Smart Network Design Methodologies

Smart Meter Data Analytics

March 2019

Contents

0 Document control

0.1 Document history

Version	Status	Issue Date	Authors
1	First Issue	18/1/19	T. Williams

0.2 Document review

Name	Responsibility	Date
Francis Shillitoe	Project Manager (NPg)	22/1/19
Alan Creighton	Technical Lead (NPg)	7/3/19

0.3 Document sign-off

Name	Responsibility	Date
Mark Nicholson	Project Sponsor	11/3/19

1 Executive Summary

This report describes the findings from the Smart Meter Data Analytics workstream. At the beginning of the workstream, in conjunction with the project stakeholders, a number of problem statements were developed. These problem statements are reproduced in Appendix B. To summarise the problem statement outcomes:

- SMA 1.1 - Disaggregating load profiles for individual customers where the smart meter load profile data has been aggregated

  Distribution Standard Licence Condition 10A (SLC 10A) requires that DNOs cannot access electricity consumption data from a domestic customer relating to a period greater than one month. We found that when electricity consumption data is aggregated and then dis-aggregated a high level of error is introduced into the data. Two disaggregation methods were tested: a “Total Consumption Method”, and a “Peak Consumption Method”. For an aggregation level of ten, the error introduced was ±57% and ±66%, respectively for each method. For an aggregation level of two, the error introduced was ±33.9% and ± 39.5% respectively. The shape and heights of consumption peaks are lost during the disaggregation process. Disaggregated consumption data recreated from aggregated consumption data for single customers must not be treated as a true representation of how that single customer would behave.

- SMA 1.2 - The application of aggregated load profiles at virtual nodes within an LV feeder network

  This problem statement set out to consider how aggregated consumption data could be applied at virtual nodes in a LV network model, and to develop rules for creation of these virtual nodes within network models. This was meant to inform the work carried out within the Novel Analysis Techniques workstream. However, within that workstream the need for the concept of a virtual node whereby time series consumption data is aggregated across different customers to develop a ‘what happened’ view, became less relevant, compared to producing building probability density functions and exceedance expectation functions of ‘what could have happened’ or ‘what might happen’. SMA3 describes how these functions can be created. Therefore, work on the problem statement work was cut short. The initial work carried out did find that for aggregation levels greater than two, the virtual nodes, on the sample networks assessed, become to course and spread out, which limits the ability to model downstream of the virtual nodes. As voltage problems tend to be experienced at the ends of feeders this could limit their application.

  Developing a method for automatically locating virtual nodes in an LV network model package may still have merits in an interim period until the novel analysis technique is implemented as this could allow the network designer to better understand power flows and voltages upstream of the virtual node.

- SMA 2 – Determining the phase connectivity of customers on an LV network when phase connectivity records are incomplete

  A number of academic papers have demonstrated that phase connectivity of low voltage customers can be carried out using smart meter voltage measurements. We wanted to use voltage measurements alone rather than consumption or a combination of voltage and consumption, in order to avoid complications surrounding access to individual consumers consumption data. We wanted to test this hypothesis on real network data. Unfortunately, due to the current low rollout numbers of SMETS2 smart meters, there were no LV feeders in Northern Powergrid’s networks where a sufficient number of SMETS2 meters were deployed. Instead, we applied time series CLNR consumption data to test LV networks built in IPSA, and ran load flows to produce simulated voltage profiles. Algorithms based on the K-means clustering principle were tested on the resultant voltage profiles but none of them produced satisfactory results.

There is currently a problem with the SMETS2 standard in that the time period over which the time step is defined does not have to be clock synchronised, unlike energy consumption. It may be worth re-examining phase identification using smart meter voltage data, but only after these voltage synchronisation issues have been resolved, or another work around has been implemented by the DNO and where there are a large number of smart meters deployed on an LV network. The K-means clustering algorithms developed in the R programming language as part of this workstream are available for reuse.

- SMA 3 – A statistical approach to generate a probability distribution of load at a given time for customers based on a suitable range of demand characteristics

  By analysing CLNR consumption data, we found that peak consumption data can be modelled using Gamma and three parameter Weibull distribution functions for specific times of day and season. These can be converted into exceedance expectation functions. These functions along with the shape and scale factors have been used in the Novel Analysis Techniques workstreams. A future development would be to use extreme value theory to better model the right hand tails of these functions.

- SMA 4 – A method to sample from the probability distribution function based on specific or general demand characteristics

  As the Novel Analysis Techniques workstream developed it was found that the need to sample from the probability distribution functions became less relevant, so this problem statement was not investigated. Details of the methodology developed for dealing with probability density functions and exceedance expectation functions can be found in the Novel Analysis Techniques workstream report.

2 SMA1.1 - Disaggregating load profiles for individual customers where the smart meter load profile data has been aggregated

2.1 Aim

The aim of this section of the project was to utilise CLNR data to develop and test techniques to dis-aggregate aggregated smart meter data to recreate load profiles for individual customers. The aggregated consumption profiles measured over 30-minute intervals are multiplied by adjustment factors determined for each customer to produce dis-aggregated consumption profiles. To calculate the adjustments, two different metrics were used to calculate the relevant weightings to apply to the aggregated consumption data. These two metrics were: 1) the total monthly consumption for each customer and 2) the peak monthly consumption for each customer. A third more simplistic method was also conducted to provide a benchmark of the accuracy of the results. This third method took the aggregated consumption and divided these by the number of customer consumption profiles that had been aggregated. The dis-aggregated consumption profiles were then compared to the original profiles to calculate the accuracy of each method.

2.2 Background Research

Previous research has suggested the acceptable levels of aggregation to preserve customer privacy and has begun to investigate the most beneficial levels of aggregation to minimise the reduction in the accuracy of the data.

[S4] The Low Carbon London study states that datasets available to a DNO that are considered personal data under Electricity Distribution Licence Standard condition 10A are: active electricity energy import, reactive electricity energy import and maximum demand. This data is only considered personal if relating to an individual customer and it relates to a period of less than one month.

Although this was the finding of the Low Carbon London study, thinking in this area is evolving as DNOs look to prepare their data privacy plans to address the SLC10A requirement. The expectation is that smart meter data can, however, can still be utilised to achieve the majority of desired outcomes by making use of the data without referring to a single customers' data for less than one month. This may be achieved by either aggregating time series energy import data for any given customer over a period of more than one month to produce a value for the total consumption for this period or aggregating time series energy import data to ensure that any network related parameter comprises at least 2 customers' data. Furthermore, maximum demand registers may be used as long as they are not reset then read more frequently than once a month.

[S9] The two Smart Meter Aggregation Assessment Reports discuss the aggregation of profiles, coupled with the development and implementation of DNO IT systems and/or business processes, to minimise the benefits reduction resulting from the aggregation whilst ensuring anonymity. The Final Report investigates, using three techniques, the level of aggregation needed to ensure anonymity of customers. The three methodologies (visual inspection, correlation analysis and clustering analysis) estimate the effect of adding additional customers consumption data into an aggregation of other customers consumption data and look to identify the point where adding additional customers makes only small changes to the aggregated profile. The three methodologies all identified this point to be between 5-10 customers.

The best methodology of correlation analysis was then used to perform more in-depth analysis to calculate visibility risk (the likelihood of an individual customer consumption profile being accurately derived from the aggregated group consumption profile). This analysis identified that aggregating two customer consumption profiles reduces the visibility risk to 22%, with further aggregation providing only a marginal further reduction in the visibility risk.

The second report looked to calculate the cumulative benefit reductions associated with aggregating customer consumption data. It found that where the aggregation is level 2, DNOs can, for a high proportion of the cases, use maximum demand (MD) data to assist them in their planning decisions. It also showed that a significant benefit loss occurs at an aggregation level of three and above and beyond an aggregation level of five the reduction in benefit was significant and the additional benefit loss associated with including further customer consumption profiles becomes marginal.

Therefore, the approach to disaggregation tested in this workstream considered three levels of aggregation (2, 5 and 10), to investigate and test methods of disaggregation. It is expected accuracy will be impacted as the level of aggregation increases.

2.3 Methodology

The CLNR dataset used for the analysis consisted of over 5,700 individual customer load profiles with data recorded every thirty minutes over a two and a half year period (01/05/2011 until 01/10/2013). This data could thus be aggregated by either combining the half hourly consumption data of numerous different customers or by calculating a single customer's monthly consumption. Importantly, the original individual customer data was also available allowing the accuracy of the disaggregation methods performed to be assessed by comparing the disaggregated load profiles to the original load profiles.

The first phase of the work required the dataset to be uploaded into R Studio including half hourly consumption (kWh), a measurement time and date stamp, each customer's MOSAIC class, and unique customer identifiers. Once uploaded into the analytics toolset a function was written that allowed multiple aggregated load profiles to be created for each half hour interval – these aggregated profiles were generated from customers with a range of consumption characteristics and consisted of aggregated consumption data from either 2, 5 or 10 customers.

2.3.1 Disaggregation Process

To disaggregate the aggregated load profile back into individual customer profiles the aggregated loads at 30-minute intervals are multiplied by a weighting factor, determined by one of three methods:

1) total monthly consumption for each customer or

2) peak monthly consumption for each customer and

3) the number of customers whose data has been aggregated.

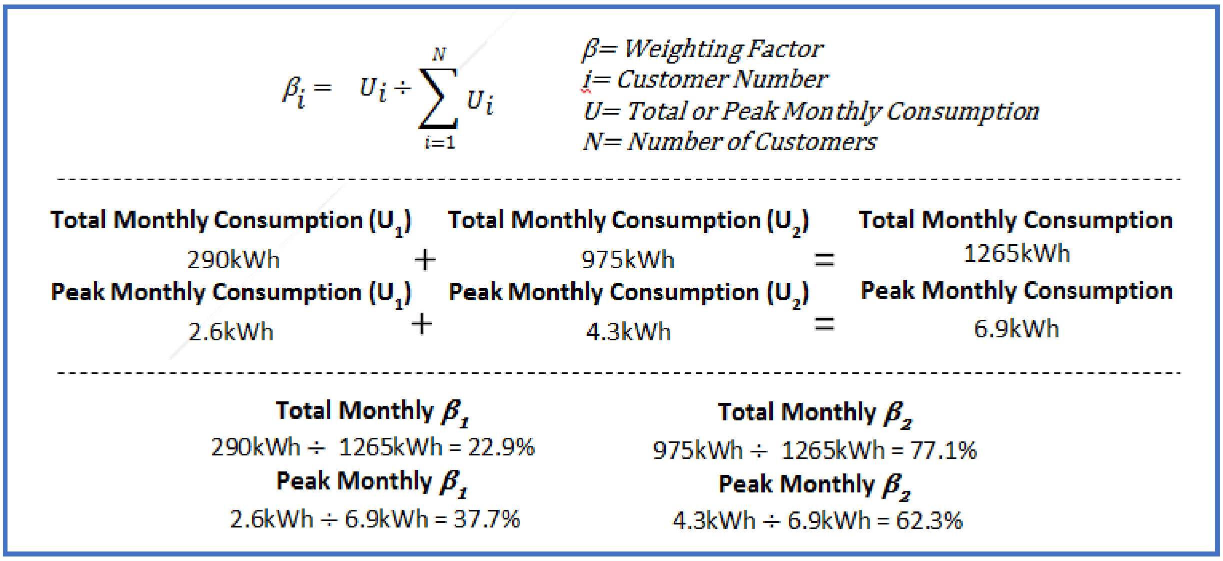
For the first two methods, the weighting factors are calculated by summing the monthly consumption or peak consumption for all customers whose load profiles have been aggregated. Next the proportion of this total that each customer's consumption represents is calculated by dividing each individual customer's peak or total consumption (dependent on the method being used) by the total for all customers whose data has been aggregated. These proportions are the weighting factors.

The third methodology simply calculates the weighting factor by dividing one by the number of customers aggregated generating the same factor for all customers.

$$\beta_i = \beta = \frac{1}{N}$$

Multiplying the aggregated load profile by these factors generates estimates of the disaggregated consumption profiles of each individual customer. These were then compared to the original data to assess the accuracy of each methodology. This was performed using mean absolute error and root mean squared error calculations.

Figure 1: Procedure for calculating weighting factors (**β**)



$$\beta_i = \; U_i \div \sum_{i=1}^{N} U_i$$

β= Weighting Factor
i= Customer Number
U= Total or Peak Monthly Consumption
N= Number of Customers

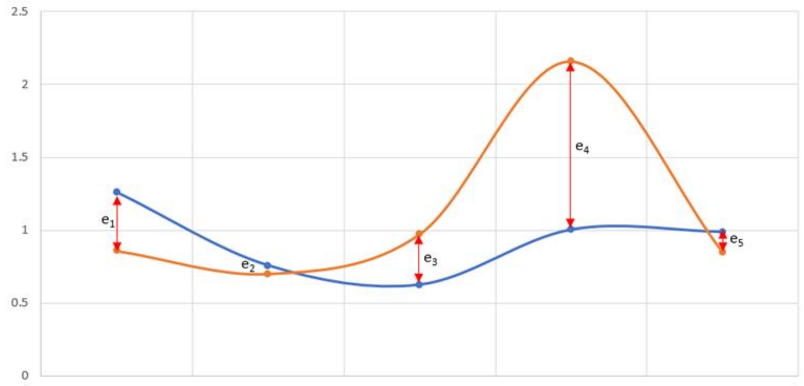
Total Monthly Consumption (U_1)		Total Monthly Consumption (U_2)		Total Monthly Consumption
290kWh	+	975kWh	=	1265kWh
Peak Monthly Consumption (U_1)		**Peak Monthly Consumption (U_2)**		**Peak Monthly Consumption**
2.6kWh	+	4.3kWh	=	6.9kWh

Total Monthly β_1
290kWh ÷ 1265kWh = 22.9%
Peak Monthly β_1
2.6kWh ÷ 6.9kWh = 37.7%

Total Monthly β_2
975kWh ÷ 1265kWh = 77.1%
Peak Monthly β_2
4.3kWh ÷ 6.9kWh = 62.3%

2.3.2 Error Calculations- Mean Absolute Error

Mean Absolute Error (MAE) is a method of calculating the accuracy of a predictive model. For each point the absolute difference (magnitude of the difference disregarding the sign) between the original signal and the model is calculated. The average of these differences is then calculated to give the mean absolute error which gives an estimate of the average error you would expect for each point you have predicted.



Equation 2: Absolute Difference

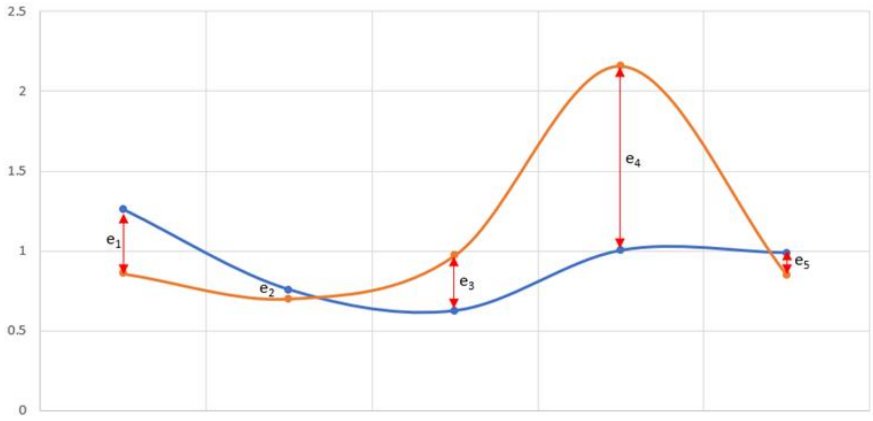
$$e_i = |y_i - x_i|$$

Equation 1: Mean Absolute Error

$$MAE = \frac{\sum_{i=1}^{N} e_i}{N}$$

2.3.3 Error Calculations- Root Mean Squared Error

The Root Mean Squared Error (RMSE) is an alternative assessment of error which has a higher sensitivity to large infrequent outliers. It is calculated using a similar process to the MAE. For each point the difference between the model and the original signal is calculated. These values are then squared to ensure all values are positive. The average of these squared errors is then taken and then square rooted to produce the RMSE.



$$e_i^2=(y_i-x_i)^2$$

$$MSE = \frac{\sum_{i=1}^{N} e_i^{\,2}}{N}$$

$$RMSE = \sqrt{MSE}$$

2.3.4 Error Calculations - Comparison between RMSE and MAE

Both error calculations express the average level of inaccuracy of a predictive model in the same units as the original values thus, for this work, the MAE and RMSE are both measured in kWh. Furthermore, both measures are negatively orientated i.e. a lower value is considered better. Where RMSE differs is that by squaring the errors before they are averaged a larger weighting is given to large infrequent outliers i.e. if a peak in consumption is underestimated the RMSE will increase to a greater extent than the MAE. This affect can be seen in Table 1 where small variances in multiple data points results in similar values for the RMSE and MAE where as one large error leads to a much greater RMSE but no change in the MAE.

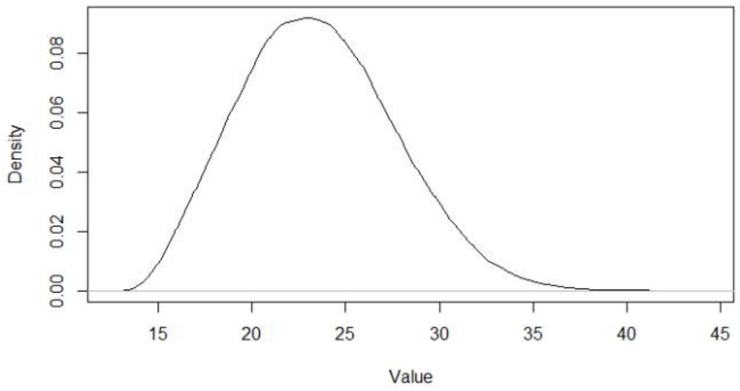
Table 1: Example MAE and RMSE calculations

ID	Error	Absolute Error	Error ^{2}
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	4	4	16
7	4	4	16
8	4	4	16
9	4	4	16
10	4	4	16
		MAE	RMSE
		2.500	2.915

ID	Error	Absolute Error	Error ^{2}
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	-25	25	625
		MAE	RMSE
		2.500	7.906

To get a better understanding of the average error in each disaggregation method the process was simulated multiple times. The outcomes of these multiple simulations were then plotted as a probability density function (an example of which can be seen in Figure 2). The probability density function is a smoothed version of a histogram of the data. It shows the likelihood of each outcome occurring where the higher the line the more likely an outcome is to occur, with the position of the peak giving the modal value. To determine the likelihood that the size of the error will be between two values, the area under the curve between the two values is calculated.

Figure 2: Example of a probability density function



2.4 Results

The following section presents results of a single test case for each level of aggregation and discusses the error in each case. For each level of aggregation multiple simulations were then conducted to provide a better understanding of the average error of both methods.

2.4.1 Two Customer Aggregation

As described in the methodology above, the weighting factors for May were calculated for both customers included within the aggregation. The aggregated signal was then multiplied by these weighting factors to recover the original signals. Figure 3 and Figure 4 show a comparison of the original load profiles and the outputs created from the two methods of disaggregation when conducted on the data from two customers.

Figure 3: Comparison of the original consumption profile of Customer 1 for the 1st May 2011 with the results of three disaggregation methods

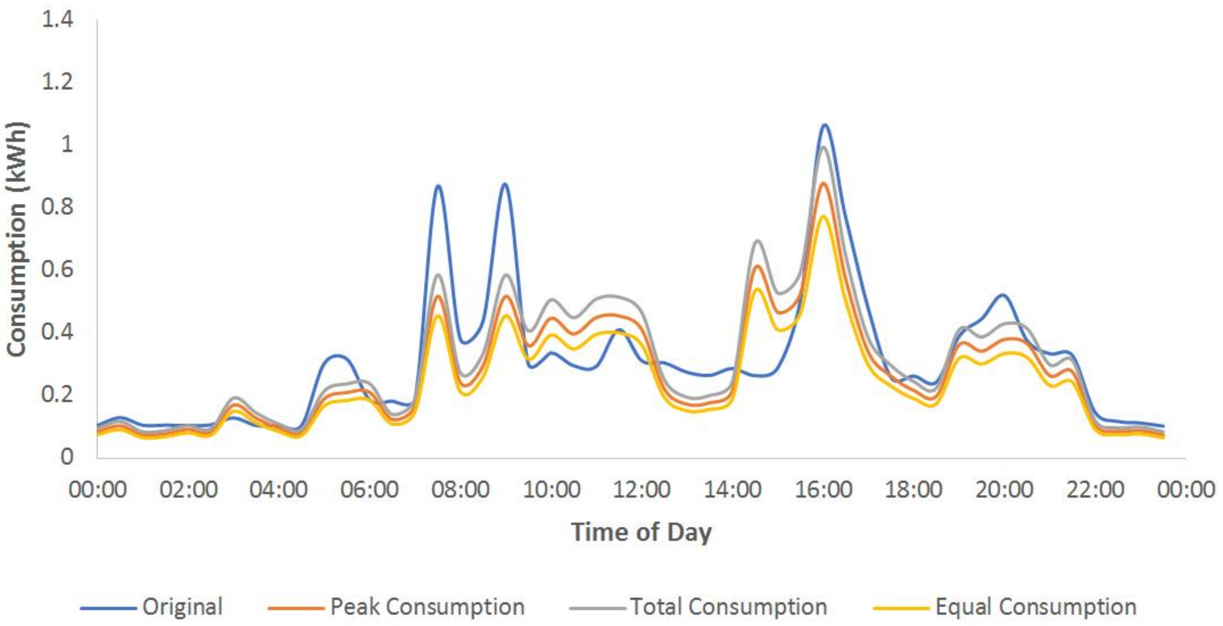
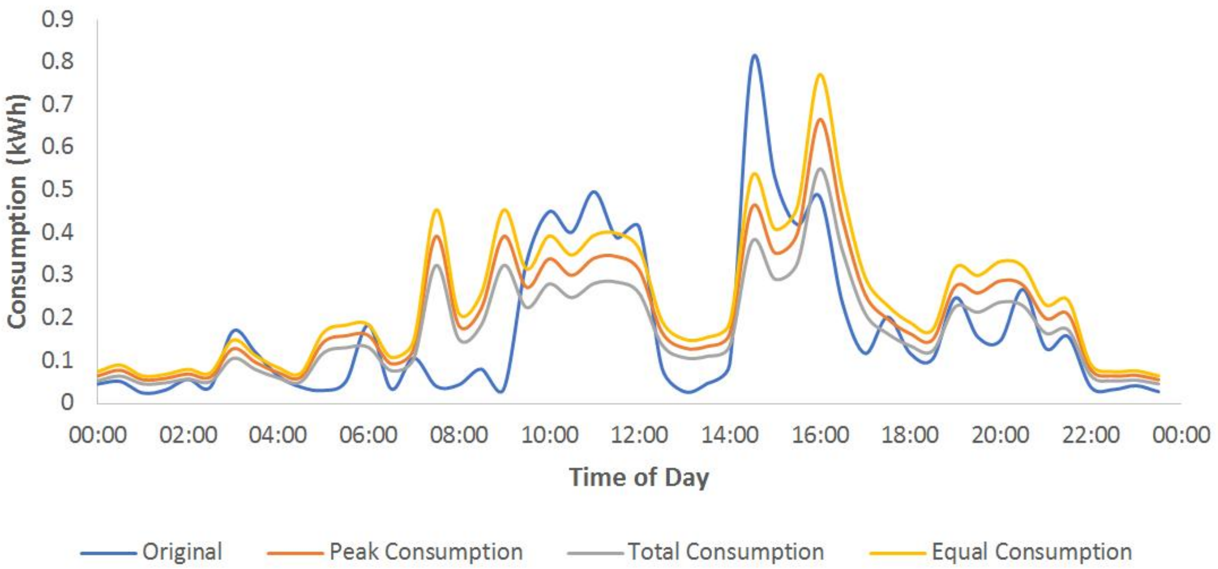


Figure 4: Comparison of the original consumption profile of Customer 2 for the 1st May 2011 with the results of three disaggregation methods



In Figure 3, the shapes of the disaggregated consumption profiles are quite similar to the original with peaks at similar positions throughout the day except for the additional peak at 14:30 and the trough between 12:30 and 14:00. For this consumption profile the disaggregated values tend to be an underestimation of this customer's consumption. In Figure 4 the disaggregated consumption profiles both follow similar patterns to the original signal however, there are again peaks in the data that weren't previously there and the maximum consumption in this period occurs at a different time (16:00 rather than 14:30). To assess the accuracy of these models the root mean squared error and mean absolute error were calculated as shown in Table 2.

