Dataport and NILMTK: A Building Data Set Designed for Non-intrusive Load Monitoring

Oliver Parson[1], Grant Fisher[2], April Hersey[2], Nipun Batra[3], Jack Kelly[4],
Amarjeet Singh[3], William Knottenbelt[4], Alex Rogers[1]
[1]University of Southampton, UK {osp, acr}@ecs.soton.ac.uk
[2]Pecan Street Inc, USA {gfisher, ahersey}@pecanstreet.org
[3]Indraprastha Institute of Information Technology Delhi, India {nipunb, amarjeet}@iiitd.ac.in
[4]Imperial College London, UK {jack.kelly, w.knottenbelt}@imperial.ac.uk

***Abstract*—Non-intrusive load monitoring (NILM), or energy disaggregation, is the process of using signal processing and machine learning to separate the energy consumption of a building into individual appliances. In recent years, a number of data sets have been released in order to evaluate such approaches, which contain both building-level and appliance-level energy data. However, these data sets typically cover less than 10 households due to the financial cost of such deployments, and are not released in a format which allows the data sets to be easily used by energy disaggregation researchers. To this end, the Dataport database was created by Pecan Street Inc, which contains 1 minute circuit-level and building-level electricity data from 722 households. Furthermore, the non-intrusive load monitoring toolkit (NILMTK) was released in 2014, which provides software infrastructure to support energy disaggregation research, such as data set parsers, benchmark disaggregation algorithms and accuracy metrics. This paper describes the release of a subset of the Dataport database in NILMTK format, containing one month of electricity data from 669 households. Through the release of this Dataport data in NILMTK format, we pose a challenge to the signal processing community to produce energy disaggregation algorithms which are both accurate and scalable.**

***Index Terms*—Smart grid, Power system measurements, Open source software**

I. Introduction

Non-intrusive load monitoring (NILM), or energy disaggregation, aims to break down a household's aggregate electricity consumption into individual appliances [1]. The motivations for such a process are threefold. First, informing a household's occupants of how much energy each appliance consumes empowers them to take steps towards reducing their energy consumption [2]. Second, personalised feedback can be provided which quantifies the savings of certain appliance-specific advice, such as the financial savings when an old inefficient appliance is replaced by a new efficient appliance. Third, if the NILM system is able to determine the time of use of each appliance, a recommender system would be able to inform the household's occupants of the savings of deferring appliance use to a time of day when electricity is cheaper.

In order to evaluate such energy disaggregation algorithms, it is necessary to collect data sets containing both the household aggregate power demand (to provide the input to the algorithm) and the power demand of each individual appliance (to provide the ground truth against which the output of the algorithm is compared). Collecting such data sets is expensive and intrusive, and as such most data sets cover fewer than 10 houses. This is a serious problem for evaluation purposes, as the variance between households means algorithms easily overfit on these small data sets.

Against this background, in this paper we present a new data set constructed from the Dataport database. Unlike the database itself, the data set is formatted following the NILMTK data set standards, and so is compatible with the statistical, preprocessing and disaggregation functions of the toolkit. The data set contains one month of data from 669 of the Dataport houses, and as such constitutes the largest energy disaggregation data set which contains both household aggregate and individual appliance power data.

The remainder of the paper is structured as follows. In Section II we describe existing data sets, whose small size motivates the use of the Dataport database, which we describe in Section III. We then describe NILMTK in detail in Section IV, before presenting the release of Dataport data in NILMTK format in Section V. In Section VI we propose the scalable disaggregation of such data as a challenge to the signal processing community, before concluding in Section VII.

II. Existing data sets

In 2011, the Reference Energy Disaggregation data set (REDD) [3] was introduced as the first publicly available data set collected specifically to aid NILM research. The data set contains both aggregate and sub-metered power data from six households, and has since become the most popular data set for evaluating energy disaggregation algorithms. In 2012, the Building-Level fUlly-labeled data set for Electricity Disaggregation (BLUED) [4] was released containing data from a single household. The data set does not include sub-metered power data, and instead records events triggered by appliance state changes. More recently, the Smart* data set [5] was released, which contains household aggregate power data from three households, while sub-metered appliance power data was only collected from a single household. The Household Electricity Survey data set [6] was also released in 2012, which contains data from 251 households. Although this was the

