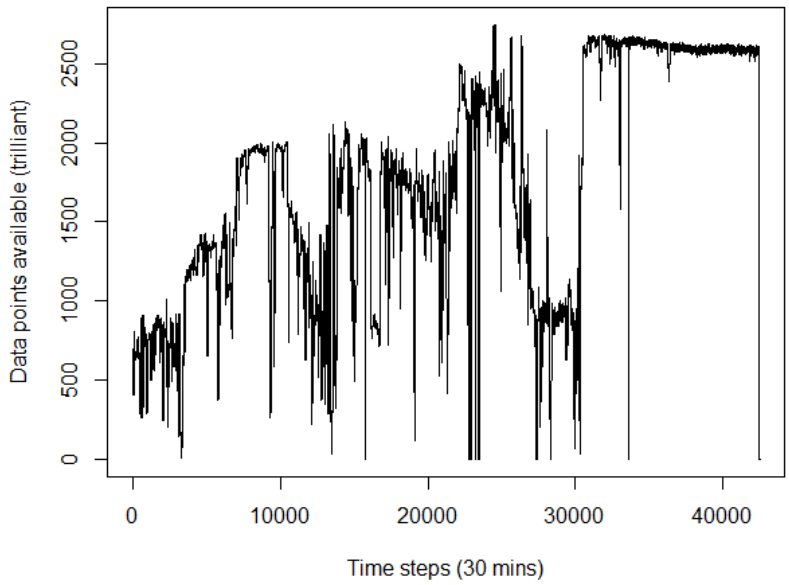


Figure A.C.3 A histogram of the proportion non-NA values returned by TC1a customers in the 'Alpha Territory' category across the trial period

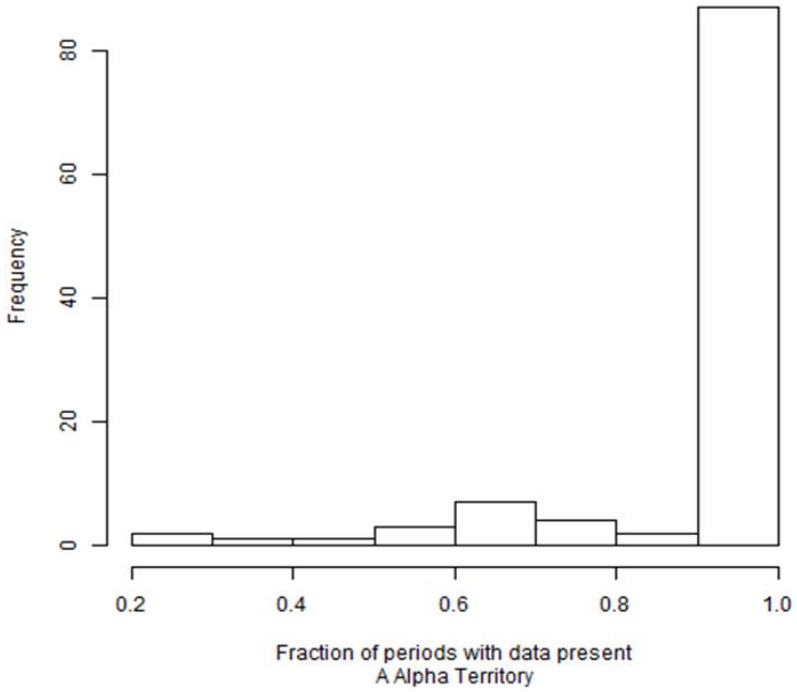


Figure A.C.3 moves on to examine the variability in the proportion of values returned by customers that were not NA, taken over the entire trial period. Customers with both types of smart meter were included, but for this particular plot, only customers in the 'Alpha Territory' category. The same plot was produced for all 15 MOSAIC types, with the results very similar and thus omitted. It can be seen that the proportions vary dramatically across the customers, ranging from poor values between 20-30%, to others between 90 – 100%. However, the latter interval is much more common than any of the others. Indeed, it was found that a total of 4000 customers have an availability (i.e. non-NA reported) value above 97%, with the number dropping fairly rapidly above that threshold. It was deemed that 4000 is a sufficient number of customers, and as such all customers with availability factors below this were excluded from the analysis.

Analysis of features in the data that can be identified as erroneous.

This part of the analysis investigated whether there were any patterns or features within some customers' series that are clearly erroneous, and could be systematically removed. Candidate features were long periods of zero consumption, periods of suspiciously constant demand, extremely high values and periods of flat negative demand. While evidence of all of these were discovered, close examination of the series did not yield any examples that were unquestionably erroneous. An example of a clearly valid time series segment for a single customer is presented in figure A.C.4 below. A relatively short period of NA was present (represented in the figure as a flat negative demand), but the remainder looks exactly as expected. Figure A.C.5 displays more suspicious-looking behaviour, given the combination of very small amplitude noise around 8kWh (16kW) for many days, followed by 'normal' patterns on a much-reduced scale, then finally a stretch that's perfectly flat at a small negative value. While the overall impression is that the data may be erroneous, there is no firm evidence of that, and so the data should be included in the analysis.