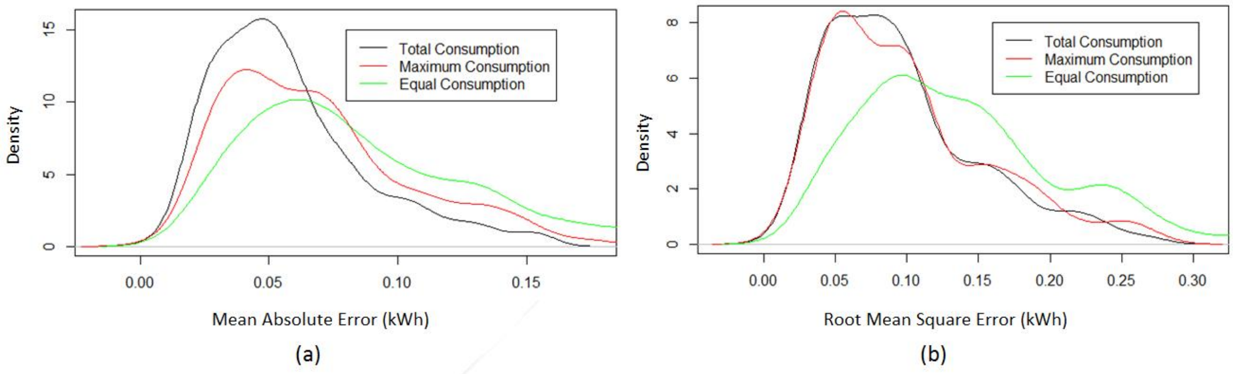
Table 2: Comparison of the Root Mean Squared Error and Mean Absolute Error for the three methods of disaggregation

Method	Root Mean Squared Error (kWh)	Mean Absolute Error (kWh)
Peak Consumption	0.12	0.09
Total Consumption	0.12	0.08
Equal Consumption	0.14	0.10

The results in Table 2 show that all three methods have similar error values. The mean absolute error is slightly smaller for the method using total monthly consumption and both values are better than the unweighted disaggregation. To contextualise these values, it is important to look at the mean consumption for the two profiles which was 0.24kWh per half an hour period. Therefore, for this case the mean absolute error represents an average error of ±33.3% for the total consumption method, an average error of ±37.5% for the peak consumption method and ±41.7% for the simple unweighted method.

The next step was to run these simulations multiple times for the same day, 1 May 2011, to get a better understanding of the average values for each of these error measurements. The results of running the same algorithms 200 times for random selections of two customers can be seen in Figure 5. Both methods using a weighting factor have a similar distribution of results however, generating weighting from the total monthly consumption had slightly lower mean values for both the root mean squared error (0.09 compared to 0.10) and the mean absolute error (0.06 compared to 0.07). These error values are both lower than the average value for just dividing the consumption by the number of customers (MAE=0.09 and RMSE= 0.14). Comparing the mean absolute error values to the mean consumption of all these scenarios (0.177 kWh) show the average percentage error is ± 33.9% for total consumption and ± 39.5% for peak consumption.

Figure 5: Distribution of values of (a) mean absolute error and (b) root mean square error for 200 simulations of the disaggregation process



(a)

(b)

2.4.1.1 Comparison to days of high and low consumption

The analysis above was then repeated for a winter's day (3rd January 2012) and a day in summer (8th August 2012). These dates were chosen to act as an example of a day of high consumption and one of low consumption respectively. The average error results for these dates are shown in Table 3.

Table 3: Comparison of error values for days throughout the year

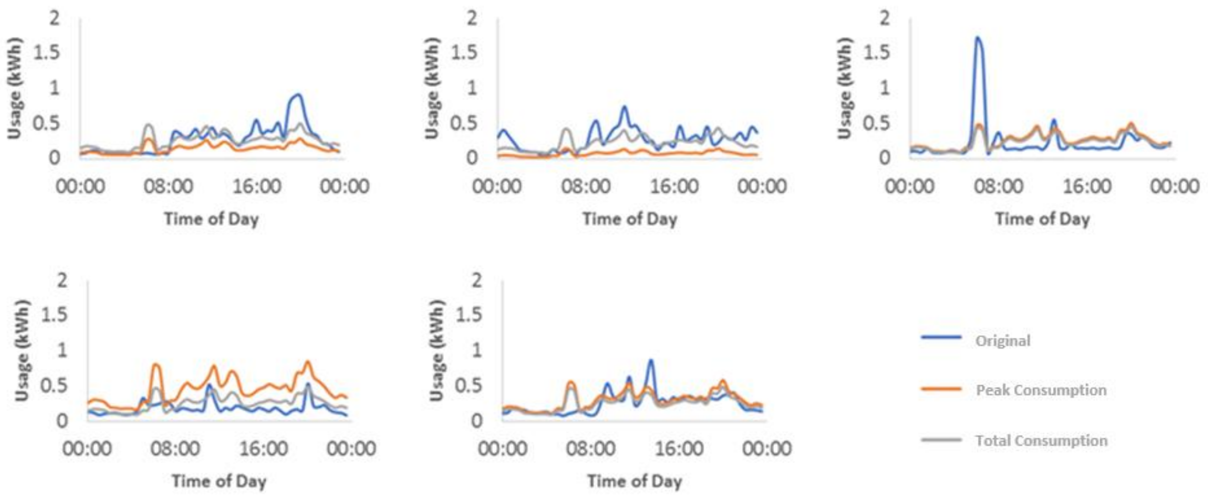
Disaggregation Method	Date	Mean Consumption	Average RMSE	Average MAE
Total Consumption	03/01/2012	0.24	0.11	0.07
	08/08/2012	0.16	0.1	0.07
	01/05/2011	0.18	0.09	0.06
Peak Consumption	03/01/2012	0.24	0.12	0.09
	08/08/2012	0.16	0.09	0.05
	01/05/2011	0.18	0.1	0.07
Equal Consumption	03/01/2012	0.24	0.18	0.12
	08/08/2012	0.16	0.12	0.08
	01/05/2011	0.18	0.14	0.09

The average error values for each method remain similar across all three time points. Therefore, when calculating the average percentage error by dividing the mean consumption by the mean absolute error the period with the highest mean consumption would have the lowest percentage error, suggesting that on days of higher consumption more accurate results may be achievable. Graphical representations of these results can be seen in Appendix A.

2.4.2 Five Customer Aggregation

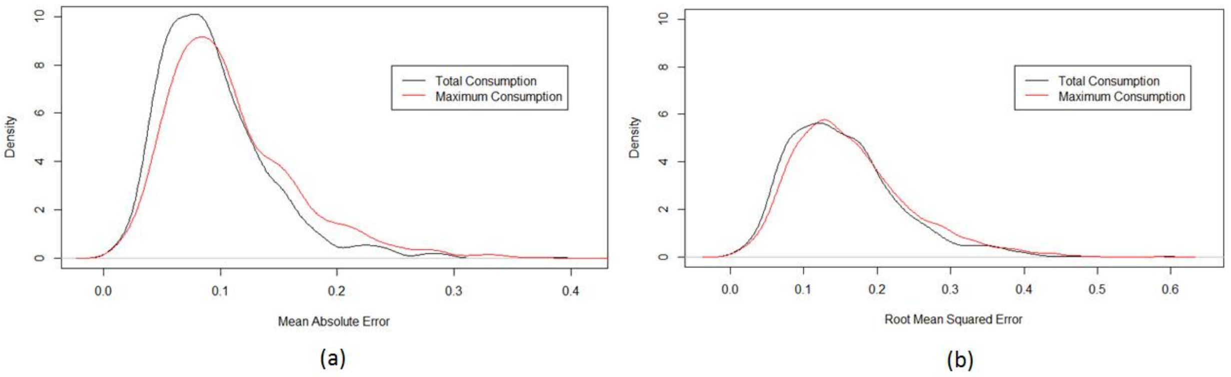
The process was then repeated for groups of five customers, also using the consumption data on 1 May 2011, so that the results are comparable. Figure 6 shows the results of the two disaggregation methods versus the original consumption profile for a sample of five customers. Additional peaks are seen within the disaggregated consumption profiles and the main peaks from the original signal can be underestimated by both methods. The graphs do however, show that there are still periods of the day where both methods mirror the original data. Due to the large inaccuracies shown for the method using equal consumption for two customers this method was not tested for groups of 5 and 10 customers.

Figure 6: Example disaggregation of the consumption profile of five customers. The graphs compare each customer's original consumption profile to the profiles created from the two methods of disaggregation



The process was again repeated 200 times to generate average values for the root mean squared error and the mean absolute error. Figure 7 shows the results of this process where the results for the accuracy of replicating each individual signal has been included i.e. for each simulation the values of the root mean squared error and the mean actual error for each customer have been taken into account.

Figure 7: Distribution of values of (a) mean absolute error and (b) root mean square error for 200 simulations of the disaggregation process



(a)

(b)

As described for the two-customer disaggregation the average value for both methods of error are very similar. The method using the total consumption had slightly lower values of the mean absolute error (0.093 compared to 0.109 for peak consumption) and the root mean squared error (0.154 compared to 0.165 for peak consumption). These values are also slightly larger than those calculated for two customer disaggregation. The average mean half hourly consumption per customer for groups of 5 customers was 0.188 kWh. Comparing the two average MAE values to the mean consumption gives an average error of ± 49.5% for using total consumption and ± 58.0% using peak consumption.

2.4.3 Ten Customer Aggregation

In Figure 8 the original consumption profiles of 10 customers, again relating to 1 May 2011, are compared to the disaggregated signals. The consumption profile of customers with lower consumption throughout the day is more accurately reproduced compared to those with large peaks throughout the day.

Figure 8: Results of the disaggregation process for 10 customers. Each graph showing the original consumption profile and the ten disaggregated profiles

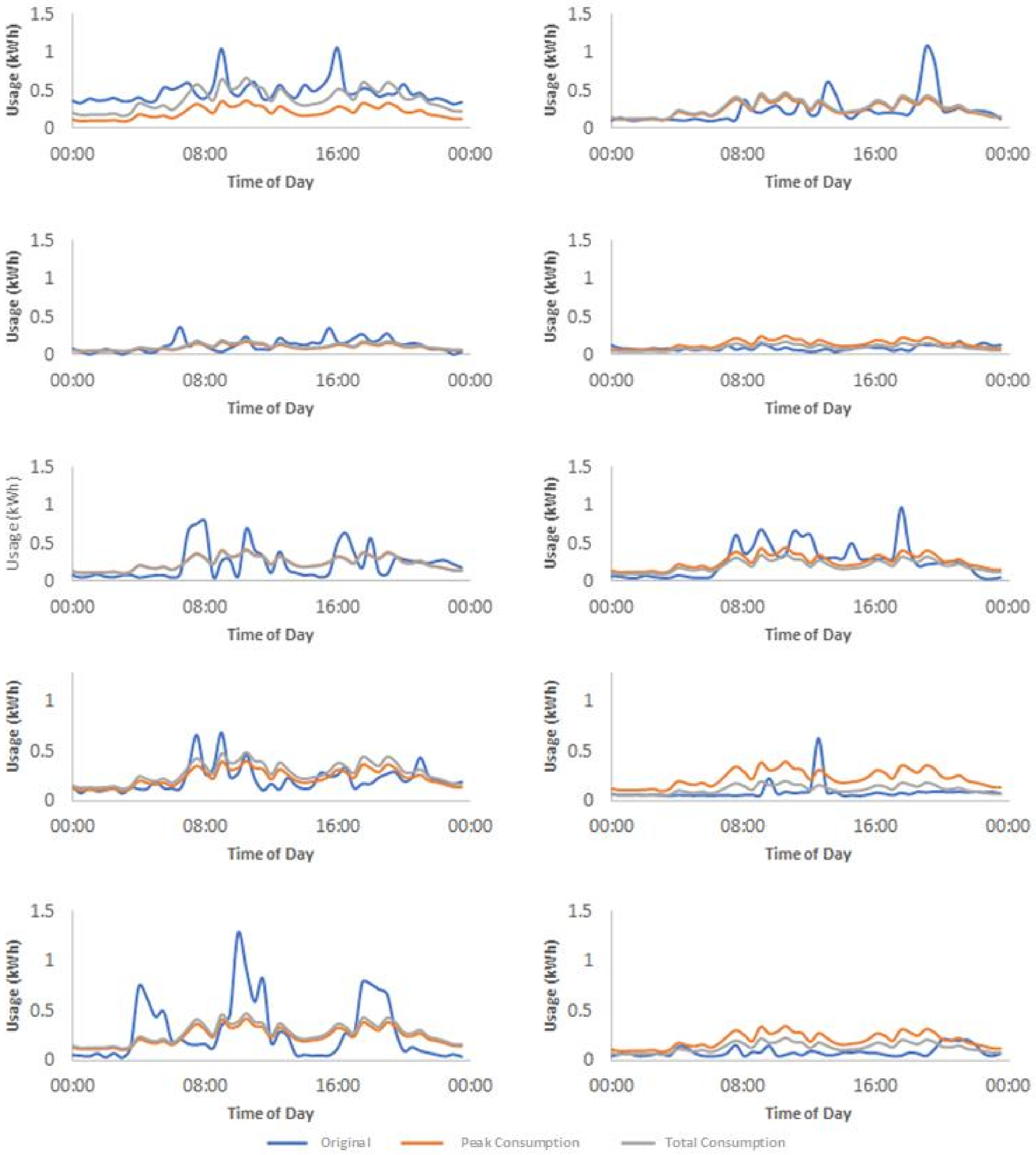
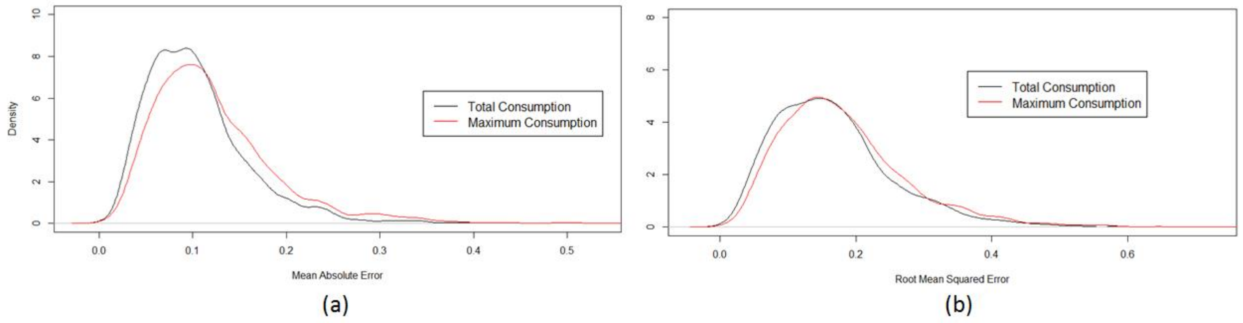


Figure 9 shows the results of running the disaggregation processes 200 times and for each individual customer assessing the accuracy of the disaggregation process. The average root mean squared error and mean absolute error for the groups of ten customers are greater than the values for customer groups of 2 and 5 customers and have a higher maximum error. However, the pattern of the slightly higher accuracy being achieved from using the total consumption to calculate the weightings continued. The mean half hourly consumption for customer within groups of 10 was 0.189 kWh when comparing this to the average mean absolute error per customer the for total consumption (0.107 kWh) and the average mean absolute error per customer for peak consumption (0.124 kWh) the percentage error is calculated to be ± 57% and ± 66% respectively.

Figure 9: Distribution of values of (a) mean absolute error and (b) root mean square error for 200 simulations of the disaggregation process



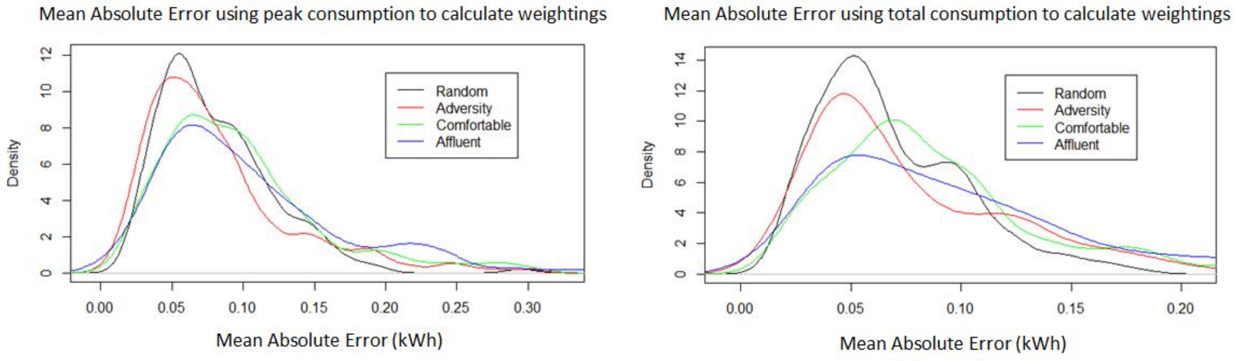
(a)

(b)

2.4.4 Two Customer Aggregation- Testing the effect of drawing customers from the same Mosaic Segment

To investigate whether recommendations could be made to better inform the aggregation process it was hypothesised that aggregating customers from the same Mosaic segment could lead to improved results of disaggregation as these customers are more likely to have similar levels of consumption. The results of conducting 200 runs, again relating to 1 May 2011 date, of the disaggregation process for samples of 2 customers taken from the same Mosaic class, “Affluent”, “Comfortable”, or “Adversity”, can be seen in Figure 10. Also shown on the graph are the results for running the process with customers chosen from random Mosaic segments.

Figure 10: Comparison of Mean Absolute Error Values when sampling customers from the same Mosaic Class



The results of this testing show that, by drawing customers from the same Mosaic class the distribution of mean absolute error values for each of the 200 simulations remains similar to when sampling from customer data regardless of Mosaic class. This suggests that further segmentation of the Mosaic segments may be required to discover customer groups with more similar consumption characteristics as the accuracy of the disaggregation techniques will be increased if the original load profiles are more alike.

2.5 Conclusions and Recommendations

Both disaggregation methods investigated have a high level of error, with the Total consumption method being slightly better than the Peak consumption method. Using the simple method of assuming that consumption is equal for each customer produces errors of ~50% when aggregating 2 customers' data suggesting that although it is a simpler calculation it would not produce the required accuracy.

Table 4: Comparison of the average errors for both disaggregation methods

Aggregation Level	Total Consumption Method Average Error	Peak Consumption Method Average Error
2	±33.9%	±39.5%
5	±49.5%	±58.0%
10	±57%	±66%

Disaggregated consumption data for single customers must not be treated as a true representation of how that single customer would behave. As reported in the two Smart Meter Aggregation Assessment Reports these results support the findings that an aggregation level of 2 sufficiently anonymises customer data on days of both high and low consumption; any higher aggregation level would introduce far higher levels of error with no practical benefit.

Aggregated consumption data could be used to build load models at virtual nodes which could be used to model the network upstream of the virtual node, but not downstream.

The aggregation and disaggregation process causes a high degree of error to be introduced. The level of information lost, particularly the shape and height of the consumption peaks raises the question of the usefulness of disaggregated data.

3 SMA1.2 – The application of aggregated load profiles at virtual nodes within an LV feeder network

SMA 1.2 set out to achieve the following objectives:

- Consider how aggregated consumption data could be applied directly at a virtual node to remove the need to disaggregate it;
- Create rules for aggregating consumption data at virtual nodes;

Some brainstorming was carried out on these objectives, however in parallel within work carried out in the Novel Analysis Techniques workstream it became clear that the need for a virtual node as defined by these objectives became less relevant. The concept of using aggregated time series consumption data directly within a network model to form a 'what happened' view became less relevant, compared to using consumption data to extract histograms to build probability density functions (PDFs) and exceedance expectation functions of 'what could have happened' or 'what might happen'. SMA3 describes how these PDFs can be created.

We did carry out some initial work to determine the loss of fidelity in a network model as the aggregation level is increased.

