Determining Load Profiles for Customers in the Indicated DSO Area based on Smart Meters: A Summary Report for European Data Incubator

Paul Monroe, Riku Arakawa, Yohei Kiguchi, Kristjan Strojan, Alberto Arias

Updated 17 February 2020

Abstract

The growth in the implementation of smart-meters presents a significant opportunity to improve operational processes in the energy sector. One such opportunity is with distribution system operators using the data to better anticipate the state of the different areas of their networks.

However, large scale smart-meter rollouts often take years to complete, and having digital solutions that can realise outcomes as data becomes increasingly available can improve the return on investment of such rollouts. In this three month project conducted as part of the European Data Incubator program, data from over 2300 metered locations in the Torun region of Poland with varying levels of frequency - ranging from every 15 minutes to readings every few months - are examined to develop a methodology for creating hourly profiles of energy consumption for every location. Numerous patterns in the unavailability of the data are identified, and a methodology is proposed to address each pattern specifically. The proposed methodology tests multiple machine-learning based approaches - including Random Forest, Decision Tree, Support Vector Regression, and clustering - and assesses their effectiveness both against a generated baseline and other similar studies from academic literature. Additionally, functional considerations for deploying the algorithmic methodology in a full commercial use case are examined.

The results achieved demonstrate comparable or superior performance against previous studies, leading to the conclusion that the methodology proposed is among the best in the industry, and several areas for future improvement are identified.

1 Background

European Data Incubator is a program funded under the European Union's Horizon 2020 programme. It matches industry parties with specific data challenges with early-stage ventures that have unique expertise in solving these challenges to develop and deploy commercial solutions.

ENERGA-OPERATOR SA is the Data Provider for the challenge "Determining Load Profiles for Customers in the Indicated DSO Area based on Smart Meters." The company is a subsidiary of ENERGA SA and distributes electricity as a distribution system operator (DSO) across multiple regions in northern Poland. The role of the DSO is to securely operate and develop electricity distribution systems, which involves connections to numerous sources of electricity generation and demand. One of the critical activities undertaken by the DSO is the tracking and attribution of energy flowing into and out of the network. It is on the basis of these energy flows that parties connected to the network are billed for their energy usage.

The role of the DSO has evolved alongside several key trends in the energy sector. Firstly, the growth of smart meters has fundamentally changed how the different parties in the energy sector operate. In previous decades, electricity usage was metered solely on the device itself and typically required regular manual readings to determine energy consumed. This approach had obvious constraints, as it was limited by human intervention and readings were typically limited to monthly intervals. As metering technology evolved, meters became digital and were capable of recording and sending data on more frequent intervals - typically every 15 to 60 minutes. As a result, these "smart" meters began producing a growing amount of granular information about energy consumption and generation all across the grid, enabling a variety of different applications and pathways for efficiency improvements for participants all along the energy value chain.

Secondly, energy generation has become increasingly decentralised. Through most of the twentieth century, energy generation was feasible mostly from large generation plants. However, decreasing costs of energy generation and storage technologies - for example, in solar panels and batteries - has enabled small-scale generation for self-consumption. In some countries, this is further incentivised by feed-in tariffs, capacity markets, and other policy mechanisms - particularly where this generation is supplied from renewable sources. However, this adds significant complexity to the operation of energy distribution networks; where previously there were only a few locations suitable for generation, there are now many potential points which may or may not be offset by local consumption.

Enabling this decentralised, digitalised grid to flourish requires addressing a number of challenges. As a DSO, ENERGA OPERATOR SA is in the process of rolling out smart meters to its customer base and wants to start making use of the data as it becomes available. This business case is useful both for ENERGA OPERATOR SA's current situation as well as other DSO's starting to install smart meters, as the rollout of smart meter infrastructure typically takes over half a decade to complete. If there are valuable applications for the data demonstrated in cases where smart meter penetration is below a certain threshold, then returns on smart meter infrastructure can be realised earlier.

SMAP ENERGY is a UK-based SME who builds digital solutions to extract value from smart meter data. It has conducted numerous academic and industry studies across various topics related to extracting commercial value from smart meter data, particularly across the energy retail use case. The most prominent example of this is the company's commercially deployed solution called Simulator, which uses limited amounts of historical smart meter data to extrapolate annual load curves so that sales teams within the utilities can estimate values of energy contracts proposed to customers. SMAP Energy has been selected to address this challenge by participation in the EDI programme.

2 Project Overview

The goal of the project was to develop an algorithm for determining the load profile for energy metering points (MP) across a wide variety of data availability scenarios. The major challenge associated with this was the intermittent quality of the data. The dataset provided by ENERGA OPERATOR SA contained energy consumption and generation data for 2317 MPs for 2018, ranging from a full set of 15 minute readings during the period to no data at all.

The requirements for the project were both pre-determined by the Data Provider at the point of submitting the project, but were also further refined over the course of the project. The list of requirements is as

follows:

- The algorithm must interpolate both the consumption and generation data for a static dataset at hourly intervals where data availability will range from fully available at 15-minute intervals to not at all (i.e. only the meter description is available, such as in the case of newly connected meter points). It is known that all meter points will record energy consumption by default, but a subset of the meters will also have data related to energy generation at the site.
- The algorithm must also be able to recompute the interpolated readings as new readings become available - for example, reducing an interpolated reading when a new actual reading suggests that actual consumption was lower than the interpolated value.
- The results of the algorithm at a daily level must be able to be viewed via a web-based user interface, which provides indications as to the values that are known and the values that are interpolated.
- The results must be able to be downloaded in a CSV format for independent investigation and confirmation.

Though not a requirement, an operational scenario would require that computation be performed daily for with data extending back for a period of one year. Additionally, it was noted that an imminent requirement of the project is that the algorithm must interpolate the missing values such that the total consumption of all readings - both actual and interpolated - do not exceed a known total for the region (underestimated values will be attributed to distribution network losses). However, as data was unavailable for this item, this requirement is postponed for future study. In lieu of this, demonstrating accuracy becomes the primary benchmark for project success.

3 Data Preparation

3.1 Dataset Description

The data presented for this project involves 3 datasets: Meter Data; Registered Data, Consumption and Generation; and Area Balancing Meters. The fields of each dataset are shown in Tables 1, 2, and 3.

