Article:
Poursharif, G., Brint, A., Holliday, J. et al. (2018) Low voltage current estimation using AMI/smart meter data. International Journal of Electrical Power and Energy Systems, 99. pp. 290-298. ISSN: 0142-0615

https://doi.org/10.1016/j.ijepes.2018.01.023

Low voltage current estimation using AMI / smart meter data

Goudarz Poursharif[1*], Andrew Brint[1], John Holliday[2], Mary Black[3] & Mark Marshall[3]

[1]Management School, The University of Sheffield, S10 1FL, UK

[2]Information School, The University of Sheffield, S1 4DP, UK

[3]Northern Powergrid plc, Castleford, WF10 5DS, UK

*Corresponding author: G.Poursharif@sheffield.ac.uk

Contribution

This paper investigates how currents on a low voltage network can be estimated when not all the advanced metering infrastructure (AMI) / smart meters are reporting their values in real-time. This is done by combining real-time monitoring at the distribution substation with historical readings from the meters. Accurate knowledge of the currents is a key element of Advanced Distribution Management Systems. It is found that the k-nearest neighbors weighted average approach performs best.

Abstract

Knowledge of the currents is a key foundation for smart grid applications. However, knowledge of low voltage currents is generally poor. The new information streams from advanced metering infrastructure (AMI) / smart meters and the monitoring of distribution substations offer the opportunity of rectifying this. Unfortunately, often not all the smart meter readings will be available in real-time. For example, this situation will arise when older (non-compliant) smart meters do not have real-time reporting capabilities. This paper investigates how knowledge of the substation currents can be combined with the available real-time AMI / smart meter readings and the historical readings from the non-real-time meters, to estimate these missing values. It is found that the k-nearest neighbor weighted average approach performs best but that the gains over using simpler methods are relatively modest.

Keywords: load allocation, smart grids, nearest neighbors, low voltage networks

1. Introduction

The smart grid technology of information gathering, control technology and distributed resources is driving the development of Advanced Distribution Management Systems for the improved operation of electricity networks (Fan and Borlase 2009, Arritt and Dugan 2011). Most of the focus has concentrated on the higher voltage levels. However, the potential impact of the technology on low voltage networks is very considerable due to their large size, the lack of previous monitoring and the current limited amount of active network management. For example, it is estimated that of all the energy supplied by the UK's electricity distribution

networks (that is 132kV and below), 4% is lost on the low voltage network, while only 3% is lost on the rest of the network combined (Sohn Associates 2009).

A key base module in these Advanced Distribution Management Systems is load estimation / load allocation, i.e. estimating the currents around the network. However, accurately estimating the real-time low voltage currents is a problem as usually the monitoring of currents has only been carried out higher up the network. For example, in the United Kingdom normally the nearest monitoring point to the low voltage network has been at the 33kV to 11kV substation. This is changing as cost reductions and technological advances mean that real-time monitoring of distribution substations is becoming a reality. Additionally, the completion of the roll out of advanced metering infrastructure (AMI) / smart meters in the UK in 2025 will potentially provide the Distribution Network Operators (DNOs) with a far more comprehensive view of the low voltage currents. However, readings from the older smart meters will not be available in real-time, i.e. the readings will be available as a batch every few weeks or months. While this is not a problem for the modelling of longer term distribution problems using smart meter data such as the fair allocation of costs (Klonari et al. 2016), it is an issue for the more active management of the low voltage network (Leão et al. 2011).

This paper investigates how to estimate these missing values so as to provide a full picture of the low voltage currents. It assumes that the currents at the distribution substation are being monitored, that the historical smart meter readings are available but that the most recent readings are not available for some of the smart meters. The performance of a number of estimation approaches are assessed on how well they estimate the total currents measured by different sized groups of smart meters. This corresponds to estimating the currents downstream of a tee junction where the real-time readings from some meters are missing from both branches. If the two groups of non-real-time meters are denoted by A and B, then the problem is:

> Given the combined current from the groups of non-real-time meters A and B, along with the historical currents from each of A and B, then what are the best estimates of the present currents for groups A and B?

The analysis found that the performance of the k-nearest neighbor weighted average approach was best. However, the performance of the much simpler approach of using the most recent matching times was sufficiently close to it to mean that this is likely to be the preferred approach in practice. The ability of these two approaches to estimate individual meter values was then assessed.

Note that the smart meters are assumed to relay the voltage as well as the load, and so the problems of load estimation and current estimation are modelled as being interchangeable.

2. Background

Knowing the magnitude of the currents is fundamental to many network management applications (Fan and Borlase 2009). The large size of the low voltage network makes it an important area for improving network efficiency, e.g. low voltage networks in the UK comprise 48% of the total length of the distribution and transmission networks (EurElectric 2013). Leão et al. (2011) note that it is essential to be able to exchange information about the demands and available generation on low voltage networks "to maintain balance and overall power quality". However, knowledge of the low voltage currents has generally been extremely limited.

2.1 Historical methods for low voltage current estimation

In the UK, usually the lowest load monitoring point on the distribution network has been of the 11kV feeders at 33kV substations. Below this the main information that has been available has been the approximately yearly consumption (billing) totals from the individual low voltage customers, and the maximum currents recorded at distribution substations have been available.

The main approach to low voltage current estimation has centred on estimating the peak currents for use in the design and extension of low voltage networks. Two approaches have been widely employed in the UK for estimating these:

- After Diversity Maximum Demand (McQueen et al. 2004) – The maximum demand for a customer of a particular type is specified. The maximum demand from a number of these customers is then taken to be this individual maximum demand times a diversity factor that is a function of the number of customers in the group.
- Customer demand curves (Carson and Cornfield 1973, McQueen et al. 2004, Vélez et al. 2014) – allowing customers with peaks at different times to be modelled.

When planning a network extension, both approaches assign a type to each existing customer and then the expected demands from each customer are adjusted in line with their annual consumptions. However, the focus of the approaches is on the maximum currents that might be experienced by the network, rather than predicting the currents at the hourly or half-hourly level that form the basis of Advanced Distribution Management Systems. Consequently, the question this paper investigates is how the estimation of the customer demands at any time of the day and year, can be improved by using data from distribution substation monitoring along with either annual consumptions or smart meter data. The problem comes down to splitting a measured demand (i.e. at the substation) between individual customers, or between groups of customers on different downstream branches. Several approaches were investigated in Kersting and Philips (2008) and Arritt et al. (2012) for splitting the substation demand between the customers using the maximum rating of the customer, e.g. "Transformer kVA Allocation" using the kVA rating of the service transformer, "Monthly Usage Allocation" using the

customer's kWh bill or "Class Loadshape Allocation" where these are combined with the customer load shapes. The billing approach was found to be much closer to the actual values "compared to other traditional methods" Arritt et al. (2012). This paper considers how smart meter data can be used to improve on these methods, how great is the improvement in the accuracy of the estimates, and how effective simple approaches based on smart meter data are. The situation investigated is where not all the smart meters are reporting their values in real-time.

