Methodology for close to real time profiling of aggregated demand using data streams from smart meters

METHODOLOGY FOR CLOSE TO REAL TIME PROFILING OF AGGREGATED DEMAND USING DATA STREAMS FROM SMART METERS

Kuanhong Li [1], *Jelena Ponoćko* [1,*], *Lingyue Zhang* [1], *Jovica V. Milanović* [1]

[1] *Electrical Energy and Power Systems Group, University of Manchester, Manchester, UK*
**jelena.ponocko@manchester.ac.uk*

Keywords: Load profile, smart meter, data streams, aggregated demand

Abstract

This paper discusses potential improvement in accuracy of estimation of load profiles at substation/aggregation point if the demand data is collected directly from smart meters rather than from balancing meters at bulk supply points. It proposes a bottom-up approach for development of daily load curves for domestic load sector by aggregating data coming as real-time data series from smart meters. In order to illustrate the concepts an assumption is made that all the smart meters in an area have the ability to measure instantaneous real power demand of each individual appliance. Following this, a probabilistic bottom-up approach is applied to generate reactive power demand at the point of aggregation. It is further assumed that the collected data streams have different sampling steps and that there are some missing data in recorded data streams. Different data conditioning methods are used to investigate the accuracy of demand aggregation at different aggregation levels not only in terms of total demand but also in terms of demand categories and controllable and uncontrollable demand.

1 Introduction

Future power networks are set to become significantly different from the existing, particularly in the areas of electricity generation and distribution. Generation will increasingly rely on low carbon technologies (LCT), i.e. renewable sources and energy storage systems scattered across all voltage levels as illustrated in Fig.1 [1].

Distribution network is becoming more active in balancing generated and consumed power through active participation of the end-users through demand side management (DSM). DSM is principally driven by market (reduction of the electricity generation price) and reliability (avoiding transmission/distribution lines congestion) objectives [2].

Load control, as direct "consequence" of DSM involves disconnection and/or shifting the connection time of controllable loads in order to provide flexibility. During the day, portion of controllable loads changes due to end user activities, hence timely assessment of the amount of available controllable load is essential for facilitating load scheduling tasks. Controllable loads, together with distributed generation from renewables, form the actual flexibility of the end-users. This flexibility is in fact the amount of consumption or production that can be shifted in time and used for mitigation of grid congestion at both distribution and transmission level, for reduce the electricity bill for consumers by following dynamic pricing and energy efficiency services, or for providing balancing services to balancing actors or directly to the transmission system operator (TSO) [3].

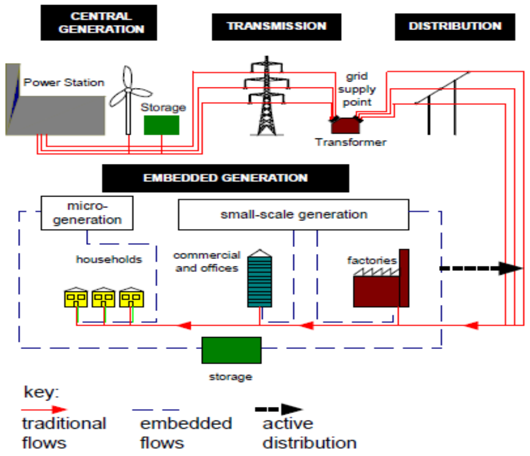

Figure 1 Distributed generation system [1]

Load and distributed generation forecast is essential for reliable flexibility assessment, as it is taken as the reference to calculate the effects of flexibility, from both network and market perspective [4]. Data used for load forecasting is currently taken from substation points where monitors measure total consumption (real and reactive power), often involving different load classes/sectors (residential, industrial and commercial). These measurements, however, do not discriminate between different load categories (induction motors, lighting, resistive loads, etc.). In this respect, smart metering system that has been increasingly developed throughout the world enables remote acquisition of information about daily load variation at users' premises. Depending on the future distribution system architecture, the low-level data acquired in this way will be aggregated either at distribution system operator (DSO) or aggregator point.

This paper therefore proposes a bottom-up approach for development of daily load curves at given aggregation point by aggregating individual load curves coming as real-time data series from smart meters. It establishes potential improvement in accuracy if the demand data is collected from smart meters rather than from balancing meters at substation (bulk) points. To illustrate the concept an assumption is made that all smart meters have the ability to measure daily demand for each individual appliance. Measured active power demand of individual appliances is then aggregated by summing up load of different appliances, differentiating between six load categories, as follows:

1) Resistive loads: hob, oven, iron, electrical water heater, etc.;
2) Switch-mode power supply (SMPS): TV, microwave, electronic devices, etc.;
3) Lighting: incandescent light bulbs, compact fluorescent bulbs, etc.;
4) Single-phase constant torque induction motors (CTIM1): washing machine, tumble dryer, dish washer, etc.;
5) Three-phase constant torque induction motor (CTIM3): electrical space heater;
6) Single-phase quadratic induction motor (QTIM1): fridge, freezer, etc.

