Energy Disaggregation for Commercial Buildings: A Statistical Analysis

Simon Henriet
LTCI, Telecom Paristech
Université Paris Saclay, France
Smart Impulse, France
simon.henriet@telecom-paristech.fr

Umut Simsekli and Gaël Richard
LTCI, Telecom Paristech
Université Paris Saclay, France
umut.simsekli@telecom-paristech.fr
gael.richard@telecom-paristech.fr

Benoit Fuentes
Smart Impulse, France
benoit.fuentes@smart-impulse.com

Abstract—**In the recent years, there have been increasing academic and industrial interests in analyzing the electrical consumption of commercial buildings. Whilst having similarities with the Non Intrusive Load Monitoring (NILM) tasks for residential buildings, performing NILM on signals that are collected from large commercial buildings exhibits additional challenges and difficulties. In this study, we explore the statistical characteristics of datasets that are collected from commercial and residential buildings. We show that they are significantly different in several aspects and we show that this information can be used for data simulation and algorithm development.**

I. Introduction

With the increasing awareness about the problem of climate change and the increasing level of energy consumption, a need for energy efficiency has emerged. At the Paris conference of the parties (COP21) [1], many countries have recognized energy efficiency as the basis of energy transition. An important step towards energy efficiency is based on reducing the energy consumption in residential and commercial buildings. To this end, one needs to measure and analyze the power consumption profiles of the devices that are installed in buildings.

There are two main research directions for electrical load monitoring: (i) full sub-metering and (ii) non-intrusive load monitoring (NILM). The former requires installing a submeter on each electrical device plugged into the network. While being accurate, this approach has important financial and computational limitations since it requires an excessive amount of measurement devices. On the contrary, the latter, the main subject of this study, involves only one sensor per building, installed at the entrance of the electrical network and therefore has a much less demanding data collection process. However, since the measured signals contain information coming from all the devices in this case, NILM requires accurate energy disaggregation algorithms for estimating the electrical consumption of each device.

Recently, there has been an increasing academic and industrial interest in applying NILM to commercial buildings. These buildings include large offices, warehouses, retails or shopping malls, and as also pointed out in [2], have fundamentally different characteristics than those of residential buildings.

An important limitation for developing disaggregation algorithms for commercial buildings is the lack of publicly available datasets that contain detailed measurements of individual devices collected from commercial buildings. To the best of our knowledge, there is only one public dataset that is collected from a commercial building, namely the COMBED dataset [2]. This dataset contains the power consumption measurements of two buildings (an academic and a library blocks) and is sampled at $1/30$ Hz. Even though it is a first step towards energy disaggregation in commercial buildings, the dataset does not include high frequency data (current or voltage) and the equipment present are not fully sub-metered.

In this paper, we aim at taking the first step towards circumventing the issues caused by the lack of knowledge for commercial buildings. We perform a statistical analysis on public residential datasets and compare them to a private dataset that is collected from real commercial buildings in France, in order to have a better understanding of the statistical differences between the two kinds of buildings. We believe that our analyzes would be useful for generating realistic synthetic data for commercial buildings, which is left as a future work.

II. Related work

NILM for commercial buildings started with Norford's work [3]. He identified three main challenges for tackling commercial buildings: (i) load detection; due to the recurrence of overlapping events (switching on or off) (ii) load estimation; due to variations in load for several devices and (iii) load identification; due to similarity in low frequency features for different devices. Batra also pointed out that the hypotheses made by existing NILM approaches, such as the "one-at-a-time" assumption (at most one device changes of state at each instant) or the "constant load" assumption (only devices of category "on/off" or "multi-state"), do not hold in this context and showed that NILM algorithms developed for residential buildings fail when applied to commercial buildings [2]. To overcome low frequency data limitations, Lee used current harmonics to separate variable speed drives from aggregate data in commercial buildings [4].

