Poster: Synthetic Dataset Generation for Non-Intrusive Load Monitoring in Commercial Buildings

Simon Henriet*
LTCI, Télécom Paristech,
Université Paris-Saclay, France

Umut Simsekli, Gaël Richard
LTCI, Télécom Paristech,
Université Paris-Saclay, France

Benoit Fuentes
Smart Impulse, France

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. One of the main difficulties is that 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 order to circumvent the issues caused by the lack of data available, we propose a model for generating realistic synthetic current waveforms by making use of both publicly available datasets and our private dataset that is collected from real commercial buildings. Our primarily experiments show that the generated data ressemble real datasets.

ACM Reference Format:
Simon Henriet, Umut Simsekli, Gaël Richard, and Benoit Fuentes. 2017. Poster: Synthetic Dataset Generation for Non-Intrusive Load Monitoring in Commercial Buildings. In *Proceedings of 4th ACM Conference on Systems for Energy-Efficient Built Environments (BuildSys'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3137133.3141458

1 INTRODUCTION

In the context of electrical load monitoring, non-intrusive load monitoring (NILM) involves the installation of only one sensor at the entrance of the electrical network and requires an accurate disaggregation algorithm for estimating the consumption of each individual device connected to the network. The majority of the current NILM literature is dedicated to residential buildings for releasing datasets and developping various disaggregation algorithms.

Recently, there have been increasing academic and industrial interests in applying NILM to commercial buildings [1]. These buildings include large offices, warehouses, retails and shopping malls. As pointed out in [1], they have fundamentally different characteristics than those of residential buildings: (i) the number of devices is much higher in commercial buildings, (ii) as opposed to residential buildings, they often contain several ‘continuously varying' devices (*e.g.* air handling units, heating pumps, inverters) whose power consumption is hard to be monitored since their behavior can vary along time, (iii) the multiplicity of devices belonging to the same category in commercial buildings (such as computers, light bulbs) can be much higher.

Due to these significant dissimilarities 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. As a result, residential NILM algorithms often fail when applied to commercial buildings. Therefore, accurate disaggregation algorithms that are tailored for commercial buildings are yet to be developed.

Apart from those existing difficulties, another important limitation for developing disaggregation algorithms for commercial buildings is the lack of publicly available datasets that contain detailed measurements of individual devices. 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. Also, 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 [1]. This dataset contains the power consumption measurements of two buildings 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 at a sampling rate > 10 kHz) and the equipments are not fully sub-metered.

In this study, we aim at circumventing the issues caused by the lack of data available in commercial buildings. We first performed statistical analysis on public residential datasets and compared them to a private dataset that was collected from real commercial buildings. Due to a lack of space, this analysis is not included in this abstract. In the light of the analysis results obtained and by making use of both publicly available datasets and our private dataset, we develop a synthetic data generation algorithm that is able to produce realistic current waveforms. Our primarily experiments show that the data generated by our algorithm resemble the data collected from commercial buildings in a statistical point of view.

2 A GENERATIVE MODEL FOR HIGH FREQUENCY CURRENT WAVEFORMS

In this section, we develop a physically-inspired data model for high frequency current and voltage that has three layers: building, category and device. Let us first introduce some notations and recall the relation between physical quantities. The digitalized voltage and current waveforms are denoted as $\mathbf{u}(n, t)$ 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

[*] Contact: simon.henriet@telecom-paristech.fr

BuildSys'17, November 8–9, 2017, Delft, The Netherlands
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5544-5/17/11.
https://doi.org/10.1145/3137133.3141458

period. The number of samples within a period of the voltage signal is supposed to be constant and is denoted by N. The mean active power (or mean power consumption or load curve) within a voltage period is then given by: $\mathbf{p}(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(n,t)\,\mathbf{i}(n,t)$.

2.1 The building model

The model put forward in this section relies on several hypotheses. First, all electrical devices are supposed to be plugged in parallel on the network: the current waveforms observed on the root node of the network are then the sum of the current waveforms of each device. The electrical network is also supposed to be in ideal conditions: wires have neither electrical resistivity nor inductance. The voltage can thus be considered as 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 stationary 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, whereas three-phase networks can be simulated in a similar fashion.

