A generative model for non-Intrusive load monitoring in commercial buildings

Simon Henriet [a,b,*], Umut Şimşekli [a], Benoit Fuentes [b], Gaël Richard [a]

[a] LTCI, Télécom Paristech, Université Paris Saclay, Paris, France
[b] Smart Impulse, Paris, France

ARTICLE INFO

ABSTRACT

In the recent years, there has been an increasing academic and industrial interest for analyzing the electrical consumption of commercial buildings. Whilst having similarities with the Non Intrusive Load Monitoring (NILM) tasks for residential buildings, the nature of the signals that are collected from large commercial buildings introduces additional difficulties to the NILM research causing existing NILM approaches to fail. On the other hand, the amount of publicly available datasets collected from commercial buildings is very limited, which makes the NILM research even more challenging for this type of large buildings. In this study, we aim at addressing these issues. We first present an extensive statistical analysis of both commercial and residential measurements from public and private datasets and show important differences. Secondly, we develop an algorithm for generating synthetic current data based on a modelization of the current flowing through an electrical device. We then demonstrate that our electrical device model fits well real measurements and that our simulations are realistic by using the quantitative metrics described in the previous section. Finally, to encourage research on commercial buildings we release a synthesized dataset called SHED that can be used to evaluate NILM algorithms.

© 2018 Elsevier B.V. All rights reserved.

1. 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 reducing greenhouse gases.

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 the buildings. In this context, depending on the particular application, one can be interested in either the estimation of the total energy consumed by devices in a certain period, the operational schedule of a particular device, the power consumption profile at a certain timestep or the estimation of electrical reliability.

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 sub-meter 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 and a complex system to centralize the data and maintain its quality. 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, NILM requires accurate energy disaggregation algorithms for estimating the electrical consumption of each device.

The majority of the current NILM literature has been focused on residential buildings using low frequency (<50 or 60 Hz) power data obtained by smart meters. 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. On top of this, the use of high frequency measurements (such as current waveforms) has been enabled by the development of low cost meters [3] and can significantly change the classical power based NILM approaches.

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. Unfortunately, collecting such data turns out to be a very challenging and expensive task since it requires installing sensors on each device in a large building, and the quality of these measurements is difficult to be maintained during a long period. 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 equipments are not fully sub-metered. As a consequence of the lack of data, developing supervised machine learning algorithms for NILM becomes even more challenging. Unsupervised algorithms require less data to be developed but still need data of good quality for evaluation purposes.

In this paper, we propose a comprehensive framework for energy disaggregation in commercial buildings. We aim at circumventing the issues caused by the lack of knowledge and data available. We first 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 differences between the two kinds of buildings. We then develop a building model for high frequency current waveforms and a synthetic data generation algorithm that is able to learn parameters on real measurements and then use it to produce new realistic data. In the light of the results obtained from our statistical analyses and by making use of both the publicly available datasets and a private dataset, we conduct various experiments for evaluating the quality of our device model and building simulations. To finally foster the NILM research for commercial buildings, we release a synthetic dataset, called SHED[1], that is generated by our algorithm.

2. Related work

NILM formalism for residential building has first been introduced in two papers by Hart and Sultanem [4,5]. Hart studied electrical devices from a low frequency power consumption point of view and classified them as: (i) on/off or constant device, (ii) multi-state and (iii) continuously varying. Sultanem worked on high frequency current measurements and introduced the notion of harmonic content to group devices into categories such as: (i) resistive, (ii) motor-driven, (iii) electronically-fed or (iv) fluorescent lighting.

Following these first papers, many NILM algorithms have been developed for residential buildings. They all share the same major features: (i) a power disaggregation step and (ii) a load classification step. In this context, they can be classified into two categories. The first category, 'event-based disaggregation' techniques, aims at detecting certain events (e.g. on/off transition, change of state) and then assign the detected events to different electrical devices [4,6–8]. It relies on two assumptions: (i) 'one at time', at most one device changes of state at each instant, and (ii) 'constant load', the consumption remains constant between two events. The second category, "source separation" techniques, assumes that the measured signal is a *mixture* of unknown *source* signals that correspond to different devices. The goal is then to recover the source signals from the mixture signal. Most of the proposed methods use low frequency power measurements [9–11] whereas Lange and Bergés are using source separation on high frequency current measurements [12].

NILM for commercial buildings started with Norford and Leeb's work [13]. They identified three main challenges for tackling commercial buildings: (i) load detection; due to the recurrence of overlapping events (two equipments starting at the same time) (ii) load estimation; due to variations in load for several devices (e.g. a variable speed engine) and (iii) load identification; due to similarity in low frequency features for different devices (two different types of light bulbs: a LED and a compact fluorescent lamp). Batra et al. also pointed out that the hypotheses made by existing "event-based" NILM approaches 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 et al. used current harmonic content to separate variable speed drives from aggregate data in commercial buildings [14]. As underlined by Zeifman and Roth [15], there is a large agreement that raising the sampling frequency also increases the probability of getting accurate estimation of individual consumption [16]. Since in commercial buildings we are facing a larger number of devices, high frequency measurements can also help in the disaggregation task.

Substantial efforts have been made to model electrical devices consumption and simulate datasets in order to evaluate NILM algorithms. In [17], Fischer et al. use "on/off" models with a probability of a device to be switched on depending on the time of the day. Other approaches [18,19] define more complicated models that can take into account uniform randomness during operation time, multi-state devices or exponentially decaying load curves. Even though these models are efficient for electrical devices in residential buildings, they lack of complexity to be used in commercial settings which contain smoothly varying devices and need high frequency measurements.

It is worth mentioning that high frequency current measurements have been studied in several papers [5,20,21]. Lam et al. used high frequency current/voltage trajectories to classify electrical devices [22]. Public datasets of high frequency current measurements of residential equipments have also been released [23,24].

Finally, Liang et al. [21] developed a simulator for high frequency current measurements but without considering long term modeling of current dynamics.

3. Statistical analysis of residential and commercial NILM datasets

In this section, we explore the statistical differences of electrical signals in commercial and residential buildings. Our goal is to develop quantitative metrics that can describe electrical power signals and differentiate both types of buildings. We first define the physical quantities and the datasets used. Then, we study the temporal structure of the power, the distribution of the temporal derivative of the power, and the harmonic content of current waveforms. This will guide us to develop the tools for understanding and differentiating the signals coming from residential and commercial buildings.

3.1. 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.

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 [2,26–31] to high frequency sampling [25–27] 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, without distinguishing the weekdays and the weekends. In addition to public data, two private datasets are used. It consists of both low frequency total power data (named SILF) and high frequency total current measurements (named SIHF) from 7 commercial buildings. All those datasets are shown in Table 1.

3.2. Definitions

Before diving into the statistical analysis, we shall introduce some notations and recall the relation between physical quantities. The digitized voltage and current waveforms are denoted: $\mathbf{u}(n, t)$ and $\mathbf{i}(n, t)$, where $t = 0, \ldots, T-1$ 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:

$$\mathbf{p}(t) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{u}(n, t)\mathbf{i}(n, t). \tag{1}$$

It is possible to down-sample or aggregate this signal by averaging several consecutive periods (this enables us to resample all the dataframe to the same frequency). For the sake of clarity, the same index t is kept regardless of the sampling frequency.