Figure A.C.4: An example time-series plot of a single customer's demand, expressed as consumption during 30-minute intervals. NA values have been replaced by -1 KWh.

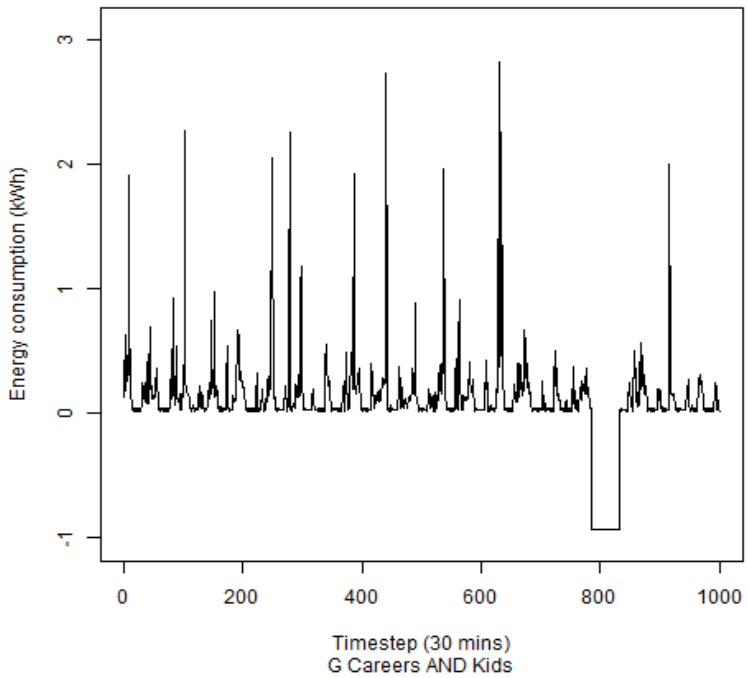
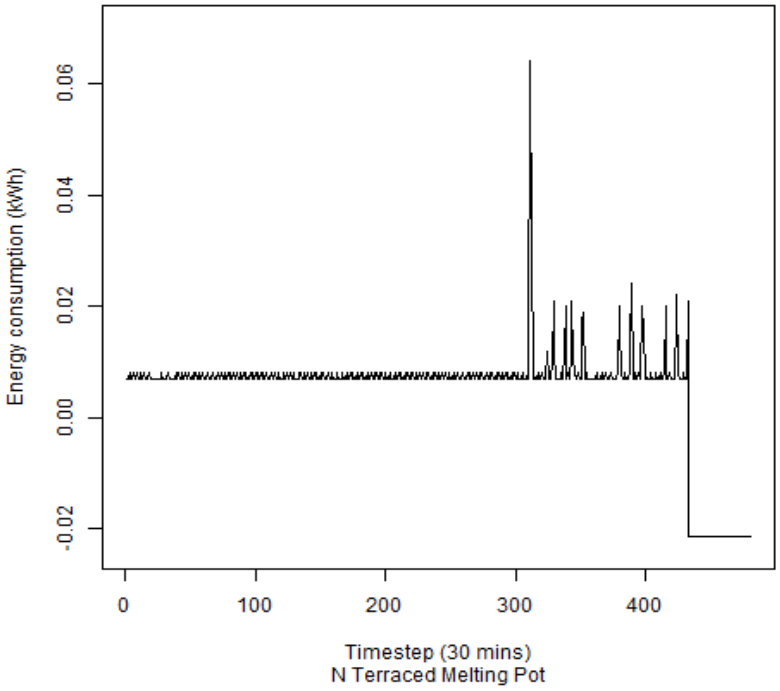


Figure A.C.5: An example time-series plot of a (different) single customer's demand, expressed as energy consumption in 30-minute intervals. The negative values are genuine.



Analysis of consumption PDF's as group averages and for individual customers

This analysis involved plotting PDFs of 30-minute energy consumption from the data to (i) establish their basic characteristics, (ii) determine how different group-average PDFs are between MOSAIC groups, and (iii) determine how different the PDFs of individual domestic customers are from each other, and from their group average. It should be noted that the PDFs calculated are entirely empirical, derived through a data smoothing technique called kernel smoothing. This is in contrast to the parametric distribution fitting that lies as the heart of our demand model. In all cases, the distribution is over all time-steps in the TC1a trial period. Figure A.C.6 shows the average result for all customers in the 'Alpha Territory category, while figures A.C.7 and A.C.8 show (different) individual customers.