It leads us to the following model for total current:

$$\mathbf{i}(n,t) = \sum_{c \in \mathrm{C}} \mathbf{i}_c(n,t) + \epsilon(n,t) \tag{1}$$

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

2.2 The category model

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

$$\mathbf{i}_c(n,t) = \sum_{d \in \mathrm{D}_c} \mathbf{i}_{c,d}(n,t) \tag{2}$$

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. Device d can be seen as an instance of a particular equipment, but it can also be used as an artificial equipment to ease the modeling of complex devices such as "multi-state" or "continuously varying".

2.3 The device model

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

$$\mathbf{i}_{c,d}(n,t) = \mathrm{s}_c(n,d)\,\mathrm{a}_c(t,d)\,, \tag{3}$$

where $\mathrm{s}_c(\cdot,d)$ and $\mathrm{a}_c(\cdot,d) \geq 0$ are called respectively the *current waveform signature* and the *activations* of device d of category c. We use the notation $\mathrm{s}_c(\cdot,d)$ for denoting an entire column of matrix s_c. The waveform signature corresponds to a fixed pattern that describes the current response to the voltage. The activation is a nonnegative magnitude. Its nature can depend on the type of devices: a 0 / 1 function for on / off devices, a piecewise constant function for multi-state devices or just any positive function for continuously varying devices.

2.4 Overall generative process

Combining the individual models (1), (2) and (3) gives us the model for the overall current, given as follows:

$$\mathbf{i}(n,t) = \sum_{c \in \mathrm{C}} \sum_{d \in \mathrm{D}_c} \mathrm{s}_c(n,d)\,\mathrm{a}_c(t,d) + \epsilon(n,t)\,. \tag{4}$$

We obtain the following formula for the mean power per category:

$$\mathbf{p}_c(t) = \sum_{d \in \mathrm{D}_c} \mathrm{a}_c(t,d)\,\frac{1}{N}\sum_{n} \mathbf{u}_0(n)\,\mathrm{s}_c(n,d)\,. \tag{5}$$

This model presents several advantages: (i) it enables the use of on/off, multi-state or continuously varying devices, (ii) it handles categories of numerous similar equipments like computers or lamps, (iii) it makes it possible to model continuously varying category as the sum of simple sub-devices (which may not have a physical meaning but a modelling purpose), (iv) it can model data that are statistically close to residential or commercial buildings.

3 SIMULATION PROCEDURE

The main idea behind our simulation algorithm is to learn the signatures from public datasets, to learn activation templates from our private dataset and finally simulate from the generative model in order to obtain a synthetic dataset.

The signatures are sampled from two public datasets of high frequency current measurements [2, 3]. The process consists of: (i) segmenting the data according to the voltage period (period index times sampling index within a period), (ii) normalizing (so that we have unitary mean active power), (iii) selecting randomly one current period to account for the device's signature.

The activations templates are learned on a private dataset collected from two large commercial buildings. Its goal is to capture the typical power consumption of a device category during a day. In order to compute such templates, we averaged the power consumptions of categories over several weeks of data. Since many equipments are programmed to switch on or off at particular moments of the day (air handling unit, heaters) or depend on building occupancy (computers), we distinguished the week days and the off days. To take inter day variability into account, a positive noise is added to the concatenated templates.

Our primarily evaluations, based on statistical distribution (kurtosis, entropy) show that simulated datasets ressemble real ones (this will be demonstrated at the conference). Several simulations can be found at https://perso.telecom-paristech.fr/shenriet/simu/.

REFERENCES

[1] Nipun Batra, Oliver Parson, Mario Berges, Amarjeet Singh, and Alex Rogers. 2014. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. arXiv preprint arXiv:1408.6595 (2014).

[2] Jingkun Gao, Suman Giri, Emre Can Kara, and Mario Bergés. 2014. 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, 198–199.

[3] 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. arXiv preprint arXiv:1611.05803 (2016).