Table 1: Meter Data

Field	Values
METERING POINT ID	8 digit INT
EVENT TIMESTAMP	YYYY/MM/DD, HH:MM:SS [0000000000]
DEVICE ID	11-digit INT
MULTIPLICAND	INT from 1-18000
TARIFF	4-digit char: B21, B23, C11, C12A, C12B, C12W, C21, C22A, C22B, C23, G11, G12, G12W, G12R
METERING POINT TYPE	String: "small","producer","industrial"

3.2 Preprocessing Procedure

In the initial dataset, there were 2317 unique Meter Point IDs and timestamps ran from 1 January 2018 to 31 December 2018. Preprocessing evaluation resulted in the removal of 26 meters from consideration. Eight of these were removed on the basis of having no value for METERING POINT TYPE. Two of these meters were found to be duplicates of the transformer records in Area Balancing Meters, effectively skewing these resulting total. Upon further examination, it was determined that the Area Balancing Meters dataset may contain incorrectly assigned meters. Given time constraints of the project, it was decided to postpone this portion of the analysis until later. 16 meters were removed on the basis of incorrect accumulation of readings for either METER READING 180 or METER READING 280. The values in these fields are cumulative readings that should only increase. However, for these meters, there is a case where some readings are

Table 2: Registered Data, Consumption and Generation

Field	Values
METERING POINT ID	8-digit INT
DATA TIMESTAMP	YYYY/MM/DD, HH:MM:SS [0000000000]
METER READING 180	Float - Cumulative total of kWh consumed
METER READING 280	Float - Cumulative total of kWh generated
METER READING 181	Float - Cumulative total of kWh consumed in first time period
METER READING 182	Float - Cumulative total of kWh consumed in second time period
METER READING 183	Float - Cumulative total of kWh consumed in third time period
METER READING 281	Float - Cumulative total of kWh generated in first time period
METER READING 282	Float - Cumulative total of kWh generated in second time period
METER READING 283	Float - Cumulative total of kWh consumed in third time period

Table 3: Area Balancing Meters

Field	Values
DATA TIMESTAMP	YYYY/MM/DD, HH:MM:SS.SSS
TR1 (kWh)	Float - Cumulative total of kWh consumed at Transformer 1
TR2 (kWh)	Float - Cumulative total of kWh consumed at Transformer 2
TOTAL	Float

randomly set to zero, typically before resuming normal operation. This occurrence is demonstrated in Fig. 1.

The primary assumption is that reported readings from a meter are always correct, but this is clearly not the case with meters exhibiting this trend. Given the comparatively small amount of meters that this pattern is observed it, these are omitted from consideration in this study but future preprocessing would involve detecting these anomalies and treating them as values to interpolate. Therefore, the total number of meters considered is 2291.

The primary fields of interest in this study is METER READING 180, as this is the cumulative total of energy consumed on the meter, and METER READING 280, as this is the cumulative total of energy generated on the meter. The data is provided in 15-minute increments; however, the challenge requirement is to provide insights at the hourly level. Therefore, there is a required preprocessing step to extract the hourly readings from this. This is done by removing the intermediate readings (where timestamp at MM value is 15, 30, or 45) and leaving only the readings directly on the hour, resulting in two 2291 by 8760 arrays - one for consumption and one for generation for 24 hours across 365 days - where each row is described further by fields for METERING POINT ID, TARIFF, and METERING POINT TYPE.

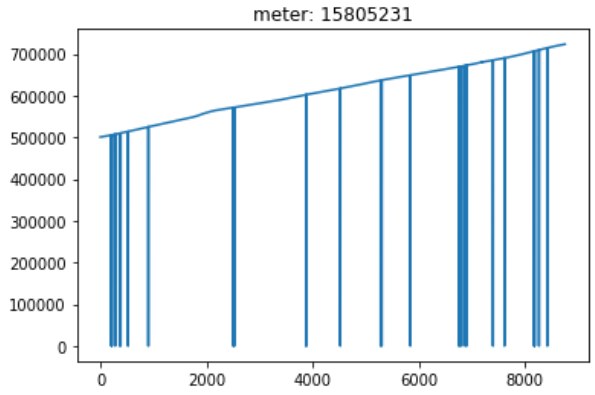

Figure 1: Example of incorrect reading accumulation in METER READING 180

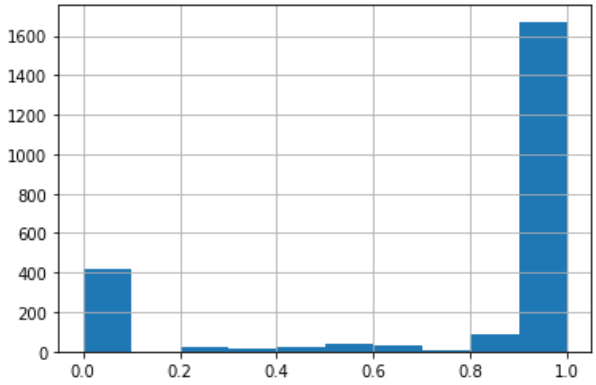

Figure 2: Histogram of METER READING 180 Data Availability (%)

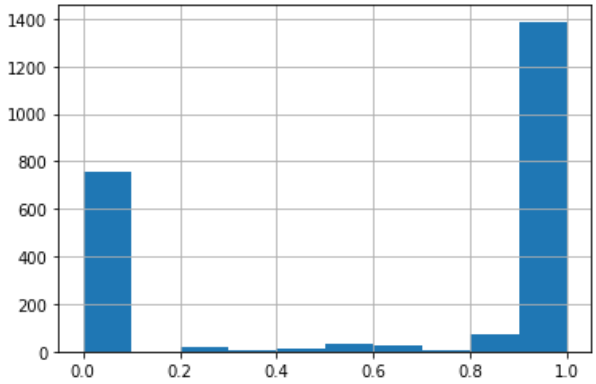

Figure 3: Histogram of METER READING 280 Data Availability (%)

3.3 Missing Data Analysis

Examination of the arrays derived in preprocessing reveals significant unavailability of data. In practice, this is expected as the meters are of different types and have different capabilities for recording data. In the METER READING 180 (consumption) data array, only 180 meters have complete values for a full year, and 411 meters have no readings. In the METER READING 280 (generation) data array, only 97 meters have complete values for one year, and 748 meters have no readings.

Histograms of the missing value counts per meter are shown in Figs. 2 and 3. From these, there are three trends immediately observable. Firstly, approximately half of the meters in both arrays have less than 1000 missing values; this represents over 90 percent data coverage for the majority of the meters. Secondly, approximately one quarter of the meters have little to no readings. Thirdly, the remainder of the meters have missing data somewhere in between these two extremes.