The figures show a consistent pattern of a highly skewed distributions, with the probability rising very quickly to a strong peak at a relatively low demand value (typically of the order of a few 100s of Wh), before falling exponentially, with the presence of vary rare but much larger than average tail values. Unsurprisingly, the group mean PDF is very smooth and the shape was found to be almost identical for all MOSAIC types, with differences of around 10-20% in the scale of the horizontal axis.

Individual customers display similar features, but almost always with a series of smaller peaks at medium demand level – a feature that can be seen clearly in figure A.C.7. The customer represented in figure A.C.8 is slightly extreme in the relatively large demand value at which the probability peaks, but it was certainly found that the horizontal axis scale varied considerably between customers.

Figure A.C.6 The empirical PDF of 30-minute energy consumption for TC1a customers in the 'Alpha Territory' category, where the distribution is over all customers, and all time-steps in the trial period

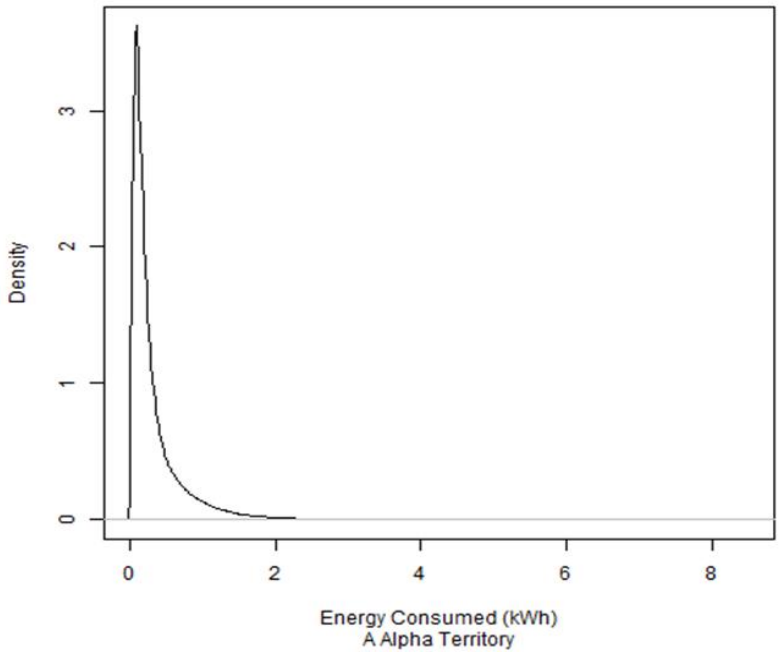


Figure A.C.7 The empirical PDF of 30-minute energy consumption for a single TC1a customer, where the distribution is over all customers, and all time-steps in the trial period

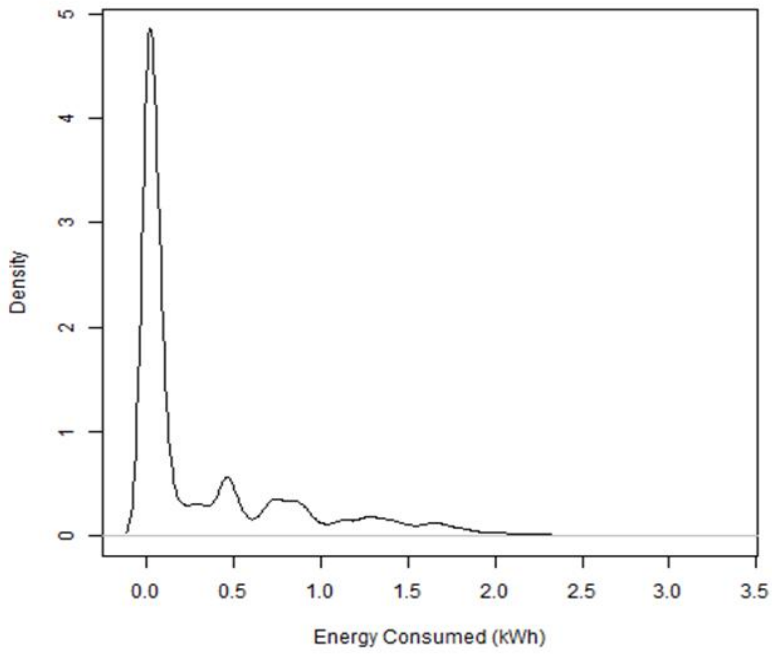


Figure A.C.8 The empirical PDF of 30-minute energy consumption for a (different) single TC1a customer, where the distribution is over all customers, and all time-steps in the trial period