3.3. Power measurements (low frequency)

In this part we study power measurements statistical properties. Particular attention is taken to time to time power changes distribution and seasonal effects. In order to discriminate residential from commercial buildings, we are particularly interested in state change events (switching on or off, different speed or heating levels) or continuous variations of electrical devices present in the building. These events result in total current signal variations and therefore, due to Eq. (1), in a time-varying power consumption. In this section, we used all the power and current datasets presented in Table 1. Power values have been calculated according to (1) for current datasets. Power time series exhibit a temporal structure defined here by high first-order autocorrelation[2] (0.92 and 0.99 for respectively residential and commercial buildings at 1/30 Hz, in average over all datasets). 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 study the power derivative
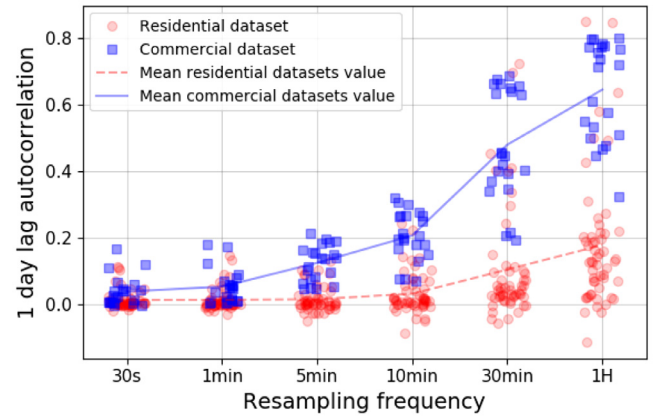

Fig. 1. Estimation of the 1 day lag autocorrelation for the power derivatives at different re-sampling frequencies for all the datasets (see Table 1).

Table 1
The public and private datasets used in this study.

Name	Data	Buildings	Phases	Frequency	Type
BLUED [25]	Current	1	2	12 kHz	Residential
UK-DALE [26]	Current	1	1	16 kHz	Residential
REDD [27]	Current	2	2/1	16.5 kHz	Residential
SIHF [private]	Current	7	3	kHz	Commercial
REDD [27]	Power	6	2	1 Hz	Residential
ECO [28]	Power	6	1	1 Hz	Residential
IAWE [29]	Power	1	1	1 Hz	Residential
UK-DALE [26]	Power	5	1	1/6 Hz	Residential
REFIT [30]	Power	20	1	1/8 Hz	Residential
RAE [31]	Power	1	2	1/15 Hz	Residential
COMBED [2]	Power	1	1	1/30 Hz	Commercial
SILF [private]	Power	7	3	1/30 Hz	Commercial

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 scales. To enable the comparison between buildings, the power derivative is normalized so that the mean is zero and the standard deviation is one.

One important structure in time series is the 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 seasonal effect is characterized here by the autocorrelation with a lag of 1 day of the power derivative (presented in Fig. 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 over all datasets). 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. Fig. 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 independent and identically distributed random variables. It can be observed in Fig. 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. In order to reflect the statistical differences between residential and commercial buildings, we have selected three statistics that can provide an accurate summary for these distributions: (i) the kurtosis, (ii) the entropy and (iii) the scale parameter of Laplace distribution.

---

[2] For a stationary signal x_t, $t = 0, \ldots, T-1$, the autocorrelation at lag τ is defined as: $R_\tau(x) = \frac{\sum_{t=0}^{T-1-\tau}(x_t-\bar{x})(x_{t+\tau}-\bar{x}_\tau)}{\sqrt{\sum_{t=0}^{T-1-\tau}(x_t-\bar{x})^2}\sqrt{\sum_{t=0}^{T-1-\tau}(x_{t+\tau}-\bar{x}_\tau)^2}}$, where $\bar{x}$ and $\bar{x}_\tau$ are the sample means of respectively x_t and $x_{t+\tau}$.
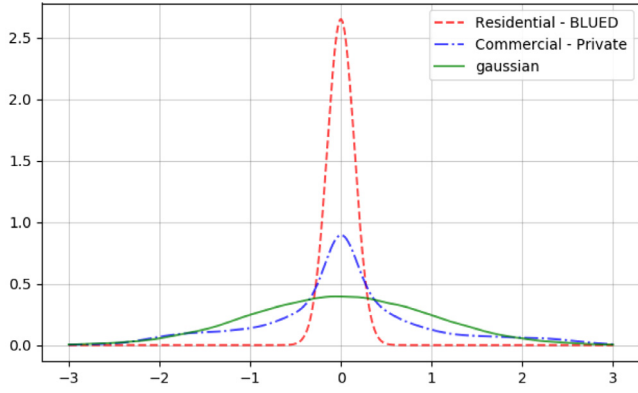
Fig. 2. Distribution of power derivatives @ 1/30 Hz for all the datasets (see Table 1).

Firstly, the kurtosis is based on a scaled version of the fourth moment of a random variable X:

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

where $\mathbb{E}$ is the mathematical expectation. In practice, we used the estimation of the kurtosis of the power derivatives given by the SciPy library [32]. The kurtosis has often been used as a measure of impulsiveness: impulsive signals typically have a high kurtosis value [33]. Fig. 3 shows a clear difference in kurtosis for the two types of building. On one hand, high kurtosis value for residential buildings can be explained by low number of devices and simple devices (ON/OFF or multi-state) which result in more impulsive power derivative signals. On the other hand, when the number of independent devices increases, the distribution of the sum of the power derivative becomes closer to a Gaussian distribution, as stated by the central limit theorem. It explains why kurtosis values for commercial buildings are closer to the standard Gaussian kurtosis value (3) than kurtosis values for residential buildings. 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 density (noted P) of a random variable X:

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

In practice, we used the estimation of the entropy of the power derivative given by the SciPy library [32]. Fig. 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, we analyze the high-kurtosis of the power derivative data by using the Laplace distribution, which is well-known for having a high kurtosis as well and is popular for modeling non-Gaussian data [34]. The Laplace distribution 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 0 for our normalized power derivatives. The scale parameter is proportional to the variance of the random variable. 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:

$$\hat{b} = \frac{1}{T} \sum_{t=0}^{T-1} |x_t - \mu|, \tag{5}$$

where x_t represents in our case the power derivatives. As shown in Fig. 3, the estimated scale parameters are higher for commercial buildings. We can finally remark that these 3 criteria promote sparseness in the data.[3]

3.4. 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:

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

where $\mathbf{I}(h,t) = \sum_{n=0}^{N-1} \mathbf{i}(n,t).\exp\left(-\frac{2j\pi}{N}hn\right)$ is the hth coefficient of the DFT of $\mathbf{i}(.,t)$. Fig. 4 shows lower values for commercial buildings that may be explained by an important proportion of linear induction motors (heating, ventilation or air conditioning) which do not create current harmonics, since in these devices, the current becomes a linear transformation of the voltage, and the voltage typically does not contain any harmonics.

4. A high frequency current waveform model

In this section, we develop a physically-inspired data model that will enable us to reproduce and simulate the behavior of the electrical network of a building in a bottom-up procedure. In our modeling strategy, we first break down the overall modeling problem into simpler subproblems by making use of the following hierarchical structure: an *individual electric device* belongs to a *category* of devices, and a *building* contains multiple devices that correspond to different categories. In agreement with this taxonomy, we first develop a generative model for an *individual building*, and for each building we then develop a model for each *device category*. These two models are based on Kirchhoff's current law. Subsequently, we propose a generative model for each device by enforcing a low-rank structure on the current waveforms. Finally, we use analytical tools to illustrate that our device model fits well real current measurements.