These trends can be further visualised in Figs. 4 and 5. In these figures, the missing data is shown in yellow and meters are sorted from top to bottom in order of increasing data availability. On this basis, the trends can be more accurately described. Firstly, there are "yearly missing" meters where all the data is unavailable. This might represent a case where a meter is newly installed. Secondly, there are "seasonally missing" meters where there is at least one period of 24 hours where the data is completely unavailable. This can correspond with cases where a meter is not "smart" and the values are being reported at weekly or monthly intervals, or where the meter was offline for a period of maintenance. Thirdly, there are "hourly missing" meters where there are intermittent periods of missing data but where there are not 24 consecutive periods of this occurring in the year. This is typical for most smart meters, as readings may be unavailable sporadically but are generally operating normally. Finally, there are "no missing" meters where every period has a reading. This is an ideal case, but clearly is an exception rather than a norm. Each meter is segmented

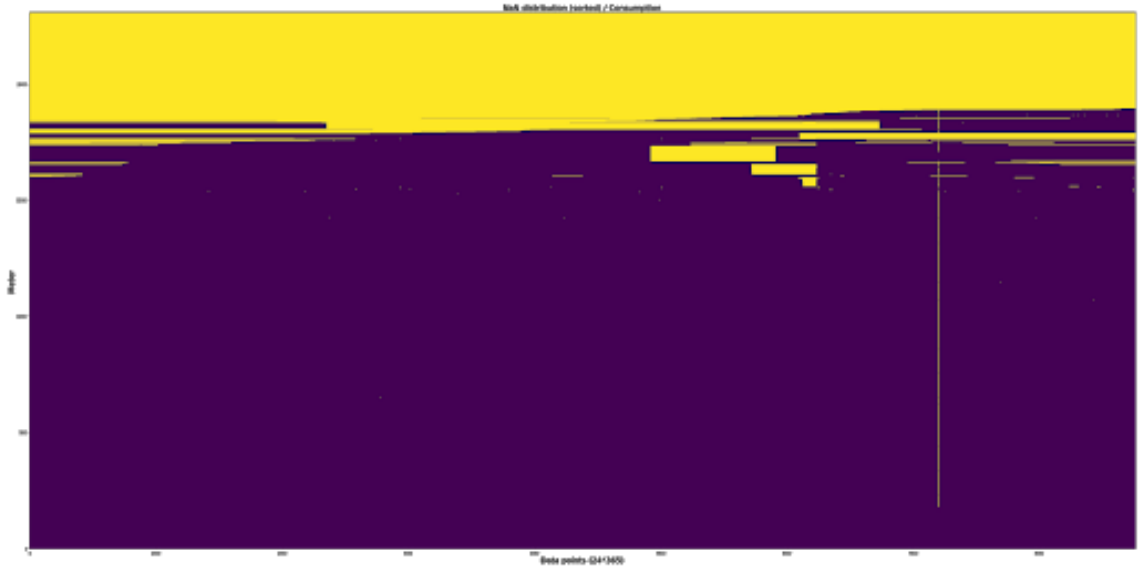

Figure 4: Missing value patterns - Consumption data

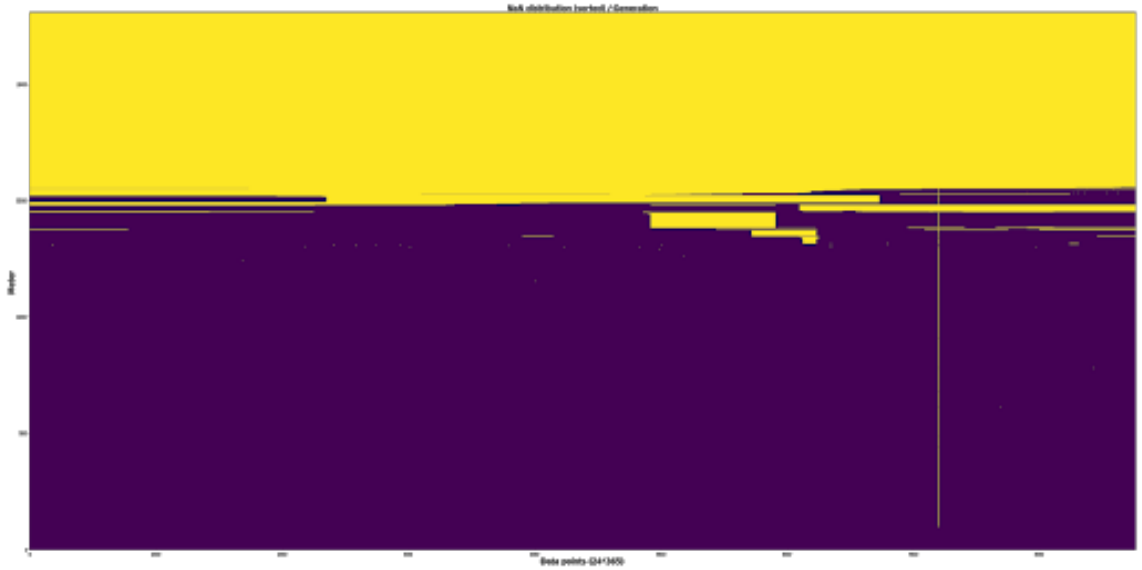

Figure 5: Missing value patterns - Generation data

into one of these trends and the counts are shown in Table 4.

Table 4: Data Unavailability Patterns

Pattern	METER READING 180 (Consumption)	METER READING 280 (Generation)
Yearly Missing	411	748
Seasonally Missing	337	245
Hourly Missing	1379	1217
No Missing	180	97

It can be observed in Fig. 4 that on 28 October there is an event where almost every meter is missing the readings for the day. When segmenting the meters, we do not take this day into account as a consecutive period and fully exclude it from consideration.

One practical note on the array for METER READING 280 is that unavailability of data is more expected than unavailability of data in the METER READING 180 array. This is because generation metering only occurs if a generating asset - such as a solar panel array - is declared, metered and connected to the grid. It is not appropriate to assume that every metered location has a generation asset, but it is appropriate to assume that every location has a meter to measure electricity consumption. In other words, every site will consume energy but not necessarily generate it.

4 Methodology

4.1 Problem Statement and Algorithm Selection

Given the distinct patterns of missing data, it is necessary to create an algorithmic methodology that is appropriate for each.The proposed method and associated baseline to test against for effectiveness is summarised in Table 5.

For small amounts of missing data - such as in the Hourly Missing pattern - it is possible to use regression techniques to interpolate the missing values. Thus, several machine learning based methods for interpolating the missing values directly are evaluated and tested. These methods include Random Forest, Decision Tree, and Support Vector Machine Regression (SVR), as these are commonly examined methods in previous studies. Decision Tree and Random Forest are typical methods evidenced in electricity forecasting case studies. SMAP Energy has used this method successfully in previous research activities, such as those undertaken in Kiguchi et al 2019 [5]. Similarly, SVR methods have been used in multiple cases similar cases related to individual household load forecasting ([3],[4]).