2.2 Spatial data statistics

Central to the approach is that for "similar times and situations" the ratios of the loads on different branches of a low voltage network are likely to be similar. For example, the split of the currents across the network exactly a week ago is likely to be a good estimate of the split now. Hence a methodology is needed for identifying these "similar times and situations", and then using the ratios on these occasions, to estimate the present ratio. A similar situation arises in spatial statistics, as often a point's value is closer to the values of geographically nearby points than to the values of more distant points (Oliver & Webster 1990, Hofstra et al. 2008). Consequently, a number of approaches have been developed for predicting the value at a point using spatial correlation. The central steps in these approaches are (i) defining the separation measure, (ii) how the data is restricted to nearby points and (iii) the estimation calculation.

In geographic settings, the definition of the separation measure (stage i) is often straightforward, e.g. using Euclidean distance. However, where a non-geographic element such as time is included in the separation, then the weighting between the elements needs to be specified (e.g. Wu et al. 2014).

In stage (ii), the data is usually restricted to nearby points. This is normally done by selecting the k-nearest neighbors (Ledolter 2013) as the alternative of choosing all points within a certain distance of the point of interest, can cause problems if the density of the points varies (Fotheringham et al. 2002). The value of k is usually set as the value that gives the best results on the training set (e.g. Eskelson et al. 2009).

In stage (iii), the separations need to be converted into weights. Fotheringham et al. (2002) looked at using Gaussian and bi-square kernels. They reported that "the results … are relatively insensitive to the choice of the weighting function but they are sensitive to the bandwidth chosen" (page 44). Other investigators have often chosen the inverse of the separation distance (e.g. Räty and Kangas 2012). Choosing a weight of one for all the k-nearest neighbors corresponds to a standard regression model.

These different choices mean that several different approaches to producing the estimated value at the target point have been reported. Li et al. (2012) used a k-nearest neighbor

weighted regression approach, while Fotheringham et al. (2002) used “geographically weighted regression” – this is similar but has some minor differences such as the bandwidth is adjusted rather than having a fixed k when selecting the data points. Wu et al. (2014) used the simple and weighted averages of the k selected points, but most people have just used the weighted average of the k selected points (e.g. Maleika 2015, Suominen et al. 2013).

3. The data

Two new sources of load information on the UK’s low voltage networks will become more widely available in the next few years:

- Real-time monitoring of the feeder phase currents at the distribution substation with the potential for monitoring at one second intervals (Lees 2014).
- AMI / smart meter data – The United Kingdom’s £12 billion programme for all domestic customers to have smart meters is due for completion in 2025. These smart meters will store and transmit half hourly consumptions and maximum demands (Lees 2014), but will have the capability to move to a shorter time interval at a future time. The values will be available to the Distribution Network Operator (DNO). However, the smart meter data will be aggregated together in groups to protect privacy (DECC 2012), the data will not identify the phase the meter is on, and the readings from a significant proportion of (the older) meters will only be available periodically (i.e. every few weeks or months rather than within 2 minutes of the half hour ending for the newer meters).

The goal is to use these two data sources to provide an estimate of the low voltage currents. The modelled scenario assumes that the currents at the substation are available in real-time but that some of the smart meter readings are not available for the most recent 24 hours. In practice, most of the smart meters will provide readings in real-time – the scenario corresponds to deleting these real-time current measurements from the substation current to just leave the current from the non-real-time meters.

Note that this scenario also matches the situation where the currents at the distribution transformer have been predicted for a future time, e.g. tomorrow, but knowledge of the low voltage currents is restricted to current and historical smart meter values. It is now required to predict the corresponding low voltage network currents for the future time. This case stems from it being easier to accurately predict the substation currents (e.g. as described in Huang et al. 2014) than the customer currents due to diversity. An interesting and very pertinent approach to substation current prediction using smart meter readings is described by Mirowski et al. (2014).

3.1 Data used

Two sets of smart meter data were used in the analysis:

Richardson & Thomson (2010)

The data comprised domestic (residential) electricity consumption data measured over approximately 12 months. The analysis was restricted to 18 of the 24 customer profiles as the others had substantial gaps in their data over the 12 month period. Figure 1 shows the values for first 4 load curves in this data set for Wednesday the 19th of March 2008.

CLNR (2017)

The data comprised readings from 200 domestic (residential) customers with time of use tariffs for a period of 3 months at the start of 2014 in the North East of England. Figure 2 shows the values of the first 4 load curves in this data set for Wednesday the 19th of March 2014. Figure 3 shows the values for the same curves a week later. Although some aspects of the load curves match between the two figures, there is a considerable difference between the two sets of curves. Corresponding maximum daily temperatures were obtained from WeatherOnline (2017).

Both data sets are available for free public download from their respective websites.

The number of meters in the first data set is low when compared with the 300 customers typically connected to a ground mounted distribution transformer in the UK. However, these 300 customers will be split across 4 feeders, and the percentage of the smart meters that cannot communicate their readings in real-time to the distribution network operator, is expected to be only in the region of 10% to 20%. Hence the number of non-real-time smart meters on a feeder is likely to be up to about 20. Therefore, the number of unknown meter readings being estimated from the ***unmatched*** part of the substation reading (i.e. after the known real-time meter readings have been removed from the feeder total) is likely to be not that much greater than the Richardson & Thomson data set size. Consequently, although the first data is small, it is likely that the number of missing readings will be of the same order of magnitude as it, and so it allows the relative accuracy of the approaches to be evaluated.

The Richardson and Thomson data set was used to analyse the performance of the different approaches when estimating the split of currents downstream of a tee junction. For each analysis, the customer profiles were randomly split into two groups A and B. These groups modelled the non-real-time meters on the two downstream branches.

The ability of the two best methods to estimate individual meter values was then assessed using the CLNR data set.

4. Estimation approaches

The underlying idea is to split the unmatched current monitored at the substation between the groups A and B of non-real-time meters, or between the individual non-real-time meters according to the most appropriate historical ratio.

4.1 Defining similar historical situations

When using similar historical situations to estimate a current value, the factors used to define the similarity between the current and historical situations need to be specified. Following this,

Stage i: The similarity measure needs to be decided upon

Stage ii: Any restriction on the number of neighbors specified

Stage iii: The similarity factor needs to be converted to a weight (Fotheringham et al. 2002).

The factors that were used were:

- *Day of the week* – Behavior patterns are likely to differ between different days, e.g. load curves on Mondays are very different from load curves on Sundays, while Tuesdays are more similar to Wednesdays than they are to Sundays.
- *Bank holiday* – Bank holidays are likely to have profiles more similar to weekends than to weekdays.
- *Half hour in the day* – Neighboring half hours are likely to have more similar load characteristics than those separated by several hours. For example, some customers use relatively more energy in the evening than in the morning than others.
- *Week in the year* – Load characteristics are likely to alter over the year due to changes in temperatures and the hours of daylight.
- *Substation load* – This acts as a proxy for other variables such as temperature that affect the demand.

Other factors could have been used but they were not present in the data set, in particular:

- *Year* – Data from several years ago will generally be less helpful than data from last year or earlier this year as customers and / or appliances may have changed.
- *Half hourly temperatures* – As the level of electric heating load is likely to vary between customers, knowing the temperature might improve the accuracy of the estimation. However, the customer load profiles in the two data sets used did not contain the temperature for each time period.
  An approximation to knowing the half hourly temperatures, is to use the daily

maximum temperature. The benefit of including this information in the modelling is investigated in Section 6.3.