4.1. The building model

The model that we put forward in this section relies on several hypotheses. First, all the electrical devices are supposed to be plugged in parallel on the network: the current waves observed on the root node of the network are then the sum of the currents of all devices. This is a direct application of the Kirchoff's current law. Then, the electrical network is supposed to be in ideal conditions: the voltage is considered to be identical on each node of the network and independent from the current. Moreover, in the following, all current signals of devices are supposed to be independent. This assumption holds only if the voltage signal is purely periodic

---

[3] For Laplace distributed random variable, entropy and the scale parameter are linked: $H[X] = \log(2be)$.

(a) Kurtosis
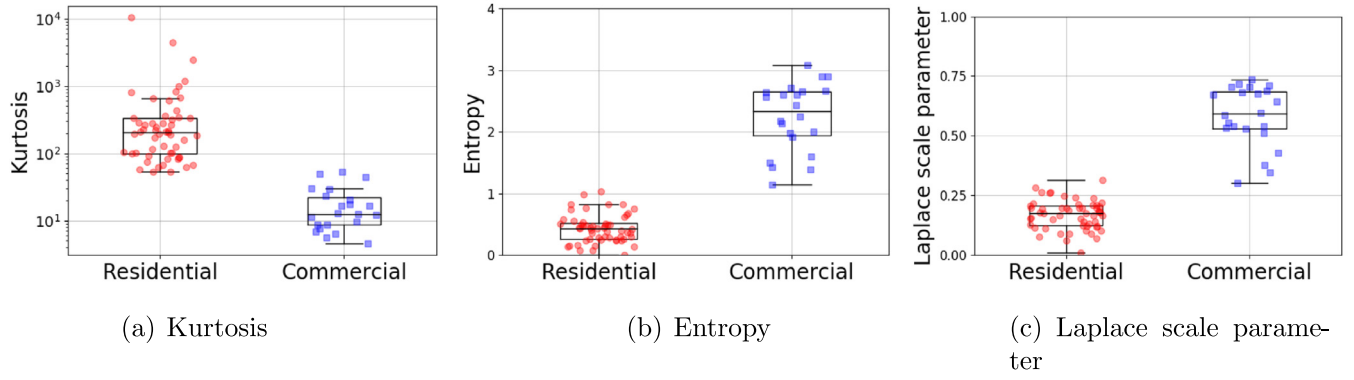
(b) Entropy

(c) Laplace scale parameter

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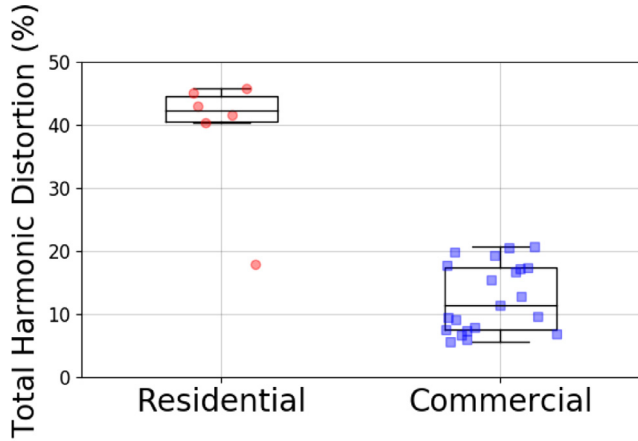

Fig. 4. Total Harmonic Distortion of current signals for all the "current" datasets (see data column in Table 1).

since the current waveform depends on the voltage waveform for most devices: $\forall t$, $\mathbf{u}(n,t) = \mathbf{u}_0(n)$.

Finally, for the sake of simplicity, only single-phase electrical networks are considered here, but three-phase networks can be simulated in a similar fashion.

These assumptions lead us to the following model for total current:

$$\mathbf{i}(n,t) = \sum_{c \in C} \mathbf{i}_c(n,t) + \epsilon(n,t) \tag{7}$$

where $\mathbf{i}$ is the total current measured at the root node of the network, $\mathbf{i}_c$ is the current signal of a category c of appliances, C is the ensemble of category indices, and $\epsilon(n,t)$ is a zero-mean Gaussian noise.

4.2. The category model

Since the number of identical equipments can be high in large buildings (*e.g.* corridors light bulb, computers or resistive heaters), it is often more important to evaluate a whole category consumption instead of each single device consumption (especially for specific NILM applications such as energy management). We then define herein a category as the aggregation of one to many similar equipments as follows:

$$\mathbf{i}_c(n,t) = \sum_{d \in D_c} \mathbf{i}_{c,d}(n,t) \tag{8}$$

where $\mathbf{i}_{c,d}$ is the current of device d belonging to category c. D_c corresponds to the set of devices belonging to category c.

Table 2
A new devices' taxonomy based on high frequency current features.

Activations signatures	Simple	Complex
Unique	On/Off or constant	Varying load
Multiple	Multi-state	Varying signature

4.3. The device model

Finally, the current of a particular device is modeled using a factorization-based approach, given as follows:

$$\mathbf{i}_{c,d}(n,t) = \mathbf{s}_{c,d}(n)\mathbf{a}_{c,d}(t) \tag{9}$$

where $\mathbf{s}_{c,d}$ and $\mathbf{a}_{c,d}$ are respectively called the *current waveform signature* and the *activation* of device d of category c. The waveform signature corresponds to a fixed pattern that describes the typical current response to the voltage. The activation is a nonnegative magnitude and its nature depends on the type of devices (0/1 function or continuously varying). As it will be demonstrated in Section 4.5, we extend our model for more complex devices by enabling the use of more than one signature in the factorization:

$$\mathbf{i}_{c,d}(n,t) = \sum_{k=1}^{K_{c,d}} \mathbf{s}_{c,d}(n,k)\mathbf{a}_{c,d}(k,t) \tag{10}$$

$K_{c,d}$ is the number of signatures and activations used to model device d. The number of components used in the model ($K_{c,d}$) and the nature of the activations ($\mathbf{a}_{c,d}$) enable us to classify the devices into 4 main classes, as shown in Table 2. In the literature [4,35], the common devices' taxonomy includes only 3 classes: (i) on/off or constant device, (ii) multi-state and (iii) continuously varying. This approach is based on low frequency features of load curves whereas we take high frequency characteristics into account. We can see in Table 2, that the main difference is that the original continuously varying class has been divided into two classes depending on the number of signatures used to model it (see Fig. 6).