III. Statistical Analysis of Residential and Commercial NILM Datasets

A. Datasets

In most commercial or residential buildings, the electric power is delivered as alternating current (AC) (sinusoidal voltage) and distributed with 1, 2 or 3 phase lines, corresponding to fix voltage phases difference. The different quantities that can be measured by the sensors are energy per period (kWh), instantaneous or average real power in watt (W) or current and voltage in ampere (A) and volt (V). These quantities are related to the notion of sampling frequency. A common definition in the literature is to consider as high frequency (HF) a measurement occurring multiple times within an electrical period (defined by the fundamental frequency of the voltage) and as low frequency (LF) a measurement that occurs at a lower frequency than the fundamental. HF measurements generally correspond to current and voltage whereas LF measurements correspond to power or energy.

TABLE I
THE PUBLIC AND PRIVATE DATASETS USED IN THIS STUDY.

Name	Data	Buildings	Phases	Freq.	Type
BLUED [5]	current	1	2	12 kHz	resid.
UK-DALE [6]	current	1	1	16 kHz	resid.
REDD [7]	current	2	2/1	16.5 kHz	resid.
SIHF [private]	current	7	3	kHz	comm.
REDD [7]	power	6	2	1 Hz	resid.
ECO [8]	power	6	1	1 Hz	resid.
IAWE [9]	power	1	1	1 Hz	resid.
UK-DALE [6]	power	5	1	1/6 Hz	resid.
REFIT [10]	power	20	1	1/8 Hz	resid.
RAE [11]	power	1	2	1/15 Hz	resid.
COMBED [2]	power	1	1	1/30 Hz	comm.
SILF [private]	power	7	3	1/30 Hz	comm.

In the last decade, we have witnessed the release of multiple publicly available datasets of different quality and with different sampling strategies. In this section, our goal is to compare residential to commercial buildings from a statistical point of view at both high and low frequency (at least $1/30$ Hz). The public datasets used for this study range from low frequency [8], [7], [10], [9], [11], [2], [6] to high frequency sampling [7], [5], [6] and correspond to measurements of individual houses (except for one which comes from an university building [2]). From each dataset we have selected houses or buildings whose measurements last longer than a week. In addition to public data, two private datasets are used. It consists of both low frequency power data (named SILF) and high frequency current measurements (named SIHF) from 7 commercial buildings. All those datasets are shown in Table I.

B. Physical preliminaries

Before getting to the statistical analysis, we shall introduce some notations and recall the relation between physical quantities. The digitalized voltage and current waveforms are denoted: $\mathbf{u}\,(n,t)$ and $\mathbf{i}\,(n,t)$, where $t=1,\ldots,T$ is the voltage period index, T denotes the total number of voltage periods, and n is the sampling index within a voltage period. The number of samples within a period of the voltage sine wave is supposed to be constant and is noted N. This segmentation according to the voltage period enables us to have a matrix representation of both current and voltage. The mean active power (or mean power consumption or load curve) for a voltage period is then given by:

$$\mathbf{p}(t)=\frac{1}{N}\sum_{n=1}^{N}\mathbf{u}(n,t)\,\mathbf{i}(n,t). \tag{1}$$

It is possible to down-sample or aggregate this signal by averaging several consecutive periods (in order to have a sample every 30 seconds for instance). For the sake of clarity, the same index t is kept regardless of the sampling frequency.

C. Power measurements (low frequency)

In order to discriminate residential buildings from commercial buildings, we are particularly interested in state change events, switching on/off events or continuous variations of electrical devices present in the building. These events result in global current signal variations and therefore, due to equation (1), in a time-varying power consumption. In this section, we used all the power datasets presented in I and power values have been calculated according to 1 for the current datasets. Power time series exhibit a strong temporal structure, characterized by high first-order autocorrelation (0.92 and 0.99 in average for respectively residential and commercial buildings at $1/30$ Hz). This can be explained by the fact that, when a device is switched ON it often remains active for several periods. This motivates us to rather study the power derivative rather than the power consumption:

$$\mathbf{p}'\,(t)=\mathbf{p}\,(t)-\mathbf{p}\,(t-1)\,, \tag{2}$$

and to characterize its structure at different time scale. To enable the comparison between buildings, the power derivative is normalized so that the mean is zero and the standard deviation is one.

