Electricity Consumption Data Sets: Pitfalls and Opportunities

Christoph Klemenjak
University of Klagenfurt, Austria
klemenjak@ieee.org

Andreas Reinhardt
TU Clausthal, Germany
reinhardt@ieee.org

Lucas Pereira
ITI, LARSyS, Técnico Lisboa, Portugal
lucas.pereira@tecnico.ulisboa.pt

Stephen Makonin
Simon Fraser University, Canada
smakonin@sfu.ca

Mario Bergés
Carnegie Mellon University, USA
marioberges@cmu.edu

Wilfried Elmenreich
University of Klagenfurt, Austria
wilfried.elmenreich@aau.at

ABSTRACT

Real-world data sets are crucial to develop and test signal processing and machine learning algorithms to solve energy-related problems. Their scope and data resolution is, however, often limited to the means required to fulfill the experimenters' objectives and moreover governed by personal experience, budgetary and time constraints, and the availability of equipment. As a result, numerous differences between data sets can be observed, e.g., regarding their sampling rates, the number of sensors deployed, their amplitude resolutions, storage formats, or the availability and extent of ground-truth annotations. This heterogeneity poses a significant problem for researchers intending to comparatively use data sets because of the required data conversion, re-sampling, and adaptation steps. In short, there is a lack of widely agreed best practices for designing, deploying, and operating electrical data collection systems. We address this limitation by dissecting the collection methodologies used in existing data sets. By offering recommendations for data collection, data storage, and data provision, we intend to foster the creation of data sets with increased usability and comparability, and thus a greater benefit to the community.

CCS CONCEPTS

• **Hardware** → *Energy metering*; *Smart grid*.

KEYWORDS

energy consumption data sets, data heterogeneity, best practices

ACM Reference Format:
Christoph Klemenjak, Andreas Reinhardt, Lucas Pereira, Stephen Makonin, Mario Bergés, and Wilfried Elmenreich. 2019. Electricity Consumption Data Sets: Pitfalls and Opportunities. In *BuildSys '19: The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, November 13–14, 2019, New York, NY, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3360322.3360867

BuildSys '19, November 13–14, 2019, New York, NY, USA

ACM ISBN 978-1-4503-7005-9/19/11…$15.00
https://doi.org/10.1145/3360322.3360867

1 INTRODUCTION

Since the concept of Non-Intrusive Load Monitoring (NILM) has been presented in [11], the analysis of electrical consumption data has received a strong research interest. In order to accelerate these research activities even more, several groups have collected and publicly released data sets. Newly designed algorithms to extract knowledge from energy data are often evaluated using these data. This strengthens their practical applicability and relevance. These data sets are invariably comprised of electrical consumption data. However, many differences can be observed in virtually all other aspects. Among many other facets, differences include the geographic areas in which data have been collected, the duration of the recorded data, and the file format used for storing them. This data set heterogeneity is a severe impediment to the development of energy analytics algorithms. Currently, extensive adaptation efforts (like the NILMTK data set converters [6]) are required to enable algorithms to work with particular data sets, if at all possible. At the same time, relying on a single data set causes issues like overfitting and the lack of generalization. We consequently argue that a methodological and widely agreed procedure for the collection and provision of data sets is crucial. The primary objectives of such a methodology are twofold: (1) Cater for *data set interoperability*, i.e., the possibility to seamlessly change the underlying data set when evaluating algorithms. (2) Enable *data set comparability*, i.e., the possibility to interpret algorithm performance measures without complex and error-prone conversion and adaptation steps. A unified data set collection methodology facilitates the benchmarking of energy analytics algorithms and accelerates their development. In an attempt to define the cornerstones of such a methodology, we summarize observations from existing data sets and analysis algorithms and derive recommendations for data collection. Moreover, we highlight several imminent open research challenges that would lead to wider applicability of data sets. This way, we guide future data set collection campaigns in making their data widely usable and comparable, and thus increase their benefit to the community.