Categories 4, 5 and 6 are considered to be controllable, loads in category 1 are partly controllable and categories 2 and 3 are uncontrollable. Since smart meters that are presently being deployed, and which will therefore remain in service for foreseeable future, most commonly do not measure reactive power consumption of end-users, a probabilistic approach is applied to generate reactive power daily demand at aggregation point using range of possible values of power factor (PF) of individual appliances.

To illustrate the approach, the input data are obtained through aggregation of low-level data generated using CREST tool [5] which allows simulation of 34 individual home appliances' daily load profiles in terms of active power, with one-minute resolution. The information provided by the CREST tool is then used to generate decomposed daily load profiles in terms of load categories in both active and reactive power. The paper builds on the methodology described in [6], where a probabilistic approach and Artificial Neural Networks (ANN) were used to estimate and predict contribution of different load categories to daily loading curve based on half hourly meetering data and bulk supply points and general information about customer composition available from customers surveys. The use of realistic data obtained thorugh aggregation of huge data streams from smart meters, as demonstrated in this paper, will certainly improve the accuracy of estimation of load composition at aggregation point as well as identification of controllable and uncontrollable portions of load which will facilitate more effective DSM.

2 Decomposed Daily Load Curves (DDLC)

Having in mind that the electricity consumption of individual residence depends on family composition, lifestyle, mixture of electrical appliances, etc., residential sector's daily electrical behaviour is quite random and therefore requires probabilistic analysis [7]. At the same time, large number of residential consumers presents a significant portion of the total power consumption in an area. In the UK, for instance, residential (domestic) sector is the largest final user of energy, presenting around 30 % of overall consumption, with industrial and commercial sector following with 26 % and 21 %, respectively [8]. This presents high potential for domestic sector involvement in future DSM programs.

2.1 Total Daily Load

Daily load pattern of the end-users is typically reconstructed according to their monthly energy consumption and typical load profile, i.e. load class they belong to [9]. However, even when they belong to the same load class or commercial code, consumers' load patterns might be very different [10]. As shown in [9], there was a limited correlation between consumers' activity type (i.e. load class) and their load pattern. It has been found that differences between individual daily load curves do not influence significantly the shape of aggregated load curve when it involves large number of consumers, for instance at substation point. However, in case of aggregators (potential new actors in future distribution grid market), the aggregation might include smaller number of customers and thus be more affected by their varieties in power demand. Therefore, a more detailed analysis should be performed over customers' load patterns in order to enhance accuracy of load estimation and forecast, especially at lower levels of aggregation.

Fig. 2 shows daily load profiles on two working days and one non-working (holiday) day, taking in account three aggregation levels: 10, 200 and 1000 houses. The load profiles were generated using CREST tool. As it can be seen from the graph, there is a significant randomness in daily load profiles at the aggregation of 10 houses, irrespectively of the day type (working day or holiday). Differences in load patterns between working days decrease as the aggregation level grows, which is very clear at the aggregation of 1000 houses. The graph also shows difference in load profile between working days and holidays. This difference, larger consumption during working hours (8 a.m. to 4 p.m.), is more observable, at higher aggregation levels (200 and 1000 houses). Therefore, in case of DSM at local level, i.e., for smaller number of consumers, the load forecasting becomes much more complex.