The separation between the estimation time and a preceding time was defined to be the product:

$$\text{Separation} = D^d \times H^h \times W^w \times S^s$$

where d, h, w and s are weighting parameters and

- D is the separation between the day types. If they are the same day type, then the value is 1. If they are different weekdays, then the value is 2. If they are different weekend days, then the value is 2. In all other cases, including bank holidays, the value is 4.
- H is the separation between the half hours. Each half hour in the day is given a value from 1 to 48 and how many half hours apart they are is determined in the range 0 to 24.
- W is how far apart the week numbers in the year are for the two situations. It takes a value in the range 0 to 26.
- S is the difference between the loads in kW at the substation for the two situations.

When converting the separation to a weight, the basic principle was that the more similar the situations were, i.e. those with lower separation scores, then the higher the weight should be. The three conversion functions that were considered were:

- The reciprocal of the separation – used in many studies including Wu et al. (2014).
- The normal distribution probability density function. The standard deviation was chosen so that the largest separation of the k-nearest neighbors corresponds to 3 standard deviations from the mean. This function is analogous to the Gaussian kernel used in Fotheringham et al. (2002).
- The reciprocal of the square root of the product of the separation and the ranking of the separations. This was considered as it reduces the dominance of the closest point when there is only one very close point.

4.2 The approaches

The case where the non-real-time smart meters were split into two groups was analysed using the Richardson and Thomson data set. The approaches investigated were based on those described in Section 2 that have been applied in analogous situations.

Although approach 3 was the best performing approach, the closeness of the simpler and computationally less demanding approach 7, meant that this could well be the choice of network operators in practice. Therefore, the approaches investigated for estimating individual meter

values using the CLNR data set concentrated on approaches 3 and 7 along with simple extensions to approach 7.

4.2.1 Estimating the currents for two groups of non-real-time smart meters

The approaches that were considered were all based on producing an initial estimate of the demands for groups A and B. These demands were then scaled by the unmatched substation load divided by the sum of the initial estimates of the A and B demands, to get the final estimated demands at A and B. This normalises the estimates for A and B so that they sum to the unmatched substation load. Hence the difference between the approaches is in the initial estimation of the demands.

The approaches considered were:

1. *The maximum demands recorded by the smart meters.* The estimated load at A was the maximum demand at A multiplied by the unmatched substation load, and then divided by the sum of the maximum demand at A and the maximum demand at B. This is similar to the maximum transformer rating approach of Arritt et al. (2012) but with the maximum demand replacing the rating of the service transformer.
2. *The total consumptions recorded by the smart meters.* The estimated load at A was the billed consumption at A multiplied by the unmatched substation load, and then divided by the sum of the billed consumptions at A and B. This is analogous to the customer consumption approach of Arritt et al. (2012).
3. *The k-nearest neighbors weighted average of previous similar situations.* The fraction of the unmatched substation load due to the load at A, was found for each historical time period. The weighted average of these fractions was calculated for the k situations nearest to the present situation. This value was multiplied by the present substation load to estimate the present load at A. This is similar to the approaches applied in other fields such as traffic flow prediction (Wu et al. 2014).
4. *The k-nearest neighbors weighted regression of previous similar situations.* The explanatory variables were the differences between the training and prediction situations for the day of the week, the week in the year, and the half hour in the day, plus the substation demand for the training case. The response variable was the ratio of the meter point load at A divided by the unmatched substation load. The weight for an observation was taken to be the same as the weight in the k-nearest neighbors weighted average approach, i.e. approach 3. Hence a separate regression was run for each prediction point. The k-nearest neighbors weighted regression was advocated for estimating traffic flows in Li et al. (2012).

5. *The k-nearest neighbors linear regression of previous similar situations.* This approach is the same as approach 4 except for the fact that all the weights were set at one. It was motivated by the weights often being taken as one when using k-nearest neighbors e.g. in Wu et al. (2014).
6. *Using the most recently available matching day* i.e. if the current day is a Tuesday, then the most recent Tuesday for which the non-real-time meter values are available. This corresponds with an approach that is likely to be widely used in practice as it is a natural approach that is straightforward to implement.
7. *Average of the same day and time over the preceding 6 weeks.* This is the same as approach 6 except for averaging over 6 weeks. 6 weeks was chosen as it allows the values for any day in March to be estimated when using the 3-month January to March period of the CLNR data set without having a problem with the Christmas – New Year break. In general, including more preceding weeks in the average gave more accurate estimates of the meter values, but the benefit of using 6 preceding weeks rather than 5 was relatively small.
   Forming a weighted average with the weeks further away from the target time having less weight than those that are closer, was investigated but gave very similar results to the non-weighted average. Therefore, only the non-weighted average case is reported.

The first two approaches are similar to the (installed) kVA and customer billing kWh approaches considered in Arritt et al. (2012). As no account is taken of the load curve shapes, the ratio of the loads between groups A and B is always the same. The remaining approaches are based around finding situations in the past that are similar to the present one, e.g. same year, day of the week, time of day, substation demand, etc. The load ratio between groups A and B is then assumed to be similar to the load ratios in these previous similar situations.

4.2.2 Estimating individual meter values

The approaches considered were approaches 3 and 7, along with 3 approaches that augment approach 7 with temperature and substation load information. Instead of averaging the historical ratios as in approach 7, approaches 8 to 10 use regression to model these ratios as depending on the temperature or substation load:

8. *Simple regression using the maximum daily temperature*. For each non-real-time meter, the response values are the ratios of the meter reading to the unmatched substation reading for the same day and time of the 6 preceding weeks. The explanatory variable is the maximum daily temperature.
9. *Simple regression using the unmatched substation value.* For each non-real-time meter, the response values are the ratios of the meter reading to the unmatched

substation reading for the same day and time of the 6 preceding weeks. The explanatory variable is the unmatched substation reading.

10. *Multiple regression using the maximum daily temperature and the unmatched substation value.* For each non-real-time meter, the response values are the ratios of the meter reading to the unmatched substation reading for the same day and time of the 6 preceding weeks. The explanatory variables are the maximum daily temperature and the unmatched substation reading.

4.3 Setting the parameters

The approaches in Section 4.2.1 involve several parameters. To set their values, the customers were split into two groups of size 9. Each group was modelled as having one collective meter value which was the sum of the 9 individual meters making up the group. One hundred times were randomly selected from the period covered by the data set

- For each of these times, the non-real-time smart meter data was restricted to no more recent than 48 hours before this time.
- Values were assigned to the weighting parameters d, h, w and s, and the number of nearest neighbors, k.
- For each of the historical times, i.e. the times 48 hours or more before the time whose demand was being estimated, the weights from the 3 conversion functions of Section 4.1 were calculated, e.g. the reciprocal of the separation.
- For each of the conversion functions, the sum of the squared differences between the loads estimated by approach 3 and the actual loads was calculated.

The sums of the squared differences between the estimated and actual collective meter values for the 100 estimation times, were added together to given an overall error score for each of the conversion functions. Finally, the steps were repeated with different values of d, h, w and s, and the values that gave the lowest overall error for each conversion function were found.