2 PROPERTIES OF EXISTING DATA SETS

Algorithms to analyze electrical consumption are inherently data-centric. Therefore, a vast amount of consumption data is necessary for their comprehensive performance evaluation. A range of data sets have been collected over the past years. The large majority of publications on the analysis of electrical consumption data, however, relies on only one or at most a few data sets. One reason for this is the heterogeneity, in some aspects even the incompatibility, of existing data sets. This obstacle expresses itself in several ways.

Table 1: Impact of the data sampling rate on the corresponding file size for 10 minutes of data (based on [16]).

Sampling Rate	Raw Size	HDF5 Size	Size Reduction
250 Hz	2.0 MB	0.8 MB	60.09 %
4 kHz	32.1 MB	12.3 MB	61.66 %
16 kHz	128.2 MB	48.8 MB	61.93 %
50 kHz	401.0 MB	156.3 MB	61.01 %

2.1 Measured Electrical Quantities

As demonstrated in a recent study [22], existing data sets differ significantly in the measurements they provide. Some data sets (e.g., [4, 5, 7, 17, 19]) contain separate traces for electrical voltages and currents. This not only allows for the computation of power (being the product of the two) but also to determine further AC power components [28]. In contrast, other data sets (e.g., [20, 25]) only report apparent electrical power. Implicitly, algorithms that require phase shift information cannot be evaluated using data of the latter type. Some data collection campaigns do not explicitly specify whether they have collected active or apparent power, but merely report whatever data the employed (mostly plug-level) sensors provide, which complicates their correct interpretation.

Suggestion 1: The provision of voltage and current measurements allows for a greater extent of analyses as well as facilitating the computation of real and reactive power consumption. We thus propose to capture and report raw voltage and current data instead of computed quantities (e.g., power) whenever possible.

Existing energy data sets can be divided into two groups [29]: macroscopic data sets with data reporting rates around 1 Hz and microscopic data sets with rates of several kHz and beyond [3]. The rates at which microscopic data sets are captured vary between virtually all collection campaigns, from 250 kHz [17] to 100 kHz [24], 44.1 kHz [12], 30 kHz [10], 16.5 kHz [15], 16 kHz [14], and 12 kHz [2]. A commonality across all data sets, however, is the fact that waveform detail is retained and allows for analysis in both the time and the frequency domains. In contrast, macroscopic data sets usually report root-mean-square (RMS) values of their monitored modalities once per second (e.g., [7, 25]) or even less often (once per minute in [19]). While microscopic data sets can be converted to their macroscopic representations, currently available data sets are rarely offered in both representations.

Suggestion 2: The collection of microscopic data should always be favoured over macroscopic data collection and, when recorded, always be reported.

Suggestion 3: In order to enable the usage of both microscopic and macroscopic data, either a down-sampled version of the data set (e.g., one sample per second, being the rate used by a large number of algorithms already) or at least an executable tool to facilitate this conversion should be provided along with the data.

High sampling rates implicitly result in a high volume of data being retrieved and thus necessitate increased efforts for data storage and management. Experiments presented in [16] show the impact of sampling frequency on the file size of a data set. We summarize a small selection of the findings in Table 1. This demonstrates the need for compression techniques and advanced file formats, which we discuss in Section 2.4.

Suggestion 4: It is preferable to capture data using sampling rates on at least the order of tens of kHz such that microscopic waveform analysis becomes possible.

Another observation addresses irregularities in data reporting rates. This is, e.g., the case for tracebase [25], where rates vary between approximately 1/4 Hz and 4 Hz. Similarly, inaccurate system or sampling clocks may lead to gradual drifts between long-term recordings. Particularly in distributed and decoupled systems, accurate time synchronization between all sensing devices is required, yet not always easy to accomplish. This diversity of data collection rates complicates the development of energy analytics algorithms, which are often tailored to consume data captured at a given (constant) sampling rate.