Figure 11 shows an LV feeder from one of our test networks. It comprises 36 single phase customers and one three phase customer. In Figure 12 we show an example of creating virtual nodes. The general approach is as follows:

- we trace up the mains cable from its end, back to the source substation. A count is kept of the number of properties connected to a phase;
- when the count meets the required aggregation level on that phase, a virtual node is dropped into the model at the service cable / mains cable joint;
- virtual nodes are created per phase;
- where there are more customers at the service cable / mains cable joint than the aggregation level, the customers are grouped together at an aggregation level equal to the number of customers;
- if the source substation is reached and there are not enough customers in the phase count to meet the aggregation level, these remaining customers are grouped with the last created virtual node on that phase.

By visual inspection of Figure 12, we see that for an aggregation level of two, a reasonable model of the mains cable can be built. However, for greater aggregation levels (five or ten in the example), the virtual nodes became too far spread out and the ability to model power flows towards the ends of the feeder become more and more limited. There are also only four customers connected to the yellow phase L2, so it is not possible to aggregate these up to five or ten.

Figure 11: Sinderby – South LV feeder showing phase allocations at properties

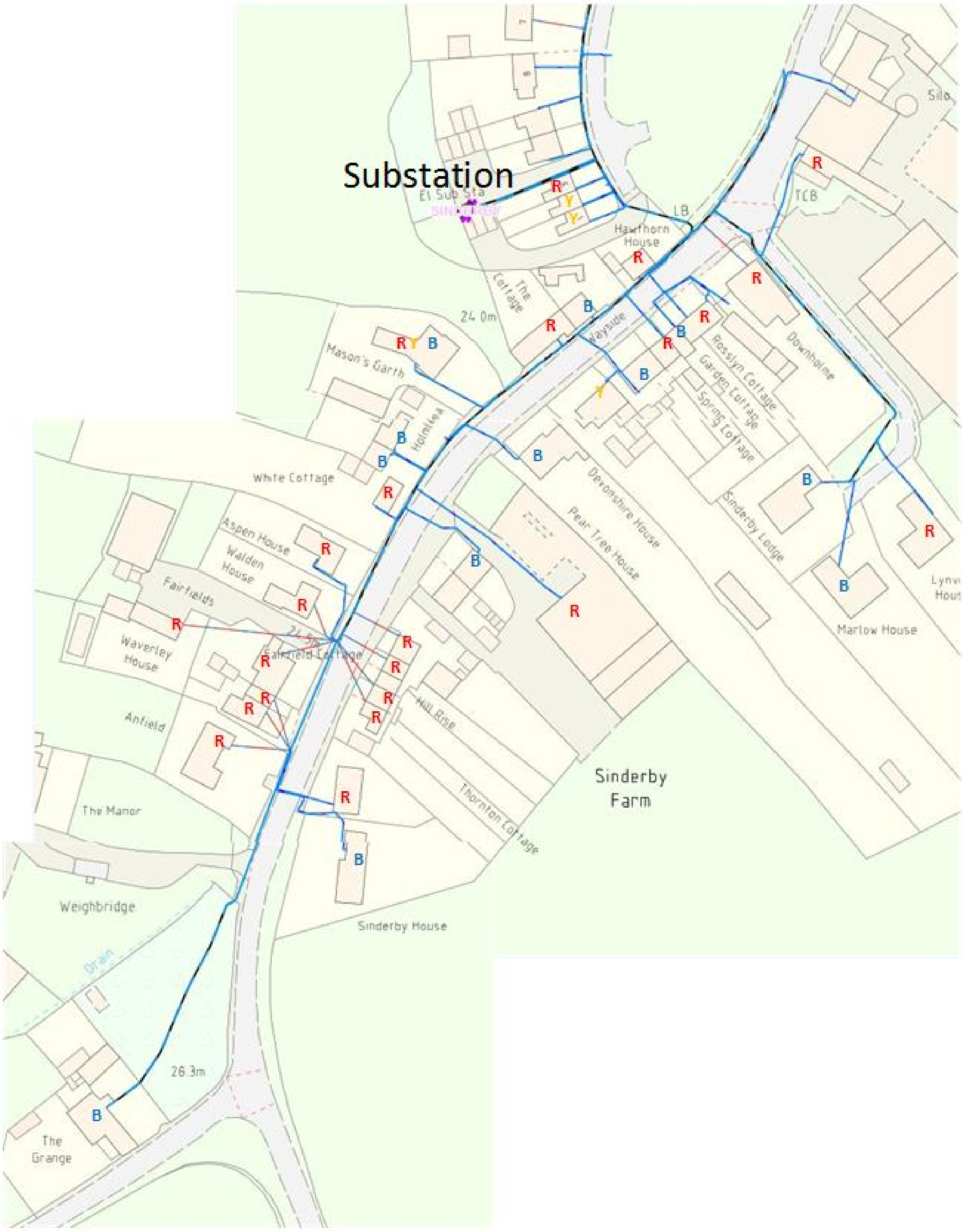
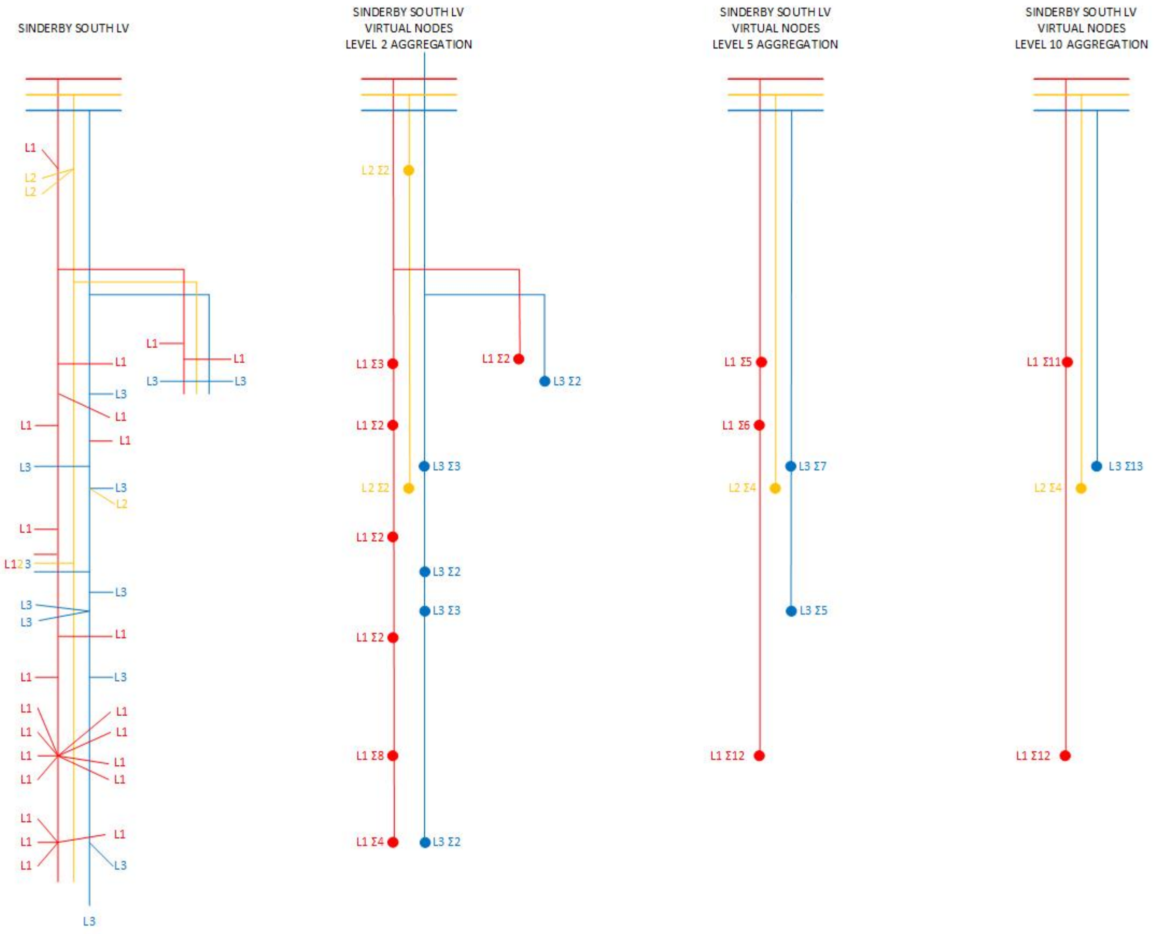


Figure 12: Creation of virtual nodes on Sinderby Southern feeder at aggregation levels 2, 5 and 10



The novel analysis technique methodology side steps the issue by using the network characterisation[1] at the point of interest (for example a cable section) and combining this with an exceedance expectation function of the downstream demand (derived from the PDF) to determine the risk of the cable section being overloaded or outside voltage limits. This is covered in detail in the Novel Analysis workstream reports.

To keep within the conditions of SLC10A, the challenge regarding access and use of a customer's smart meter data now becomes one of how consumption histograms can be created (which can be converted to PDFs) rather than aggregating different customer's time series consumption data. In SMA3 and the Novel Analysis Techniques workstream, our approach looks at creating 192 PDFs for each of the 48 half-hour time periods in a day for each of the four seasons. In this way the data is not related to a specific time period on a specific day within a particular month. The limited analysis carried out, given the change in emphasis of this aspect of the project, concluded that :

- For an aggregation level of two, a reasonable network model could be created using virtual nodes, particularly where detailed network analysis is only required upstream of the virtual node, as is the case for many of the assessments carried out in the design function.

[1] Network characterisation is the relationship between loading and voltage at a cable section against total downstream demand

- It was possible to develop an algorithm for automatically positioning the virtual nodes, for a given aggregation level, in a network model and to create a schedule of consumption data required to populate these virtual nodes.
- It was reasonable for the functional specification for the LV network modelling tool include such a requirement, potentially as an option that could be implemented as an interim solution pending the implementation of assessment based on the novel assessment technique.

The possibility of applying the virtual node concept as an incremental development of a traditional LV modelling approach might be worth of exploration.

4 SMA 2 – Determining the phase connectivity of customers on an LV network when phase connectivity records are incomplete

4.1 Aim

In this section we examine whether smart meter data can be used to identify the phase connectivity of customers when this information is not present in DNO records. At the time of this study there were no LV networks within NPg's licence areas that had an extensive deployment of SMETS2 meters, and SMETS1 meters were not accessible by NPg. Therefore, we have adopted an approach whereby real test networks have been selected but dummy half-hour power flow data (taken from data captured during CLNR) is applied to customers on these networks, and load flows run (using the IPSA unbalanced load flow algorithm), to produce simulated voltages at customer properties. Two test networks are used. The first, Sinderby, is a rural network comprising 58 customers on two feeders. The second, Cranwood, is an urban network with 675 customers on six feeders. Sinderby is in NPg's Northeast licence area, where phase connectivity information has not been recorded as standard. Field measurements were taken at each property using a Hasys phase identification unit, to identify the property phase connections. Cranwood is in NPg's Yorkshire licence area where the majority of properties have their phase connectivity recorded in NPg's asset database.

In this study, algorithms have been developed in the R programming language to attempt to identify the phase connectivity of customers using the simulated voltages from the load flow results at each customer property. These are compared against the actual phase connectivity records to determine how successful the algorithm is.

4.2 Background Research

A literature review identified several classes of method that could be used to determine phase connectivity based on smart meter measurements. These can be divided into the following classes:

- Voltage correlation. This method relies on the fact that neighbouring properties on the same phase will have similar voltage profiles.
- Power flow summation. These generally rely on a full rollout of smart meters on a network, and require power flow measurements on feeder ways at secondary distribution substations. Rollout of smart meters in Great Britain is incremental and it is not currently mandated for customers to have smart meters installed. The majority of secondary distribution substations do not have measurement equipment installed on feeder ways. The methods may also be compromised in Great Britain due to the requirements for aggregation of consumption data under Standard Licence Condition 10A.
- Other. Several academic papers looked at using signals that are not available on SMETS2 meters such as harmonic measurements or phase angles.

Therefore, we have explored voltage correlation algorithms in this study. The most promising algorithms were published in [S2][2]. In this paper K-means and Gaussian Mixture Model clustering algorithms are developed in MATLAB and tested on a European Low Voltage Test feeder modelled in Open DSS. The algorithms had varying levels of success depending upon the statistical markers used. The most successful statistical marker was where the voltage differences between each time step was used, and a 100% success rate was claimed. However, the voltage measurements were taken with a 1 minute resolution. The default

[2] F. Ni, J. Q. Liu, F. Wei, C. D. Zhu – Tellhow Sci-Tech Co., Ltd., S. X. Xie – Eindhoven University of Technology, "Phase Identification in Distribution Systems by Data Mining Methods", IEEE 2017

resolution with SMETS2 smart meters is 30 minutes, however different resolutions, down to at least one minute can be configured on SMETS2 meters by DNOs.

The approach taken in [S105][3] uses time series voltage measurements from each location and partitions these into clusters, also using K-means. The method uses similar clustering techniques to those in [S2], however the Pearson correlation distance rather than the Euclidean distance is used within the K-means clustering algorithm. The paper discusses the reduced accuracy of aggregating data over longer time periods through the comparison of results with readings every 10 seconds, 30 seconds, 1 minute and 5 minutes.

4.3 Methodology

4.3.1 Sinderby Test Network

The Sinderby network, Figure 13, comprises two feeders with 16 properties connected to the north feeder, Figure 14 and 42 connected to the south feeder, Figure 11. There are a couple of properties where there is a service termination, but no electricity is currently supplied. The network is slightly unusual as it used to be a split phase network; there are far fewer customers connected to the yellow phase than red or blue on the South Feeder. On the south Feeder there is one three phase customer, which was ignored in the study, as the consumption data set used was for single phase customers. During the visit by field staff, there was a small number of properties where they could not positively correlate the property to the records in the eAM Spatial database (which was still under development); these properties were ignored from the study. This omission did not have a material effect on the findings of the analysis. This gives a total number of modelled customers as 16 on the north feeder and 32 on the south feeder. The phase connectivity of customers on the north feeder way is shown in Figure 14.

[3] Arya, V., Mitra, R., 2013. “Voltage-based clustering to identify connectivity relationships in distribution networks”. In: Proceedings of 4th IEEE International Conference on Smart Grid Communications. https://doi.org/10.1109/ SmartGridComm.2013.6687925.

Figure 13: Sinderby LV network

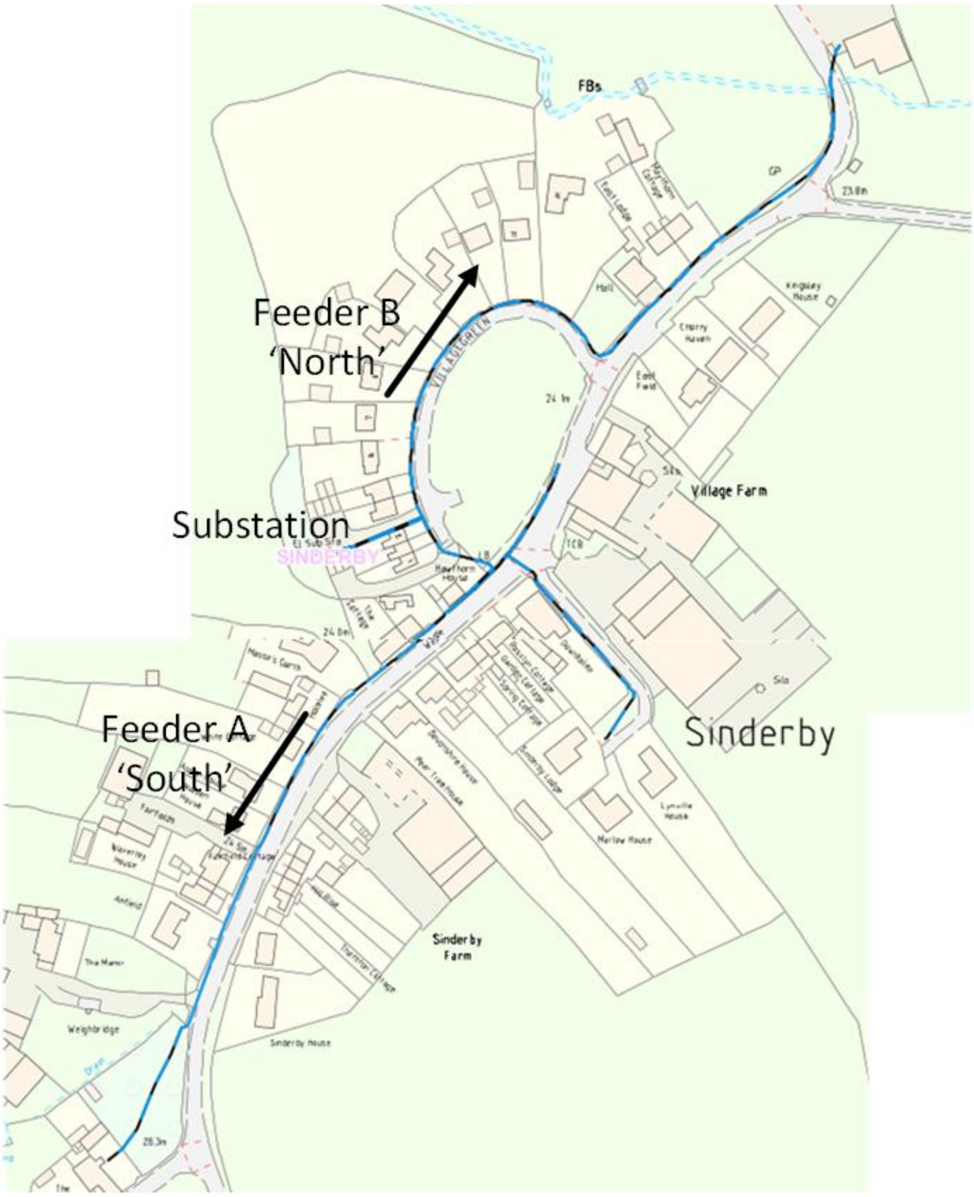
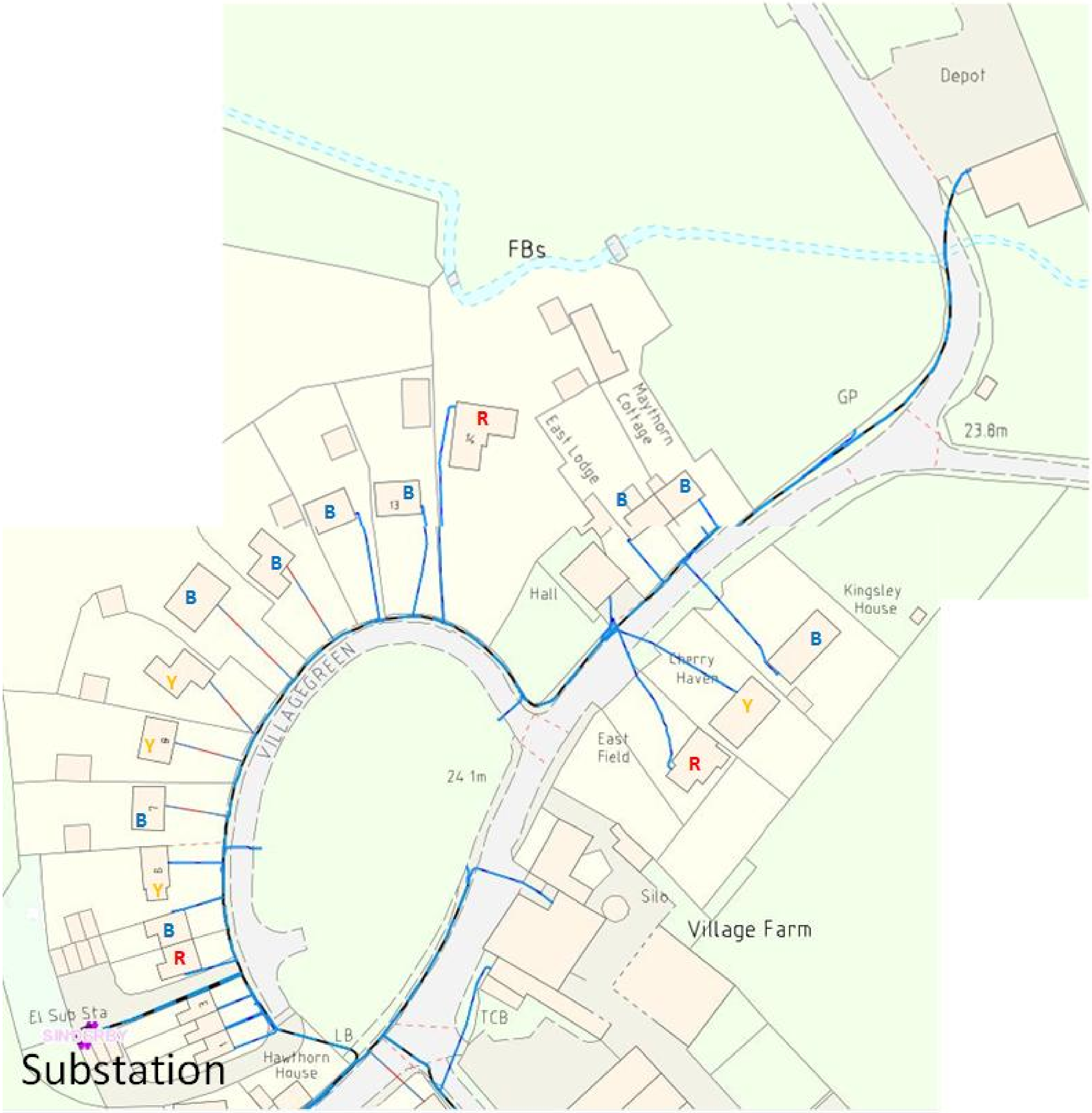


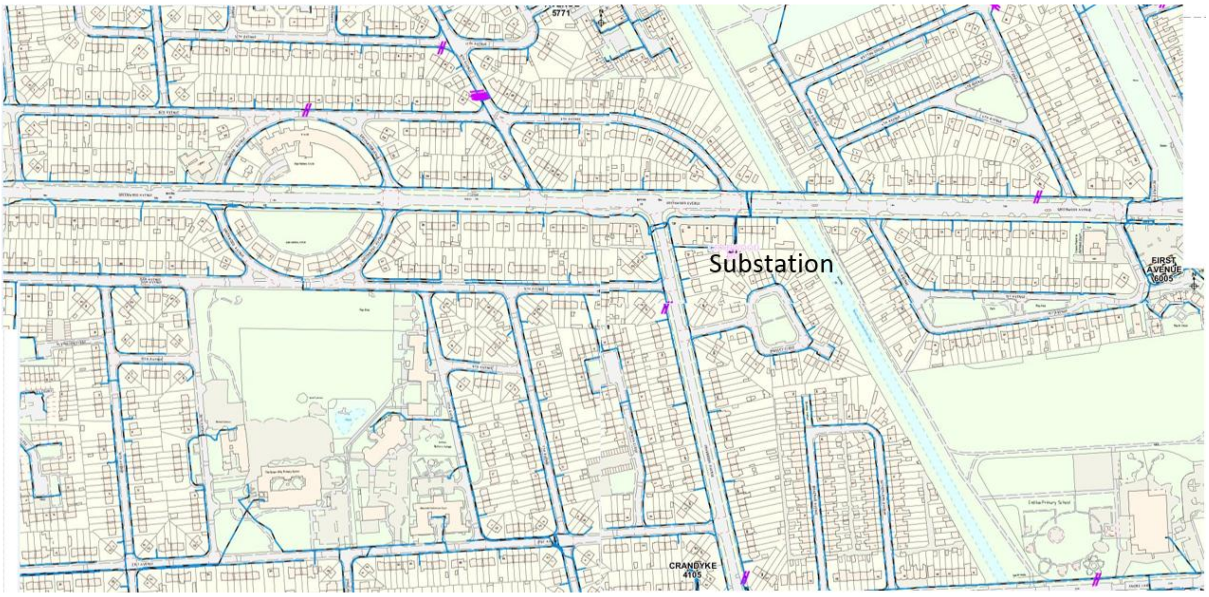
Figure 14: Sinderby – North LV feeder



4.3.2 Cranwood Test Network

Cranwood is an urban network with 675 customers on six feeders, Figure 15. However, 11 of these customers could not be matched to an output of the IPSA model therefore, the results of 664 customers have been used within testing. This is due to data quality issues in eAM Spatial which has only recently gone live; the omission of these customers was not considered material.