Data set	Institution	Location	Duration per house	Number of houses	Appliance sample frequency	Aggregate sample frequency
REDD (2011)	MIT	MA, USA	3-19 days	6	3 sec	1 sec & 15 kHz
BLUED (2012)	CMU	PA, USA	8 days	1	N/A*	12 kHz
Smart* (2012)	UMass	MA, USA	3 months	3	1 sec	1 sec
HES (2012)	DECC, DEFRA, EST	UK	1 or 12 months	251	2 or 10 min	2 or 10 min
AMPds 2 (2013)	Simon Fraser University	BC, Canada	2 years	1	1 min	1 min
iAWE (2013)	IIIT Delhi	Delhi, India	73 days	1	1 or 6 sec	1 sec
UK-DALE (2014)	Imperial College	London, UK	3-26 months	5	6 sec	1-6 sec & 16 kHz
ECO (2014)	ETH Zurch	Switzerland	8 months	6	1 sec	1 sec
GREEND (2014)	Alpen-Adria-U. Klagenfurt	Italy & Austria	12 months	9	1 sec	N/A
SustData (2014)	University of Madeira	Madeira, Portugal	5-21 months	50	N/A	50 Hz
Dataport (2014)	Pecan Street Inc	TX, USA	up to 3.25 years	722	1 min	1 min
DRED (2015)	TU Delft	Netherlands	2 months	1	1 sec	1 sec

TABLE I: Comparison of household energy data sets. *BLUED labels state transitions for each appliance.

largest data set of appliance-level data, household aggregate data was only collected from 14 households, and as such the energy consumption which was not monitored by the appliance-level meters is unknown for the majority of houses.

In 2013, the Almanac of Minutely Power data set (AMPds) [7] was released, which contained one year of aggregate and sub-metered power data from a single household, and a subsequent release (AMPds 2) extended this to two years of data. The Indian data for Ambient Water and Electricity Sensing (iAWE) [8] was also released in 2013, which contains both aggregate and sub-metered power data from a single house. In 2014, the UK Domestic Appliance-Level Electricity data set [9] (UK-DALE) was released which contains data from five households using both aggregate meters and individual appliance sub-meters. Later that year, the ECO data set [7] was released, which contains 1 second data for six houses in Switzerland. More recently, the GREEND data set [10] was released, which contains 1 second appliance data from nine houses in Italy and Austria. Most recently, the SustData data set [11] and DRED data set [12] have been released containing data from 50 and 1 home respectively. Table I summarises these data sets.

However, with the exception of HES, all data sets cover less than 10 households due to the financial cost and installation disruption of sub-metering individual appliances. As a result, there is a severe problem of overfitting energy disaggregation methods to such small sets of houses. Consequently, the scalability of energy disaggregation algorithms cannot be evaluated with such data sets. In contrast, the HES data set contains appliance-level data for 251 households. However, since household aggregate data was only collected in 14 houses, of which only five are reliable [13], it is not known how much energy was not sub-metered in most households. As a result, many households might present a simplified version of the energy disaggregation problem, in which only a small subset of appliances were monitored in each house. These disadvantages motivated the collection of the Dataport database, which contains both building aggregate and circuit-level electricity measurements for 722 households, which we describe in Section III. In addition, subtle differences in the aims of each data set have led to completely different data formats being used. This motivated the release of the non-intrusive load monitoring toolkit, described in Section IV.

III. Dataport database

The Dataport database[1] is the world's largest source of disaggregated customer energy data. The database contains electricity data collected from 722 houses in the US; 631 in Texas, 49 in Colorado and 42 in California. The houses monitored include 501 single-family homes, 183 apartments, 35 town homes and 3 mobile homes. A range of houses were monitored, as shown by Figure 1 (a) which shows a histogram of the building construction date, while Figure 1 (b) shows a histogram of the house size in square feet. The installation of electricity monitors began in 2011 and is currently still in progress. Figure 1 (c) shows the installation date for each home, for which data is still being collected for 624 houses. Access to the portal is free for members of universities, while commercial access is limited to members of Pecan Street's Industry Advisory Council.

The houses were installed with at least one eGauge 2, EG3000, EG2010 or EG2011 meter,[2] each of which allow up to 12 electrical circuits to be monitored via current transformer clamps. Both mains circuits and individual appliance circuits were monitored in most houses. Since households typically contain about 10 circuits, with most large appliances occurring in isolation on a single circuit, the circuit-level data can effectively be considered as appliance-level data and used as the ground truth for testing energy disaggregation algorithms. Furthermore, the mains circuits can be used to calculate the proportion of energy which was sub-metered by the circuit-level meters. The average power demand of each circuit was measured at one minute intervals.