To fix the inherent scale ambiguity of the multiplicative model of (10) (every scalar multiplication of a column of the signature can be canceled out by the same scalar division of the corresponding row of the activations), we normalize the signatures such that:

$$\forall c, d, k \quad \frac{1}{N} \sum_{n=1}^{N} \mathbf{s}_{c,d}(n,k).\mathbf{u}_0(n) = 1 \tag{11}$$

It has the double advantage to fix the multiplicative ambiguity and to directly link the activations to the consumed power. Indeed, the

power consumption of device d is given by:

$$\mathbf{p}_{c,d}(t) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{i}_{c,d}(n,t).\mathbf{u}_0(n)$$

$$= \sum_{k=1}^{K_{c,d}}\mathbf{a}_{c,d}(k,t)\frac{1}{N}\sum_{n=1}^{N}\mathbf{s}_{c,d}(n,k).\mathbf{u}_0(n) \qquad (12)$$

$$= \sum_{k=1}^{K_{c,d}}\mathbf{a}_{c,d}(k,t)$$

We can notice that in the case of a device with a single component ($K_{c,d} = 1$), the activation becomes the power consumption. Otherwise, the power equals the sum of the activations of each component.

4.4. The overall model

Combining the individual models (7), (8) and (10) gives the model for the total current:

$$\mathbf{i}(n,t) = \sum_{c\in C}\sum_{d\in D_c}\sum_{k=1}^{K_{c,d}}\mathbf{s}_{c,d}(n,k)\mathbf{a}_{c,d}(k,t) + \epsilon(n,t). \qquad (13)$$

Finally, taking into account equations (1) and (12), we obtain the following formula for the power per category:

$$\mathbf{p}_c(t) = \sum_{d\in D_c}\sum_{k=1}^{K_{c,d}}\mathbf{a}_{c,d}(k,t). \qquad (14)$$

4.5. Device model evaluation

In this section, we investigate the accuracy and quality of the proposed generative model. The proposed models for buildings and categories are inherently accurate since they are based on Kirchhoff's current law. Therefore, in this evaluation section we only focus on the device model that is proposed in Section 4.3.

In order to build an evaluation framework, we first use the observation that our device model for current waveforms is based on real-valued signatures and nonnegative activations. From this point of view, our model has close connections with the semi nonnegative matrix factorization (SNMF) algorithm [36], which aims to approximate an observation matrix as a product of a real-valued matrix and a non-negative matrix. For the problem at hand, SNMF attempts to solve the following optimization problem for a given device d and a category c:

$$\hat{\mathbf{s}}, \hat{\mathbf{a}} = \arg\min_{\mathbf{s},\mathbf{a}\geq 0}\frac{1}{2}\sum_{n,t}\left(\mathbf{i}(n,t) - \sum_{k=1}^{K}\mathbf{s}(n,k)\mathbf{a}(k,t)\right)^2 \qquad (15)$$

where $\mathbf{i}\in\mathbb{R}^{N\times T}$ is the current observation matrix of a device d in category c, $\mathbf{s}\in\mathbb{R}^{N\times K}$ is the signature matrix, and $\mathbf{a}\in\mathbb{R}_+^{K\times T}$ is the activation matrix. In our context, the goal of the SNMF algorithm becomes expressing the data matrix $\mathbf{i}$ (corresponding to a single device) in terms of the device model that is proposed in Section 4.3. Therefore, we can use the SNMF algorithm to assess the quality of the proposed device model: once the algorithm blindly estimates the signature and the activation matrices for a given device, we can use the residual $\sum_{n,t}\left(\mathbf{i}(n,t) - \sum_{k=1}^{K}\hat{\mathbf{s}}(n,k)\hat{\mathbf{a}}(k,t)\right)$ for quantifying the quality of our generative model that assumes $\mathbf{i}(n,t) \approx \sum_{k=1}^{K}\hat{\mathbf{s}}(n,k)\hat{\mathbf{a}}(k,t)$.

For measuring the goodness of fit of our model on real datasets, we use the Signal to Noise Ratio (SNR) between the true current signals and the residual that is provided by SNMF. The SNR is formally defined as follows:

$$\text{SNR} = 10\times\log_{10}\left(\frac{\sum_{n,t}\left(\hat{\mathbf{i}}(n,t)\right)^2}{\sum_{n,t}\left(\mathbf{i}(n,t) - \hat{\mathbf{i}}(n,t)\right)^2}\right), \qquad (16)$$

where $\mathbf{i}(n,t)$ is the real current measurement and $\hat{\mathbf{i}}(n,t) = \sum_{k=1}^{K}\hat{\mathbf{s}}(n,k)\hat{\mathbf{a}}(k,t)$ is the model reconstruction using the estimated signatures ($\hat{\mathbf{s}}$) and activations ($\hat{\mathbf{a}}$) given by the solution of (15).

The efficiency of our device model is evaluated using two high frequency public datasets which correspond to a few seconds of high frequency current measurements following the switching ON of a device [23,24] and a third private dataset which corresponds to high frequency current measurements over several days for devices in commercial buildings.

Fig. 5 shows that for several device categories only one component ($K = 1$) results in very high values of the SNR, which means a good data reconstruction. For more complex devices, Fig. 5 also shows the required number of components to reach a SNR value of at least 50 dB (which means an amplitude ratio of ≈ 316) between the model and the error). Fig. 6 illustrates the factorizations learned on the four kinds of devices as presented in Table 2. This experiment demonstrates the capacity of our model to catch the features of real current measurements. It also validates our choice to separate simple ($K = 1$) from complex ($K > 1$) devices in our models. It can be noticed that complex devices are for the majority found in commercial buildings and not in residential ones (e.g. air handling unit, lift, split, inverter).

5. A generative procedure for dataset simulations

In order to be able to simulate new datasets, we need to solve two more problems. First of all, the SNMF model used to estimate factors (signatures and activations) is analytical and do not provide any generating procedure to simulate new data. Secondly, the lack of publicly available high frequency datasets of individual equipments makes it difficult to learn both signature and activations on the same dataset. To circumvent these issues, we first propose separate generative models for signatures and activations. Then, we estimate their parameters and simulate new data independently for signatures and activations using different datasets: (i) short high frequency current measurements for signatures [23,24] and (ii) long low frequency power measurements for activations ([37] and our private dataset).