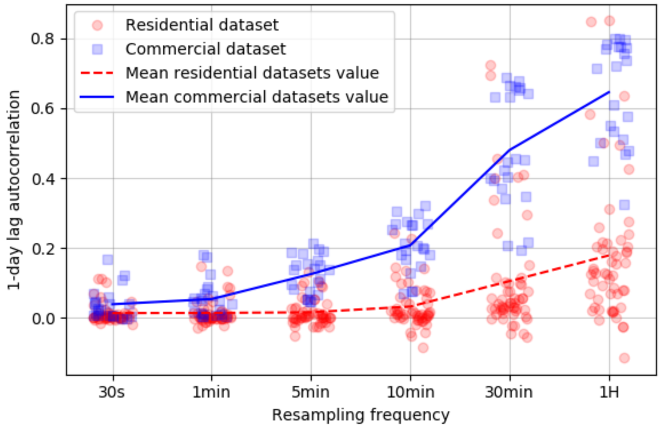


Fig. 1. Estimation of the 1 day lag autocorrelation for the power derivatives at different re-sampling frequencies for all the datasets (see Table I)

One important structure in time series is seasonality. It is a weak assumption to state that the power consumption and thus its derivative can show daily seasonality due to the habits of the people and time-scheduled equipments. The serial autocorrelation with a lag of 1 day of the power derivative is presented in

Figure 1. It first shows that the derivative of hourly aggregated power discriminates the two kinds of buildings, since the seasonal effect is higher for the commercial ones than for the residential ones (0.65 vs 0.18 in average). This can be interpreted by the fact that the consumption patterns are more periodical in commercial buildings than in residential: (i) many equipments are programmed and have recurrent patterns, (ii) the average behavior of occupants is more recurrent than individual behaviors. Figure 1 also shows that the seasonal effect is more intense at higher time scale.

At a $1/30$ Hz sampling frequency, the power derivative has almost no temporal structure (zero first-order autocorrelation) and can thus be studied as realizations of independant and indentically distributed random variables.

It can be observed in Figure 2 that the distribution of the power derivative for a residential building can be more peaky around zero and has a heavier tail than the one of a commercial building. Additionnaly, 3 statistics that accurately reflect the difference in distribution are presented: (i) the kurtosis, (ii) the entropy and (iii) the scale parameter of Laplace distribution.

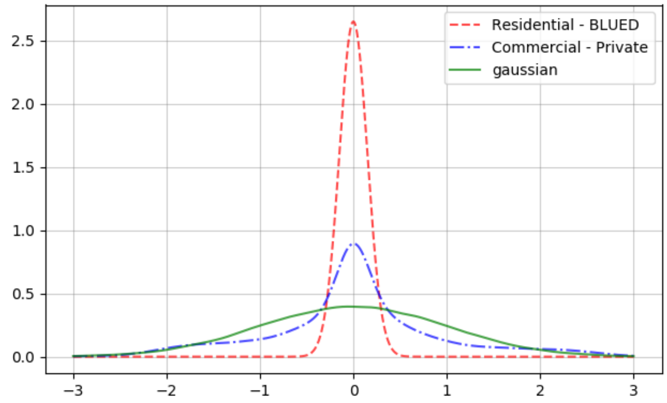


Fig. 2. Distribution of derivative power @ $1/30$Hz for all the datasets (see Table I)

Firstly, the kurtosis is based on a scaled version of the fourth moment of a distribution:

$$\mathrm{Kurt}[X] = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^2}, \tag{3}$$

where $\mathbb{E}$ is the mathematical expectation and X a random variable. It can be noted here that the kurtosis has often been used as a measure of impulsiveness: impulsive signals typically have a high kurtosis value [12]. Figure 3 shows a clear difference in kurtosis for the two types of building. On the one hand, high kurtosis value for residential building can be explained by low number of devices and simple devices (ON/OFF or multistate) which result in more impulsive power derivative signals. On the other hand, due to the Central Limit Theorem, the more independant individual devices there are, the closer the random variable resulting from the sum is to a Gaussian and the closer its kurtosis is to the one of the Gaussian distribution (i.e. equal to 3 for standard Gaussian). It can however be observed that the kurtosis for commercial buildings remains high compared to the kurtosis of the standard Gaussian distribution, and this characteristic can still be used in NILM algorithms.