Figure 15: Cranwood LV network (other networks also partially shown)



4.3.3 CLNR data and IPSA load flow

Unbalanced network models for the Sinderby and Cranwood test networks have been built in IPSA. Consumption data was taken from the TC1a CLNR data set. Each customer was allocated a nearly sequential series of 1000 half-hourly active power measurements (approximately 21 days' worth of data) from 1 May 2011. The presence of missing data from too many customers meant that some half hours had to be. A script was developed to carry out an unbalanced load flow on each of the 1000 time periods to produce a set of simulated voltage measurements at each customer property for each of the 1000 half-hour periods. These measurements were then fed into the K-means clustering algorithms developed algorithms in R.

4.3.4 K-means Clustering

K-means clustering is a type of unsupervised learning, which aims to group data based on the similarity of the specified variables, the number of groups the data is segmented into is set by the value K. The algorithm works iteratively to assign each individual row of data to one of K groups based on the values of each variable, clustering those with the most similar values together.

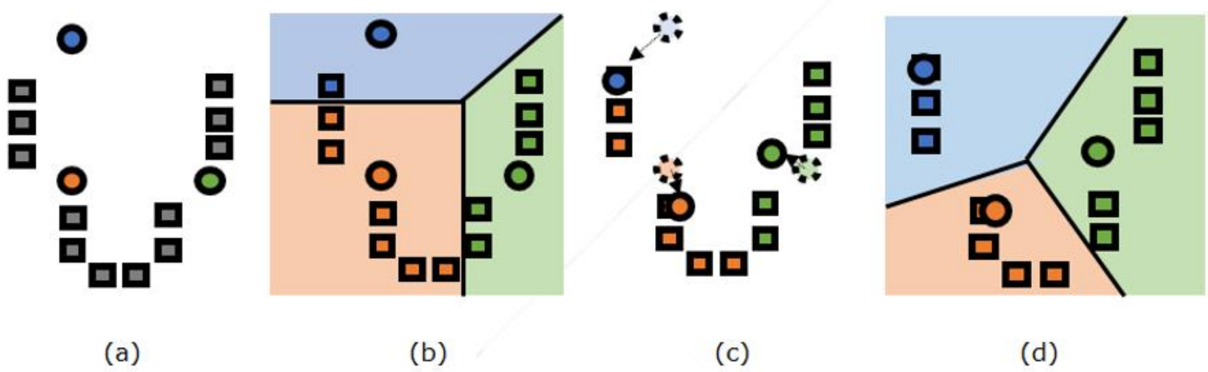
These principles have been applied to phase identification of single phase customers. Utilising the hypothesis that customers on the same phase will have similar voltage characteristics, the clustering algorithm groups customers based upon these characteristics of their voltage profiles. If the method is successful it should lead to those customers connected to the same phase being grouped together. The value of K will be set to three as we want to segment the customers into three groups, one for each phase.

Worked example

Taking the input to the K-means algorithm to be two variables (for example the mean and median of the 1000 voltages for each customer) points can be plotted in x-y space with the position of each point dependent on the value of each input variable. Setting the value of K to equal three (for the three phases), the algorithm chooses three of the points to act as the first guess of the centres of each group, Figure 16(a). Then going customer by customer, the distance to each centre point is calculated. The customer is then assigned to the centre that is closest and grouped with all the other customers who are also closest to that centre, Figure 16(b).

Each of the centre points are then moved to the middle of the points that are grouped together, Figure 16(c), and the process of calculating the nearest centre is repeated for each customer. This may lead to customers changing which group they are assigned to due to a new centre now being closer. The whole process of moving the centres is then repeated multiple times until the centres no longer move as they have found their optimum position, Figure 16(d).

Figure 16: K-means clustering algorithm process



The two papers investigated use a variety of descriptive features, calculated from each customer's voltage profile, in various combinations as inputs to the K-means algorithm. The first paper investigates five combinations of descriptive features that are used as the input variables into the K-means algorithms. The six combinations, or methods, all using the Euclidean distance to calculate which cluster the customer should be assigned to, are:

1. Mean and Median
2. Mean, Mode and Median
3. Mean, Mode, Median and Standard Deviation
4. Mean, Mode, Median, Standard Deviation, Maximum, Minimum, and Variance
5. First 50 values in the voltage data set of each customer
6. First 50 differences in the voltage data set for each customer

For the first four methods, to calculate the above descriptive metrics for each customer (mean, median, mode, standard deviation, maximum, minimum and variance), all 1000 data points are included within the calculation and a single value returned i.e. the mean voltage for each customer is calculated by taking all 1000 data points relating to each customer in turn and calculating the average of each set of 1000 values, returning the mean for each customer.

In method 5 as per the methodology of Ni et al [S2] only the first 50 data points are used as inputs to the K-means algorithms. This is also the case for method 6 where the first 50 differences for each customer are used as the input, where a difference is calculated as the voltage change from one time point to the next. The time points are contiguous. The methodology was tested using all 1000 data points however, there was no discernible difference to the results therefore the original methodology of using only the first 50 data points was chosen to reduce computational time.

A seventh method presented in Arya and Mitra [S105] again uses the first 50 values for each customer however, it calculates distance using the Pearson distance rather than the Euclidean distance.

When the number of input variables surpasses three i.e. in methods 3 to 6, due to the large number of dimensions the results can no longer be depicted graphically, however the same principles remain that the distance between each point and the cluster centroids are minimised and the cluster centres are set to minimise the average distance between each point and its centre. However, the distance between each point and its cluster centre can be calculated using a variety of methods.

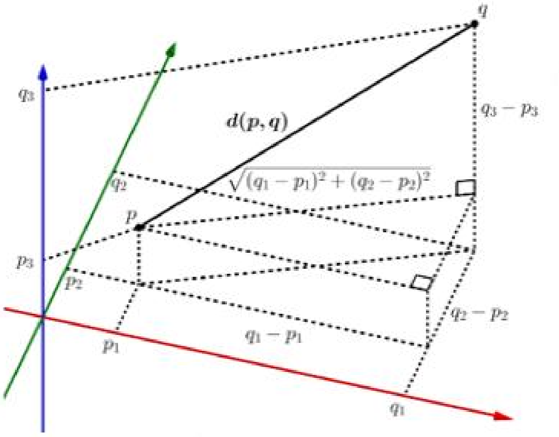
Different distance calculations

The two methodologies explored differ by the calculation used to identify which centre each customer is closest to. The two calculations used are the Euclidean distance and the Pearson correlation distance.

Euclidean distance

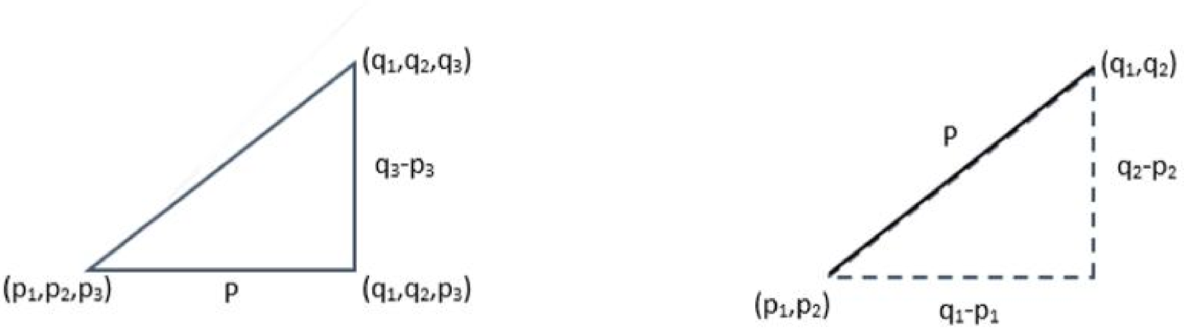
This Euclidean distance is calculated as the true straight-line distance between two points. The following example details how this calculation looks in three-dimensional space. Considering two points p and q where p is situated at (p1, p2, p3) and q is situated at (q1, q2, q3) as shown in Figure 17.

Figure 17: How to calculate Euclidean distance



The equation for the Euclidian distance is derived by using Pythagoras' Theorem. In 3D space we form a right-angled triangle between the two points as shown in Figure 18. The vertical line has length z_2-z_1 as the x and y co-ordinates are the same. The two points on the horizontal have the same z coordinate therefore can be looked at in the x-y plane alone. This results in a second triangle that can be used to calculate this horizontal distance. The horizontal distance of the second triangle is given by x_2-x_1 and the vertical distance by y_2-y_1. This leads to the length of the hypotenuse (P) equalling $\sqrt{(q_1-p_1)^2+(q_2-p_2)^2}$ which is equal to the horizontal distance in the first triangle. Using this value and Pythagoras' theorem we can calculate the Euclidean distance in 3D space as $\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+(q_3-p_3)^2}$.

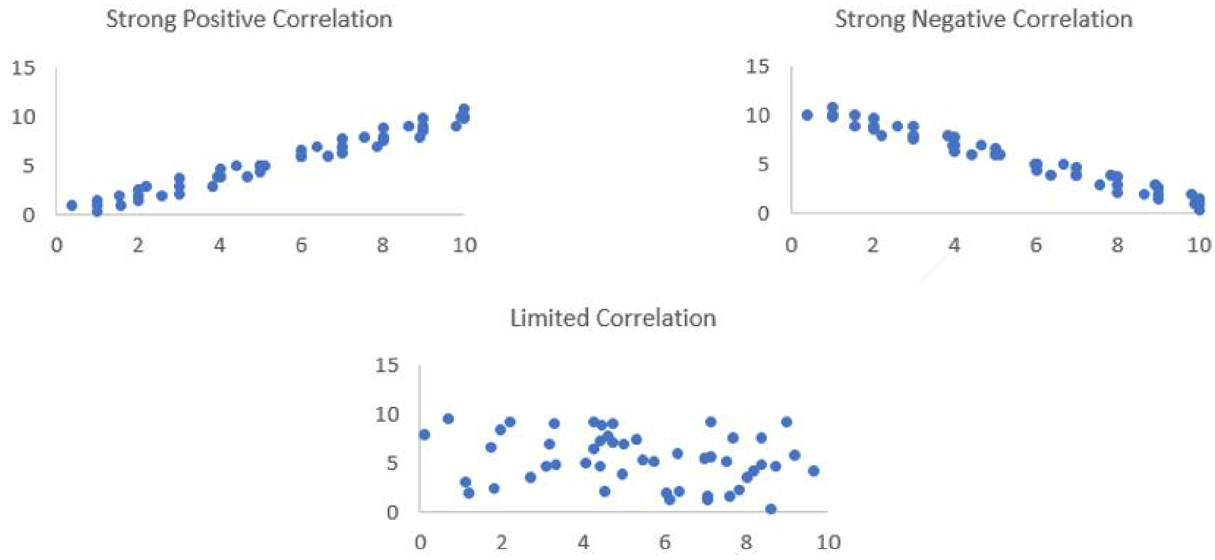
Figure 18: Further calculations explain how to calculate Euclidean distance



Pearson Correlation Distance

The Pearson correlation distance is a measure of the linear correlation between metrics. The purpose of a linear correlation is to calculate whether, and how strong, a relationship there is between two sets of variables. It can show whether there is a positive, negative or no correlation. Examples of these scenarios can be shown by using a scatter plot of the two variables against each other, Figure 19.

Figure 19: Examples of data with various strength of correlation



The value of the Pearson correlation coefficient serves as a numeric measure of the strength of the correlation. It is measured on a scale of -1 to 1 where -1 indicates a perfect negative correlation, 0 indicates no correlation and, +1 indicates a perfect positive correlation.

The Pearson distance is then calculated by subtracting the Pearson correlation coefficient from 1. Thus, the Pearson distance ranges from 0 to 2.

In the context of the customer voltage profiles used as inputs for our testing, for each customer the correlation between the first 50 values and the 50 values for each of the centroids are calculated. The Pearson coefficients found from these comparisons are then used to calculate the Pearson distances. The customer is then grouped with the centroid with the Pearson distance closest to 2.

4.3.5 Analysing Sinderby feeders separately

The Sinderby North and Sinderby South feeders were analysed separately i.e. the voltage profiles for the customers on the North feeder were fed through the K-means algorithms separately from the customers on the South feeder. This was done as we thought the results would be better by treating the feeders separately. This was not performed for Cranwood, as we did not have a process available to automatically determine which of the six feeders each customer was connected to. For Sinderby, due to the small number of customers, we could do this manually.

4.4 Results

As previously described CLNR data was used within an IPSA model to create voltage profiles for each of these customers. These voltage profiles were, subsequently, used to calculate the relevant input variables required by each methodology. The method used to calculate these input values and the results of the K-means clustering are described below.

The clustering methodology segments the customers into three groups however, it does not identify which phase each cluster represents. To test the accuracy of each method each cluster is assigned one of the phases. This leads to 6 possibilities shown in Table 5.

Table 5: Possible combinations of Phases to be assigned to each cluster

Cluster 1	Cluster 2	Cluster 3
Red	Yellow	Blue
Red	Blue	Yellow
Yellow	Red	Blue
Yellow	Blue	Red
Blue	Red	Yellow
Blue	Yellow	Red

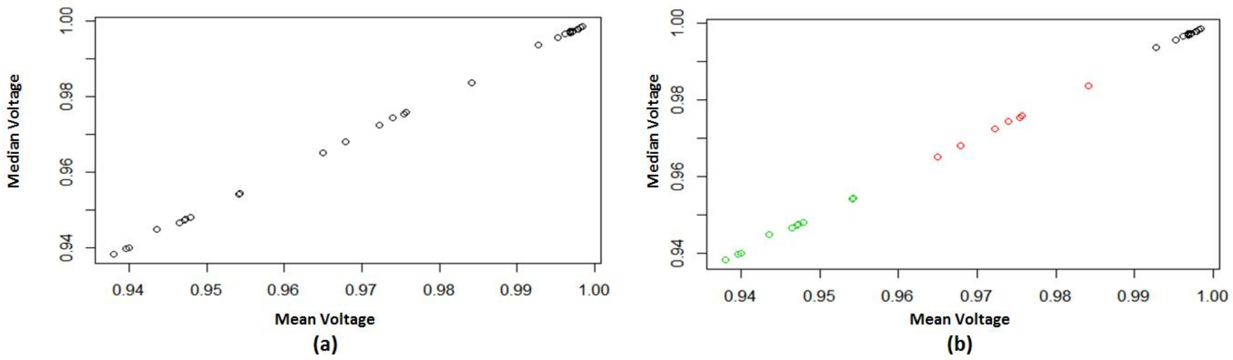
Considering each of these scenarios in turn the proportion of customers that are correctly assigned can be calculated, leading to six percentage values. The maximum and average values were then both calculated with the average taken as the accuracy of the methodology. A worked example is shown in Method 1: Mean and Median. In subsequent results, only the maximum value is shown

4.4.1 Method 1: Mean and Median

The first K-means method uses the mean and median values (i.e. the mean and median of 1000 voltage 'snapshots') for each customer as its input. Thus, on a feeder by feeder basis each customer's half hourly voltage values were averaged and the median value found. The K-means clustering algorithm was then used to group the customers appropriately. A plot of the results for the South Feeder is shown in Figure 20(a).

The results of applying the K-means algorithm in this scenario are shown in Figure 20(b) with each cluster being identified by a different colour.

Figure 20: Process conducted by K-means clustering algorithm



The results for each customer after completion of the clustering on the South Feeder are shown in Table 6.

.

Table 6: Results of clustering method 1 on the South Feeder - The table shows the actual phase colour of each customer and the cluster each customer was grouped into

Customer Number	Phase	Cluster	Customer Number	Phase	Cluster	Customer Number	Phase	Cluster
CNPT-00973839	BLUE	1	CNPT-00978234	RED	2	CNPT-00965886	RED	3
CNPT-00977617	YELLOW	1	CNPT-00930647	RED	2	CNPT-00916450	RED	3
CNPT-00957074	YELLOW	1	CNPT-00976592	RED	2	CNPT-00930648	RED	3
CNPT-00953460	BLUE	1	CNPT-00998641	RED	2	CNPT-00987902	RED	3
CNPT-00994326	YELLOW	1	CNPT-00973110	RED	2	CNPT-00928320	RED	3
CNPT-00962885	YELLOW	1	CNPT-00924691	RED	2	CNPT-01520084	RED	3
CNPT-00991144	RED	1	CNPT-00964965	RED	2	CNPT-01494924	RED	3
CNPT-01498238	BLUE	1	CNPT-00983391	RED	2			
CNPT-00919786	BLUE	1	CNPT-00908699	RED	2			
CNPT-00998176	BLUE	1	CNPT-00981009	RED	2			
CNPT-00942341	BLUE	1						
CNPT-00902199	BLUE	1						
CNPT-01520084	YELLOW	1						
CNPT-01520084	BLUE	1						
CNPT-00903862	BLUE	1						

Calculating the accuracy of the method is completed as previously described. Each permutation of Red, Blue and Yellow is assigned to each cluster and the number of correctly assigned customers in each scenario is identified, which is then converted to a percentage. The outcomes of these calculations are shown in Table 7.

Table 7: Accuracy calculations for each permutation of phase assignment on Sinderby South Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	1	31	3%
Red	Blue	Yellow	1	31	3%
Yellow	Red	Blue	15	17	47%
Yellow	Blue	Red	12	20	38%
Blue	Red	Yellow	19	13	59%
Blue	Yellow	Red	16	16	50%

The highest accuracy is achieved when cluster 1 is assigned to blue, cluster 2 to red and cluster 3 to yellow (59%). However, cluster 1 only has one customer connected to the red phase. This explains the low scores for both occasions where cluster 1 is assigned as customers connected to the red phase. The average accuracy is calculated to be 33%.

The tests were also completed on the North Feeder. Of the 16 customers 9 are connected to the blue phase, 4 to the yellow phase and 3 to the red phase.

Table 8: Accuracy results for method 1 using customers on the Sinderby North Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	2	14	13%
Red	Blue	Yellow	3	13	19%
Yellow	Red	Blue	2	14	13%
Yellow	Blue	Red	2	14	13%
Blue	Red	Yellow	12	4	75%
Blue	Yellow	Red	11	5	69%

The two scenarios where the clustering produces the most accurate results is when cluster 1 is assigned to the blue phase, Table 8. This is due to cluster 1 containing 8 customers all of which are connected to the blue phase (Table 9). However, the remaining two clusters are not unique and contain a variety of phase connections.

Table 9: Results of clustering method 1 on the North Feeder - The table shows the actual phase colour of each customer and the cluster each customer was grouped into

Customer	Phase	Cluster
CNPT-01596783	BLUE	1
CNPT-01425745	BLUE	1
CNPT-00988267	BLUE	1
CNPT-00929869	BLUE	1
CNPT-01492063	BLUE	1
CNPT-01516259	BLUE	1
CNPT-01585009	BLUE	1
CNPT-01428755	BLUE	1
CNPT-00982812	RED	2
CNPT-00951731	RED	2
CNPT-01581989	BLUE	2
CNPT-01443200	YELLOW	2
CNPT-00904064	YELLOW	2
CNPT-01551394	YELLOW	3
CNPT-01583548	YELLOW	3
CNPT-01427132	RED	3

From the results of both feeders this method has limited accuracy. In both cases it can segment the majority of customers that are on the most commonly connected phase but struggles to separate out the phases with limited customers connected.

The same clustering approach was then conducted using data from the Cranwood network. Utilising all 1000 data points for each of the 664 customers the mean and median were calculated to be used as inputs to the K-means algorithm. The results showed that in all permutations of the phase assignment to different clusters there were more incorrectly assigned customers than those correctly assigned, Table 10.

.

Table 10: Accuracy of clustering Cranwood customers using method 1

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	276	388	42%
Red	Blue	Yellow	200	464	30%
Yellow	Red	Blue	302	362	45%
Yellow	Blue	Red	151	513	23%
Blue	Red	Yellow	237	447	36%
Blue	Yellow	Red	162	482	24%

4.4.2 Method 2: Mean, Median and Mode

In the second method tested the additional variable of the mode of all 1000 data points for each customer is included as an input to the clustering algorithm. This changes the clustering process from a two dimensional to a three-dimensional problem.

Repeating the testing conducted in Method 1 each customer is assigned to a cluster. The accuracy of the clustering is again calculated by assigning each possible combination of phases to the clusters and calculating the proportion of customers who have been correctly assigned. The results for the South Feeder are shown in Table 11.