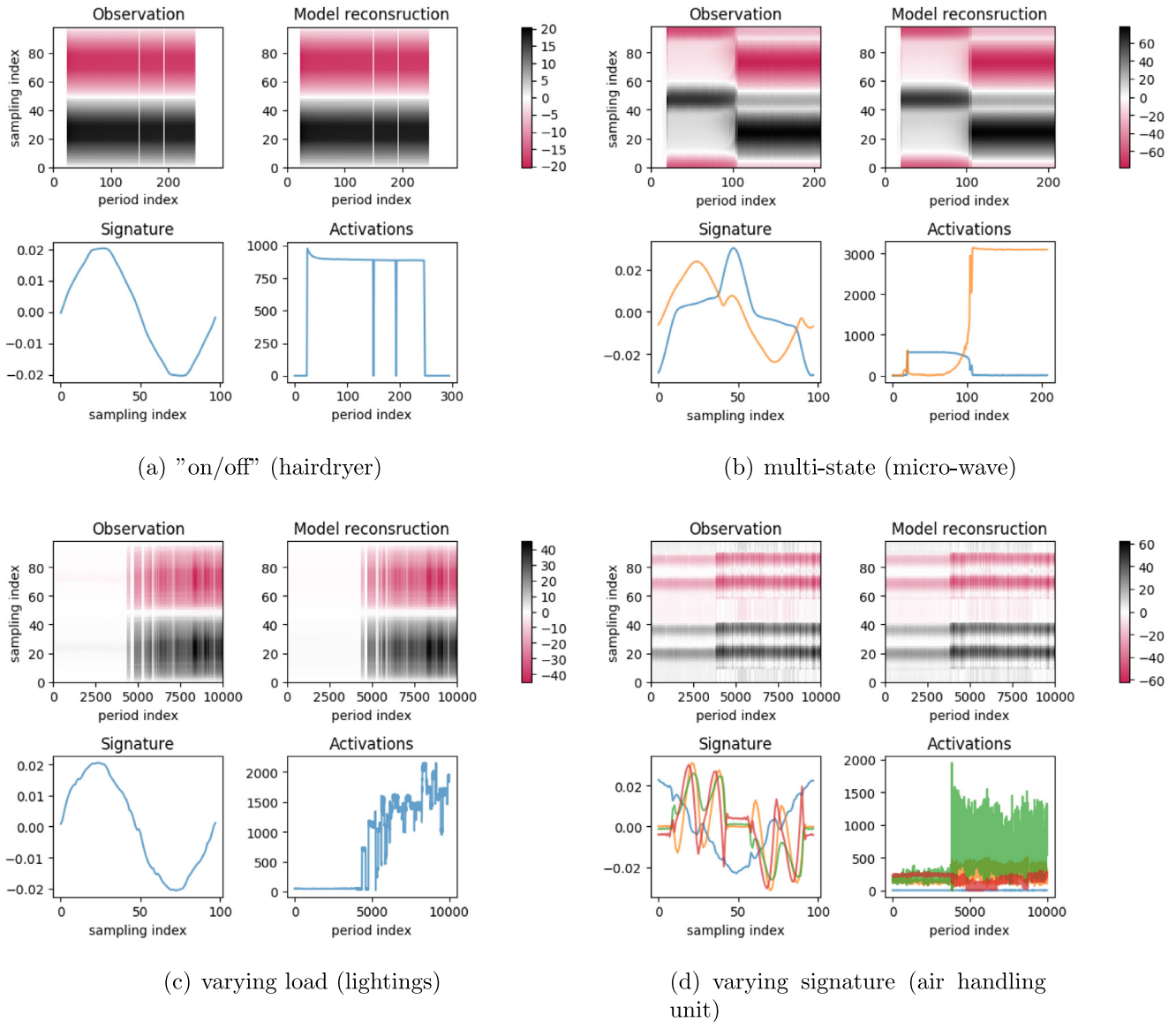### 5.1. Signature sampling algorithm

In order to generate signatures, we first decompose the current waveforms that correspond to an individual device, by using the SNMF algorithm that was defined in Section 4.5. Then, we use the output of the SNMF algorithm for generating random signatures. More precisely, we generate a signature $\mathbf{s}^{new}$, by using the following approach:

$$\mathbf{s}^{new}(n,k) \sim \mathcal{N}(\hat{\mathbf{s}}(n,k), \sigma^2) \qquad (17)$$

where $\mathcal{N}(\mu,\sigma^2)$ denotes the univariate Gaussian distribution with mean μ and variance σ^2, and $\hat{\mathbf{s}} = \arg\min_{\mathbf{s},\mathbf{a}\geq 0}\frac{1}{2}\sum_{n,t}\left(\mathbf{i}(n,t) - \sum_{k=1}^{K}\mathbf{s}(n,k)\mathbf{a}(k,t)\right)^2$ is a learned signature template by using the SNMF algorithm on the high frequency current signals belonging to individual devices. For this task we use the datasets [23,24]. The variance σ^2 is chosen as an hyperparameter. Fig. 6 shows four examples of learned signatures for different devices.
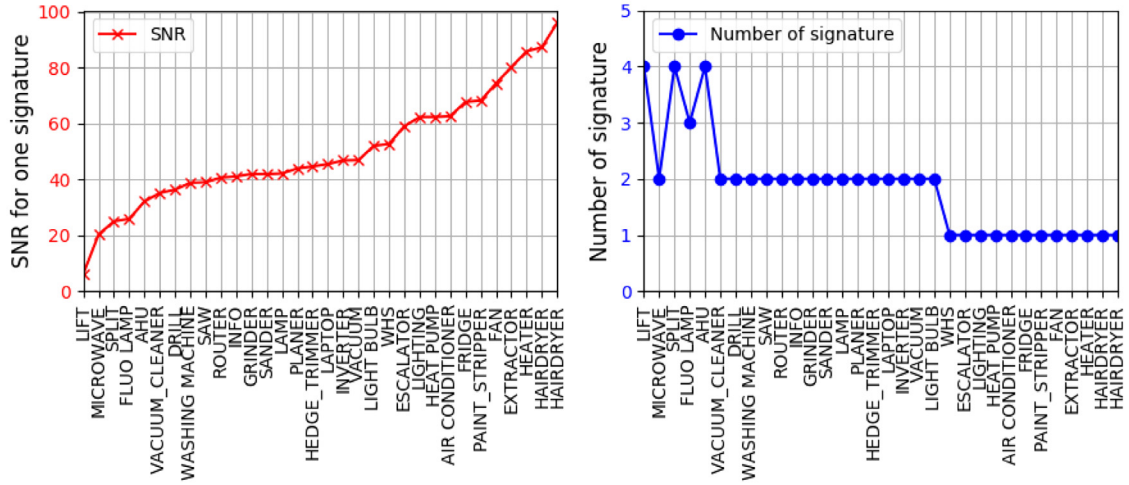
Fig. 5. Device model evaluation: (left) shows the reconstruction SNR with one signature and (right) shows the minimum number of signatures to use for reaching a SNR of at least 50 dB.

(a) "on/off" (hairdryer)

(b) multi-state (micro-wave)

(c) varying load (lightings)

(d) varying signature (air handling unit)

Fig. 6. Learned factorizations for the 4 device classes, each of them is composed of (top left) the observations in matrix shape (sampling index × period index), (top right) the model reconstruction (sampling index × period index), (bottom left) the signature matrix: each line corresponds to a column of the matrix and (bottom right) the activation matrix: each line corresponds to a row of the matrix. The number of components (K) is selected such that the SNR > 50 dB.
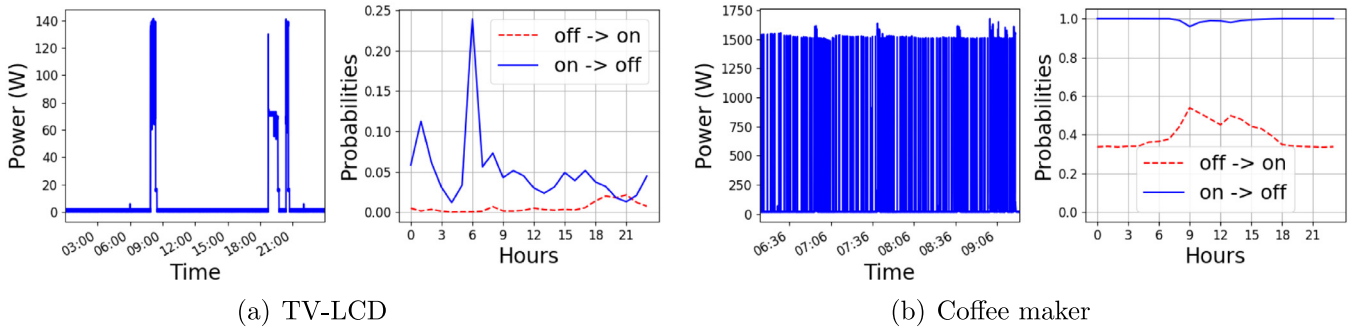
(a) TV-LCD