The scope for the regression model is initially limited to the readings immediately before and after, though this can be expanded in future studies. The time of day and the month are also relevant, as are the TARIFF and METERING POINT TYPE identifiers. Therefore, for interpolating a reading for meter m at time t in the consumption dataset C, the inputs to this model $C_m(t-1)$, $C_m(t-2)$, t, month, and a cluster ID combination derived from the combination of TARIFF and METERING POINT TYPE.

For large amounts of missing data, however, these techniques will not be appropriate as they would be very difficult to accurately assess effectiveness and could potentially introduce larger errors in a trained algorithm. Instead, extracting and applying standard load curves via clustering are the investigated methods for the Yearly Missing and Seasonally Missing segments. The inputs to this clustering selection can be fixed known "demographic" identifiers - such as TARIFF and METERING POINT TYPE values - or statistical elements - such as average daily load in a given month. We compare both methods in the case of Seasonally Missing data, but can only examine the Demographic clustering case in the Yearly Missing segment as there is no other data to examine statistically as an input. In both cases, it is useful to compare against a zero-baseline.

The methods selected for evaluation of the Hourly Missing meters can be summarised as follows:

- Average: The total value between two known measurements is distributed evenly over the intermediate unknown periods.

Table 5: Algorithmic Methodology for Segment

Pattern	Proposed Method	Baseline
Yearly Missing	Use representative curve from cluster ("Demographic")	Zero Padding
Seasonally Missing	Use representative curve from cluster ("Demographic" or "Statistical" methods)	Zero Padding
Hourly Missing	Machine Learning Regression	Linear Interpolation

- Random Forest: A Random Forest regression algorithm is trained and used to interpolate missing values.
- Decision Tree: A Decision Tree regression algorithm is trained and used to interpolate missing values.
- Support Vector Machine Regression (SVR): A Support Vector Machine regression algorithm is trained and used to interpolate missing values.

For the Yearly Missing and Seasonally Missing meters, a different method must be employed because of the large amounts of missing data. However, there is a strong precedent for this as distribution network operators have effectively been operating without this data for decades, relying instead on statistical estimates. The proposed methodology in this case will work similarly, relying on constructing average load curves under different categorical combinations.

- Zero Padding: All missing values are set to zero.
- Clustering - Demographic: An average load profile is generated for a cluster set by known variables; in this study, TARIFF and METERING POINT TYPE are used.
- Clustering - Statistical: An average load profile is generated for a cluster set by applying K-means to a selected statistical variable; in this study, average daily load is used.

For the demographic clustering approach, given the data available there are 90 potential load curves for each month generated from the combination of the following variables:

- Day type - 3 values (Weekday, Weekend, Holiday)
- TARIFF - 10 values (B21, B23, C11, C12A, C12B, C12W, G11, G12, G12W, G12R)
- METERING POINT TYPE - 3 values (small, producer, industrial)

An example visual of the generated clusters is shown in Fig. 6, where example curves for January are shown across three TARIFF/METERING POINT TYPE combinations. The top three images show energy consumption data, whereas the bottom three images show generation data. The red line is the average of the values, and the blue window is the range of values observed. Several trends can be immediately observed - for example, the producer average consumption drops in midday as generation rises. This can be suggestive of a portion of the generation being onsite being used to offset consumption during the day.

For the statistical clustering approach, clusters are generated for each month using average daily load as the input. K-means clustering is applied and the elbow method approach is used to select an optimal number of clusters, and this is demonstrated in Figure 7. Using the elbow method, 3 clusters is initially viewed to be optimal number, effectively creating a "low","medium" and "high" daily consumption clusters. This matches inherently with our understanding of the data, as we have known high-energy users amongst the "Industrial" class in METERING POINT TYPE and the "B" TARIFF value classes. However, there is a greater degree of variation in the "low" and "medium" clusters that is not reflected by selecting only 3 clusters. This also makes sense inherently, as individual consumption patterns are more diverse than that reflected by large industrial users. To accommodate for this fully, 10 clusters is selected as the preferred limit.

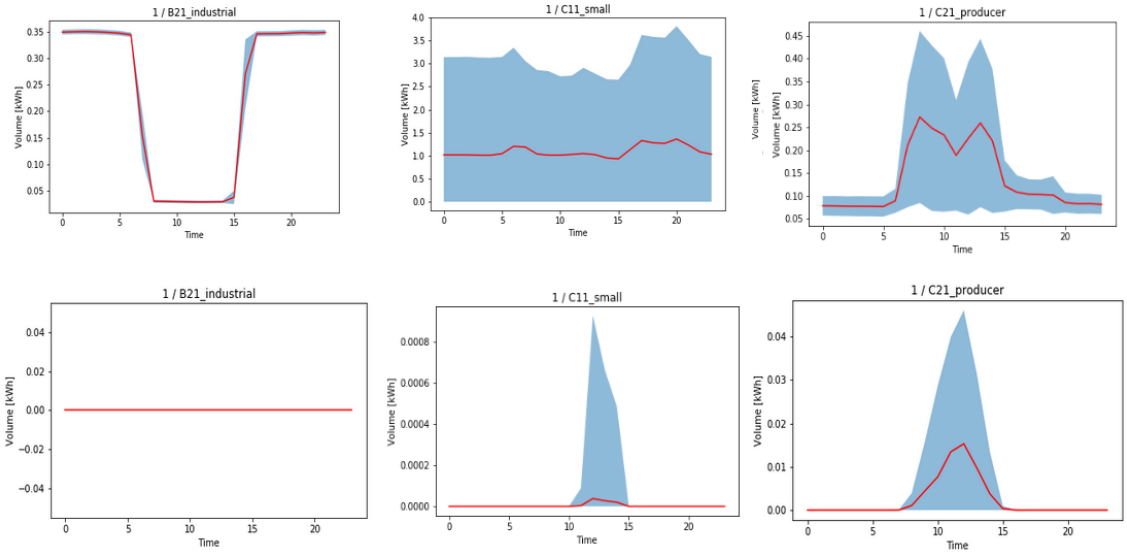

Figure 6: Demographic cluster selection.