Table 11: Accuracy calculations for clustering customers on the Sinderby South Feeder using method 2

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	1	31	3%
Red	Blue	Yellow	2	30	6%
Yellow	Red	Blue	14	18	44%
Yellow	Blue	Red	11	21	34%
Blue	Red	Yellow	20	12	63%
Blue	Yellow	Red	16	16	50%

The results again show quite a variation in the accuracy of findings dependent on which phase each cluster is assigned to however, the maximum result for this method is only 63% suggesting that the results of including the mode detract from the performance. This result equates to only 20 of the 32 customers being assigned to the correct phase.

The results for the North Feeder were unchanged from the previous method, Table 8, with all customers being sorted into the same clusters as when the mode is not included.

The results for Cranwood using the mean, mode and median as inputs to the clustering algorithm were a maximum accuracy of 45% which corresponds to the successful phase identification of 301 of the 664 customers, Table 12.

Table 12: Accuracy of clustering customers on the Cranwood network using method 2

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	196	468	30%
Red	Blue	Yellow	268	396	40%
Yellow	Red	Blue	152	512	23%
Yellow	Blue	Red	301	363	45%
Blue	Red	Yellow	167	482	25%
Blue	Yellow	Red	244	435	37%

4.4.3 Method 3: Mean, Median, Mode and Standard deviation

In the third method tested the additional variable of the standard deviation of all 1000 data points for each customer is included as an input to the clustering algorithm. This changes the clustering process from a three-dimensional problem to a four-dimensional one.

Table 13: Accuracy results for the Sinderby South Feeder using clustering method 3

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	15	17	47%
Red	Blue	Yellow	12	20	38%
Yellow	Red	Blue	18	14	56%
Yellow	Blue	Red	1	31	3%
Blue	Red	Yellow	16	16	50%
Blue	Yellow	Red	2	30	6%

The results for the South Feeder, Table 13, show that the inclusion of the standard deviation again results in similar results to the two previous methodologies with only a small number of customers changing cluster. The percentages have a smaller range of values because there are two clusters that are predominantly red (17 out of 18 customers in clusters 1 and 2 are red phase customers). This leads to either 7 or 10 customers being correctly assigned for scenarios where cluster 1 is assigned to red and where cluster 2 is assigned to red respectively which drives the similarity in accuracies for each of these scenarios. For the North Feeder the results of the first two methods are again replicated for this method.

For Cranwood, the maximum accuracy achieved for this data set was 45%. The range of results for the accuracy calculations was lower than in the Sinderby tests showing that each cluster contains a range of customers connected to each phase, Table 14.

Table 14: Accuracy of clustering customers on the Cranwood network using method 3

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	197	467	30%
Red	Blue	Yellow	269	395	41%
Yellow	Red	Blue	152	512	23%
Yellow	Blue	Red	300	364	45%
Blue	Red	Yellow	167	482	25%
Blue	Yellow	Red	243	436	37%

4.4.4 Method 4: Mean, Mode, Median, Standard Deviation, Maximum, Minimum, and Variance

The fourth method tested increases the number of variables used as input to the clustering algorithm to seven due to the inclusion of the maximum, minimum and variance. The results for the South Feeder for this methodology were identical to those found in Method 3 as displayed in Table 13.

Whereas the previous three methods have produced the same outputs for the North Feeder, method 4 leads to a variation in the outcomes. The results remain that one of the clusters is predominantly made up of blue phase customers (Cluster 3 has 7 blue phase customers and 1 yellow phase customer) however, the remaining two clusters vary compared to previous results. Cluster 2 only contains two customers, one connected to the red phase and one to the yellow phase and Cluster 1 contains 4 yellow phase customers and 1 customer for each of the other phases. This leads to the accuracy results shown in Table 15.

Table 15: Accuracy results for the Sinderby North Feeder clustered using method 4

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	8	8	50%
Red	Blue	Yellow	2	14	13%
Yellow	Red	Blue	12	4	75%
Yellow	Blue	Red	6	10	38%
Blue	Red	Yellow	2	14	13%
Blue	Yellow	Red	2	14	13%

For Cranwood, again, the accuracies are below 50% for all 6 permutations of phase assignment to each of the clusters as shown in Table 16.

Table 16: Accuracy of clustering customers on the Cranwood network using method 4

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	290	374	44%
Red	Blue	Yellow	237	427	36%
Yellow	Red	Blue	272	392	41%
Yellow	Blue	Red	172	492	26%
Blue	Red	Yellow	202	444	30%
Blue	Yellow	Red	155	527	23%

4.4.5 Method 5: First 50 values of each customer

The fifth method tested uses only the first 50 data points for each customer in contrast to the previous four methods where each descriptive variable is calculated using all 1000 data points. These first 50 data points are used as the inputs to the K-means algorithm for each customer.

For both the Sinderby North and South Feeder this method results in the same clustering results as using the first method where the inputs are the full set of 1000 mean and median values.

For Cranwood the results were slightly changed from previous testing however, the accuracy scores remained comparable to the previous methods tested with a maximum accuracy of 44%, Table 17.

Table 17: Accuracy of clustering customers on the Cranwood network using method 5

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	256	408	39%
Red	Blue	Yellow	250	414	38%
Yellow	Red	Blue	290	374	44%
Yellow	Blue	Red	172	492	26%
Blue	Red	Yellow	236	476	36%
Blue	Yellow	Red	124	492	19%

4.4.6 Method 6: First 50 differences for each customer

Like in the previously described method the sixth method tested uses only the first 50 data points for each customer. To re-iterate, the data points are a contiguous set of time series half-hour measurements. The difference in voltage between each pair of time points is calculated as shown in Table 18.

Table 18: Demonstration of the difference calculations used in method 6

Voltage 1 (pu)		Voltage 2 (pu)		Voltage 3 (pu)		Voltage 4 (pu)		Voltage 5 (pu)
0.99982	Diff 1 (pu)	0.99985	Diff 2 (pu)	0.99983	Diff 3 (pu)	0.99981	Diff 4 (pu)	0.99983
	0.00003		-0.00002		-0.00002		0.00002	

The first 50 of these differences are then used as the inputs to the K-means algorithm. Using these values initially produces one outlier i.e. a value that is clustered on its own. Removing this value and repeating the clustering produces the results displayed in Table 19.

Table 19: Accuracy of Method 6 using customer data from the Sinderby South Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	3	28	10%
Red	Blue	Yellow	2	29	6%
Yellow	Red	Blue	16	15	52%
Yellow	Blue	Red	7	24	23%
Blue	Red	Yellow	21	10	68%
Blue	Yellow	Red	13	18	42%

Running the same analysis on the Sinderby North Feeder, results in a maximum accuracy of 81% where only three customers are assigned to the incorrect cluster, Table 20. The three incorrectly assigned customers are all found within cluster 2 which should be all yellow phase customers however, it also includes one red phase customer and two blue phase customers. This does mean though that two of the clusters contain only customers from one phase.

Table 20: Accuracy of Method 6 using customer data from Sinderby North Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	13	3	81%
Red	Blue	Yellow	4	12	25%
Yellow	Red	Blue	8	8	50%
Yellow	Blue	Red	2	14	13%
Blue	Red	Yellow	1	15	6%
Blue	Yellow	Red	4	12	25%

As with the other five methods, the results for Cranwood lead to lower accuracy levels than for Sinderby. The maximum accuracy recorded for Cranwood, for this method was 46%, Table 21.

Table 21: Accuracy of clustering customers on Cranwood using method 6

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	199	465	30%
Red	Blue	Yellow	304	360	46%
Yellow	Red	Blue	234	430	35%
Yellow	Blue	Red	165	499	25%
Blue	Red	Yellow	300	594	45%
Blue	Yellow	Red	126	308	19%

4.4.7 Method 7: First 50 values of each customer using Pearson distance

As with method 5, method 7 uses only the first 50 data points for each customer. These first 50 data points are used as the inputs to the K-means algorithm for each customer however, in contrast to method 5 this method uses the Pearson distance to calculate the distances to be minimised to choose which cluster each customer should be grouped in as opposed to the Euclidean distance used previously.

The first attempt at running the algorithm resulted in one of the clusters only containing one customer (as in method 6). This suggests that this customer is an outlier in comparison to the remaining customer data. Removing this outlier and recalculating the results we find that the best scenario results in 71% of customers being correctly grouped, Table 22. The results lead to cluster 3 only containing customers connected to the red phase and cluster 1 containing customers predominantly connected to the blue phase. However, cluster 2 contains a mixture of all three customer phases and only two of the five customers connected to the yellow phase.

Table 22: Accuracy results for method 7 using customer data from the South Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	3	28	10%
Red	Blue	Yellow	1	30	3%
Yellow	Red	Blue	8	23	26%
Yellow	Blue	Red	14	17	45%
Blue	Red	Yellow	14	17	45%
Blue	Yellow	Red	22	9	71%

The Sinderby North Feeder also had a cluster with only one value however, due to the limited number of customers on the feeder this customer was retained within the analysis. A maximum accuracy was achieved when cluster 1 was assigned to blue, cluster 2 to red and cluster 3 to yellow. In this scenario only 3 customers were incorrectly grouped together resulting in an accuracy of 81%. However, cluster 3 does include customers connected to each of the 3 phases, Table 23.

Table 23: Accuracy results for method 7 using customer data from the North Feeder

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	1	15	6%
Red	Blue	Yellow	4	12	25%
Yellow	Red	Blue	2	14	13%
Yellow	Blue	Red	2	14	13%
Blue	Red	Yellow	13	3	81%
Blue	Yellow	Red	10	6	63%

Using this method to test the data from Cranwood produced the highest maximum accuracy score for this network. The maximum accuracy was 48% equating to the successful assignment of 318 of the customers on the network, Table 24.

Table 24: Accuracy of clustering customers on the Cranwood network using method 7

Cluster 1	Cluster 2	Cluster 3	Correct	Incorrect	Percentage
Red	Yellow	Blue	164	500	25%
Red	Blue	Yellow	223	441	34%
Yellow	Red	Blue	271	393	41%
Yellow	Blue	Red	318	346	48%
Blue	Red	Yellow	182	482	27%
Blue	Yellow	Red	170	494	26%

4.4.8 Defining the starting centroids

To try and improve the accuracy of the results generated by the above methodologies, the above methodologies were retested with the centroids predefined. The values to define each of the three centroids were chosen to be the values of one customer from each phase i.e. centroid 1 was set to the values of a customer connected to the red phase, centroid 2 was set to the values of a customer connected to the yellow phase and centroid 3 set to the values of a customer connected to the blue phase.

The above methods for clustering all customers were then repeated. However, conducting these tests resulted in no discernible changes to which clusters each customer was assigned to.

4.5 Conclusions

We could not find a method that was suitable for phase identification. It could be that this is due to the input consumption data from CLNR being half-hourly. The paper [S105][4] did find that accuracy became worse when time steps were increased from 10 seconds to 5 minutes. The default time step for recording average voltage measurements on GB SMETS2 meters is 30 minutes, although these can be set to lower values. Another reason could be that in the IPSA model the voltages across the three phases on the slack bus (located at the 11kV busbar on the 11kV/LV transformer) are set to the same value. Any natural variation in the voltage across the phases would help in the clustering process.

We originally intended to examine the effects of smart meter voltage measurement accuracy on the phase identification methods. This was not pursued as the methods were unsuccessful with no error assigned.

There is currently a problem with the SMETS2 standard in that the time period over which the time step is defined does not have to be clock synchronised, unlike energy consumption. This means that one smart meter could be recording an average between say 11:06 and 11:36, whilst a neighbouring smart meter could be recording between 11:23 and 11:53. This problem has been observed on the small amount of SMETS2 meters deployed on Northern Powergrid's network. However, this observation would not have affected the findings of this assessment as the voltage profiles used were calculated.

In order to develop thinking in this area, it is suggested that:

- The analysis conducted in this project is repeated using sampled data that is more reflective of times of the day / year when network demand is higher and voltage drop on the network will be higher. In this scenario voltage variations along a feeder will be more pronounced and it should be easier to detect patterns.
- It may be worth re-examining phase identification methods using 'real' smart meter voltage data, where there are a large number of smart meters deployed on a representative sample of LV feeders.
- Any future work needs to consider the voltage variations on individual feeders rather than combination of feeders supplied from the same substation. Where the LV feeder comprises branches then each branch could be analysed individually.
- The effects of the smart meter voltage measurement accuracy on the phase identification methods be assessed.

[4] Arya, V., Mitra, R., 2013. "Voltage-based clustering to identify connectivity relationships in distribution networks". In: Proceedings of 4th IEEE International Conference on Smart Grid Communications. https://doi.org/10.1109/ SmartGridComm.2013.6687925.

5 SMA 3 – A statistical approach to generate a probability distribution of load at a given time for customers based on a suitable range of demand characteristics

5.1 Aim

The aim of this section of the work was to produce probability density functions (PDFs) and cumulative density functions (CDFs) that allow the estimation of the most likely values of the shape factors used to define the parametric distributions that model the consumption characteristics of a random group of customers for a given time of day and season.

5.2 Background Research

The research undertaken has provided valuable guidance on how measured data can be segmented to individual customer's consumption data.

[S1] The CLNR study looks to better understand customer demand by profiling customers by three socio-economic classes (affluent, comfortable and adversity). This work can be developed further to create probability distributions of load by customer group for modelling variances in load between customers.

[S5] The New Thames Valley Vision report proposes a categorisation of customers by energy demand, based on Council Tax band, profile class and presence of generation. The clustering was based on two stages, the first based on daily mean usages and the second on intra-day features. Again, this demonstrates the significant contributors to loads between customer groups. The paper also proposes a bootstrapping technique to generate sample loads from measured data, which could offer a more straightforward approach to creating probability distributions and sampling from them.

A paper (D. Toffanin, 2016, Generation of customer load profiles based on smart-metering time series, building-level data and aggregated measurements) references evidence that segmenting the day into four time periods provides a meaningful characterisation of customers' demand. These periods are selected according to the model peaks in the load distribution.

5.3 Methodology

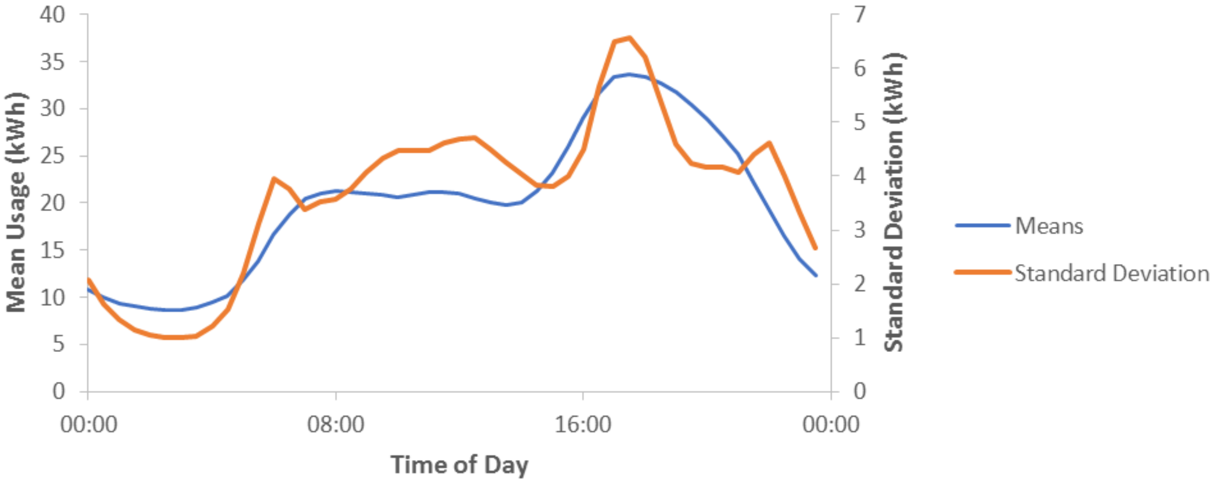
Once again, the same CLNR data was used to conduct the necessary analysis to achieve the aim of this section of work. Initially the data was uploaded into R Studio and segmented by season, Table 25.

Table 25: Season segmentation

Season	Months
Spring	March, April, May
Summer	June, July, August
Autumn	September, October, November
Winter	December, January, February

Once the data was uploaded the next step was to understand the mean consumption and variance per half hour for each of the seasons for differing numbers of customers (5, 10, 50, and 100). To generate these plots of the mean aggregated consumption for each thirty-minute segment of the day, 1000 scenarios were run for each group size and season. Subsequently plots were made of mean half-hourly consumption and standard deviation as shown in Figure 21.

Figure 21: Mean and variance of consumption in autumn for 1000 groups of 100 customers



From Figure 21 it can be seen that average consumption throughout the day is variable, therefore matching a parametric distribution with a constant mean to the entire day would not accurately model consumption. To account for this, two steps were taken- 1) segment the day into periods with similar mean consumption and 2) use distributions that can be scaled either by addition or multiplication to adjust the mean to fit the distribution for the whole period and then scaled back to produce half-hourly distributions.

Two distributions that can be scaled are the gamma distribution and the three parameter Weibull distribution. The gamma distribution can be scaled by multiplication; therefore, it is best to model segments with a changeable mean and standard deviation. The 3 parameter Weibull distribution is scaled by addition therefore, it is best to model sections of changeable mean and constant variance.

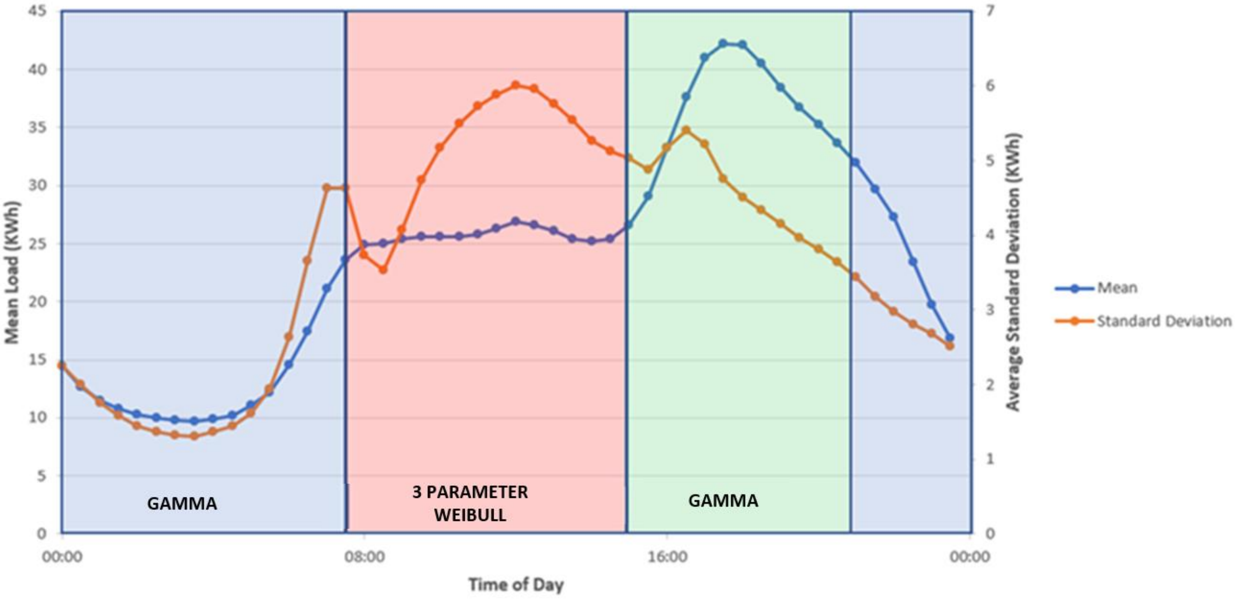
Having chosen the two possible distributions that would model the likelihood of consumption the next step was to segment the day and identify the best distribution for each period. The processes involved in completing this task are detailed below for groups of 100 customers in Winter.

5.3.1 Worked Example- Groups of 100 customers in Winter

5.3.1.1 Step 1 - Segmenting the day and choosing distributions

Having run 1000 simulations of the consumption of groups of 100 customers in winter and plotted the mean half hourly consumption, the first task was to choose the best distribution to match each section of the day. The first section to be modelled was the section where there was least variation in the mean between 07:30 and 15:00. During this period the standard deviation had some fluctuation however, as the fluctuations didn't match the fluctuations of the mean and were small the three parameter Weibull distribution was chosen. The remainder of the day has a fluctuating mean and standard deviation however, as they follow similar patterns the gamma distribution should model these the most accurately. To improve the accuracy of the fit, the peak and trough have been isolated into separate sections one from 15:30 to 22:00 and one from 22:30 to 07:00 as can be seen in Figure 22.

Figure 22: Average half hourly consumption for 100 customers in Winter



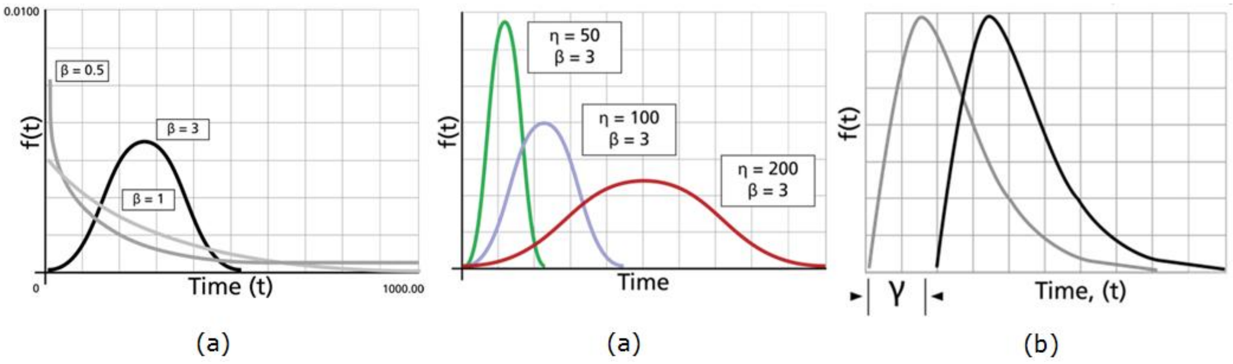
Three Parameter Weibull Distribution

The probability density function of the 3 parameter Weibull distribution is described by the equation:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$$

Where $\beta = shape\ factor\ \eta = scale\ factor, and\ \gamma = threshold\ factor$. The scale factor calculated to normalise the mean is added to γ. The effect of changing these parameters can be seen in Figure 23[5].