This process was repeated for five values of k, namely 10, 15, 20, 30 and 50. The resulting means of the overall error scores are given in Table 1. The performance of the weight conversion functions was similar. As the reciprocal of the separation is the simplest of these functions, this was used in all the subsequent analyses. Table 1 also shows that k=20 gave the best or close to the best predictions for approach 3. The best values for d, h, w and s for k=20 were 0.07, 0.14, 0.05 and 0.10 respectively. These were the values used in the rest of the analyses. However, it should be noted that if a different data set of load curves was used, then these values would be likely to change.

5. Comparing the nearest neighbor approaches

The customers were partitioned so that there were two groups of nine non-real-time meters and as before, one hundred times were randomly selected from the period covered by the data set. The three nearest neighbor approaches (i.e. approaches 3, 4 and 5) were then used to estimate the load from group A at these times for different values of k. Each setting was repeated 80 times and the mean values for the squared differences between the estimated and the observed loads from group A, are given in Table 2 and plotted in Figure 4.

The k-nearest neighbor weighted average approach gives poorer predictions as the number of nearest neighbors increases. It seems that the key information is in a small number of the most similar situations. Conversely, the nearest neighbor regression approaches improve as the number of nearest neighbors increases (as the robustness of the regression coefficient calculations improve with the data set size).

Therefore, the only nearest neighbor approach used in the following analyses was the weighted average approach (approach 3) with k taking the value 20.

6. Selecting the best approach

The performances of approaches 1, 2 and 3 were also compared using one hundred times randomly selected from the period covered by the data set. For the modelling, the non-real-time meter values were not available for the 48 hours preceding each estimation time. The results are shown in Table 3. The k-nearest neighbors weighted average has the lowest error in this table while the two non-load shape approaches performed similarly to each other.

Carrying out a t test to assess whether the errors from the k-nearest neighbors weighted average (approach 3) were lower than those from the ratio of the annual consumptions (approach 2), gave a p value below 10^{-5}. This can be interpreted as there is a less than a 10^{-5} probability of the observed error values from approach 3 being this much lower than those from approach 2 if approach 3 was actually no better than approach 2. Hence, the conclusion is that the k-nearest neighbors weighted average approach gives better predictions than the non-smart meter data approaches.

6.1 Altering the group size

As data from smart meters in the UK will be aggregated together so as to protect customer privacy (DECC 2012), an important issue is how the accuracy of the predictions alters with the size of the group being predicted. Figure 5 shows the Mean Absolute Percentage Error (MAPE) for 100 randomly selected prediction times for different sizes of group A when the total number of customers was 18 (i.e. groups A and B combined had 18 customers). MAPE was used rather than a squared error term as the group currents being predicted increase in size as the number of customers in the group increases.

For approach 6, the performance using each of the previous 10 weeks was analysed. The regression model with the number of customers in A and the number of the previous week (i.e. 1 to 10) as the explanatory variables, and MAPE as the response variable, gave the coefficient of the previous week variable as 0.80 with a standard error of 0.24. Consequently, how many weeks earlier the approach is based on, does not have a huge effect on the accuracy of the prediction. Therefore, to reduce the noise in a single week's observation, the average of the same day in the preceding 6 weeks was used, i.e. approach 7.

As approach 3 was better than nearest neighbor approaches 4 and 5 in Figure 4, only approaches 1, 2, 3 and 7 were analysed. The values for approach 1 for group sizes below 5 were 812%, 364%, 141% and 124%. These are not shown in Figure 5 so as to avoid compressing the vertical scale.

The conclusions from Figure 5 are that

- Approach 1 (maximum demand) gives inaccurate results, especially for small group sizes, e.g. below 8 customers.
- The performance of the other approaches become similar as the group size becomes large, e.g. 8 customers or more.
- Approach 3 (k-nearest neighbors) gives the best results but averaging over the same day in recent weeks (approach 7) is not that much worse.
- As the group size increases, the percentage error between the predicted load value and the actual load value decreases. This is in line with the general principle that the total load from a group of customers becomes less volatile as the group size increases because of diversity amongst the customers.

6.2 Equal group sizes

Figure 6 shows the results from 100 predictions when the sizes of groups A and B were the same, ranging from two groups of 1 individual each to two groups of 9 individuals each. While there is more volatility in the values stemming from the reduced number of individuals being modelled, the results are broadly in line with Figure 5 in that approaches 3 (k-nearest neighbors) and 7 (average of the previous 6 weeks) seem to perform better than approach 2 (total consumption).

6.3 Estimating individual meter values

The CLNR data set was used to assess the ability to estimate the values of individual meters. This data set had the advantage of a much larger number of load profiles than the Richardson and Thomson data set, but the monitoring period for the data used in the current analysis was much shorter at 3 months rather than the 12 months of the Richardson and Thomson data set.

The analysis using the Richardson and Thomson data set led to approaches 3 and 7 being selected for this analysis. This stemmed from the overall performance of approach 3 in Sections 5, 6.1 and 6.2, and the good performance of approach 7 coupled with its straightforward nature. A variant of approach 7 where the ratio of a meter's value divided by the unmatched substation value is regressed against the daily maximum temperature for the 6 preceding matching days, to give an estimated ratio on the target date, was also analysed. The rationale for this approach was that the relative amount of electricity used by different customers was likely to vary with how hot or cold the day was, and so the ratio should be modelled as varying with the maximum daily temperature.

As a similar conjecture can be put forward for the relative electricity usage at times of high and low substation loads, the individual ratios were also regressed against the unmatched substation value in approach 9, and both the unmatched substation value and the daily maximum temperature in approach 10.

The analysis involved subsets of 10, 20, 30 and 40 meters. These were randomly chosen from the 200 meters. This random selection was repeated 40 times to give 40 sets of meters for each of the 10, 20, 30 and 40 meters cases. The approaches were compared using the Mean Absolute Error between a meter's estimated value and its actual value.

The analysis was carried out for 7 days in March 2014 – each one being a different day of the week. The days were spread out through the month so that any customer behavior over a 2 or 3-day period, only affected one of these 7 days. For each of the days, the meter values were estimated for the 5 times: 08:15, 11:15, 14:15, 17:15 and 20:15. These were chosen as being equally spaced out during the main daily demand period. The results are given in Table 4. Approach 3 performs slightly better than approach 7. Figure 7 gives the histogram of the average absolute error for each meter where the averaging is over the 35 target times.

Table 4 indicates that the extra information from using the daily maximum temperature and the substation value, provides little if any benefit in estimating individual meter values. The histograms for approaches 8, 9 and 10 were very similar to Figure 7 except for slightly more meters being to the right of the 0.56 interval for approaches 8 and 10. It is these meters that led to approaches 8 and 10 performing worse.

6.4 Evaluating the approaches

Besides the accuracy of the estimates, in practice it is desirable that an approach is

a) Straightforward to implement

b) Computationally inexpensive as the objective is to estimate the real-time power flows on the low voltage network.