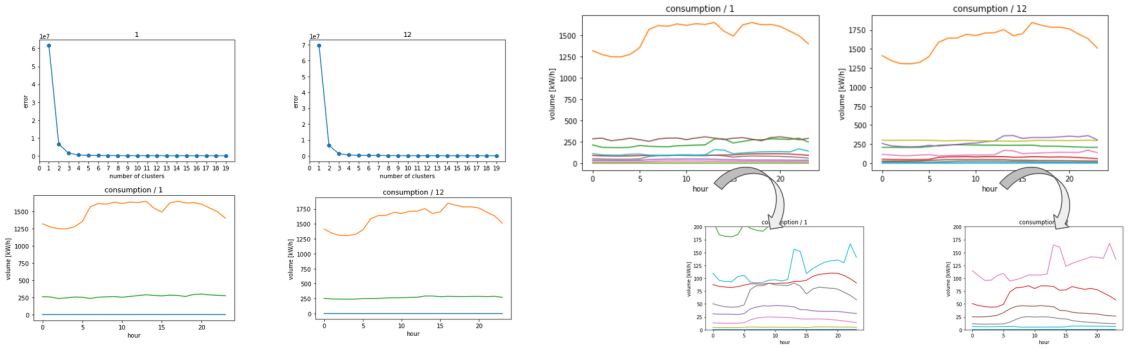

Figure 7: Statistical clustering method cluster selection.

4.2 Evaluation

To evaluate the performance of the algorithm, a subset of the data needs to be created so that the model can be compared against known data. After determining the missing data segment each meter exhibits, the "Yearly Missing" meters are removed, creating a pool of meters where data availability is much higher. Segment labels are then reset and assigned randomly in proportion with original dataset. These relabelled meters are used for testing, where in each period we compare the interpolated value against the known value. For every individual meter, the mean average error (MAE) and mean average percentage error (MAPE) are calculated and compared, forming the basis for assessing how accurate the methodology is for each individual meter point. This is also repeated at the aggregate level - where interpolated values for the meters are summed and compared against the sum of the known readings. This aggregate calculation is also important from the standpoint of the DSO, as it helps understand the potential error of consumption at the substation level and may be used in future risk calculations.

5 Results

5.1 Results Tables - Hourly Missing

The MAE and MAPE results of each method at the individual meter level are shown in Table 6 and at the aggregate level in Table 7. In this case, Random Forest is the preferred model given that it has the lowest MAPE across the consumption meters. As such it is the only model assessed at the aggregate level. As expected, the MAPE across the aggregate case is significantly improved as errors in positive/negative directions cancel out.

Table 6: MAE (kWh) and MAPE - Individual, Hourly Missing

Meter set	Baseline Average	Random Forest	Decision Tree	SVR
Consumption Meters	0.213 (48.6%)	0.150 (34.2%)	0.207 (47.2%)	0.181 (41.3%)
Generation Meters	9.95e-4 (56.2%)	4.53e-4 (25.6%)	4.37e-4 (24.7%)	5.21e-4 (29.4%)

Table 7: MAE (kWh) and MAPE - Aggregate, Hourly Missing

Meter set	Baseline Average	Random Forest
Consumption Meters	3629.97 (31.2%)	1186.72 (10.2%)
Generation Meters	18.24 (38.8%)	6.91 (14.7%)

5.2 Results Tables - Seasonally Missing

The MAE and MAPE results of each method at the individual meter level are shown in Table 8 and at the aggregate level in Table 9. The statistical approach to clustering is determined to be the most effective given that it has the lowest MAPE. It is noted that overall, this approach has a noticeably higher MAPE than compared to the results obtained in the Hourly Missing investigation, but this is to be expected. It is also noted that when considering the zero-padding benchline, MAPE is typically not appropriate to consider as it will always be 100%. Nevertheless, the MAE is still relevant to record here to understand the scope of error.

Table 8: MAE (kWh) and MAPE - Individual, Seasonally Missing

Meter Set	Zero Padding	Clustering - Demographic	Clustering - Statistical
Consumption Meters	0.481 (100%)	0.321 (69.5%)	0.264 (57.2%)
Generation Meters	0.002 (100%)	1.23e-4 (72.1%)	1.14e-3 (66.6%)

Table 9: MAE (kWh) and MAPE - Individual, Seasonally Missing

Meter Set	Clustering - Demographic	Clustering - Statistical
Consumption Meters	2916.69 (40.1%)	2560.29 (35.2%)
Generation Meters	19.53 (52.6%)	15.89 (47.7%)

5.3 Results Tables - Yearly Missing

The MAE and MAPE results of each method at the individual meter level are shown in Table 10 and at the aggregate level in Table 11. Unlike the Seasonally Missing data segment, only the demographic clustering approach can be used here as there is not other potential input. The results in terms of MAPE are generally comparable to those achieved in the Seasonally Missing segment.

Table 10: MAE (kWh) and MAPE - Individual, Yearly Missing

Meter Set	Zero Padding	Clustering
Consumption Meters	0.360 (100%)	0.194 (54.9%)
Generation Meters	0.009 (100%)	5.24e-3 (67.1%)

Table 11: MAE (kWh) and MAPE - Aggregate, Yearly Missing

MAE (kWh) and MAPE - Aggregate, Yearly Missing	Clustering
Consumption Meters	3538.74 (37.7%)
Generation Meters	56.22 (49.5%)

5.4 Benchmarking and Discussion

In order to understand the effectiveness of the results achieved in this study, we compare against the results of several similar academic studies performed over the past few years.

Gajowniczek et al [4] was a study conducted at Warsaw University wherein the forecasting of day-ahead energy consumption of a single Polish household was examined. The data in this study involved two years of one-minute readings of the subject's energy consumption and include sub-circuit monitoring of 20 outlets tied to known household devices and room types. 10 different models - including Random Forest and SVR - were tested, and included data enhancements in the form of identifying behavioural patterns in the device usage. The best MAPE achieved was 23.6% under a neural network model; Random Forest and SVR achieved 29.41% 26.78% MAPE as their best results respectively. This experiment was also repeated on a dataset of 46 US households with 1-hour readings and sub-circuit monitoring of 24 outlets. MAPE for each household observed was widely variable - ranging from 21.28% to 99.46%. The average MAPE for all households was 41.7%.

In Gajowniczek et al [4], the most comparable part of the study involved day-ahead forecasting of 46 households where one-hour data was available for sub-circuit monitoring of 20 outlets. In our study, we similarly have hourly values but are limited to only one reading for consumption, and therefore cannot enhance our data with behaviour in the same way. However, this does not significantly affect the model performance; at the individual level for hourly-missing consumption data, our Random Forest model outperforms the method in this study by delivering 34.2% MAPE compared to 41.7%. This is not only evidence for the effectiveness of the Random Forest model developed in this project, but also suggests that the benefit achieved from sub-circuit monitoring may not be relevant in all cases.

Ding et al [3] investigated 15-minute-ahead forecasting for 21 households distributed across Japan. The dataset involved one-minute readings of energy consumption for a period of 3 months and a supplementary dataset of self-reported daily activities. Using SVR, the best achieved mean MAPE was 42.1%. Similarly to Gajowniczek et al, the additional dataset did contribute to a decrease in MAPE, but the overall performance did not outperform the MAPE achieved in this study.