Figure 23: Weibull Distribution - Showing the effects of varying (a) the shape factor, (b) the scale factor and (c) the threshold factor on the probability density function



[5] http://reliawiki.org/index.php/The_Weibull_Distribution

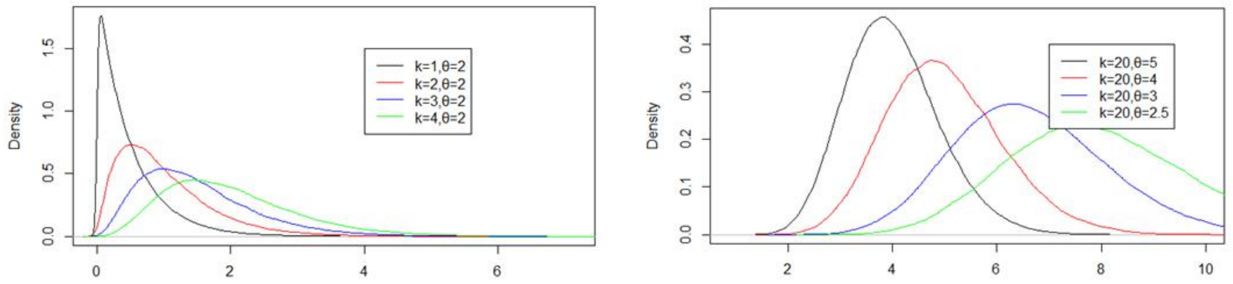
Gamma Distribution

The probability density function of the Gamma distribution is described by the equation:

$$f(t) = \frac{1}{\Gamma(k)\theta^k} t^{k-1} e^{-\frac{t}{\theta}}$$

Where $k = shape\,factor\ and\ \theta = scale\,factor$. The scale factor calculated to normalise the mean is multiplied by k. The effect of changing these parameters can be seen in Figure 24.

Figure 24: Gamma Distribution - Showing the effects of altering (a) the shape factor and (b) the scale factor on the probability density function



5.3.1.2 Step 2 - Calculating Scale Factors

Having decided which distributions are to be used, the next step is to take a single group of 100 customers and isolate the data for one of these distributions. For example, take just data for all days in winter between 07:30 and 15:00. As the distribution used to model this period will have constant mean and standard deviation we must adjust the data by a scale factor to achieve this. The scale factors calculated to make this adjustment can later be applied to the model probability density functions to return a probability density function for each half hour period. For the gamma distribution this requires a multiplier whereas the Weibull distribution requires the difference to be calculated.

To calculate the scale factors the first step is to calculate the mean aggregated load for each half hour period. These means are then averaged to find the average aggregated load for the entire period. The scale factors are then calculated by finding the multiplier or difference that maps each half hourly average on to the average for the period. The relevant equations are shown below.

$$Gamma\ Scale\ Factors = \frac{Average\ Aggregated\ Value}{Period\ Mean}$$

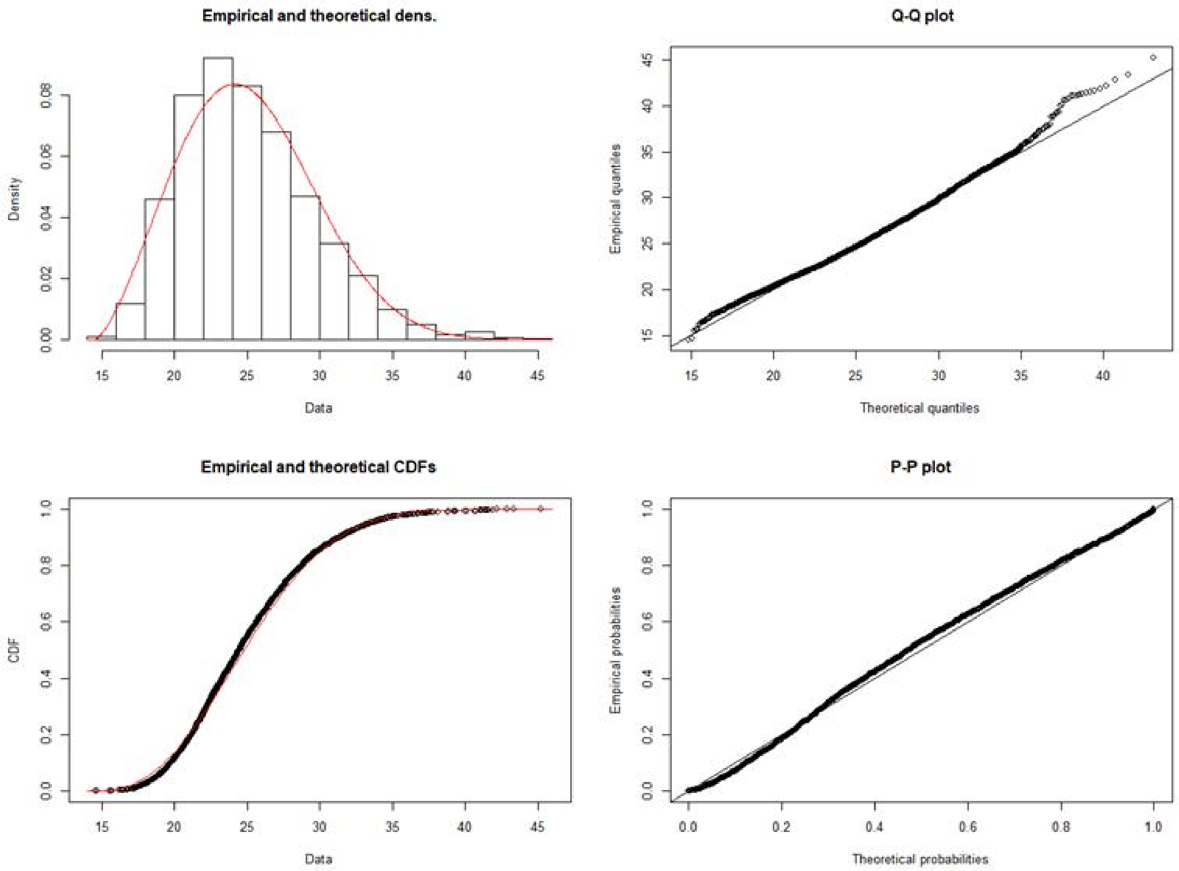
$$Weibull\ Scale\ Factors = Average\ Aggregated\ Value - Period\ Mean$$

This gives a vector of scale factors for each half an hour period. These are applied respectively to the aggregated data. This results in an adjusted data set that has constant mean for each half an hour. This adjusted data can now be matched to a model parametric distribution producing a model probability density function and cumulative distribution function for the total period. Then, to convert this to half hourly specific distributions the scaling factors can be applied to the graphs for the distribution for the period.

5.3.1.3 Step 3 - Testing the distributions

To test the chosen distributions the accuracy of the fit of the model distributions needed to be checked using the scaled aggregated load profiles. This process included a comparison of the original means and standard deviations to those of the model and looking at the PDF and CDF of the model as well as the Q-Q and P-P plots of the data and the model. The Q-Q plot compares the quartiles of the data distribution against the theoretical PDF. The P-P plot compares the empirical CDF of the data with the theoretical CDF.

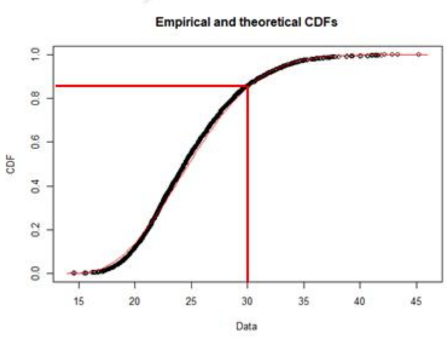
Figure 25: Example goodness of fit plots



As can be seen from Figure 25, there are four plots comparing the theoretical modelled distribution to the empirical values (the scaled load profiles for the chosen period). The first shows the histogram of the original data and compares this to the probability density function of the model distribution. The first check is to ensure the peaks and troughs of these two graphs follow similar patterns.

Below the histogram is a comparison between the cumulative distribution function of the original data and the cumulative distribution function of the model (bottom-left graph in Figure 25). A cumulative distribution function gives the probability of getting a value less than or equal to the value you are interested in (an example is given in Figure 26).

Figure 26: Example of using a Cumulative Distribution Function to calculate the probability that x is less than or equal to 30 kWh



Using a cumulative distribution plot to calculate probabilities:

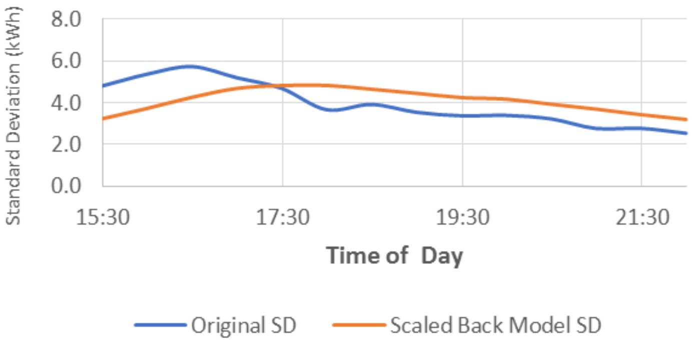
1) Draw a line from x=30kWh to the curve of the CDF
2) Find the y value of this point of intersection
3) That value of y is the probability that x is less than or equal to 30kWh

The aim is to ensure that the cumulative distribution function of the model matches that of the original data allowing the correct probability that a specific load will or won't be exceeded to be calculated. To better visualise how well the two CDFs match a P-P plot (bottom-right graph in Figure 25) is produced by plotting the probability of getting a value in the model against the probability of getting the same value in the actual data. If these values were always identical all points of the P-P plot would lie on the line y=x and the CDFs would be identical therefore, the accuracy of the model CDF can be gauged from how far points deviate from this line.

The final plot is a Q-Q plot (top-right graph in Figure 25) which shows a comparison of the quantiles of the original values against quantiles of the model. Once again, a perfect match would lead to all points being plotted on the line y=x. The Q-Q plot provides more of a focus on the fit of the tails of the distribution. In the top-right graph in Figure 25, the plotted points begin to move away from the line of y=x as the x value increases towards the 35th quantile and above. This shows that the right-hand tail (i.e. the higher values of the distribution) isn't as accurately modelled as the values for lower quantiles which, for the most part, lie on the line of y=x.

Having checked these goodness-of-fit plots the next check to make was a comparison of the original means and standard deviations to the scaled values of the mean and standard deviation of the model distribution. For each half an hour period the original mean and standard deviation are plotted on the same graph as the mean and standard deviation of the model distribution scaled by the relevant half-hourly scale factor. The mean values are reproduced identically however, there are slight variations in the standard deviation which will lead to slight inaccuracies (some overestimates and some underestimates) in the modelling of the extreme values of the data, however, because of the distributions' use within the novel analytics technique where probabilities are summed for each half hour period these errors should average out to be negligible.

Figure 27: Example of comparing original with scaled back standard deviation



5.3.1.4 Generating a distribution of shape and scale factors

Having concluded that the model distributions adequately match the actual distributions, the next stage of the process is to run the algorithms multiple times. For each individual customer group this gives an output of the shape factors that describe the model distributions and the scale factors needed to scale the probability density functions for each half an hour period. Collating the outputs from these multiple runs and plotting the probability density functions for these variables, the most common shape and scale factors and the likelihood of these factors describing the model distribution can be found. These can then be used directly within the novel analytics techniques workstream.

5.4 Results

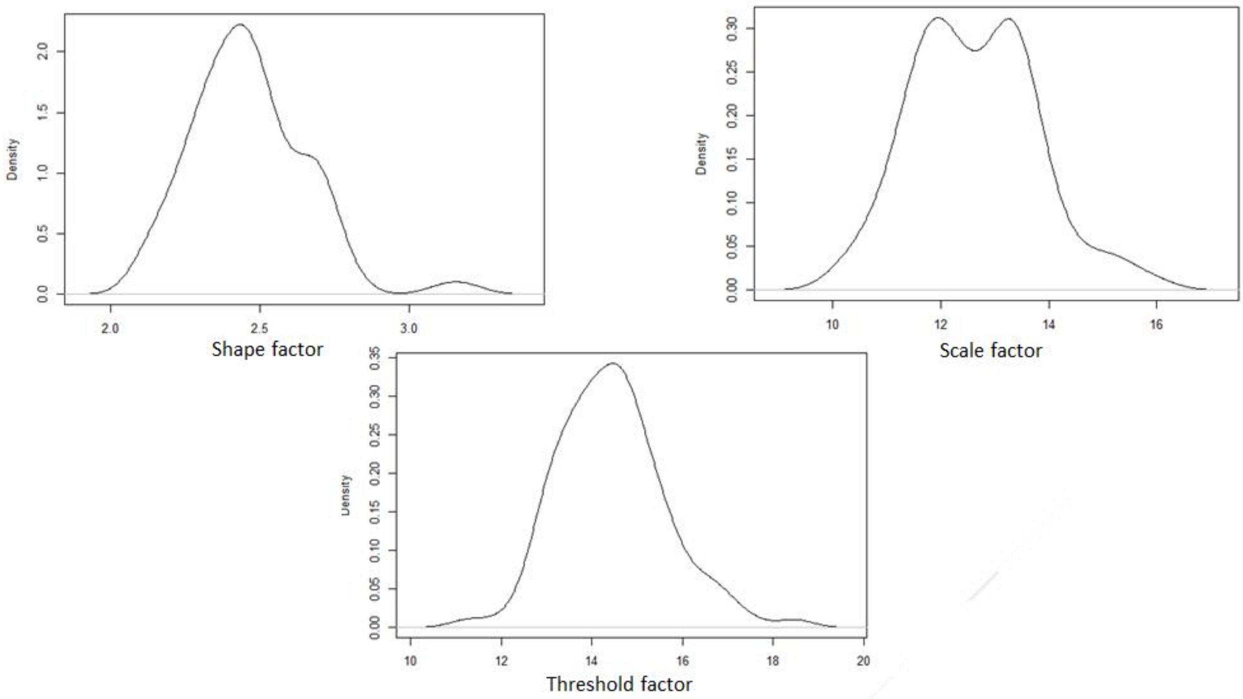
The following section details the outputs of running the algorithms to produce distributions for 100 customers across the year. These include the probability density functions for the shape factors that describe the distribution for each third of a day in each of the seasons as well as the mean half hourly scale factors to be applied to the distributions to scale back the distributions to match the original means. The same process has been replicated for groups of 5, 10 and 50 customers.

Winter - 100 simulations of groups of 100 customers

07:30 to 15:00

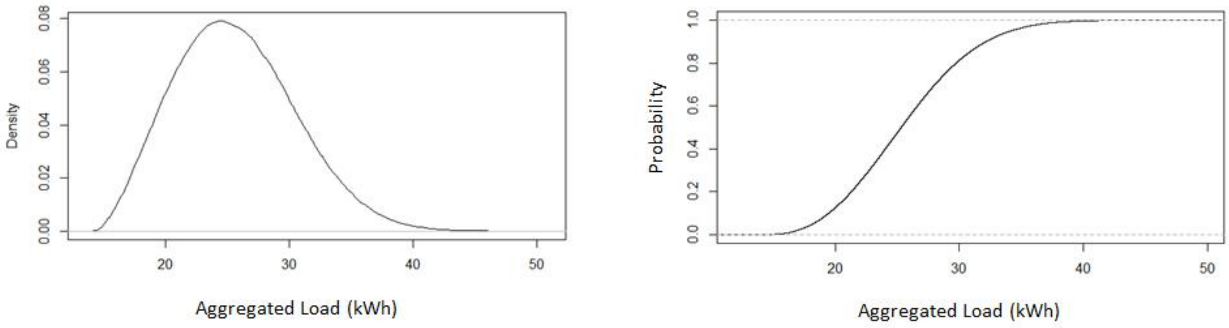
This period of the day is modelled by a three parameter Weibull distribution hence there are three probability density functions one for each of the 3 parameters. The first is used to describe the shape of the distribution (shape factor), the second the scale of the distribution (scale factor) and the third the threshold value which calculates how far along the x axis the distribution should be shifted (threshold factor).

Figure 28: Probability Density Function plots for the three shape parameters of the three parameter Weibull distribution used to model winter consumption between 07:30 and 15:00



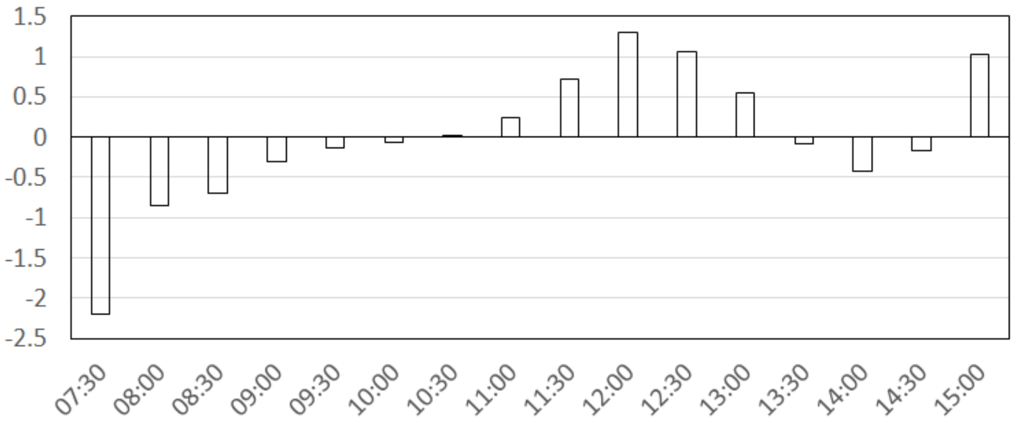
All three plots in Figure 28 are almost symmetrical around a central mean value. This lends themselves to be modelled by a normal distribution. To demonstrate how these values can be used to generate a cumulative distribution function, the mean values for each of the three factors was calculated. These factors are subsequently used to define a three parameter Weibull distribution that is most likely to represent the probability of consumption within this period. This distribution will have a probability density function and cumulative distribution function, Figure 29, which can be used within the novel analytics techniques workstream.

Figure 29: Probability density function and cumulative distribution function of the three parameter Weibull defined by the average parameters for this segment (07:30 to 15:00 in winter)



The scale factors to apply to these distributions to refine the results for each half an hour period are shown in Figure 24. For the Weibull distribution these factors are added to the threshold factor to shift the probability density function and alter the cumulative distribution function.

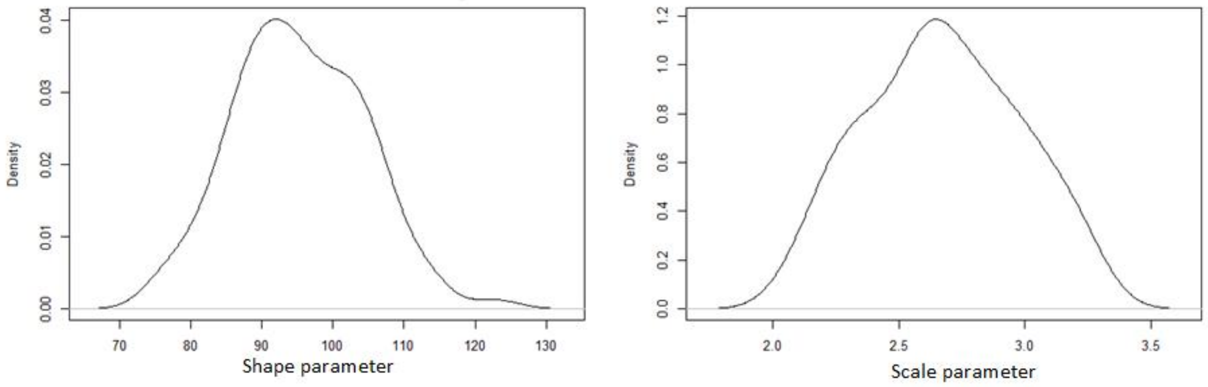
Figure 30: Average scale factors for each half hour period



15:30 to 22:00

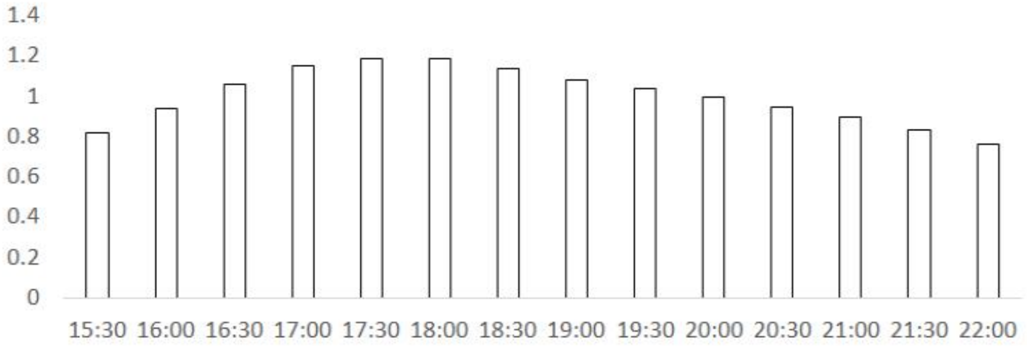
This period of the day is modelled by a gamma distribution and subsequently, has two probability density functions one for the scale parameter and one for the shape parameter. These distributions are shown in Figure 31.

Figure 31: Probability distribution functions for the shape parameters of the gamma distribution used to model consumption between 15:30 to 22:00 in winter



Once again it also important to have an understanding of the scale factors to apply to the scale parameter to refine these outputs for each half hour period. The average scale factors for these 100 runs are plotted in Figure 32.