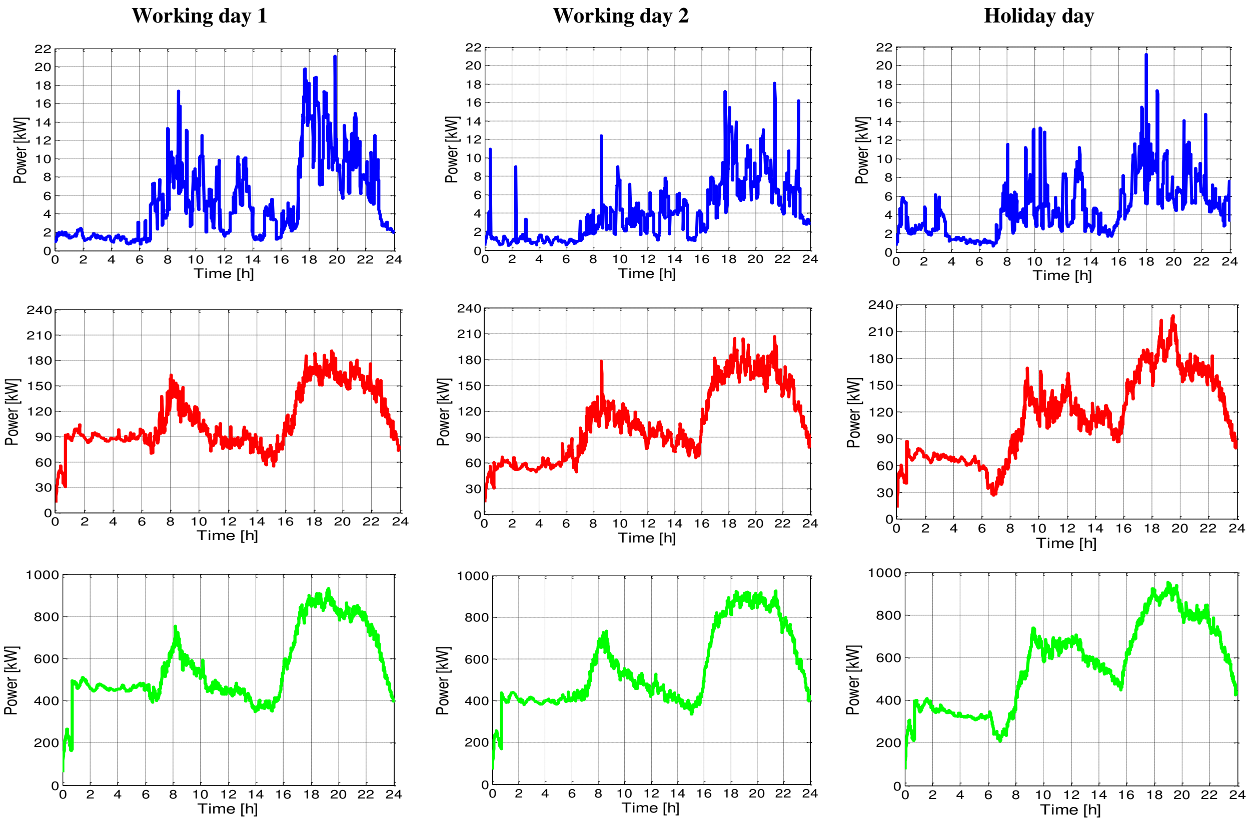

Figure 2 DLC for aggregation levels of: 10 houses (blue), 200 houses (red) and 1000 houses (green)

2.2 Load Decomposition

Going further from total load prediction, even more chalenging is to estimate impacts of different load categories (induction motors, lighting, resistive loads, power electronics, etc.) at the aggregation point level. Potential flexibility in terms of different load types by areas or groups of customers has only been assessed through surveys, which is time demanding and not necessarily sufficently accurate. Since flexibility depends on the availability of controllable loads, there is a need for isolating (disaggregating) impacts of different categories/types of load at the aggregation point according to available measurements. With the assessment/prediction of available demand flexibility it is possible to adjust the level of incentives that need to be offered in order to attract more end-users to participate in DSM actions and such provide required support to distribution network. This would result in shifting (washing machines, dryers, dishwashers, boilers) or curtailing (AC, space-heaters) [2] parts of the load when needed.

Fig. 3, adopted from [11], shows the expected format of the output, where load curves of different load types (a) are summarised into corresponding load categories (b). Further grouping should show percentages of controllable and uncontrollable load, obtained by summing the percentages of controllable/uncontrollable load categories.

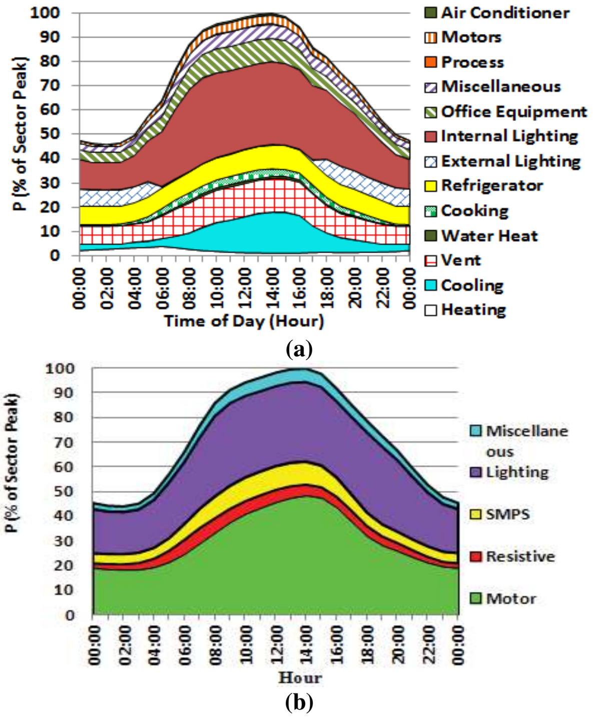

Figure 2 Decomposition of load curve in terms of load types (a) and load categories (b) [11]

Another benefit from information about demand (real and reactive power) composition in terms of load categories is that it can be further utilised for the estimation and prediction of the dynamic response of demand [12]. In case of a voltage disturbance in the network, i.e. voltage step change, there is a dynamic response of the load (real and reactive power) which may influence the voltage and angular stability of the system. This response is highly dependent on the composition of load (shares of different load categories in total demand at given point in time) and the voltage change. Therefore, it would be highly useful to assess the time change of real and reactive power and composition of loads in order to be able to predefine the desired composition of load categories whose demand response would not harm stability of the system in case of a voltage disturbance (small and large).