Table 3 and Figures 5 and 6 show that approaches 1 and 2 give poor estimates compared with approach 3. Overall approach 3 generally gives the best estimates but it is in the middle in terms of ease of implementation, computational expense, and data requirements. It also has the problem that the weights d, h, w and s in Section 4.3 may need to be recalculated for different low voltage networks. Approaches 4 and 5 are less attractive than approach 3 due to taking considerably longer to run. Approaches 6 and 7 are straightforward to implement in a spreadsheet and evaluate quickly. Although approach 7's estimates are usually worse than those of approach 3 in Figures 5 and 6 and Table 4, they are close enough to mean that its simplicity makes it a good choice in practice. Approaches 8, 9 and 10 seem to provide little if any benefit over approach 7.

7. Concluding remarks

The problem investigated was estimating the loads around a low voltage network when some of the smart meters do not provide their readings in real-time. It was found that the k-nearest neighbors weighted average approach (approach 3) gave the best results. However, simply using the average for the same day over the most recently available weeks (approach 7), does not perform that much worse than approach 3. It is also a very simple approach to implement. Hence it is the approach that seems the most suitable for use in practice. Extra information such as the maximum daily temperatures seems of limited value when estimating the values of individual meters.

An area where further work would be of benefit is in determining confidence intervals for estimates of the missing meter values. For example, if the load from a non-real-time group of meters is estimated to be 10kW, what is the corresponding 95% confidence interval? Sahlin et al. (2014) have suggested using the weighted average of the prediction error of each nearest neighbor, but how practical this is for the current situation is unclear.

8. References

Arritt, RF, & Dugan, RC (2011) Distribution system analysis and the future smart grid *IEEE Transactions on Industry Applications* **47**(6) 2343-2350

Arritt, RF, Dugan, RC, Uluski, RW & Weaver, TF (2012) "Investigating load estimation methods with the use of AMI metering for distribution system analysis", *IEEE Rural electric power conference*, B3, 1-9 (IEEE xplore conference publication: http://ieeexplore.ieee.org/Xplore/home.jsp)

Carson, MJ & Cornfield, G (1973) Design of low-voltage distribution networks – interactive method based on calculus of variations *Proceedings of the Institution of Electrical Engineers* **120**(5) 585-592

CLNR (2017) Customer Led Network Revolution Project Data http://www.networkrevolution.co.uk/resources/project-data/

DECC (2012) Smart metering implementation programme: Data access and privacy – Government response to consultation *Department of Energy and Climate Change* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/43046/7225-gov-resp-sm-data-access-privacy.pdf

Eskelson, BNI, Temesgen, H, Lemay, V, Barrett, TM, & Crookston, NL (2009) The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases *Scandinavian Journal of Forest Research* **24** (3) 235-246

EurElectric (2013) Power distribution in Europe facts and figures http://www.eurelectric.org/media/113155/dso_report-web_final-2013-030-0764-01-e.pdf

Fan, J & Borlase, S (2009) The evolution of distribution *IEEE power and energy magazine* 7(2) 63-68

Fotheringham, AS, Brunsdon, C, & Charlton, M (2002) *Geographically weighted regression* Chichester: Wiley

Hofstra, N, Haylock, M, New, M, Jones, P & Frei, C (2008) Comparison of six methods for the interpolation of daily, European climate data *Journal of Geophysical Research. Atmospheres* **113**(D21) 1-19

Huang, J, Pilgrim, JA, Lewin, PL, Scott, D & Morrice, D (2014) Use of day-ahead load forecasting for predicted cable rating *5th IEEE PES Smart Grid Technologies Europe* (IEEE xplore conference publication: http://ieeexplore.ieee.org/Xplore/home.jsp)

Kersting, WH, & Philips, WH (2008) "Load allocation based upon automatic meter readings", *IEEE Transmission and distribution conference and exposition*, 1-8 (IEEE xplore conference publication: http://ieeexplore.ieee.org/Xplore/home.jsp)

Klonari, V, Toubeau, J-F, Grève, Z De, Durieux, O, Lobry, J, & Vallée, F (2016) Probabilistic simulation framework for balanced and unbalanced low voltage networks *International Journal of Electrical Power and Energy Systems* **82** 439-451

Leão, RPS, Barroso, GC, Sampaio, RF, Almada, JB, Lima, CFP, Rego, MCO, & Antunes, FLM (2011) The future of low voltage networks: Moving from passive to active *International Journal of Electrical Power and Energy Systems* **33** 1506-1512

Ledolter, J (2013) Data mining and business analytics with R New Jersey: Wiley

Lees, M (2014) Enhanced network monitoring report Customer-Led Network Revolution report CLNR-L232 http://www.networkrevolution.co.uk/resources/project-library/

Li, S, Shen, Z, & Xiong, G (2012) A k-nearest neighbour locally weighted regression method for short-term traffic flow forecasting Proceedings of the 15th IEEE Conference on Intelligent Transportation Systems, Anchorage, Alaska, USA, September 16-19, 1596-1601 (IEEE xplore conference publication: http://ieeexplore.ieee.org/Xplore/home.jsp)

McQueen, DHO, Hyland, PR & Watson, SJ (2004) Monte Carlo simulation of residential electricity demand for forecasting maximum demand on distribution networks *IEEE Transactions on Power Systems* **19**(3) 1685-1689

Maleika, W (2015) Moving average optimization in digital terrain model generation based on test multibeam echosounder data *Geo-Marine Letters* **35**(1) 61-68

Mirowski, P, Chen, S, Ho, TK, & Yu, C-N (2014) Demand forecasting in smart grids *Bell Labs Technical Journal* **18**(4) 135-158

Oliver, MA & Webster, R (1990) Kriging: a method of interpolation for geographical information systems *International Journal of Geographical Information Systems* **4**(3) 313-332

Räty, M, & Kangas, A (2012) Comparison of k-MSN and kriging in local prediction *Forest Ecology and Management* **263** 47-56

Richardson, I. & Thomson, M., (2010) *One-Minute Resolution Domestic Electricity Use Data, 2008-2009* [computer file]. Colchester, Essex: UK Data Archive [distributor], October 2010. SN: 6583, http://dx.doi.org/10.5255/UKDA-SN-6583-1

Sahlin, U, Jeliazkova, N, & Öberg, T (2014) Applicability domain dependent predictive uncertainty in QSAR regressions *Molecular Informatics* **33**(1) 26-35

Sohn Associates (2009). Electricity distribution system losses: Non-technical overview https://www.ofgem.gov.uk/ofgem-publications/43519/sohn-overview-losses-final-internet-version.pdf

Suominen, L, Ruokolainen, K, Tuomisto, H, Llerena, N, & Higgins, MA (2013) Predicting soil properties from floristic composition in western Amazonian rain forests: performance of k-nearest neighbour estimation and weighted averaging calibration *Journal of Applied Ecology* **50**(6) 1441-1499

Vélez, VM, Hincapíe, RA, & Gallego, RA (2014) Low voltage distribution system planning using diversified demand curves *International Journal of Electrical Power and Energy Systems* **61** 691-700

WeatherOnline (2017) http://www.weatheronline.co.uk/

Wu, S, Yang, Z, Zhu, X, & Yu, B (2014) Improved k-nn for Short-Term Traffic Forecasting Using Temporal and Spatial Information *Journal of Transportation Engineering* **140**(7) 04014026 1 - 9

Table 1: The mean error scores for the comparison of the functions for converting the separation to a weight

	Number of nearest neighbors k				
	10	15	20	30	50
reciprocal of the separation	0.64	0.59	0.60	0.62	0.67
normal distribution probability density	0.65	0.62	0.60	0.63	0.65
reciprocal of the square root of the separation times the rank	0.67	0.65	0.62	0.63	0.66

Table 2: How the error score varies with the number of nearest neighbors for the different approaches. The values are the average of 80 runs.