Veit et al [8] examined day-ahead forecasting in hourly intervals. It examined two datasets: the TUM dataset, which comprised of one-minute readings from an individual household in Germany for a period of 8 months; and the REDD dataset, which comprised of three-second readings from 6 US households for a period of 18 days. The best MAPE achieved in this study was 25% and 46% respectively. As with Gajowniczek et al [4], our RF model MAPE results outperform this when multiple households are examined, though not for a study of a single household with a comparatively deep amount of data.

Barbour et al [1] presents perhaps the most comparable case to the study undertaken in this project. A day-ahead forecasting case was examined for 326 households in Austin, TX, USA where 15-minute readings for 10 months were available. The data was also supplemented by temperature data and assumed a known day-ahead weather forecast. Firstly, a multiple linear regression model was used to forecast the load at an aggregate level, and an MAPE of 9.32% was achieved. In comparison, the RF model developed in our project reported 10.2% achieved when applied to aggregate consumption on hourly-missing segment. Secondly, a clustering model was applied to the households and used to create normalised loads for each cluster. This was used as an input to develop several alternative models for forecasting load at the individual level. The best median MAPE achieved under all examined models is 51.3%. Even though median MAPE is reported here as opposed to average MAPE in our study and others, it is clear that the 34.2% average MAPE achieved with the Random Forest model developed in our study outperforms this.

5.5 Conclusions

In each of these studies, data availability is typically much greater, with several studies having data resolution on the order of seconds. The most comparable case in this regard is comparing against the hourly-missing segment results, and in this regard the Random Forest model developed in this project demonstrates better performance in nearly all cases. However, there are two exceptions to this.

Firstly, on the basis of comparing average MAPE the Random Forest model will typically underperform against studies where there is a single household with a high resolution of data. This is to be expected however, as the models trained in these studies are highly fitted to a single household and have much deeper data to extract potential features from. When the studies are expanded to cases where there is more than one household or the data is not as prevalent, then the Random Forest model will perform better in comparison.

Secondly, in the case of aggregate forecasting in the study with Barbour et al [1], the multiple linear regression model in this study performs slightly better than the Random Forest model developed in this EDI project. It is anticipated that this is primarily due to the impact from including a day-ahead temperature forecast. Temperature forecasts are generally highly correlated with overall energy consumption as devices like air conditioning units and heat pumps are generally energy intensive, and, as the study notes, are generally regarded as accurate within a 24-hour timescale [1]. Rather than indicating a weakness in the Random Forest model developed in this EDI project, this presents a promising area for future enhancement of the model.

It is on this basis that we conclude that the Random Forest model developed in this EDI project is not only sufficiently accurate for the purposes of interpolating hourly-missing data, but represents a significant enhancement compared to these selected studies.

When comparing performance of the clustering methodologies developed for Seasonally and Yearly Missing data segments, the MAPE range achieved for these are generally higher. Comparing to Barbour et al [1] which achieved median MAPE of 51.3%, average MAPE for consumption in seasonally missing and yearly missing segments is 69.5% and 54.9% respectively. The increased MAPE is expected given the inherent difficulty of the task. A reasonable comparison would be comparing month-ahead forecasts with hourly granularity as this replicates a scenario where an electricity consumer is reporting a meter reading each month but it is desired to understand intraday consumption at the hourly level. However, such models are very limited limited as they are focused on market price prediction, work on daily or weekly granularity, or system level forecasting as opposed to individual. Rahman et al [7] confirms this limitation and summarises a handful of studies where mid to long term forecasts at one-hour resolution demonstrated relative errors in excess of 40%-50% ([2], [6], [10]). However, these studies are further limited in their relevance given that they primarily consider single instances of nonresidential buildings. As observed in the selection of clusters in Fig. 7 there are greater variations in the load patterns for residential buildings, which could be a factor behind larger observed MAPE.

In this study, the statistical clustering methodology outperformed the demographic clustering methodology. Average daily energy consumption was selected as the primary variable to perform this clustering, but in practice it could be any number of statistical descriptors. O'Neill and Weeks [[9]] studied the variable importance for 41 categorical variables determined from a survey - including age, income, type of home, and number of different types of devices - and 34 usage variables derived from energy consumption measured via smart meter - including mean, variance, min, and max during peak and off peak times, and relevant ratios of consumption in certain periods - in determining responsiveness to time-based electricity pricing. They determined that the usage variables generally held stronger variable importance than the survey variables. Among these, mean and variance of monthly peak usage were typically among the variables exhibiting the greatest importance This finding is relevant for two reasons. Firstly, it adds to the evidence that data collection outside of energy consumption data is comparatively less important in terms of training models and may in many cases be unnecessary. Secondly, it provides a basis for future investigations, in that elements of mean and variance as well as designations for peak periods may be relevant to incorporate into the model.

It is therefore determined that the clustering methodology developed here is generally performing in line with similar assessments. However, there are a significant number of potential improvements that could be made in future studies to improve the accuracy of this method further.

6 Research Commercialisation

An important part of the European Data Incubator program is taking the resulting research from academic investigation to commercially deployed products. In this section, items related to necessary considerations for deploying the product are addressed.

6.1 User Interface

A requirement of the project is to create a web-based user interface wherein the user is able to select a day and view the full matrix of values - both interpolated and known. Additionally, these values need to be downloadable to a CSV format so that they can be inspected in greater detail. Figure 8 shows a screenshot of the UI meeting these requirements. The date is able to be manually selected, and the results will update accordingly. The data can be downloaded by selecting the download button, and the data will appear in a CSV identical to that shown on the interface but with rows for all meters and additional columns with boolean values indicating whether the data is actual (0) or interpolated (1).

Clicking on the meter ID will open a visual for the day's load curve, and this is demonstrated in Figs. 9 and 10. Figure 9 shows and example for a meter of the METER POINT TYPE "producer". It is immediately observable that generation grows during the day, seemingly offsetting consumption at the same time. Figure 10 shows and example for a meter of the METER POINT TYPE "industrial". Though not a design requirement, the interface colours the relevant peak periods, the times of which are extracted from the TARIFF type.

6.2 Reliability

As addressed in the conclusions section, the results of the algorithm perform at or above a level of acceptability in terms of accuracy.

6.3 Scalability

This study is conducted across a static dataset of approximately 2300 meters with one-year history of hourly readings. An operational scenario would require filling this dataset daily; the current model is able to accommodate this requirement as it completes in 13 hours on cheapest AWS instance and accuracy can be confirmed in 8 hours. Scaling the application further will be subject to a number of variables, including:

- Number of meters and amount of missing data - More meters and more missing data will increase the time.
- Adjusting processing budget - Faster servers can be purchase to decrease time.