3 Case studies

3.1 Processing and Conditioning Imperfect Data

According to [13], up to 20 % of active load measurements at substation points are inaccurate. This consequently affects to a large extent the accuracy of load forecast. This case study focuses on the effect of missing data on the accuracy of demand decomposition. Some studies have treated missing data by simple elimination and consequent data size reduction or by imputing mean values of available data in places where the data is missing [14, 15].

Several techniques have been investigated in this study in order to improve the accuracy of data restoration. Three data imputation methods, namely, simple linearization, locally weighted scatterplot smoothing (LOESS) [18] and K-nearest neighbor (kNN) [16, 17] are illustrated here as representatives of different classes of possible methods. Simple linearization is taken as the simplest of methods for restoring missing data by simply connecting existing samples by a linear line whenever there is a missing sample/set of samples between them. LOESS can be used in cases when data streams have “NaN” values, although it shows higher errors for large portions of missing data (relative error can be up to 30 % in case when 20 % of data is missing in a data stream). kNN method requires a set of training data which is then used to impute missing samples in the test data using distance (e.g., Euclidean) minimization. Part of this study will adopt an improved version of kNN, weight adjusted kNN (WAkNN) [19] - in this case, if a training object has smaller distance from the test object, this training object will have higher weight. Eventually, the missing part will be replaced with the sum of weighted training objects that were closest to the test object. Furthermore, two ways of applying KNN method are considered:

- kNN based on total samples: all available samples in a daily load curve (DLC) of a house are set as the test object;
- kNN based on adjacent samples: only a set of two closest bordering samples around the missing part of a DLC is taken as the test object (in this case WAkNN method is applied).

3.1.1 Methodology

An assumption is made that perfect data presents fully accurate smart meter data with one-minute sampling step. Another assumption is that in an aggregation area some smart meters have different sampling step of 10, 30 and 60 minutes including that some have missing values in their data streams. In order to present realistic issues with aggregation of incoming data streams, total number of smart meters in an aggregation area is divided into four groups:

- Group A: 1-minute resolution smart meters with 20-30 times, 10-30 samples missing from a total of 1440 samples (daily load);
- Group B: 10-minute resolution smart meters with 10-20 times, 1-6 samples missing from a total of 144 samples;
- Group C: 30-minute resolution smart meters with 1-3 times, 2-6 samples missing from a total of 48 samples;
- Group D: 60-minute resolution smart meters with 1-2 times, 2-4 samples missing from a total of 24 samples.

3.1.2 Results

Following the random distribution of the smart meter groups (A to D) within an aggregation of 1000 homes (i.e. 1000 daily load profiles) on a working day in January (which was set in CREST tool), five different approaches are used to restore missing data. They are summarised in Table 1. Fig. 4 illustrates the original DLC for 1000 houses and five restored DLCs using approaches described in Table 1.

Table 1 Approaches used for restoration of missing data

	Different Sampling Rates	**Missing data**
1	Linearization	Replace with zero
2	Linearization	Linearization
3	LOESS	LOESS
4	kNN based on total samples	kNN based on total samples
5	kNN based on two adjacent samples	kNN based on two adjacent samples

Apart from the approach 1 (the simplest), all other data restoration methods result in acceptable (visual) resemblance to the original data set. In order to support this conclusion, a comparison of errors is presented in Table 2, showing maximum (E_max) and average (E_ave) values of relative errors across 1440 samples of the original data, as well as root mean square error (RMSE) for different levels of aggregation: 1000, 200 and 50 homes. RMSE is normalized based on the mean daily power value of the original data set.