Suggestion 5: Conduct a manual cross-check before releasing a data set and propose (leveraging the expert knowledge available to the data set collector) how to deal with gaps and irregularities in the data, e.g., by suggesting interpolation methods.

In some data sets, considerable changes to the voltage signals can be observed (i.e., standard deviations of up to 30 V in [10, 14]), despite the general expectation of the voltage to be comparably constant. In part, this can be attributed to their operation in varying locations within the electrical distribution grid. To another extent, however, sensor accuracy may also depend on the environmental conditions and the accuracy class of the device. Some data sets specify (e.g., COOLL [24]) calibration factors for each contained trace to this end, in order to compensate for such deviations.

Suggestion 6: Whenever possible, the employed transducers should be characterized empirically by using a reference power supply and load to determine their linearity. Measurements without any load should also be made, to quantify transducer noise.

2.2 Sensor Placement and Campaign Duration

Despite the fact that energy data set collection campaigns were conducted with rather similar aims, we can observe significant differences in the way they were conducted. Virtually all published data sets are heterogeneous with regard to the duration of their collection campaign. For instance, while AMPds2 [19] features several years of data, ECO [7] covers eight months, and REDD [15] only spans about one month. The PLAID data set [10] only captures a few seconds of each appliance's initial current demand. Thus, the large majority of existing data sets cannot fully capture particularities of all seasons because they span less than one year.

Suggestion 7: In order to capture gradual changes in appliance usage patterns and human behaviour (e.g., due to seasonal changes or equipment degradation), we suggest to collect data over time intervals of at least one year. Exceptions apply to campaigns investigating specific phenomena, if a shorter duration suffices to capture all relevant features in the required detail.

The placement of sensing devices is equally diverse between data sets. Some data sets are based on capturing the aggregate load of a complete building (also known as *single-point sensing*), e.g., [2]. More fine-grained instrumentation deployments exist, e.g., to monitor individual circuits (e.g., [19]) or even each load separately (like

in [4]). Besides the diversity of how many sensing devices are deployed, and where in the electrical circuit they are being used, a second issue exists. The inconsistent deployment of submetering devices, i.e., monitoring only a few appliances or circuits leads to situations where only partial data is available. Even meticulously planned deployments of sensing infrastructure to collect data for each appliance can be rendered inaccurate when power strips with multiple attached appliances are attached to wall outlets. Similarly, in order to correctly track mobile appliances (such as vacuum cleaners), their connection points must be constantly kept up-to-date.

Suggestion 8: Data sets should contain both the aggregate electrical quantities of the monitored environment and data from the appliances attached to the same circuit(s). We acknowledge that the partial instrumentation of buildings with sensing infrastructure can be inevitable due to physical limitations or privacy policies. We strongly recommend the application of auxiliary sensing equipment in such corner cases to infer information about the operation of appliances whose electrical consumption is not monitored directly.

2.3 Metadata Annotations

Besides the monitored electrical quantities, supplementary features are often logged as part of measurement campaigns and used to validate algorithms, e.g., user activities [1, 2, 7].

Suggestion 9: During data collection, the type tags or stickers of all monitored appliances should be photographed (like in [24]) as well as relevant documentation stored in a digital manner. In cases where this is not possible, at least an indication of the appliance's nominal power consumption should be provided.

Suggestion 10: For some application areas, the existence of labelled events (such as present in [2]) is essential to train algorithms. Therefore, we suggest to log activities and events during the campaign in some sort of diary, video footage, or other adequate forms that allow for the annotation of traces with event/activity information (like in [1, 23]). In case an appliance has been removed or replaced, this should also be logged.

Suggestion 11: The geographic and ambient features during the collection (e.g., in [12]), as well as user demographics and properties of the building(s) under consideration should be logged.

Different proposals for the provision of metadata have been made. One example, NILM metadata [13], is a metadata schema for representing appliances, meters, buildings, data sets, prior knowledge about appliances and appliance models. An alternative format, following a similar objective, is EMD-DF [21].