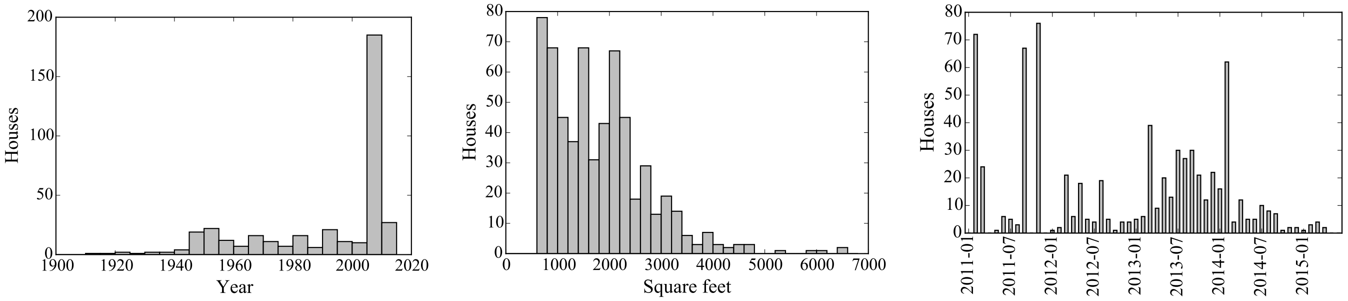
IV. Non-intrusive load monitoring toolkit

The non-intrusive load monitoring toolkit (NILMTK) was first released (v0.1) as open source software[3] in April 2014 [14]. The toolkit was designed specifically to enable easy access to and comparative analysis of energy disaggregation algorithms across diverse data sets. NILMTK provides a complete pipeline from data sets to accuracy metrics, thereby

[1] https://dataport.pecanstreet.org

[2] http://egauge.net/products

[3] http://nilmtk.github.io



(a) Building construction

(b) House size

(c) Installation date

Fig. 1: Metadata of monitored houses.

lowering the entry barrier for researchers to implement a new algorithm and compare its performance against the current state of the art. The toolkit contains data set parsers, data set analysis statistics, preprocessors for reformatting data sets, benchmark disaggregation algorithms and accuracy metrics.

A second version of the toolkit (v0.2) was later released in July 2014 [15]. Crucially, this version of the toolkit provides support for out-of-core processing, allowing data sets to be processed which are too large to fit into volatile memory. In addition, v0.2 adds rich metadata support via the NILM Metadata project [16], allowing arbitrarily complex metering hierarchies to be described.

V. Dataport as a NILMTK data set

In this paper, we describe the new release of a subset of the Dataport database in NILMTK format, which is available as an HDF5 file via the Dataport portal.[4] The data set was constructed using the NILMTK Dataport data set converter, therefore allowing users to download a different subset of the Dataport database as well as download the pre-constructed data set. The pre-constructed HDF5 file is 1.09 GB in size, and contains one month of data from 669 of the Dataport houses, which were selected as they contain at least 8 meters. In each house, the circuit name has been converted from the Dataport names to the NILM Metadata controlled vocabulary. The dataset can be easily analysed using tools described in the NILMTK documentation.[5]

Figure 2 shows an example of the electricity data recorded over a 24 hour period for a single house. The power demand of the house aggregate site meter is shown, as well as the power demand of each individual circuit. Electrical signatures of individual appliances can be seen, such as the high short spikes drawn by the air conditioning, and the low but longer blocks drawn by the refrigerator. It is exactly these patterns which a disaggregation algorithm must use in order to separate the power demand of each appliance.