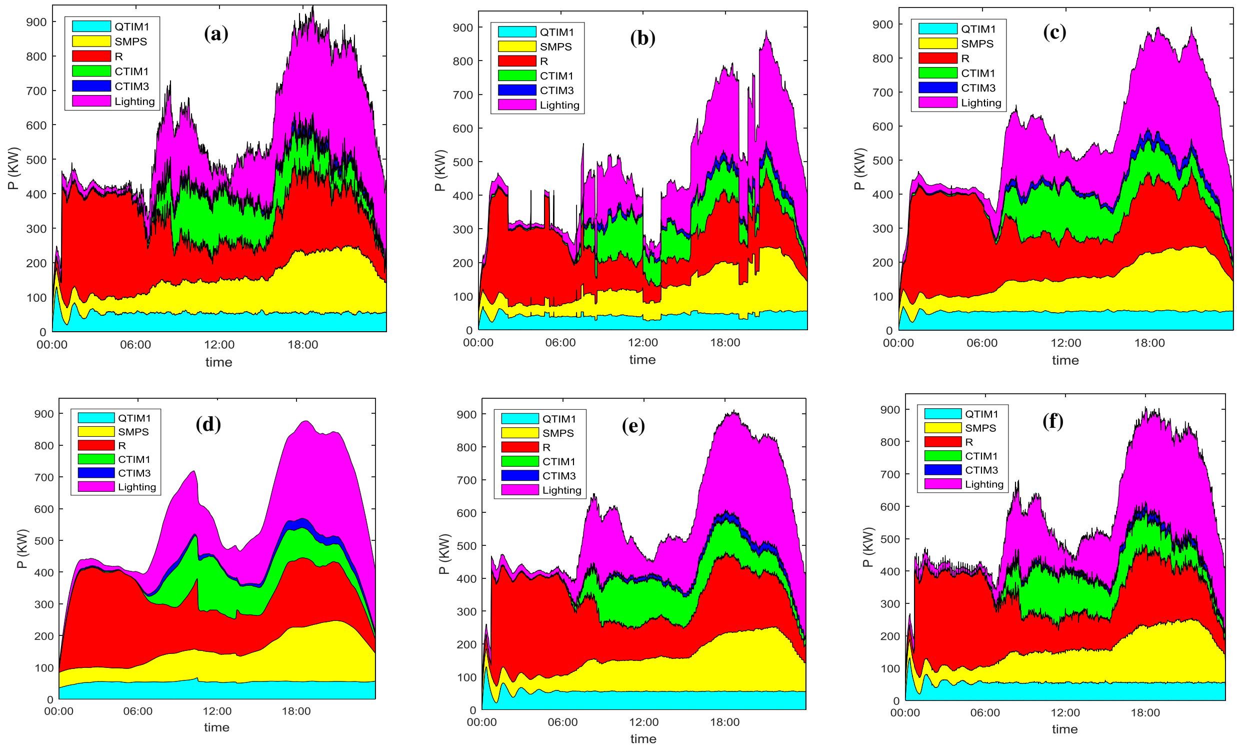

Figure 4 DDLCs of original full data streams (a) and processed data streams by treatment 1 (b), treatment 2 (c), treatment 3 (d), treatment 4 (e) and treatment 5 (f) for the sum of 1000 houses based on 6 load categories; Legend: QTIM1 - Single-phase quadratic induction motors, SMPS - Switch-mode power supply, R - Resistive loads, CTIM1 - Single-phase constant torque induction motors, CTIM3 - Three-phase constant torque induction motors, Lighting

Even though the approach 5 showed the best results for the aggregation of 1000 end users, in cases of aggregation at 200 and 50 house level, the approach 2 (simple linearization) showed higher accuracy. This brings the conclusion that for lower level of aggregation, there is no need for time consuming data mining methods to deal with missing data. The highest error for all three aggregation levels resulted from replacing the missing data with zero values, i.e. not restoring missing data at all.

Since the main aim of the decomposed DLC is to estimate the amount of controllable load, the division into controllable/uncontrollable load is performed over aggregation of 1000 homes using results of data restoration approaches 1 and 5, to illustrate maximum and minimum errors, respectively. Fig. 5 presents the original data set with categories classified into controllable and uncontrollable load (a), followed by the same classification done after treatments 1 (b) and 5(c) and corresponding time-varying relative errors (d and f) for total load and controllable/uncontrollable load. As seen from the figure, in case of approach 1, relative errors of total load and controllable/uncontrollable load are quite high (around 50 %). Relative error for total load curve over the whole time frame in case of treatment 5 is drastically smaller (up to around 10 %) compared to 1. On the other hand, errors in assessment of controllable load are larger, reaching around 25 % at some time steps. Further analysis should be performed to investigate if there is any correlation between the period of the day and increased error in estimation of the controllable load share.

Table 2 Three types of errors for 5 treatments in case of 1000, 200 and 50 houses

Treatment	E_max (%)	E_ave (%)	RMSE (%)
1000 houses			
1	49.73	18.08	20.94
2	18,18	3.70	4.97
3	37,59	5.72	7.67
4	23.33	2.72	3.39
5	**14.07**	**2.57**	**3.23**
200 houses			
1	66.67	16.71	23.16
2	**34.17**	**6.55**	9.58
3	40.45	7.73	10.78
4	66.59	7.89	9.29
5	48.53	6.65	**8.45**
50 houses			
1	78.88	18.39	27.13
2	**54.27**	**7.24**	**12.29**
3	84.68	10.87	15.73
4	95.58	14.66	17.47
5	100	10.37	14.67