(b) Coffee maker

Fig. 7. Activation probabilities learned on public dataset (right) and a few hours of the measurements (left).
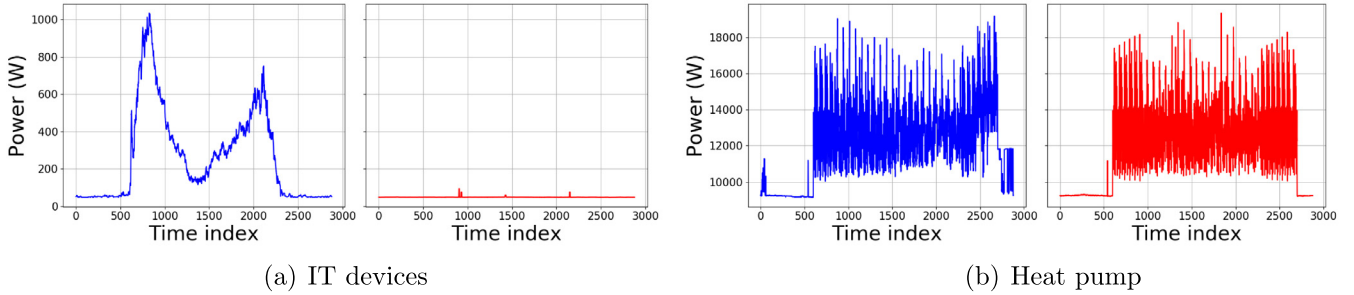

(a) IT devices

(b) Heat pump

Fig. 8. Activation templates learned on the private dataset, the templates correspond to one day (timestep = 30 s): week-day (left) and day-off (right).

5.2. Activation sampling algorithm

We describe here two different algorithms to simulate the activations: one for simple activations (on/off) and one for complex activations (continuously varying devices).

5.2.1. Simple activations

As mentioned in Section 2, a key feature of the activations is their temporal structure. Dinesh et al. [38] introduced a time-of-day usage pattern for a device defined by the probability of being activated at different periods of the day. These 'periods of the day' are defined as subsets of a partition of the time. In this study, we follow a similar procedure and partition the time in hours. For instance, one subset (a *period of the day* noted S_τ) may correspond to the slot 10 am to 11 am for every day. The total number of subsets is hence 24.

The approach that was proposed in [38] assumes that the activations only depend on the time of the day, and therefore it does not take the temporal dependence of the activations into account. We extend that approach by providing a generative model for on/off device activations that take into account the previous state of the device. We are considering here 0 or 1 activations and use a deterministic switching mode 2-state Markov chain to model the device's activation where the transition probability is defined as:

$$\forall \tau, \forall t \in S_\tau, \forall i, j \in \{0, 1\}^2 \quad P[\mathbf{a}(t) = i | \mathbf{a}(t-1) = j] = \gamma_\tau(i, j),$$
(18)

where t is the time index, S_τ is a period of the day and γ_τ the transition matrix for period of the day S_τ. This model enables us, first, to infer the transition probabilities depending on the period of the day from databases and, second, to generate new activations. Using maximum likelihood inference, the γ parameter is estimated by the following equation:

$$\hat{\gamma}_\tau(i, j) = \frac{\sum_{t \in S_\tau} \mathbb{1}_{(\mathbf{a}(t)=i, \mathbf{a}(t-1)=j)}}{\#S_\tau},$$
(19)

where $\#S_\tau$ is the size of subset S_τ. Intuitively, this estimation corresponds to counting the number of ON-to-OFF and OFF-to-ON events occurring during the period of the day S_τ. We are using the TRACEBASE dataset [37] gathering power measurements for individual devices for several days to estimate the parameters. Firstly, we transform the power time series into on/off time series using a simple thresholding mechanism: $\tilde{x}(t) = \mathbb{1}_{(x(t) > 20)}$. Secondly, we estimate the model parameters using (19). Finally, the learned parameters are used to generate new data:

$$\mathbf{a}^{new}(t) \sim \mathcal{B}er(\hat{\gamma}_\tau(1, \mathbf{a}^{new}(t-1)))$$
(20)

where $\mathcal{B}er$ is the Bernoulli distribution (which is the natural choice in two state Markov chains).

Fig. 7 shows two examples of simple activations data and the learned activations parameters. The learned activations show that the probability of switching ON is highest at 8 a.m. and 7 p.m. for the TV. It also shows that for the coffee maker, the probability of switching ON is quite high all day long and that once ON it immediately switches OFF.

5.2.2. Complex activations

In this part, we are considering generating activations by learning 'activation templates' on a private dataset due to the lack of public dataset for complex devices (load varying and signature varying, see Table 2). The private data is collected from two large commercial buildings in two different cities in France. It contains 11 device categories and is recorded during several weeks at low sampling frequency. The goal of the templates is to catch the typical power consumption of a device category during a *period of the day* and thus account for the daily seasonal effects shown in commercial buildings (see Section 3). Since many equipments are programmed to switch on or off on particular days (air handling unit, heaters) or depend on building occupancy (computers), we distinguish the week days and the days off. In this part the partition of the time is made with period of 30 s. The total number of subsets is then 5760 (2880 periods of 30 s per week days and days off). In order to compute such templates, we simply average the power