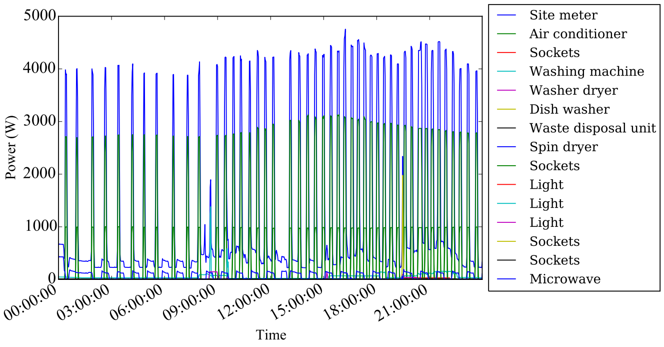


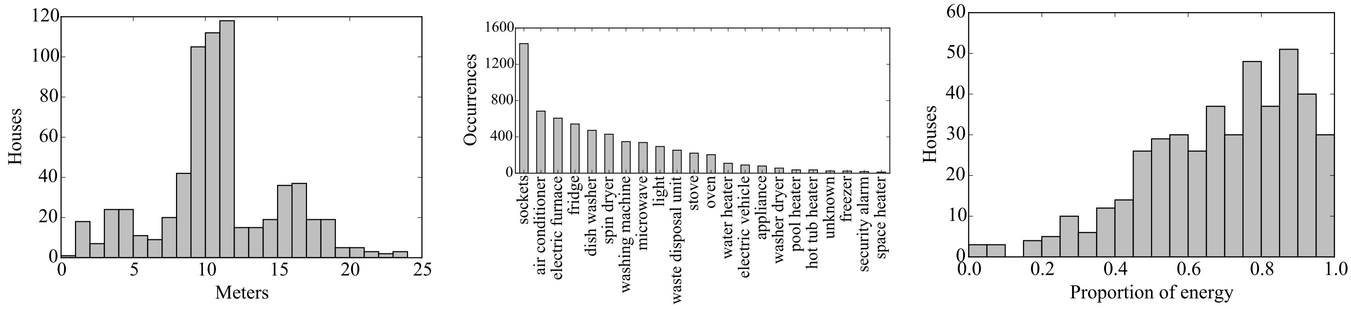
Fig. 2: Example day of data from one house.

Different numbers of current transformer clamps were installed in each house. Figure 3 (a) shows a histogram of the number of circuits which were monitored in each house, including the aggregate circuits. It can be seen that on average approximately 10 circuits were monitored in each house, with some houses containing up to 23 circuit meters. However, it should be noted that 74 houses contain less than five meters, 19 of the houses contain only aggregate-level meters, and 13 houses contain only circuit-level meters, as not all meters were present for the monitored month. Figure 3 (b) shows a bar chart representing the number of occurrences of each appliance category across all houses. It can be seen that the sockets category is the most common, since it occurs multiple times in each house. This presents a slight problem for the training and evaluation of energy disaggregation algorithms, as it is not known which appliances are connected to such circuits. However, other appliances such as the air conditioner, electric furnace, fridge and dishwasher are monitored separately in most houses. Figure 3 (c) shows a histogram of the proportion of household aggregate energy also recorded by individual circuits. It can be seen that the majority of energy was sub-metered in most houses, although less energy was sub-metered in houses with fewer sub-meters.

Figure 4 shows a boxplot of the proportion of energy consumed for each circuit category across all households. Only the houses containing at least one circuit matching the appliance were used for each category. It can be seen that the washing machine and washer dryer consistently consume

[4] https://dataport.pecanstreet.org/data/database?hdf5

[5] https://github.com/nilmtk/nilmtk/tree/master/docs/manual



(a) Number of circuit meters per house.

(b) Number of appliance occurrences.

(c) Proportion of energy sub-metered.

Fig. 3: Analysis of HDF5 data set.

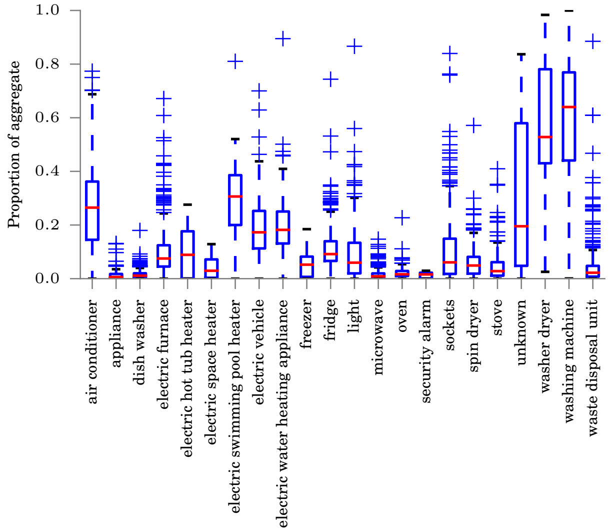


Fig. 4: Proportion of energy per appliance. The red line represents the median, the boxed area represents the 25th and 75th percentiles, and points beyond the whiskers represent the outlying circuits greater than 1.5 times the inter-quartile range.

a high proportion of the total consumption in the houses in which they were present, while the microwave and oven loads consume much less energy. Such proportions should be taken into account to weight the relative importance of disaggregating various appliances. However, it should be noted that most data is from May 2015, and seasonal loads such as air conditioning and heating will vary widely over the year.