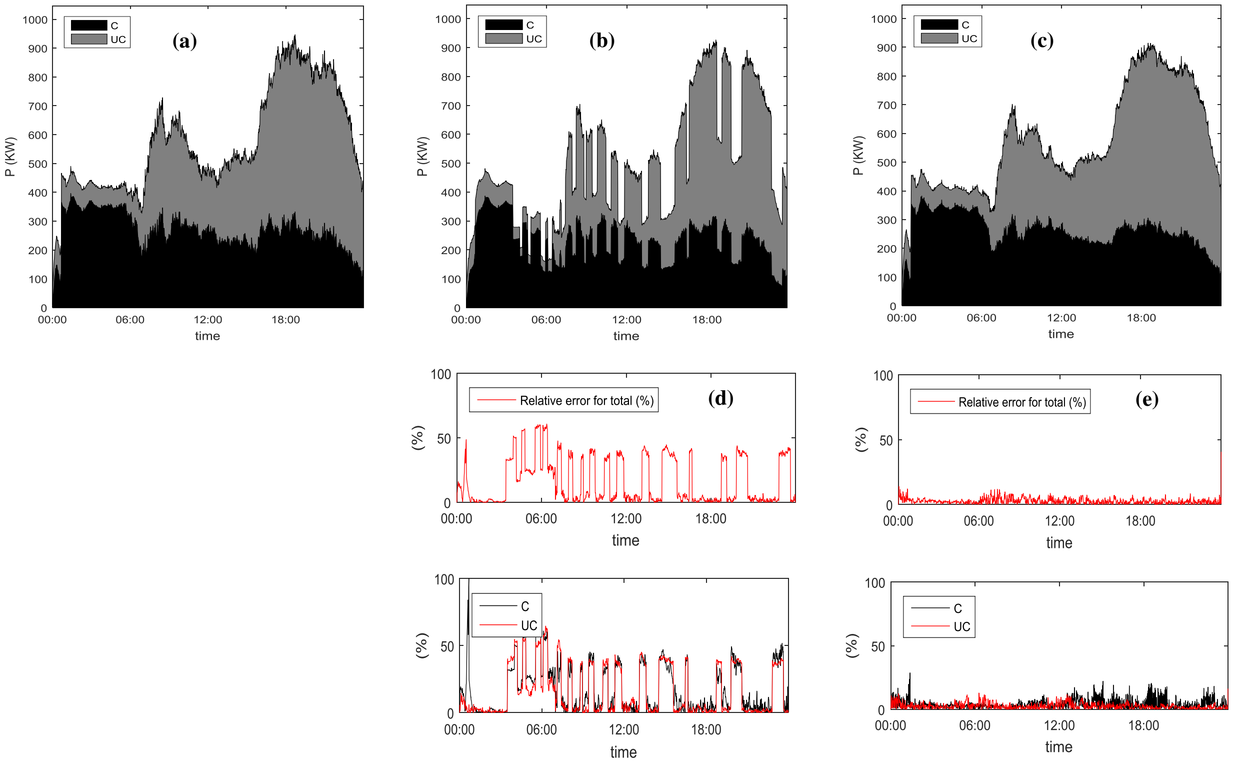

Figure 5 DDLCs of: (a) perfect data streams, (b) incomplete data streams conditioned using approach 1 and (c) incomplete data streams conditioned using approach 5 for the aggregation of 1000 houses, followed by corresponding errors (d and e) for total load and controllable/uncontrollable (C/UC) load

3.2 Probabilistic Generation of Reactive Power Load

In order to make a complete profile of aggregated load in an area, both active and reactive load measurements are needed. In most cases, smart meters do not collect reactive power data, which brings the need for probabilistic assessment. A bottom-up approach is followed in this case too, by considering the range of possible power factors (PF) for different home appliances in CREST library. In order to present reactive power consumption more realistically, with variable PF across appliances of the same type, PF value for each appliance in each time step is randomized 100 times for all the 1000 customers analyzed. By aggregation of probabilistic ranges of Q for each device in every household, a range of probabilistic daily reactive load curves is obtained.

Following this a method needs to be adopted to generate reactive power for all the appliances. The following four options are considered:

1) average value, based on average PF taken from typical range of PF for each appliance,
2) most probable value of the probabilistic range of reactive power obtained with randomization,
3) mean value of the probabilistic range of reactive power,
4) typical value - most commonly used PF from the typical range of PF for each appliance

Typical range of PF for each appliance is adopted by considering values provided at different manufacturers' websites. Taking lighting load as an example, Fig. 6 shows four different reactive load curves for an aggregation of 1000 customers. As shown in the figure, the biggest discrepancies occur when the most typical value of PF is adopted for each appliance.