Suggestion 12: Metadata should be annotated and formatted in a machine-readable way in order to facilitate simple processing. This is particularly true for event annotations, which represent an important input data for many NILM algorithms.

Ambient information recorded during the collection of electrical data may also serve as a foundation for the explanation of observed phenomena (e.g., to determine the relationship between air conditioning use and outside temperature). Other sensing modalities include brightness, ambient noise, motion, statistics related to weather, and building occupancy [7, 8]. When annotated properly, these metadata annotations simultaneously foster the collection of region-specific features. Such features are particularly interesting for comparative case studies, e.g., to identify electricity usage and wastage across different geographic areas.

Suggestion 13: The collection of ambient features, as well as aspects that are characteristic for the collection site(s), should be logged and provided as metadata.

2.4 File Formats

Current data sets come in a variety of file formats. Comma-separated values (CSV) are widely used to store macroscopic data (cf. Section 2.1). However, more sophisticated formats (HDF5 in [17], FLAC in [12], WAVE in [21], or Matroska media containers in [27]), and in parts also non-relational databases, have established themselves for microscopic data. Converging on a file format for energy data sets still involves a series of trade-offs: Supported computing frameworks, inclusion of metadata, error correction, and chunking [18]. The Hierarchical Data Format 5 (HDF5) [9] has emerged as a viable candidate. It supports metadata annotations, efficient data storage, data transformations, and libraries for most scientific computing frameworks. NILMTK-DF is a data format tailored to the needs of NILMTK and internally relies on HDF5 with a custom metadata structure [6]. However, the fact that NILMTK-DF has not found wide acceptance in the scientific community is underpinned by the observation that no data set is provided in this format currently.

Suggestion 14: Data set creators should make an informed decision on the format using which they release a data set, in order to maximize its compatibility with NILM tools and energy analytics algorithms.

With regard to file sizes, data compression plays an important role. As researchers find in [26], macroscopic smart meter data can be compressed with average compression rates between 75 % and 95 %. This claim is supported for microscopic data by the two rightmost columns in Table 1, which show the savings achievable through the use of HDF5 compression [18]. Data compression should thus be considered when deciding on a data set file format, in order to also reduce its download time.

2.5 Access and Use of Data Sets

In principle, two strategies exist in order to make a data set publicly available: self-hosting or third-party hosting. Self-hosting solutions tend to represent a threat to the long-term availability of data sets due to local infrastructure and personnel fluctuations. Prominent examples of energy data sets (e.g., SMARTENERGY.KOM [1]) suffer from this issue. This is particularly problematic because contributions evaluated on these data sets can no longer be validated once the data has become unavailable.

Suggestion 15: Consider the use of public hosting services[1] to ensure long-time availability of a data set.

Two final crucial aspects to consider when publishing a data set are licensing and user privacy protection, both of which have an impact on usage limitations. Only very few data sets clearly state the conditions under which data may be used and distributed, such as the Pecan Street Dataport data set.

[1] E.g., Harvard Dataverse (dataverse.harvard.edu) or IEEE DataPort (ieee-dataport.org)

Suggestion 16: Data should be provided with a license that is as permissive as possible. Data set collectors must take all necessary precautions to protect the privacy of the users concerned.

2.6 Re-Releasing Existing Data Sets

While our aforementioned propositions mainly apply to new data collection campaigns, we have derived these recommendations from the currently available data sets. They carry an important value, as they have been used for the evaluation of many energy analytics papers. Many of our suggestions, however, can be applied to existing data sets as well.

Suggestion 17: The creators of existing data sets should consider a re-release of their data that take our previously expressed suggestions into account, where applicable.