Number of nearest neighbors, k	k-nearest neighbor weighted average (approach 3)	k-nearest neighbor regression (approach 5)	k-nearest neighbor weighted regression (approach 4)
20	0.61	2.09	2.55
40	0.65	1.42	1.48
60	0.68	1.08	1.09
80	0.71	1.01	1.02
100	0.72	0.90	0.92
140	0.75	0.82	0.82
200	0.78	0.85	0.78
300	0.78	0.87	0.75

Table 3: The performance of the different prediction approaches when the customers are split into two groups of size 9.

	Squared difference between actual and predicted kW	
	Mean	Standard error
Ratio of the maximum demands (approach 1)	1.07	0.10
Ratio of the annual consumptions (approach 2)	1.04	0.10
k-nearest neighbor weighted average (approach 3)	0.58	0.07

Table 4: The mean absolute prediction error per meter in kW where the average is taken over the different samples of meters and the 35 target times for the CLNR data set

Number of meters being estimated	Approach				
	3	7	8	9	10
10	0.27	0.28	0.33	0.24	0.35
20	0.33	0.35	0.39	0.36	0.43
30	0.33	0.35	0.39	0.36	0.45
40	0.34	0.37	0.41	0.38	0.45

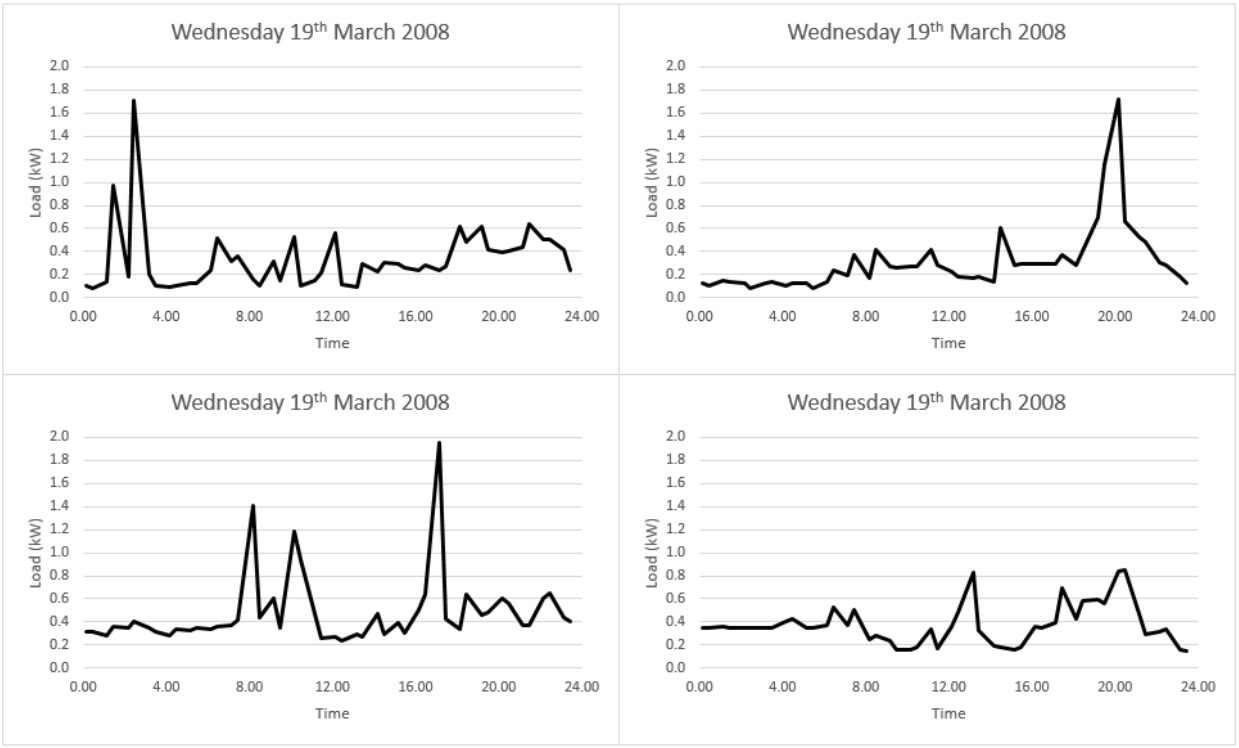

Figure 1: Four load profiles from the middle of March 2008 from the Richardson & Thomson (2010) data set.

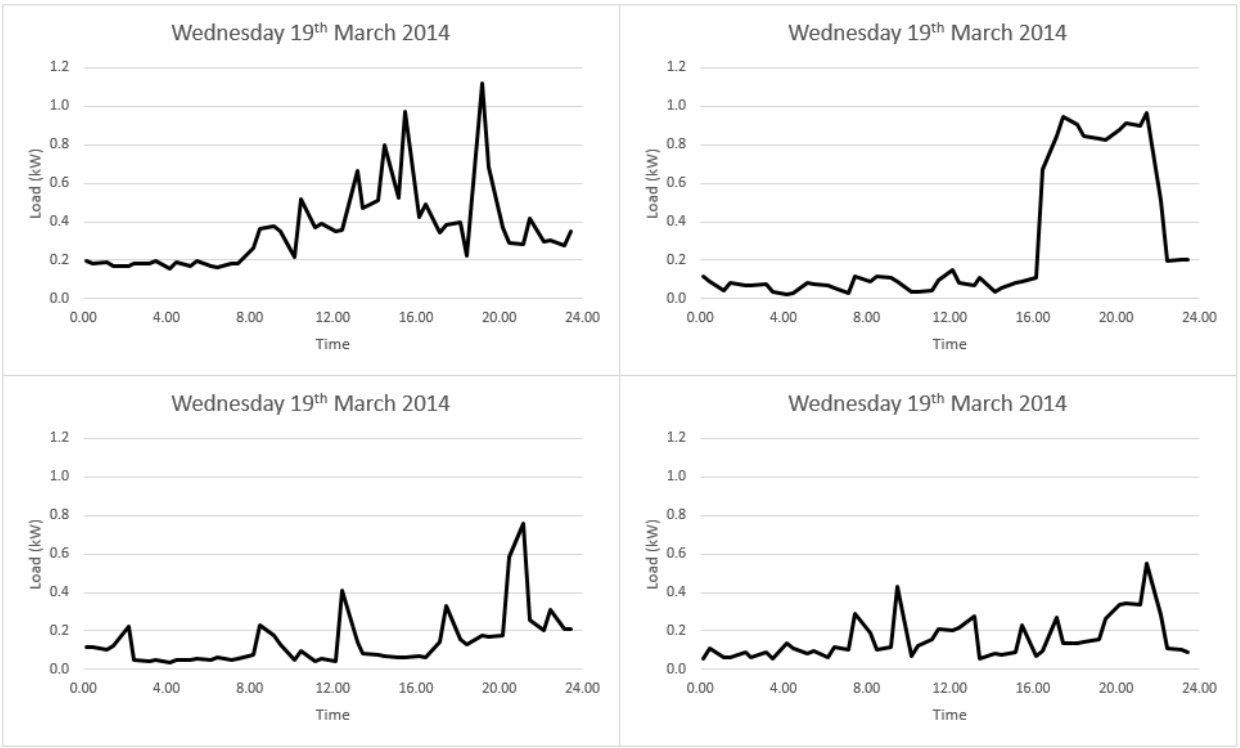

Figure 2: Four load profiles from the middle of March 2014 from the CLNR (2017) data set.