Secondly, entropy is defined as the average amount of information produced by a stochastic source of data. It is based on the logarithm of the probability distribution:

$$\mathrm{H}[X] = \mathbb{E}\left[-\ln(P(X))\right], \tag{4}$$

Figure 3 shows that entropy values are higher for commercial buildings. This results from the fact that commercial datasets contain more devices and thus more information, which is more complex to encode. This can also come from the fact that there are much more devices with varying power in commercial buildings than in residential ones.

Finally, Laplace distribution is a high kurtosis distribution (such as our empirical distributions shown in Figure 2) that has two parameters: a location (μ) and a scale (b). The location parameter equals the mean of the distribution and is of less interest because it is close to 0 for power derivatives. In order to compare the datasets, we estimate the scale parameter considering the distributions as Laplace and then compare the estimated parameters. A maximum likelihood estimator of the scale parameter is given by:

$$\hat{b} = \frac{1}{N}\sum_{n=1}^{N}|x_i - \mu|, \tag{5}$$

As shown in Figure 3, the estimated scale parameters are higher for commercial buildings. We can finally remark that these 3 criteria promote sparseness in the data.[1]

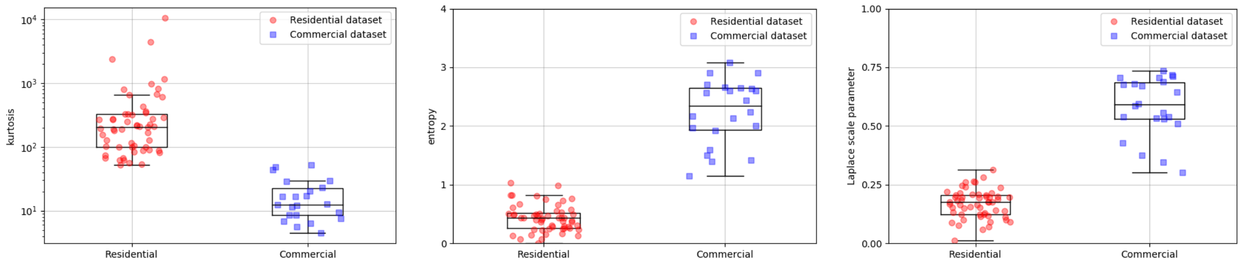
D. Current measurements (high frequency)

In buildings the voltage can be considered as pure sine wave. In frequency domain this is characterized by a signal with energy only on the fundamental frequency and no energy on harmonic frequencies. On the contrary, the current signal shows relatively important energies on harmonic frequencies due to non linear devices present on the network. This property can be measured with the total harmonic distortion (THD). It is based on the coefficients of the discrete Fourier transform (DFT) of the current signal. The DFT and the THD are computed for every period:

$$\mathrm{THD}(t) = 100 \times \frac{\sqrt{\sum_{h=2}^{N} \mathbf{I}(h,t)^2}}{\sqrt{\sum_{h=1}^{N} \mathbf{I}(h,t)^2}}, \tag{6}$$

where $\mathbf{I}(h,t)$ is the h^{th} coefficient of the DFT of $\mathbf{i}(.,t)$. Figure 4 shows lower values for commercial buildings that may be explained by the presence of big linear induction motors (heating, ventilation or air conditioning) which do not have harmonics energy.

[1] For Laplace distributed random variable, entropy and the scale parameter are linked: $\mathrm{H}[X] = \log(2\mathrm{b}\mathrm{e})$.