3 CONCLUSIONS

We have elaborated on a number of issues that directly impact the usefulness of electricity consumption data sets with respect to the development and testing of signal processing and machine learning algorithms. Due to the constant rise of smart metering, the number of available data sets have significantly increased in the last years, which is important to evaluate algorithms on a broad basis and to reduce the risk of overfitting. Many data sets turn out to be significantly heterogeneous in aspects like the measured quantities, their sampling rates, the coverage of metered data, and their campaign durations. Existing tools (such as NILMTK [6]) provide data set converters to allow algorithm validation with different data sets. However, if an algorithm requires a certain measurand that is not part of the data set or cannot be deduced from it, the data set cannot be used for the algorithm's evaluation. Based on our analysis of more than a dozen data sets, we brought forward 17 suggestions. We expect these suggestions to be a basis for the planning of future measurement campaigns and, in consequence, the release of new data sets. In addition, existing data sets can benefit from a conversion according to the provided guidelines in order to achieve a wider use in the development of better energy analytics algorithms.

ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft grant no. RE 3857/2-1, Portugal FCT grant UID/EEA/50009/2019, and by Lakeside Labs via the Smart Microgrid Lab.

REFERENCES

[1] Alaa Alhamoud, Felix Ruettiger, Andreas Reinhardt, Frank Englert, Daniel Burgstahler, Doreen Boehnstedt, Christian Gottron, and Ralf Steinmetz. 2014. SMARTENERGY.KOM: An Intelligent System for Energy Saving in Smart Home. In *3rd Workshop on GlObal Trends in Smart Cities (goSMART)*.

[2] Kyle Anderson, Adrian Ocneanu, Diego Benitez, Derrick Carlson, Anthony Rowe, and Mario Berges. 2012. BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research. In *2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*.

[3] Toktam Babaei, Hamid Abdi, Chee Peng Lim, and Saeid Nahavandi. 2015. A Study and a Directory of Energy Consumption Data Sets of Buildings. *Energy and Buildings* 94 (2015).

[4] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. 2012. Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes. In *Workshop on Data Mining Applications in Sustainability (SustKDD)*.

[5] Nipun Batra, Manoj Gulati, Amarjeet Singh, and Mani B Srivastava. 2013. It's Different: Insights into Home Energy Consumption in India. In *5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys)*.

[6] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. 2014. NILMTK: An Open Source Toolkit for Non-Intrusive Load Monitoring. In *5th ACM International Conference on Future Energy Systems (e-Energy)*.

[7] Christian Beckel, Wilhelm Kleiminger, Romano Cicchetti, Thorsten Staake, and Silvia Santini. 2014. The ECO Data Set and the Performance of Non-Intrusive Load Monitoring Algorithms. In *1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys)*.

[8] Mario Berges, Lucio Soibelman, and H. Scott Matthews. 2010. Leveraging Data From Environmental Sensors to Enhance Electrical Load Disaggregation Algorithms. In *13th International Conference on Computing in Civil and Building Engineering (ICCCBE)*.

[9] Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal, and Dana Robinson. 2011. An Overview of the HDF5 Technology Suite and its Applications. In *EDBT/ICDT Workshop on Array Databases*.

[10] Jingkun Gao, Suman Giri, Emre Can Kara, and Mario Bergés. 2014. PLAID: A Public Dataset of High-resolution Electrical Appliance Measurements for Load Identification Research: Demo Abstract. In *1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys)*.

[11] George W. Hart. 1985. *Prototype Nonintrusive Appliance Load Monitor*. Technical Report. MIT Energy Laboratory and Electric Power Research Institute.

[12] Matthias Kahl, Anwar Ul Haq, Thomas Kriechbaumer, and Hans-Arno Jacobsen. 2016. WHITED – A Worldwide Household and Industry Transient Energy Data Set. In *3rd International Workshop on Non-Intrusive Load Monitoring (NILM)*.

[13] Jack Kelly and William Knottenbelt. 2014. Metadata for Energy Disaggregation. In *38th International Computers, Software, and Applications Conference (COMPSAC) Workshops*.