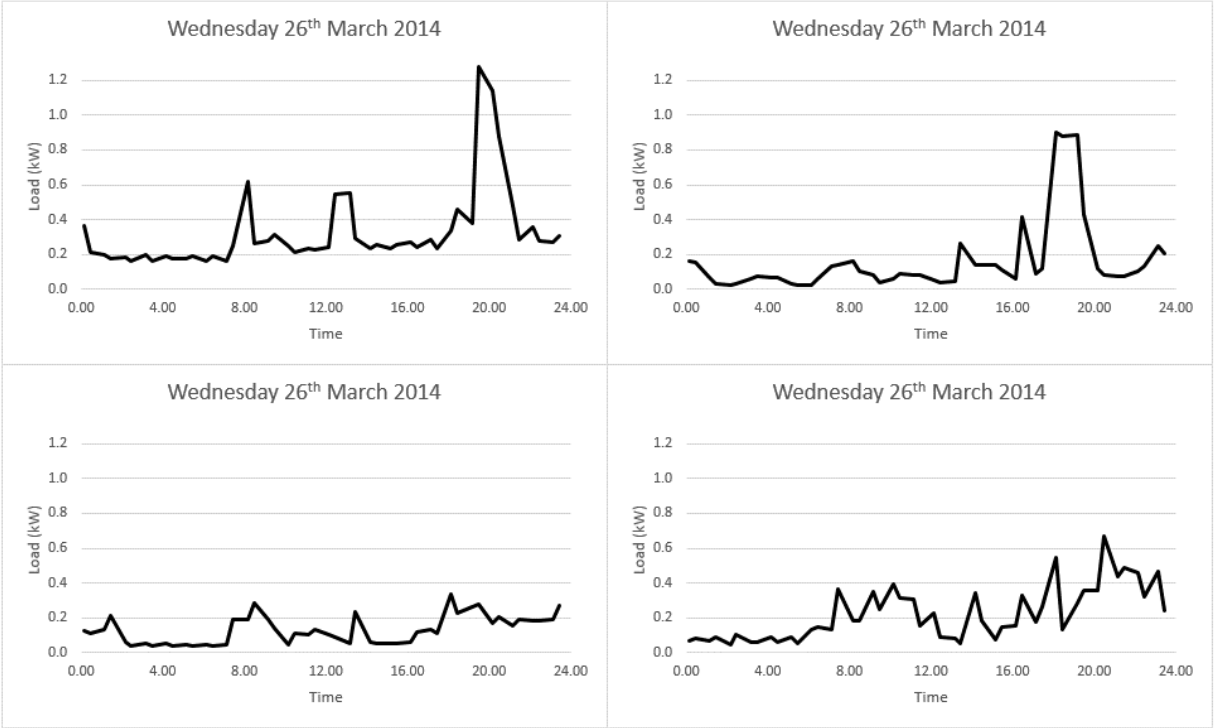

Figure 3: The four load profiles shown in Figure 2 one week later.

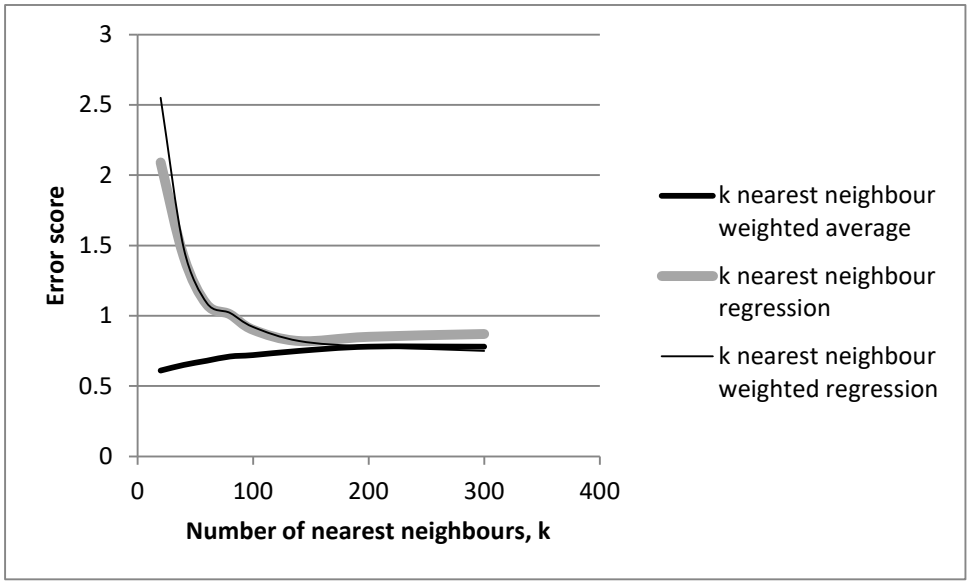

Figure 4: How the squared difference between the actual and the predicted loads from a group of nine customers varies with the number of nearest neighbors.