(a) Autocorrelation

(b) Kurtosis

(c) Entropy

(d) Laplace scale parameter

(e) THD
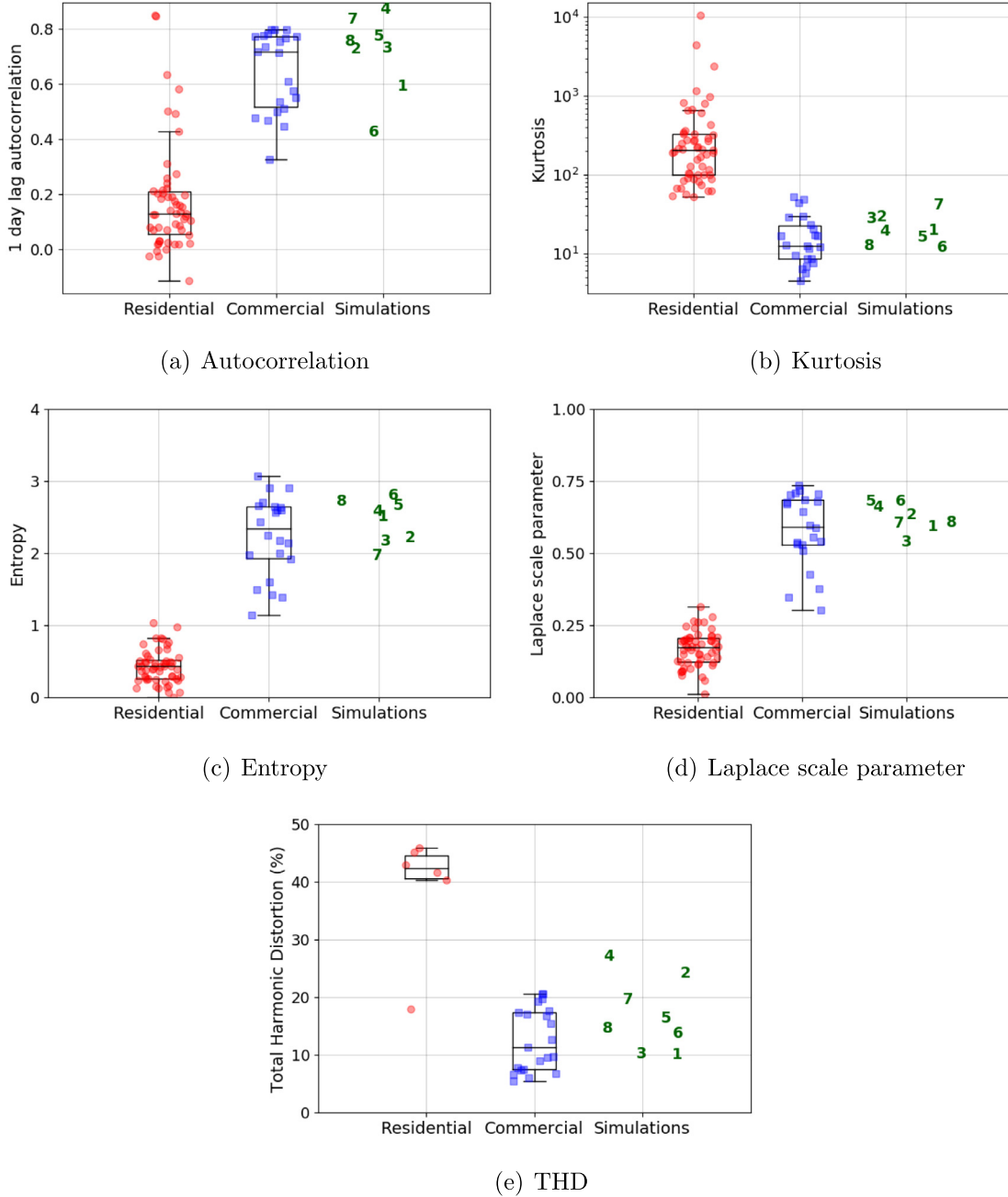
Fig. 9. Quantitative evaluation of the simulated datasets: comparison of the statistical metrics of simulations and real datasets. Every circle or square corresponds to one building. Numbers in simulations columns correspond to building indexes in the SHED dataset.

consumptions of individual devices over several weeks of data per period of the day:

$$\hat{\mathbf{a}}(\tau) = \frac{\sum_{t \in \mathcal{S}_\tau} \mathbf{p}(t)}{\#\mathcal{S}_\tau} \qquad (21)$$

The learned templates are illustrated in Fig. 8. We can observe that IT devices are switched off during day off and have smooth load curves whereas the heat pump has a more noisy consumption.

To generate new data, we multiply a positive noise with the templates to take the day to day variability into account:

$$\forall \tau, \forall t \in \mathcal{S}_\tau \quad \mathbf{a}^{new}(t) = \hat{\mathbf{a}}(\tau) \times \exp(\epsilon(t)), \qquad (22)$$

where ϵ follows an ARMA(p,q) process [39]. An ARMA process is a linear time serie model involving lagged values of itself and of a white noise. It is well used in time series modeling because

of its stationary property and its ability to model autocorrelation at different lags. It is defined as follows: $\epsilon(t) = \sum_{i=1}^{p} \theta_i \epsilon(t - i) + \sum_{i=1}^{q} \beta_i \gamma(t - i) + \gamma(t)$, where $\gamma(t)$ is a white noise and $\theta = (\theta_1, \ldots, \theta_p)$, $\beta = (\beta_1, \ldots, \beta_q)$ are respectively the autoregressive and moving average parameters, which we consider as hyperparameters of the model. We used it to add smooth variations to the templates from one day to another.

In Section 4.3, we defined two kinds of devices with complex activations: (i) single signature or (ii) multiple signatures. While the former has just been addressed, we need to find a generative process for the latter. The proposed generative method uses the same process as before and considers a random convex combination of the activations:

$$\forall \tau, \forall t \in \mathcal{S}_\tau \quad \mathbf{a}^{new}(k, t) = \hat{\mathbf{a}}(\tau) \times \exp(\epsilon(t)) \times \delta(k), \qquad (23)$$

Table 3
Devices used to simulate the buildings in the SHED dataset: On/Off (A), Multi-state (B), Varying load (C), Varying signature (D).

Class	A	B	C	D	Total
Building 1	4	0	2	3	9
Building 2	1	4	2	3	10
Building 3	0	2	2	3	7
Building 4	2	0	4	3	9
Building 5	0	3	4	1	8
Building 6	3	0	3	4	10
Building 7	0	0	3	2	5
Building 8	0	0	4	4	8

where δ is a K-dimensional Dirichlet-distributed random variable whose parameter $\alpha = (\alpha_1, \ldots, \alpha_K)$ controls the activation components proportion and α is considered as an hyperparameter. Note that $\sum_k \delta(k) = 1$ and $\delta(k) \geq 0$ for all k.

6. A synthetic high-frequency energy disaggregation (SHED) dataset for commercial buildings

In order to enable high frequency NILM algorithm evaluation, we release a synthetic dataset called SHED. The purpose of our simulations is to evaluate the disaggregation performance of NILM algorithms (i.e. the capability to separate individual consumptions from a mixture). Our simulation procedure does not allow the evaluation of classification performance (assigning every disaggregated load curve to a particular category).

The SHED dataset consists of 8 buildings. For each building, it includes the total current consumption, as well as the individual consumptions corresponding to different categories. For buildings 1 to 6, the individual consumptions consist of low frequency power measurements and for buildings 7 and 8 they consist of high frequency current measurements. One current waveform is recorded at every 30 s and for every current waveform 200 points are sampled. Power measurements are also sampled at 1/30Hz. The choice of the classes of the devices and the number of categories enables us to control the complexity of each building: the details of the buildings are described in Table 3. The features of the buildings have been selected in such a way that they would correspond to commercial buildings. The simulation parameters used in this section have been learned on 3 public datasets (PLAID [23], COOLL [24] and Tracebase [37]) and one private dataset.

After having evaluated the quality of the device model in Section 4.5, we evaluate here the total current of the building. We use the metrics introduced in Section 3 to check the quality of the simulations. Fig. 9 shows clearly that the simulated datasets share very similar statistical properties as real commercial datasets. It provides a strong justification that our simulations are realistic. We can however notice that the THD values of simulations are more spread than for commercial buildings. It may be explained by the fact that the public datasets used for simulating signatures mostly correspond to residential equipments.

Finally, the SHED dataset can be downloaded at https://nilm.telecom-paristech.fr/shed/.

7. Conclusion and discussion

We addressed the task of non-intrusive load monitoring for commercial buildings by analyzing existing datasets, developing a synthetic data generation process, and releasing an evaluation dataset.

We produced an extensive data analysis on public and private datasets that showed that commercial and residential buildings have significantly different characteristics. The study of the power

derivative distribution illustrated that the residential distributions are more peaky at zero than the commercial ones. On top of this, we showed that the kurtosis, the entropy and the Laplace scale parameter of the power derivative are good discriminative indicators for residential and commercial buildings. We explained this difference by a higher amount of devices in commercial buildings and the presence of complex categories of devices (continuously varying equipment, multitude of similar 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 for the power variations at every instant) would be more realistic.