[14] Jack Kelly and William Knottenbelt. 2015. The UK-DALE Dataset, Domestic Appliance-level Electricity Demand and Whole-House Demand from Five UK Homes. *Scientific Data* 2, 150007 (2015). https://doi.org/10.1038/sdata.2015.7

[15] J Zico Kolter and Matthew J Johnson. 2011. REDD: A Public Data Set for Energy Disaggregation Research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD)*, Vol. 25.

[16] Thomas Kriechbaumer. 2019. *Methodologies for Distributed Acquisition and Collection of Electrical Energy Data*. Dissertation. Technische Universität München.

[17] Thomas Kriechbaumer and Hans-Arno Jacobsen. 2018. BLOND, a Building-level Office Environment Dataset of Typical Electrical Appliances. *Scientific Data* 5, 180048 (2018).

[18] Thomas Kriechbaumer, Daniel Jorde, and Hans-Arno Jacobsen. 2019. Waveform Signal Entropy and Compression Study of Whole-Building Energy Datasets. In *10th ACM International Conference on Future Energy Systems (e-Energy)*.

[19] Stephen Makonin, Bradley Ellert, Ivan V. Bajic, and Fred Popowich. 2016. Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014. *Scientific Data* 3, 160037 (2016).

[20] Andrea Monacchi, Dominik Egarter, Wilfried Elmenreich, Salvatore D'Alessandro, and Andrea M Tonello. 2014. GREEND: An Energy Consumption Dataset of Households in Italy and Austria. In *IEEE International Conference on Smart Grid Communications (SmartGridComm)*.

[21] Lucas Pereira. 2017. EMD-DF: A Data Model and File Format for Energy Disaggregation Datasets. In *4th ACM International Conference on Systems for Energy-efficient Built Environments (BuildSys)*.

[22] Lucas Pereira and Nuno Nunes. 2018. Performance Evaluation in Non-Intrusive Load Monitoring: Datasets, Metrics, and Tools – A Review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 6 (2018).

[23] Lucas Pereira, Miguel Ribeiro, and Nuno Nunes. 2017. Engineering and Deploying a Hardware and Software Platform to Collect and Label Non-Intrusive Load Monitoring Datasets. In *5th IFIP Conference on Sustainable Internet and ICT for Sustainability (SustainIT)*.

[24] Thomas Picon, Mohamed Nait Meziane, Philippe Ravier, Guy Lamarque, Clarisse Novello, Jean-Charles Le Bunetel, and Yves Raingeaud. 2016. COOLL: Controlled On/Off Loads Library, a Public Dataset of High-Sampled Electrical Signals for Appliance Identification. *CoRR* abs/1611.05803 (2016).

[25] Andreas Reinhardt, Paul Baumann, Daniel Burgstahler, Matthias Hollick, Hristo Chonov, Marc Werner, and Ralf Steinmetz. 2012. On the Accuracy of Appliance Identification Based on Distributed Load Metering Data. In *2nd IFIP Conference on Sustainable Internet and ICT for Sustainability (SustainIT)*.

[26] Martin Ringwelski, Christian Renner, Andreas Reinhardt, Andreas Weigel, and Volker Turau. 2012. The Hitchhiker's Guide to Choosing the Compression Algorithm for your Smart Meter Data. In *IEEE International Energy Conference and Exhibition (ENERGYCON)*.

[27] Benjamin Völker, Philipp M. Scholl, and Bernd Becker. 2019. Semi-Automatic Generation and Labeling of Training Data for Non-intrusive Load Monitoring. In *10th ACM International Conference on Future Energy Systems (e-Energy)*.

[28] Jacques L. Willems. 2010. The IEEE Standard 1459: What and Why?. In *IEEE International Workshop on Applied Measurements for Power Systems (AMPS)*.

[29] Michael Zeifman and Kurt Roth. 2011. Nonintrusive Appliance Load Monitoring: Review and Outlook. *IEEE Transactions on Consumer Electronics* 57, 1 (2011).