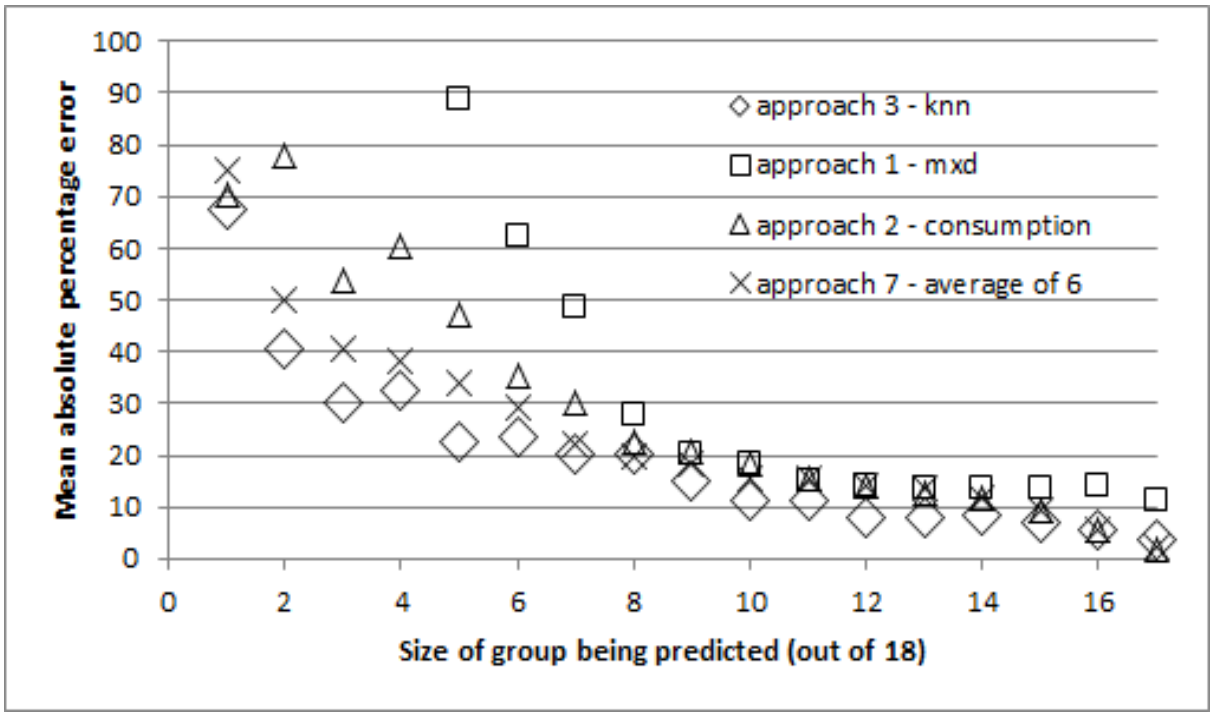

Figure 5: The variation in the predictive accuracy of the k-nearest neighbor weighted average approach as the prediction group size changes but the overall number of customers is fixed at 18. The approaches are described in Section 4.2.

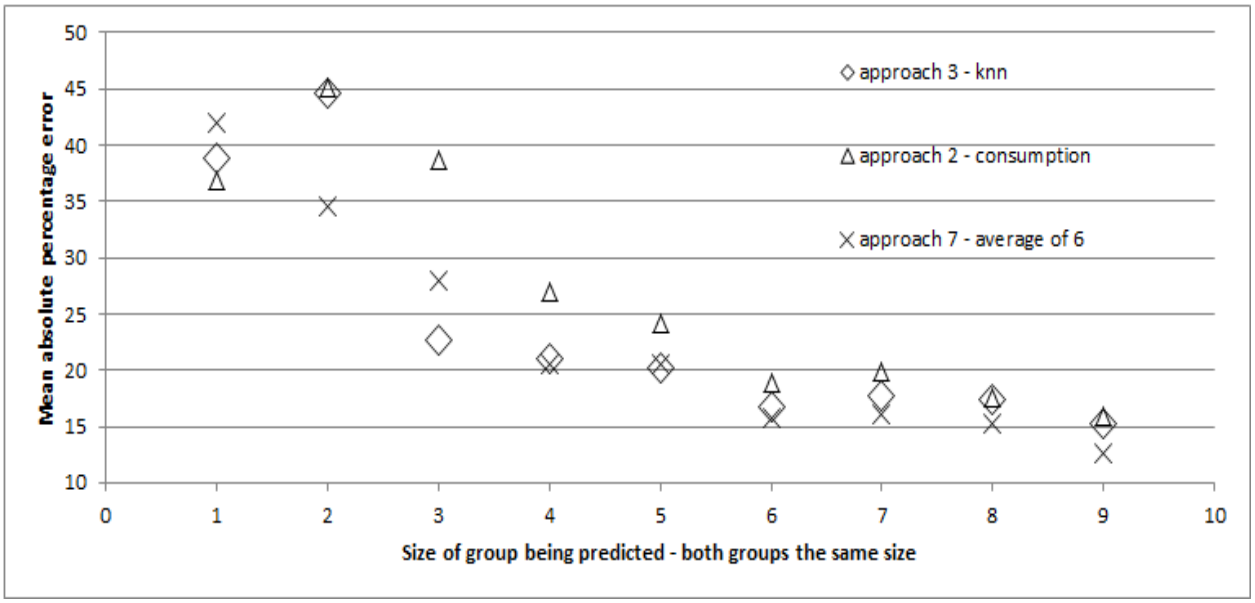

Figure 6: The variation in the predictive accuracy as the prediction group size changes and the overall number of customers is twice the prediction group size. The approaches are described in Section 4.2.

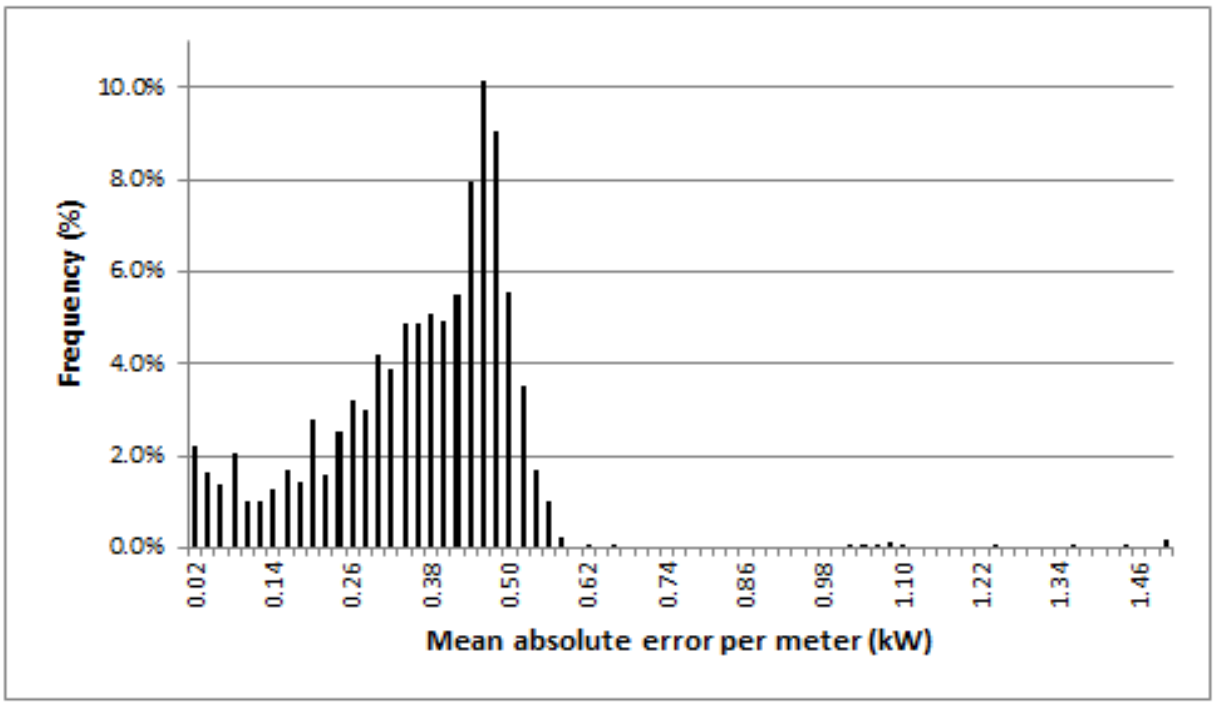

Figure 7: The mean percentage error for each meter for the 40 meters case using approach 7 (Section 6.3), where the average is over the 7 target days and 5 target times. The frequencies come from repeating the analysis with the 40 samples of 40 meters.