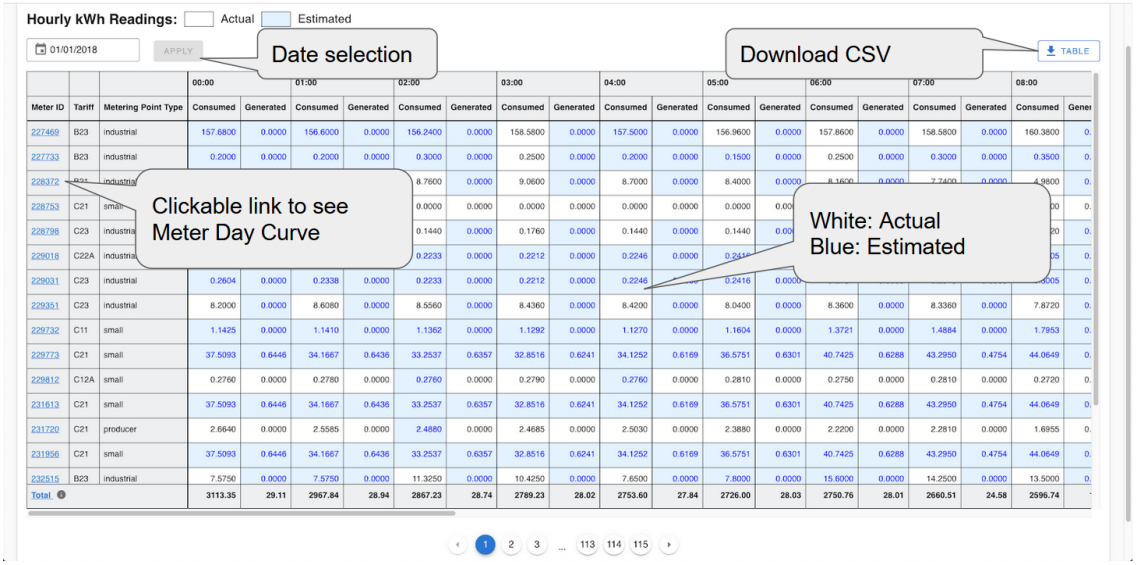

Figure 8: Resulting Data Table Example

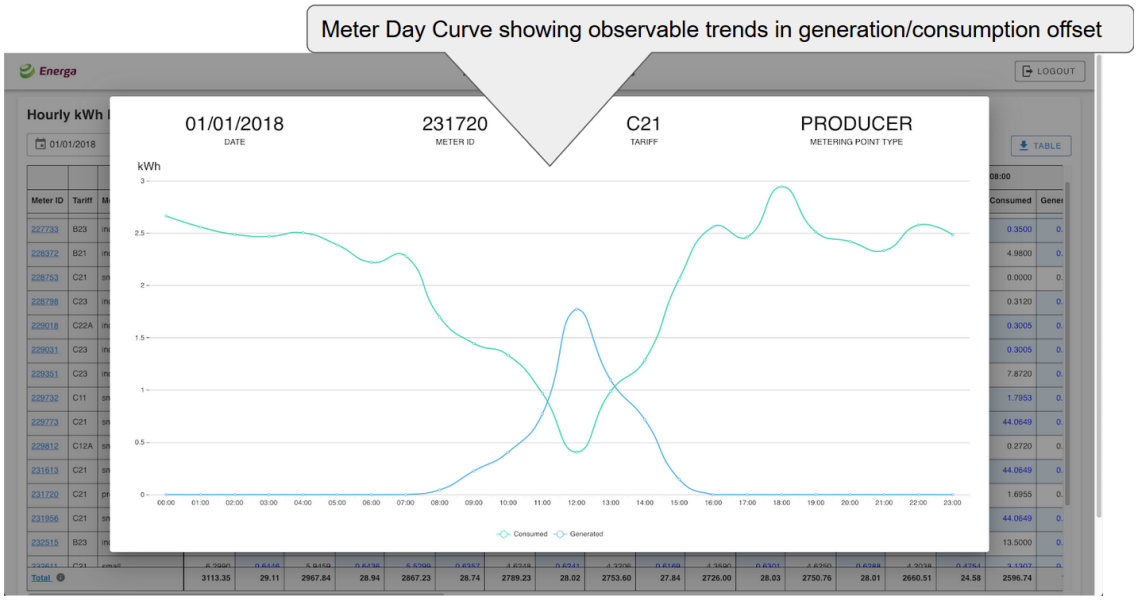

Figure 9: Selected Producer Meter Example

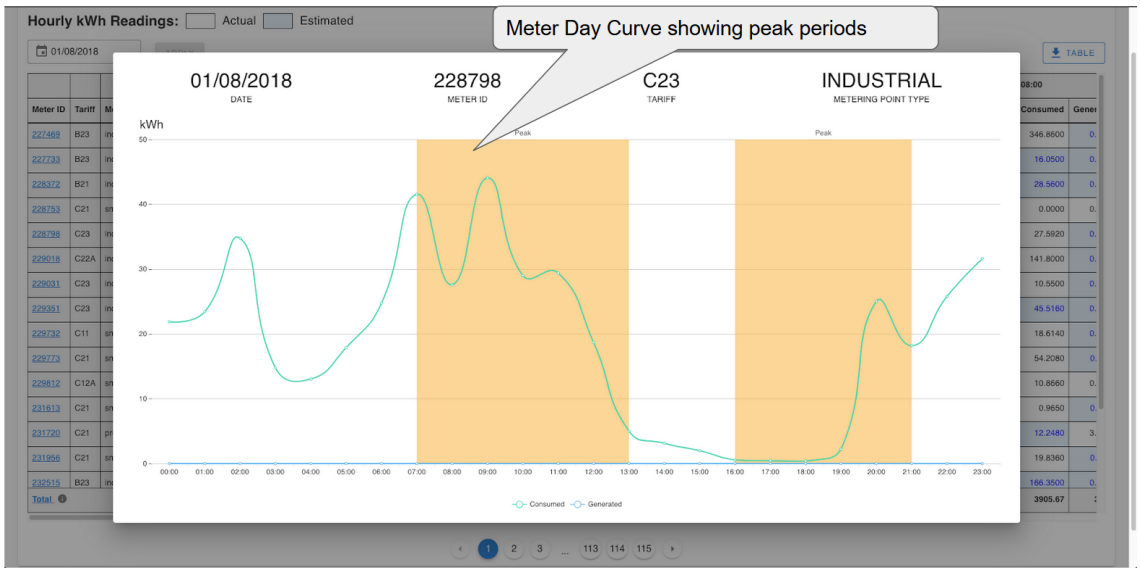

Figure 10: Selected Industrial Meter Example

- Adjusting necessary historical time frame - Decreasing the necessary historical time frame, for example from on year to 9 months or less, will decrease time.
- Building on different language - We have currently built on Python, but using a language more suited for big data applications like Scala or Julia will decrease processing time. However, this will require specific investment and may come at the expense of making the application more unwieldy for future data science investigations.
- Considering Total condition - As noted above, the requirement to ensure the sum of the readings each hour is less than or equal to a known total it not currently addressed. Adding this will involve extra iterations over the data as the condition needs to be checked and, if needed, interpolated values will need to be scaled accordingly.
- Additional datasets added (Total balancing, day-ahead weather, etc) - Incorporating extra datasets may increase the time given the extra data inputs that need to be considered.