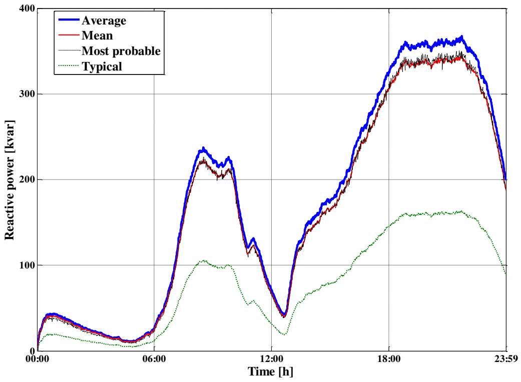

Figure 6 Different probabilistic reactive load curves

As for the other three solutions, due to relatively small differences between them, the most probable reactive load curve is adopted as the "original" reactive load curve. Based on this, a decomposed reactive load curve is obtained and illustrated in Fig.7. It shows participation of 6 load categories (including notionally resistive loads - which are not modelled as purely resistive) -Fig. 7 (a), and controllable/uncontrollable load - Fig. 7 (b), within the total daily reactive load.

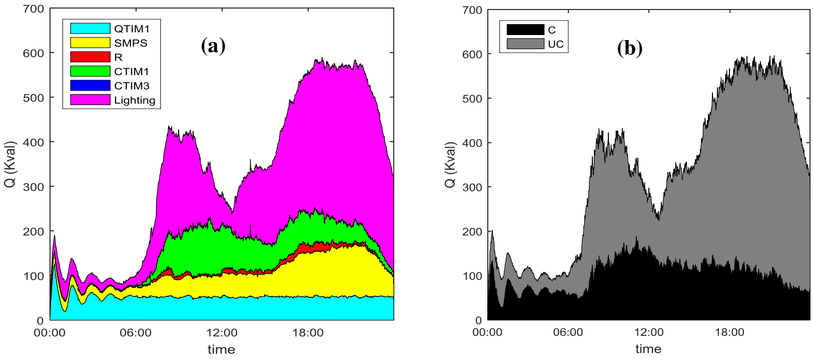

Figure 7 DLCs of generated reactive power decomposed into: (a) categories; (b) controllable/uncontrollable loads

4 Analysis of Results

Figure 8 illustrates RMSE for different aggregation levels (50, 200 and 1000 customers) in cases of incomplete active power data before any processing (a), incomplete active power data processed using approach 5 (b) and reactive power derived probabilistically from active power data after it was processed using approach 5 (c). The errors are shown for total load, controllable/uncontrollable (C/UC) load and six load categories. Incomplete and processed active load curves were compared to the original, complete active load curve.

RMSE_P1 presents RMSE of incomplete data and RMSE_P2 presents RMSE after using approach 5. Regarding reactive power, two curves are compared to obtain RMSE_Q:

1) Reactive load estimated using the full data set of active load curve; adopting the most probable reactive power curve as the “original” one;
2) Reactive load estimated using active load curve with missing data restored by approach 5; the most probable curve was adopted for reactive load curve.

In case of incomplete active load data (Fig. 8a), the error does not decrease monotonously with higher level of aggregation. The reason is randomness in selection of homes (i.e. smart meters) for different aggregation levels, which is why there were probably more smart meters with full (minute-based) data in 200 homes dataset. In cases of incomplete active load data processed using approach 5 (Fig. 8b), there is a clear decrease in RMSE over the whole range of load categories. The largest improvement in accuracy is shown for the case of aggregation of data from 1000 homes, which justifies the application of kNN method for missing data at higher aggregation levels. Furthermore, the application of kNN method also improves the accuracy of estimation of controllable/uncontrollable load.

Analysing the distribution of RMSE over the range of load categories, the highest errors are notable in resistive load category, both before and after data treatment. The most probable reason for this is substantial randomness in the use of resistive domestic appliances. On the other hand, electrical space heating (CTIM3 category) has shown the smallest errors due to seldom use of this type of loads in the analysed residential district. As for the reactive load (Fig. 8c), dataset derived from conditioned incomplete active load data gave acceptable RMSE of less than 5 % for most of the load categories.