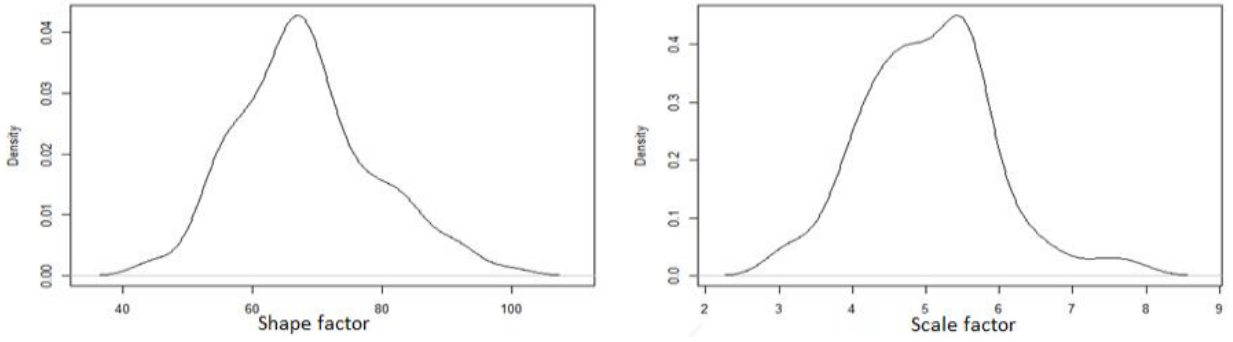
Figure 32: Average scale factor for each half hour period between 15:30 and 22:00 to scale the relevant probability density functions



22:30 to 07:00

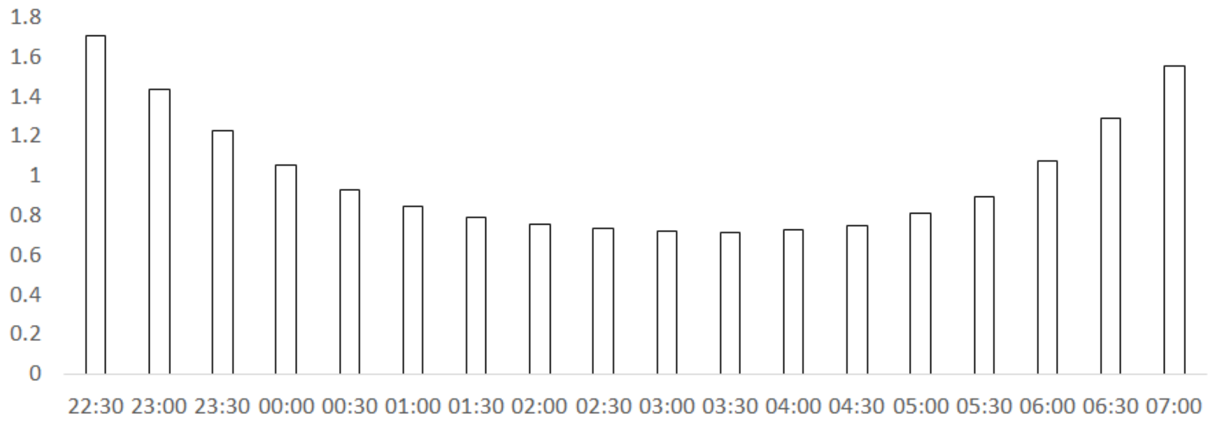
This period of the day is, once again, modelled by a gamma distribution and subsequently, has two probability density functions one for the scale parameter and one for the shape parameter, Figure 33.

Figure 33: Probability distribution functions for the shape parameters of the gamma distribution used to model consumption between 22:30 to 07:00 in winter



Compared to earlier in the day this period has a wider range of scale factor values suggesting that the consumption in the early hours of the morning and overnight is more variable than in the evening. This could be due to the larger time period and the greater variation in the average scale factors which can be seen in Figure 34.

Figure 34: Average scale factor for each half hour period between 22:30 and 07:00 to scale the relevant probability density functions



These graphs and distributions have been generated for each season and the scaling factors for each half an hour period. However, the methodology developed in the novel analysis workstream plans to reduce the number of scale factors by grouping periods with negligible differences in their scale factors.

Use within the novel analytics technique

These probability density functions can be sampled from to choose a set of shape factors that model the distribution for a set customer group. With these shape factors defined, the probability density function and cumulative distribution function of the model distribution can be easily found. From network modelling, the consumption level that, if exceeded will lead to an overloaded network can be found. Then the likelihood of this level of consumption occurring can be read directly from the cumulative distribution function of the model distribution.

Testing the effect of drawing customers from the same Mosaic class

To investigate the effects of sampling customers from the same Mosaic class, the above analysis was repeated, drawing customers from a specific customer Mosaic class, rather than at random. For each Mosaic class the average distribution parameters were then calculated for 100 simulations, these were then used to generate a cumulative distribution function for each Mosaic class for each half hour period in winter.

The results for the periods between 07:30 and 15:00 and 15:30 and 22:00 (Figure 35 and Figure 36 respectively) show that the results of drawing from each Mosaic class are very similar to each other as well as the results of drawing a random selection of customers. Figure 37 shows the results for the third period of the day (22:30 to 07:00) which shows more variation in the results with the more affluent customers slightly more likely to consume less during this period.

The above simulations were conducted to assess whether grouping customers from the same Mosaic class produce cumulative distribution functions that vary significantly from each other. If this were to be the case, then where possible attempts should be made to group customers by Mosaic class, so their aggregate consumption can be better predicted. However, these results imply that drawing results from the same Mosaic class does not make a substantial difference to the results and as such there is little benefit of drawing customers from the same Mosaic class compared to the grouping of customers regardless of their Mosaic class.

Figure 35: Cumulative distribution functions for the aggregated usage of 100 customers in Winter between 07:30 and 15:00 drawn from specific Mosaic classes

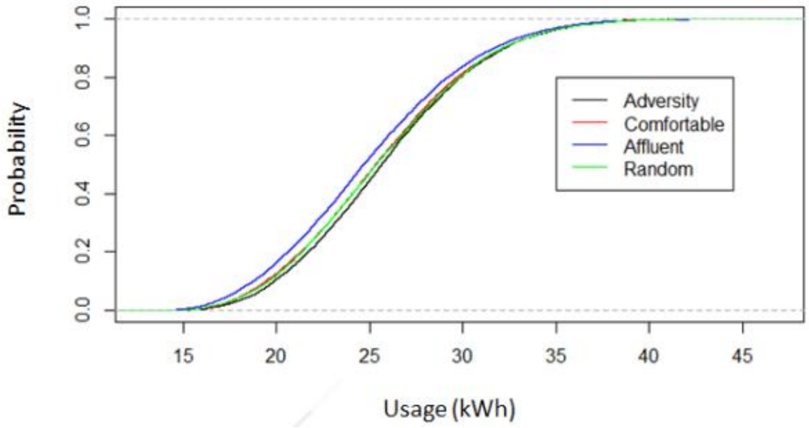


Figure 36: Cumulative distribution functions for the aggregated usage of 100 customers in Winter between 15:30 and 22:00 drawn from specific Mosaic classes

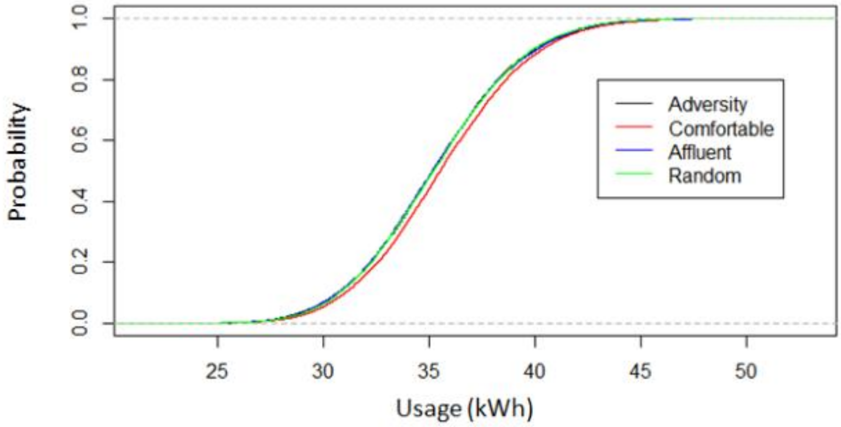
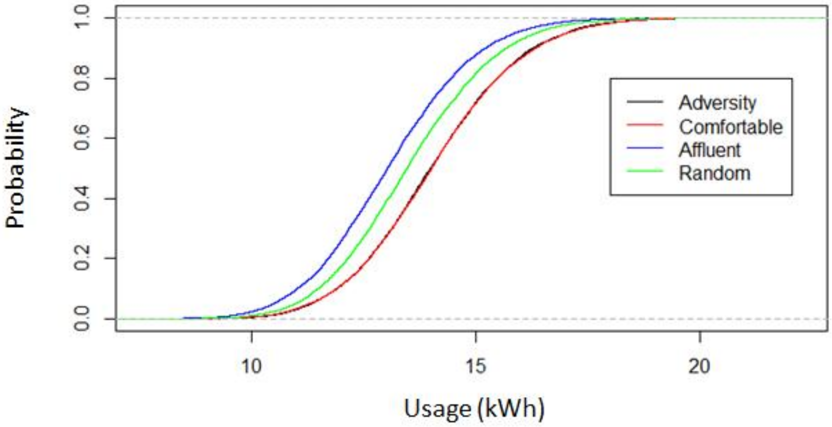


Figure 37: Cumulative distribution functions for the aggregated usage of 100 customers in Winter between 22:30 and 07:00 drawn from specific Mosaic classes



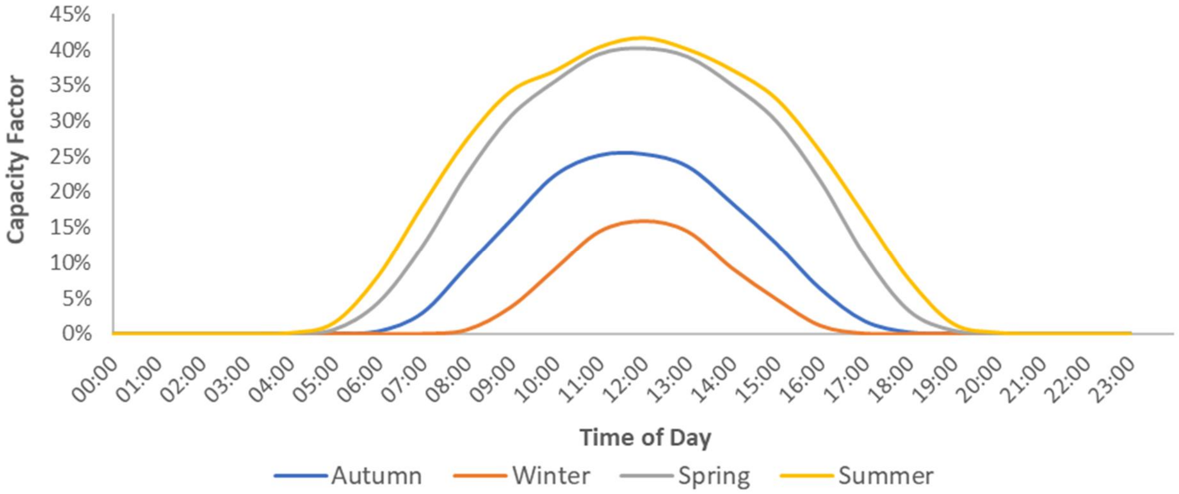
Embedded Generation

So far, this work has only considered customers' consumption. Increasingly however, houses across the UK are choosing to invest in small-scale renewable and low-carbon electricity generation technologies. Understanding the implications of generation capabilities on a network are therefore, becoming ever more important.

Within the CLNR data utilised for the above methodology there were no markers of generation. Therefore, to model embedded generation, an alternative dataset was identified - Renewables.ninja. This online tool allows you to run simulations of the hourly power output from solar power installations located anywhere in the world based upon accurate historical weather data.

Figure 38 shows, by season, the average Capacity Factor per hour in the UK. Capacity factors are calculated as the ratio of the actual output over a given period to the maximum output over that period. These averages were calculated using data from 2011 to 2013 (the period covered by the CLNR data). Data is taken from 438 sites across the UK and aggregated to a country level as described in sections 2.3 and 3.1 of Pfenniger and Staffell[6]. As would be expected the highest capacity factor and thus the highest generation is recorded during the summer months and will therefore be the focus of the further tests.

Figure 38: Average power generated across an average day for each season



Test Case – Sinderby

The Sinderby network consists of 66 customers. Conducting 1,000 simulations and sampling 66 random customer load profiles from the CLNR data allows an estimate of the average half hourly aggregate consumption of these customers to be calculated.

To see the effect of generation on the net consumption of these customers the Renewables.ninja online database was utilised. Renewables.ninja requires the following inputs; location, timeframe, capacity, system loss, tilt angle and azimuth. For this test case the location used was Sinderby, with the average power generated per summer day being calculated from data collected between 2011 and 2013. The average generation capacity of 3.5kW was calculated from the three sources of generation currently installed on the network. Losses within the PV installation are taken to be 10% and the values for tilt angle and azimuth were taken as the average values reported in the tools' documentation based upon the variety seen in real world installations (25° and 180° respectively).

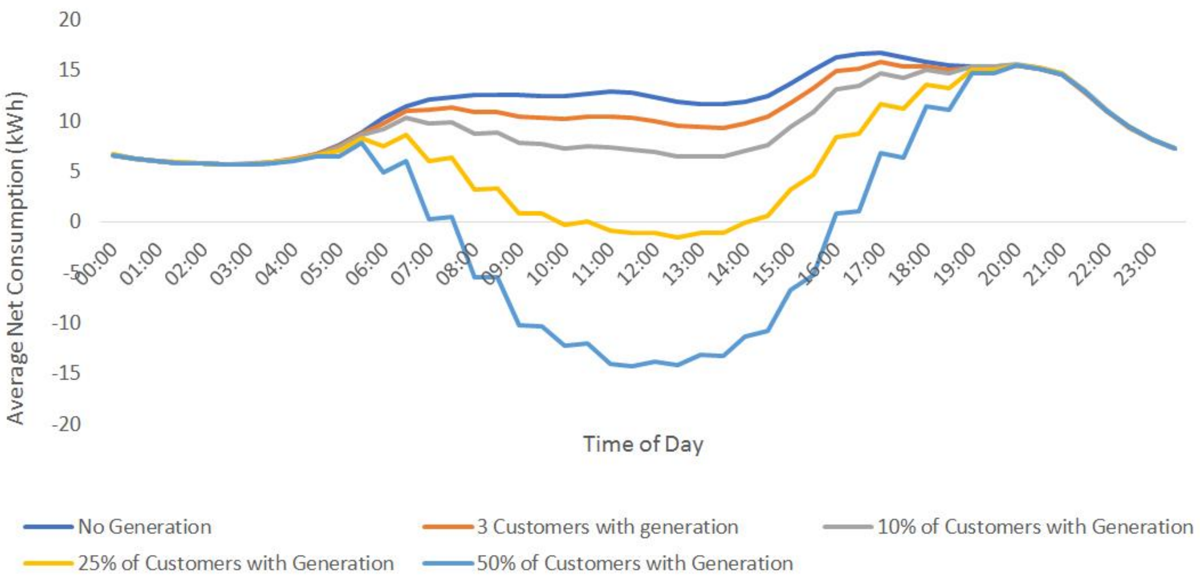
[6] https://www.sciencedirect.com/science/article/pii/S0360544216311744?via%3Dihub

The data extracted from Renewables.ninja reports the hourly generation in kW however, to align this data to the CLNR consumption data, half hourly generation data was required. To create half hourly figures, the generation capacity for each half hour was set to that of the previous hour i.e. 11:00 has a generation of 2.062 kW this will be the same for 11:30. These values were then converted to kWh so they could be subtracted from the consumption of individuals to calculate net consumption.

Four scenarios were modelled to see the effect that PV generation has on the average aggregated net consumption of the customers in Sinderby. The first is the current scenario where three of the 66 customers have generation capability; the second is that 10% of customers (7 customers) have generation capability; scenario 3 increases the proportion of customers to 25% (17 customers) and finally 50% of customers (33 customers).

The results of these 4 scenarios are shown in Figure 39. When generation is limited to only a small proportion of customers (scenario 1 and 2) the overall shape of the average profile remains similar however, as generation capacity is increased to 25 % of customers and beyond the mean consumption is no longer approximately constant and at certain time points the net consumption has become negative.

Figure 39: Average net consumption across the day of 66 customers with a varying proportion of PV generation



The previous distributions used to model the likelihood of aggregated consumption (three parameter Weibull and gamma distributions) are unable to model negative values. Therefore, the negative section will have to be modelled in a separate way. This could be using an alternative distribution or by shifting the partitions of the day to isolate all the negative values into one section. Once isolated, these values can then be negated so they are positive, modelled in the same way and then negated back to their original values.

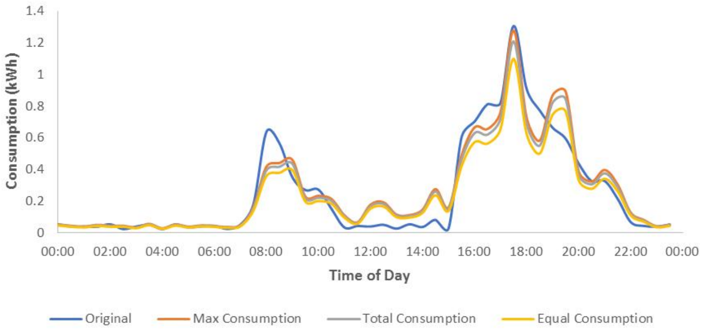
5.5 Future Developments

As these functions are to be used to predict the likelihood of extreme loads it is important that the extreme values are accurately modelled. At present, there is still the propensity for improved modelling of these extreme values. To increase the accuracy of the modelling of the right-hand tail of the probability density functions (which model the likelihood of highest consumptions are modelled) extreme value theory could be utilised. This process involves the fitting of a separate distribution to the most extreme values to improve accuracy of modelling of the tails of the distributions.

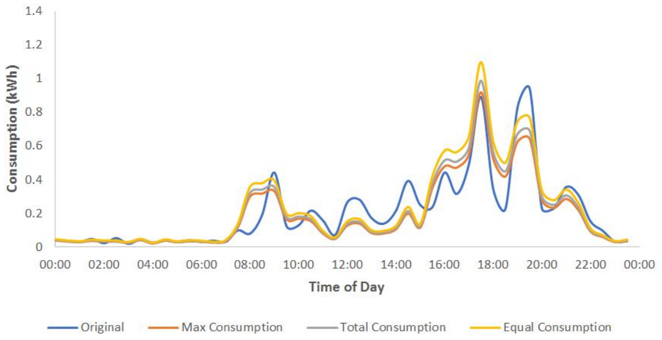
Appendix A – Results of further testing of disaggregation methods

Two Customer Disaggregation- January 3rd, 2012

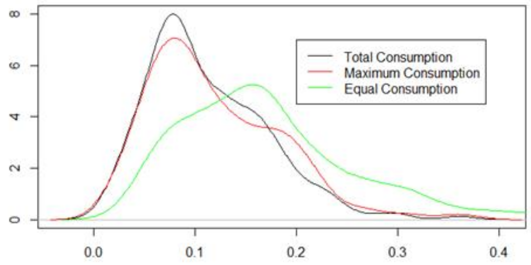
Example results of a disaggregation of customer profiles for a winter's day.



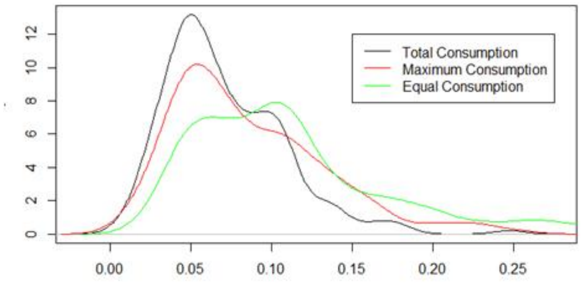
z



Probability density function of root mean square error values for January 3rd, 2012

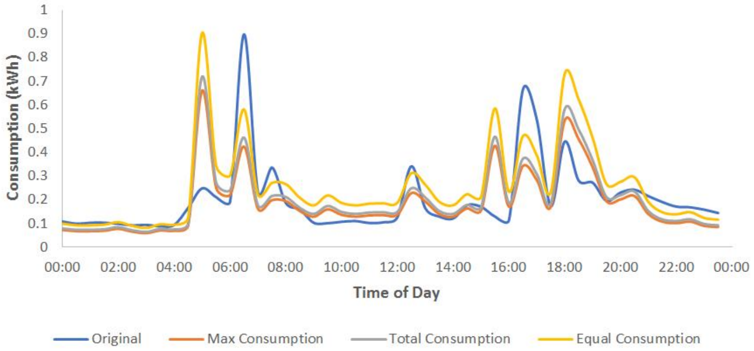


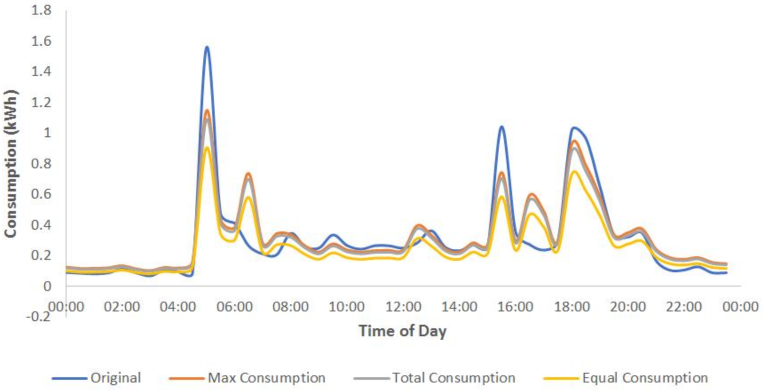
Probability density function of mean absolute error values for January 3rd, 2012



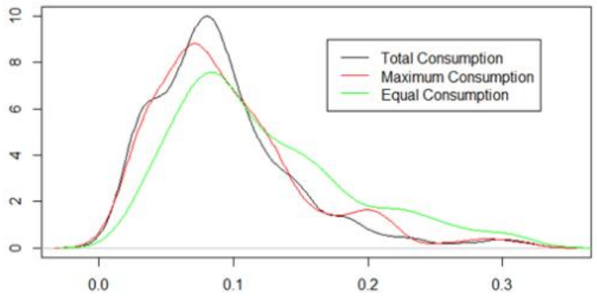
Two Customer Disaggregation - August 8th, 2012

Example results of a disaggregation of customer profiles for a summer's day.