VI. Challenge to signal processing community

By releasing this data set, we intend to make the domain of energy disaggregation more accessible to researchers from new communities, such as the signal processing community. A key challenge in the energy disaggregation field is producing an approach which is both accurate and practical. This difficulty arises due to the large variance between houses, in that different appliances are present in each house, and also the appliances are used in different ways. To overcome this, some approaches require appliance-level data from the house in which disaggregation is to be performed for training the disaggregation model. However this is not a practical assumption and severely limits the scalability of such approaches. Instead, a more compelling scenario consists of the blind disaggregation of a specific house, in which no information is known about which appliances are present in a house, and no appliance-level data is available for training. To overcome the lack of appliance-level training data from the disaggregation house, approaches must make use of appliance-level data from houses other than the disaggregation house. We believe this to be the scenario of highest real-world impact.

An additional challenge arises due to the nature of the electricity data in the Dataport database. Similar to smart meters, the Dataport database contains measurements of the average power over a period of time (analogous to the energy consumed over an interval of time). This is in contrast to many other data sets, which often capture the instantaneous power demand (rate of energy consumption) at a single point in time. As such, the effects of sharp step changes in the power demand caused by the switching on or off of appliances are often averaged over consecutive readings in smart meter data. This poses a particular challenge to disaggregation algorithms, as patterns in the average power (energy) data are not as consistent as for instantaneous power data. This is another key challenge which we encourage new research to address.

VII. Conclusions and future work

In this paper we have presented a new data set constructed from the Dataport database. Unlike the database itself, the data set is formatted following the NILMTK data set standard, and as such is compatible with the statistical, preprocessing and disaggregation functions of the toolkit. The data set contains one month of data from 669 houses, and as such constitutes the largest energy disaggregation data set which contains both aggregate and appliance electricity data. Our immediate future work will focus on the release of a larger subset of the Dataport database in NILMTK HDF5 format. Specifically, we expect to release longer durations of data from the 669 houses. In the longer-term, we would like to include additional metadata for each circuit, such as the manufacturer and model of each appliance attached to an individual circuit.

References

[1] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[2] S. Darby, "The effectiveness of feedback on energy consumption," *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 2006.

[3] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Proceedings of 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, San Diego, CA, USA, 2011.

[4] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Bergés, "BLUED: A fully labeled public dataset for Event-Based Non-Intrusive load monitoring research," in *Proceedings of 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012, pp. 12–16.

[5] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," in *Proceedings of 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012.

[6] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evans, "Household Electricity Survey. A study of domestic electrical product usage," DEFRA, Tech. Rep. R66141, May 2012.

[7] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic, "AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research," in *IEEE Electrical Power and Energy Conference*, Halifax, NS, Canada, 2013.

[8] N. Batra, M. Gulati, A. Singh, and M. B. Srivastava, "It's Different: Insights into home energy consumption in India," in *Proceedings of the Fifth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (ACM BuildSys)*, 2013.

[9] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, 2015.

[10] A. Monacchi, D. Egarter, W. Elmenreich, S. DAlessandro, and A. M. Tonello, "GREEND: An Energy Consumption Dataset of Households in Italy and Austria," in *the 5th IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Venice, Italy, Nov. 2014.

[11] L. Pereira, F. Quintal, R. Gonçalves, and N. J. Nunes, "SustData: A public dataset for ICT4S electric energy research," in *ICT for Sustainability 2014 (ICT4S-14)*. Atlantis Press, 2014.

[12] Akshay Uttama Nambi S.N., A. R. Lua, and R. V. Prasad, "Loced: Location-aware energy disaggregation framework," in *Proceedings of the Second ACM International Conference on Embedded Systems For Energy-Efficient Built Environments (ACM BuildSys)*, 2015.

[13] "Household Electricity Survey: Cleaning the Data," Cambridge Architectural Research Limited, Tech. Rep. Reference 475/09/2012, distributed with HES data set, March 2013.

[14] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring," in *Fifth International Conference on Future Energy Systems (ACM e-Energy)*, Cambridge, UK, 2014.

[15] J. Kelly, N. Batra, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK v0.2: A Non-intrusive Load Monitoring Toolkit for Large Scale Data Sets," in *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (ACM BuildSys)*, 2014, pp. 182–183.

[16] J. Kelly and W. Knottenbelt, "Metadata for Energy Disaggregation," in *The 2nd IEEE International Workshop on Consumer Devices and Systems (CDS 2014)*, Västerås, Sweden, Jul. 2014.