(a) Kurtosis

(b) Entropy

(c) Laplace scale parameter

Fig. 3. Statistical analysis of power changes at a 1/30 Hz sampling frequency for all the datasets (see Table I)

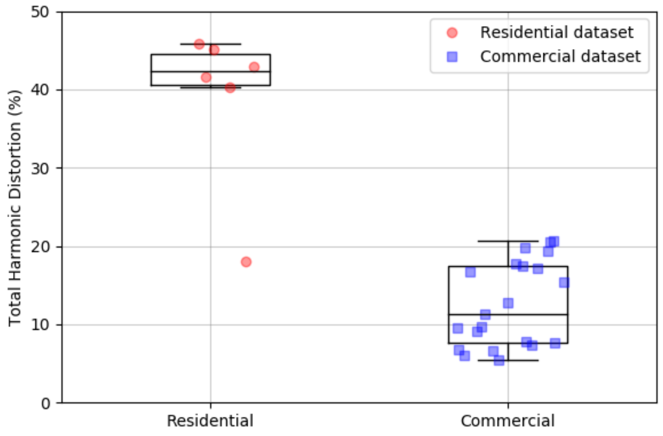


Fig. 4. Total Harmonic Distortion of current signals for all the ”current” datasets (see data column in Table I)

IV. Conclusion and Discussion

We have provided a data analysis on public and private datasets that showed significant differences between commercial and residential buildings. The study of the power derivative illustrated that residential distributions are more peaky at zero than commercial ones. We also showed that the kurtosis, entropy and Laplace scale parameter of the power derivative are good discriminative indicators for residential and commercial buildings. We have explained these differences by a higher amount of devices in commercial buildings and the presence of complex categories of devices.

These statistical characteristics are in contradiction with the hypothesis used for residential NILM algorithms (”one at a time” and ”constant load”). In this context, detecting a single event on the power curve is a difficult task and this explains why residential NILM algorithms fail when applied to commercial buildings. The statistical metrics used in our study suggest that using a soft version of the ”one at a time” hypothesis such as ”few at a time” (only a few devices are responsible of the power variations at every instant) would be more realistic.

The lack of NILM datasets (consumption measures for the entire building and for individual devices) is a major difficulty in testing and evaluating diaggregation algorithms. Developping a building simulator is one solution to this issue. This analysis can provide the basis for a statistical evaluation of the quality of simulated data.

References

[1] P. Protocol, “Report of the conference of the parties,” in United Nations Framework Convention on Climate Change (UNFCCC), 2015.

[2] N. Batra, O. Parson, M. Berges, A. Singh, and A. Rogers, “A comparison of non-intrusive load monitoring methods for commercial and residential buildings,” arXiv preprint arXiv:1408.6595, 2014.

[3] L. K. Norford and S. B. Leeb, “Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms,” Energy and Buildings, vol. 24, no. 1, pp. 51–64, 1996.

[4] K. D. Lee, S. B. Leeb, L. K. Norford, P. R. Armstrong, J. Holloway, and S. R. Shaw, “Estimation of variable-speed-drive power consumption from harmonic content,” IEEE Transactions on Energy Conversion, vol. 20, no. 3, pp. 566–574, 2005.

[5] A. Filip, “Blued: A fully labeled public dataset for event-based non-intrusive load monitoring research,” in 2nd Workshop on Data Mining Applications in Sustainability (SustKDD), 2011.

[6] 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.

[7] J. Z. Kolter and M. J. Johnson, “Redd: A public data set for energy disaggregation research,” in Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, vol. 25, 2011, pp. 59–62.

[8] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini, “The eco data set and the performance of non-intrusive load monitoring algorithms,” in Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings. ACM, 2014, pp. 80–89.

[9] N. Batra, M. Gulati, A. Singh, and M. B. Srivastava, “It’s different: Insights into home energy consumption in india,” in Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings. ACM, 2013, pp. 1–8.

[10] D. Murray, L. Stankovic, and V. Stankovic, “An electrical load measurements dataset of united kingdom households from a two-year longitudinal study,” Scientific data, vol. 4, p. 160122, 2017.

[11] S. Makonin, Z. J. Wang, and C. Tumpach, “Rae: The rainforest automation energy dataset for smart grid meter data analysis,” arXiv preprint arXiv:1705.05767, 2017.

[12] Z. Liang, J. Wei, J. Zhao, H. Liu, B. Li, J. Shen, and C. Zheng, “The statistical meaning of kurtosis and its new application to identification of persons based on seismic signals,” Sensors, vol. 8, no. 8, pp. 5106–5119, 2008.