Motivated by the lack of data for commercial buildings, we developed a generative model for synthesizing high frequency current waveforms. Inspired by physical realities, it is compound of three layers: devices, categories and buildings. Our device model is based on a matrix factorization approach, breaking down high frequency current waveforms into signatures and activations components. The model efficiency has been validated with real data. It also enabled us to introduce a new device taxonomy taking the high frequency features of the devices into account.

Finally, we proposed a simulation procedure that enables us to learn parameters on real data and then simulate new synthetic data. Our quantitative evaluation experiments showed that the simulated datasets share the same statistical properties as real datasets. To enable algorithms testing and comparison, a simulated dataset called SHED is released at https://nilm.telecom-paristech.fr/shed/.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.enbuild.2018.07.060.

References

[1] P. Protocol, Report of the conference of the parties, United Nations Framework Convention on Climate Change (UNFCCC), 2015.
[2] N. Batra, O. Parson, M. Berges, A. Singh, A. Rogers, A comparison of non-intrusive load monitoring methods for commercial and residential buildings, arXiv:1408.6595(2014).
[3] M. Quintana, H. Lange, M. Bergés, Demo: Design and implementation of a low–cost arduino-based high-frequency ac waveform meter board for the raspberry pi, in: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments, ACM, 2017, pp. 249–250.
[4] G.W. Hart, Nonintrusive appliance load monitoring, in: Proceedings of the IEEE, 1992.
[5] F. Sultanem, Using appliance signatures for monitoring residential loads at meter panel level, IEEE Trans. Power Delivery 6 (4) (1991) 1380–1385.
[6] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, P. Armstrong, Power signature analysis, IEEE Power Energy Mag. 99 (2) (2003) 56–63.
[7] A.A. Girmay, C. Camarda, Simple event detection and disaggregation approach for residential energy estimation, in: Proceedings of the 3rd International Workshop on Non-Intrusive Load Monitoring, 2016.
[8] K.S. Barsim, B. Yang, Toward a semi-supervised non-intrusive load monitoring system for event-based energy disaggregation, in: Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on, IEEE, 2015, pp. 58–62.
[9] J.Z. Kolter, T. Jaakkola, Approximate inference in additive factorial hmms with application to energy disaggregation, in: Artificial Intelligence and Statistics, 2012, pp. 1472–1482.
[10] J.Z. Kolter, S. Batra, A.Y. Ng, Energy disaggregation via discriminative sparse coding, in: Advances in Neural Information Processing Systems, 2010, pp. 1153–1161.
[11] A. Rahimpour, H. Qi, D. Fugate, T. Kuruganti, Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint, IEEE Trans. Power Syst. (2017).
[12] H. Lange, M. Bergés, Bolt: Energy disaggregation by online binary matrix factorization of current waveforms, in: Proceedings of the 3rd ACM International

Conference on Systems for Energy-Efficient Built Environments, ACM, 2016, pp. 11–20.

[13] L.K. Norford, S.B. Leeb, Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms, Energy Build. 24 (1) (1996) 51–64.

[14] K.D. Lee, S.B. Leeb, L.K. Norford, P.R. Armstrong, J. Holloway, S.R. Shaw, Estimation of variable-speed-drive power consumption from harmonic content, IEEE Trans. Energy Convers. 20 (3) (2005) 566–574.

[15] M. Zeifman, K. Roth, Nonintrusive appliance load monitoring: review and outlook, IEEE Trans. Consum. Electron. 57 (1) (2011) 76–84, doi:10.1109/TCE.2011.5735484.

[16] R. Dong, L. Ratliff, H. Ohlsson, S.S. Sastry, Fundamental limits of nonintrusive load monitoring, in: Proceedings of the 3rd International Conference on High Confidence Networked Systems, ACM, 2014, pp. 11–18.

[17] D. Fischer, A. Härtl, B. Wille-Haussmann, Model for electric load profiles with high time resolution for german households, Energy Build. 92 (2015) 170–179.

[18] N. Buneeva, A. Reinhardt, Ambal: realistic load signature generation for load disaggregation performance evaluation, in: Smart Grid Communications (SmartGridComm), 2017 IEEE International Conference on. IEEE, 2017, pp. 1–9.

[19] S. Barker, S. Kalra, D. Irwin, P. Shenoy, Empirical characterization and modeling of electrical loads in smart homes, in: Green Computing Conference (IGCC), 2013 International, IEEE, 2013, pp. 1–10.

[20] W. Lee, G. Fung, H. Lam, F. Chan, M. Lucente, Exploration on load signatures, in: International Conference on Electrical Engineering (ICEE), vol. 152, 2004.

[21] J. Liang, S.K. Ng, G. Kendall, J.W. Cheng, Load signature studypart I: basic concept, structure, and methodology, IEEE Trans. Power Delivery 25 (2) (2010) 551–560.

[22] H.Y. Lam, G.S.K. Fung, W.K. Lee, A novel method to construct taxonomy electrical appliances based on load signatures of, IEEE Trans. Consum. Electron. 53 (2) (2007) 653–660, doi:10.1109/TCE.2007.381742.

[23] J. Gao, S. Giri, E.C. Kara, M. Bergés, Plaid: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract, in: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, ACM, 2014, pp. 198–199.

[24] T. Picon, M.N. Meziane, P. Ravier, G. Lamarque, C. Novello, J.-C.L. Bunetel, Y. Raingeaud, Cooll: Controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification, arXiv:1611.05803.

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

[26] J. Kelly, W. Knottenbelt, The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes, Sci. Data 2 (150007) (2015), doi:10.1038/sdata.2015.7.

[27] J.Z. Kolter, 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.

[28] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, 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.

[29] N. Batra, M. Gulati, A. Singh, 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.

[30] D. Murray, L. Stankovic, V. Stankovic, An electrical load measurements dataset of united kingdom households from a two-year longitudinal study, Sci. Data 4 (2017) 160122.

[31] S. Makonin, Z.J. Wang, C. Tumpach, Rae: the rainforest automation energy dataset for smart grid meter data analysis, arXiv:1705.05767

[32] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open Source Scientific Tools for Python, 2001. [Online; accessed today] URL http://www.scipy.org/.

[33] Z. Liang, J. Wei, J. Zhao, H. Liu, B. Li, J. Shen, C. Zheng, The statistical meaning of kurtosis and its new application to identification of persons based on seismic signals, Sensors 8 (8) (2008) 5106–5119.

[34] S. Kotz, T. Kozubowski, K. Podgorski, The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance, Springer Science & Business Media, 2012.

[35] C. Klemenjak, P. Goldsborough, Non-intrusive load monitoring: A review and outlook, arXiv:1610.01191.

[36] C.H. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 45–55.

[37] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, R. Steinmetz, On the accuracy of appliance identification based on distributed load metering data, in: Sustainable Internet and ICT for Sustainability (SustainIT), 2012, IEEE, 2012, pp. 1–9.

[38] C. Dinesh, S. Makanin, I.V. Bajic, Incorporating time-of-day usage patterns into non-intrusive load monitoring. URL http://sigport.org/2303.

[39] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2015.