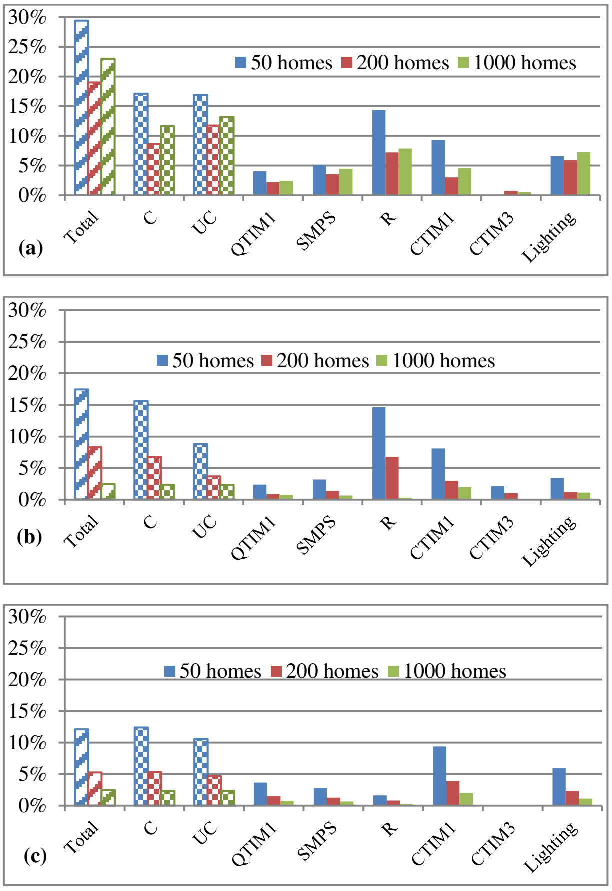

Figure 8 RMSE for incomplete load data: (a) RMSE_P1, (b) RMSE_P2, (c) RMSE_Q

5 Conclusion

This paper presented a bottom-up approach for development of daily load curves by aggregating individual load curves coming as real-time data series from smart meters in residential load sector. Aggregation was done on appliance-level, followed by load category-level, assuming that smart meters could measure active power consumption of each appliance. In order to analyse a realistic situation, some smart meters were chosen to have different sampling step, as well as randomly missing data of different size. From several data restoration/conditioning methods considered, the kNN method resulted in the highest accuracy in estimation of both, total demand and demand composition for aggregation of large number of individual measurements.

Following the assumption that smart meters cannot measure reactive power consumption, a probabilistic approach was developed to generate corresponding reactive load data. Two datasets were generated, one based on full active load data and one, more realistic, based on load curve with restored missing data. It was demonstrated that sufficiently accurate decomposed daily loading curves for reactive power can be developed from real power data sets after restoration of missing data.

The case studies presented in this paper were based on simulated realistic measurement data, but not on real measurements. Once fully developed and validated, the methodology will be tested at later stage using real data streams coming from smart meters. Future work will focus in particular on conditioning of data streams in real-time, to facilitate short-term load forecast and real-time load decomposition, as basic services of DSM.

6 Acknowledgments

This research is partly supported by the EU Horizon 2020 project “Nobel Grid”, contract number 646184.

7 References

[1]"UK Electricity Networks, Postnote, Parliamentary Office of Science and Technology ", [Online].Available: http://www.parliament.uk/documents/post/pn163.pdf

2001.

[2]I. Cobelo, "Active Control of Distribution Networks," PhD thesis, School of Electrical and Electronic Engineering, The University of Manchester, 2005.

[3]"Consolidated View on the ETP SG (European Technology Platform on Smart Grids) on Research, Development and Demonstration Needs in the Horizon 2020 Work Programme 2016-2017," 2015.

[4]G.Chicco, "A Multi-faceted View on the Characterisation of Electrical Demand," *Presentation at the University of Manchester,* 2015.

[5]I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy and Buildings,* vol. 42, pp. 1878-1887, 2010.

[6]Y. Xu, "Probabilistic Estimation and Prediciton of the Dynamic Response of Demand at Bulk Supply Points," PhD thesis, School of Electrical and Electronic Engineering, The Univeristy of Manchester, 2015.

[7]E. Carpaneto and G. Chicco, "Probabilistic characterisation of the aggregated residential load patterns," *Generation, Transmission & Distribution, IET,* vol. 2, pp. 373-382, 2008.

[8]"Digest of United Kingdom Energy Statistics 2015," Department of Energy and Climate Change2015.

[9]D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible consumers," in *Power Tech Conference Proceedings, 2003 IEEE Bologna*, 2003, p. 5 pp. Vol.1.

[10]"Load Profiles and Their Use in Electricity Settlement," Elexon2013.

[11]X. Yizheng and J. V. Milanovic, "Developmnet of probabilistic daily demand curves for different categories of customers," in *Electricity Distribution (CIRED 2013), 22nd International Conference and Exhibition on*, 2013, pp. 1-4.

[12]X. Yizheng and J. V. Milanovic, "Framework for estimation of daily variation of dynamic response of aggregate load," in *Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES*, 2013, pp. 1-5.

[13]X. Chen, C. Kang, X. Tong, Q. Xia, and J. Yang, "Improving the Accuracy of Bus Load Forecasting by a Two-Stage Bad Data Identification Method," *IEEE Transactions on Power Systems,* vol. 29, pp. 1634-1641, 2014.

[14]M. M. Kantardzic and A. Kumar, "Toward Autonomic Distributed Data Mining with Intelligent Web Services," in *IKE*, 2003, pp. 544-552.

[15]K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied Intelligence,* vol. 11, pp. 259-275, 1999.

[16]G. E. Batista and M. C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method," *HIS,* vol. 87, p. 48, 2002.

[17]G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence,* vol. 17, pp. 519-533, 2003.

[18]R. A. Cohen, "An introduction to PROC LOESS for local regression," in *Proceedings of the 24th SAS users group international conference, Paper*, 1999.

[19]E.-H. S. Han, G. Karypis, and V. Kumar, *Text categorization using weight adjusted k-nearest neighbor classification*: Springer, 2001.