6.4 Flexibility

It is expected that this application can be adapted to address other use cases in the DSO context. Potential examples of this include:

- Anomaly detection - In the interpolation process, it is feasible to identify outliers in the data to potentially indicate meter error or potentially fraud.
- Forecasting - Applying the algorithm in the forward direction to estimate future consumption and generation is common and important need for the energy sector.
- Device detection - Specific patterns of consumption and generation can be correlated with the installation of specific devices, such as solar panels. One example of this is in Figure 9.

6.5 Security

The datasets provided in this study have undergone anonymisation and do not contain sufficiently identifiable characteristics as to be classified as personal data. ENERGA has also made this data - minus the area balancing dataset - publicly available for study. Nonetheless, the platform is GDPR compliant and utilises

secure technology, "private by design" principles, and organisational procedures to ensure security and compliance. When handling personal data of any form, SMAP Energy works with the client to ensure that incoming data is as anonymous as possible. This often includes steps such as removing unnecessary metadata (e.g. addresses), as well as giving data points an ID that while relevant to the client cannot be used to associate with any more data than absolutely necessary for performance of the application. Data stored for use with SMAP Energy applications is done so in ways to fully comply with GPDR, as well as our own Access Control, Information Security, and Network Security and Retention policies.

7 Future Work

Future work could proceed from several perspectives. This is broadly examined from the perspective of improving the algorithmic performance of the application and from the perspective of improving the application as a commercial product.

7.1 Improving algorithmic performance

The scope of the investigation in this study was primarily focused on identifying causes of missing data and developing a methodology to handle each. As noted previously, results that are competitive or superior to methods described in academic literature are achieved in this study, but there is room for further improvement.

Firstly, external data sources to provide variables for inputs such as temperature, weather, or the regional total energy consumption, could also be included in the scope for future investigation. These are noted as having value in significant improvements the methods in other studies (Barbour et al [1] and it is widely available for incorporation in a cost-effective manner.

Secondly, internal statistical data may be used to improve performance as well - particularly as it relates to the statistical clustering methodologies. As noted in O'neill and Weeks ([9]), statistical variables related to mean, variance, and max consumption are among the most important variables to consider on a model.

Thirdly, the current algorithm utilises one hourly reading before and after the target period to generate the interpolated values. The impact of expanding this range could be examined in the future.

Fourthly, it is recognised that in some cases, our preprocessing decisions will neglect potentially valuable interim information that could be obtained from available 15-minute interval readings. Incoporating these readings in some form may be of value as well.

Fifthly, there are a number of methods that could possibly be explored in addition to these - including artificial neural networks, etc. We initially do not consider these as our previous in this investigation in order to adhere to the time constraints of the program and because previous studies have shown that other models like Random Forest typically outperform these methods (Kiguchi et al [5]). However, it may still be worth considering these in future work.

7.2 Product Enhancements

This application is developed under the general criteria of a minimum viable product, and further work is needed to deploy this application in a fully operational scenario. Firstly, the application is operational on a static dataset, but an operational scenario would require recomputing and updating results daily. Establishing the full data pipeline - including addressing the newly identified preprocessing steps such as those described in Figure 1 - is an important and time-intensive step that may involve not only further development of the software architecture but also specific reporting subapplications.

Secondly, there are UI enhancements that can be added such as meter filtering, ordering, and searching may be required in the future as well. These enhancements would be prioritised following the guidance and needs of trial clients.

Thirdly, optimising the backend performance of the application would be valuable in the long term as it will minimise costs incurred in processing. This would require a specific technical investment - for example, by rewriting parts of the app in languages like Scala which are better suited for big data processing - and timing this investment will be important so as to not prohibit flexibility in adapting the platform to client requirements.

8 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779790. Miroslaw Matusewicz and Krzysztof Profis from ENERGA SA provided useful insight on the dataset and commentary on the project. Tom Haake, Laia Tarragona, Dr. Diego López-de-Ipiña González-de-Artaza, and numerous other members of the EDI and University of Deusto staff were actively involved in oversight and administration for the project.

9 Relevant Links

The UI demo is available to approved parties at the following link: https://edi-energa.smapenergy.com/. To request access, please send an introduction email to info@smapenergy.com.

References

[1] Barbour E and Gonzalez M; "Enhancing household-level load forecasts using daily load profile clustering", 2018, *The 5th ACM International Conference on Systems for the Built Environment.*

[2] Fumo N and M. Rafe Biswas M; "Regression analysis for prediction of residential energy consumption", 2015, *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343.

[3] Ding Y, Borges J, Neumann M, and Beigl M; "Sequential pattern mining - A study to understand daily activity patterns for load forecasting enhancement", 2015, *IEEE First International Smart Cities Conference.*

[4] Gajowniczek K and Zabkowski T; "Electricity forecasting on the individual household level enhanced based on activity patterns", 2017, *PLoS One.*

[5] Kiguchi Y, Heo Y, Weeks M, and Choudhary R; "Predicting intra-day load profiles under time-of-use tariffs using smart meter data", 2019, *Energy*, vol. 173, pp. 959-970.

[6] Mocanu E, Nguyen PH, Gibescu M, and Kling WL; "Deep learning for estimating building energy consumption", 2016, *Sustainable Energy*, Gridsand Networks vol. 6, pp. 91–99.22

[7] Rahman A, Srikumar V, and Smith A; "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks", 2017, *Applied Energy*, pp. 1-24.

[8] Veit A, Goebel C, Tidke R, Doblander C, and Jacobsen H-A; "Household electricity demand forecasting: benchmarking state-of-the-art methods", 2014, *Proceedings of the 5th international conference on Future energy systems.*

[9] O'Neill E and Weeks M; "Causal Treee Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes", 2018.

[10] Yun K, Luck R, Mago PJ, and Cho H, "Building hourly thermal load prediction using an indexed ARX model", 2012, *Energy and Buildings*, vol. 54, pp. 225–233.