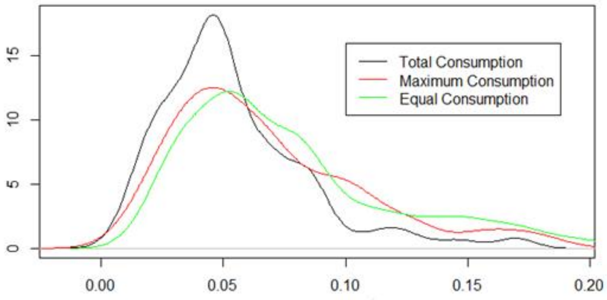




Probability density function of root mean square error values for August 8th, 2012



Probability density function of mean absolute error values for August 8th, 2012



Appendix B – Problem Statements

SMA 1.1					
Reference	SMA1.1	Version	1.0	Date	12-06-2018
Owner	A Creighton			Status	Draft
Problem Statement	Disaggregating load profiles for individual customers where the smart meter load profile data has been aggregated.				
Relevant Use Cases	1.1, 1.2, 2.1, 2.2, 3.1.1, 3.1.2, 3.2.1, 3.2.2, 4, 6				
Relevant Research	[S4] The Low Carbon London study states that datasets available to a DNO that are considered personal data under Electricity Distribution Licence Standard condition 10A are: active electricity energy import, reactive electricity energy import, maximum demand. This data is only considered personal if relating to an individual customer and if it relates to a period of less than one month. All uses of smart meter data in the report are able to make use of the smart meter data without referring to a single customers' data for less than one month. This can be achieved by: - Aggregating time series energy import data over a period of more than one month to produce consumption profiles - Aggregating time series consumption data to ensure that any network related parameter comprises at least 2 customers' data - Maximum demand registers are not reset more frequently than once a month. It is proposed that monthly maximum demand could be used to determine weighted averages that can be used to disaggregate aggregated data. [S10] The Smart Meter Aggregation Assessment Reports discussed the aggregation of two profiles, coupled with the development and implementation of DNO IT systems and/or business processes, to minimise the benefits reduction resulting from the aggregation whilst ensuring anonymity. Best performing aggregation method was then assessed against the 'visibility risk' metric to evaluate the most suitable aggregation levels. It proposes an optimum aggregation of Level 2. However, the approach to disaggregation will consider any level of aggregation, though accuracy will be impacted as the level increases.				

Summary of	This approach is to dis-aggregate smart meter data to create load profiles for individual customers. The aggregated loads at 30-minute intervals are to be divided

Approach	by the number of profiles aggregated, then adjusted upwards or downwards depending on the monthly consumption for that customer to generate a weighted average. The CLNR dataset shall be used for the analysis as it allows us to aggregate individual load profiles (which would not be available if SMETS2 data was to be used), and to use the technique documented below to dis-aggregate them and test for accuracy.
Data Description	Use the CLNR data and extract the following datasets: ▪ Datasets A: Smart meter energy consumption data at half-hourly intervals in KWh with associated unique customer identifier. ▪ Datasets B: Smart meter energy consumption data with associated unique customer identifier for a customer over a month. A review of the data has been conducted and records with poor data quality have been removed. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2b will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Network Description	An LV network model is not required for this approach.
Pre-requisites	▪ Signed-off literature review, data review and use cases. ▪ Receipt of SMETS2 DNO applicable DUIS / GBCS use cases / SMETS2 functional requirements. ▪ CLNR data to be available in an accessible form, with data of sufficient quality to derive useful insight to effectively test the approach. ▪ Input from NPG and other stakeholders on any lessons learnt regarding the applicability and usability of the data.
Tasks	The tasks required as part of this approach are as follows: ▪ Dataset to be uploaded into an analytics toolset or database with the following data items as a minimum - half hourly consumption (KWh), measurement time and date stamp, MOSAIC class, generation capability (Y/N) and unique customer identifier ▪ Create a set of aggregated load profiles (based on the individual load profiles in the CLNR dataset) for each half hour interval over a given month – these sets should be generated based on a range of consumption and generation characteristics (MOSAIC class and Generation capability - Y/N) and an aggregation level of 2, 5 and 10. ▪ Dis-aggregate the aggregated load profiles for each half-hourly interval by comparing the aggregated monthly consumption over multiple MPANs for each half-hour interval with the sum of measured monthly consumption for the equivalent MPANs. ▪ Take a weighted average of the load factor for each 30-minute interval for each MPAN, weighted on their monthly consumption. ▪ Investigate alternative methods for devising a weighted monthly consumption such as MDI. ▪ Calculate the variance between the dis-aggregated data against the original dataset to propose rules for dis-aggregation that minimises any inaccuracies within the dis-

	aggregation process. ▪ Compare aspects of the data dictionary from CLNR dataset relevant to aggregation with the data available from Siemens IP and note any additional data requirements not met by Siemens IP. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2 will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Outcomes	The outcomes from this activity are: ▪ An approach to disaggregated and adjusted consumption profiles for each customer at 30-minute intervals. ▪ Proposed rules for dis-aggregation to minimise any loss of accuracy. ▪ A view of the accuracy of this approach, together with any data characteristics that are to be accounted for to minimise any loss of accuracy. ▪ Jointly with TNEI recommend how smart metering could be effectively utilised in network design and planning by integrating with LV and multi-voltage level network modelling. ▪ Any additional data requirements not met by NPg's current systems.
Assumptions	▪ The approach is valid when number of aggregated customers (*m*) exceeds 2, though NPg have proposed an aggregation of 2 to optimise accuracy whilst maintaining data privacy. ▪ The monthly consumption per customer is used to average the aggregated load profile for each 30-minute interval, hence this approach suggests that if Household A uses twice as much as Household B during the month, the load profile for Household A is consistently twice that of Household B. The validity of this assumption will be assessed during our analysis. ▪ Data will be aggregated on extraction from DCC using the same rules as will be applied in Live Service. ▪ Monthly consumption per customer is available in a dis-aggregated form directly from the SiemensIP system. ▪ The data to be available in Siemens IP reflects the SMETS2 data objects (as provided by NPg)
Notes	The final narrative should include discussion of how success is measured and how the method to make this assessment was chosen.

SMA 1.2					
Reference	SMA1.2	Version	1.0	Date	12-06-2018
Owner	A Creighton			Status	Draft
Problem Statement	Application of aggregated load profiles at virtual nodes within an LV feeder network.				
Relevant Use	1.1, 1.2, 2.1, 2.2, 3.1.1, 3.1.2, 3.2.1, 3.2.2, 4, 6				

Cases	
Relevant Research	
Summary of Approach	Applying load profiles to virtual nodes on the LV network based on each virtual node representing a group of customers. The rules applied to create the aggregated load profiles would need to match the customer mix at the virtual node. The CLNR dataset shall be used for the analysis as it allows us to aggregate individual load profiles (which would not yet be available if SMETS2 data was to be used), and have access to information describing the generation and consumption demand characteristics.
Data Description	Use the CLNR data to create the following datasets: ▪ Datasets A: Smart meter energy consumption data at half-hourly intervals in KWh with associated unique customer identifier. ▪ Dataset B: Metadata relating to the individual customer including MOSAIC class and generation flag, to be linked to Dataset A by a unique customer identifier. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2b will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Network Description	▪ The Test LV network shall be used to test the applicability of an aggregated load profile at a virtual node using smart meter data, and test this approach against traditional methods of applying a deterministic max and min.
Pre-requisites	▪ Signed-off literature review, data review and use cases. ▪ Receipt of SMETS2 DNO applicable DUIS / GBCS use cases / SMETS2 functional requirements ▪ CLNR data to be available in an accessible form, with data of sufficient quality to derive useful insight to effectively test the approach. ▪ Input from NPG and other stakeholders on any lessons learnt regarding the applicability and usability of the data.
Tasks	The tasks that are required as part of this approach are as follows: ▪ Dataset to be uploaded into an analytics toolset or database with the following data items as a minimum - half hourly consumption (KWh), measurement time and date stamp, MOSAIC class, generation capability (Y/N) and unique customer identifier. ▪ Create a set of aggregated load profiles (based on the individual load profiles in the CLNR dataset) for each half hour interval over a given month, based on the configuration of individual customers at each (virtual) node within the test LV feeder network. ▪ Consider how this aggregated load profiles can be best applied to virtual nodes. It is expected that the level of aggregation required will be a function of the network design. ▪ Inform tnei of any requirements for the generation of virtual nodes and how these will be represented in the LV model. ▪ Document any recommendations on rules that could be applied to aggregate the load profile information on extraction from the Siemens IP system. For example, a recommendation could be to create virtual nodes of customers from the same MOSAIC

	classes to optimise the accuracy of the load profile to be applied at the node. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2 will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Outcomes	The outcome of this approach is to consider the use of aggregated consumption data directly to a (virtual) node in a LV feeder network, as opposed to having to disaggregate the data and potentially lose accuracy. The outcomes from this activity are: ▪ A method for applying an aggregated load profile to represent a level of consumption at a node on a LV feeder network. ▪ Appropriate business rules to apply when extracting data from the Siemens IP system to minimise loss of accuracy. ▪ Any additional data requirements not met by NPG's current systems. ▪ Jointly with TNEI recommend how smart metering could be effectively utilised in network design and planning by integrating with LV and multi-voltage level network modelling.
Assumptions	▪ Business rules on data aggregation could be incorporated in the Siemens IP system in accordance to this approach
Notes	We will consider how the approach impacts on the use of virtual nodes.

SMA2					
Reference	SMA2	Version	1.0	Date	12-06-2018
Owner	A Creighton			Status	Draft
Problem Statement	Determining the phase connectivity of customers on an LV network when phase connectivity records are incomplete.				
Relevant Use Cases	1.1, 2.1, 3.1.1, 3.2.1, 4, 6				
Relevant Research	[S2] A paper on the Phase Identification in Distribution Systems by Data Mining Methods analyses clustering methods, including K-means and GMM as promising candidates, but are relatively expensive to compute and therefore requires additional research to be more efficient. The approach is complex and utilises the OpenDSS simulation tool and Matlab for the analytics, and the research indicates it is likely to be too computationally extensive to be achieved in practice.				

	[S3] An algorithm is described which uses voltage profile correlation analysis. kWh as well as voltage measurements at customer smart meters along with knowledge of the service cable resistance are used to estimate the voltage profiles at Points of Common Coupling along the mains cables. The algorithm assumes a voltage drop along the whole of the mains cable i.e. no embedded generation is present. PCC voltages are compared and those with the highest correlation are assumed to be neighbours and on the same phase. The technique is mathematically simple compared to K-means, although its reliance on kWh measurements is a disadvantage. The reliance on voltage drop in the technique could be mitigated against by analysing measurements at night where no photo-voltaic generation is present. The method described shall be investigated further due to its simplicity. [S8] No new techniques are explored in this paper, although a summary of existing techniques is given. [S5] Concluded that data on phase connectivity should be gathered at source where detailed analysis is required as poor quality measured data significantly impacts the validity of any analysis. A further unpublished study 'Using grouped smart meter data in phase identification' proposes matching loads from smart meters with loads measured at different phases as a substation can assist with phase identification, however the approach requires a reasonable level of coverage and significantly reduces in applicability once the smart meter data is aggregated across different phases. Poursharif, G 2018 Investigating the Ability of Smart Electricity Meters to Provide Accurate Low Voltage Network Information to the UK Distribution Network operators, PhD Thesis, University of Sheffield The paper below describes using K-means to cluster nodes based on time-series voltage measurements. Arya, V., Mitra, R., 2013. Voltage-based clustering to identify connectivity relationships in distribution networks. In: Proceedings of 4th IEEE International Conference on Smart Grid Communications (Available on IEEE Xplore). https://doi.org/10.1109/SmartGridComm.2013.6687925.
Summary of Approach	To utilise smart meter data to identify the phase a customer is connected to by analysing voltage profile data.
Data Description	Use the CLNR data and extract the following dataset: ▪ Datasets A: Smart meter energy consumption data at half-hourly intervals in KWh with associated unique customer identifier. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2b will include a review of the content and structure of the SMETS2 data to check the data can be

	used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Network Description	▪ It is proposed to use the part of the Yorkshire unbalanced test LV network for this problem statement, if suitable.
Pre-requisites	▪ Signed-off literature review, data review and use cases. ▪ CLNR data to be available in an accessible form, with data of sufficient quality to derive useful insight to effectively test the approach. ▪ Input from NPG and other stakeholders on any lessons learnt regarding the applicability and usability of the data. ▪ Yorkshire unbalanced IPSA test network available to generate a set of synthesised voltage profiles with specified meter errors and time resolution.
Framework methodology	We require TNEI to implement the following to generate a synthesised set of voltage data: ▪ Apply good quality CLNR consumption time series to customers in chosen scenarios. ▪ Run the data through the model to create voltage profiles along the feeder network and at the substations (voltage measurements are required at each of the three transformers at the sub-station). ▪ Conduct sensitivity analysis to reflect the errors in the meters (fixed and random) and measurements at different time resolutions. (see below). For the profile correlation analysis: ▪ Perform a correlation between the transformer voltage and node voltage ▪ Select the phase allocation scenario that has the highest correlation. For the correlation analysis the following correlation methods could be considered: ▪ Linear correlation for the series ▪ Linear correlation of the differences (changes in voltages between time steps) ▪ Applying a rank correlation The appropriateness of the methods shall be considered, taking into account varying levels of error in the smart meter voltage data (2%, 0.5% and 0.1%, based on both fixed biased error and random error) and time resolution (30mins, 10mins, 1min). The methodologies will be assessed and qualified by determining which approach best predicts which phases customers are allocated to. For example, if Method A correctly predicts 18 out of 20 households, and Method B correctly predicts 16 out of 20, then Method A is the most accurate. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2 will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report
Outcomes	The outcome of this approach is an evaluation of methods to determine the phase connectivity of customers.

	The outcomes from this activity are: ▪ An approach to use smart data for applying voltage correlation techniques to determine the phase connectivity of customers. ▪ Sensitivity analysis to investigate the effect of errors on voltage meter readings and time resolution on the applicability and accuracy of the approach. ▪ Jointly with TNEI recommend how smart metering could be effectively utilised in network design and planning by integrating with LV and multi-voltage level network modelling. ▪ The data requirements to implement this approach and any additional data requirements not met by NPg's current systems.
Assumptions	▪ The voltage profile of an individual customer is well correlated to the voltage profile on the same phase of the transformer. ▪ Voltage data will be available in a dis-aggregated form from the Siemens IP system. ▪ Following consultation with a metering consultant, NPg have stated that the voltage measurement accuracy should be better than 2%, probably better than 0.5% and 0.1% should be achievable with modern metering chips.
Notes	

SMA3					
Reference	SMA3	Version	1.0	Date	12-06-2018
Owner	A Creighton			Status	Draft
Problem Statement	A statistical approach to generate a probability distribution of load at a given time for customers based on a suitable range of demand characteristics.				
Relevant Use Cases	1.1, 1.2, 2.1, 2.2, 3.1.1, 3.1.2, 3.2.1, 3.2.2, 5, 6				

Relevant Research	The research undertaken has provided valuable guidance on how measured data can be segmented into individual customer's consumption data. [S1] The CLNR study looks to better understand customer demand by profiling customers by three socio-economic classes (affluent, comfortable and adversity). This work can be developed further to create probability distributions of load by customer group for modelling variances in load between customers. [S5] The New Thames Valley Vision report proposes a categorisation of customers by energy demand, based on Council Tax band, profile class and presence of generation. The clustering was based on two stages, the first based on daily mean usages and the second on intra-day features. Again, this demonstrates the significant contributors to loads between customer groups. The paper also proposes a bootstrapping technique to generate sample loads from measured data, which could offer a more straightforward approach to creating probability distributions and sampling from them. A paper (D. Toffanin, 2016, Generation of customer load profiles based on smart-metering time series, building-level data and aggregated measurements) references evidence that segmenting the day into four time periods provides a meaningful characterisation of customers demand. These periods are selected according to the model peaks in the load distribution.
Summary of Approach	This method produces cumulative distribution functions for reaching a specified level of consumption over certain periods for customers within defined groups, representing their demand generation and consumption characteristics.
Data Description	The CLNR dataset shall be used to develop the algorithms and to conduct an initial test and validation of the approach: ▪ Dataset A: Smart meter energy consumption data at half-hourly intervals in KWh with an associated unique customer identifier. ▪ Dataset B: Customer metadata representing demographic class (MOSAIC categories), and whether there is a generation capability (Y/N), with a key (unique customer identifier) to link to Dataset A. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2b will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.

Network Description	▪ An LV network model is not required for this analysis.
Pre-requisites	▪ Signed-off literature review, data review and use cases. ▪ Receipt of SMETS2 DNO applicable DUIS / GBCS use cases / SMETS2 functional requirements ▪ CLNR data to be available in an accessible form, with data of sufficient quality to derive useful insight to effectively test the approach. ▪ Input from NPG and other stakeholders on any lessons learnt regarding the applicability and usability of the data.
Tasks	▪ Dataset to be uploaded into an analytics toolset or database with the following data items as a minimum - half hourly consumption (KWh), Measurement time and date stamp, MOSAIC class, generation capability (Y/N) and unique customer identifier. ▪ Segment the half-hourly consumption (KWh) data by the following attributes: o Time of Year (Winter, Spring, Summer, High Summer, Autumn) o Time of Day (0630-0900, 0900-1530, 1530-2230, 2230-0630) o MOSAIC Class (Affluent [A-E], Comfortable [F-J], Adversity [K-Q]) o Generation capability (Yes, No) ▪ Create a Probability Density Function (PDF) using an appropriate technique (e.g. Kernel-density estimation) to represent half-hourly consumption for each group. ▪ Generate a Cumulative Distribution Function (CDF) to represent half-hourly consumption for each group. ▪ Using the CDF, for an individual customer, construct an exceedance expectation function - that is the expected number of time-steps per year when a specified consumption rate (*x*) is exceeded. ▪ Compare the data dictionary from CLNR dataset with the data available from NPg legacy systems and those in development, specifically e-Spatial and Siemens IP, and note any additional data requirements not due to be incorporated into NPg's systems. Assuming sufficient SMETS2 data available in Q1 2019, Phase 2 will include a review of the content and structure of the SMETS2 data to check the data can be used in an equivalent way to the CLNR data for the documented approach. If there are clear discrepancies, these will be noted in the final report.
Outcomes	This approach produces a set of probability functions using smart meter data that represent consumption at different seasons, times or day and for a range of statistically significant customer characteristics. This information can be used as a 'stepping stone' towards deriving a probabilistic approach to network modelling, or to produce useful analytics on variation in consumption. The outcomes from this activity are: ▪ A set of PDFs, CDFs and Exceedance Expectation function based on an individual customer and their consumption characteristics. ▪ A methodology to generate the above outcomes based on the expected data available. ▪ Jointly with TNEI recommend how smart metering could be effectively utilised in network design and planning by integrating with LV and multi-voltage level network modelling. ▪ Any additional data requirements not met by NPg's current systems.

Assumptions	
Notes	

<table>
<tr><th colspan="6">SMA4</th></tr>
<tr><td>Reference</td><td>SMA4</td><td>Version</td><td>1.0</td><td>Date</td><td>12-06-2018</td></tr>
<tr><td>Owner</td><td colspan="3">A Creighton</td><td>Status</td><td>Draft</td></tr>
<tr><td>Problem Statement</td><td colspan="5">A method to sample from the probability distribution function based on specific or general demand characteristics.</td></tr>
<tr><td>Relevant Use Cases</td><td colspan="5">1.1, 1.2, 2.1, 2.2, 3.1.1, 3.1.2, 3.2.1, 3.2.2, 5, 6</td></tr>
<tr><td>Relevant Research</td><td colspan="5">[S4] The LCL study discusses how to consider load diversity when considering a cluster of individuals within a node of an LV network. The demand for the group will be less than the aggregated peak loads due to averaging effects and peak coincidence, hence it is not appropriate to sum the peak demand across individual customers to produce a peak demand for the customer.
The study calculated the diversity factors based on the number of customers in each cluster and it is suggested that the conclusions in Section 3.6 are compared with diversity factors calculated for a subset of the groups.
The Peak Demand After Diversity shall be used as input to the LV Model to represent ‘peak’ demand, rather than the traditional method of applying a deterministic max and min.</td></tr>
<tr><td>Summary of Approach</td><td colspan="5">This builds on SMA3.1 and is an approach for generating ‘peak’ demand values (using smart meter data) for a defined group of individual customers with certain, known demand characteristics within an LV model node.
A sampling technique shall be developed to support the probabilistic modelling approach as part of the Novel Analysis Techniques workstream. The smart Data Analytics workstream shall test the approach developed using representative smart meter data.</td></tr>
<tr><td>Data Description</td><td colspan="5">The CLNR dataset shall be used to test the algorithms developed as part of the Novel Analysis Techniques workstream. The CLNR data will be used to create the distributions outlined in SDA3 and this problem statement proposed a methodology to sample from those distributions or other distributions created from smart meter</td></tr>
</table>

	data.
Network Description	An LV network model is not required to test an approach for sampling from a distribution.
Pre-requisites	▪ Signed-off literature review, data review and use cases. ▪ An approach developed as part of the Novel Analysis Techniques workstream to employ a probabilistic approach to network planning, rather than the traditional method of applying a deterministic max and min.
Tasks	An outcome from the Novel Analytics workstream is to define an algorithm to support a probabilistic approach to network modelling, rather than the traditional approach of applying a deterministic max and min. This approach is expected to involve the following: ▪ A joint probability distribution function (expressed as an Exceedance Expectation) for each node within the given LV network, where each node represents a cluster of individual customers with different attributes. That is, for a given *y*, what is the likely 'maximum' demand. For example, if *y*=1 half-hour period in a year, x represents a level of consumption that is only expected to be exceeded for 1 half-hour period per year for that node. This joint distribution calculates a value for *x* (the Peak Demand After Diversity). ▪ Generation of a joint probability distribution, using a dataset that is representative of the characteristics of the cluster and use the values to create a PDF. The tasks associated with this Problem Statement shall be detailed once the approach and algorithm specified as part of the Novel Analytics workstream is understood. In summary, it is expected that CLNR data shall be used to test the approach and generate a 'Peak Demand after Diversity' value (or set of values) that can be used in the LV Model.
Outcomes	The outcome of this approach is to determine how smart meter data can be used as a basis for sampling technique developed as part of the Novel Modelling Techniques to support a probabilistic approach to network planning, rather than the traditional method of applying a deterministic max and min. The outcomes from this activity are: ▪ Review the approach identified as part of the Novel Modelling techniques workstream and produce an algorithm to ensure smart meter data can be used to produce the Peak Demand After Diversity for a cluster of individual customers. ▪ Recommendations on how smart metering could be effectively utilised in network design and planning by integrating with LV and multi-voltage level network modelling.
Assumptions	

Notes	